Statistical Game Theory

Arun Sai Suggala

August 2021
CMU-ML-21-109

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee

Pradeep Ravikumar, Chair	*Carnegie Mellon University*
Tuomas Sandholm	*Carnegie Mellon University*
Larry Wasserman	*Carnegie Mellon University*
Robert Schapire	*Microsoft Research*

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy.

This research was supported by: the Air Force Research Laboratory award FA87501720152, the National Institutes of Health award R01GM117594, the National Science Foundation awards IIS1661755 and IIS1909816; and the U.S. Army Research Office contract W911NF1210390.

Keywords: statistics, machine learning, game theory, online learning, bandits, minimax estimation, boosting

Dedicated to my parents who supported me through out my life and helped me pursue my dreams. ♡

Abstract

Game theory and statistics are two huge scientific disciplines that have played a significant role in the development of a wide variety of fields, including computer science, natural sciences, and social sciences. Traditionally, game theory has been used for decision making in strategic environments where multiple agents interact with each other. Statistics, on the other hand, is traditionally used for reasoning in non-adversarial settings where the samples are assumed to be generated by some stationary non-reactive source. Due to the contrasting settings in which game theory and statistics are often studied, these two disciplines have traditionally been regarded as disparate research areas. However, there is a great degree of commonality between the two fields. A surprisingly wide range of problems in classical and modern statistics have a game theoretic component to them. Classically, the mathematical philosophy of statistics, particularly frequentist statistics, posits that the source of samples is potentially adversarial. This resulted in the rich theory of minimax statistical games and estimation. Boosting algorithms, which are often regarded as best off-the-shelf classifiers, can be viewed as playing a zero-sum game against a weak learner. To allow for various departures of "test environment" from "train environments", the emerging field of robust machine learning allows for adversarial manipulation of the train or test environments. Finally, an emerging class of density estimators in modern machine learning use an adversarial "critic" of the density estimator to improve the final density estimation. The common theme among these classical and modern developments is an interplay between statistical estimation and multiplayer games.

Statistical game theory is a unified analytical and algorithmic framework underlying all these classical and modern developments. This thesis aims to lay the foundations of statistical game theory to address the above-mentioned (and many more) statistical problems. While our primary focus in this thesis is on minimax statistical estimation and boosting, the tools and techniques developed here are broadly applicable and are useful for studying other problems such as robust learning, and adversarial density estimation.

Our work on minimax statistical estimation aims to provide efficient techniques for algorithmically building minimax optimal estimators. These techniques automate the process of designing minimax estimators and can aid statisticians in building these estimators. For various fundamental problems such as mean estimation, and entropy estimation, our algorithmic minimax estimators match, if not beat, the performance of existing minimax estimators designed by statisticians. Our work on boosting aims to improve its performance and bring it closer to neural networks' performance. To this end, we develop a generalized boosting framework that combines weak classifiers using more complex forms of aggregation than additive combinations considered in traditional boosting. Our generalized boosting algorithms have better performance than traditional boosting and have performance close to neural networks.

Acknowledgments

I am incredibly grateful to my advisor Pradeep Ravikumar for his unconditional support over the years, and for giving me the freedom to pursue a variety of topics during my Ph.D. This thesis would not have been possible without the close collaboration I have had with him over the years. Pradeep was very patient with me during the early days of my Ph.D., and taught me many valuable lessons on how to be a good researcher. He encouraged me to think big, develop a broad vision, and pursue impactful research directions. My interactions with him have profoundly shaped this work and more generally, the way I do research.

I would like to thank Tuomas Sandholm, Larry Wasserman, and Robert Schapire for serving on my dissertation committee. Their insightful questions and valuable feedback during my proposal and defense have greatly helped improve my work.

The foundations of this thesis were laid during my time at Microsoft Research in the summer of 2018, where I had the opportunity to work with Praneeth Netrapalli and Prateek Jain. Praneeth and Prateek are incredibly smart people and are really good mentors who bring the best out of me. I enjoyed working with them, and over time, we developed strong connections and collaborated on numerous research projects.

This thesis wouldn't have been possible without my close collaborators who have contributed to the projects I pursued during my Ph.D., including Kush Bhatia, Kartik Gupta, David I Inouye, Mladen Kolar, Bingbin Liu, Adarsh Prasad, Eunho Yang, Chih-Kuan Yeh. I would like to give special mention to Adarsh Prasad for the numerous hours we have spent discussing various research problems and having non-research conversations.

I would also like to thank my wonderful friends and colleagues at CMU who made by graduate school both enjoyable and memorable. This includes Sushma Akoju, Alnur Ali, Ifigeneia Apostolopoulou, Shaojie Bai, Jonathon Byrd, Sebastian Caldas, Jeremiah Cohen, Amanda Coston, Christoph Dann, Avinava Dubey, Dhivya Eswaran, Ben Eysenbach, Chirag Gupta, Juyong Kim, Paul Liang, Leqi Liu, Vaishnavh Nagarajan, Ojash Neopane, Eyan Noronha, Biswajit Paria, Emilio Parisotto, Pratik Patil, Anthony Platanios, Charvi Rastogi, Elan Rosenfeld, Mrinmaya Sachan, Otilia Stretcu, Mariya Toneva, Ian En-Hsu Yen. I would like to thank CMU for providing me with a vibrant and collaborative research environment to pursue my graduate studies. Finally, I would like to thank Diane Stidle for help with administrative work throughout my time at CMU.

Last but certainly not the least, I would like to thank all of my family members for supporting me and providing me with immense amount of love throughout my life. I would like to especially thank my parents, who were also my mentors and teachers during my childhood, for providing me with good education. It is them who identified my interest in mathematics and encouraged me to pursue a career in computer science and engineering. Next, I would like thank my sister Anusha, and my cousins Anil, Asha, Pavan, Prasad, for bringing so much joy to my life. Finally, I would like to thank my amazing wife Mounica for her unwavering moral support during the final few years of my PhD. She has immense patience to listen to my petty problems and helped me stay sane during stressful times. Whether I'm happy or sad, angry or anxious, she is my go to person.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Game theory and statistics are two huge scientific disciplines that have in turn played a significant role in the development of a wide variety of fields, including computer science, natural sciences, and social sciences. Traditionally, game theory has been used for decision-making in strategic environments where multiple agents interact with each other. For example, in economics, it is often used for designing auctions and for decision-making in competitive markets. In computer science, it has found applications in numerous sub-fields such as distributed computing, network security, robotics, self-driving cars, and in general where multiple self-interested parties interact with each other.

Unlike game theory, statistics has traditionally been used for reasoning in non-strategic and non-adversarial environments. In particular, statistics is concerned with the analysis and interpretation of data generated by some stationary non-reactive source. For example, in numerous fields such as astronomy, biostatistics, business analytics, epidemiology, finance, statistical analysis, and estimation is often performed on data generated from non-reactive sources. Due to the contrasting settings in which game theory and statistics are often studied, these two disciplines have traditionally been regarded as disparate research areas. However, there is a great degree of commonality between the two fields. A surprising range of developments in classical and modern statistics have a game theoretic component to them:

- **Classical Developments.** Classically, the mathematical philosophy of statistics, in particular frequentist statistics, was concerned about strategic considerations. It posits that the source of samples seen by the statistician is potentially adversarial. This resulted in the rich theory of minimax statistical estimation and games [Wal49]. In these games, statistical estimation problems are framed as two-player games in which nature adversarially selects a distribution that makes it difficult for a statistician to perform the estimation. Boosting algorithms, which are often regarded as best off-the-shelf classifiers, can be viewed as playing a zero-sum game against a weak learner [FS96].
- **Modern Developments.** Modern statistical and machine learning applications are increasingly moving towards multi-agent learning as illustrated by the following examples. To allow for various departures of "test environment" from "train environments", the emerging field of robust machine learning allows for adversarial manipulation of the train

or test environments [Pra+20; Sze+13]. An emerging class of density estimators in modern machine learning use an adversarial "critic" of the density estimator to improve the final density estimation [Goo+14]. Finally, approaches for algorithmic fairness [Has+18; DN18], uncertainty quantification (*e.g.,* calibration, prediction intervals) [Gup+21], can be framed as finding the equilibrium of two-player games.

The common theme among these classical and modern developments is an interplay between statistical estimation and two-player games. Moving beyond two-player games, several emerging problems in statistics and machine learning naturally lead to multi-player games. Due to various privacy concerns, data used in many modern statistical applications in healthcare and advertising is often collected in a decentralized manner by multiple local actors, each with their own self-interests. Statistical inference in such scenarios naturally leads to an interplay with multi-player game theory. All these examples show that the intersection of statistics and game theory is becoming an increasingly relevant sub-field.

In this thesis, we aim to bring together statistics and game theory and study the interplay between the two fields. In particular, we are interested in studying statistical problems from a game theoretic perspective and understand how game theory can advance statistics. Despite the many commonalities between the two fields, the game theoretic perspective of many statistical problems is often ignored due to various analytical and computational reasons:

- **Large Domains.** One of the unpleasant facts about many games arising in statistics is that they are generally much too big and too difficult to solve than those typically arising in economics and computer science. For example, consider the problem of minimax statistical estimation, which can be viewed as a game between statistician and nature. The action space of the statistician in this game is the set of all functions, which is an infinite-dimensional space. Existing algorithmic tools from game theory are inefficient for solving this game. Consequently, statisticians have often ignored the game theoretic viewpoint while designing minimax estimators.
- **Nonconcave Utilities.** Another unpleasant fact about games arising in statistics and machine learning is that the utility functions of the players often turn out to be nonconcave. For example, this is the case in many modern statistical applications such as robust machine learning, Generative Adversarial Networks (GANs) that rely on deep neural networks. This is the case even in classical statistical problems such as minimax statistical estimation. Existing analytical and algorithmic tools from game theory, which primarily focus on concave utility functions, are inadequate for studying such games.

Setting aside these analytical and computational caveats, the game theoretic perspective provides tremendous value and comes with several benefits. It can help us reason about and construct optimal solutions for the wide range of statistical problems described above. As an example, consider again the problem of minimax statistical estimation. Existing approaches for designing minimax estimators often rely on prior knowledge and require a deeper understanding of the problem at hand. This process is very time-consuming, and often requires decades of research on the problem; for example, designing the popular LASSO estimator required decades of research on sparse estimation. In contrast, the game theoretic perspective can help us come up with algorithmic approaches that can automate the process of designing minimax estimators. Such algorithmic approaches can be of tremendous value to statisticians, as they can aid them in building minimax estimators. As another example, consider robust machine learning. Existing approaches for constructing robust models often rely on heuristics and are not guaranteed to return an optimal solution. In contrast, the game theoretic viewpoint of robust machine learning provides us a wide array of tools for constructing robust models that can withstand adversarial manipulations better than existing approaches.

The sub-field of statistical game theory provides an analytical and algorithmic framework for addressing the above issues and helps us study statistical problems from a game theoretic perspective. In this thesis, we aim to lay the foundations of statistical game theory and study some of the above-described statistical applications. Our primary focus in this thesis is on minimax statistical estimation and boosting. However, the tools and techniques developed here are broadly applicable and are useful in other areas such as contextual bandits, robust machine learning, and adversarial density estimation.

Here is the organization of the thesis. In Parts I and II, we develop necessary algorithmic tools in online learning, game theory, and optimization that help us study several statistical problems from a game theoretic perspective. In Part III of the thesis, we study the problem of minimax statistical estimation. Here, we utilize the tools in Part I to develop efficient techniques for algorithmically building minimax optimal estimators. In Part IV, we study the problem of boosting. Here, we develop new techniques to improve the performance of boosting and bring it closer to neural networks' performance.

1.1 Part I: Online Learning with Full Information Feedback

As previously mentioned, a major challenge in studying games that arise in statistical applications is that they come with nonconcave utility functions. For such games, there need not exist a Nash equilibrium (NE)[1], that is, there need not exist situations where all the players are satisfied with their actions. So, it is crucial to first understand the type of solutions we should target when studying these games. Several works have studied alternatives to Nash equilibrium in zero-sum games with nonconcave utilities. Two such popular solution concepts are local Nash equilibrium [JNJ20; DSZ21] and mixed strategy Nash equilibrium. Of these two concepts, mixed strategy NE is much more suitable for statistical applications. This is because, for problems such as minimax statistical estimation, local NE solutions can lead to highly sub-optimal estimators[2]. Consequently, in this thesis, we primarily focus on studying mixed strategy NE of games with nonconcave utility functions.

In Part I of the thesis, we present efficient algorithms for computing mixed strategy NE of games with nonconcave utility functions. A popular and widely used approach for

[1] Nash equilibrium is a very popular notion that is often used to analyze games and multi-agent systems.

[2] In certain applications such as GANs, adversarial training, it could be possible that local NE solutions might suffice.

computing NE of games is to rely on online learning algorithms [Haz16; CL06]. In this thesis, we take this approach and develop efficient algorithms for online nonconvex learning that achieve optimal regret. This in turn gives us efficient algorithms for solving games with nonconcave utility functions.

In Chapter 2, we show that the classical Follow-the-Perturbed-Leader (FTPL) algorithm is optimal for online learning with nonconvex losses, and is (oracle) efficient. In particular, we show that it achieves the optimal $O(T^{-1/2})$ regret even when the sequence of loss functions chosen by the adversary is nonconvex. In each iteration, the FTPL algorithm makes a single call to an offline optimization oracle to choose its next action. Given that offline optimization is well understood for a number of problems of interest [HP13], this entails an efficient algorithm for online nonconvex learning. We note that the result in this chapter is of independent interest and has several consequences beyond the setting of nonconvex-nonconcave games considered here. The most important of these is its applications to online learning in bandit settings and contextual bandits.

While the $O(T^{-1/2})$ regret guarantees achieved by FTPL is optimal, these guarantees are derived under the assumption that the sequence of loss functions encountered by the learner could be adversarial. However, when online learning is used in the context of two-player games, this assumption becomes invalid. So a natural question in this context is whether there exist algorithms that can achieve better regret guarantees when the sequence of loss functions is benign and predictable. We answer this question in the affirmative. In Chapter 3, we show that an optimistic variant of FTPL can achieve better regret guarantees when the sequence of losses is predictable. In the context of two-player games, we show that Optimistic FTPL (OFTPL) converges at a faster rate to a Nash equilibrium than vanilla FTPL.

1.2 Part II. Bandit Optimization

Studying games with nonconcave utility functions invariably leads us to the question of maximizing nonconcave functions. Unfortunately, this is a very hard problem (in fact, it is known to be NP-hard). In addition, in many statistical applications of interest, we are faced with two more challenges: (a) we only have zeroth-order access (a.k.a bandit feedback) to the functions we want to maximize, and (b) evaluating the function at any given point is computationally expensive (see Chapter 5 for details). So we ideally want derivative-free optimization techniques that satisfy the following desiderata

1. handle nonconcave objectives
2. require as few function evaluations as possible
3. scale well to high dimensional problems

Unfortunately, none of the existing derivative-free optimization techniques satisfy all these requirements. Gaussian Process Optimization [Sri+09], perhaps the most popular derivative-free optimization technique, doesn't scale well to high dimensional problems. Random walk based approaches such as simulated annealing [VA87] require too many function evaluations, thus making them inefficient even for low dimensional problems. So, our aim is to develop derivative-free order optimization techniques that satisfy the above desiderata.

In Chapter 4, we take a *small* step towards this goal by studying the above question for convex quadratic loss functions (albeit in the much harder adversarial setting). Surprisingly, there are no efficient derivative-free optimization techniques known even in this simple setting. In Chapter 4, we design a regularized bandit Newton method which achieves the optimal $\tilde{O}\left(T^{-1/2}\right)$ regret guarantee in this setting and is computationally efficient. In ongoing work, we are relying on the insights gained from studying quadratic losses to design efficient derivative-free optimization techniques for nonconvex losses.

1.3 Part III: Minimax Statistical Estimation

For decades, minimax statistical estimation has been crucial for the development of frequentist statistics, as it aids statisticians in picking estimators that work well even under the worst circumstances. Traditional approaches for solving this statistical game are usually problem-specific. In these approaches, an estimator is first proposed for a specific problem, and then its optimality is certified by showing its worst-case risk matches the known lower bounds for the minimax value of the game. However, such approaches can be time-consuming, require a deeper understanding of the problem, and do not often provide concrete guidelines for designing minimax optimal estimators for general problems. So algorithmic approaches that automate this process can be of immense help to statisticians.

In Chapter 5, we aim to develop algorithmic techniques for solving minimax statistical games. As previously described, a critical distinction of statistical games, in contrast to the typical zero-sum games studied in economics and computer science, is that the set of all possible moves of the statistician is extremely large, and importantly, the game need not have concave utility functions. To handle these technical caveats, we rely on algorithmic tools developed in Part I of the thesis.

Solving Minimax Statistical Games. A standard technique for computing a NE of the game relies on online learning algorithms. Here, the minimization player and the maximization player play a repeated game against each other, with both relying on online learning algorithms to choose their actions in each round of the game, and with the objective of minimizing their respective regret. Whenever the algorithms used by both the players guarantee vanishing regret, it can be shown that the repeated game play converges to a NE. Equipped with the FTPL, OFTPL algorithms developed in Chapters 2, 3, in Chapter 5, we rely on this technique to solve the statistical game. The resulting algorithm requires access to two subroutines: a Bayes estimator subroutine that outputs a Bayes estimator corresponding to any given prior, and a subroutine that computes the (perturbed) worst-case risk of any given estimator. Given access to these two subroutines, we show that our algorithm outputs both a minimax estimator and a least favorable prior (LFP). For problems where the two subroutines are efficiently implementable, our algorithm provides an efficient technique to construct minimax estimators. While implementing the subroutines can be computationally hard in general, we show that the computational complexity can be significantly reduced for a wide range of problems satisfying certain invariance properties.

To demonstrate the power of this technique, we use it to construct provably minimax estimators for the classical problems of finite-dimensional Gaussian sequence model and linear regression. Furthermore, for the fundamental problems of covariance and entropy estimation, we present empirical evidence showing that our algorithmically constructed estimators match the performance of existing minimax estimators designed by statisticians.

1.4 Part IV: Boosting

In the final part of the thesis, we focus on boosting. Boosting is a widely used learning technique in machine learning for solving classification problems. Over the years, boosting based methods have shown tremendous success in many real-world applications. Moreover, boosting based methods are easy to train and understand from a theoretical standpoint, thus making it easier to adopt these methods in critical applications such as healthcare. However, this success is mostly limited to classification tasks involving structured or tabular data with hand-engineered features. On classification problems involving low-level features and complex decision boundaries, boosting tends to perform poorly. One example where this is evident is the image classification task, where the decision boundaries are often complex and the features are low-level pixel intensities. This drawback stems from the fact that boosting builds an additive model of weak classifiers, each of which has very little predictive power. Since such additive models with any reasonable number of weak classifiers are usually not powerful enough to approximate complex decision boundaries, the models' output by boosting tend to have poor performance. This then brings us to the following question:

can we generalize boosting to allow for more complex forms of aggregation than linear combinations of weak classifiers?

Such a generalized boosting algorithm can have several benefits. For example, if we can develop boosting algorithms that combine weak classifiers through function compositions, it entails a simple and easy-to-understand algorithm for learning neural networks. Moreover, such an algorithm can make neural network training transparent and easy to adopt in critical applications.

The above question can be studied from two different perspectives: one based on the statistical view of boosting, where boosting is viewed as greedy stagewise optimization [FHT+00], and the other based on the game theoretic view, where boosting algorithms are viewed as playing a game against a weak learner [FS95]. In Chapter 6, we study the above question from the statistical viewpoint. In particular, we develop greedy stagewise optimization algorithms which allow for more complex forms of aggregation than additive combinations that are considered by traditional stagewise optimization techniques. Our algorithms improve upon traditional boosting and bridge the gap in performance between traditional boosting and neural networks.

The algorithms we developed in Chapter 6 don't yet match the performance of end-to-end trained neural networks. To truly bridge the gap in performance between boosting and neural networks, we hypothesize that one has to look at the game theoretic viewpoint of boosting. Historically, the game theoretic perspective has been much more successful in developing boosting algorithms with good generalization guarantees, than the statistical perspective. For example, consider the problem of multiclass boosting. Numerous boosting algorithms have been developed for this problem from the statistical perspective. However, many of these algorithms often perform poorly in practice. When viewed from a game theoretic perspective, many of these algorithms actually turn out to be sub-optimal [MS13]. Furthermore, the game theoretic viewpoint has played a crucial role in designing optimal algorithms for multiclass boosting. Consequently, in future, we aim to develop generalized boosting algorithms from a game theoretic perspective.

1.5 Summary of publications

The content of Chapter 2 appears in:

> [SN20b] Arun Sai Suggala and Praneeth Netrapalli. “Online Non-Convex Learning: Following the Perturbed Leader is Optimal”. In: ed. by Aryeh Kontorovich and Gergely Neu. Vol. 117. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, Aug. 2020, pp. 845–861. URL: http://proceedings.mlr.press/v117/suggala20a.html

The content of Chapter 3 appears in:

> [SN20a] Arun Sai Suggala and Praneeth Netrapalli. “Follow the Perturbed Leader: Optimism and Fast Parallel Algorithms for Smooth Minimax Games”. In: *Advances in Neural Information Processing Systems 33*. 2020. URL: https://arxiv.org/abs/2006.07541

The content of Chapter 4 appears in:

> [SRN21] Arun Sai Suggala, Pradeep Ravikumar, and Praneeth Netrapalli. “Efficient Bandit Convex Optimization: Beyond Linear Losses”. In: *Conference on Learning Theory*. 2021

The content of Chapter 5 appears in:

> [Gup+20] Kartik Gupta, Arun Sai Suggala, Adarsh Prasad, Praneeth Netrapalli, and Pradeep Ravikumar. “Learning Minimax Estimators via Online Learning”. In: *arXiv preprint arXiv:2006.11430* (2020)

The content of Chapter 6 appears in:

> [SLR20] Arun Sai Suggala, Bingbin Liu, and Pradeep Ravikumar. “Generalized Boosting”. In: *Advances in Neural Information Processing Systems 33*. 2020

Non-thesis research: I have also pursued the following research directions on robust machine learning during my Ph.D. These are excluded from the remainder of this thesis.

> [Sug+19a] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. “Adaptive Hard Thresholding for Near-optimal Consistent Robust Regression”. In: *Conference on Learning Theory*. 2019, pp. 2892–2897
>
> [Sug+19b] Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, and Pradeep Ravikumar. “Revisiting adversarial risk”. In: *The 22nd Interna-*

tional Conference on Artificial Intelligence and Statistics. 2019, pp. 2331–2339

Part I

Online Learning with Full Information Feedback

Chapter 2

Following the Perturbed Leader for Nonconvex Losses

In this chapter, we study the problem of online learning with non-convex losses, where, in each iteration, the learner chooses an action and observes a loss which could potentially be non-convex. The goal of the learner is to choose a sequence of actions which minimize the cumulative loss suffered over the course of learning. The paradigm of online learning has been studied in a number of fields, including game theory, machine learning, statistics and has several practical applications. In recent years a number of efficient algorithms have been developed for online learning. Convexity of the loss functions has played a central role in the development of many of these techniques. In this chapter, we consider a more general setting, where the sequence of loss functions encountered by the learner could be non-convex. Such a setting has numerous applications in machine learning, especially in adversarial training [Sze+13], robust optimization and training of Generative Adversarial Networks (GANs) [Goo+14].

As mentioned above, most of the existing works on online optimization have focused on convex loss functions [Haz16]. A number of computationally efficient approaches have been proposed for regret minimization in this setting. However, when the losses are non-convex, minimizing the regret is computationally hard. Recent works on learning with non-convex losses get over this computational barrier by either working with a restricted class of loss functions such as approximately convex losses [GLZ18] or by optimizing a computationally tractable notion of regret [HSZ17]. Consequently, the techniques studied in these papers do not guarantee vanishing regret for general non-convex losses. Another class of approaches consider general non-convex losses, but assume access to a sampling oracle [MM10; Kri+15] or an offline optimization oracle [AGH19]. Of these, assuming access to an offline optimization oracle is reasonable, given that in practice, simple heuristics such as stochastic gradient descent seem to be able to find approximate global optima reasonably fast even for complicated tasks such as training deep neural networks.

In a recent work Agarwal, Gonen, and Hazan [AGH19] take this later approach, where they assume access to an offline optimization oracle, and show that the classical Follow the Perturbed Leader (FTPL) algorithm achieves $O(T^{-1/3})$ regret for general non-convex losses which are Lipschitz continuous. In this chapter, we improve upon this result and show that FTPL in fact achieves optimal $O(T^{-1/2})$ regret.

2.1 Problem Setup and Main Results

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the set of all possible moves of the learner. In the online learning framework, on each round t, the learner makes a prediction $\mathbf{x}_t \in \mathcal{X}$ and the nature/adversary simultaneously chooses a loss function $f_t \colon \mathcal{X} \to \mathbb{R}$ and observe each others actions. The goal of the learner is to choose a sequence of actions $\{\mathbf{x}_t\}_{t=1}^T$ such that the following notion of regret is small

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \frac{1}{T}\inf_{\mathbf{x}\in\mathcal{X}}\sum_{t=1}^{T} f_t(\mathbf{x}).$$

We assume that $\mathcal{X}$ is bounded and has ℓ_∞ diameter of D, which is defined as

$$D = \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|_\infty.$$

Moreover, we assume that the sequence of loss functions f_t chosen by the adversary are L-Lipschitz with respect to ℓ_1 norm, that is, for all $\mathbf{x},\mathbf{y}\in\mathcal{X}$, $|f_t(\mathbf{x}) - f_t(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|_1$.

Approximate Optimization Oracle. Our results rely on an offline optimization oracle which takes as input a function $f \colon \mathcal{X} \to \mathbb{R}$ and a d-dimensional vector σ and returns an approximate minimizer of $\mathbf{x} \mapsto f(\mathbf{x}) - \langle\sigma,\mathbf{x}\rangle$. An optimization oracle is called "(α,β)-approximate optimization oracle" if it returns $\mathbf{x}^* \in \mathcal{X}$ such that

$$f(\mathbf{x}^*) - \langle\sigma,\mathbf{x}^*\rangle \leq \inf_{\mathbf{x}\in\mathcal{X}}\left[f(\mathbf{x}) - \langle\sigma,\mathbf{x}\rangle\right] + (\alpha + \beta\|\sigma\|_1),$$

We denote such an optimization oracle with $\mathcal{O}_{\alpha,\beta}\left(f - \langle\sigma,\cdot\rangle\right)$.

FTPL. Given access to an (α,β)-approximate offline optimization oracle, we study the FTPL algorithm which is described by the following prediction rule (see Algorithm 1).

$$\mathbf{x}_t = \mathcal{O}_{\alpha,\beta}\left(\sum_{i=1}^{t-1} f_i - \langle\sigma_t,\cdot\rangle\right), \tag{2.1}$$

where $\sigma_t \in \mathbb{R}^d$ is a random perturbation such that $\sigma_{t,j}$, the j^{th} coordiante of σ_t, is sampled from $\mathrm{Exp}(\eta)$, the exponential distribution with parameter η[1]. We note that one can also generate the perturbations from other probability distributions such as uniform distribution and achieve similar regret bounds as presented in this chapter.

[1] Recall, Z is an exponential random variable with parameter η if $P(Z \geq s) = \exp(-\eta s)$

Algorithm 1 Follow the Perturbed Leader (FTPL)

```
Input: Parameter of exponential distribution $\eta$, approximate optimization oracle $\mathcal{O}_{\alpha,\beta}$
for $t = 1 \ldots T$ do
    Generate random vector $\sigma_t$ such that $\{\sigma_{t,j}\}_{j=1}^{d} \overset{i.i.d}{\sim} \mathrm{Exp}(\eta)$
    Predict $\mathbf{x}_t$ as
        $\mathbf{x}_t = \mathcal{O}_{\alpha,\beta}\left(\sum_{i=1}^{t-1} f_i - \langle \sigma_t, \cdot \rangle\right).$
    Observe loss function $f_t$
end for
```

2.1.1 Main Result

We present our main result for an oblivious adversary who fixes the sequence of losses $\{f_t\}_{t=1}^{T}$ ahead of the game. Following [CL06], one can show that any algorithm that is guaranteed to work against an oblivious adversary also works for a non-oblivious adversary, whose actions are allowed to depend on the past predictions of the algorithm. For the sake of completeness, we present a proof of this reduction from non-oblivious to oblivious adversary model in Appendix A.2.

Theorem 1 (Non-Convex FTPL). *Let D be the ℓ_∞ diameter of $\mathcal{X}$. Suppose the losses encountered by the learner are L-Lipschitz w.r.t ℓ_1 norm. Moreover, suppose the optimization oracle used by Algorithm 1 is a "(α, β)-approximate" optimization oracle. For any fixed η, the predictions of Algorithm 1 satisfy the following regret bound*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \frac{1}{T}\inf_{\mathbf{x}\in\mathcal{X}}\sum_{t=1}^{T} f_t(\mathbf{x})\right] \leq O\left(\eta d^2 D L^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \beta d L\right).$$

The above result shows that for appropriate choice of η, FTPL achieves $O(d^{\frac{3}{2}}T^{-\frac{1}{2}} + \alpha + \beta d^{\frac{3}{2}}T^{\frac{1}{2}})$ regret. This also shows that when $\alpha = O(T^{-\frac{1}{2}}), \beta = O(T^{-1})$, FTPL achieves the optimal $O(T^{-\frac{1}{2}})$ regret. This improves upon the $O(T^{-\frac{1}{3}})$ regret bound obtained by Agarwal, Gonen, and Hazan [AGH19]. We note that the above result can be generalized to infinite-dimensional spaces such as ℓ^1 space of sequences. To do this we assume that the domain $\mathcal{X}$ is bounded and can be enclosed in a hyper-rectangle with edge length D_i along the i^{th} standard basis vector. Through a more careful analysis we can obtain regret bounds that depend on the *effective dimension* of $\mathcal{X}$, which is defined as $\frac{\sum_{i=1}^{d} D_i}{\max_i D_i}$, instead of d.

Before we conclude the section we point out that as an immediate consequence of the above regret bounds, we obtain algorithms for approximating the mixed strategy Nash equilibria of general non-convex non-concave saddle point problems of the form $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}, \mathbf{y})$. This follows from the observation that saddle point problems can be solved by playing two online optimization algorithms against each other [CL06; Haz16].

2.2 Background

In this section we briefly review the relevant literature on online learning in both convex and non-convex settings.

Online Convex Optimization. When the domain $\mathcal{X}$ and the loss functions f_t encountered by the learner are convex, a number of efficient algorithms for regret minimization have been studied. Most of these algorithms fall into three broad categories, namely Follow the Regularized Leader (FTRL), Online Mirror Descent (OMD) [Haz16] and Follow the Perturbed Leader (FTPL) [KV16]. FTRL algorithms make a prediction in each iteration by minimizing $\text{argmin}_{\mathbf{x}} \sum_{i=1}^{t-1} f_i(\mathbf{x}) + R(\mathbf{x})$, where R is a strongly convex regularizer. The regularization R plays a crucial role in the performance of the algorithm and helps avoid overfitting to the observed loss functions. Similar to FTRL, OMD also relies on explicit regularization to guarantee vanishing regret. In fact, under certain settings, both OMD and FTRL algorithms are known to be equivalent [McM11]. For a broad class of online convex optimization problems, FTRL and OMD are known to achieve optimal regret guarantees.

FTPL algorithms rely on random perturbation of loss functions to guarantee vanishing regret. This random perturbation can be viewed as having a similar role as the explicit regularization used in FTRL and OMD. In a recent work Abernethy, Lee, and Tewari [ALT16] use duality to connect FTPL and FTRL. They show that every instance of FTPL is also an instance of FTRL.

Online Non-Convex Optimization. A natural question that arises in the context of online non-convex learning is whether there exist counterparts of FTRL and OMD which achieve vanishing regret. Unfortunately, the answer is no. As we show in the following Proposition, there exists no deterministic algorithm that can achieve vanishing regret when the losses are non-convex.

Proposition 1. *No deterministic algorithm can achieve $o(1)$ regret in the setting of online non-convex learning.*

The above Proposition shows that only randomized algorithms can achieve vanishing regret. Recent works of Maillard and Munos [MM10] and Krichene, Balandat, Tomlin, and Bayen [Kri+15] consider the natural extension of Exponential Weight Algorithm to continuous domains and show that the resulting algorithm has vanishing regret in the setting of online non-convex learning. The algorithms studied in these works rely on an offline sampling oracle which can generate samples from any given probability distribution. In another line of work, Agarwal, Gonen, and Hazan [AGH19] study the classical FTPL algorithm with access to a certain offline optimization oracle and show that it achieves $O(T^{-1/3})$ regret. As an immediate consequence of this result, the authors show that both online adversarial learning model and statistical learning model are computationally equivalent.

2.3 Non-Convex FTPL

In this section, we present a proof of Theorem 1. Since we are in the oblivious adversary setting, it suffices to work with a single random vector σ, instead of generating a new random vector in each iteration. The first step in the proof involves relating the expected regret to the stability of prediction, which is a standard step in the analysis of many online learning algorithms.

Lemma 2. *The regret of Algorithm 1 can be upper bounded as*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \inf_{\mathbf{x}\in\mathcal{X}}\sum_{t=1}^{T} f_t(\mathbf{x})\right] \leq \frac{L}{T}\sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_1\right]}_{\textit{Stability}} + \frac{d(\beta T + D)}{\eta T} + \alpha. \tag{2.2}$$

In the rest of the proof we focus on bounding the stability term $\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_1\right]$. The randomness used in the algorithm is crucial for bounding its stability. The more randomness we add, the more stable the algorithm is. However, there is a price we pay for adding randomness. It causes the algorithm to make poor predictions, which leads to worse regret. This is evident in the second term in the upper bound in Equation (2.2), which increases as η decreases.

We first provide an brief sketch of the proof in the 1-dimensional case. Similar to the proof of Agarwal, Gonen, and Hazan [AGH19], our proof relies on showing certain monotonicity properties of the predictions of the algorithm. Letting $\mathbf{x}_t(\sigma)$ be the prediction in the t^{th} iteration of FTPL with random perturbation σ, we show that the predictions are monotonic functions of σ

$$\forall t, c > 0, \quad \mathbf{x}_t(\sigma + c) \geq \mathbf{x}_t(\sigma).$$

Moreover, we show that

$$\forall c > L, \quad \min\left\{\mathbf{x}_t(\sigma + c), \mathbf{x}_{t+1}(\sigma + c)\right\} \geq \max\left\{\mathbf{x}_t(\sigma), \mathbf{x}_{t+1}(\sigma)\right\}.$$

Since the domain is bounded, these two properties imply that the functions $\mathbf{x}_t(\sigma), \mathbf{x}_{t+1}(\sigma)$ should be close to each other for sufficiently large values of σ (see Figure 2.1 for an illustration). The closeness of these two functions immediately implies the stability of the algorithm. In what follows, we formalize this argument and extend it to the high-dimensional case.

Lemma 3 (Monotonicity 1)**.** *Let $\mathbf{x}_t(\sigma)$ be the prediction of FTPL in iteration t, with random perturbation σ. Let $\mathbf{e}_i$ denote the i^{th} standard basis vector and $\mathbf{x}_{t,i}$ denote the i^{th} coordinate of $\mathbf{x}_t$. Then the following monotonicity property holds for any $c > 0$*

$$\mathbf{x}_{t,i}(\sigma + c\mathbf{e}_i) \geq \mathbf{x}_{t,i}(\sigma) - \frac{2(\alpha + \beta\|\sigma\|_1)}{c} - \beta.$$

Proof. Let $f_{1:t}(\mathbf{x}) = \sum_{i=1}^{t} f_i(\mathbf{x})$ and $\sigma' = \sigma + c\mathbf{e}_i$. Moreover, let $\gamma(\sigma) = \alpha + \beta\|\sigma\|_1$ be the approximation error of the offline optimization oracle. From the approximate optimality

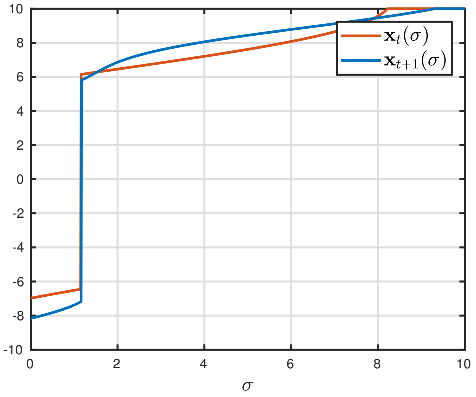


Figure 2.1: Illustration of monotonicity properties of the predictions of FTPL on a 1-dimensional example with $D = 10, L = 2$.

of $\mathbf{x}_t(\sigma)$ we have

$$\begin{aligned}
&f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma, \mathbf{x}_t(\sigma)\rangle \\
&\quad \leq f_{1:t-1}(\mathbf{x}_t(\sigma')) - \langle \sigma, \mathbf{x}_t(\sigma')\rangle + \gamma(\sigma) \\
&\quad = f_{1:t-1}(\mathbf{x}_t(\sigma')) - \langle \sigma', \mathbf{x}_t(\sigma')\rangle + c\mathbf{x}_{t,i}(\sigma') + \gamma(\sigma) \\
&\quad \overset{(a)}{\leq} f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma', \mathbf{x}_t(\sigma)\rangle + c\mathbf{x}_{t,i}(\sigma') + \gamma(\sigma) + \gamma(\sigma') \\
&\quad = f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma, \mathbf{x}_t(\sigma)\rangle + c\left(\mathbf{x}_{t,i}(\sigma') - \mathbf{x}_{t,i}(\sigma)\right) + \gamma(\sigma) + \gamma(\sigma'),
\end{aligned}$$

where (a) follows from the approximate optimality of $\mathbf{x}_t(\sigma')$. Combining the first and last terms in the above expression, we get $\mathbf{x}_{t,i}(\sigma') \geq \mathbf{x}_{t,i}(\sigma) - \frac{2\gamma(\sigma)}{c} - \beta$. □

Lemma 4 (Monotonicity 2)**.** *Let $\mathbf{x}_t(\sigma)$ be the prediction of FTPL in iteration t, with random perturbation σ. Let $\mathbf{e}_i$ denote the i^{th} standard basis vector and $\mathbf{x}_{t,i}$ denote the i^{th} coordinate of $\mathbf{x}_t$. Suppose $\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1 \leq 10d \cdot |\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|$. For $\sigma' = \sigma + 100Ld\mathbf{e}_i$, we have*

$$\begin{aligned}
\min\left(\mathbf{x}_{t,i}(\sigma'), \mathbf{x}_{t+1,i}(\sigma')\right) \geq\ &\max\left(\mathbf{x}_{t,i}(\sigma), \mathbf{x}_{t+1,i}(\sigma)\right) - \frac{1}{10}|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| \\
&- \frac{3(\alpha + \beta\|\sigma\|_1)}{100Ld} - \beta.
\end{aligned}$$

Proof. Let $f_{1:t}(\mathbf{x}) = \sum_{i=1}^t f_i(\mathbf{x})$ and let $\gamma(\sigma) = \alpha + \beta\|\sigma\|_1$ be the approximation error of the offline optimization oracle. From the approximate optimality of $\mathbf{x}_t(\sigma)$, we have

$$\begin{aligned}
&f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma, \mathbf{x}_t(\sigma)\rangle + f_t(\mathbf{x}_t(\sigma)) \\
&\quad \leq f_{1:t-1}(\mathbf{x}_{t+1}(\sigma)) - \langle \sigma, \mathbf{x}_{t+1}(\sigma)\rangle + f_t(\mathbf{x}_t(\sigma)) + \gamma(\sigma) \\
&\quad \overset{(a)}{\leq} f_{1:t-1}(\mathbf{x}_{t+1}(\sigma)) - \langle \sigma, \mathbf{x}_{t+1}(\sigma)\rangle + f_t(\mathbf{x}_{t+1}(\sigma)) + L\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1 + \gamma(\sigma) \\
&\quad \overset{(b)}{\leq} f_{1:t-1}(\mathbf{x}_{t+1}(\sigma)) - \langle \sigma, \mathbf{x}_{t+1}(\sigma)\rangle + f_t(\mathbf{x}_{t+1}(\sigma)) + 10Ld|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| + \gamma(\sigma),
\end{aligned}$$

where (a) follows from the Lipschitz property of $f_t(\cdot)$ and (b) follows from our assumption on $\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1$. Next, from the optimality of $\mathbf{x}_{t+1}(\sigma')$, we have

$$\begin{aligned}
&f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma, \mathbf{x}_t(\sigma)\rangle + f_t(\mathbf{x}_t(\sigma))\\
&\quad= f_{1:t-1}(\mathbf{x}_t(\sigma)) - \langle \sigma', \mathbf{x}_t(\sigma)\rangle + f_t(\mathbf{x}_t(\sigma)) + \langle 100Ld\mathbf{e}_i, \mathbf{x}_t(\sigma)\rangle\\
&\quad\geq f_{1:t-1}(\mathbf{x}_{t+1}(\sigma')) - \langle \sigma', \mathbf{x}_{t+1}(\sigma')\rangle + f_t(\mathbf{x}_{t+1}(\sigma')) + 100Ld\mathbf{x}_{t,i}(\sigma) - \gamma(\sigma')\\
&\quad= f_{1:t-1}(\mathbf{x}_{t+1}(\sigma')) - \langle \sigma, \mathbf{x}_{t+1}(\sigma')\rangle + f_t(\mathbf{x}_{t+1}(\sigma')) + 100Ld(\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma')) - \gamma(\sigma')\\
&\quad\geq f_{1:t-1}(\mathbf{x}_{t+1}(\sigma)) - \langle \sigma, \mathbf{x}_{t+1}(\sigma)\rangle + f_t(\mathbf{x}_{t+1}(\sigma)) + 100Ld(\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma')) - \gamma(\sigma') - \gamma(\sigma),
\end{aligned}$$

where the last inequality follows from the optimality of $\mathbf{x}_{t+1}(\sigma)$. Combining the above two equations, we get

$$\mathbf{x}_{t+1,i}(\sigma') - \mathbf{x}_{t,i}(\sigma) \geq -\frac{1}{10}|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| - \frac{3\gamma(\sigma)}{100Ld} - \beta.$$

A similar argument shows that

$$\mathbf{x}_{t,i}(\sigma') - \mathbf{x}_{t+1,i}(\sigma) \geq -\frac{1}{10}|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| - \frac{3\gamma(\sigma)}{100Ld} - \beta.$$

Finally, from the monotonicity property in Lemma 3 we know that

$$\mathbf{x}_{t+1,i}(\sigma') - \mathbf{x}_{t+1,i}(\sigma) \geq -\frac{3\gamma(\sigma)}{100Ld} - \beta, \quad \mathbf{x}_{t,i}(\sigma') - \mathbf{x}_{t,i}(\sigma) \geq -\frac{3\gamma(\sigma)}{100Ld} - \beta.$$

Combining the above four inequalities gives us the required result. □

Proof of Theorem 1. We now proceed to the proof of Theorem 1. We use the same notation as in Lemmas 3, 4. First note that $\mathbb{E}\left[\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1\right]$ can be written as

$$\mathbb{E}\left[\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1\right] = \sum_{i=1}^{d} \mathbb{E}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right]. \tag{2.3}$$

To bound $\mathbb{E}\left[\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1\right]$ we derive an upper bound for $\mathbb{E}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right], \forall i \in [d]$. For any $i \in [d]$, define $\mathbb{E}_{-i}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right]$ as

$$\mathbb{E}_{-i}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right] \coloneqq \mathbb{E}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| \middle| \{\sigma_j\}_{j\neq i}\right],$$

where σ_j is the j^{th} coordinate of σ. Let $\mathbf{x}_{max,i}(\sigma) = \max\left(\mathbf{x}_{t,i}(\sigma), \mathbf{x}_{t+1,i}(\sigma)\right)$ and $\mathbf{x}_{min,i}(\sigma) = \min\left(\mathbf{x}_{t,i}(\sigma), \mathbf{x}_{t+1,i}(\sigma)\right)$. Then $\mathbb{E}_{-i}\left[|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right] = \mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - \mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right]$. Define event $\mathcal{E}$ as

$$\mathcal{E} = \left\{\sigma : \|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1 \leq 10d \cdot |\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)|\right\}.$$

Consider the following

$$\begin{aligned}
\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right] &= \mathbb{P}(\sigma_i < 100Ld)\underbrace{\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)|\sigma_i < 100Ld\right]}_{T_1}\\
&\quad + \underbrace{\mathbb{P}(\sigma_i \geq 100Ld)\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)|\sigma_i \geq 100Ld\right]}_{T_2}.
\end{aligned} \tag{2.4}$$

We now try to lower bound T_1, T_2 in the above equation. Since the domain of i^{th} coordinate lies within some interval of length D and since T_1 and $\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right]$ are points in this interval, their difference is bounded by D. So T_1 is lower bounded by $\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - D$. We next rewrite T_2 as follows.

$$\begin{aligned} T_2 = \mathbb{P}(\sigma_i \geq 100Ld)\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)|\sigma_i \geq 100Ld\right] &= \int_{\sigma_i=100Ld}^{\infty} \mathbf{x}_{min,i}(\sigma)P(\sigma_i)d\sigma_i \\ &= \int_{\sigma_i=100Ld}^{\infty} \mathbf{x}_{min,i}(\sigma)\eta e^{-\eta\sigma_i}d\sigma_i \end{aligned}$$

We now do a *change of variables* in the above integration. Let $\sigma_i = \sigma'_i + 100Ld$ and $\sigma' = [\sigma_1, \dots \sigma_{i-1}, \sigma'_i, \sigma_{i+1}, \dots]$ be the vector obtained by replacing the i^{th} coordinate of σ with σ'_i. Rewriting the RHS in terms of σ'_i and σ', we get

$$\begin{aligned} \int_{\sigma_i=100Ld}^{\infty} \mathbf{x}_{min,i}(\sigma)\eta e^{-\eta\sigma_i}d\sigma_i &= \int_{\sigma'_i=0}^{\infty} \mathbf{x}_{min,i}(\sigma' + 100Lde_i)\eta e^{-\eta(\sigma'_i+100Ld)}d\sigma'_i \\ &= e^{-100\eta Ld}\int_{\sigma'_i=0}^{\infty} \mathbf{x}_{min,i}(\sigma' + 100Lde_i)\eta e^{-\eta\sigma'_i}d\sigma'_i \\ &= e^{-100\eta Ld}\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma' + 100Ld\mathbf{e}_i)\right]. \end{aligned}$$

This shows that $T_2 = e^{-100\eta Ld}\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma + 100Ld\mathbf{e}_i)\right]$. Substituting the lower bounds for T_1, T_2 in Equation (2.4), we get

$$\begin{aligned} \mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right] \quad &\geq \quad (1 - \exp(-100\eta Ld))\left(\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - D\right) \\ &\quad + \exp(-100\eta Ld)\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma + 100Ld\mathbf{e}_i)\right], \end{aligned}$$

We can further lower bound $\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right]$ as follows

$$\begin{aligned} \mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right] \quad &\geq \quad (1 - \exp(-100\eta Ld))\left(\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - D\right) \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E})\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma + 100Ld\mathbf{e}_i)|\mathcal{E}\right] \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E}^c)\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma + 100Ld\mathbf{e}_i)|\mathcal{E}^c\right], \end{aligned}$$

where $\mathbb{P}_{-i}(\mathcal{E})$ is defined as $\mathbb{P}_{-i}(\mathcal{E}) \coloneqq \mathbb{P}\left(\mathcal{E}\middle|\{\sigma_j\}_{j\neq i}\right)$. We now use the monotonicity properties proved in Lemmas 3, 4 to further lower bound $\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right]$. Let $\gamma(\sigma) = \alpha + \beta\|\sigma\|_1$ be the approximation error of the offline optimization oracle. Then

$$\begin{aligned} \mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right] \quad &\geq \quad (1 - \exp(-100\eta Ld))\left(\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - D\right) \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E})\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma) - \tfrac{1}{10}|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| - \tfrac{3\gamma(\sigma)}{100Ld} - \beta\middle|\mathcal{E}\right] \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E}^c)\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma) - \tfrac{2\gamma(\sigma)}{100Ld} - \beta|\mathcal{E}^c\right] \\ &\geq \quad (1 - \exp(-100\eta Ld))\left(\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right] - D\right) \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E})\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma) - \tfrac{1}{10}|\mathbf{x}_{t,i}(\sigma) - \mathbf{x}_{t+1,i}(\sigma)| - \tfrac{3\gamma(\sigma)}{100Ld} - \beta\middle|\mathcal{E}\right] \\ &\quad + \exp(-100\eta Ld)\mathbb{P}_{-i}(\mathcal{E}^c)\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma) - \tfrac{1}{10d}\|\mathbf{x}_t(\sigma) - \mathbf{x}_{t+1}(\sigma)\|_1 - \tfrac{2\gamma(\sigma)}{100Ld} - \beta\middle|\mathcal{E}^c\right], \end{aligned}$$

where the first inequality follows from Lemmas 3, 4, the second inequality follows from the definition of $\mathcal{E}^c$. Rearranging the terms in the RHS and using $\mathbb{P}_{-i}(\mathcal{E}) \leq 1$ gives us

$$\begin{aligned}
\mathbb{E}_{-i}\left[\mathbf{x}_{min,i}(\sigma)\right] \quad \geq \quad & (1-\exp(-100\eta Ld))\left(\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right]-D\right)\\
& +\exp(-100\eta Ld)\mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)-\tfrac{3\gamma(\sigma)}{100Ld}-\beta\right]\\
& -\exp(-100\eta Ld)\mathbb{E}_{-i}\left[\tfrac{1}{10}|\mathbf{x}_{t,i}(\sigma)-\mathbf{x}_{t+1,i}(\sigma)|+\tfrac{1}{10d}\|\mathbf{x}_t(\sigma)-\mathbf{x}_{t+1}(\sigma)\|_1\right]\\
\geq \quad & \mathbb{E}_{-i}\left[\mathbf{x}_{max,i}(\sigma)\right]-100\eta LdD-\tfrac{3\gamma(\sigma)}{100Ld}-\beta\\
& -\mathbb{E}_{-i}\left[\tfrac{1}{10}|\mathbf{x}_{t,i}(\sigma)-\mathbf{x}_{t+1,i}(\sigma)|+\tfrac{1}{10d}\|\mathbf{x}_t(\sigma)-\mathbf{x}_{t+1}(\sigma)\|_1\right],
\end{aligned}$$

where the last inequality uses the the fact that $\exp(x) \geq 1+x$. Rearranging the terms in the last inequality gives us

$$\begin{aligned}
\mathbb{E}_{-i}\left[|\mathbf{x}_{t,i}(\sigma)-\mathbf{x}_{t+1,i}(\sigma)|\right] \leq & \frac{1}{9d}\mathbb{E}_{-i}\left[\|\mathbf{x}_t(\sigma)-\mathbf{x}_{t+1}(\sigma)\|_1\right]\\
& +\frac{1000}{9}\eta LdD+\frac{\mathbb{E}_{-i}\left[\gamma(\sigma)\right]}{30Ld}+\frac{10}{9}\beta.
\end{aligned}$$

Since the above bound holds for any $\{\sigma_j\}_{j\neq i}$, we get the following bound on the unconditioned expectation

$$\begin{aligned}
\mathbb{E}\left[|\mathbf{x}_{t,i}(\sigma)-\mathbf{x}_{t+1,i}(\sigma)|\right] \leq & \frac{1}{9d}\mathbb{E}\left[\|\mathbf{x}_t(\sigma)-\mathbf{x}_{t+1}(\sigma)\|_1\right]\\
& +\frac{1000}{9}\eta LdD+\frac{\mathbb{E}\left[\gamma(\sigma)\right]}{30Ld}+\frac{10}{9}\beta.
\end{aligned}$$

Plugging this in Equation (2.3) gives us the following bound on stability of predictions of FTPL

$$\mathbb{E}\left[\|\mathbf{x}_t(\sigma)-\mathbf{x}_{t+1}(\sigma)\|_1\right] \leq 125\eta Ld^2D+\frac{\beta d}{20\eta L}+2\beta d+\frac{\alpha}{20L}. \tag{2.5}$$

Plugging the above bound in Equation (2.2) gives us the required bound on regret.

2.4 Discussion

In this chapter, we considered the problem of online learning with non-convex losses and showed that the classical FTPL algorithm with access to an offline optimization oracle achieves optimal regret rate of $O(T^{-1/2})$. The problem of online non-convex learning has several important applications in machine learning. In particular, the algorithms studied in this chapter can lead to improved training procedures for adversarial training and training of Generative Adversarial Networks, which currently rely on algorithms from online convex learning to solve the non-convex non-concave games in their training objectives. Moreover, we believe the algorithms in this chapter have applications to online learning in bandit setting, and contextual bandits. Both these problems often involve computing an unbiased estimate of the unknown loss function and reducing the problem to online learning in the full-information setting. Hedge is a popular algorithm that is typically used for the latter step [AHK12]. However, Hedge can be computationally expensive when the action space of the learner is huge. FTPL, on the other hand, can be efficiently implemented for a number of problems of interest, even when the action space is huge.

Chapter 3

Optimistic Follow the Perturbed Leader for Convex & Nonconvex Losses

In this chapter, we study optimistic variants of FTPL, which can achieve better regret guarantees than FTPL when the sequence of loss functions encountered by the learner is not adversarial. While the primary focus of this thesis is on nonconvex losses, in this chapter, we also study optimistic variants of FTPL for convex losses.

As mentioned in Chapter 2, the various algorithms that have been developed for online learning can be classified into two broad categories, namely, Follow the Regularized Leader (FTRL) [McM17] and FTPL [KV05] style algorithms. When the sequence of loss functions encountered by the learner are convex, both these algorithms are known to achieve the optimal $O\left(T^{1/2}\right)$ worst case regret [CL06; Haz16]. While these algorithms have similar regret guarantees, they differ in their computational aspects. Each iteration of FTRL involves optimization of a non-linear convex function over the action space (also called the projection step). In contrast, each step of FTPL involves solving a linear optimization problem, which can be implemented efficiently for many problems of interest [GH13; GJL16; HM20]. For example, if the action space is an ℓ_p ball for some $p \notin \{1, 2, \infty\}$, then projecting onto this set is much more computationally expensive than performing linear optimization over this set. As another example, consider the scenario where the action space is the set of all positive semidefinite matrices. Then projecting onto this set requires performing expensive singular value decompositions. Whereas, linear optimization only requires computation of the leading eigenvector. This crucial difference between FTRL and FTPL makes the latter algorithm more attractive in practice. Even in the more general nonconvex setting, where the loss functions encountered by the learner can potentially be nonconvex, FTPL algorithms are attractive. In this setting, FTPL requires access to an offline optimization oracle which computes the perturbed best response, and achieves $O\left(T^{1/2}\right)$ worst case regret. Furthermore, these optimization oracles can be efficiently implemented for many problems by leveraging the rich body of work on global optimization [HP13].

Despite the importance and popularity of FTPL, it has been mostly studied for the worst case setting, where the loss functions are assumed to be adversarially chosen. In a number of applications of online learning, the loss functions are actually benign and predictable [RS12]. In such scenarios, FTPL can not utilize the predictability of losses to achieve tighter regret bounds. While Rakhlin and Sridharan [RS12] study variants of FTPL which can make use of predictability, they consider restricted settings (see Section 3.1 for more details). This is unlike FTRL, where optimistic variants that can utilize the predictability of loss functions have been well understood [RS12; RS13] and have been shown to provide faster convergence rates in applications such as minimax games. In this chapter, we aim to bridge this gap and study a variant of FTPL called Optimistic FTPL (OFTPL), which can achieve better regret bounds, while retaining the optimal worst case regret guarantee for unpredictable sequences. The main challenge in obtaining these tighter regret bounds is handling the stochasticity and optimism in the algorithm, which requires different analysis techniques to those commonly used in the analysis of FTPL. In this chapter, we rely on the dual view of perturbation as regularization to derive regret bounds of OFTPL.

To demonstrate the usefulness of OFTPL, we consider the problem of solving minimax games. A widely used approach for solving such games relies on online learning algorithms [CL06]. In this approach, both the minimization and the maximization players play a repeated game against each other and rely on online learning algorithms to choose their actions in each round of the game. In our algorithm for solving games, we let both the players use OFTPL to choose their actions. For solving smooth convex-concave games, our algorithm only requires access to a linear optimization oracle. For Lipschitz and smooth nonconvex-nonconcave games, our algorithm requires access to an optimization oracle which computes the perturbed best response. In both these settings, our algorithm solves the game up to an accuracy of $O\left(T^{-1/2}\right)$ using T calls to the optimization oracle. While there are prior algorithms that achieve these convergence rates [HH15; SN20b], an important feature of our algorithm is that it is highly parallelizable and requires only $O(T^{1/2})$ iterations, with each iteration making $O\left(T^{1/2}\right)$ parallel calls to the optimization oracle. We note that such parallelizable algorithms are especially useful in large-scale machine learning applications such as training of GANs, adversarial training, which often involve huge datasets such as ImageNet [Rus+15].

3.1 Preliminaries and Background Material

In this chapter, we use a similar notation as in Chapter 2. In round t of online learning, the learner makes a prediction $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ for some compact set $\mathcal{X}$, and the adversary simultaneously chooses a loss function $f_t : \mathcal{X} \to \mathbb{R}$ and observe each others actions. The goal of the learner is to choose a sequence of actions $\{\mathbf{x}_t\}_{t=1}^T$ so that the following notion of regret is minimized: $\sum_{t=1}^T f_t(\mathbf{x}_t) - \inf_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})$.

Online Convex Learning. When the domain $\mathcal{X}$ and loss functions f_t are convex, a number of efficient algorithms for regret minimization have been studied. Some of these include deterministic algorithms such as Online Mirror Descent, Follow the Regularized Leader (FTRL) [Haz16; McM17], and stochastic algorithms such as Follow the Perturbed Leader (FTPL) [KV05]. In FTRL, one predicts $\mathbf{x}_t$ as $\mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{t-1} \langle \nabla_i, \mathbf{x}\rangle + R(\mathbf{x})$, for some strongly convex regularizer R, where $\nabla_i = \nabla f_i(\mathbf{x}_i)$. FTRL is known to achieve the optimal $O(T^{1/2})$ worst case regret in the convex setting [McM17]. In FTPL, one predicts $\mathbf{x}_t$ as $m^{-1}\sum_{j=1}^{m}\mathbf{x}_{t,j}$, where $\mathbf{x}_{t,j}$ is a minimizer of the following linear optimization problem: $\mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\langle\sum_{i=1}^{t-1}\nabla_i - \sigma_{t,j}, \mathbf{x}\rangle$. Here, $\{\sigma_{t,j}\}_{j=1}^{m}$ are independent random perturbations drawn from some appropriate probability distribution such as exponential distribution or uniform distribution in a hyper-cube. Various choices of perturbation distribution gives rise to various FTPL algorithms. When the loss functions are linear, Kalai and Vempala [KV05] show that FTPL achieves $O\left(T^{1/2}\right)$ expected regret, irrespective of the choice of m. When the loss functions are convex, Hazan [Haz16] showed that the deterministic version of FTPL (*i.e.*, as $m \to \infty$) achieves $O\left(T^{1/2}\right)$ regret. While projection free methods for online convex learning have been studied since the early work of [HK12], surprisingly, regret bounds of FTPL for finite m have only been recently studied [HM20]. Hazan and Minasyan [HM20] show that for Lipschitz and convex functions, FTPL achieves $O\left(T^{1/2} + m^{-1/2}T\right)$ expected regret, and for smooth convex functions, the algorithm achieves $O\left(T^{1/2} + m^{-1}T\right)$ expected regret.

Online Learning with Optimism. When the sequence of loss functions is convex and predictable, Rakhlin and Sridharan [RS12] and Rakhlin and Sridharan [RS13] study optimistic variants of FTRL which can exploit the predictability to obtain better regret bounds. Let g_t be our guess of $\nabla f_t(\mathbf{x}_t)$ at the beginning of round t. Given g_t, we predict $\mathbf{x}_t$ in Optimistic FTRL (OFTRL) as $\mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\langle\sum_{i=1}^{t-1}\nabla f_i(\mathbf{x}_i) + g_t, \mathbf{x}\rangle + R(\mathbf{x})$. Note that when $g_t = 0$, OFTRL is equivalent to FTRL. [RS12; RS13] show that the regret bounds of OFTRL only depend on $(g_t - \nabla f_t(\mathbf{x}_t))$. Moreover, these works show that OFTRL provides faster convergence rates for solving smooth convex-concave games. In contrast to FTRL, the optimistic variants of FTPL have been less well understood. Rakhlin and Sridharan [RS12] studies OFTPL for linear loss functions. But they consider restrictive settings and their algorithms require the knowledge of sizes of deviations $(g_t - \nabla f_t(\mathbf{x}_t))$. When the sequence of loss functions is nonconvex and predictable, there are no prior works which study OFTPL.

Projection Free Learning. Projection free optimization algorithms are those algorithms which only involve solving linear optimization problems in each iteration. They are attractive because for many problem of interest linear optimization problems are very easy to solve. Two broad classes of projection free techniques have been considered for online convex learning and convex-concave minimax games, namely, Frank-Wolfe (FW) methods and FTPL based methods. Garber and Hazan [GH13] consider the problem of online learning when the action space $\mathcal{X}$ is a *polytope*. They provide a FW method which achieves $O\left(T^{1/2}\right)$ regret using T calls to the linear optimization oracle. Hazan and Kale [HK12] provide a FW technique which achieves $O\left(T^{3/4}\right)$ regret for general online convex learning with Lipschitz losses and uses T calls to the linear optimization oracle. In a recent work, Hazan and Minasyan [HM20] show that FTPL achieves $O\left(T^{2/3}\right)$ regret for online convex learning with smooth losses, using T calls to the linear optimization oracle. This translates to $O\left(T^{-1/3}\right)$ rate of convergence for solving smooth convex-concave games. Note that, in contrast, our algorithm achieves $O\left(T^{-1/2}\right)$ convergence rate in the same setting. Gidel, Jebara, and Lacoste-Julien [GJL16] study FW methods for solving convex-concave games. When the constraint sets $\mathcal{X}, \mathcal{Y}$ are *strongly convex*, the authors show geometric convergence of their algorithms. In a recent work, He and Harchaoui [HH15] propose a FW technique for solving smooth convex-concave games which converges at a rate of $O\left(T^{-1/2}\right)$ using T calls to the linear optimization oracle. We note that our simple OFTPL based algorithm achieves these rates, with the added advantage of parallelizability. That being said, He and Harchaoui [HH15] achieve dimension free convergence rates in the Euclidean setting, where the smoothness is measured w.r.t $\|\cdot\|_2$ norm. In contrast, the rates of convergence of our algorithm depend on the dimension.

Notation. $\|\cdot\|$ is a norm on some vector space, which is typically $\mathbb{R}^d$ in our work. $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, which is defined as $\|\mathbf{x}\|_* = \sup\{\langle \mathbf{u}, \mathbf{x}\rangle : \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\| \leq 1\}$. We use Ψ_1, Ψ_2 to denote norm compatibility constants of $\|\cdot\|$, which are defined as $\Psi_1 = \sup_{\mathbf{x}\neq 0} \|\mathbf{x}\|/\|\mathbf{x}\|_2,\ \Psi_2 = \sup_{\mathbf{x}\neq 0}\|\mathbf{x}\|_2/\|\mathbf{x}\|$.

We use the notation $f_{1:t}$ to denote $\sum_{i=1}^t f_i$ and ∇_i to denote $\nabla f_i(\mathbf{x}_i)$. In some cases, when clear from context, we overload the notation $f_{1:t}$ and use it to denote the set $\{f_1, f_2 \dots f_t\}$. For any convex function f, $\partial f(\mathbf{x})$ is the set of all subgradients of f at $\mathbf{x}$. For any function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, $f(\cdot, \mathbf{y}), f(\mathbf{x}, \cdot)$ denote the functions $\mathbf{x} \rightarrow f(\mathbf{x}, \mathbf{y}), \mathbf{y} \rightarrow f(\mathbf{x}, \mathbf{y})$. For any function $f : \mathcal{X} \rightarrow \mathbb{R}$ and any probability distribution P, we let $f(P)$ denote $\mathbb{E}_{\mathbf{x}\sim P}\left[f(\mathbf{x})\right]$. Similarly, for any function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ and any two distributions P, Q, we let $f(P, Q)$ denote $\mathbb{E}_{\mathbf{x}\sim P, \mathbf{y}\sim Q}\left[f(\mathbf{x}, \mathbf{y})\right]$. For any set of distributions $\{P_j\}_{j=1}^m$, $\frac{1}{m}\sum_{j=1}^m P_j$ is the mixture distribution which gives equal weights to its components. We use $\text{Exp}(\eta)$ to denote the exponential distribution, whose CDF is given by $P(Z \leq s) = 1 - \exp(-s/\eta)$.

3.2 Dual view of Perturbation as Regularization

In this section, we present a key result which shows that when the sequence of loss functions is convex, every FTPL algorithm is an FTRL algorithm. Our analysis of OFTPL relies on this dual view to obtain tight regret bounds. This duality between FTPL and FTRL was originally studied by Hofbauer and Sandholm [HS02], where the authors show that any FTPL algorithm, with perturbation distribution admitting a strictly positive density on $\mathbb{R}^d$, is an FTRL algorithm w.r.t some convex regularizer. However, many popular perturbation distributions such as exponential and uniform distributions don't have a strictly positive density. In a recent work, Abernethy, Lee, and Tewari [ALT16] point out that the duality between FTPL and FTRL holds for very general perturbation distributions. However, the authors do not provide a formal theorem showing this result. Here, we provide a proposition formalizing the claim of [ALT16].

Proposition 2. *Consider the problem of online convex learning, where the sequence of loss functions $\{f_t\}_{t=1}^T$ encountered by the learner are convex. Consider the deterministic version of FTPL algorithm, where the learner predicts $\mathbf{x}_t$ as $\mathbb{E}_\sigma\left[\text{argmin}_{\mathbf{x}\in\mathcal{X}}\langle \nabla_{1:t-1} - \sigma, \mathbf{x}\rangle\right]$. Suppose the perturbation distribution is absolutely continuous w.r.t the Lebesgue measure. Then there exists a convex regularizer $R : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$, with domain $dom(R) \subseteq \mathcal{X}$, such that*

Algorithm 2 Convex OFTPL

```
Input: Perturbation Distribution $P_{\text{PRTB}}$, number of samples $m$, number of iterations $T$
Denote $\nabla_0 = 0$
for $t = 1 \dots T$ do
    Let $g_t$ be the guess for $\nabla_t$
    for $j = 1 \dots m$ do
        Sample $\sigma_{t,j} \sim P_{\text{PRTB}}$
        $\mathbf{x}_{t,j} \in \text{argmin}_{\mathbf{x} \in \mathcal{X}} \langle \nabla_{0:t-1} + g_t - \sigma_{t,j}, \mathbf{x} \rangle$
    end for
    Play $\mathbf{x}_t = \frac{1}{m} \sum_{j=1}^{m} \mathbf{x}_{t,j}$
    Observe loss function $f_t$
end for
```

$\mathbf{x}_t = \text{argmin}_{\mathbf{x} \in \mathcal{X}} \langle \nabla_{1:t-1}, \mathbf{x} \rangle + R(\mathbf{x})$. *Moreover,* $-\nabla_{1:t-1} \in \partial R(\mathbf{x}_t)$, *and* $\mathbf{x}_t = \partial R^{-1} \left(-\nabla_{1:t-1} \right)$, *where* ∂R^{-1} *is the inverse of* ∂R *in the sense of multivalued mappings.*

3.3 Online Learning with OFTPL

3.3.1 Online Convex Learning

In this section, we present the OFTPL algorithm for online convex learning and derive an upper bound on its regret. The algorithm we consider is similar to the OFTRL algorithm (see Algorithm 2). Let $g_t[f_1 \dots f_{t-1}]$ be our guess for ∇_t at the beginning of round t, with $g_1 = 0$. To simplify the notation, in the sequel, we suppress the dependence of g_t on $\{f_i\}_{i=1}^{t-1}$. Given g_t, we predict $\mathbf{x}_t$ in OFTPL as follows. We sample independent perturbations $\{\sigma_{t,j}\}_{j=1}^{m}$ from the perturbation distribution P_{PRTB} and compute $\mathbf{x}_t$ as $m^{-1} \sum_{j=1}^{m} \mathbf{x}_{t,j}$, where $\mathbf{x}_{t,j}$ is a minimizer of the following linear optimization problem

$$\mathbf{x}_{t,j} \in \underset{\mathbf{x} \in \mathcal{X}}{\text{argmin}} \langle \nabla_{1:t-1} + g_t - \sigma_{t,j}, \mathbf{x} \rangle.$$

We now present our main theorem which bounds the regret of OFTPL. A key quantity the regret depends on is the *stability* of predictions of the deterministic version of OFTPL. Intuitively, an algorithm is stable if its predictions in two consecutive iterations differ by a small quantity. To capture this notion, we first define function $\nabla\Phi : \mathbb{R}^d \to \mathbb{R}^d$ as: $\nabla\Phi\left(g\right) = \mathbb{E}_{\sigma}\left[\text{argmin}_{\mathbf{x} \in \mathcal{X}} \left\langle g - \sigma, \mathbf{x} \right\rangle\right]$. Observe that $\nabla\Phi\left(\nabla_{1:t-1} + g_t\right)$ is the prediction of the deterministic version of OFTPL. We say the predictions of OFTPL are stable, if $\nabla\Phi$ is a Lipschitz function.

Definition 3.3.1 (Stability)**.** The predictions of OFTPL are said to be β-stable w.r.t some norm $\|\cdot\|$, if

$$\forall g_1, g_2 \in \mathbb{R}^d \quad \|\nabla\Phi\left(g_1\right) - \nabla\Phi\left(g_2\right)\|_* \leq \beta\|g_1 - g_2\|.$$

Theorem 5. *Suppose the perturbation distribution* P_{PRTB} *is absolutely continuous w.r.t Lebesgue measure. Let* D *be the diameter of* $\mathcal{X}$ *w.r.t* $\|\cdot\|$, *which is defined as* $D =$

$\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}} \|\mathbf{x}_1 - \mathbf{x}_2\|$. *Let* $\eta = \mathbb{E}_\sigma\left[\|\sigma\|_*\right]$, *and suppose the predictions of OFTPL are* $C\eta^{-1}$*-stable w.r.t* $\|\cdot\|_*$, *where* C *is a constant that depends on the set* $\mathcal{X}$. *Finally, suppose the sequence of loss functions* $\{f_t\}_{t=1}^T$ *are Holder smooth and satisfy*

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \quad \|\nabla f_t(\mathbf{x}_1) - \nabla f_t(\mathbf{x}_2)\|_* \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|^\alpha,$$

for some constant $\alpha \in [0, 1]$. *Then the expected regret of Algorithm 2 satisfies*

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right] \leq \eta D + \sum_{t=1}^T \frac{C}{2\eta}\mathbb{E}\left[\|\nabla_t - g_t\|_*^2\right] - \sum_{t=1}^T \frac{\eta}{2C}\mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right] + LT\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha}.$$

where $\mathbf{x}_t^\infty = \mathbb{E}\left[\mathbf{x}_t | g_t, f_{1:t-1}, \mathbf{x}_{1:t-1}\right]$ *and* $\tilde{\mathbf{x}}_{t-1}^\infty = \mathbb{E}\left[\tilde{\mathbf{x}}_{t-1} | f_{1:t-1}, \mathbf{x}_{1:t-1}\right]$ *and* $\tilde{\mathbf{x}}_{t-1}$ *denotes the prediction in the* t^{th} *iteration of Algorithm 2, if guess* $g_t = 0$ *was used. Here,* Ψ_1, Ψ_2 *denote the norm compatibility constants of* $\|\cdot\|$.

Regret bounds that hold with high probability can be found in Appendix B.7. The above Theorem shows that the regret of OFTPL only depends on $\|\nabla_t - g_t\|_*$, which quantifies the accuracy of our guess g_t. In contrast, the regret of FTPL depends on $\|\nabla_t\|_*$ [Haz16]. This shows that for predictable sequences, with an appropriate choice of g_t, OFTPL can achieve better regret guarantees than FTPL. As we demonstrate in Section 5.2, this helps us design faster algorithms for solving minimax games.

Note that the above result is very general and holds for any absolutely continuous perturbation distribution. The key challenge in instantiating this result for any particular perturbation distribution is in showing the stability of predictions. Several past works have studied the stability of FTPL for various perturbation distributions such as uniform, exponential, Gumbel distributions [KV05; Haz16; HM20]. Consequently, the above result can be used to derive tight regret bounds for all these perturbation distributions. As one particular instantiation of Theorem 5, we consider the special case of $g_t = 0$ and derive regret bounds for FTPL, when the perturbation distribution is the uniform distribution over a ball centered at the origin.

Corollary 1 (FTPL). *Suppose the perturbation distribution is equal to the uniform distribution over* $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq (1 + d^{-1})\eta\}$. *Let* D *be the diameter of* $\mathcal{X}$ *w.r.t* $\|\cdot\|_2$. *Then* $\mathbb{E}_\sigma\left[\|\sigma\|_2\right] = \eta$, *and the predictions of OFTPL are* $dD\eta^{-1}$*-stable w.r.t* $\|\cdot\|_2$. *Suppose, the sequence of loss functions* $\{f_t\}_{t=1}^T$ *are G-Lipschitz and satisfy* $\sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x})\|_2 \leq G$. *Moreover, suppose* f_t *satisfies the Holder smooth condition in Theorem 5 w.r.t* $\|\cdot\|_2$ *norm. Then the expected regret of Algorithm 2 with guess* $g_t = 0$, *satisfies*

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right] \leq \eta D + \frac{dDG^2T}{2\eta} + LT\left(\frac{D}{\sqrt{m}}\right)^{1+\alpha}.$$

This recovers the regret bounds of FTPL for general convex loss functions derived by Hazan and Minasyan [HM20].

Algorithm 3 Nonconvex OFTPL

```
Input: Perturbation Distribution $P_{\text{PRTB}}$, number of samples $m$, number of iterations $T$
Denote $f_0 = 0$
for $t = 1 \dots T$ do
    Let $g_t$ be the guess for $f_t$
    for $j = 1 \dots m$ do
        Sample $\sigma_{t,j} \sim P_{\text{PRTB}}$
        $\mathbf{x}_{t,j} \in \text{argmin}_{\mathbf{x} \in \mathcal{X}} f_{0:t-1}(\mathbf{x}) + g_t(\mathbf{x}) - \sigma_{t,j}(\mathbf{x})$
    end for
    Let $P_t$ be the empirical distribution over $\{\mathbf{x}_{t,1}, \mathbf{x}_{t,2} \dots \mathbf{x}_{t,m}\}$
    Play $\mathbf{x}_t$, a random sample generated from $P_t$
    Observe loss function $f_t$
end for
```

3.3.2 Online Nonconvex Learning

We now study OFTPL in the nonconvex setting. In this setting, we assume the sequence of loss functions belong to some function class $\mathcal{F}$ containing real-valued measurable functions on $\mathcal{X}$. Some popular choices for $\mathcal{F}$ include the set of Lipschitz functions, the set of bounded functions. The OFTPL algorithm in this setting is described in Algorithm 3. Similar to the convex case, we first sample random perturbation functions $\{\sigma_{t,j}\}_{j=1}^m$ from some distribution P_{PRTB}. Some examples of perturbation functions that we have considered in Chapter 2 include $\sigma_{t,j}(\mathbf{x}) = \langle \bar{\sigma}_{t,j}, \mathbf{x} \rangle$, for some random vector $\bar{\sigma}_{t,j}$ sampled from exponential or uniform distributions. Another popular choice for $\sigma_{t,j}$ is the Gumbel process, which results in the continuous exponential weights algorithm [MTM14]. Letting, g_t be our guess of loss function f_t at the beginning of round t, the learner first computes $\mathbf{x}_{t,j}$ as $\text{argmin}_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^{t-1} f_i(\mathbf{x}) + g_t(\mathbf{x}) - \sigma_{t,j}(\mathbf{x})$. We assume access to an optimization oracle which computes a minimizer of this problem. We often refer to this oracle as the *perturbed best response* oracle. Let P_t denote the empirical distribution of $\{\mathbf{x}_{t,j}\}_{j=1}^m$. The learner then plays an $\mathbf{x}_t$ which is sampled from P_t. Algorithm 3 describes this procedure. We note that for the online learning problem, $m = 1$ suffices, as the expected loss suffered by the learner in each round is independent of m; that is $\mathbb{E}\left[f_t(\mathbf{x}_t)\right] = \mathbb{E}\left[f_t(\mathbf{x}_{t,1})\right]$. However, the choice of m affects the rate of convergence when Algorithm 3 is used for solving nonconvex nonconcave minimax games.

Before we present the regret bounds, we introduce the *dual space* associated with $\mathcal{F}$. Let $\|\cdot\|_{\mathcal{F}}$ be a seminorm associated with $\mathcal{F}$. For example, when $\mathcal{F}$ is the set of Lipschitz functions, $\|\cdot\|_{\mathcal{F}}$ is the Lipschitz seminorm. Various choices of $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ induce various distance metrics on $\mathcal{P}$, the set of all probability distributions on $\mathcal{X}$. We let $\gamma_{\mathcal{F}}$ denote the Integral Probability Metric (IPM) induced by $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$, which is defined as

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}, \|f\|_{\mathcal{F}} \leq 1} \Big| \mathbb{E}_{\mathbf{x} \sim P}\left[f(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x} \sim Q}\left[f(\mathbf{x})\right] \Big|.$$

We often refer to $(\mathcal{P}, \gamma_{\mathcal{F}})$ as the dual space of $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$. When $\mathcal{F}$ is the set of Lips-

$\gamma_{\mathcal{F}}(P, Q)$	$\|f\|_{\mathcal{F}}$	$\mathcal{F}$
Dudley Metric	$\text{Lip}(f) + \|f\|_\infty$	$\{f : \text{Lip}(f) + \|f\|_\infty < \infty\}$
Kantorovich Metric (or) Wasserstein-1 Metric	$\text{Lip}(f)$	$\{f : \text{Lip}(f) < \infty\}$
Total Variation (TV) Distance	$\|f\|_\infty$	$\{f : \|f\|_\infty < \infty\}$
Maximum Mean Discrepancy (MMD) for RKHS $\mathcal{H}$	$\|f\|_{\mathcal{H}}$	$\{f : \|f\|_{\mathcal{H}} < \infty\}$

Table 3.1: Table showing some popular Integral Probability Metrics. Here $\text{Lip}(f)$ is the Lipschitz constant of f which is defined as $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} |f(\mathbf{x}) - f(\mathbf{y})|/\|\mathbf{x}-\mathbf{y}\|$ and $\|f\|_\infty$ is the supremum norm of f.

chitz functions and when $\|\cdot\|_{\mathcal{F}}$ is the Lipschitz seminorm, $\gamma_{\mathcal{F}}$ is the Wasserstein distance. Table 3.1 presents examples of $\gamma_{\mathcal{F}}$ induced by some popular function spaces. Similar to the convex case, the regret bounds in the nonconvex setting depend on the stability of predictions of OFTPL.

Definition 3.3.2 (Stability)**.** Suppose the perturbation function $\sigma(\mathbf{x})$ is sampled from P_{PRTB}. For any $f \in \mathcal{F}$, define random variable $\mathbf{x}_f(\sigma)$ as $\text{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x})$. Let $\nabla\Phi\left(f\right)$ denote the distribution of $\mathbf{x}_f(\sigma)$. The predictions of OFTPL are said to be β-stable w.r.t $\|\cdot\|_{\mathcal{F}}$ if

$$\forall f, g \in \mathcal{F} \quad \gamma_{\mathcal{F}}(\nabla\Phi\left(f\right), \nabla\Phi\left(g\right)) \leq \beta\|f - g\|_{\mathcal{F}}.$$

Theorem 6. *Suppose the sequence of loss functions $\{f_t\}_{t=1}^T$ belong to $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$. Suppose the perturbation distribution P_{PRTB} is such that $\text{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x})$ has a unique minimizer with probability one, for any $f \in \mathcal{F}$. Let $\mathcal{P}$ be the set of probability distributions over $\mathcal{X}$. Define the diameter of $\mathcal{P}$ as $D = \sup_{P_1,P_2\in\mathcal{P}} \gamma_{\mathcal{F}}(P_1, P_2)$. Let $\eta = \mathbb{E}\left[\|\sigma\|_{\mathcal{F}}\right]$. Suppose the predictions of OFTPL are $C\eta^{-1}$-stable w.r.t $\|\cdot\|_{\mathcal{F}}$, for some constant C that depends on $\mathcal{X}$. Then the expected regret of Algorithm 3 satisfies*

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right] \leq \eta D + \sum_{t=1}^T \frac{C}{2\eta}\mathbb{E}\left[\|f_t - g_t\|_{\mathcal{F}}^2\right] - \sum_{t=1}^T \frac{\eta}{2C}\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\right],$$

where $P_t^\infty = \mathbb{E}\left[P_t|g_t, f_{1:t-1}, P_{1:t-1}\right]$, $\tilde{P}_t^\infty = \mathbb{E}\left[\tilde{P}_{t-1}|f_{1:t-1}, P_{1:t-1}\right]$ and $\tilde{P}_{t-1}$ is the empirical distribution computed in the t^{th} iteration of Algorithm 3, if guess $g_t = 0$ was used.

As in the convex case, the key challenge in instantiating the above result for any particular perturbation distribution is in showing the stability of predictions. In Chapter 2 we considered linear perturbation functions $\sigma(\mathbf{x}) = \langle\bar{\sigma}, \mathbf{x}\rangle$, for $\bar{\sigma}$ sampled from exponential distribution, and showed stability of FTPL. We now instantiate the above theorem for this setting.

Corollary 2. *Consider the setting of Theorem 6. Let $\mathcal{F}$ be the set of Lipschitz functions and $\|\cdot\|_{\mathcal{F}}$ be the Lipschitz seminorm, which is defined as $\|f\|_{\mathcal{F}} = \sup_{\mathbf{x}\neq\mathbf{y} \text{ in } \mathcal{X}} |f(\mathbf{x}) - f(\mathbf{y})|/\|\mathbf{x}-\mathbf{y}\|_1$. Suppose the perturbation function is such that $\sigma(\mathbf{x}) = \langle\bar{\sigma}, \mathbf{x}\rangle$, where $\bar{\sigma} \in \mathbb{R}^d$ is a random vector whose entries are sampled independently from Exp(η). Then*

$\mathbb{E}_\sigma\left[\|\sigma\|_{\mathcal{F}}\right] = \eta \log d$, *and the predictions of OFTPL are* $O\left(d^2 D\eta^{-1}\right)$*-stable w.r.t* $\|\cdot\|_{\mathcal{F}}$. *Moreover, the expected regret of Algorithm 3 is upper bounded by*

$$O\left(\eta D \log d + \sum_{t=1}^{T} \frac{d^2 D}{\eta}\mathbb{E}\left[\|f_t - g_t\|_{\mathcal{F}}^2\right] - \sum_{t=1}^{T}\frac{\eta}{d^2 D}\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\right]\right).$$

We note that the above regret bounds are tighter than the regret bounds of FTPL derived in Chapter 2, where we derived the following bound

$$O\left(\eta D \log d + \sum_{t=1}^{T}\frac{d^2 D}{\eta}\mathbb{E}\left[\|f_t\|_{\mathcal{F}}^2\right]\right).$$

These tigher bounds help us design faster algorithms for solving minimax games in the nonconvex setting (see Section 3.4 for a more detailed discussion).

3.4 Minimax Games

We now consider the problem of solving minimax games of the following form

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}). \tag{3.1}$$

Nash equilibria of such games can be computed by playing two online learning algorithms against each other [CL06; Haz16]. In this chapter, we study the algorithm where both the players employ OFTPL to decide their actions in each round.

Convex-Concave games. For convex-concave games, both the players use the OFTPL algorithm described in Algorithm 2 (see Algorithm 13 in Appendix B.4). The following theorem derives the rate of convergence of this algorithm to a Nash equilibirum (NE).

Theorem 7. *Consider the minimax game in Equation (3.1). Suppose both the domains* $\mathcal{X}, \mathcal{Y}$ *are compact subsets of* $\mathbb{R}^d$, *with diameter*

$$D = \max\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}}\|\mathbf{x}_1-\mathbf{x}_2\|_2, \sup_{\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}}\|\mathbf{y}_1-\mathbf{y}_2\|_2\}.$$

Suppose f *is convex in* $\mathbf{x}$, *concave in* $\mathbf{y}$ *and is smooth w.r.t* $\|\cdot\|_2$

$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}',\mathbf{y}')\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}',\mathbf{y}')\|_2 \leq L\|\mathbf{x}-\mathbf{x}'\|_2 + L\|\mathbf{y}-\mathbf{y}'\|_2.$$

Suppose Algorithm 13 is used to solve the minimax game. Suppose the perturbation distributions used by both the players are the same and equal to the uniform distribution over $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq (1+d^{-1})\eta\}$. *Suppose the guesses used by* $\mathbf{x},\mathbf{y}$ *players in the* t^{th} *iteration are* $\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1}), \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})$, *where* $\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1}$ *denote the predictions of* $\mathbf{x},\mathbf{y}$ *players in the* t^{th} *iteration, if guess* $g_t = 0$ *was used. If Algorithm 13 is run with* $\eta = 6dD(L+1), m = T$, *then the iterates* $\{(\mathbf{x}_t,\mathbf{y}_t)\}_{t=1}^T$ *satisfy*

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y}\right) - f\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\right] = O\left(\frac{dD^2(L+1)}{T}\right).$$

Table 3.2: Table comparing oracle complexities of various projection free techniques for finding an ϵ-approximate NE of smooth convex-concave games.

Method	Total calls to linear optimization oracle	Number of iterations	Parallel calls to oracle in each iteration
FTPL	$O\left(\epsilon^{-3}\right)$	$O\left(\epsilon^{-2}\right)$	$O\left(\epsilon^{-1}\right)$
He and Harchaoui [HH15]	$O\left(\epsilon^{-2}\right)$	$O\left(\epsilon^{-2}\right)$	1
OFTPL	$O\left(\epsilon^{-2}\right)$	$O\left(\epsilon^{-1}\right)$	$O\left(\epsilon^{-1}\right)$

Rates of convergence which hold with high probability can be found in Appendix B.7. We note that Theorem 7 can be extended to more general noise distributions and settings where gradients of f are Holder smooth w.r.t non-Euclidean norms, and $\mathcal{X}, \mathcal{Y}$ lie in spaces of different dimensions (see Theorem 27 in Appendix). We now discuss the above result.

- Theorem 7 shows that for smooth convex-concave games, Algorithm 13 converges to a NE at $O\left(T^{-1}\right)$ rate using $4T^2$ calls to the linear optimization oracle. Moreover, the algorithm runs in T iterations, with each iteration making $4T$ parallel calls to the optimization oracle. In contrast, FTPL makes $2T^3$ calls to the linear optimization oracle to achieve $O\left(T^{-1}\right)$ rates of convergence and runs for T^2 iterations, with each iteration making $2T$ parallel calls to the optimization oracle. This can be obtained by setting $m = \sqrt{T}, \alpha = 1$, and $\eta = O\left(\sqrt{T}\right)$ in Corollary 1.
- The Frank-Wolfe technique of He and Harchaoui [HH15] achieves the same convergence rates as our algorithm; that is, it achieves $O\left(T^{-1}\right)$ rates using T^2 calls to the linear optimization oracle. However, unlike [HH15], our algorithm is parallelizable and can be run in T iterations.
- He and Harchaoui [HH15] achieve dimension free convergence rates in the Euclidean setting, where the smoothness is measured w.r.t $\|\cdot\|_2$ norm. In contrast, the rates of convergence of our algorithm depend on the dimension. We believe the dimension dependence in the rates can be removed by appropriately choosing the perturbation distributions based on domains $\mathcal{X}, \mathcal{Y}$ (see Appendix B.6).
- Note that OFTRL also achieves $O\left(T^{-1}\right)$ rates of convergence after T iterations. However, each iteration of OFTRL involves optimization of a non-linear convex function over the domains $\mathcal{X}, \mathcal{Y}$, which can be quite expensive in practice.

Nonconvex-Nonconcave games. We now consider the more general nonconvex - nonconcave games. In this case, both the players use the nonconvex OFTPL algorithm described in Algorithm 3 to choose their actions. Instead of generating a single sample from the empirical distribution P_t computed in t^{th} iteration of Algorithm 3, the players now play the entire distribution P_t (see Algorithm 14 in Appendix B.5). Letting $\{P_t\}_{t=1}^T, \{Q_t\}_{t=1}^T$, be the sequence of iterates generated by the $\mathbf{x}$ and $\mathbf{y}$ players, the following theorem shows that $\left(\frac{1}{T}\sum_{t=1}^T P_t, \frac{1}{T}\sum_{t=1}^T Q_t\right)$ converges to a NE.

Theorem 8. *Consider the minimax game in Equation (3.1). Suppose the domains $\mathcal{X}, \mathcal{Y}$ are compact subsets of $\mathbb{R}^d$ with diameter $D = \max\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}} \|\mathbf{x}_1-\mathbf{x}_2\|_1, \sup_{\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}} \|\mathbf{y}_1-\mathbf{y}_2\|_1\}$.*

Suppose f is Lipschitz w.r.t $\|\cdot\|_1$ and satisfies

$$\max\left\{\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y})\|_\infty, \sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})\|_\infty\right\}\leq G.$$

Moreover, suppose f satisfies the following smoothness property

$$\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{x}}f(\mathbf{x}',\mathbf{y}')\|_\infty+\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f(\mathbf{x}',\mathbf{y}')\|_\infty\leq L\|\mathbf{x}-\mathbf{x}'\|_1+L\|\mathbf{y}-\mathbf{y}'\|_1.$$

Suppose both $\mathbf{x}$ and $\mathbf{y}$ players use Algorithm 14 to solve the game with linear perturbation functions $\sigma(\mathbf{z})=\langle\bar{\sigma},\mathbf{z}\rangle$, where $\bar{\sigma}\in\mathbb{R}^d$ is such that each of its entries is sampled independently from Exp(η). Suppose the guesses used by $\mathbf{x}$ and $\mathbf{y}$ players in the t^{th} iteration are $f(\cdot,\tilde{Q}_{t-1}), f(\tilde{P}_{t-1},\cdot)$, where $\tilde{P}_{t-1},\tilde{Q}_{t-1}$ denote the predictions of $\mathbf{x},\mathbf{y}$ players in the t^{th} iteration, if guess $g_t=0$ was used. If Algorithm 14 is run with $\eta=10d^2D(L+1), m=T$, then the iterates $\{(P_t,Q_t)\}_{t=1}^T$ satisfy

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T P_t,\mathbf{y}\right)-f\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^T Q_t\right)\right]=O\left(\frac{d^2D^2(L+1)\log d}{T}\right)$$
$$+O\left(\min\left\{D^2L,\frac{d^2G^2\log T}{LT}\right\}\right).$$

More general versions of the Theorem, which consider other function classes and general perturbation distributions, can be found in Appendix B.5. The above result shows that Algorithm 14 converges to a NE at $\tilde{O}\left(T^{-1}\right)$ rate using T^2 calls to the perturbed best response oracle. This matches the rates of convergence of FTPL derived in Chapter 2. However, the key advantage of our algorithm is that it is highly parallelizable and runs in $O\left(T\right)$ iterations, in contrast to FTPL, which runs in $O\left(T^2\right)$ iterations.

3.5 Discussion

We studied an optimistic variant of FTPL which achieves better regret guarantees when the sequence of loss functions is predictable. As one specific application of our algorithm, we considered the problem of solving minimax games. For solving convex-concave games, our algorithm requires access to a linear optimization oracle and for nonconvex-nonconcave games our algorithm requires access to a more powerful perturbed best response oracle. In both these settings, our algorithm achieves $O\left(T^{-1/2}\right)$ convergence rates using T calls to the oracles. Moreover, our algorithm runs in $O\left(T^{1/2}\right)$ iterations, with each iteration making $O\left(T^{1/2}\right)$ parallel calls to the optimization oracle. We believe our improved algorithms for solving minimax games are useful in a number of modern machine learning applications such as training of GANs, adversarial training, which involve solving nonconvex-nonconcave minimax games and often deal with huge datasets.

Part II

Bandit Optimization

Chapter 4

Efficient Bandit Optimization for Convex Quadratic Losses

In this chapter, we study the problem of online learning with bandit feedback, which can be viewed as a repeated game between a learner and an adversary. In round t of this game, the learner chooses an action $\mathbf{x}_t$ from a known domain $\mathcal{X} \subset \mathbb{R}^d$. The adversary simultaneously selects a loss function $f_t : \mathcal{X} \to$ and reveals the loss suffered by the learner $f_t(\mathbf{x}_t)$. The performance of the learner at the end of T rounds is measured using regret

$$\mathrm{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

For this problem, we would like to design an algorithm for choosing $\mathbf{x}_t$ that satisfies the following key criteria: (a) (**optimal regret**) achieves regret bounds which have optimal dependence on T, and which hold in high-probability against adaptive adversaries, and (b) (**computational efficiency**) the run-time of *each iteration* of algorithm should have a small dependence on dimension d, *e.g.,* polynomial dependence with a small exponent, and *independent* of the number of rounds T (ideally, we would like it to have similar run-time as efficient algorithms for online learning in the full information setting).

The framework of bandit optimization is extremely general and has found numerous practical applications in fields such as computer science, economics, game theory, and medical decision making. Some of these applications include design of clinical trials, market pricing, online ad placement, and recommender systems [Kle05; BC12; Haz16]. Owing to its importance, there has been extensive work on designing low-regret algorithms for bandit optimization. Early works on this problem have focused on finite action space $\mathcal{X}$, in which case the problem is called multi-armed bandit (MAB) problem. This problem has been well studied and several efficient algorithms achieving the optimal high-probability regret rate of $O\left(T^{1/2}\right)$ have been developed [Aue+02; AB10; Lee+20]. Later works on bandit optimization have turned to continuous action spaces and convex loss functions. Seminal works along this line have developed online gradient descent style algorithms for regret minimization [FKM04; Kle05]. When the loss functions f_t are convex and bounded, these algorithms achieve $O\left(T^{5/6}\right)$ regret in *expectation*. Improving upon these regret guarantees has remained an open problem until the works of Hazan and Li [HL16] and Bubeck, Lee, and Eldan [BLE17]. The algorithms developed in these latter works achieve the optimal $\tilde{O}\left(T^{1/2}\right)$ regret, albeit they are computationally expensive and are not efficiently implementable in practice. In particular, the run time of the algorithm of Hazan and Li [HL16] depends exponentially on the dimension, and the algorithm of Bubeck, Lee, and Eldan [BLE17] involves minimization of an approximately convex function over a nonconvex set, which is non-trivial in practice[1].

Despite years of research, designing efficient algorithms for bandit convex optimization (BCO) has turned out to be challenging. This can be attributed to the extremely limited information available to the learner about the loss functions chosen by the adversary. Consequently, several works have focused on sub-cases of BCO. These works can be classified into two broad categories. One category imposes parametric assumptions such as linearity on the loss functions. The other category imposes structural assumptions such as strong convexity. The most popular parametric assumption that is studied in the literature is the linearity assumption [AHR09]. Recent works have designed efficient algorithms achieving optimal regret guarantees under this assumption [Lee+20]. However, apart from linearity, to the best of our knowledge, no other parametric assumption has been considered in the literature. When it comes to structural assumptions, several works have considered assumptions such as Lipschitzness [FKM04], smoothness [ST11], and strong convexity, smoothness [HL14; Ito20]. Perhaps surprisingly, among all these assumptions, computationally efficient and optimal algorithms are only known for strongly convex, smooth functions [HL14]. While these results are interesting, it should be noted that strong convexity is a restrictive assumption which rarely holds in practice. Consequently, it is important to relax this assumption. However, the oracle lower bounds of Hu, Prashanth, György, and Szepesvari [Hu+16] suggest that designing optimal algorithms for non-strongly convex, smooth functions might require new and different algorithmic techniques to those used in existing works. In particular, all existing works which design computationally efficient algorithms first estimate the gradient of the loss function from one-point feedback, and then use Online Mirror Descent (OMD) style updates to choose the next action. The lower bounds of Hu, Prashanth, György, and Szepesvari [Hu+16] suggest that such techniques will not be able to achieve the optimal $\tilde{O}\left(T^{1/2}\right)$ regret for non-strongly convex, smooth functions. So, to make progress along this line, we need new algorithmic techniques. Unfortunately, it is unclear how to come up with such techniques for general convex functions.

This Chapter. In this chapter, we make progress on this problem by designing an efficient algorithm for convex, quadratic loss functions that achieves optimal high-probability regret guarantees against an adaptive adversary. To be precise, our algorithm achieves a regret of $\tilde{O}\left(d^{16}\sqrt{T}\right)$, which is known to be optimal in T (see Dani, Hayes, and Kakade

[1] Although Bubeck, Lee, and Eldan [BLE17] present a modified algorithm for polytopes which can be implemented in polynomial time, the *per iteration* runtime of this algorithm has a large polynomial dependence on d and a linear dependence on T.

[DHK07] for lower bounds on the regret). In terms of computation, the key computational bottleneck of our algorithm involves generating uniform samples from a convex set. This is a well studied problem and several efficient MCMC algorithms such as Hit-and-run algorithms have been developed for this problem [LV03; Bel+15; LLV20]. For action sets which are polytopes with m constraints, the amortized time complexity of each iteration of our algorithm is $\tilde{O}\left(\frac{m^2d^3+md^6}{T} + md^4 + m^2d\right)$. In comparison, the only existing computationally efficient and optimal algorithm for this setting has a time complexity of $\tilde{O}\left(\text{poly}(dm)T\right)$ with a much larger exponent on d [BLE17]. Moreover, the runtime of each iteration of this algorithm has a linear dependence on T, thus making it extremely inefficient for large T.

Furthermore, our algorithm is robust to model mis-specification: if each loss function f_t is ϵ-close to a convex, quadratic function in $\|\cdot\|_\infty$ norm, the regret of our algorithm is bounded by $\tilde{O}\left(\epsilon T + d^{16}\sqrt{T}\right)$. We believe robustness is necessary for algorithms which focus on sub-cases of BCO, as the assumptions on loss functions do not typically hold in practice. However, most existing works do not study this property. To the best of our knowledge, ours is the first algorithm for BCO with quadratic functions, which is computationally efficient, robust and achieves optimal regret guarantees.

Techniques. Our algorithm is a regularized Newton's method with self concordant barrier of $\mathcal{X}$ as the regularizer. It involves estimation of gradients and Hessians of the loss functions from single point feedback. This is unlike most existing computationally efficient algorithms which rely *only* on the gradient estimates to choose their actions [ST11; HL14]. As previously mentioned, gradient information alone doesn't suffice to design algorithms achieving optimal regret for nonlinear losses (see Section 4.4.1 for empirical evidence). So, in this chapter, we estimate both the gradient and Hessian of the unknown loss function and use the estimates in a regularized Newton method. However, estimating the Hessian comes with its own challenges. The variance of the Hessian estimates is typically very large. Consequently, we need new techniques to cancel the effect of variance. In this chapter, we crucially rely on "focus regions" to handle the variance. This technique is inspired by Bubeck, Lee, and Eldan [BLE17], who use similar ideas to design an optimal, albeit computationally inefficient, algorithm for general convex functions. At a high level, the variance of the Hessian estimates can only be controlled in a small region, which we call focus region. So, we restrict ourselves to this region and always choose actions within this region. However, the resulting algorithm only ensures low regret with respect to (w.r.t) points in the focus region. To ensure low regret even w.r.t points outside the focus region, we perform a test every iteration called "restart condition". Intuitively, this test checks if the minimizer of the cumulative loss over the entire domain falls well within the focus region. If yes, we continue the algorithm, as having a low regret w.r.t points in the focus region ensures the overall regret is low. The test fails when the minimizer gets too close to the boundary of the focus region. In this case, we show that the regret of our actions until now is negative, and restart the algorithm.

While the ideas of focus region and restart condition appeared in Bubeck, Lee, and Eldan [BLE17], we note that new techniques are needed to make this approach computationally efficient. Restricting to quadratics doesn't automatically make Bubeck, Lee, and Eldan [BLE17]'s approach computationally efficient. To make our algorithm efficient, we move away from the exponential weights update scheme used by Bubeck, Lee, and Eldan [BLE17] and instead rely on Newton method and OMD framework. Moreover, we design a new test for the restart condition that is much more computationaly efficient than the test of Bubeck, Lee, and Eldan [BLE17] (see Section 4.4 for more details).

Before we proceed, we note that our algorithm requires access to a self-concordant barrier (SCB) of $\mathcal{X}$ which satisfies certain assumption on the behavior of its Hessian (see Assumption 1). If $\mathcal{X} \subset \mathbb{R}$, then any SCB satisfies this property (see Proposition 4). Moreover, we show that the log-barrier of any polyhedral set satisfies this property. We conjecture that any SCB of any convex action set $\mathcal{X} \subset \mathbb{R}^d$ satisfies this property.

Paper Organization. Section 4.1 presents necessary background. Section 4.2 presents our main results. Section 4.3 discusses some of the related works. Section 4.4 presents our algorithm and Section 4.5 discusses the key ideas used in the algorithm. In Section 4.6 we discuss the computational aspects of our algorithm. We conclude with Section 4.7. Due to the lack of space, most proofs are presented in the appendix.

4.1 Problem Setting and Background

Notation. Throughout the paper, we denote vectors by bold-faced letters ($\mathbf{x}$), and matrices by capital letters (A). $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$ and $\|\cdot\|_A$ is the weighted Euclidean norm, *i.e.*, $\|\mathbf{x}\|_A = \langle A\mathbf{x}, \mathbf{x}\rangle^{1/2}$, where A is a positive definite matrix. We let $B_r(\mathbf{x})$ denote an ℓ_2 ball of radius r centered at $\mathbf{x}$, *i.e.*, $B_r(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y}-\mathbf{x}\| \leq r\}$. We let $B_{r,A}(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{x}-\mathbf{y}\|_A \leq r\}$. For any strictly convex twice differentiable function f, we define the local norm at $\mathbf{x}$ as $\|\mathbf{v}\|_{\mathbf{x},f} = \langle \mathbf{v}, \nabla^2 f(\mathbf{x})\mathbf{v}\rangle^{1/2}$. $\partial\mathcal{X}$ denotes the boundary of a set $\mathcal{X}$. $b = \tilde{O}\left(a\right)$ implies $b \leq Ca\log a$ for a large enough constant C independent of a.

A function $f : \mathcal{X} \to \mathbb{R}$ is ϵ-close to a function $g : \mathcal{X} \to \mathbb{R}$ if $\sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - g(\mathbf{x})| \leq \epsilon$. A function f is a quadratic function if it can be parameterized as $f(\mathbf{x}; A, \mathbf{b}, c) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \langle \mathbf{b}, \mathbf{x}\rangle + c$, for some $A \in \mathbb{R}^{d\times d}, \mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$. In addition, if $(A + A^T)$ is positive semidefinite, then f is called a *convex* quadratic function. Note that the set of *linear* functions is a subset of the set of convex quadratic functions. We let $\mathbb{E}_t$ denote the conditional expectation conditioned on all randomness in the first $t-1$ rounds. We use $\mathbb{B}^d, \mathbb{S}^{d-1}$ to denote the d-dimensional unit ball and unit sphere w.r.t Euclidean norm. We let $\mathbf{u} \sim \mathbb{B}^d, \mathbf{v} \sim \mathbb{S}^{d-1}$ denote the random variables chosen uniformly from these sets.

Problem Setting. In this chapter, we assume that the action space $\mathcal{X}$ is convex, compact and has non-empty interior. Without loss of generality, we assume $\mathcal{X}$ contains an Euclidean ball of radius 1, and has an ℓ_2 diameter of D, *i.e.*, $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\| \leq D$. We assume that each loss function f_t is ϵ-close to a convex quadratic function q_t which is bounded and Lipschitz, *i.e.*, $\sup_{\mathbf{x}\in\mathcal{X}} |q_t(\mathbf{x})| \leq B$, and for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $|q_t(\mathbf{x}) - q_t(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|$. Finally, we assume the adversary is adaptive, *i.e.*, the decisions of the adversary can depend on the learner's previous actions.

4.1.1 One-point Gradient and Hessian Estimates

A major component of our algorithm involves estimating the gradient and Hessian of the unknown loss function from one-point feedback provided by the adversary. These estimates are then used in OMD to pick the next move of the learner. In this chapter, we rely on the following randomized sampling scheme to compute these estimates.

Proposition 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a quadratic function. Let $C \in \mathbb{R}^{d\times d}$ be any symmetric positive definite matrix. Then*

$$\nabla f(\mathbf{x}) = d\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1 f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right],$$
$$\nabla^2 f(\mathbf{x}) = \frac{d^2}{2}\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right].$$

To generate unbiased estimates of the gradient and Hessian of f at $\mathbf{x}$, we first randomly sample $\mathbf{v}_1$ and $\mathbf{v}_2$ from uniform distribution on $\mathbb{S}^{d-1}$, and get the one-point feedback from the adversary about $f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)$, and then rely on the above proposition. We note that one can also rely on Gaussian smoothing to estimate this information (see Proposition 16 in Appendix). For the simplicity and clarity of analysis, in this chapter, we use the above sampling scheme instead of Gaussian smoothing. However, our algorithm and its analysis can be modified in a straightforward way to rely on Gaussian smoothing.

4.1.2 Self Concordant Barriers

Self Concordant Barriers (SCBs) play a crucial role in our algorithm and its analysis. So, in this section, we define SCB and present some of its useful properties.

Definition 4.1.1. Let $\mathcal{X} \subseteq^d$ be a closed convex set with non-empty interior. A function $R : \text{int}(\mathcal{X}) \to$ is called a ν-self-concordant barrier of $\mathcal{X}$, if

1. (Barrier Property) R is three times continuously differentiable with $R(\mathbf{x}_k) \to \infty$ along every sequence $\{\mathbf{x}_k \in \text{int}(\mathcal{X})\}$ converging to a boundary point of $\mathcal{X}$, as $k \to \infty$
2. R satisfies the following for all $\mathbf{x} \in \text{int}(\mathcal{X}), h \in^d$,

$$|\nabla^3 R(\mathbf{x})[h,h,h]| \le 2|\nabla^2 R(\mathbf{x})[h,h]|^{3/2}, \quad |\langle \nabla R(\mathbf{x}), h\rangle| \le \sqrt{\nu}|\nabla^2 R(\mathbf{x})[h,h]|^{1/2}.$$

Without loss of generality, we assume $\min_{\mathbf{x}\in\mathcal{X}} R(\mathbf{x}) = 0$. It is well known that R satisfies the following properties (see Appendix C.7 for a more comprehensive review)

- **(P1)** *Dikin Ellipsoid*: For any $\mathbf{x} \in \text{int}(\mathcal{X})$, the Dikin ellipsoid centered at $\mathbf{x}$, $B_{1,\nabla^2 R(\mathbf{x})}(\mathbf{x})$, is entirely contained in $\mathcal{X}$.
- **(P2)** For any $\mathbf{x} \in \text{int}(\mathcal{X})$, and $\mathbf{y} \in B_{1,\nabla^2 R(\mathbf{x})}(\mathbf{x})$, we have

$$(1-\|\mathbf{x}-\mathbf{y}\|_{\nabla^2 R(\mathbf{x})})^2\nabla^2 R(\mathbf{x}) \preceq \nabla^2 R(\mathbf{y}) \preceq \frac{1}{(1-\|\mathbf{x}-\mathbf{y}\|_{\nabla^2 R(\mathbf{x})})^2}\nabla^2 R(\mathbf{x}). \tag{4.1}$$

In this chapter, we assume that $\mathcal{X}$ has an SCB which satisfies the following additional property.

Assumption 1. *For any $\mathbf{x}, \mathbf{y} \in int(\mathcal{X})$ such that $\|\mathbf{y} - \mathbf{x}\|_{\nabla^2 R(\mathbf{x})} \leq \lambda$*

$$\nabla^2 R(\mathbf{y}) \succeq \frac{1}{(1+\lambda)^2} \nabla^2 R(\mathbf{x}). \tag{4.2}$$

The following propositions show that a wide range of action spaces have SCBs which satisfy this property. We conjecture that any SCB satisfies this property.

Proposition 4. *Suppose $\mathcal{X} \subseteq \mathbb{R}$. Then any SCB of $\mathcal{X}$ satisfies Assumption 1.*

Proposition 5. *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is polyhedral,* i.e., *it is the intersection of a finite number of closed half spaces. Then the logarithmic barrier of $\mathcal{X}$ is an SCB which satisfies Assumption 1.*

4.2 Main Results

Theorem 9 (Approximately quadratic losses). *Suppose f_t is ϵ-close to a convex, quadratic function $q_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A_t \mathbf{x} + \langle \mathbf{b}_t, \mathbf{x} \rangle + c_t$, for $\epsilon = O\left(dBT^{-1/2}\right)$. Let R be a ν-self-concordant barrier of $\mathcal{X}$ that satisfies Assumption 1. Suppose Algorithm 4 is run for T iterations with appropriate choice of hyper-parameters. Suppose the diameter of $\mathcal{X}$ is bounded by T, and the Lipschitz constants of $\{q_t\}_{t=1}^T$ are bounded by T. Then with probability at least $1-\delta$, the regret of the algorithm is upper bounded by $\tilde{O}\left(d^{11}(d+\nu)^5\sqrt{T}\right)$.*

Remarks. We now briefly discuss the above result. See Table 4.1 for a detailed comparison of our algorithm with other related algorithms.

- Our algorithm achieves the optimal regret guarantees in high probability, against adaptive adversaries. In comparison with Bubeck, Lee, and Eldan [BLE17], our regret bound has similar dependence on T and slightly worse dependence on dimension d. We believe the dimension dependence of our regret can be improved to d^8 using a tighter analysis. Also note that the OMD based algorithm of Saha and Tewari [ST11], which only relies on gradient estimates of loss functions, achieves a sub-optimal regret of $\tilde{O}\left(T^{2/3}\right)$.
- There are two key computational bottlenecks in our approach: (a) (**uniform sampling**) on an average, each iteration of our algorithm involves generating $\tilde{O}\left(\frac{d}{T}\right)$ samples from uniform distribution over a convex set. This is a well studied problem and several efficient algorithms are known for uniform sampling from various classes of convex sets. To derive concrete runtime bounds, we consider the special case of the action set $\mathcal{X}$ being a polytope with m constraints. By relying on the algorithm of Laddha, Lee, and Vempala [LLV20], we can generate a single sample in $\tilde{O}\left(m^2d^2 + (m+d)d^4\right)$ time. (b) (**Newton update**) The Newton update in our algorithm involves minimization of a convex objective. This objective can be minimized using plethora of convex optimization techniques that have been developed. For the special case of action set being a polytope with m constraints, this objective can be minimized in $\tilde{O}\left(m^2d + (m+d)d^3\right)$ time using interior point methods (IPM).
- Our algorithm is robust to model mis-specification. In particular, even if each loss function f_t is $O\left(T^{-1/2}\right)$ away from a convex, quadratic function, our algorithm achieves the optimal regret. This result can be improved in a straightforward fashion: suppose

Paper	Regret	Adversary	amortized time complexity of each iteration (dependence on d, T)
Hazan and Li [HL16]	$\tilde{O}\left(2^{d^4}(\log T)^{2d}T^{1/2}\right)$ (h.p)	adaptive	$O\left((\log T)^{\text{poly}(d)}\right)$
Bubeck, Lee, and Eldan [BLE17]	$\tilde{O}\left(d^{9.5}T^{1/2}\right)$ (h.p)	adaptive	$O\left(2^d\right)$
Bubeck, Lee, and Eldan [BLE17] (computationally efficient variant)	$\tilde{O}\left(d^{10.5}T^{1/2}\right)$ (h.p)	adaptive	$\tilde{O}\left(\text{poly}(dm)T\right)$
Saha and Tewari [ST11]	$\tilde{O}\left(d^{2/3}T^{2/3}\right)$ (exp)	oblivious	involves minimization of a self concordant function
Flaxman, Kalai, and McMahan [FKM04]	$\tilde{O}\left(d^{1/2}T^{3/4}\right)$ (exp)	oblivious	involves projecting a point onto a convex set
This chapter (instantiation for polytopes)	$\tilde{O}\left(d^{16}T^{1/2}\right)$ (h.p)	adaptive	$\tilde{O}\left(\frac{m^2d^3+(m+d)d^5}{T}\right)$ $+\tilde{O}\left((m+d)d^3+m^2d\right)$

Table 4.1: Comparison of various approaches for BCO with quadratic losses. "h.p", "exp" in the second column denote high probability and expected regret bounds respectively. m in the last column denotes the number of constraints in the polytope.

each f_t is ϵ_t close to a convex, quadratic function. Then our algorithm achieves the optimal regret as long as $\sum_{t=1}^{T} \epsilon_t = O\left(T^{1/2}\right)$.

4.3 Related Work

In this section, we present a review of bandit optimization that is necessarily incomplete but is relevant to the current work. Multi-armed bandits is perhaps the simplest and most well studied sub-case of bandit optimization. Several efficient and optimal algorithms have been proposed for this problem [AB+09; AB10; ALT15; Lee+20]. These algorithms first estimate the unknown loss function from one-point feedback, and then rely on Follow-the-Regularized-Leader (FTRL) framework with appropriate regularizer to choose the next action.

Moving beyond MAB, several recent works on bandit optimization have focused on BCO. For bounded, convex functions, Flaxman, Kalai, and McMahan [FKM04] and Kleinberg [Kle05] developed online gradient descent style algorithms which achieve $\tilde{O}\left(T^{5/6}\right)$ regret. Recent works of Bubeck, Lee, and Eldan [BLE17] and Hazan and Li [HL16] improved upon this result and developed algorithms which achieve the optimal $\tilde{O}\left(T^{1/2}\right)$ regret (also see Lattimore [Lat20] for information-theoretic upper bounds). However, these algorithms are computationally expensive. Moreover, the regret bounds of Hazan and Li [HL16] have exponential dependence on dimension. As previously mentioned, several works have studied sub-cases of BCO. The most popular among these sub-cases is bandit linear optimization. For this problem, Abernethy, Hazan, and Rakhlin [AHR09] provided the first efficient algorithm with optimal $O\left(T^{1/2}\right)$ regret in expectation (see Dani, Hayes, and Kakade [DHK07] for lower bounds on regret for linear losses). This algorithm uses one-point estimate of the gradient and relies on OMD with SCB of $\mathcal{X}$ as the regularizer to choose the next action. Subsequent works have attempted to develop algorithms which achieve optimal regret in high-probability [Bar+08; AR09]. However, this turned out to be a difficult problem. It is only recently that an efficient and optimal algorithm for this problem was designed [Lee+20]. A related line of work studied generalizations of linear bandits in euclidean space to the framework of Reproducing Kernel Hilbert Spaces (RKHS) [CPB19; TS20]. As an application of this general framework, Chatterji, Pacchiano, and Bartlett [CPB19] study convex quadratic losses. However, their algorithm, which is based on exponential weights update scheme, is computationally inefficient as it involves sampling from non log-concave distributions, which is NP-hard in general. Moving beyond linear losses, Flaxman, Kalai, and McMahan [FKM04] provided an algorithm with $O\left(T^{3/4}\right)$ regret for convex, Lipschitz loss functions. Saha and Tewari [ST11] provided an algorithm for convex, smooth loss functions with $\tilde{O}\left(T^{2/3}\right)$ regret. For strongly convex, smooth functions, Hazan and Levy [HL14] and Ito [Ito20] provide algorithms which achieve the optimal $\tilde{O}\left(T^{1/2}\right)$ regret (see Shamir [Sha13] for lower bounds on regret for strongly convex losses).

Another active line of research on bandit optimization has focused on handling weaker adversary models. One such popular model is the stochastic adversary model, where it is assumed that the loss functions seen by the learner are independent samples generated from an unknown but fixed distribution [LR85; AG12; Fil+10; Kve+20; Aga+11; Sri+09]. Recently, there has been a flurry of research on designing computationally efficient and optimal regret algorithms for this setting. However, these algorithms usually have poor performance in the stronger adversary model considered in this chapter. Yet another line of research on bandit optimization has focused on multi-point feedback models where the player can query each loss function at multiple points. Several recent works have designed efficient algorithms for this setting [ADX10; Duc+15; Sha17]. These works show that it is possible to achieve similar regret guarantees in this setting as in the full-information setting.

4.4 Regularized Bandit Newton Algorithm

In this section we describe our algorithm for BCO (see Algorithm 4). At a high level, our algorithm tries to estimate the missing information (*i.e.*, gradient and Hessian) about the unknown loss function and pass it to the OMD framework, which chooses the next action.

Gradient and Hessian estimation. To estimate the gradient and Hessian of f_t at $\mathbf{x}_t$, we rely on the following randomized sampling scheme. We first randomly sample a point from the uniform distirbution on a ellipsoid with mean $\mathbf{x}_t$ and whose covariance matrix depends on the Hessian estimates of the past loss functions $\{f_s\}_{s=1}^{t-1}$. Next, we get one-point feedback from the adversary about the loss value at the sampled point, and use it to estimate the gradient and Hessian (see lines 6-13 of Algorithm 4). Our choice of the covariance matrix ensures that the sampling scheme adapts to the geometry of the cumulative loss $\sum_{s=1}^{t-1} f_s(\mathbf{x})$. In particular, it reduces exploration along directions which are strongly convex, and increases exploration along directions which are linear. This choice of exploration helps us achieve the right balance between exploration and exploitation, and plays a crucial role in achieving optimal regret guarantees.

Focus Region. Once we have an estimate of the gradient and Hessian, we construct a quadratic approximation of f_t around $\mathbf{x}_t$ (see line 14 of Algorithm 4). One caveat with this approximation, however, is that it is *not* guaranteed to have a low variance. To see this, first note that the variance of our estimate $\hat{f}_t(\mathbf{x})$ scales with $\|\mathbf{x}-\mathbf{x}_t\|_{M_t}$ (look at line 6 for definition of M_t). If $\mathbf{x}_t$ gets too close to the boundary of $\mathcal{X}$, then $\|\nabla^2 R(\mathbf{x}_t)\|_2$ and $\|M_t\|_2$ become very large. This in turn increases the variance of $\hat{f}_t(\mathbf{x})$, for $\mathbf{x}$ far away from $\mathbf{x}_t$. Consequently, we can not directly plug in the estimate $\hat{f}_t(\mathbf{x})$ into the OMD framework to choose the next action. To handle this issue, we rely on focus regions. In each iteration of the algorithm, we maintain a focus region F_t which satisfies the following key property: *the variance of the quadratic approximation within F_t is small and bounded.* To this end, we choose an F_t such that $\|\mathbf{x}-\mathbf{x}_t\|_{M_t}$ is bounded for any $\mathbf{x}\in F_t$. When picking the next action $\mathbf{x}_{t+1}$ using OMD, we restrict ourselves to the focus region F_t.

At the beginning of the algorithm, we set F_1 to $\mathcal{X}_\xi$, a scaled version of $\mathcal{X}$, which is defined as $\mathcal{X}_\xi = \xi\mathbf{x}_1 + (1-\xi)\mathcal{X}$, where $\xi = T^{-4}$ and $\mathbf{x}_1$ is the minimizer of $R(\mathbf{x})$ over $\mathcal{X}$. We use $\mathcal{X}_\xi$ instead of $\mathcal{X}$ purely for theoretical reasons, as it simplifies our proofs. In practice, one can set F_1 to $\mathcal{X}$. To ensure F_t satisfies the above mentioned property on low variance, we perform a check in each iteration of the algorithm (see lines 20-25). Intuitively, this checks if the current focus region has a large overlap with $B_{\alpha,M_t}(\mathbf{x}_t)$, the region of low variance of the quadratic approximation. If yes, we do not change the focus region. If not, we shrink the focus region so that it overlaps with the low variance region. Moreover, we simultaneously increase the learning rate (η_t) of OMD. This learning rate change ensures that the algorithm can quickly adapt to any changes of the adversary. If the adversary attempts to move the minimizer of $\min_{\mathbf{x}\in\mathcal{X}_\xi}\sum_{s=0}^{t} f_s(\mathbf{x})$ outside of the focus region, then increasing the learning rate helps us quickly detect this change. This plays a crucial role in the restart condition, which we explain next. Several recent works have used the idea of increasing learning schedule for various purposes [Aga+17; BLE17; Lee+20].

Restart Condition. By relying on focus regions, we can only guarantee low regret w.r.t points within the focus region. To ensure low regret even w.r.t points outside the focus region, we perform another test every iteration, which we call "restart condition" (see lines 15-18). Intuitively, this test checks if the minimizer of $\min_{\mathbf{x}\in\mathcal{X}}\sum_{s=1}^{t} f_s(\mathbf{x})$ is well within the focus region. If yes, we continue the algorithm, as having a low regret w.r.t points in the focus region ensures the overall regret is low. If instead the test fails, then it usually implies that the minimizer is too close to the boundary of the focus region $\partial F_t \cap \operatorname{int}(\mathcal{X})$. In this case we show that the regret of our actions until now is negative. So, we can safely restart the algorithm. That is, we act as if time step $t+1$ is time step 1 and run the algorithm for $T-t$ steps.

We note that the ideas of focus region and restart condition appeared in the work of Bubeck, Lee, and Eldan [BLE17]. However, their approach is computationally expensive, even after restricting the loss functions to convex quadratics. There are two main reasons for this:

1. the algorithm of Bubeck, Lee, and Eldan [BLE17] relies on exponential weights update scheme. Each iteration of this algorithm involves generating $\tilde{O}(d)$ samples from an approximately log-concave distribution, which can be computationally expensive in high dimensions. In contrast, we rely on OMD framework in our work, which doesn't require

access to an approximately log-concave sampler.

2. the restart condition of Bubeck, Lee, and Eldan [BLE17] involves optimization of an approximately convex objective over a *non-convex* set. To be precise, the authors use the following restart condition

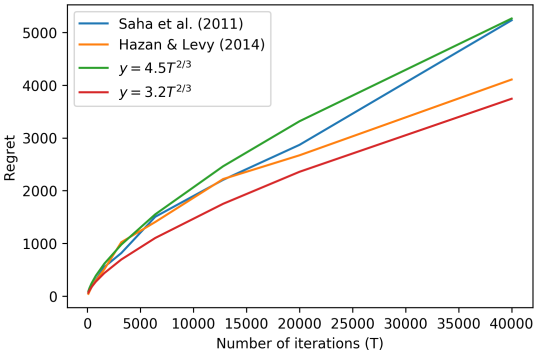
$$\min_{\mathbf{y}\in\partial F_t\cap\text{int}(\mathcal{X})}\sum_{s=0}^{t}\hat{f}_s(\mathbf{y}) - \min_{\mathbf{y}\in F_t}\sum_{s=0}^{t}\hat{f}_s(\mathbf{y}) \le \frac{1}{\eta_1}.$$

   Implementing this can be NP-hard in general because the domain of the first optimization is a nonconvex set. While the authors present a modified algorithm to handle this issue, it is still computationally expensive (the runtime of each iteration is $\tilde{O}\left(d^a T\right)$ for some large a). Moreover, the modified algorithm only works for constraint sets which are polytopes and whose coefficients in the constraints are rational numbers with absolute values of numerators and denominators bounded by poly(T). In contrast, the restart condition we use only involves minimization of $\min_{\mathbf{x}\in F_t}\sum_{s=0}^{t}\hat{f}_s(\mathbf{x})$, which we show is approximately convex and can be optimized efficiently (see Section 4.6).

4.4.1 Importance of Hessian Estimates

In this section we empirically demonstrate that existing OMD algorithms that only rely on gradient information don't achieve optimal regret bounds for quadratic losses [AHR09; ST11; HL14].

Lets consider a simple example where the adversary always selects the following loss function in each iteration: $f_t(\mathbf{x}) = \sum_{i=1}^{d/2} x_i^2 + \sum_{i=1}^{d} x_i$. Here, we choose the domain $\mathcal{X}$ to be $\mathbb{B}^d$. In this case, all the three algorithms mentioned above get sub-optimal regret of $\Omega(T^{2/3})$ (see image for empirical evidence). This is because M_t (defined in line 6 of Algorithm 4), which controls the exploration, is not chosen appropriately by these algorithms. Ideally, we should explore the first $d/2$ directions less and the last $d/2$ directions more. This is because the expected regret of these algorithms depends on the following term: $\mathbb{E}[f_t(\mathbf{y}_t) - f_t(\mathbf{x}_t)] = \mathbb{E}[f_t(\mathbf{x}_t + M_t^{-1/2}\mathbf{v}_t) - f_t(\mathbf{x}_t)] = \sum_{i=1}^{d/2}\mathbb{E}[(M_t^{-1/2}\mathbf{v}_t)_i^2]$. So a good choice of M_t should ensure $\mathbb{E}[(M_t^{-1/2}\mathbf{v}_t)_i^2]$ is low for the first $d/2$ coordinates. We achieve this in our algorithm by relying on Hessian estimates, which tell us how much exploration to do in each direction. For the example considered here, M_t in our algorithm is approximately equal to $\nabla^2 R(\mathbf{x}_t) + \sum_{s=0}^{t-1} 2\eta_s \nabla^2 f_s(\mathbf{x})$. For this choice of M_t, $\mathbb{E}[(M_t^{-1/2}\mathbf{v}_t)_i^2]$ goes down with t along the first $d/2$ directions. As a result, our algorithm performs less exploration along directions with large curvature, and more exploration along directions with small curvature, and achieves the optimal trade-off between exploration and exploitation. If we do uniform exploration in all directions (similar to existing algorithms), then we don't achieve the optimal regret.



4.5 Analysis

In this section we provide an outline of the proof of our main result stated in Theorem 9. We prove the following Theorem from which Theorem 9 follows readily.

Theorem 10 (Regret)**.** *Consider the setting of Theorem 9. Suppose Algorithm 4 is run for T iterations with the following hyper-parameters*

$$\lambda = \frac{1}{4},\ \alpha = c_1(\nu + d)d\log^2 dT,\ \beta = d\log dT,\ \gamma = \frac{c_2}{d\log T},\ \eta_1 = \frac{c_3}{d^7(B+\epsilon)\alpha^4\sqrt{T}\log T},$$

for some universal constants $c_1, c_2, c_3 > 0$. Let $\mathcal{T}$ be the minimum between T and the first time at which the algorithm restarts. Then with probability at least $1-\delta$

$$\sum_{t=1}^{\mathcal{T}} f_t(\mathbf{y}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{\mathcal{T}} f_t(\mathbf{x}) \leq \begin{cases} \tilde{O}\left(d^{11}(d+\nu)^5\sqrt{T}\right) & \text{if } \mathcal{T} = T \\ 0 \quad \text{otherwise} \end{cases}.$$

Proof. (Sketch) We first consider the case where the restart condition triggered for the first time at iteration $\mathcal{T} < T$. Then we show that the regret of the learner until $\mathcal{T}$ is negative. There are several key steps involved in showing this result:

1. We first show that the minimizer of the cumulative loss $\sum_{s=0}^{\mathcal{T}} f_s(\mathbf{x})$ over the entire domain $\mathcal{X}$ lies in $F_{\mathcal{T}}$; that is, $\min_{\mathbf{x}\in\mathcal{X}} \sum_{s=0}^{\mathcal{T}} f_s(\mathbf{x}) = \min_{\mathbf{x}\in F_{\mathcal{T}}} \sum_{s=0}^{\mathcal{T}} f_s(\mathbf{x})$. This immediately entails that the regret after $\mathcal{T}$ iterations satisfies.

$$\mathrm{Reg}_{\mathcal{T}} = \sum_{s=0}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in F_{\mathcal{T}}} \sum_{s=0}^{\mathcal{T}} f_s(\mathbf{x}).$$

2. Next, consider the following for any $\mathbf{x} \in F_{\mathcal{T}}$

$$\sum_{s=0}^{\mathcal{T}} f_s(\mathbf{y}_s) - \sum_{s=0}^{\mathcal{T}} f_s(\mathbf{x}) = \sum_{s=0}^{\mathcal{T}} [f_s(\mathbf{y}_s) - f_s(\mathbf{x}_s)] + \sum_{s=0}^{\mathcal{T}} \left[f_s(\mathbf{x}_s) - f_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) + \hat{f}_s(\mathbf{x})\right] + \sum_{s=0}^{\mathcal{T}} \left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right].$$

Recall $\mathbf{y}_t - \mathbf{x}_t = \lambda M_t^{-1/2}(\mathbf{v}_{1,t} + \mathbf{v}_{2,t})$. Relying on standard martingale concentration inequalities, the first term in the RHS above can be bounded as $\tilde{O}\left(d\eta_1^{-1}\right)$. To bound the second term, we rely on a key property of our loss estimates $\{\hat{f}_t\}_{t=1}^{T}$: the cumulative loss estimate concentrates well around the true cumulative loss (see Proposition 6). Using this property, the second term can be bounded as $O\left(\eta_1^{-1}\right)$. To bound the last term, we rely on the definition of restart condition which says that $\sum_{s=0}^{\mathcal{T}} \hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x}) \leq -\beta\eta_1^{-1}$. Combining these bounds shows that the regret after $\mathcal{T}$ iterations is negative.

Next, consider the case where the restart condition never triggered. Here, we can again show that the minimizer of the cumulative loss over the entire domain lies in the focus region F_T. So it suffices to bound $\sum_{s=0}^T f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in F_T}\sum_{s=0}^T f_s(\mathbf{x})$. Consider the same decomposition of regret as above. We use the same arguments as above to bound the first two terms in the decomposition. To bound the thrid term, we consider the following

$$\sum_{s=0}^{T}\left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right] = \sum_{s=0}^{T}\left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x}_{s+1})\right] + \sum_{s=0}^{T}\left[\hat{f}_s(\mathbf{x}_{s+1}) - \hat{f}_s(\mathbf{x})\right].$$

The first term in the RHS can be upper bounded using stability of the iterates $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{M_t}$ (in our proof we show that $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{M_t}$ is upper bounded by $\tilde{O}\left(\eta_t\right)$). The second term is the regret of Be-The-Regularized-Leader and can be upper bounded as $\tilde{O}\left(\eta_1^{-1}\right)$. Combining these two bounds, we show that the regret is $\tilde{O}\left(T^{1/2}\right)$. □

The proof of Theorem 10 relies on several crucial properties of the iterates produced by our algorithm. First, we need to ensure that the matrix M_t is positive definite and the iterates $\mathbf{y}_t$ produced by our algorithm lie within the domain $\mathcal{X}$. Second, we need to show that the algorithm is stable, *i.e.,* $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{M_t}$ is small. The following proposition plays a crucial role in showing these properties. It is concerned about concentration of the Hessian estimates $\{\hat{H}_t\}_{t=1}^T$, and the loss estimates $\{\hat{f}_t\}_{t=1}^T$ computed by the Algorithm.

Proposition 6. *Consider the setting of Theorem 10. Let $\mathcal{T}$ be the minimum between T and the first time at which the algorithm restarts. Then for any $t \leq \mathcal{T}$, the following properties hold with probability at least $1 - T^{-2}$*

- *Let $H_t = \frac{1}{2}(A_t + A_t^T)$ be the Hessian of $q_t(\mathbf{x})$, and let $\tilde{M}_t = \nabla^2 R(\mathbf{x}_t) + \sum_{s=0}^{t-1}\eta_s H_s$. Then M_t defined in line 6 of Algorithm 4 satisfies*

$$\|\tilde{M}_t^{-1/2}(\tilde{M}_t - M_t)\tilde{M}_t^{-1/2}\|_2 = O\left(\alpha^2\eta_1\lambda^{-2}d^5 B\sqrt{T\log(dT)}\right).$$

- *The cumulative loss estimate $\sum_{s=1}^t \hat{f}_s(\mathbf{x})$ satisfies*

$$\sup_{\mathbf{x}\in F_t}\left|\sum_{s=1}^{t}\eta_1(\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - q_s(\mathbf{x}) + q_s(\mathbf{x}_s))\right| \leq O\left(\alpha^2\eta_1\lambda^{-2}Bd^{4.5}\sqrt{T\log dT}\right).$$

4.6 Implementation

In this section, we discuss the implementation aspects of our algorithm.

Focus region update. To estimate the ratio $\frac{\mathrm{Vol}(F_t\cap B_{\alpha,M_{t+1}}(\mathbf{x}_{t+1}))}{\mathrm{Vol}(F_t)}$, we generate sufficiently many independent uniformly distributed samples in F_t and count what fraction of them fall in $F_t \cap B_{\alpha,M_{t+1}}(\mathbf{x}_{t+1})$. By sampling just $O(\log T)$ points, we can show that with probability at least $1 - \frac{1}{T^4}$, the focus region gets updated whenever the true ratio is less than $\frac{1}{4}$ and doesn't get updated whenever the true ratio is greater than $\frac{3}{4}$. The intermediate values don't effect our argument. Next, note that we need not generate the samples every iteration. It suffices to generate them only when the focus region gets updated. We can reuse the old samples in rest of the iterations. In Appendix C.5 we show that the focus region doesn't get updated more than $O\left(d\log T\right)$ times (see Lemma 49). So, over T iterations of the Algorithm, we only need to generate $O\left(d\log^2 T\right)$ samples.

As previously mentioned, uniform sampling from a convex set is a well studied problem. For the special case of the action set being a polytope with m constraints, we rely on the recent work of Laddha, Lee, and Vempala [LLV20] which uses Dikin walk for sampling. The authors show that the Dikin walk mixes in $O\left(d\bar{\nu}\right)$ steps, where $\bar{\nu}$ is the strong self concordant parameter of the set. For our problem, $\bar{\nu}$ is $O(m + O\left(d\log T\right))$ (this follows from the fact that each of our focus regions is an intersection of $O\left(d\log T\right)$ elliposoids and a polytope). Moreover, each iteration of Dikin walk takes $O\left(d^3\log T + md + d^2\right)$ time. So generating a single sample from uniform distribution in F_t takes $\tilde{O}\left(m^2d^2 + (m+d)d^4\right)$ time.

OMD Update. Our results in Appendix C.5 entail that the objective in line 19 is strictly convex (see Lemma 48). So we can use IPM to solve the objective. As a concrete example, lets again consider the case of action set being a polytope with m constraints. Since there are $O\left(d\log T\right)$ elliposoidal constrains and m linear constraints, the self concordant parameter of the entire objective is $m + O\left(d\log T\right)$. So, the number of Newton updates we perform is $\tilde{O}\left(m+d\right)$. Moreoever, performing each newton update takes $O\left(d^3\log T + md\right)$ time. So, the overall compute time of IPM is $\tilde{O}\left(m^2d + (m+d)d^3\right)$.

Restart Condition. Checking the restart condition involves minimizing $\sum_{s=0}^{t}\hat{f}_s(\mathbf{y})$ over the focus region F_t. We note that this need not be a convex function. However, it is pointwise close to the following convex function: $\sum_{s=0}^{t}\hat{f}_s(\mathbf{y}) + (d^2\alpha^2\eta_1)^{-1}(\mathbf{y}-\mathbf{x}_t)^T\nabla^2 R(\mathbf{x}_t)(\mathbf{y}-\mathbf{x}_t)$ (see Remark C.5.1 in Appendix for a discussion on the convexity of this objective). To see why this objective is pointwise close to $\sum_{s=0}^{t}\hat{f}_s(\mathbf{y})$, first note that our choice of F_t always ensures $\|\mathbf{y}-\mathbf{x}_t\|_{\nabla^2 R(\mathbf{x}_t)} \leq O\left(d\alpha\right)$ for any $\mathbf{y}\in F_t$ (see Lemma 48). So the modified objective is $O\left(\eta_1^{-1}\right)$ close to the original objective. So we can rely on IPM to solve the modified objective and obtain $O\left(\eta_1^{-1}\right)$-approximate solution to the original objective (note that an approximate solution suffices for our argument). The computational complexity of IPM in this case is same as the complexity of OMD update described above.

4.7 Discussion

In this chapter, we presented a new algorithm for bandit optimization with convex (approximately) quadratic functions. Our algorithm achieves the optimal regret rate of $\tilde{O}(\sqrt{T})$ and is computationally much more efficient than any other known algorithms for this problem. To obtain these results, we (i) estimate the Hessian of the loss functions and use it in a controlled fashion to minimize the effect of variance in this estimation and (ii) develop new algorithmic ideas to implement this efficiently.

While our work focuses on the convex quadratic setting, we believe our ideas can be extended to other convex, parameteric loss functions such as generalized linear models. However, extending the idea of using Hessian (or more generally k^{th} order derivatives for $k>1$) estimates to obtain efficient algorithms with optimal regret rates seems challenging, even for highly smooth functions as the estimates of higher order derivatives come with high variance and new ideas seem necessary to make effective use of them. This is an interesting future direction to explore. Finally, we believe the dimension dependence in our regret bound can improved to d^8 by tightening the Hessian concentration result in

Proposition 6. We base this claim on the results in Appendix C.4, where we show that Algorithm 4 achieves $\tilde{O}\left(d^{5.5}\sqrt{T}\right)$ regret when the Hessian of f_t is known to the learner ahead of round t.

Algorithm 4 Regularized Bandit Newton Algorithm

```
**Input:** $\nu$-self-concordant barrier $R$, initial learning rate $\eta_1$, number of iterations $T$, radius of initial focus region $\alpha$, learning rate increment $\gamma$, exploration parameter $\lambda$, $\beta$.
Denote $\hat{g}_0 = 0, \hat{H}_0 = 0, \eta_0 = 0, \xi = T^{-4}$
Let $\mathbf{x}_1 = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} R(\mathbf{x})$
Focus Region $F_1 = \mathcal{X}_\xi$, where $\mathcal{X}_\xi = \xi\mathbf{x}_1 + (1-\xi)\mathcal{X}$
**for** $t = 1 \ldots T$ **do**
    Let $M_t = \left(\nabla^2 R(\mathbf{x}_t) + \sum_{s=0}^{t-1} \eta_s \hat{H}_s\right)$.
    Sample $\mathbf{v}_{1,t}, \mathbf{v}_{2,t} \sim \mathbb{S}^{d-1}$, and compute $\mathbf{y}_t = \mathbf{x}_t + \lambda M_t^{-1/2}(\mathbf{v}_{1,t} + \mathbf{v}_{2,t})$
    **if** $\mathbf{y}_t \in \mathcal{X}$ **then**
        Play $\mathbf{y}_t$ and observe $f_t(\mathbf{y}_t)$
        Estimate gradient and Hessian of $f_t$ at $\mathbf{x}_t$ as
            $$\hat{g}_t = \lambda^{-1} d f_t(\mathbf{y}_t) M_t^{1/2} \mathbf{v}_{1,t}, \quad \hat{H}_t = \frac{\lambda^{-2}}{2} d^2 f_t(\mathbf{y}_t) M_t^{1/2} \left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right) M_t^{1/2}$$
    **else**
        Play $\mathbf{x}_t$ and set $\hat{g}_t = 0, \hat{H}_t = 0$.
    **end if**
    Let $\hat{f}_t(\mathbf{x}) = \langle \hat{g}_t - \hat{H}_t\mathbf{x}_t, \mathbf{x}\rangle + \frac{1}{2}\mathbf{x}^T \hat{H}_t \mathbf{x}$ be the quadratic approximation of $f_t$ at $\mathbf{x}_t$
    //restart condition
    **if** $\sum_{s=0}^{t} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_t} \sum_{s=0}^{t} \hat{f}_s(\mathbf{y}) \le -\frac{\beta}{\eta_1}$ **then**
        Restart
    **end if**
    Compute $\mathbf{x}_{t+1}$ using OMD
        $$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in F_t} \eta_t \langle \hat{g}_t, \mathbf{x}\rangle + \Phi_{R_{t+1}}(\mathbf{x}, \mathbf{x}_t).$$
    Here $\Phi_{R_{t+1}}$ is Bregman divergence w.r.t $R_{t+1}(\mathbf{x}) \overset{\text{def}}{=} R(\mathbf{x}) + \sum_{s=0}^{t} \frac{\eta_s}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s)$
    //Update focus region
    **if** $\mathrm{Vol}(F_t \cap B_{\alpha, M_{t+1}}(\mathbf{x}_{t+1})) \le \frac{1}{2}\mathrm{Vol}(F_t)$ **then**
        $F_{t+1} = F_t \cap B_{\alpha, M_{t+1}}(\mathbf{x}_{t+1})$ and $\eta_{t+1} = (1+\gamma)\eta_t$
    **else**
        $F_{t+1} = F_t$ and $\eta_{t+1} = \eta_t$
    **end if**
**end for**
```

Part III

Minimax Statistical Estimation

Chapter 5

Learning Minimax Estimators via Online Learning

Estimating the properties of a probability distribution is a fundamental problem in machine learning and statistics. In this problem, we are given observations generated from an unknown probability distribution P belonging to a class of distributions $\mathcal{P}$. Knowing $\mathcal{P}$, we are required to estimate certain properties of the unknown distribution P, based on the observations. Designing good and "optimal" estimators for such problems has been a fundamental subject of research in statistics. Over the years, statisticians have considered various notions of optimality to compare the performance of estimators and to aid their search of good estimators. Some popular notions of optimality include admissibility, minimax optimality, Bayesian optimality, asymptotic efficiency [Fer14; LC06]. Of these, minimax optimality is the most popular notion and has received wide attention in frequentist statistics. This notion of optimality has led to the minimax estimation principle, where the goal is to design estimators with the minimum worst-case risk. Let $R(\hat{\theta}, \theta(P))$ be the risk of an estimator $\hat{\theta}$ for estimating the property $\theta(P)$ of a distribution P, where an estimator is a function which maps observations to the set of possible values of the property. Then the worst-case risk of $\hat{\theta}$ is defined as $\sup_{P \in \mathcal{P}} R(\hat{\theta}, \theta(P))$. The goal in minimax estimation principle is to design estimators with worst-case risk close to the best worst-case risk, which is defined as $R^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} R(\hat{\theta}, \theta(P))$, where the infimum is computed over the set of all estimators. Such estimators are often referred to as minimax estimators [Tsy08].

Classical Estimators A rich body of work in statistics has focused on studying the minimax optimality properties of classical estimators such as the maximum likelihood estimator (MLE), Bayes estimators, and minimum contrast estimators (MCEs) [IH81; Le 12; Vaa98; Bir83; BM93; YB99]. Early works in this line have considered parametric estimation problems and focused on the asymptotic setting, where the number of observations approaches infinity, for a fixed problem dimension. In a series of influential works, Hájek and Le Cam showed that under certain regularity conditions on the parametric estimation problem, MLE is asymptotically minimax whenever the risk is measured with respect to a convex loss function [Le 12; IH81]. Later works in this line have considered both parametric and non-parametric estimation problems in the non-asymptotic setting and studied the minimax rates of estimation. In a series of works, Birgé [Bir83; BM93] showed that under certain regularity conditions on the model class $\mathcal{P}$ and the estimation problem, MLE and MCEs are approximately minimax w.r.t Hellinger distance.

While these results paint a compelling picture of classical estimators, we highlight two key problem settings where they tend to be rate inefficient (that is, achieve sub-optimal worst-case risk) [Wel15; BM93]. The first is the so-called high dimensional sampling setting, where the number of observations is comparable to the problem dimension, and under which, classical estimators can be highly sub-optimal. In some recent work, Jiao, Venkat, Han, and Weissman [Jia+15] considered the problem of entropy estimation in discrete distributions and showed that the MLE (plug-in rule) is sub-optimal in the high dimensional regime. Similarly, Cai and Low [CL11] considered the problem of estimation of non-smooth functional $\frac{1}{d}\sum_{i=1}^{d}|\theta_i|$ from an observation $Y \sim \mathcal{N}(\theta, I_d)$ and showed that the MLE is sub-optimal. The second key setting where classical estimators tend to be sub-optimal is when the risk $R(\hat{\theta}, \theta(P))$ is measured w.r.t "non-standard" losses that have a very different behavior compared to standard losses such as Kullback-Leibler (KL) divergence. For example, consider the MLE, which can be viewed as a KL projection of the empirical distribution of observations onto the class of distributions $\mathcal{P}$. By its design, we expect it to be minimax when the risk is measured w.r.t KL divergence and other related metrics such as Hellinger distance [BM93]. However, for loss metrics which are not aligned with KL, one can design estimators with better performance than MLE, by taking the loss into consideration. This phenomenon is better illustrated with the following toy example. Suppose $\mathcal{P}$ is the set of multivariate normal distributions in $\mathbb{R}^d$ with identity covariance, and suppose our goal is to estimate the mean of a distribution $P \in \mathcal{P}$, given n observations drawn from it. If the risk of estimating θ as $\tilde{\theta}$ is measured w.r.t the following loss $\|\tilde{\theta}-\theta-c\|_2^2$, for some constant c, then it is easy to see that MLE has a worst-case risk greater than $\|c\|_2^2$. Whereas, the minimax risk R^* is equal to d/n, which is achieved by an estimator obtained by shifting the MLE by c. While the above loss is unnatural, such a phenomenon can be observed with natural losses such as ℓ_q norms for $q \in (0, 1)$ and asymmetric losses.

Bespoke Minimax Estimators For problems where classical estimators are not optimal, designing a minimax estimator can be challenging. Numerous works in the literature have attempted to design minimax estimators in such cases. However the focus of these works is on specific problems [CL11; VV11; Jia+15; But+18], and there is no single estimator which is known to be optimal for a wide range of estimation problems. For example, Jiao, Venkat, Han, and Weissman [Jia+15] and Wu and Yang [WY16] considered the problem of entropy estimation for discrete distributions and provided a minimax estimator in the high-dimensional setting. Cai and Low [CL11] considered the problem of estimating a non-smooth functional in high dimensions and provided a minimax estimator. While these results are impressive, the techniques used in these works are tailored towards specific problems and do not extend to other problems. So, a natural question that arises in this context is, how should one go about constructing minimax estimators for problems where none of the classical estimators are optimal? Unfortunately, our current understanding of minimax estimators does not provide any concrete guidelines on designing such estimators.

Minimax Estimation via Solving Statistical Games In this chapter, we attempt to tackle the problem of designing minimax estimators from a game-theoretic perspective. Instead of the usual two-step approach of first designing an estimator and then certifying its minimax optimality, we take a more direct approach and attempt to directly solve the following min-max statistical game: $\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} R(\hat{\theta}, \theta(P))$. Since the resulting estimators are solutions to the min-max game, they are optimal by construction. Such a direct approach for construction of minimax estimators has certain advantages over the classical estimators. First, the technique itself is very general and can *theoretically* be used to construct minimax estimators for any estimation problem. Second, a direct approach often results in *exact* minimax estimators with $R^* + o(1)$ worst-case risk. In contrast, classical estimators typically achieve $O(1)R^*$ worst-case risk, which is constant factors worse than the direct approach. Finally, a direct approach can make effective use of any available side information about the problem, to construct estimators with better worst-case risk than classical estimators. For example, consider the problem of mean estimation given samples drawn from an unknown Gaussian distribution. If it is known a priori that the true mean lies in a bounded set, then a direct approach for solving the min-max statistical game results in estimators with better performance than classical estimators. Several past works have attempted to directly solve the min-max game associated with the estimation problem [see Ber85, and references therein]. We discuss these further in Section 5.1 after providing some background, but in gist, existing approaches either focus on specific problems or are applicable only to simple estimation problems.

This Chapter In this chapter, we rely on recent advances in online learning and game theory to directly solve the min-max statistical game. Recently, online learning techniques have been widely used for solving min-max games. For example, Freund and Schapire [FS96] relied on these techniques to find equilibria in min-max games that arise in the context of boosting. Similar techniques have been explored for robust optimization by Chen, Lucier, Singer, and Syrgkanis [Che+17] and Feige, Mansour, and Schapire [FMS15]. In this chapter, we take a similar approach and provide an algorithm for solving statistical games. A critical distinction of statistical games, in contrast to the typical min-max games studied in the learning and games literature, is that the domain of all possible measurable estimators is extremely large, the set of possible parameters need not be convex, and the loss function need not be convex-concave. We show that it is nonetheless possible to finesse these technical caveats and solve the statistical game, provided we are given access to two subroutines: a Bayes estimator subroutine which outputs a Bayes estimator corresponding to any given prior, and a subroutine which computes the worst-case risk of any given estimator. Given access to these two subroutines, we show that our algorithm outputs both a minimax estimator and a least favorable prior (LFP). The minimax estimator output by our algorithm is a randomized estimator which is an ensemble of multiple Bayes estimators. When the loss function is convex - which is the case for a number of commonly used loss functions - the randomized estimator can be transformed into a deterministic minimax estimator. For problems where the two subroutines are efficiently implementable, our algorithm provides an efficient technique to construct minimax estimators. While implementing the subroutines can be computationally hard in general, we show that the computational complexity can be significantly reduced for a wide range of problems satisfying certain invariance properties.

To demonstrate the power of this technique, we use it to construct provably minimax estimators for the classical problems of finite dimensional Gaussian sequence model and linear regression. In the problem of Gaussian sequence model, we are given a single sample drawn from a normal distribution with mean θ and identity covariance, where $\theta \in \mathbb{R}^d, \|\theta\|_2 \leq B$. Our goal is to estimate θ well under squared-error loss. This problem has received much attention in statistics because of its simplicity and connections to non-parametric regression [Joh02]. Surprisingly, however, the exact minimax estimator is unknown for the case when $B \geq 1.16\sqrt{d}$ [Bic81; Ber90; MP02]. In this chapter, we show that our technique can be used to construct provably minimax estimators for this problem, for general B. To further demonstrate that our technique is widely applicable, we present empirical evidence showing that our algorithm can be used to construct estimators for covariance and entropy estimation which match the performance of existing minimax estimators.

On Criticisms of Minimaxity Perhaps it is important to note that sometimes minimax estimators are deemed to be unnecessarily pessimistic, as they are driven by the worst case risk. Nonetheless they occupy a unique position in the statistical estimation literature. They have undoubtedly been a major source of intellectual curiosity as exemplified by many settings under which minimax estimators have already been constructed. In scenarios where other notions of optimality might be desired, minimax estimators and LFPs can still be used to validate the performance of the constructed estimators against worst case scenario. We could also use model selection over minimax estimators for varying subclasses of models to mitigate some of their undesirable properties. This makes constructing minimax estimators a worthy pursuit. Given the difficulty of constructing such estimators in new scenarios, it is important to find methods to automate this process. While previous work has tried to do this under very restricted assumptions [Nel66; Kem87], in this chapter we show the feasibility of this approach in a fairly general context. We are able to achieve this because of the recent advancements in online learning, as we highlight in later sections.

Outline We conclude this section with a brief outline of the rest of the paper. In Section 5.1, we provide necessary background on online learning and minimax estimation. In Section 5.2, we introduce our algorithm for solving statistical games. In Sections 5.3, 5.4, 5.5 we utilize our algorithm to construct provably minimax estimators for finite dimensional Gaussian sequence model and linear regression. In Section 5.8 we study the empirical performance of our algorithm on a variety of statistical estimation problems. We defer technical details to the Appendix. Finally, we conclude in Section 5.9 with a discussion of future directions and some open problems.

5.1 Background and Problem Setup

In this section, we formally introduce the problem of minimax statistical estimation and review the necessary background on online learning.

5.1.1 Minimax Estimation and Statistical Games

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ be a parametric family of distributions. In this chapter, we assume Θ is a compact set. Let $\mathbb{X}^n = \{X_1, \dots X_n\} \in \mathcal{X}^n$ be n independent samples drawn from some unknown distribution $P_\theta \in \mathcal{P}$. Given $\mathbb{X}^n$, our goal is to estimate the unknown parameter θ. A deterministic estimator $\hat{\theta}$ of θ is any measurable function from $\mathcal{X}^n$ to Θ. We denote the set of deterministic estimators by $\mathcal{D}$. A randomized estimator is given by a probability measure on the set of deterministic estimators. Given $\mathbb{X}^n$, the unknown parameter θ is estimated by first sampling a deterministic estimator according to this probability measure and using the sampled estimator to predict θ. Since any randomized estimator can be identified by a probability measure on $\mathcal{D}$, we denote the set of randomized estimators by $\mathcal{M}_{\mathcal{D}}$, the set of all probability measures on $\mathcal{D}$. Let $M : \Theta \times \Theta \to \mathbb{R}$ be a measurable loss function such that $M(\theta', \theta)$ measures the cost of an estimate θ' when the true parameter is θ. Define the risk of an estimator $\hat{\theta}$ for estimating θ as $R(\hat{\theta}, \theta) \overset{\text{def}}{=} \mathbb{E}\left[M(\hat{\theta}(\mathbb{X}^n), \theta)\right]$, where the expectation is taken with respect to randomness from $\mathbb{X}^n$ and the estimator $\hat{\theta}$. The worst-case risk of an estimator $\hat{\theta}$ is defined as $\sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$ and the minimax risk is defined as the best worst-case risk that can be achieved by any estimator

$$R^* \overset{\text{def}}{=} \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta). \tag{5.1}$$

Any estimator whose worst case risk is equal to the minimax risk is called a minimax estimator. We refer to the above min-max problem as a *statistical game*. Often, we are also interested in deterministic minimax estimators, which are defined as estimators with worst case risk equal to

$$\inf_{\hat{\theta} \in \mathcal{D}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta). \tag{5.2}$$

From the perspective of game theory, the optimality notion in Equation (5.1) is referred to as the *minmax* value of the game. This is to be contrasted with the *maxmin* value of the game $\sup_{\theta \in \Theta} \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta)$. In general, these two quantities are **not** equal, but the following relationship always holds:

$$\sup_{\theta \in \Theta} \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta) \leq \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta). \tag{5.3}$$

In statistical games, for typical choices of loss functions, $\sup_{\theta \in \Theta} \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta) = 0$, whereas $\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) > 0$; that is, the minmax value is strictly greater than maxmin value of the game. So we cannot in general reduce computing the minmax value to computing the maxmin value.

Linearized Statistical Games Without any additional structure such as convexity, computing the values of min-max games is difficult in general. So it is common in game theory to consider a *linearized game* in the space of probability measures, which is in general better-behaved. To set up some notation, for any probability distribution P, define $R(\hat{\theta}, P)$ as $\mathbb{E}_{\theta\sim P}\left[R(\hat{\theta},\theta)\right]$. In the context of statistical games, a linearized game has the following form:

$$\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}}\sup_{P\in\mathcal{M}_{\Theta}} R(\hat{\theta}, P), \tag{5.4}$$

where $\mathcal{M}_\Theta$ is the set of all probability measures on Θ. The minmax and maxmin values of the linearized game and the original game in Equation (5.1) are related as follows

$$\sup_{\theta\in\Theta}\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta},\theta) \le \sup_{P\in\mathcal{M}_{\Theta}}\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta},P) \le \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}}\sup_{P\in\mathcal{M}_{\Theta}} R(\hat{\theta},P) \overset{(a)}{=} \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}}\sup_{\theta\in\Theta} R(\hat{\theta},\theta),$$

where (a) holds because for any estimator $\hat{\theta}$, $\sup_{P\in\mathcal{M}_\Theta} R(\hat{\theta},P)$ is equal to $\sup_{\theta\in\Theta} R(\hat{\theta},\theta)$. Thus, the minmax values of the original and linearized statistical games are equal. Any estimator whose worst-case risk is equal to the minmax value of the linearized game is a minimax estimator. The maxmin values of the original and linearized statistical games are however in general different. In particular, as discussed above, the maxmin value of the original statistical game is usually equal to zero. The maxmin value of the *linearized game* however has a deep connection to Bayesian estimation.

Note that $R(\hat{\theta},P)$ is simply the integrated risk of the estimator $\hat{\theta}$ under prior $P \in \mathcal{M}_\Theta$. Any estimator which minimizes $R(\hat{\theta},P)$ is called the Bayes estimator for P, and the corresponding minimum value is called Bayes risk. Though the set of all possible measurable estimators is in general vast, in what might be surprising from an optimization or game-theoretic viewpoint, the Bayes estimator can be characterized simply as follows. Letting $P(\cdot|\mathbb{X}^n)$ be the posterior distribution of θ given the data $\mathbb{X}^n$, a Bayes estimator of P can be found by minimizing the posterior risk

$$\hat{\theta}_P(\mathbb{X}^n) \in \operatorname*{argmin}_{\tilde{\theta}\in\Theta} \mathbb{E}_{\theta\sim P(\cdot|\mathbb{X}^n)}\left[M(\tilde{\theta},\theta)\right]. \tag{5.5}$$

Certain mild technical conditions need to hold for $\hat{\theta}_P$ to be measurable and for it to be a Bayes estimator [Ber85]. We detail these conditions in Appendix D.1, which incidentally are all satisfied for the problems considered in this chapter. A least favourable prior is defined as any prior which maximizes the Bayes risk; that is, $\tilde{P}$ is LFP if $\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta},\tilde{P}) = \sup_{P\in\mathcal{M}_\Theta}\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta},P)$. Thus, LFPs solve for the maxmin value of the linearized statistical game. Any prior whose Bayes risk is equal to the maxmin value of the linearized game is an LFP.

Nash Equilibrium Directly solving for the minmax or maxmin values of the (linearized) min-max games is in general computationally hard, in large part because: (a) these values need not be equal, which limits the set of possible optimization algorithms, and (b) the optimal solutions need not be stable, which makes it difficult for simple optimization problems. It is thus preferable that the two values are equal[1], and the solutions be stable, which is formalized by the game-theoretic notion of a *Nash equilibrium* (NE).

For the original statistical game in Equation (5.1), a pair $(\hat{\theta}^*, \theta^*) \in \mathcal{M}_{\mathcal{D}} \times \Theta$ is called a pure strategy NE, if the following holds

$$\sup_{\theta \in \Theta} R(\hat{\theta}^*, \theta) \leq R(\hat{\theta}^*, \theta^*) \leq \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta^*) = \inf_{\hat{\theta} \in \mathcal{D}} R(\hat{\theta}, \theta^*),$$

where the equality follows since the optimum of a linear program over a convex hull can always be attained at an extreme point. Intuitively, this says that there is no incentive for any player to change their strategy while the other player keeps hers unchanged. Note that whenever a pure strategy NE exists, the minmax and maxmin values of the game are equal to each other:

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) \leq \sup_{\theta \in \Theta} R(\hat{\theta}^*, \theta) \leq R(\hat{\theta}^*, \theta^*) \leq \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta^*) \leq \sup_{\theta \in \Theta} \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, \theta).$$

Since the RHS is always upper bounded by the LHS from (5.3), the inequalities above are all equalities.

As we discussed above, the maxmin and minmax values of the statistical game in Equation (5.1) are in general not equal to each other, so that a pure strategy NE will typically not exist for the statistical game (5.1). Instead what often exists is a mixed strategy NE, which is precisely a pure strategy NE of the linearized game. That is, $(\hat{\theta}^*, P^*) \in \mathcal{M}_{\mathcal{D}} \times \mathcal{M}_{\Theta}$ is called a mixed strategy NE of statistical game (5.1), if

$$\sup_{\theta \in \Theta} R(\hat{\theta}^*, \theta) = \sup_{P \in \mathcal{M}_{\Theta}} R(\hat{\theta}^*, \theta) \leq R(\hat{\theta}^*, P^*) \leq \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, P^*) = \inf_{\hat{\theta} \in \mathcal{D}} R(\hat{\theta}, P^*).$$

As with the original game, if $(\hat{\theta}^*, P^*)$ is a pure strategy NE of the linearized game of (5.1), aka, a mixed strategy NE of (5.1), then the minmax and maxmin values of the linearized game are equal to each other, and, moreover $\hat{\theta}^*$ is a minimax estimator and P^* is an LFP. Conversely, if $\hat{\theta}^*$ is a minimax estimator, and P^* is an LFP, and the minmax and maxmin values of (5.4) are equal to each other, then $(\hat{\theta}^*, P^*)$ is a mixed strategy NE of (5.1). These just follow from similar sandwich arguments as with the original game, which we add for completeness in Appendix D.2.

In gist, it might be computationally easier to recover the mixed strategy NE of the statistical game, assuming they exist, and doing so, would recover minimax estimators and LFPs. In this chapter, we are thus interested in imposing mild conditions on the statistical game so that a mixed strategy NE exists, and under this setting, develop tractable algorithms to estimate the mixed strategy NE.

Existence of NE We now briefly discuss sufficient conditions for the existence of NE. As discussed earlier, a pure strategy NE does not exist for statistical games in general. So, here we focus on existence of mixed strategy NE. In a seminal work, Wald [Wal49]

[1] John Von Neumann, a founder of game theory, has said he could not foresee there even being a theory of games without a theorem that equates these two values

studied the conditions for existence of a mixed strategy NE, and showed that a broad class of statistical games have mixed strategy NE. Suppose every distribution in the model class $\mathcal{P}$ is absolutely continuous, Θ is compact, and the loss M is a bounded, non-negative function. Then minmax and maxmin values of the linearized game are equal. Moreover, a minimax estimator with worst-case risk equal to R^* exists. Under the additional condition of compactness of $\mathcal{P}$, [Wal49] showed that an LFP exists as well. Thus, based on our previous discussion, this implies the game has a mixed strategy NE. In this chapter, we consider a different and simpler set of conditions on the statistical game. We assume that Θ is compact and the risk $R(\hat{\theta}, \theta)$ is Lipschitz in its second argument. Under these assumptions, we show that the minmax and maxmin values of the linearized game in Equation (5.4) are equal to each other. Such results are known as minimax theorems and have been studied in the past [VMK07; Yan74; Wal49]. However, unlike past works that rely on fixed point theorems, we rely on a constructive learning-style proof to prove the minimax theorem, where we present an algorithm which outputs an approximate NE of the statistical game. Under the additional condition that the risk $R(\hat{\theta}, \theta)$ is bounded, we show that the statistical game has a minimax estimator and an LFP.

Computation of NE Next, we discuss previous numerical optimization techniques for computing a mixed strategy NE of the statistical game. Note that this is a difficult computational problem: minimizing over the domain of all possible estimators, and maximizing over the set of all probability measures on Θ. Nonetheless, several works in statistics have attempted to tackle this problem [Ber85]. One class of techniques involves reducing the set of estimators $\mathcal{D}$ via admissibility considerations to a small enough set. Given this restricted set of estimators, they can then directly calculate a minimax test for some testing problems; see for instance Hald [Hal71]. A drawback of these approaches is that they are restricted to simple estimation problems for which the set of admissible estimators are easy to construct. Another class of techniques for constructing minimax estimators relies on the properties of LFPs [CB94; Joh02]. When the parameter set Θ is a compact subset of $\mathbb{R}$, and when certain regularity conditions hold, it is well known that LFPs are supported on a finite set of points [Gho64; Ber85]. Based on this result, Nelson [Nel66] and Kempthorne [Kem87] propose numerical approaches to determine the support points of LFPs and the probability mass that needs to be placed on these points. However, these approaches are restricted to 1-dimensional estimation problems and are not broadly applicable. In a recent work, Luedtke, Carone, Simon, and Sofrygin [Lue+20] propose heuristic approaches for solving statistical games using deep learning techniques. In particular, they use neural networks to parameterize the statistical game and solve the resulting game using local search techniques such as alternating gradient descent. However, these approaches are not guaranteed to find minimax estimators and LFPs and can lead to undesirable equilibrium points. They moreover parameterize estimators via neural networks whose inputs are a simple concatenation of all the samples, which is not feasible for large n.

In our work, we develop numerical optimization techniques that rely on online learning algorithms. Though the domains as well as the setting of the statistical game are far more challenging than typically considered in learning and games literature, we reduce the problem of designing minimax estimators to a purely computational problem of efficient implementation of certain optimization subroutines. For the wide range of problems where these subroutines can be efficiently implemented, our algorithm provides an efficient and scalable technique for constructing minimax estimators.

5.2 Minimax Estimation via Online Learning

In this section, we present our algorithm for computing a mixed strategy NE of the statistical game in Equation (5.1) (equivalently a pure strategy NE of the linearized game in Equation (5.4)). A popular and widely used approach for solving min-max games is to rely on online learning algorithms [Haz16; CL06]. In this approach, the minimization player and the maximization player play a repeated game against each other. Both the players rely on online learning algorithms to choose their actions in each round of the game, with the objective of minimizing their respective regret. The following proposition shows that this repeated game play converges to a NE.

Proposition 7. *Consider a repeated game between the minimization and maximization players in Equation (5.4). Let $(\hat{\theta}_t, P_t)$ be the actions chosen by the players in iteration t. Suppose the actions are such that the regret of each player satisfies*

$$\sum_{t=1}^{T} R(\hat{\theta}_t, P_t) - \inf_{\hat{\theta}\in\mathcal{D}} \sum_{t=1}^{T} R(\hat{\theta}, P_t) \leq \epsilon_1(T),$$
$$\sup_{\theta\in\Theta} \sum_{t=1}^{T} R(\hat{\theta}_t, \theta) - \sum_{t=1}^{T} R(\hat{\theta}_t, P_t) \leq \epsilon_2(T).$$

Let $\hat{\theta}_{\text{RND}}$ denote the randomized estimator obtained by uniformly sampling an estimator from the iterates $\{\hat{\theta}_t\}_{t=1}^{T}$. Define the mixture distribution P_{AVG} as $\frac{1}{T}\sum_{i=1}^{T} P_i$. Then $(\hat{\theta}_{\text{RND}}, P_{\text{AVG}})$ is an approximate mixed strategy NE of Equation (5.1)

$$R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}) \leq \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG}}) + \frac{\epsilon_1(T) + \epsilon_2(T)}{T},$$
$$R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}) \geq \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND}}, \theta) - \frac{\epsilon_1(T) + \epsilon_2(T)}{T}.$$

Note that the above proposition doesn't specify an algorithm to generate the iterates $(\hat{\theta}_t, P_t)$. All it shows is that as long as both the players rely on algorithms which guarantee sub-linear regret, the iterates converge to a NE. There exist several algorithms such as FTRL, FTPL, Best Response (BR), which guarantee sub-linear regret. It is important to choose these algorithms appropriately as our choices impact the rate of convergence to a NE and also the computational complexity of the resulting algorithm. First, consider the minimization player, whose domain $\mathcal{M}_{\mathcal{D}}$ is the set of all probability measures over $\mathcal{D}$. Note that $\mathcal{D}$, the set of all deterministic estimators, is an infinite dimensional space. So, algorithms such as FTRL, FTPL, whose regret bounds depend on the dimension of the domain, can not guarantee sub-linear regret. So the minimization player is forced to rely on BR, which has 0 regret. Recall, in order to use BR, the minimization player requires the knowledge of the future action of the opponent. This can be made possible in the context of min-max games by letting the minimization player choose her action after the maximization player reveals her action. Next, consider the maximization player. Since the minimization player is relying on BR, the maximization player has to rely on either FTRL or FTPL to choose her action[2]. In this chapter, we choose the FTPL and OFTPL algorithms studied in Chapters 2, 3. Our choice is mainly driven by the computational aspects of the algorithm. Each iteration of the FTRL algorithm of Krichene, Balandat, Tomlin, and Bayen [Kri+15] involves sampling from a general probability distribution. Whereas, each iteration of the FTPL algorithm requires minimization of a non-convex objective. While both sampling and optimization are computationally hard in general, the folklore is that optimization is relatively easier than sampling in many practical applications.

We now describe our algorithm for computing a pure strategy NE of Equation (5.4). In iteration t, the maximization player chooses distribution P_t using FTPL. P_t is given by the distribution of the random variable $\theta_t(\sigma)$, which is generated by first sampling a random vector $\sigma \in \mathbb{R}^d$ from exponential distribution and then computing an optimizer of

$$\sup_{\theta\in\Theta} \sum_{i=1}^{t-1} R(\hat{\theta}_i, \theta) + \langle \sigma, \theta \rangle. \tag{5.6}$$

The minimization player chooses $\hat{\theta}_t$ using BR, which involves computing a minimizer of the integrated risk under prior P_t

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_t). \tag{5.7}$$

Very often, computing exact optimizers of the above problems is infeasible. Instead, one can only compute approximate optimizers. To capture the error from this approximation, we introduce the notion of approximate optimization oracles/subroutines.

Definition 5.2.1 (Maximization Oracle)**.** A function $\mathcal{O}^{\max}_{\alpha,\beta}(\cdot)$ is called (α, β)-approximate maximization oracle, if for any set of estimators $\{\hat{\theta}_i\}_{i=1}^T$ and perturbation σ, it returns $\theta' \in \Theta$ which satisfies the following inequality

$$\sum_{i=1}^{T} R\left(\theta', \theta\right) + \langle \sigma, \theta' \rangle \geq \sup_{\theta\in\Theta} \sum_{i=1}^{T} R(\hat{\theta}_i, \theta) + \langle \sigma, \theta \rangle - \left(\alpha + \beta\|\sigma\|_1\right).$$

We denote the output θ' by $\mathcal{O}^{\max}_{\alpha,\beta}\left(\{\hat{\theta}_i\}_{i=1}^T, \sigma\right)$.

Definition 5.2.2 (Minimization Oracle)**.** A function $\mathcal{O}^{\min}_{\alpha}(\cdot)$ is called α-approximate minimization oracle, if for any probability measure P, it returns an approximate Bayes estimator $\hat{\theta}'$ which satisfies the following inequality

$$R(\hat{\theta}', P) \leq \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P) + \alpha.$$

We denote the output $\hat{\theta}'$ by $\mathcal{O}^{\min}_{\alpha}(P)$.

[2] If both the players use BR, then both will wait for the other player to pick an action first. As a result, the algorithm will never proceed.

Algorithm 5 FTPL for statistical games

```
1: Input: Parameter of exponential distribution $\eta$, approximate optimization oracles $\mathcal{O}^{\max}_{\alpha,\beta}(\cdot), \mathcal{O}^{\min}_{\alpha'}(\cdot)$ for problems (5.6), (5.7) respectively
2: for $t = 1 \dots T$ do
3:     Let $P_t$ be the distribution of random variable $\theta_t(\sigma)$, which is generated as follows:
         (i) Generate a random vector $\sigma$ such that $\{\sigma_j\}_{j=1}^{d} \overset{i.i.d}{\sim} \text{Exp}(\eta)$
         (ii) Compute $\theta_t(\sigma)$ as
               $\theta_t(\sigma) = \mathcal{O}^{\max}_{\alpha,\beta}\left(\{\hat{\theta}_i\}_{i=1}^{t-1}, \sigma\right).$
4:     Compute $\hat{\theta}_t$ as
               $\hat{\theta}_t = \mathcal{O}^{\min}_{\alpha'}(P_t).$
5: end for
6: Output: $\{\hat{\theta}_1, \dots \hat{\theta}_T\}, \{P_1, \dots P_T\}$.
```

Given access to subroutines $\mathcal{O}^{\max}_{\alpha,\beta}(\cdot), \mathcal{O}^{\min}_{\alpha'}(\cdot)$ for approximately solving the optimization problems in Equations (5.6), (5.7), the algorithm alternates between the maximization and minimization players who choose P_t and $\hat{\theta}_t$ in each iteration. We summarize the overall algorithm in Algorithm 5. The following theorem shows that Algorithm 5 converges to an approximate NE of the statistical game.

Theorem 11 (Approximate NE). *Consider the statistical game in Equation (5.1). Suppose $\Theta \subseteq \mathbb{R}^d$ is compact with ℓ_∞ diameter D, i.e., $D = \sup_{\theta_1,\theta_2\in\Theta} \|\theta_1 - \theta_2\|_\infty$. Suppose $R(\hat{\theta}, \theta)$ is L-Lipschitz in its second argument w.r.t ℓ_1 norm:*

$$\forall \hat{\theta}, \theta_1, \theta_2 \quad |R(\hat{\theta}, \theta_1) - R(\hat{\theta}, \theta_2)| \le L\|\theta_1 - \theta_2\|_1.$$

Suppose Algorithm 5 is run for T iterations with approximate optimization subroutines $\mathcal{O}^{max}_{\alpha,\beta}(\cdot)$, $\mathcal{O}^{min}_{\alpha'}(\cdot)$ for solving the maximization and minimization problems. Let $\hat{\theta}_{\text{RND}}$ be the randomized estimator obtained by uniformly sampling an estimator from the iterates $\{\hat{\theta}_t\}_{t=1}^{T}$. Define the mixture distribution P_{AVG} as $\frac{1}{T}\sum_{i=1}^{T} P_i$. Then $(\hat{\theta}_{\text{RND}}, P_{\text{AVG}})$ is an approximate mixed strategy NE of the statistical game in Equation (5.1)

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND}}, \theta) - \epsilon \le R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}) \le \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG}}) + \epsilon,$$

where $\epsilon = O\left(\eta d^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha'\right)$.

As an immediate consequence of Theorem 11, we show that the minmax and maxmin values of the statistical game in Equation (5.4) are equal to each other. Moreover, when the risk is bounded, we show that the statistical game (5.1) has minimax estimators and LFPs.

Corollary 3 (Minimax Theorem). *Consider the setting of Theorem 11. Then*

$$\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \sup_{P\in\mathcal{M}_\Theta} R(\hat{\theta}, P) = \sup_{P\in\mathcal{M}_\Theta} \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, P) =: R^*.$$

Furthermore, suppose the risk $R(\hat{\theta}, \theta)$ is a bounded function and Θ is compact w.r.t the following metric: $\Delta_M(\theta_1, \theta_2) = \sup_{\theta\in\Theta} |M(\theta_1, \theta) - M(\theta_2, \theta)|$. Then there exists a minimax *estimator $\hat{\theta}^* \in \mathcal{M}_{\mathcal{D}}$ whose worst-case risk satisfies*

$$\sup_{\theta\in\Theta} R(\hat{\theta}^*, \theta) = R^*,$$

and there exists a least favorable prior $P^* \in \mathcal{M}_\Theta$ whose Bayes risk satisfies

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P^*) = R^*.$$

We note that the assumption on compactness of Θ w.r.t Δ_M is mild and holds whenever Θ is compact w.r.t ℓ_2 norm and M is a continuous function. As another consequence of Theorem 11, we show that Algorithm 5 outputs approximate minimax estimators and LFPs.

Corollary 4. *Consider the setting of Theorem 11. Suppose Algorithm 5 is run with $\eta = \sqrt{\frac{1}{dL^2T}}$. Then the worst-case risk of $\hat{\theta}_{\text{RND}}$ satisfies*

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND}}, \theta) \leq R^* + O(d^{\frac{3}{2}}LT^{-\frac{1}{2}} + \alpha + \alpha' + \beta d^{\frac{3}{2}}LT^{\frac{1}{2}}).$$

Moreover, P_{AVG} is approximately least favorable with the associated Bayes risk satisfying

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG}}) \geq R^* - O(d^{\frac{3}{2}}LT^{-\frac{1}{2}} + \alpha + \alpha' + \beta d^{\frac{3}{2}}LT^{\frac{1}{2}}).$$

In addition, suppose the loss M used in the computation of risk is convex in its first argument. Let $\hat{\theta}_{\text{AVG}}$ be the deterministic estimator which is equal to the mean of the probability distribution associated with $\hat{\theta}_{\text{RND}}$. Then the worst-case risk of $\hat{\theta}_{\text{AVG}}$ satisfies

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{AVG}}, \theta) \leq R^* + O(d^{\frac{3}{2}}LT^{-\frac{1}{2}} + \alpha + \alpha' + \beta d^{\frac{3}{2}}LT^{\frac{1}{2}}),$$

and $\hat{\theta}_{\text{AVG}}$ is an approximate Bayes estimator for prior P_{AVG}.

Remark 5.2.1 (Near Optimal Estimator)**.** *Corollary 4 shows that when the approximation error of the optimization oracles is sufficiently small and when T is large enough, Algorithm 5 outputs a minimax estimator with worst-case risk $(1+o(1))R^*$. This improves upon the approximate minimax estimators that are usually designed in statistics, which have a worst-case risk of $O(1)R^*$.*

Remark 5.2.2 (Deterministic Minimax Estimators)**.** *For general non-convex loss functions, Algorithm 5 only provides a randomized minimax estimator. Given this, a natural question that arises is whether there exist efficient algorithms for finding a deterministic minimax estimator. Unfortunately, even with access to the optimization subroutines used by Algorithm 5, finding a deterministic minimax estimator can be NP-hard [see Theorem 9 of Che+17]*

Remark 5.2.3 (Implementation Details)**.** *Note that the estimators $\{\hat{\theta}_i\}_{i=1}^T$ and distributions $\{P_i\}_{i=1}^T$ output by Algorithm 5 are infinite dimensional objects and can not in general be stored using finitely many bits. However, in practice, we use independent samples generated from P_i as its proxy and only work with these samples. Since $\hat{\theta}_i$ is a Bayes estimator* *for prior P_i, it can be approximately computed using samples from P_i. This process of approximating P_i with its samples introduces some approximation error and the number of samples used in this approximation need to be large enough to ensure Algorithm 5 returns a minimax estimator. For the problems of finite Gaussian sequence model and linear regression studied in Sections 5.4, 5.5, we show that poly(d) samples suffice to ensure a minimax estimator.*

Remark 5.2.4 (Computation of the Oracles)**.** *We now consider the computational aspects involved in the implementation of optimization oracles used by Algorithm 5. Recall that the maximization oracle, given any estimator, computes its worst-case risk with some linear perturbation. Since this objective could potentially be non-concave, maximizing it can take exponential time in the worst-case. And recall that the minimization oracle computes the Bayes estimator given some prior distribution. Implementation of this minimization oracle can also be computationally expensive in the worst case. While the worst case complexities are prohibitive, for a number of problems, one can make use of* the problem structure *to efficiently implement these oracles in polynomial time.*

In particular, we leverage symmetry and invariance properties of the statistical games to reduce the complexity of optimization oracles, while controlling their approximation errors; see Section 5.3. We further consider the case where there is no structure in the problem, other than the existence of finite-dimensional sufficient statistics for the statistical model. This allows one to reduce the computational complexity of the minimization oracle by replacing the optimization over $\mathcal{D}$ in Equation (5.7) with universal function approximators such as neural networks. Moreover, one can use existing global search techniques to implement the maximization oracle. While such a heuristic approach can reduce the computational complexity of the oracles, bounding their approximation errors can be hard (recall, the worst-case risk of our estimator depends on the approximation error of the optimization oracles). Nevertheless, in later sections, we empirically demonstrate that the estimators from this approach have superior performance over many existing estimators which are known to be approximately minimax.

We briefly discuss some classical work that can be leveraged for efficient implementation of optimization oracles, albeit for specific models or settings. For several problems, it can be shown that there exists an approximate minimax estimator in some restricted space of estimators such as linear or polynomial functions of the data [DLM90; CL11; PW19]. Such results can be used to reduce the space of estimators in the statistical game (5.1). By replacing $\mathcal{M}_{\mathcal{D}}$ in Equation (5.1) with the restricted estimator space, one can greatly reduce the computational complexity of the optimization oracles. Another class of results relies on analyses of convergence of posterior distributions. As a key instance, when the number of samples n is much larger than the dimension d, it is well known that the posterior distribution behaves like a normal distribution, whenever the prior has sufficient mass around the true parameter [Har83]. Such a property can be used to efficiently implement the minimization oracle.

5.3 Invariance of Minimax Estimators and LFPs

In this section, we show that whenever the statistical game satisfies certain invariance properties, the computational complexity of the optimization oracles required by Algorithm 5 can be greatly reduced. We first present a classical result from statistics about the invariance properties of minimax estimators.When the statistical game in Equation (5.2) is invariant to group transformations, the *invariance theorem* says that there exist minimax estimators which are also invariant to these group transformations [Kie+57; Ber85]. Later, we utilize this result to reduce the computational complexity of the oracles required by Algorithm 5.

We first introduce the necessary notation and terminology to formally state the invariance theorem. We note that the theorem stated here is tailored for our setting and more general versions of the theorem can be found in Kiefer et al. [Kie+57]. Let G be a compact group of transformations on $\mathcal{X} \times \Theta$ which acts component wise; that is, for each $g \in G$, $g(X, \theta)$ can be written as $(g_1X, g_2\theta)$, where g_1, g_2 are transformations on $\mathcal{X}, \Theta$. With a slight abuse of notation we write $gX, g\theta$ in place of $g_1X, g_2\theta$. We assume that the group action is continuous, so that the functions $(g, X) \to gX$ and $(g, \theta) \to g\theta$ are continuous. Finally, let μ be the unique left Haar measure on G with $\mu(G) = 1$. We now formally define "invariant statistical games", "invariant estimators" and "invariant probability measures".

Definition 5.3.1 (Invariant Game)**.** A statistical game is invariant to group transformations G, if the following two conditions hold for each $g \in G$

- for all $\theta \in \Theta$, $g\theta \in \Theta$. Moreover, the probability distribution of gX is $P_{g\theta}$, whenever the distribution of X is P_θ.
- $M(g\theta_1, g\theta_2) = M(\theta_1, \theta_2)$, for all $\theta_1, \theta_2 \in \Theta$.

Definition 5.3.2 (Invariant Estimator)**.** A deterministic estimator $\hat{\theta}$ is invariant if for each $g \in G$, $\hat{\theta}(g\mathbb{X}^n) = g\hat{\theta}(\mathbb{X}^n)$, where $g\mathbb{X}^n = \{gX_1, \dots gX_n\}$.

Definition 5.3.3 (Invariant Measure)**.** Let $\mathcal{B}(\Theta)$ be the Borel σ-algebra corresponding to the parameter space Θ. A measure ν on $(\Theta, \mathcal{B}(\Theta))$ is invariant if for all $g \in G$ and any measurable set $A \in \mathcal{B}(\Theta)$, $\nu(gA) = \nu(A)$.

Example 5.3.1. Consider the problem of estimating the mean of a Gaussian distribution. Given n samples $X_1, \dots X_n$ drawn from $\mathcal{N}(\theta, I_{d\times d})$, our goal is to estimate the unknown parameter θ. Suppose the parameter space is given by $\Theta = \{\theta' : \|\theta'\|_2 \leq B\}$ and the risk of any estimator is measured w.r.t squared L_2 loss. Then it is easy to verify that the problem is invariant to transformations of the orthogonal group $\mathbb{O}(d) = \{U : UU^T = U^TU = I\}$.

We now present the main result concerning the existence of invariant minimax estimators.A more general version of the result can be found in [Kie+57].

Theorem 12 (Invariance)**.** *Consider the statistical game in Equation (5.1). Suppose the game is invariant to group transformations G. Suppose the loss metric M is convex in its first argument. Then for any deterministic estimator $\hat{\theta}$, there exists an estimator $\hat{\theta}_G$ which is invariant to group transformations G, with worst-case risk no larger than the worst-case risk of $\hat{\theta}$*

$$\sup_{\theta\in\Theta} R(\hat{\theta}_G, \theta) \leq \sup_{\theta\in\Theta} R(\hat{\theta}, \theta).$$

This shows that there exists a minimax estimator which is invariant to group transformations. We now utilize this invariance property to reduce the complexity of the optimization oracles. Let $\Theta = \bigcup_\beta \Theta_\beta$ be the partitioning of Θ into equivalence classes under the equivalence $\theta_1 \sim \theta_2$, if $\theta_1 = g\theta_2$ for some $g \in G$. The quotient space of Θ is defined as the set of equivalence classes of the elements of Θ under the above defined equivalence and is given by $\Theta/G = \{\Theta_\beta\}_\beta$. For an invariant estimator $\hat{\theta}$, we define $R_G(\hat{\theta}, \Theta_\beta)$ as $R(\hat{\theta}, \theta_\beta)$ for any $\theta_\beta \in \Theta_\beta$. Note that this is well defined because for invariant estimators $R(\hat{\theta}, \theta_1) = R(\hat{\theta}, \theta_2)$ whenever $\theta_1 \sim \theta_2$ (see Lemma 52). Our main result shows that Equation (5.1) can be reduced to the following simpler objective

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{\Theta_\beta \in \Theta/G} R_G(\hat{\theta}, \Theta_\beta), \tag{5.8}$$

where $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to group transformations G. This shows that the outer minimization over the set of all estimators in Equation (5.1) can be replaced with a minimization over just the invariant estimators. Moreover, the inner maximization over the entire parameter space Θ can be replaced with a maximization over the smaller quotient space Θ/G , which in many examples we study here is a one or two-dimensional space, irrespective of the dimension of Θ.

Theorem 13. *Suppose the statistical game in Equation (5.1) is invariant to group transformations G. Moreover, suppose the loss metric M is convex in its first argument. Then,*

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{\Theta_\beta \in \Theta/G} R_G(\hat{\theta}, \Theta_\beta).$$

Moreover, given any ϵ-approximate mixed strategy NE of the reduced statistical game (5.8), one can reconstruct an ϵ-approximate mixed strategy NE of the original statistical game (5.1).

We now demonstrate how Theorem 13 can be used on a variety of fundamental statistical estimation problems.

5.3.1 Finite Gaussian Sequence Model

In the finite Gaussian sequence model, we are given a single sample $X \in \mathbb{R}^d$ sampled from a Gaussian distribution $\mathcal{N}(\theta, I)$. We assume the parameter θ has a bounded L_2 norm and satisfies $\|\theta\|_2 \leq B$. Our goal is to design an estimator for θ which is minimax with respect to squared-error loss. This results in the following min-max problem

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\|\theta\|_2 \leq B} R(\hat{\theta}, \theta) \equiv \mathbb{E}_{X \sim \mathcal{N}(\theta, I)} \left[\|\hat{\theta}(X) - \theta\|_2^2 \right]. \tag{5.9}$$

Theorem 14. *Let $\mathbb{O}(d) = \{U : UU^T = U^T U = I\}$ be the group of $d \times d$ orthogonal matrices with matrix multiplication as the group operation. The statistical game in Equation (5.9) is invariant under the action of $\mathbb{O}(d)$, where the action of $g \in \mathbb{O}(d)$ on (X, θ) is defined as $g(X, \theta) = (gX, g\theta)$. Moreover, the quotient space $\Theta/\mathbb{O}(d)$ is homeomorphic to the real interval $[0, B]$ and the reduced statistical game is given by*

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{b \in [0,B]} R(\hat{\theta}, b\mathbf{e}_1), \tag{5.10}$$

where $\mathbf{e}_1$ is the first standard basis vector in $\mathbb{R}^d$ and $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to the actions of orthogonal group.

The theorem shows that the supremum in the reduced statistical game (5.8) is over a bounded interval on the real line. So the maximization oracle in this case can be efficiently implemented using grid search over the interval $[0, B]$. In Section 5.4 we use this result to obtain estimators for Gaussian sequence model which are provably minimax and can be computed in polynomial time.

Estimating a few co-ordinates Here, we again consider with the Gaussian sequence model described above, but we are now interested in the estimation of only a subset of the co-ordinates of θ. Without loss of generality, we assume these are the first k coordinates. The loss M is the squared L_2 loss on the first k coordinates. The following Theorem presents the invariance properties of this problem. It relies on the group $\mathbb{O}(k) \times \mathbb{O}(d-k)$, which is defined as the set of orthogonal matrices of the form $g = \begin{bmatrix} g_1 & 0 \\ 0 & g_2 \end{bmatrix}$ where $g_1 \in \mathbb{O}(k)$ and $g_2 \in \mathbb{O}(d-k)$.

Theorem 15. *The statistical game described above is invariant under the action of the group $\mathbb{O}(k) \times \mathbb{O}(d-k)$. Moreover, the quotient space $\Theta/\mathbb{O}(k) \times \mathbb{O}(d-k)$ is homeomorphic to the ball of radius B centered at origin in $\mathbb{R}^2$ and the reduced statistical game is given by*

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{b_1^2 + b_2^2 \leq B^2} R(\hat{\theta}, [b_1 \mathbf{e}_{1,k}, b_2 \mathbf{e}_{1,d-k}]), \tag{5.11}$$

where $\mathbf{e}_{1,k}$ is the first standard basis vector in $\mathbb{R}^k$ and $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to the actions of orthogonal group.

5.3.2 Linear Regression

In the problem of linear regression with random design we are given n independent samples $D_n = \{(X_i, Y_i)\}_{i=1}^n$ generated from a linear model $Y_i = X_i^T \theta^* + \epsilon_i$, where $X_i \sim \mathcal{N}(0, I)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$. We assume the true regression vector is bounded and satisfies $\|\theta^*\|_2 \leq B$. Our goal is to design minimax estimator for estimating θ^* from D_n, w.r.t squared error loss. This leads us to the following min-max problem

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{||\theta||_2 \leq B} R(\hat{\theta}, \theta) \equiv \mathbb{E}_{D_n} \left[||\hat{\theta}(D_n) - \theta||_2^2 \right]. \tag{5.12}$$

Theorem 16. *The statistical game in Equation (5.12) is invariant under the action of the orthogonal group $\mathbb{O}(d)$, where the action of $g \in \mathbb{O}(d)$ on $((X, Y), \theta)$ is defined as $g((X, Y), \theta) = ((gX, Y), g\theta)$. Moreover, the quotient space $\Theta/\mathbb{O}(d)$ is homeomorphic to the interval $[0, B]$ and the reduced statistical game is given by*

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{b \in [0, B]} R(\hat{\theta}, b\mathbf{e}_1), \tag{5.13}$$

where $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to the actions of orthogonal group.

5.3.3 Normal Covariance Estimation

In the problem of normal covariance estimation we are given n independent samples $\mathbb{X}^n = \{X_i\}_{i=1}^n$ drawn from $N(0, \Sigma)$. Here, we assume that the true Σ has a bounded operator norm and satisfies $\|\Sigma\|_2 \leq B$. Our goal is to construct an estimator for Σ which is minimax w.r.t the entropy loss, which is defined as

$$M(\Sigma_1, \Sigma_2) = \mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2\right) - \log|\Sigma_1^{-1}\Sigma_2| - d.$$

This leads us to the following min-max problem

$$\inf_{\hat{\Sigma}\in\mathcal{M}_{\mathcal{D}}} \sup_{\Sigma\in\Xi} R(\hat{\Sigma}, \Sigma) \equiv \mathbb{E}_{\mathbb{X}^n}\left[M(\hat{\Sigma}(\mathbb{X}^n), \Sigma)\right], \tag{5.14}$$

where $\Xi = \{\Sigma : ||\Sigma||_2 \leq B\}$.

Theorem 17. *The statistical game defined by normal covariance estimation with entropy loss is invariant under the action of the orthogonal group $\mathbb{O}(d)$, where the action of $g \in \mathbb{O}(d)$ on (X, Σ) is defined as $g(X_i, \Sigma) = (gX_i, g\Sigma g^T)$. Moreover the quotient space $\Xi/\mathbb{O}(d)$ is homeomorphic to $\Xi_G = \{\lambda \in \mathbb{R}^d : B \geq \lambda_1 \geq \dots \lambda_d > 0\}$ and the reduced statistical game is given by*

$$\inf_{\hat{\Sigma}\in\mathcal{M}_{\mathcal{D},G}} \sup_{\lambda\in\Xi_G} R(\hat{\Sigma}, Diag(\lambda)), \tag{5.15}$$

where $Diag(\lambda)$ is the diagonal matrix whose diagonal entries are given by λ and $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to the actions of orthogonal group.

The theorem shows that the maximization problem over Ξ can essentially be reduced to an optimization problem over a d-dimensional space.

5.3.4 Entropy estimation

In the problem of entropy estimation, we are given n samples $\mathbb{X}^n = \{X_1, \dots X_n\}$ drawn from a discrete distribution $P = (p_1, \dots p_d)$. Here, the domain of each X_i is given by $\mathcal{X} = \{1, 2, \dots d\}$. Our goal is to estimate the entropy of P, which is defined as $f(P) = -\sum_{i=1}^d p_i \log_2 p_i$, under the squared error loss. This leads us to the following min-max problem

$$\inf_{\hat{f}\in\mathcal{M}_{\mathcal{D}}} \sup_{P\in\mathcal{P}} R(\hat{f}, P) \equiv \mathbb{E}_{\mathbb{X}^n}\left[\left(\hat{f}(\mathbb{X}^n) - f(P)\right)^2\right], \tag{5.16}$$

where $\mathcal{P}$ is the set of all probability distributions supported on d elements.

Theorem 18. *The statistical game in Equation (5.16) is invariant to the action of the permutation group $\mathbb{S}_d$. The quotient space $\mathcal{P}/\mathbb{S}_d$ is homeomorphic to $\mathcal{P}_G = \{P \in \mathbb{R}^d : 1 \geq p_1 \geq \ldots \geq p_d \geq 0,\ \sum_i p_i = 1\}$ and the reduced statistical game is given by*

$$\inf_{\hat{f}\in\mathcal{M}_{\mathcal{D},G}} \sup_{P\in\mathcal{P}_G} R(\hat{f}, P), \tag{5.17}$$

where $\mathcal{M}_{\mathcal{D},G}$ represents the set of randomized estimators which are invariant to the actions of permutation group.

5.4 Finite Gaussian Sequence Model

In this section we consider the finite Gaussian sequence model described in Section 5.3.1 and use Algorithm 5 to construct a provably minimax estimator, which can be computed in polynomial time. This problem has received a lot of attention in statistics because of its simplicity, relevance and its connections to non-parametric regression [see Chapter 1 of Joh11]. When the radius of the domain B is smaller than $1.15\sqrt{d}$, Marchand and Perron [MP02] show that the Bayes estimator with uniform prior on the boundary is a minimax estimator for the problem. For larger values of B, the exact minimax estimator is unknown. Several works have attempted to understand the properties of LFP in such settings [CS81] and constructed approximate minimax estimators [Bic81]. In this chapter, we rely on Algorithm 5 to construct an exact minimax estimator and an LFP, for any value of B, d.

Recall, in Theorem 14 we showed that the original min-max statistical game can be reduced to the simpler problem in Equation (5.10) To use Algorithm 5 to find a Nash equilibrium of the reduced game, we need efficient implementation of the required optimization oracles and a bound on their approximation errors. The optimization problems corresponding to the oracles in Equations (5.6), (5.7) are given as follows

$$\hat{\theta}_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right], \quad b_t(\sigma) \leftarrow \operatorname*{argmax}_{b\in[0,B]} \sum_{i=1}^{t-1} R(\hat{\theta}_i, b\mathbf{e}_1) + \sigma b,$$

where $\mathcal{D}_G$ is the set of deterministic invariant estimators and P_t is the distribution of random variable $b_t(\sigma)$. We now present efficient techniques for implementing these oracles (Algorithms 6, 7). Since the maximization problem is a 1 dimensional optimization problem, grid search can be used to compute an approximate maximizer. The approximation error of the resulting oracle depends on the grid width and the number of samples used to compute the expectation in the risk $R(\hat{\theta}, b\mathbf{e}_1)$. Later, we show that poly(d, B) grid points and samples suffice to have a small approximation error. The minimization problem, which requires finding an invariant estimator minimizing the integrated risk under any prior P_t, can also be efficiently implemented. As shown in Proposition 8 below, the minimizer has a closed-form expression which depends on P_t and modified Bessel functions. To compute an approximate minimizer of the problem, we approximate P_t with its samples and rely on the closed-form expression. The approximation error of this oracle depends on the number of samples used to approximate P_t. We again show that poly(d, B) samples suffice to have a small approximation error.

Proposition 8. *The optimizer $\hat{\theta}_t$ of the minimization problem defined above has the following closed-form expression*

$$\hat{\theta}_t(X) = \left(\frac{\mathbb{E}_{b\sim P_t}\left[b^{3-d/2}e^{-b^2/2}I_{d/2}(b\|X\|_2)\right]}{\mathbb{E}_{b\sim P_t}\left[b^{2-d/2}e^{-b^2/2}I_{d/2-1}(b\|X\|_2)\right]}\right)\frac{X}{\|X\|_2},$$

where I_ν is the modified Bessel function of first kind of order ν.

Algorithm 6 Maximization Oracle

```
Input: Estimators $\{\hat{\theta}_i\}_{i=1}^{t-1}$, perturbation $\sigma$, grid width $w$, number of samples for computation
    of expected risk $R(\hat{\theta}, \theta)$: $N_1$
Let $\{b_1, b_2 \dots b_{B/w}\}$ be uniformly spaced points on $[0, B]$
for $j = 1 \dots B/w$ do
    for $i = 1 \dots t-1$ do
        Generate $N_1$ independent samples $\{X_k\}_{k=1}^{N_1}$ from the distribution $\mathcal{N}(b_j \mathbf{e}_1, I)$
        Estimate $R(\hat{\theta}_i, b_j \mathbf{e}_1)$ as $\frac{1}{N_1}\sum_{k=1}^{N_1} \|\hat{\theta}_i(X_k) - b\mathbf{e}_1\|_2^2$.
    end for
    Evaluate the objective at $b_j$ using the above estimates
end for
Output: $b_j$ which maximizes the objective
```

Algorithm 7 Minimization Oracle

```
Input: Samples $\{b_i\}_{i=1}^{N_2}$ generated from distribution $P_t$.
For any $X$, compute $\hat{\theta}_t(X)$ as
    $$\left(\frac{\sum_{i=1}^{N_2} w_i b_i A(b_i \|X\|_2)}{\sum_{i=1}^{N_2} w_i}\right) \frac{X}{\|X\|_2},$$
    where $A(\gamma) = \frac{I_{d/2}(\gamma)}{I_{d/2-1}(\gamma)}$, $w_i = b_i^{2-d/2} e^{-b_i^2/2} I_{d/2-1}(b_i \|X\|_2)$, and $I_\nu$ is the modified Bessel
    function of the first kind of order $\nu$.
```

We now show that using Algorithm 5 for solving objective (5.10) with Algorithms 6, 7 as optimization oracles, gives us a provably minimax estimator and an LFP for finite Gaussian sequence model.

Theorem 19. *Suppose Algorithm 5 is run for T iterations with Algorithms 6, 7 as the maximization and minimization oracles. Suppose the hyper-parameters of these algorithms are set as $\eta = \frac{1}{B(B+1)\sqrt{T}}$, $w = \frac{B}{T^{3/2}}$, $N_1 = \frac{T^3}{(B+1)^2}$, $N_2 = \frac{T^4}{(B+1)^2}$. Let $\hat{P}_t$ be the approximation of probability distribution P_t used in the t^{th} iteration of Algorithm 5. Moreover, let $\hat{\theta}_t$ be the output of Algorithm 7 in the t^{th} iteration of Algorithm 5.*

1. *Then the averaged estimator $\hat{\theta}_{avg}(X) = \frac{1}{T}\sum_{i=1}^{T} \hat{\theta}_i(X)$ is approximately minimax and satisfies the following worst-case risk bound with probability at least $1-\delta$*

$$\sup_{\theta : \|\theta\|_2 \le B} R(\hat{\theta}_{avg}, \theta) \le R^* + \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right),$$

where $\tilde{O}(.)$ hides log factors and R^* is the minimax risk.

2. *Define the mixture distribution $\hat{P}_{\text{AVG}}$ as $\frac{1}{T}\sum_{i=1}^{T} \hat{P}_i$. Let $\hat{P}_{LFP}$ be a probability distribution over $\mathbb{R}^d$ with density function defined as $\hat{p}_{LFP}(\theta) \propto \|\theta\|_2^{1-d} \hat{P}_{\text{AVG}}(\|\theta\|_2)$, where $\hat{P}_{\text{AVG}}(\|\theta\|_2)$ is the probability mass placed by $\hat{P}_{\text{AVG}}$ at $\|\theta\|_2$. Then $\hat{P}_{LFP}$ is approximately*

least favorable and satisfies the following with probability at least $1-\delta$

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{LFP}) \geq R^* - \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right),$$

where the infimum is over the set of all estimators.

We believe the polynomial factors in the bounds can be improved with a tighter analysis of the algorithm. The above Theorem shows that Algorithm 5 learns an approximate minimax estimator in poly(d, B) time. To the best our knowledge, this is the first result providing provable minimax estimators for finite Gaussian sequence model, for any value of B.

5.5 Linear Regression

In this section we consider the linear regression problem described in Section 5.3.2 and provide a provably minimax estimator. Recall, in Theorem 16 we showed that the original min-max statistical game can be reduced to the simpler problem in Equation (5.13). We now provide efficient implementations of the optimization oracles required by Algorithm 5 for finding a Nash equilibrium of this game. The optimization problems corresponding to the two optimization oracles are as follows

$$\hat{\theta}_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right], \quad b_t(\sigma) \leftarrow \operatorname*{argmax}_{b\in[0,B]} \sum_{i=1}^{t-1} R(\hat{\theta}_i, b\mathbf{e}_1) + \sigma b,$$

where $\mathcal{D}_G$ is the set of deterministic invariant estimators and P_t is the distribution of random variable $b_t(\sigma)$. Similar to the Gaussian sequence model, the maximization oracle can be efficiently implemented via a grid search over $[0, B]$ (Algorithm 8). The solution to the minimization problem has a closed-form expression in terms of the mean and normalization constant of Fisher-Bingham distribution, which is a distribution obtained by constraining multivariate normal distributions to lie on the surface of unit sphere [KW05]. Letting $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, the probability density of a random variable Z distributed according to Fisher-Bingham distribution is given by

$$p(Z; A, \gamma) = C(A, \gamma)^{-1} \exp\left(-Z^T A Z + \langle\gamma, Z\rangle\right),$$

where $Z \in \mathbb{S}^{d-1}$, and $\gamma \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$ are the parameters of the distribution with A being positive semi-definite and $C(A, \gamma)$ is the normalization constant. Note that the mean of Fisher-Bingham distribution is given by $C(A, \gamma)^{-1}\frac{\partial}{\partial\gamma}C(A, \gamma)$. The following proposition obtains a closed-form expression for $\hat{\theta}_t$ in terms of $C(A, \gamma)$ and $\frac{\partial}{\partial\gamma}C(A, \gamma)$.

Proposition 9. *The optimizer $\hat{\theta}_t$ of the minimization problem defined above has the following closed-form expression*

$$\hat{\theta}_t(D_n) = \frac{\mathbb{E}_{b\sim P_t}\left[b^2 \frac{\partial}{\partial\gamma} C\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, \gamma\right)\Big|_{\gamma=b\mathbf{X}^T\mathbf{Y}}\right]}{\mathbb{E}_{b\sim P_t}\left[bC\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, b\mathbf{X}^T\mathbf{Y}\right)\right]},$$

Algorithm 8 Regression Maximization Oracle

```
Input: Estimators $\{\hat{\theta}_i\}_{i=1}^{t-1}$, perturbation $\sigma$, grid width $w$, number of samples for computation of expected risk $R(\hat{\theta}, \theta)$: $N_1$
Let $\{b_1, b_2 \dots b_{B/w}\}$ be uniformly spaced points on $[0, B]$
for $j = 1 \dots B/w$ do
    for $i = 1 \dots t-1$ do
        Generate $N_1$ independent datasets $\{D_{n,k}\}_{k=1}^{N_1}$ from the linear model with true regression vector $b_j\mathbf{e}_1$
        Estimate $R(\hat{\theta}_i, b_j\mathbf{e}_1)$ as $\frac{1}{N_1}\sum_{k=1}^{N_1} \|\hat{\theta}_i(D_{n,k}) - b\mathbf{e}_1\|_2^2$.
    end for
    Evaluate the objective at $b_j$ using the above estimates
end for
Output: $b_j$ which maximizes the objective
```

Algorithm 9 Regression Minimization Oracle

```
Input: Samples $\{b_i\}_{i=1}^{N_2}$ generated from distribution $P_t$
For any $D_n$, compute $\hat{\theta}_t(D_n)$ as
    $$\hat{\theta}_t(D_n) = \frac{\sum_{i=1}^{N_2} b_i^2 \frac{\partial}{\partial \gamma} C\left(2^{-1}b_i^2\mathbf{X}^T\mathbf{X}, \gamma\right)\Big|_{\gamma = b_i\mathbf{X}^T\mathbf{Y}}}{\sum_{i=1}^{N_2} b_i C\left(2^{-1}b_i^2\mathbf{X}^T\mathbf{X}, b_i\mathbf{X}^T\mathbf{Y}\right)},$$
    where $\mathbf{X} = [X_1, X_2 \dots X_n]^T$ and $\mathbf{Y} = [Y_1, Y_2 \dots Y_n]$.
```

where $\mathbf{X} = [X_1, X_2 \dots X_n]^T$ *and* $\mathbf{Y} = [Y_1, Y_2 \dots Y_n]$.

We note that there exist a number of efficient techniques for computation of the mean and normalization constant of Fisher-Bingham distribution [KW05; Imh61]. In our experiments we rely on the technique of Kume and Wood [KW05] (we relegate the details of this technique to Appendix D.6.2). To compute an approximate optimizer of the minimization problem, we approximate P_t with its samples and rely on the above closed-form expression. Algorithm 9 describes the resulting minimization oracle. We now show that using Algorithm 5 for solving objective (5.13) with Algorithms 8, 9 as optimization oracles, gives us a provably minimax estimator and an LFP for linear regression.

Theorem 20. *Suppose Algorithm 5 is run for T iterations with Algorithms 8, 9 as the maximization and minimization oracles. Suppose the hyper-parameters of these algorithms are set as $\eta = \frac{1}{B(B\sqrt{n}+1)\sqrt{T}}$, $w = \frac{B}{T^{3/2}}$, $N_1 = \frac{T^3}{(B\sqrt{n}+1)^2}$, $N_2 = \frac{T^4}{(B\sqrt{n}+1)^2}$. Let $\hat{P}_t$ be the approximation of probability distribution P_t used in the t^{th} iteration of Algorithm 5. Moreover, let $\hat{\theta}_t$ be the output of Algorithm 9 in the t^{th} iteration of Algorithm 5.*

1. *Then the averaged estimator $\hat{\theta}_{avg}(D_n) = \frac{1}{T}\sum_{i=1}^{T}\hat{\theta}_i(D_n)$ is approximately minimax and satisfies the following worst-case risk bound with probability at least $1-\delta$*

$$\sup_{\theta:\|\theta\|_2 \leq B} R(\hat{\theta}_{avg}, \theta) \leq R^* + \tilde{O}\left(B^2(B+1)\sqrt{\frac{n}{T}}\right).$$

2. *Define the mixture distribution $\hat{P}_{\text{AVG}}$ as $\frac{1}{T}\sum_{i=1}^{T}\hat{P}_i$. Let $\hat{P}_{LFP}$ be a probability distribution over $\mathbb{R}^d$ with density function defined as $\hat{p}_{LFP}(\theta) \propto \|\theta\|_2^{1-d}\hat{P}_{\text{AVG}}(\|\theta\|_2)$, where $\hat{P}_{\text{AVG}}(\|\theta\|_2)$ is the probability mass placed by $\hat{P}_{\text{AVG}}$ at $\|\theta\|_2$. Then $\hat{P}_{LFP}$ is approximately least favorable and satisfies the following with probability at least $1-\delta$*

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{LFP}) \geq R^* - \tilde{O}\left(B^2(B+1)\sqrt{\frac{n}{T}}\right).$$

5.6 Normal Covariance Estimation

In this section, we consider the problem of normal covariance estimation. Recall, in Section 5.3.3 we showed that the problem is invariant to the action of the orthogonal group and can be reduced to the simpler problem in Equation (5.15). The optimization problems corresponding to the oracles in Equations (5.6), (5.7) are as follows

$$\hat{\Sigma}_t \leftarrow \operatorname*{argmin}_{\hat{\Sigma}\in\mathcal{D}_G} \mathbb{E}_{\lambda\sim P_t}\left[R(\hat{\Sigma}, \text{Diag}(\lambda))\right], \quad \lambda_t(\sigma) \leftarrow \operatorname*{argmax}_{\lambda\in\Xi_G} \sum_{i=1}^{t-1} R(\hat{\Sigma}_i, \text{Diag}(\lambda)) + \langle\lambda, \sigma\rangle,$$

where $\mathcal{D}_G$ is the set of deterministic invariant estimators and P_t is the distribution of random variable $\lambda_t(\sigma)$. Note that the maximization problem involves optimization of a non-concave objective in d-dimensional space. So, implementing a maximization oracle with low approximation error can be computationally expensive, especially in high dimensions. Moreover, unlike finite Gaussian sequence model and linear regression, the minimization problem doesn't have a closed form expression, and it is not immediately clear how to efficiently implement a minimization oracle with low approximation error. In such scenarios, we show that one can rely on a combination of heuristics and problem structure to further reduce the computational complexity of the optimization oracles. Although relying on heuristics comes at the expense of theoretical guarantees, in later sections, we empirically demonstrate that the resulting estimators have superior performance over classical estimators. We begin by showing that the domain of the outer minimization in Equation (5.15) can be reduced to a smaller set of estimators. Our reduction relies on Blackwell's theorem, which shows that for convex loss functions M, there exists a minimax estimator which is a function of the sufficient statistic [IH81]. We note that Blackwell's theorem is very general and can be applied to a wide range of problems, to reduce the computational complexity of the minimization oracle.

Proposition 10. *Consider the problem of normal covariance estimation. Let S_n be the empirical covariance matrix which is defined as $\frac{\sum_{i=1}^{n} X_i X_i^T}{n}$ and let $U\Delta U^T$ be the eigen decomposition of S_n. Then there exists a minimax estimator which can be approximated arbitrarily well using estimators of the form $\hat{\Sigma}_{f,g}(\mathbb{X}^n) = U\tilde{\Sigma}_{f,g}(\Delta)U^T$, where $\tilde{\Sigma}_{f,g}(\Delta)$ is a diagonal matrix whose i^{th} diagonal entry is given by*

$$\tilde{\Sigma}_{f,g,i}(\Delta) = f\left(\Delta_i, \sum_{j\neq i} g(\Delta_i, \Delta_j)\right),$$

for some functions $f : \mathbb{R}^{d+1} \to \mathbb{R}, g : \mathbb{R}^2 \to \mathbb{R}^d$. Here, Δ_i is the i^{th} diagonal entry of Δ. Moreover, the optimization problem in Equation (5.15) can be reduced to the following simpler problem

$$\inf_{\hat{\Sigma} \in \mathcal{M}_{f,g}} \sup_{\lambda \in \Xi_G} R(\hat{\Sigma}, \mathit{Diag}(\lambda)) = R^*, \tag{5.18}$$

where $\mathcal{M}_{f,g}$ is the set of probability distributions over estimators of the form $\hat{\Sigma}_{f,g}$.

We now use Algorithm 5 to solve the statistical game in Equation (5.18). The optimization problems corresponding to the two optimization oracles are given by

$$\hat{f}_t, \hat{g}_t \leftarrow \operatorname*{argmin}_{f,g} \mathbb{E}_{\lambda \sim P_t} \left[R(\hat{\Sigma}_{f,g}, \mathrm{Diag}(\lambda)) \right],$$
$$\lambda_t(\sigma) \leftarrow \operatorname*{argmax}_{\lambda \in \Xi_G} \sum_{i=1}^{t-1} R(\hat{\Sigma}_{\hat{f}_i, \hat{g}_i}, \mathrm{Diag}(\lambda)) + \langle \lambda, \sigma \rangle.$$

We rely on heuristics to efficiently implement these oracles. To implement the minimization oracle, we use neural networks (which are universal function approximators) to parameterize functions f, g. Implementing the minimization oracle then boils down to the finding the parameters of these networks which minimize the objective. To implement the maximization oracle, we rely on global search techniques. In our experiments, we use DragonFly [Kan+19], which is a zeroth order optimization technique, to implement this oracle. Note that these heuristics do not come with any guarantees and as a result the oracles are not guaranteed to have a small approximation error. Despite this, we empirically demonstrate that the estimators learned using this approach have good performance.

5.7 Entropy Estimation

In this section, we consider the problem of entropy estimation. Recall, in Section 5.3.4 we showed that the problem is invariant to the action of permutation group and can be reduced to the simpler problem in Equation (5.17). Similar to the problem of covariance estimation, implementing the optimization oracles for this problem, with low approximation error, can be computationally expensive. So we again rely on heuristics and problem structure to reduce the computational complexity of optimization oracles.

Proposition 11. *Consider the problem of entropy estimation. Let $\hat{P}_n = (\hat{p}_1, \dots \hat{p}_d)$ be the observed empirical probabilities. Then there exists a minimax estimator which can be approximated arbitrarily well using estimators of the form $\hat{f}_{g,h}(\hat{P}_n) = g(\sum_{i=1}^d h(\hat{p}_i))$, for some functions $g : \mathbb{R}^{d+1} \to \mathbb{R}, h : \mathbb{R} \to \mathbb{R}^{d+1}$. Moreover, the optimization problem in Equation (5.17) can be reduced to the following problem*

$$\inf_{\hat{f} \in \mathcal{M}_{g,h}} \sup_{P \in \mathcal{P}_G} R(\hat{f}, P) = R^*, \tag{5.19}$$

where $\mathcal{M}_{g,h}$ is the set of probability distributions over estimators of the form $\hat{f}_{g,h}$.

The proof of this proposition is presented in Appendix D.8.1. We now use Algorithm 5 to solve the statistical game in Equation (5.19). The optimization problems corresponding to the two optimization oracles are given by

$$\hat{g}_t, \hat{h}_t \leftarrow \operatorname*{argmin}_{g,h} \mathbb{E}_{P\sim P_t}\left[R(\hat{f}_{g,h}, P)\right], \quad P_t(\sigma) \leftarrow \operatorname*{argmax}_{P\in\mathcal{P}_G} \sum_{i=1}^{t-1} R(\hat{f}_{\hat{g}_i,\hat{h}_i}, P) + \langle P, \sigma\rangle,$$

where P_t is the distribution of random variable $P_t(\sigma)$. To implement the minimization oracle, we use neural networks to parameterize functions g, h. To implement the maximization oracle, we rely on DragonFly.

5.8 Experiments

In this section, we present experiments showing performance of the proposed technique for constructing minimax estimators. While our primary focus is on the finite Gaussian sequence model and linear regression for which we provided provably minimax estimators, we also present experiments on other problems such as covariance and entropy estimation. For each of these problems, we begin by describing the setup as well as the baseline algorithms, before proceeding to a discussion of the experimental findings.

5.8.1 Finite Gaussian Sequence Model

In this section, we focus on experiments related to the finite Gaussian sequence model. We first consider the case where the risk is measured with respect to squared error loss, *i.e.*, $M(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2$.

Proposed Technique We use Algorithm 5 with optimization oracles described in Algorithms 6, 7 to find minimax estimators for this problem. We set the hyper-parameters of our algorithm as follows: number of iterations of FTPL $T = 500$, grid width $w = 0.05 \times B$, number of samples for computation of $R(\hat{\theta}, \theta)$ in Algorithm 6 $N_1 = 1000$, number of samples generated from P_t in Algorithm 7 $N_2 = 1000$. We note that these are default values and were not tuned. The randomness parameter η in Algorithm 5 was tuned using a coarse grid search. We report the performance of the following two estimators constructed using the iterates of Algorithm 5: (a) Averaged Estimator $\hat{\theta}_{\text{AVG}}(X) = \frac{1}{T}\sum_{i=1}^{T} \hat{\theta}_i(X)$, (b) Bayes estimator for prior $\frac{1}{T}\sum_{i=1}^{T} \hat{P}_i$ which we refer to as "Bayes estimator for avg. prior". The performance of the randomized estimator $\hat{\theta}_{\text{RND}}$ is almost identical to the performance of $\hat{\theta}_{\text{AVG}}$. So we do not report its performance.

Baselines We compare our estimators with various baselines: (a) standard estimator $\hat{\theta}(X) = X$, (b) James Stein estimator $\hat{\theta}(X) = (1 - (d-3)/\|X\|_2^2)^+ X$, where $c^+ = \max(0, c)$, (c) projection estimator (MLE) $\hat{\theta}(X) = \min(\|X\|_2, B)\frac{X}{\|X\|_2}$, (d) Bayes estimator for uniform prior on the boundary; this estimator is known to be minimax for $B \leq 1.15\sqrt{d}$.

Table 5.1: Worst-case risk of various estimators for finite Gaussian sequence model. The risk is measured with respect to squared error loss. The worst-case risk of the estimators from Algorithm 5 (last two rows) is smaller than the worst-case risk of baselines. The numbers in the brackets for Averaged Estimator represent the duality gap.

	Worst-case Risk								
	$\mathbf{B = \sqrt{d}}$			**$\mathbf{B = 1.5\sqrt{d}}$**			**$\mathbf{B = 2\sqrt{d}}$**		
Estimator	**d = 10**	**d = 20**	**d = 30**	**d = 10**	**d = 20**	**d = 30**	**d = 10**	**d = 20**	**d = 30**
Standard	10	20	30	10	20	30	10	20	30
James Stein	6.0954	11.2427	16.073	7.9255	15.0530	21.3410	8.7317	16.6971	24.7261
Projection	8.3076	17.4788	26.7873	10.3308	20.3784	30.2464	10.1656	20.2360	30.3805
Bayes estimator for uniform prior on boundary	**4.8559**	**9.9909**	**14.8690**	11.7509	23.4726	35.2481	24.5361	49.0651	73.3158
Averaged Estimator	**4.7510** (0.1821)	**9.7299** (0.2973)	**14.8790** (0.0935)	**6.7990** (0.0733)	**13.8084** (0.2442)	**20.5704** (0.0087)	**7.8504** (0.3046)	**15.6686** (0.2878)	**23.8758** (0.6820)
Bayes estimator for avg. prior	**4.9763**	**10.1273**	**14.8128**	**6.7866**	**13.8200**	**20.3043**	**7.8772**	**15.6333**	**23.5954**

Worst-case Risk We compare the performance of various estimators based on their worst-case risk. The worst-case risk of the standard estimator is equal to d. The worst case risk of all the other estimators is computed as follows. Since all these estimators are invariant to orthogonal group transformations, the risk $R(\hat{\theta}, \theta)$ only depends on $\|\theta\|_2$ and not its direction. So the worst-case risk can be obtained by solving the following optimization problem: $\max_{b \in [0,B]} R(\hat{\theta}, b\mathbf{e}_1)$, where $\mathbf{e}_1$ is the first standard basis vector. We use grid search to solve this problem, with $0.05 \times B$ grid width. We use 10^4 samples to approximately compute $R(\hat{\theta}, b\mathbf{e}_1)$ for any $\hat{\theta}, b$.

Duality Gap For estimators derived from our technique, we also present the duality gap, which is defined as $\sup_{\theta \in \Theta} R(\hat{\theta}_{\text{AVG}}, \theta) - \inf_{\hat{\theta} \in \mathcal{D}} R(\hat{\theta}, \frac{1}{T}\sum_{i=1}^{T} \hat{P}_i)$. Duality gap quantifies the closeness of $(\hat{\theta}_{\text{AVG}}, \frac{1}{T}\sum_{i=1}^{T} \hat{P}_i)$ to a Nash equilibrium. Smaller the gap, closer we are to an equilibrium.

Results Table 5.1 shows the performance of various estimators for various values of d, B along with the duality gap for our estimator. For $B = \sqrt{d}$, the estimators obtained using Algorithm 5 have similar performance as the "Bayes estimator for uniform prior on boundary", which is known to be minimax. For $B = 2\sqrt{d}, 3\sqrt{d}$ for which the exact minimax estimator is unknown, we achieve better performance than baselines. Finally, we note that the duality gap numbers presented in the table can be made smaller by running our algorithm for more iterations. When the dimension $d = 1$, Donoho, Liu, and MacGibbon [DLM90] derived lower bounds for the minimax risk, for various values of B. In Table 5.2, we compare the worst risk of our estimator with these established lower bounds. It can be seen that the worst case risk of our estimator is close to the lower bounds.

Table 5.2: Comparison of the worst case risk of $\hat{\theta}_{\text{AVG}}$ with established lower bounds from [DLM90] for finite Gaussian sequence model with $d = 1$.

	$\mathbf{B = 1}$	$\mathbf{B = 2}$	$\mathbf{B = 3}$	$\mathbf{B = 4}$
Worst case risk of Averaged Estimator	0.456	0.688	0.799	0.869
Lower bound	0.449	0.644	0.750	0.814

Estimating a few coordinates

In this section we again consider the finite Gaussian sequence model, but with a different risk. We now measure the risk on only the first k coordinates: $M(\theta_1, \theta_2) = \sum_{i=1}^{k}(\theta_1(i) - \theta_2(i))^2$. We present experimental results for $k = 1, d/2$.

Proposed Technique Following Theorem 15, the original min-max objective can be reduced to the simpler problem in Equation (5.11). We use similar optimization oracles as in Algorithms 6, 7, to solve this problem. The maximization problem is now a 2D optimization problem for which we use grid search. The minimization problem, which requires computation of Bayes estimators, can be solved analytically and has similar expression as the Bayes estimator in Algorithm 7 (see Appendix D.5.3 for details). We use a 2D grid of $0.05B$ width and length in the maximization oracle. We use the same hyper-parameters as above and run FTPL for 10000 iterations for $k = 1$ and 4000 iterations for $k = d/2$.

Worst-case Risk We compare our estimators with the same baselines described in the previous section. For the case of $k = 1$, we also compare with the best linear estimator, which is known to be approximately minimax with worst case risk smaller than 1.25 times the minimax risk [Don94]. Since all these estimators, except the best linear estimator, are invariant to the transformations of group $\mathbb{O}(k) \times \mathbb{O}(d-k)$, the max risk of these estimators can be written as $\max_{b_1^2+b_2^2 \leq B^2} R(\hat{\theta}, [b_1\mathbf{e}_{1,k}, b_2\mathbf{e}_{1,d-k}])$. We solve this problem using $2D$ grid search. The worst case risk of best linear estimator has a closed form expression.

Results Table 5.3 shows the performance of various estimators for various values of d, B. It can be seen that for $B = \sqrt{d}$, our estimators have better performance than other baselines. The performance difference goes down for large B, which is as expected. In order to gain insights about the estimator learned by our algorithm, we plot the contours of $\hat{\theta}_{\text{AVG}}(X)$ in Figure 5.1, for the $k = 1$ case, where the risk is measured on the first coordinate. It can be seen that when $X(1)$ is close to 0, irrespective of other coordinates, the estimator just outputs $X(1)$ as its estimate of $\theta(1)$. When $X(1)$ if far from 0, by looking along the corresponding vertical line, the estimator can be seen as outputting a shrinked version of $X(1)$, where the amount of shrinkage increases with the norm of $X(2:d)$. Note that this is unlike James Stein estimator which shrinks vectors with smaller norm more than larger norm vectors.

Table 5.3: Worst-case risk of various estimators for bounded normal mean estimation when the risk is evaluated with respect to squared loss on the first k coordinates.

	Worst-case Risk								
	$\mathbf{k = 1, B = \sqrt{d}}$			$\mathbf{k = 1, B = 2\sqrt{d}}$			$\mathbf{k = 1, B = 3\sqrt{d}}$		
Estimator	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$
Standard Estimator	1	1	1	1	1	1	1	1	1
James-Stein Estimator	2.3796	4.9005	7.3489	2.5087	4.9375	7.3760	2.4288	4.8951	7.3847
Projection Estimator	1.0055	1.4430	2.0424	1.0263	1.1051	1.5077	1.0288	1.0310	1.0202
Best Linear Estimator	0.9091	0.9524	0.9677	0.9756	0.9877	0.9917	0.9890	0.9945	0.9963
Bayes Estimator for average prior	**0.7955**	**0.8565**	**0.8996**	**0.9160**	**0.9496**	0.9726	0.9611	1.0007	1.0172
Averaged Estimator	**0.7939**	**0.8579**	**0.8955**	**0.9104**	**0.9497**	0.9724	0.9640	1.0003	1.0101

	Worst-case Risk								
	$\mathbf{k = d/2, B = \sqrt{d}}$			$\mathbf{k = d/2, B = 2\sqrt{d}}$			$\mathbf{k = d/2, B = 3\sqrt{d}}$		
Estimator	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$	$\mathbf{d = 10}$	$\mathbf{d = 20}$	$\mathbf{d = 30}$
Standard Estimator	5	10	15	5	10	15	5	10	15
James-Stein Estimator	4.1167	7.9200	11.6892	5.0109	9.7551	14.6568	5.0281	10.0155	14.9390
Projection Estimator	7.1096	15.8166	24.8158	30.3166	66.1806	103.0456	73.4834	156.5076	241.1031
Bayes Estimator for average prior	**3.2611**	**6.5834**	**9.8189**	**4.2477**	**8.6564**	**13.0606**	**4.6359**	**9.2773**	**13.9678**
Averaged Estimator	**3.2008**	**6.4763**	**9.7763**	**4.2260**	**8.6421**	**13.0353**	**4.6413**	**9.2760**	**13.9446**

5.8.2 Linear Regression

In this section we present experimental results on linear regression. We use Algorithm 5 with optimization oracles described in Algorithms 8, 9 to find minimax estimators for this problem. We use the same hyper-parameter settings as finite Gaussian sequence model, and run Algorithm 5 for $T = 500$ iterations. We compare the worst-case risk of minimax estimators obtained using our algorithm for various values of (n, d, B), with ordinary least squares (OLS) and ridge regression estimators. Since all the estimators are invariant to the transformations of orthogonal group $\mathbb{O}(d)$, the max risk can be written as $\max_{b \in [0,B]} R(\hat{\theta}, b\mathbf{e}_1)$, which can be efficiently computed using grid search. Table 5.4 presents the results from this experiment. It can be seen that we achieve better performance than ridge regression for small values of n/d, B. For large values of n/d, B, the performance

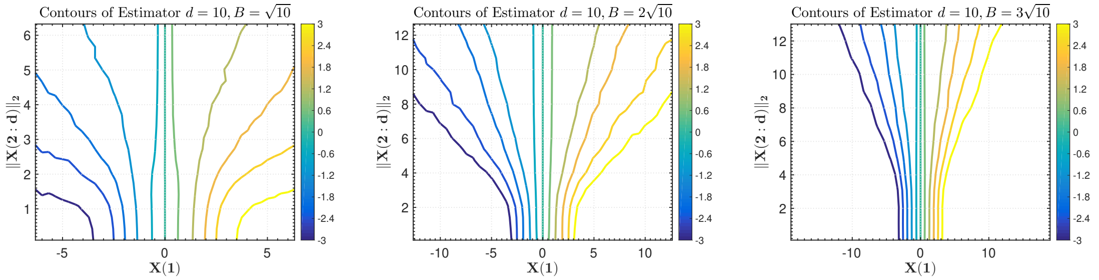


Figure 5.1: Contour plots of the estimator learned using Algorithm 5 when the risk is evaluated on the first coordinate. x axis shows the first coordinate of X, which is the input to the estimator. y axis shows the norm of the rest of the coordinates of X. The contour bar shows $\hat{\theta}(1)$, the first co-ordinate of the output of the estimator.

of our estimator approaches ridge regression. The duality gap numbers presented in the Table suggest that the performance of our estimator can be improved for larger values of $n/d, B$, by choosing better hyper-parameters.

Table 5.4: Worst-case risk of various estimators for linear regression. The performance of ridge is obtained by choosing the best regularization parameter. The numbers in the brackets for Averaged Estimator represent the duality gap.

Estimator	**Worst-case Risk**							
	$\mathbf{n = 1.5 \times d, B = 0.5 \times \sqrt{d}}$				$\mathbf{n = 1.5 \times d, B = \sqrt{d}}$			
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$
OLS	5.0000	2.5000	2.5000	2.2222	5.0000	2.5000	2.5000	2.2222
Ridge regression	0.6637	0.9048	1.1288	1.1926	1.3021	1.4837	1.6912	1.6704
Averaged Estimator	**0.5827** (0.0003)	**0.8275** (0.0052)	**0.9839** (0.0187)	**1.0946** (0.0404)	**1.2030** (0.0981)	1.4615 (0.1145)	1.6178 (0.1768)	1.6593 (0.1863)
Bayes estimator for avg. prior	**0.5827**	**0.8275**	**0.9844**	**1.0961**	**1.1750**	1.4621	1.6265	1.6674

Estimator	**Worst-case Risk**							
	$\mathbf{n = 2 \times d, B = 0.5 \times \sqrt{d}}$				$\mathbf{n = 2 \times d, B = \sqrt{d}}$			
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 15}$	$\mathbf{d = 20}$
OLS	1.2500	1.1111	1.0714	1.053	1.2500	1.1111	1.0714	1.053
Ridge regression	0.5225	0.6683	0.7594	0.8080	0.8166	0.8917	0.9305	0.9608
Averaged Estimator	0.4920 (0.0038)	**0.5991** (0.0309)	**0.6873** (0.0485)	**0.7339** (0.0428)	0.8044 (0.0647)	0.8615 (0.0854)	0.9388 (0.0996)	0.9621 (0.1224)
Bayes estimator for avg. prior	0.4894	**0.6004**	**0.6879**	**0.7320**	0.8140	0.8618	0.9375	0.9656

5.8.3 Normal Covariance Estimation

In this section we present experimental results on normal covariance estimation.

Minimization oracle In our experiments we use neural networks, which are universal function approximators, to parameterize functions f, g in Equation (5.18). To be precise, we use two layer neural networks to parameterize each of these functions. Implementing the minimization oracle then boils down to finding the parameters of these networks which minimize $\mathbb{E}_{\lambda \sim P_t}\left[R(\hat{\Sigma}_{f,g}, \text{Diag}(\lambda))\right]$. In our experiments, we use stochastic gradient descent to learn these parameters.

Baselines We compare the performance of the estimators returned by Algorithm 5 for various values of (n, d, B), with empirical covariance S_n and the James Stein estimator [JS92] which is defined as $K_n \Delta_{JS} K_n^T$, where K_n is a lower triangular matrix such that $S_n = K_n K_n^T$ and Δ_{JS} is a diagonal matrix with i^{th} diagonal element equal to $\frac{1}{n+d-2i+1}$.

Results We use worst-case risk to compare the performance of various estimators. To compute the worst-case risk, we again rely on DragonFly. We note that the worst-case computed using this approach may be inaccurate as DragonFly is not guaranteed to return a global optimum. So, we also compare the risk of various estimators at randomly generated Σ's (see Appendix D.9). Table 5.5 presents the results from this experiment. It can be seen that our estimators outperform empirical covariance for almost all the values of n, d, B and outperform James Stein estimator for small values of n/d, B. For large values of n/d, B, our estimator has similar performance as JS. In this setting, we believe the performance of our estimators can be improved by running the algorithm with better hyper-parameters.

Table 5.5: Worst-case risk of various estimators for covariance estimation for various configurations of (n, d, B). The worst-case risks are obtained by taking a max of the worst-case risk estimate from DragonFly and the risks computed at randomly generated Σ's.

Estimator	**Worst-case Risk**							
	$\mathbf{n = 1.5 \times d, B = 1}$		$\mathbf{n = 1.5 \times d, B = 2}$		$\mathbf{n = 1.5 \times d, B = 4}$		$\mathbf{n = 1.5 \times d, B = 8}$	
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$
Empirical Covariance	2.5245	5.1095	2.5245	5.1095	2.5245	5.1095	2.5245	5.1095
James-Stein Estimator	2.1637	4.1704	2.1637	4.1704	2.1637	4.1704	2.1637	4.1704
Averaged Estimator	**1.8686**	**3.1910**	**1.9371**	**3.7019**	2.0827	4.2454	2.1416	**3.9864**

Estimator	**Worst-case Risk**							
	$\mathbf{n = 2 \times d, B = 1}$		$\mathbf{n = 2 \times d, B = 2}$		$\mathbf{n = 2 \times d, B = 4}$		$\mathbf{n = 2 \times d, B = 8}$	
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$
Empirical Covariance	1.8714	3.4550	1.8714	3.4550	1.8714	3.4550	1.8714	3.4550
James-Stein Estimator	1.6686	2.9433	1.6686	2.9433	1.6686	2.9433	1.6686	2.9433
Averaged Estimator	**1.2330**	**2.1944**	1.5237	**2.6471**	1.6050	3.0834	1.6500	2.9907

Estimator	**Worst-case Risk**							
	$\mathbf{n = 3 \times d, B = 1}$		$\mathbf{n = 3 \times d, B = 2}$		$\mathbf{n = 3 \times d, B = 4}$		$\mathbf{n = 3 \times d, B = 8}$	
	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$	$\mathbf{d = 5}$	$\mathbf{d = 10}$
Empirical Covariance	1.1425	2.1224	1.1425	2.1224	1.1425	2.1224	1.1425	2.1224
James-Stein Estimator	1.0487	1.9068	1.0487	1.9068	1.0487	1.9068	**1.0487**	1.9068
Averaged Estimator	**0.8579**	**1.3731**	0.9557	**1.7151**	1.0879	1.9174	1.2266	2.0017

5.8.4 Entropy Estimation

In this section, we consider the problem of entropy estimation described in Section 5.3.4. Similar to covariance estimation, we use two layer neural networks to parameterize functions g, h in Equation (5.19). Implementing the minimization oracle then boils down to finding the parameters of these networks which minimize $\mathbb{E}_{P \sim P_t}\left[R(\hat{f}_{g,h}, P)\right]$. We use stochastic gradient descent to solve this optimization problem.

Baselines We compare the performance of the estimators returned by Algorithm 5 for various values of (n, d), with the plugin MLE estimator $-\sum_{i=1}^{d} \hat{p}_i \log \hat{p}_i$, and the minimax rate optimal estimator of Jiao, Venkat, Han, and Weissman [Jia+15] (JVHW). The plugin estimator is known to be sub-optimal in the high dimensional regime, where $n < d$ [Jia+15].

Results We compare the performance of various estimators based on their worst-case risk computed using DragonFly. Since DragonFly is not guaranteed to compute the worst-case risk, we also compare the estimators based on their risk at randomly generated distributions (see Appendix D.9). Table 5.6 presents the worst-case risk numbers. It can be seen that the plugin MLE estimator has a poor performance compared to JVHW and our estimator. Our estimator has similar performance as JVHW, which is the best known minimax estimator for entropy estimation. We believe the performance of our estimator can be improved with better hyper-parameters.

Table 5.6: Worst-case risk of various estimators for entropy estimation, for various values of (n, d). The worst-case risks are obtained by taking a max of the worst-case risk estimate from DragonFly and the risks computed at randomly generated distributions.

	Worst-case Risk									
	d = 10		**d = 20**		**d = 40**			**d = 80**		
Estimator	**n = 10**	**n = 20**	**n = 20**	**n = 40**	**n = 10**	**n = 20**	**n = 40**	**n = 20**	**n = 40**	**n = 80**
Plugin MLE	0.2895	0.1178	0.2512	0.0347	2.1613	0.8909	0.2710	2.2424	0.9142	0.2899
JVHW [Jia+15]	0.3222	0.0797	**0.1322**	0.0489	0.6788	0.2699	0.0648	**0.3751**	**0.1755**	0.0974
Averaged Estimator	**0.1382**	0.0723	0.1680	0.0439	**0.5392**	**0.2320**	0.0822	0.5084	0.2539	0.0672

5.9 Discussion

We introduced an algorithmic approach for constructing minimax estimators, where we attempt to directly solve the min-max statistical game associated with the estimation problem. This is unlike the traditional approach in statistics, where an estimator is first proposed and then its minimax optimality is certified by showing its worst-case risk matches the known lower bounds for the minimax risk. Our algorithm relies on techniques from online non-convex learning for solving the statistical game and requires access to certain optimization subroutines. Given access to these subroutines, our algorithm returns a minimax estimator and a least favorable prior. This reduces the problem of designing minimax estimators to a purely computational question of efficient implementation of these subroutines. While implementing these subroutines is computationally expensive in the worst case, we showed that one can rely on the structure of the problem to reduce their computational complexity. For the well studied problems of finite Gaussian sequence model and linear regression, we showed that our approach can be used to learn provably minimax estimators in poly(d) time. For problems where provable implementation of the optimization subroutines is computationally expensive, we demonstrated that our framework can still be used together with heuristics to obtain estimators with better performance than existing (up to constant-factor) minimax estimators. We empirically demonstrated this on classical problems such as covariance and entropy estimation. We believe our approach could be especially useful in high-dimensional settings where classical estimators are sub-optimal and not much is known about minimax estimators. In such settings, our approach can provide insights into least favourable priors and aid statisticians in designing minimax estimators.

There are several avenues for future work. The most salient is a more comprehensive understanding of settings where the optimization subroutines can be efficiently implemented.

In this chapter, we have mostly relied on invariance properties of statistical games to implement these subroutines. As described in Section 5.2, there are several other forms of problem structure that can be exploited to implement these subroutines. Exploring these directions can help us construct minimax estimators for several other estimation problems. Another direction for future work would be to modify our algorithm to learn an approximate minimax estimator (*i.e.,* a rate optimal estimator), instead of an exact minimax estimator. There are several reasons why switching to approximate rather than exact minimaxity can be advantageous. First, with respect to our risk tolerance, it may suffice to construct an estimator whose worst-case risk is constant factors worse than the minimax risk. Second, by switching to approximate minimaxity, we believe one can design algorithms requiring significantly weaker optimization subroutines than those required by our current algorithm. Third, the resulting algorithms might be less tailored or over-fit to the specific statistical model assumptions, so that the resulting algorithms will be much more broadly applicable. Towards the last point, we note that our minimax estimators could always be embedded within a model selection sub-routine, so that for any given data-set, one could select from a suite of minimax estimators using standard model selection criteria. Finally, it would be of interest to modify our algorithm to output a single estimator which is simultaneously minimax for various values of n, the number of observations.

Part IV

Boosting

Chapter 6

Generalized Boosting

Boosting is a widely used learning technique in machine learning for solving classification problems. Boosting aims to improve the performance of a weak learner by combining multiple weak classifiers to produce a strong classifier with good predictive performance. Since the seminal works of Schapire [Sch90] and Freund [Fre95], a number of practical algorithms such as AdaBoost [FS+96], gradient boosting [Mas+00], XGBoost [CG16], have been proposed for boosting. Over the years, boosting based methods such as XGBoost in particular, have shown tremendous success in many real-world classification problems, as well as competitive settings such as Kaggle competitions. However, this success is mostly limited to classification tasks involving structured or tabular data with hand-engineered features. On classification problems involving low-level features and complex decision boundaries, boosting tends to perform poorly [BSW14; Pon+17] (also see Section 6.4). One example where this is evident is the image classification task, where the decision boundaries are often complex and the features are low-level pixel intensities. This drawback stems from the fact that boosting builds an additive model of weak classifiers, each of which has very little predictive power. Since such additive models with any reasonable number of weak classifiers are usually not powerful enough to approximate complex decision boundaries, the models' output by boosting tend to have poor performance.

In this chapter, we aim to overcome this drawback of traditional boosting by considering a generalization of boosting which allows for more complex forms of aggregation than linear combinations of weak classifiers. To achieve this goal, we work in the feature representation space and boost the performance of *weak feature transformers*. Working in the representation space allows for more flexible combinations of weak feature transformers. This is unlike traditional boosting which works in the label space and builds an additive model on the predictions of the weak classifiers. The starting point for our approach is the greedy view of boosting, originally studied by Friedman, Hastie, Tibshirani, et al. [FHT+00] and Mason, Baxter, Bartlett, and Frean [Mas+00]. Letting $\widehat{R}_S(f)$ be the risk of a classifier f on training samples S, boosting techniques aim to approximate the minimizer of $\widehat{R}_S$ in terms of linear combinations of elements from a set of weak classifiers $\mathcal{F}$. Many popular boosting algorithms including AdaBoost, XGBoost, rely on greedy techniques to find such an approximation. In our generalized framework for boosting, we take this greedy view, but differ in how we aggregate the weak learners. We approximate the minimizer of $\widehat{R}_S$ using models of the form $f_T = W\phi_T$, where $\phi_T = \sum_{t=0}^{T} g_t$, and $\{g_t\}_{t=0}^{T}$ are feature transformations learned in each iteration of the greedy algorithm, and W is the linear classifier on top of the feature transformation. Unlike additive boosting, where each g_t comes from a fixed weak feature transformer class $\mathcal{G}$, in our framework each g_t comes from a class $\mathcal{G}_t$ which evolves over time t and is allowed to depend on the past iterates $\{\phi_i\}_{i=0}^{t-1}$. Some potential choices for $\mathcal{G}_t$ that could be of interest are $\{g \circ \phi_{t-1} \text{ for } g \in \mathcal{G}\}$, $\{g \circ ([\phi_0, \dots, \phi_{t-1}]) \text{ for } g \in \mathcal{G}\}$, where $g \circ \phi(\mathbf{x}) = g(\phi(\mathbf{x}))$ denotes function composition of g and ϕ, and $\mathcal{G}$ is a weak feature transformer class. Note that the former choice of $\mathcal{G}_t$ is connected to layer-by-layer training of models with ResNet architecture [He+16].

As one particular instantiation of our framework, we consider weak feature transformers that are neural networks and use function compositions to combine them; that is, we use $\mathcal{G}_t$'s constructed using function compositions. We show that for certain choices of $\mathcal{G}_t$, our framework recovers the layer-by-layer training techniques developed in deep learning [Ben+07; Hua+17a]. Greedy layer-by-layer training techniques have seen a revival in recent years [Che+18; Hua+17a; BEO18; NS18; LOV19]. One reason for this revival is that greedy techniques consume less memory than end-to-end training of deep networks, as they do not perform end-to-end back-propagation. Consequently, they can accommodate much larger models in limited memory. As a primary contribution of the work, we identify several drawbacks of existing layer-by-layer training techniques, and show that the choice of $\mathcal{G}_t$ used by these algorithms can lead to a drop in performance. We propose alternative choices for $\mathcal{G}_t$ which fix these issues and empirically demonstrate that the resulting algorithms have superior performance over existing layer-by-layer training techniques, and in some cases achieve performance close to that of end-to-end trained DNNs. Moreover, we show that the proposed algorithms perform much better than traditional additive boosting algorithms, on a variety of classification tasks.

As the second contribution of the work, we provide excess risk bounds for models learned using our generalized boosting framework. Our results depend on a certain weak learning condition on feature transformer classes $\{\mathcal{G}_t\}_{t=1}^{T}$, which is a natural generalization of the weak learning condition that is typically imposed in traditional boosting. The resulting risk bounds are modular and depend on the generalization bounds of $\{\mathcal{G}_t\}_{t=1}^{T}$. An advantage of such modular bounds is that one can rely on the best-known generalization bounds for weak transformation classes $\{\mathcal{G}_t\}_{t=1}^{T}$ and obtain tight risk bounds for boosting. As an immediate consequence of this result, we obtain excess risk bounds for existing greedy layer-by-layer training techniques.

Related Work. Several works have proposed generalizations of traditional boosting [GB11; CMS14; Cor+17; Hua+17a]. Cortes, Mohri, and Syed [CMS14] propose a boosting algorithm where the hypothesis set of weak classifiers is chosen adaptively. However, the resulting models are still additive models of weak classifiers and usually perform poorly on hard classification problems. Several recent works have attempted to learn neural networks greedily based on boosting theory. Cortes, Gonzalvo, Kuznetsov, Mohri, and Yang [Cor+17] propose a boosting-style algorithm to learn both the structure and weights of neural networks in an adaptive way. However, the algorithms developed are restricted to feed forward neural networks and are mostly theoretical in nature. The experimental evidence in the chapter is a proof-of-concept and only considers small scale binary classification tasks. Huang, Ash, Langford, and Schapire [Hua+17a] and Nitanda and Suzuki [NS18] use ideas from classical boosting to learn neural networks in a layer-by-layer fashion. As we show later, these algorithms are specific instances of our generalized framework, and have certain drawbacks arising from the choice of $\mathcal{G}_t$ they use.

6.1 Preliminaries

In this section, we set up the notation and review the necessary background on additive boosting. A consolidated list of notation can be found in Appendix E.1.

Notation. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ denote a feature-label pair following a probability distribution P. Let P^X, P^Y denote the marginal distributions of X and Y. In this chapter, we consider the multi-class classification problem where $\mathcal{Y} = \{0, \dots K-1\}$, and assume $\mathcal{X} \subseteq \mathbb{R}^d$. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n i.i.d samples drawn from P. Let P_n be the empirical distribution of S and P_n^X, P_n^Y be the marginal distributions of $\{\mathbf{x}_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$.

In classification, our goal is to find a predictor that can well predict the label of any feature from just the samples S. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote a score-based classifier which assigns X to class $\operatorname{argmax}_i f_i(X)$. The expected classification risk of f is defined as $\mathbb{E}_{X,Y}\left[\ell_{0-1}(f(X), Y)\right]$, where $\ell_{0-1}(f(X), Y) = 0$ if $\operatorname{argmax}_i f_i(X) = Y$, and 1 otherwise. Since optimizing 0/1 risk is computationally intractable, we consider convex surrogates of $\ell_{0-1}(f(X), Y)$, which we denote by $\ell(f(X), Y)$; typical choices for ℓ include the logistic loss and the exponential loss. The population risk of f is then defined as $R(f) = \mathbb{E}_{X,Y}\left[\ell(f(X), Y)\right]$. Since directly optimizing the population risk is impossible, we approximate it with the empirical risk $\widehat{R}_S(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i)$ and try to find its minimizer.

We consider classifiers of the form $f(X) = W\phi(X)$, where $\phi : \mathcal{X} \to \mathbb{R}^D$ is the feature transformer and $W \in \mathbb{R}^{K \times D}$ is the linear classifier on top. A popular choice for ϕ is a neural network. We denote the population and empirical risks of such an f as $R(W, \phi), \widehat{R}_S(W, \phi)$. We usually work in the space of feature transforms. Let $L_2(P)$ denote the space of square integrable functions w.r.t P, and define the inner product between $\phi_1, \phi_2 \in L_2(P)$ as $\langle \phi_1, \phi_2 \rangle_P = \mathbb{E}_{X \sim P}\left[\langle \phi_1(X), \phi_2(X) \rangle\right]$. We denote with $\nabla_\phi R(W, \phi)$ the functional gradient of $R(W, \phi)$ w.r.t ϕ in the $L_2(P^X)$ space, which is defined as $\nabla_\phi R(W, \phi)(\mathbf{x}) = \mathbb{E}_{Y|\mathbf{x}}\left[W^T \nabla \ell(W\phi(\mathbf{x}), Y)\right]$, where $\nabla \ell(W\phi(\mathbf{x}), y)$ denotes the gradient of ℓ w.r.t its first argument, evaluated at $W\phi(\mathbf{x})$. Similarly, we let $\nabla_\phi \widehat{R}_S(W, \phi)$ denote the functional gradient of $\widehat{R}_S(W, \phi)$ in the $L_2(P_n^X)$ space

$$\nabla_\phi \widehat{R}_S(W, \phi)(\mathbf{x}) = \begin{cases} W^T \nabla \ell(W\phi(\mathbf{x}_i), y_i), & \text{if } \mathbf{x} = \mathbf{x}_i, \\ 0 & \text{otherwise} \end{cases}.$$

Additive Boosting. In this chapter, we refer to traditional boosting as additive boosting, as it constructs additive models of weak classifiers. Let $\mathcal{F}$ be a hypothesis class of weak classifiers, a typical example being decision trees of bounded depth. Additive boosting aims to find an element in the linear span of $\mathcal{F}$ which minimizes the empirical risk

$\widehat{R}_S(f)$. As previously mentioned, there exists a duality between boosting and greedy algorithms [FHT+00; Fri01; Mas+00]. Many popular boosting algorithms use a greedy forward stagewise approach to find a minimizer of $\widehat{R}_S(f)$, and solve the following in each iteration:

$$\eta_t, f_t = \mathrm{argmin}_{\eta \in \mathbb{R}, f \in \mathcal{F}} \widehat{R}_S \left(\sum_{i=1}^{t-1} \eta_i f_i + \eta f \right),$$

where η is the learning rate. Various algorithms differ in how they solve this optimization problem. In gradient boosting, one uses a linear approximation of $\widehat{R}_S$ around $\sum_{i=1}^{t-1} \eta_i f_i$ [Mas+00]. In this chapter, we take this greedy view of boosting to design the generalized boosting framework.

Additive Representation Boosting. In this chapter, we perform boosting in the representation space, contrasting with traditional boosting which works in the output space. Let $\mathcal{G}$ be a hypothesis class of *weak feature transformers*, whose examples include the set of one layer neural networks of bounded width and a set of vector-valued polynomials of bounded degree. More generally, $\mathcal{G}$ can be any set of non-linear transformations. In additive representation boosting, we aim to find a strong feature transform ϕ in the linear span of $\mathcal{G}$, and a linear predictor $W \in \mathcal{W} \subseteq \mathbb{R}^{K \times D}$ that minimizes $\widehat{R}_S(W, \phi)$. To this end, we consider greedy algorithms that solve the following problem each iteration:

$$W_t, g_t = \mathrm{argmin}_{W \in \mathcal{W}, g \in \mathcal{G}} \widehat{R}_S \left(W, \phi_{t-1} + \eta_t g \right), \tag{6.1}$$

where $\phi_t = \phi_0 + \sum_{i=1}^{t} \eta_i g_i$ with ϕ_0 being the initial feature transformation, and $\{\eta_i\}_{i=1}^{\infty}$ is a predefined learning rate schedule.

6.2 Generalized Boosting

The starting point for our generalized boosting framework is the additive representation boosting described in Section 6.1. Typically, linear combinations of weak feature transformations are not powerful enough to model complex decision boundaries. Consequently, the minimizer of $\widehat{R}_S(W, \phi)$ over the linear span of $\mathcal{G}$ tends to have a high risk. A simple workaround for this issue would be to perform additive boosting with a complex hypothesis class $\mathcal{G}$. For example, if the weak feature transformers are one layer neural networks, then one could increase the complexity of $\mathcal{G}$ by using deeper networks. However, such an alternative has several drawbacks both from an optimization and generalization perspective and defeats the purpose of boosting, which aims to convert weak learners into strong learners. From an optimization perspective, moving to complex $\mathcal{G}$ makes each greedy step harder to optimize. For example, compared to deep neural networks, shallow networks are easier to optimize, require fewer resources, and are easier to analyze or interpret [BEO18]. From a generalization perspective, since the generalization bounds of boosting depend on the complexity of $\mathcal{G}$, larger hypothesis classes can lead to overfitting and poor performance on unseen data.

In this chapter, we are interested in other approaches for increasing the complexity of models produced by boosting, while ensuring the boosting/greedy steps are easy to implement. One way to achieve this is by considering more complex combinations of weak feature transformers than the linear combinations considered in additive representation boosting. Formally, let $\mathcal{G}_t$ denote the hypothesis class of feature transformations used in the t^{th} iteration of boosting. In additive boosting, $\mathcal{G}_t = \mathcal{G}$ for all t. In our generalized boosting framework, we increase the complexity of $\mathcal{G}_t$ by letting it depend on the past iterates $\{\phi_i\}_{i=0}^{t-1}$. Here are some potential choices for $\mathcal{G}_t$, other than the ones stated in the introduction: $\{g \circ (\sum_{i=0}^{t-1} \alpha_i \phi_i), \text{ for } g \in \mathcal{G}, \alpha_i \in \mathbb{R}\}$, $\{g \circ \phi_{t-1} \circ \phi_{t-2} \cdots \circ \phi_0, \text{ for } g \in \mathcal{G}\}$. Depending on the problem domain, one could consider several other ways of constructing $\mathcal{G}_t$ using the past iterates. Note that even with these complex choices of $\mathcal{G}_t$, the greedy steps are easy to implement and only need a weak learner which can identify an element in $\mathcal{G}$ that best fits the data. As a result, this remains in the spirit of boosting and at the same time ensures the models learned are complex enough for real world problems.

We now present our algorithm for generalized boosting (see Algorithm 10). Similar to additive representation boosting, our algorithm proceeds in a greedy fashion. In the t^{th} iteration of the algorithm, we aim to solve the following optimization problem:

$$W_t, g_t = \underset{W \in \mathcal{W}, g \in \mathcal{G}_t}{\operatorname{argmin}} \widehat{R}_S \left(W, \phi_{t-1} + \eta_t g\right). \tag{6.2}$$

We provide two approaches for solving this problem. One is the *exact greedy approach*, which directly solves the optimization problem (Algorithm 11). For problems where direct optimization of Equation (6.2) is difficult[1], we provide an approximate technique which performs functional gradient descent on the objective. In this approach, which we call *gradient greedy approach*, we approximate the objective with the linear approximation of $\widehat{R}_S$ around ϕ_{t-1} (Algorithm 12):

$$\widehat{R}_S \left(W, \phi_{t-1} + \eta_t g\right) \approx \widehat{R}_S \left(W, \phi_{t-1}\right) + \eta_t \langle \nabla_\phi \widehat{R}_S(W, \phi_{t-1}), g \rangle_{P_n^X}.$$

To optimize the linear approximation, we first fix W to W_{t-1} and find a minimizing $g_t \in \mathcal{G}_t$. Intuitively, this step can be seen as finding a g which best aligns with the negative functional gradient of empirical risk at the current iterate. For appropriate choice of learning rate η, moving along g_t results in reduction of $\widehat{R}_S$. Next, we fix g_t and find a linear predictor W which minimizes the empirical risk $\widehat{R}_S(W, \phi_t)$. This alternating optimization of g and W makes the algorithm easy to implement in practice. Moreover, this algorithm is more stable than joint optimization of g and W. We note that such gradient greedy approaches have been developed for traditional boosting [Mas+00].

6.2.1 Compositional Boosting

As one particular instantiation of our framework, we consider $\mathcal{G}_t$'s constructed by composing elements from a weak feature transformer class $\mathcal{G}$ with the past iterates $\{\phi_i\}_{i=0}^{t-1}$ and study the resulting boosting algorithms. We refer to such boosting algorithms as *compositional boosting* algorithms since the strong feature transformer is constructed from weak feature transformer via function composition. When $\mathcal{G}_t = \{g \circ \phi_{t-1} \text{ for } g \in \mathcal{G}\}$, the

[1] Such scenarios can potentially arise if the feature transformations are non-differentiable functions.

models in our framework have the ResNet architecture and can be defined recurrently as $\phi_t = \phi_{t-1} + \eta_t g_t \circ \phi_{t-1}$. Moreover, Algorithm 10 with this choice of $\mathcal{G}_t$ and Algorithm 11 as update routine give us the greedy layer-wise supervised training technique proposed by Bengio, Lamblin, Popovici, and Larochelle [Ben+07] and recently revisited by Belilovsky, Eickenberg, and Oyallon [BEO18]. In another recent work, Huang, Ash, Langford, and Schapire [Hua+17a] propose a boosting-based algorithm for learning ResNets (see Algorithm 17 in Appendix). We now show that their approach is equivalent to the greedy technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07], and thus can be seen as an instance of our general framework. We note that such a connection is not known previously.

Proposition 12. *Suppose the classification loss ℓ is the exponential loss. Then the greedy technique of Huang, Ash, Langford, and Schapire [Hua+17a] for learning ResNets is equivalent to the greedy layer-wise supervised training technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07].*

In another recent work, Nitanda and Suzuki [NS18] propose a gradient boosting technique to greedily learn a ResNet. This algorithm is closely related to the gradient greedy approach described in Algorithm 12, with $\mathcal{G}_t = \{g \circ \phi_{t-1} \text{ for } g \in \mathcal{G}\}$.

Algorithm 10 Generalized Boosting

```
Input: Training data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, iterations $T$, initial linear predictor $W_0$, initial feature transformer $\phi_0$, learning rates $\{\eta_i\}_{i=1}^T$, Update-routine: UPDATE
$t \leftarrow 1$
while $t \leq T$ do
    Construct feature transformer class $\mathcal{G}_t$ based on past iterates $\{(W_i, \phi_i)\}_{i=0}^{t-1}$
    $W_t, \phi_t, g_t \leftarrow$ UPDATE $(S, W_{t-1}, \phi_{t-1}, \eta_t, \mathcal{G}_t)$
    $t \leftarrow t + 1$
end while
Return: $W_T, \phi_T$
```

Algorithm 11 Exact Greedy Update

```
Input: Training data $S$, previous iterate $(W, \phi)$, learning rate $\eta$, feature transformer class $\mathcal{G}$
$W^+, g^+ \leftarrow \operatorname*{argmin}_{\widetilde{W} \in \mathcal{W}, \tilde{g} \in \mathcal{G}} \widehat{R}_S(\widetilde{W}, \phi + \eta \tilde{g})$
$\phi^+ \leftarrow \phi + \eta g^+$
Return: $W^+, \phi^+, g^+$
```

Algorithm 12 Gradient Greedy Update

```
Input: Training data $S$, previous iterate $(W, \phi)$, learning rate $\eta$, feature transformer class $\mathcal{G}$
// Pick a descent direction
$g^+ \leftarrow \operatorname{argmin}_{\tilde{g} \in \mathcal{G}} \langle \nabla_\phi \widehat{R}_S(W, \phi), \tilde{g} \rangle_{P_n^X}$
$\phi^+ \leftarrow \phi + \eta g^+$
// Update the linear predictor
$W^+ \leftarrow \operatorname{argmin}_{\widetilde{W} \in \mathcal{W}} \widehat{R}_S(\widetilde{W}, \phi^+)$
Return: $W^+, \phi^+, g^+$
```

We now highlight certain drawbacks of the existing greedy layer-wise training techniques, which arise from the particular choice of $\mathcal{G}_t$ used by these algorithms. Since $\{g \circ \phi_{t-1} \text{ for } g \in \mathcal{G}\}$ is constructed solely based on the past iterate ϕ_{t-1}, any mistake in ϕ_{t-1} is propagated to all the future iterates. As a result, these algorithms can not recover from their past mistakes. As an example, consider the following scenario where two points $\mathbf{x}_1, \mathbf{x}_2$ belonging to two different classes are placed close to each other in the feature space, after 1^{st} iteration of greedy; that is $\phi_1(\mathbf{x}_1) \approx \phi_1(\mathbf{x}_2)$. In such a scenario, the future iterates $\{\phi_t\}_{t=2}^{\infty}$ generated by existing greedy algorithms will always place $\mathbf{x}_1, \mathbf{x}_2$ close to each other in the representation space. As a result, the algorithm will always misclassify at least one of $\mathbf{x}_1, \mathbf{x}_2$. Another issue with existing greedy techniques is that they do not guarantee that the complexity of $\mathcal{G}_t$ increases with time t. In such scenarios, Algorithm 10 doesn't make much progress in each iteration and can result in poor models. As an example, consider the setting where $\mathcal{G}$ is the set of all linear transformations. Suppose ϕ_0 is the identity transform and ϕ_1 is such that its range lies in a low dimensional subspace. Then, it is evident that $\mathcal{G}_1 \supseteq \mathcal{G}_t$ for all $t \geq 2$.

To fix these issues, we propose two new compositional boosting algorithms obtained with a more careful choice of $\mathcal{G}_t$. In our first algorithm, which we call DenseCompBoost, we choose $\mathcal{G}_t$ as follows

$$\mathcal{G}_t = \left\{ g \circ \left(\mathrm{Id} + \sum_{i=0}^{t-1} \alpha_i \phi_i \right), \text{ for } g \in \mathcal{G}, \alpha_i \in \mathbb{R} \right\}, \tag{6.3}$$

where $\mathrm{Id}(\cdot)$ is the identify function. Such a choice of $\mathcal{G}_t$ helps us recover from the past mistakes. For example, if ϕ_1 is a constant function, then the algorithm can still learn a good feature transformer by relying on the input $\mathbf{x}$ and the initial feature transform ϕ_0. Moreover, our choice of $\mathcal{G}_t$ ensures its complexity grows with t and satisfies: $\mathcal{G}_{t-1} \subseteq \mathcal{G}_t$, for all t. We call our algorithm DenseCompBoost, since the resulting model for this choice of $\mathcal{G}_t$ resembles a DenseNet [Hua+17b], where each layer is allowed to be connected to all the previous layers. That being said, the models output by DenseCompBoost differ from DenseNet in how they aggregate the previous layers. DenseNet concatenates the features from previous layers, whereas DenseCompBoost adds the features. Our second algorithm, which we call CmplxCompBoost, tries to increase the complexity of $\mathcal{G}_t$ in each iteration as follows

$$\mathcal{G}_t = \left\{ g \circ \phi_{t-1}, \text{ for } g \in \widetilde{\mathcal{G}}_t \right\}, \tag{6.4}$$

where $\widetilde{\mathcal{G}}_t$ is a weak feature transformer class and satisfies $\widetilde{\mathcal{G}}_{t-1} \subset \widetilde{\mathcal{G}}_t$ for all t. In the case of one layer neural networks, such $\widetilde{\mathcal{G}}_t$'s can be constructed by increasing the layer width with t. We note that the $\widetilde{\mathcal{G}}_t$ in this algorithm is independent of the past iterates. By increasing the complexity of $\widetilde{\mathcal{G}}_t$ with t, we expect the complexity of $\mathcal{G}_t$ to increase and Algorithm 10 to make more progress in each iteration. While not immediately evident, we note that this technique can also fix the mistakes made by past iterates. For example, suppose ϕ_1 is such that it places two points $\mathbf{x}_1, \mathbf{x}_2$ from different classes, close to each other in the feature space. Then having a more complex $\widetilde{\mathcal{G}}_2$ can help recover from this mistake, as one can potentially find a $g \in \widetilde{\mathcal{G}}_2$ which can separate these two points. In Section 6.4, we present empirical evidence showing that our new boosting algorithms have superior performance over existing additive and compositional boosting algorithms. Further empirical evidence corroborating the issues we identified with existing layer-wise training techniques can be found in Appendix E.10.1.

6.3 Excess Risk Bounds

In this section, we provide excess risk bounds for the models' output by the generalized boosting framework. Our results depend on a *weak learning condition* on the hypothesis class $\mathcal{G}_t$ used in the t^{th} iteration of Algorithm 10. This condition is a way to quantify the relative strength of $\mathcal{G}_t$ and roughly says that there always exists an element in $\mathcal{G}_t$ which has an acute angle with the negative functional gradient at the current iterate. Such a condition ensures progress in each iteration of boosting.

Definition 6.3.1. Let $\beta \in (0, 1], \epsilon \geq 0$ be constants. $\mathcal{G}_{t+1}$ is said to satisfy the (β, ϵ)-weak learning condition for a dataset S, if there exists a $g \in \mathcal{G}_{t+1}$ such that

$$\langle g, -\nabla_\phi \widehat{R}_S(W_t, \phi_t)\rangle_{P_n^X} \geq \beta B(\mathcal{G}_{t+1}) \|\nabla_\phi \widehat{R}_S(W_t, \phi_t)\|_{P_n^X} - \epsilon,$$

where $B(\mathcal{G}_{t+1}) = \sup_{g\in\mathcal{G}_{t+1}} \|g\|_{P_n^X}$, and P_n is the empirical distribution of S.

In traditional boosting, such conditions are typically referred to as the *edge* of a weak learner and play a crucial role in the convergence analysis. For example, Freund and Schapire [FS95] assume that for any set of weights over the training set S, there exists a classifier in the hypothesis class of weak classifiers which has better than random accuracy on the weighted samples. The following proposition shows that their condition is closely related to Definition 6.3.1.

Proposition 13. *For binary classification, the weak learning condition of Freund and Schapire [FS95] satisfies the empirical weak learning condition in Definition 6.3.1, albeit in the label space.*

For binary classification problems, it is well known that the weak learning condition of [FS95] is the weakest condition under which boosting is possible [FS96; RW05]. This, together with the above proposition, suggests that our weak learning condition in Definition 6.3.1 cannot be weakened for binary classification problems.

To begin with, we derive excess risk bounds for the gradient greedy approach. Our analysis crucially relies on the observation that it can be viewed as performing inexact gradient descent on the population risk R. Several recent works have analyzed inexact gradient descent on convex objectives [SRB11; Tem14; DGN14; BWY+17]. However, the condition on the inexact gradient imposed by these works is different from ours and in many cases is stronger than our condition. For example, the condition of Balakrishnan, Wainwright, Yu, et al. [BWY+17] translates to $\|g + \nabla_\phi R(W, \phi)\|_{P^X} \leq \epsilon$ in our setting, which is stronger than our weak learning condition. So the core of our analysis focuses on understanding inexact gradient descent with descent steps satisfying the weak learning condition in Definition 6.3.1. In our analysis, we consider a sample-splitting variant of the algorithm, where in each iteration we use a fresh batch of samples. This is mainly done to simplify the analysis by avoiding complex statistical dependencies between the iterates of the algorithm. Let $\tilde{n} = \lfloor \frac{n}{T} \rfloor$, we split the training dataset S into T subsets $\{S_t\}_{t=1}^T$ of size $\tilde{n}$, where $S_t = \{(\mathbf{x}_{t,i}, y_{t,i})\}_{i=1}^{\tilde{n}}$. We work with the subset S_t in the t^{th} iteration of Algorithm 10. We are now ready to state our main result on the excess risk bounds of the iterates of Algorithm 12. Our results depend on the Rademacher complexity terms related to the hypothesis sets $\mathcal{W}, \mathcal{G}_t$

$$\mathcal{R}\left(\mathcal{W}, \mathcal{G}_t\right) = \mathbb{E}\left[\sup_{\substack{W \in \mathcal{W}, \\ g \in \mathcal{G}_t}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \rho_{ik}[Wg(\mathbf{x}_{t,i})]_k\right], \ \mathcal{R}\left(\mathcal{G}_t\right) = \mathbb{E}\left[\sup_{g \in \mathcal{G}_t} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{D} \rho_{ij}[g(\mathbf{x}_{t,i})]_j\right],$$

where $[\mathbf{u}]_k$ denotes the k^{th} entry of a vector $\mathbf{u}$, and the expectation is taken w.r.t the randomness from S_t and the Rademacher random variables ρ_{ij}'s.

Theorem 21 (Gradient Greedy)**.** *Suppose the classification loss ℓ is L-Lipschitz and M-smooth w.r.t the first argument. Let the hypothesis set of linear predictors $\mathcal{W}$ be s.t. any $W \in \mathcal{W}$ satisfies $\lambda_{min}\left(WW^T\right) \geq \sigma^2_{min} > 0$ and $\lambda_{max}\left(WW^T\right) \leq \sigma^2_{max}$. Moreover, suppose for all t, $\mathcal{G}_t$ satisfies the (β, ϵ_t)-weak learning condition of Definition 6.3.1 for any dataset S_t. Finally, suppose any $g \in \mathcal{G}_t$ is bounded with $\sup_X \|g(X)\|_2 \leq B$. Let the learning rates $\{\eta_t\}_{t=1}^{\infty}$ be chosen as $\eta_t = ct^{-s}$, for some $s \in \left(\frac{\beta+1}{\beta+2}, 1\right)$ and positive constant c. If Algorithm 10 is run for T iterations with Algorithm 12 as update routine, then (W_T, ϕ_T), the T^{th} iterate output by the algorithm, satisfies the following risk bound for any W^*, ϕ^* and $\alpha \in (0, \beta(1-s))$, with probability at least $1-\delta$ over datasets of size n*

$$R(W_T, \phi_T) \leq R(W^*, \phi^*) + O\left(\frac{1}{T^{\alpha}} + T^{2-s}\sqrt{\frac{\log\frac{T}{\delta}}{\tilde{n}}}\right) + 2\sum_{t=1}^{T}\eta_t\left(L\mathcal{R}\left(\mathcal{W}, \mathcal{G}_t\right) + L\mathcal{R}\left(\mathcal{G}_t\right) + \epsilon_t\right).$$

Proof Sketch. We first show that Algorithm 12 can be viewed as performing inexact gradient descent on the population risk R. Specifically, we show that with high probability, the t^{th} iterate g_t satisfies

$$\langle g_t, -\nabla_{\phi}R(W_{t-1}, \phi_{t-1})\rangle_P \geq \beta B\|\nabla_{\phi}R(W_{t-1}, \phi_{t-1})\|_P - \epsilon_t - \zeta_t,$$

for some $\zeta_t > 0$. This follows from the weak learning condition satisfied by $\mathcal{G}_t$. Ignoring ϵ_t, ζ_t, the above equation shows that g_t makes acute angle with the population functional gradient at ϕ_{t-1}. Consequently, we would expect the population risk to decrease, if we move along g_t. This is indeed the case, and the final step in the proof formalizes this intuition. □

Remarks: We now briefly discuss the above result. See Appendix E.4 for more discussion.

- The reference classifier (W^*, ϕ^*) in the above bound can be any classifier, as long as $\|W^*\|_2 < \infty, \|\phi^*\|_{P^X} < \infty$. In particular, if there exists a Bayes optimal classifier satisfying this condition, then the above Theorem provides an excess risk bound w.r.t the Bayes optimal classifier.
- The $T^{-\alpha}$ term in the bound corresponds to the *optimization error*. The $\eta_t\epsilon_t$ term corresponds to the *approximation error* and the rest of the terms correspond to the *generalization error*. As T increases, the optimization error goes down, and as $\tilde{n}$ increases, the generalization error goes down. If there is no approximation error, that is $\epsilon_t = 0$ for all t, then the excess risk goes down to 0 as $\tilde{n}, T \to \infty$ at appropriate rate.

- If $\beta = 1$, then for appropriate choice of step size the optimization error goes down as $O\left(T^{-1/3+\gamma}\right)$, for some arbitrarily small $\gamma > 0$. This rate is slower than the $O(T^{-1})$ rates for inexact gradient descent obtained by Schmidt, Roux, and Bach [SRB11] and Devolder, Glineur, and Nesterov [DGN14]. However, we note that unlike our work, these works assume that the level sets of the objective are bounded. Under the assumption that the level sets of population risk are bounded, the optimization error in Theorem 21 can be improved to $O(T^{-1})$. However, such a condition need not hold in the our setting.
- Note that the risk bounds are modular and only depend on the Rademacher complexity terms $\mathcal{R}(\mathcal{W}, \mathcal{G}_t), \mathcal{R}(\mathcal{G}_t)$ which capture the complexity of $\mathcal{G}_t$. To instantiate Theorem 21 for specific choices of $\mathcal{G}_t$, we need to bound these two complexity terms.

We now extend the analysis of Theorem 21 to the exact greedy approach.

Corollary 5 (Exact Greedy)**.** *Consider the setting of Theorem 21. Suppose Algorithm 10 is run with Algorithm 11 as update routine. Then (W_T, ϕ_T), the T^{th} iterate output by the algorithm, satisfies the same risk bounds as gradient greedy algorithm in Theorem 21.*

In the rest of the section, we instantiate Theorem 21 for specific choices of $\mathcal{G}_t$. We first consider the additive representation boosting algorithm.

Corollary 6. *Consider the setting of Theorem 21 and consider the additive representation boosting algorithm, where $\mathcal{G}_t = \mathcal{G}$ for all t. Suppose $\mathcal{G}$ is the set of one layer neural networks with sigmoid activation functions: $\mathcal{G} = \left\{\sigma(C\mathbf{x}),\ \text{for } C \in \mathbb{R}^{D\times d}, \|C_{i,*}\|_1 \le \Lambda, \forall i\right\}$. Moreover, suppose the feature domain $\mathcal{X}$ is a subset of $[0,1]^d$. Then the T^{th} iterate output by Algorithm 10, with Algorithm 11 or 12 as update routine, satisfies the following risk bound for any (W^*, ϕ^*), with probability at least $1-\delta$*

$$R(W_T, \phi_T) \le R(W^*, \phi^*) + O\left(\frac{1}{T^\alpha}\right) + 2\sum_{t=1}^{T} \eta_t \epsilon_t + O\left(\frac{KD\Lambda T^{1-s}\log D}{\sqrt{\tilde{n}}} + T^{2-s}\sqrt{\frac{\log\frac{T}{\delta}}{\tilde{n}}}\right).$$

Next, we consider the layer-by-layer fitting technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07].

Corollary 7. *Consider the setting of Corollary 6 and consider the layer-by-layer training technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07], where $\mathcal{G}_t = \{g \circ \phi_{t-1}$ for $g \in \mathcal{G}\}$. Suppose $\mathcal{G}$ is the set of one layer neural networks with sigmoid activation functions: $\mathcal{G} = \left\{\sigma(C\mathbf{x}),\ \text{for } C \in \mathbb{R}^{D\times D}, \|C_{i,*}\|_1 \le \Lambda, \forall i\right\}$. Then the T^{th} iterate output by Algorithm 10, with Algorithm 11 or 12 as update routine, satisfies the following risk bound for any (W^*, ϕ^*) with probability at least $1-\delta$*

$$R(W_T, \phi_T) \le R(W^*, \phi^*) + O\left(\frac{1}{T^\alpha}\right) + 2\sum_{t=1}^{T} \eta_t \epsilon_t + O\left(\frac{KD\Lambda T^{2-2s}\log D}{\sqrt{\tilde{n}}} + T^{2-s}\sqrt{\frac{\log\frac{T}{\delta}}{\tilde{n}}}\right).$$

Note that the generalization and optimization errors for both additive feature boosting and layer-by-layer fitting have similar dependence on $T, \tilde{n}$. However, the latter tends to have a smaller approximation error (ϵ_t) as it is able to build complex $\mathcal{G}_t$'s over time. So one would expect layer-by-layer fitting to output models with a better population risk, which our empirical results in fact verify.

6.4 Experiments

In this section, we present experiments comparing the performance of various boosting techniques on both simulated and benchmark datasets.

Baselines. We compare our proposed boosting techniques with XGBoost, AdaBoost, additive representation boosting (discussed in Corollary 6) and greedy layer-by-layer training technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07] (Corollary 7). XGBoost uses decision trees as weak classifiers. For AdaBoost, we use 1 hidden layer neural networks as weak classifiers. We use two kinds of neural networks, based on the dataset. For tabular datasets, we use fully connected networks and for image datasets, we use convolutional networks (CNN) with the convolution block made up of *Convolution, BatchNorm, ReLU* layers arranged sequentially. For additive representation boosting (Additive Feature Boost from now on) and layer-by-layer fitting (StdCompBoost from now on), the weak feature transformer class $\mathcal{G}$ consists of one layer neural network transformations. Similar to AdaBoost, we use two kinds of transformations: a) fully connected transformations of the form $g(\mathbf{x}) = \text{ReLU}(C\mathbf{x} + \mathbf{d})$, and b) convolutional transformations with *Convolution, BatchNorm, ReLU* blocks arranged sequentially. Finally, we also compare against end-to-end training of ResNets.

Proposed Techniques. For DenseCompBoost, we use a slight variant of $\mathcal{G}_t$ defined in Equation (6.3) : $\mathcal{G}_t = \{h + g \circ (\sum_{i=0}^{t-1} \alpha_i \phi_i),$ for $h \in \mathcal{H}, g \in \mathcal{G}, \alpha_i \in \mathbb{R}\}$, where $\mathcal{H}, \mathcal{G}$ are weak feature transformer classes. We use this variant because the dimensions of the input feature space and the representation space need not be the same, and as a consequence $\mathcal{G}_t$ in Equation (6.3) can not always be used. Similar to StdCompBoost, we consider two choices for $\mathcal{H}, \mathcal{G}$: one based on fully connected blocks and the other based on convolution blocks. For CmplxCompBoost, we again consider two choices for the weak transformer class $\tilde{\mathcal{G}}_t$ in Equation (6.4): a) $\text{ReLU}(C\mathbf{x} + \mathbf{d})$ with $C \in \mathbb{R}^{D_t \times D_{t-1}}$, where $D_t = D_{t-1} + \Delta$ for some positive constant Δ, and b) convolution blocks with number of output channels equal to the number of input channels plus a constant Δ. This choice of feature transformers ensures the complexity of $\tilde{\mathcal{G}}_t$ increases with t. We use exact greedy updates (Algorithm 11) for both of our proposed methods and set learning rate η_t to 1. We do not present experimental results for Algorithm 12, which we noticed has marginally worse performance than Algorithm 11.

6.4.1 Simulated Datasets

Datasets. In this section we compare the techniques described above on simulated datasets. We generated 3 synthetic binary classification datasets in $\mathbb{R}^{32}$. Simulation 1 is a concentric ellipsoids dataset, where a point $\mathbf{x}$ is classified based on $\mathbf{x}^T A\mathbf{x}$, for some randomly generated positive semidefinite matrix A. Simulations 2, 3 are datasets whose classification boundaries are polynomials of degrees 8 and 9 respectively. For each of these datasets, we generated 10^6 samples for training and testing.

Hyper-parameters. We used hold-out set validation to pick the best hyper-parameters for all the methods. We used 20% of the training data as validation data and picked the best parameters using grid search, based on validation accuracy. After picking the best parameters, we train on the entire training data and report performance on the test data.

For all the greedy techniques based on neural networks, we used fully connected blocks and tuned the following parameters: weight decay, width of weak feature transformers, number of boosting iterations T, which we upper bound by 15. For CmplxCompBoost, we set $\Delta = D_0/5$. For end-to-end training, we tuned weight decay, width of layers, depth. We used SGD for optimization of all these techniques. The number of epochs and step size schedule of SGD are chosen to ensure convergence. For XGBoost, we tuned the number of trees, depth of each tree, learning rate. The exact values of hyper-parameters tuned for each of the methods can be found in Appendix E.10.

Results. Table 6.1 presents the results from our experiments. Both CmplxCompBoost and StdCompBoost largely outperform the additive boosting methods, with CmplxCompBoost being slightly better due to the increasing complexity in $\tilde{G}_t$. Notably, DenseCompBoost performs significantly better than the rest and is able to bridge the gap between StdCompBoost and End-to-End. We attribute its success to its ability to recover from earlier mistakes: while StdCompBoost or CmplxCompBoost necessarily accumulate errors at each layer, DenseCompBoost is further connected to earlier layers, allowing it to undo its past mistakes.

6.4.2 Benchmark Datasets

Datasets. In this section, we compare various techniques on the following image datasets: CIFAR10, MNIST, FashionMNIST [XRV17], MNIST-rot-back-image [Lar+07], convex [XRV17], SVHN [Net+11], and the following tabular datasets from UCI repository [BM98]: letter recognition [FS91], forest cover type (covtype), connect4. The convex dataset involves classifying shapes in images as either convex or non-convex.

Hyper-parameters. For covtype dataset, which doesn't come with a test set, we randomly sample 20% of the original data and use it as the test set. We use a similar hyper-parameter selection technique as above and tune the same set of hyper-parameters as described above. We use convolution blocks for CIFAR10, SVHN, FashionMNIST, convex, MNIST-rot-back-image and fully connected blocks for the rest. We limit the width of fully connected blocks to 4096, and the number of output channels in convolution blocks to 128 while tuning the hyper-parameters for the composition boosting techniques and end-to-end training. For AdaBoost and additive representation boosting, we set these limits to 16000 and 350 respectively. For CmplxCompBoost with convolution blocks, we set $\Delta = D_0/8$. We *do not* use data augmentation in our experiments.

Results. Table 6.2 presents the results from our experiments. It can be seen that on image classification tasks, additive boosting techniques have poor performance. Among compositional boosting methods, StdCompBoost performs the worst. While DenseCompBoost performs comparably to CmplxCompBoost on image datasets, it is better on tabular data. We believe a hybrid of DenseCompBoost and CmplxCompBoost algorithms can achieve better performance than either of the algorithms.

Table 6.1: Test accuracy of various boosting techniques on synthetic datasets. Numbers in bold indicate the best performance among various greedy techniques.

Technique	Simulation 1	Simulation 2	Simulation 3
XGBoost (Trees)	84.40	97.59	50.10
AdaBoost (1 NN)	67.90	93.73	72.64
Additive Feature Boost	88.49	93.91	73.13
StdCompBoost	91.53	96.95	82.49
DenseCompBoost	**93.55**	**98.35**	**95.70**
CmplxCompBoost	91.97	97.22	82.52
End-to-End	93.88	98.35	99.09

Table 6.2: Test accuracy of various boosting techniques on benchmark datasets. We use convolution blocks for the first 5 datasets and fully connected blocks for the other datasets.

Technique	SVHN	FashionMNIST	CIFAR10	Convex	MNIST-rot-back-image	MNIST	Letter	CovType	Connect4
XGBoost (Trees)	77.72	90.34	58.34	82.29	53.89	97.96	96.16	**97.46**	86.63
AdaBoost (1 NN)	82.88	88	72.78	86.17	50.02	98.27	92.08	90.95	86.39
Additive Feature Boost	83.36	89.95	74.33	89.30	54.31	98.27	90.86	93.12	86.58
StdCompBoost	90.81	92.77	81.93	98.19	73.17	98.37	96.43	95.61	86.33
DenseCompBoost	91.03	**93.17**	82.31	**98.6**	73.1	98.34	**96.96**	96.28	**86.85**
CmplxCompBoost	**91.25**	**93.18**	**82.43**	98.52	**74.32**	98.34	96.66	95.92	86.49
End-to-End	94.82	93.49	86.88	98.81	82.69	98.95	97.67	96.86	87.37

6.5 Discussion

We proposed a generalized framework for boosting, which allows for more complex forms of aggregation of weak learners than traditional boosting. Our generalized framework allows to derive learning algorithms that (a) have performance close to that of end-to-end trained DNNs, and (b) come with strong theoretical guarantees. Additive boosting algorithms do not satisfy property (a), while DNNs do not satisfy property (b). In particular, additive boosting algorithms, even with small neural networks as their weak classifiers, do not have the strong performance of end-to-end trained DNNs. Improving their performance requires the hypothesis space to increase in complexity while not increasing sample complexity of each boosting step too greatly, which can be achieved by our generalized boosting framework. One particular instantiation of our framework is aggregation using function compositions. A number of existing greedy techniques for learning neural networks fall into our framework, and our analysis allowed us to delineate some of their key flaws, then consequently, propose new techniques which improve upon them. We believe our work opens up a new line of inquiry for greedy learning of highly flexible models with rigorous theoretical guarantees, by leveraging the theory of boosting and generalized greedy algorithms in function spaces. We moreover believe our work has the potential to bridge the gap in performance between existing greedy layer-by-layer training techniques and end-to-end training of deep networks.

Chapter 7

Conclusion and Future Work

Statistical Game Theory. This thesis aims to lay the foundations of statistical game theory to study several classical and modern statistical problems. Our algorithmic contributions to statistical game theory are primarily driven by two challenges that often arise while studying statistical games: (a) large domains, and (b) nonconcave utility functions. To this end, in Chapters 2, 3, we designed computationally efficient algorithms for finding mixed strategy Nash equilibrium of games with nonconcave utilities. A number of open questions need to be solved to adequately address the above two challenges. The most important of these is the need for efficient derivative-free optimization techniques that scale well to high dimensional problems, and require very few function evaluations. Existing techniques for derivative-free optimization do not satisfy these desiderata: random walk based approaches such as Simulated Annealing require too many function evaluations and model-based approaches such as Gaussian Process optimization don't scale well to high dimensional problems. In Chapter 4, we took a step towards this goal by designing efficient derivative-free optimization techniques for convex quadratic loss functions. However, this is a very restrictive setting and a lot of work remains to be done to derive efficient derivative-free optimization techniques for nonconvex losses. To this end, in an ongoing work, we are relying on the insights gained from studying quadratic losses to develop bandit Newton methods for nonconvex losses.

There are several other avenues for future research in statistical game theory. Depending on the specific statistical application that is being studied, one might face various algorithmic and analytic challenges. It is important to identify these challenges and come up with appropriate tools to handle them. For example, for the problem of minimax statistical estimation, it often suffices to learn approximate minimax estimators (*i.e.,* rate optimal estimators), instead of exact minimax estimators. To construct such approximate minimax estimators, we need new game-theoretic tools that help us find an approximate NE whose value is constant factors away from the minimax value of the game. As another example, consider the problem of robust machine learning. Machine learning practitioners often prefer deterministic classifiers over randomized classifiers. In such cases, pure strategy equilibrium turns out to be a more appropriate solution concept to study than mixed strategy NE considered in this thesis. Studying this solution concept would require new game-theoretic tools that can compute pure strategy equilibria in games with nonconcave utility functions.

Statistical Applications. In this thesis, we mainly focused on the following two classical statistical applications: minimax statistical estimation and boosting. In minimax statistical estimation, we utilized our game-theoretic tools to construct minimax estimators for various problems such as mean estimation, regression, and entropy estimation. We believe our algorithmic tools (in combination with problem structure) can help construct minimax estimators for numerous other problems. In an ongoing work, we are designing algorithmic minimax estimators for fundamental problems such as sparse mean estimation, sparse linear regression. For the problem of boosting, the algorithms we developed in Chapter 6 don't yet match the performance of end-to-end trained neural networks. To truly bridge the gap in performance between boosting and neural networks, we hypothesize that one has to look at the game-theoretic viewpoint of boosting. Consequently, in an ongoing work, we are developing generalized boosting algorithms from a game-theoretic perspective.

There are several other emerging problems in modern machine learning that can be studied from a game-theoretic perspective. Some of these include robustness, GANs, and algorithmic fairness. Many existing algorithms for these problems rely on heuristics to solve the associated games. These heuristic approaches are not always guaranteed to find an optimal solution. So it is important to understand the drawbacks of these heuristics and come up with algorithms that improve upon them.

Bibliography

This bibliography contains 169 references.

[AB+09] Jean-Yves Audibert, Sébastien Bubeck, et al. "Minimax Policies for Adversarial and Stochastic Bandits." In: *COLT*. Vol. 7. 2009, pp. 1–122.

[AB10] Jean-Yves Audibert and Sébastien Bubeck. "Regret bounds and minimax policies under partial monitoring". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2785–2836.

[ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. "Optimal Algorithms for Online Convex Optimization with Multi-Point Bandit Feedback." In: *COLT*. Citeseer. 2010, pp. 28–40.

[AG12] Shipra Agrawal and Navin Goyal. "Analysis of thompson sampling for the multi-armed bandit problem". In: *Conference on learning theory*. 2012, pp. 39–1.

[Aga+11] Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. "Stochastic convex optimization with bandit feedback". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1035–1043.

[Aga+17] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. "Corralling a band of bandit algorithms". In: *Conference on Learning Theory*. PMLR. 2017, pp. 12–38.

[AGH19] Naman Agarwal, Alon Gonen, and Elad Hazan. "Learning in Non-convex Games with an Optimization Oracle". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 18–29. URL: http://proceedings.mlr.press/v99/agarwal19a.html.

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. "The multiplicative weights update method: a meta-algorithm and applications". In: *Theory of Computing* 8.1 (2012), pp. 121–164.

[AHR09] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. "Competing in the dark: An efficient algorithm for bandit linear optimization". In: (2009).

[ALL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers". In: *Advances in neural information processing systems*. 2019, pp. 6155–6166.

[ALT15] Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. "Fighting bandits with a new kind of smoothness". In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 2197–2205.

[ALT16] Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. "Perturbation techniques in online learning and optimization". In: *Perturbations, Optimization, and Statistics* (2016), p. 233.

[AR09] Jacob Abernethy and Alexander Rakhlin. "Beating the adaptive bandit with high probability". In: *2009 Information Theory and Applications Workshop*. IEEE. 2009, pp. 280–289.

[Aue+02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. "The nonstochastic multiarmed bandit problem". In: *SIAM journal on computing* 32.1 (2002), pp. 48–77.

[Ban+05] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. "Clustering on the unit hypersphere using von Mises-Fisher distributions". In: *Journal of Machine Learning Research* 6.Sep (2005), pp. 1345–1382.

[Bar+08] Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. "High-probability regret bounds for bandit online linear optimization". In: *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008.* Omnipress. 2008, pp. 335–342.

[BC12] Sébastien Bubeck and Nicolo Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *arXiv preprint arXiv:1204.5721* (2012).

[Bel+15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. "Escaping the local minima via simulated annealing: Optimization of approximately convex functions". In: *Conference on Learning Theory.* 2015, pp. 240–265.

[Ben+07] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. "Greedy layer-wise training of deep networks". In: *Advances in neural information processing systems.* 2007, pp. 153–160.

[BEO18] Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. "Greedy Layerwise Learning Can Scale to ImageNet". In: *arXiv preprint arXiv:1812.11446* (2018).

[Ber73] Dimitri P Bertsekas. "Stochastic optimization problems with nondifferentiable cost functionals". In: *Journal of Optimization Theory and Applications* 12.2 (1973), pp. 218–231.

[Ber85] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 1985.

[Ber90] J Calvin Berry. "Minimax estimation of a bounded normal mean vector". In: *Journal of Multivariate Analysis* 35.1 (1990), pp. 130–139.

[Bic81] PJ Bickel. "Minimax estimation of the mean of a normal distribution when the parameter space is restricted". In: *The Annals of Statistics* 9.6 (1981), pp. 1301–1309.

[Bir83] Lucien Birgé. "Approximation dans les espaces métriques et théorie de l'estimation". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 65.2 (1983), pp. 181–237.

[BLE17] Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. "Kernel-based methods for bandit convex optimization". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 72–85.

[BM02] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[BM93] Lucien Birgé and Pascal Massart. "Rates of convergence for minimum contrast estimators". In: *Probability Theory and Related Fields* 97.1-2 (1993), pp. 113–150.

[BM98] Catherine L Blake and Christopher J Merz. *UCI repository of machine learning databases, 1998*. 1998.

[BP73] Lawrence D Brown and R Purves. "Measurable selections of extrema". In: *The annals of statistics* 1.5 (1973), pp. 902–912.

[BSW14] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. "Searching for exotic particles in high-energy physics with deep learning". In: *Nature communications* 5 (2014), p. 4308.

[But+18] Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. "Variable selection with Hamming loss". In: *The Annals of Statistics* 46.5 (2018), pp. 1837–1875.

[BWY+17] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *The Annals of Statistics* 45.1 (2017), pp. 77–120.

[CB94] Bertrand S Clarke and Andrew R Barron. "Jeffreys' prior is asymptotically least favorable under entropy risk". In: *Journal of Statistical planning and Inference* 41.1 (1994), pp. 37–60.

[CG16] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[Che+17] Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. "Robust optimization for non-convex objectives". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4705–4714.

[Che+18] Chang Chen, Zhiwei Xiong, Xinmei Tian, and Feng Wu. "Deep boosting for image denoising". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–18.

[CL06] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CL11] T Tony Cai and Mark G Low. "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional". In: *The Annals of Statistics* 39.2 (2011), pp. 1012–1041.

[CMS14] Corinna Cortes, Mehryar Mohri, and Umar Syed. "Deep Boosting". In: ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1179–1187. URL: http://proceedings.mlr.press/v32/cortesb14.html.

[Cor+17] Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. "Adanet: Adaptive structural learning of artificial neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 874–883.

[CPB19] Niladri Chatterji, Aldo Pacchiano, and Peter Bartlett. "Online learning with kernel losses". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 971–980.

[CS81] George Casella and William E Strawderman. "Estimating a bounded normal mean". In: *The Annals of Statistics* (1981), pp. 870–878.

[DGN14] Olivier Devolder, Franccois Glineur, and Yurii Nesterov. "First-order methods of smooth convex optimization with inexact oracle". In: *Mathematical Programming* 146.1-2 (2014), pp. 37–75.

[DHK07] Varsha Dani, Thomas P Hayes, and Sham M Kakade. "The price of bandit information for online optimization". In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. 2007, pp. 345–352.

[DLM90] David L Donoho, Richard C Liu, and Brenda MacGibbon. "Minimax risk over hyperrectangles, and implications". In: *The Annals of Statistics* (1990), pp. 1416–1437.

[DN18] John Duchi and Hongseok Namkoong. "Learning models with uniform performance via distributionally robust optimization". In: *arXiv preprint arXiv:1810.08750* (2018).

[Don94] David L Donoho. "Statistical estimation and optimal recovery". In: *The Annals of Statistics* (1994), pp. 238–270.

[DSZ21] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. "The complexity of constrained min-max optimization". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 1466–1478.

[Duc+15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. "Optimal rates for zero-order convex optimization: The power of two function evaluations". In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2788–2806.

[Fer14] Thomas S Ferguson. *Mathematical statistics: A decision theoretic approach*. Vol. 1. Academic press, 2014.

[FHT+00] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[Fil+10] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. "Parametric bandits: The generalized linear case". In: *Advances in Neural Information Processing Systems*. 2010, pp. 586–594.

[FKM04] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *arXiv preprint cs/0408007* (2004).

[FMS15] Uriel Feige, Yishay Mansour, and Robert Schapire. "Learning and inference in the presence of corrupted inputs". In: *Conference on Learning Theory*. 2015, pp. 637–657.

[Fre95] Yoav Freund. "Boosting a weak learning algorithm by majority". In: *Information and computation* 121.2 (1995), pp. 256–285.

[Fri01] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[FS+96] Yoav Freund, Robert E Schapire, et al. "Experiments with a new boosting algorithm". In: *icml*. Vol. 96. Citeseer. 1996, pp. 148–156.

[FS91] Peter W Frey and David J Slate. "Letter recognition using Holland-style adaptive classifiers". In: *Machine learning* 6.2 (1991), pp. 161–182.

[FS95] Yoav Freund and Robert E Schapire. "A desicion-theoretic generalization of on-line learning and an application to boosting". In: *European conference on computational learning theory*. Springer. 1995, pp. 23–37.

[FS96] Yoav Freund and Robert E Schapire. "Game theory, on-line prediction and boosting". In: *COLT*. Vol. 96. Citeseer. 1996, pp. 325–332.

[GB11] Alexander Grubb and J Andrew Bagnell. "Generalized boosting algorithms for convex optimization". In: *arXiv preprint arXiv:1105.2054* (2011).

[GH13] Dan Garber and Elad Hazan. "Playing non-linear games with linear oracles". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. 2013, pp. 420–428.

[Gho64] MN Ghosh. "Uniform approximation of minimax point estimates". In: *The Annals of Mathematical Statistics* (1964), pp. 1031–1047.

[GJL16] Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. "Frank-Wolfe algorithms for saddle point problems". In: *arXiv preprint arXiv:1610.07797* (2016).

[GLZ18] Xiand Gao, Xiaobo Li, and Shuzhong Zhang. "Online learning with nonconvex losses and non-stationary regret". In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 235–243.

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[Gup+20] Kartik Gupta, Arun Sai Suggala, Adarsh Prasad, Praneeth Netrapalli, and Pradeep Ravikumar. "Learning Minimax Estimators via Online Learning". In: *arXiv preprint arXiv:2006.11430* (2020).

[Gup+21] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. "Online multivalid learning: Means, moments, and prediction intervals". In: *arXiv preprint arXiv:2101.01739* (2021).

[Hal71] Anders Hald. “The size of bayes and minimax tests as function of the sample size and the loss ratio”. In: *Scandinavian Actuarial Journal* 1971.1-2 (1971), pp. 53–73.

[Har83] JA Hartigan. “Asymptotic normality of posterior distributions”. In: *Bayes theory*. Springer, 1983, pp. 107–118.

[Has+18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. “Fairness without demographics in repeated loss minimization”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1929–1938.

[Haz16] Elad Hazan. “Introduction to online convex optimization”. In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[HH15] Niao He and Zaid Harchaoui. “Semi-proximal mirror-prox for nonsmooth composite minimization”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 3411–3419.

[HK12] Elad Hazan and Satyen Kale. “Projection-free online learning”. In: *arXiv preprint arXiv:1206.4657* (2012).

[HKZ+12] Daniel Hsu, Sham Kakade, Tong Zhang, et al. “A tail inequality for quadratic forms of subgaussian random vectors”. In: *Electronic Communications in Probability* 17 (2012).

[HL14] Elad Hazan and Kfir Levy. “Bandit convex optimization: Towards tight bounds”. In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 784–792.

[HL16] Elad Hazan and Yuanzhi Li. “An optimal algorithm for bandit convex optimization”. In: *arXiv preprint arXiv:1603.04350* (2016).

[HM20] Elad Hazan and Edgar Minasyan. “Faster Projection-free Online Learning”. In: *CoRR* abs/2001.11568 (2020). arXiv: 2001.11568. URL: https://arxiv.org/abs/2001.11568.

[HP13] Reiner Horst and Panos M Pardalos. *Handbook of global optimization*. Vol. 2. Springer Science & Business Media, 2013.

[HS02] Josef Hofbauer and William H Sandholm. “On the global convergence of stochastic fictitious play”. In: *Econometrica* 70.6 (2002), pp. 2265–2294.

[HSZ17] Elad Hazan, Karan Singh, and Cyril Zhang. “Efficient regret minimization in non-convex games”. In: *arXiv preprint arXiv:1708.00075* (2017).

[Hu+16] Xiaowei Hu, LA Prashanth, András György, and Csaba Szepesvari. “(Bandit) convex optimization with biased noisy gradient oracles”. In: *Artificial Intelligence and Statistics*. PMLR. 2016, pp. 819–828.

[Hua+17a] Furong Huang, Jordan Ash, John Langford, and Robert Schapire. “Learning deep resnet blocks sequentially using boosting theory”. In: *arXiv preprint arXiv:1706.04964* (2017).

[Hua+17b] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[IH81] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. New York: springer, 1981.

[Imh61] Jean-Pierre Imhof. “Computing the distribution of quadratic forms in normal variables”. In: *Biometrika* 48.3/4 (1961), pp. 419–426.

[Ito20] Shinji Ito. “An Optimal Algorithm for Bandit Convex Optimization with Strongly-Convex and Smooth Loss”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2229–2239.

[Jia+15] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. “Minimax estimation of functionals of discrete distributions”. In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2835–2885.

[Jin+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. “A short note on concentration inequalities for random vectors with subgaussian norm”. In: *arXiv preprint arXiv:1902.03736* (2019).

[JNJ20] Chi Jin, Praneeth Netrapalli, and Michael Jordan. “What is local optimality in nonconvex-nonconcave minimax optimization?” In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4880–4889.

[Joh02] Iain M Johnstone. “Function estimation and gaussian sequence models”. In: *Unpublished manuscript* 2.5.3 (2002).

[Joh11] Iain M Johnstone. “Gaussian estimation: Sequence and wavelet models”. In: *Unpublished manuscript* (2011).

[JS92] William James and Charles Stein. “Estimation with quadratic loss”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 443–460.

[Kan+19] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R. Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. “Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly”. In: *arXiv preprint arXiv:1903.06694* (2019).

[Kem87] Peter J Kempthorne. “Numerical specification of discrete least favorable prior distributions”. In: *SIAM Journal on Scientific and Statistical Computing* 8.2 (1987), pp. 171–184.

[Kie+57] Jack Kiefer et al. “Invariance, minimax sequential estimation, and continuous time processes”. In: *The Annals of Mathematical Statistics* 28.3 (1957), pp. 573–601.

[Kle05] Robert D Kleinberg. “Nearly tight bounds for the continuum-armed bandit problem”. In: *Advances in Neural Information Processing Systems*. 2005, pp. 697–704.

[Kri+15] Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. “The hedge algorithm on a continuum”. In: *International Conference on Machine Learning*. 2015, pp. 824–832.

[KST09] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. "On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization". In: *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf* 2.1 (2009).

[KV05] Adam Kalai and Santosh Vempala. "Efficient algorithms for online decision problems". In: *Journal of Computer and System Sciences* 71.3 (2005), pp. 291–307.

[KV16] Adam Kalai and Santosh Vempala. "Efficient algorithms for on-line optimization". In: *Journal of Computer and System Sciences* 71 (2016).

[Kve+20] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. "Randomized exploration in generalized linear bandits". In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 2066–2076.

[KW05] Alfred Kume and Andrew TA Wood. "Saddlepoint approximations for the Bingham and Fisher–Bingham normalising constants". In: *Biometrika* 92.2 (2005), pp. 465–476.

[Lar+07] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. "An empirical evaluation of deep architectures on problems with many factors of variation". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 473–480.

[Lat20] Tor Lattimore. "Improved regret for zeroth-order adversarial bandit convex optimisation". In: *arXiv preprint arXiv:2006.00475* (2020).

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[Le 12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[Lee+20] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. "Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs". In: *arXiv preprint arXiv:2006.08040* (2020).

[LLV20] Aditi Laddha, Yin Tat Lee, and Santosh Vempala. "Strong self-concordance and sampling". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 1212–1222.

[LOV19] Sindy Löwe, Peter O'Connor, and Bastiaan Veeling. "Putting An End to End-to-End: Gradient-Isolated Learning of Representations". In: *Advances in Neural Information Processing Systems*. 2019, pp. 3033–3045.

[LR85] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[Lue+20] Alex Luedtke, Marco Carone, Noah Simon, and Oleg Sofrygin. "Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures". In: *Science Advances* 6.9 (2020). DOI: 10.1126/sciadv.aaw2140. eprint: https://advances.sciencemag.org/content/6/9/eaaw2140.full.pdf. URL: https://advances.sciencemag.org/content/6/9/eaaw2140.

[LV03] László Lovász and Santosh Vempala. *Where to start a geometric random walk.* 2003.

[Mas+00] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. "Boosting algorithms as gradient descent". In: *Advances in neural information processing systems*. 2000, pp. 512–518.

[Mau16] Andreas Maurer. "A vector-contraction inequality for rademacher complexities". In: *International Conference on Algorithmic Learning Theory*. Springer. 2016, pp. 3–17.

[McM11] Brendan McMahan. "Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 525–533.

[McM17] H Brendan McMahan. "A survey of algorithms and analysis for adaptive online learning". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 3117–3166.

[MJ09] Kanti V Mardia and Peter E Jupp. *Directional statistics.* Vol. 494. John Wiley & Sons, 2009.

[MM10] Odalric-Ambrym Maillard and Rémi Munos. "Online learning in adversarial lipschitz environments". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2010, pp. 305–320.

[MP02] Éric Marchand and Franccois Perron. "On the minimax estimator of a bounded normal mean". In: *Statistics & probability letters* 58.4 (2002), pp. 327–333.

[MS13] Indraneel Mukherjee and Robert E Schapire. "A theory of multiclass boosting". In: *Journal of Machine Learning Research* 14.Feb (2013), pp. 437–497.

[MTM14] Chris J Maddison, Daniel Tarlow, and Tom Minka. "A* sampling". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3086–3094.

[Nel66] Wayne Nelson. "Minimax solution of statistical decision problems by iteration". In: *The Annals of Mathematical Statistics* (1966), pp. 1643–1657.

[Nem04] Arkadi Nemirovski. "Interior point polynomial time methods in convex programming". In: *Lecture notes* (2004).

[Nes18] Yurii Nesterov. *Lectures on convex optimization.* Vol. 137. Springer, 2018.

[Net+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. "Reading digits in natural images with unsupervised feature learning". In: (2011).

[NS18] Atsushi Nitanda and Taiji Suzuki. "Functional gradient boosting based on residual network perception". In: *arXiv preprint arXiv:1802.09031* (2018).

[Pon+17] Natalia Ponomareva, Thomas Colthurst, Gilbert Hendry, Salem Haykal, and Soroush Radpour. "Compact multi-class boosted trees". In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 47–56.

[Pra+20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. "Robust estimation via robust gradient estimation". In: *Journal* *of the Royal Statistical Society: Series B (Statistical Methodology)* 82.3 (2020), pp. 601–627.

[PW19] Yury Polyanskiy and Yihong Wu. "Dualizing Le Cam's method, with applications to estimating the unseens". In: *arXiv preprint arXiv:1902.05616* (2019).

[Roc70] R Tyrrell Rockafellar. *Convex analysis*. 28. Princeton university press, 1970.

[RS12] Alexander Rakhlin and Karthik Sridharan. "Online learning with predictable sequences". In: *arXiv preprint arXiv:1208.3728* (2012).

[RS13] Sasha Rakhlin and Karthik Sridharan. "Optimization, learning, and games with predictable sequences". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3066–3074.

[Rus+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[RW05] Gunnar Rätsch and Manfred K Warmuth. "Efficient margin maximizing with boosting". In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 2131–2152.

[Sch90] Robert E Schapire. "The strength of weak learnability". In: *Machine learning* 5.2 (1990), pp. 197–227.

[Sha07] Shai Shalev-Shwartz. "Thesis submitted for the degree of "Doctor of Philosophy"". In: (2007).

[Sha13] Ohad Shamir. "On the complexity of bandit and derivative-free stochastic convex optimization". In: *Conference on Learning Theory*. PMLR. 2013, pp. 3–24.

[Sha17] Ohad Shamir. "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1703–1713.

[SLR20] Arun Sai Suggala, Bingbin Liu, and Pradeep Ravikumar. "Generalized Boosting". In: *Advances in Neural Information Processing Systems 33*. 2020.

[SN20a] Arun Sai Suggala and Praneeth Netrapalli. "Follow the Perturbed Leader: Optimism and Fast Parallel Algorithms for Smooth Minimax Games". In: *Advances in Neural Information Processing Systems 33*. 2020. URL: https://arxiv.org/abs/2006.07541.

[SN20b] Arun Sai Suggala and Praneeth Netrapalli. "Online Non-Convex Learning: Following the Perturbed Leader is Optimal". In: ed. by Aryeh Kontorovich and Gergely Neu. Vol. 117. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, Aug. 2020, pp. 845–861. URL: http://proceedings.mlr.press/v117/suggala20a.html.

[SRB11] Mark Schmidt, Nicolas L Roux, and Francis R Bach. "Convergence rates of inexact proximal-gradient methods for convex optimization". In: *Advances in neural information processing systems*. 2011, pp. 1458–1466.

[Sri+09] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. "Gaussian process optimization in the bandit setting: No regret and experimental design". In: *arXiv preprint arXiv:0912.3995* (2009).

[SRN21] Arun Sai Suggala, Pradeep Ravikumar, and Praneeth Netrapalli. "Efficient Bandit Convex Optimization: Beyond Linear Losses". In: *Conference on Learning Theory*. 2021.

[ST11] Ankan Saha and Ambuj Tewari. "Improved regret guarantees for online smooth convex optimization with bandit feedback". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 636–642.

[Sug+19a] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. "Adaptive Hard Thresholding for Near-optimal Consistent Robust Regression". In: *Conference on Learning Theory*. 2019, pp. 2892–2897.

[Sug+19b] Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, and Pradeep Ravikumar. "Revisiting adversarial risk". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2331–2339.

[Sze+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[Tem14] Vladimir Nikolaevich Temlyakov. "Greedy expansions in convex optimization". In: *Proceedings of the Steklov Institute of Mathematics* 284.1 (2014), pp. 244–262.

[Tro12] Joel A Tropp. "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4 (2012), pp. 389–434.

[TS20] Sho Takemori and Masahiro Sato. "Approximation Methods for Kernelized Bandits". In: *arXiv preprint arXiv:2010.12167* (2020).

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387790519, 9780387790510.

[VA87] Peter JM Van Laarhoven and Emile HL Aarts. "Simulated annealing". In: *Simulated annealing: Theory and applications*. Springer, 1987, pp. 7–15.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[VMK07] John Von Neumann, Oskar Morgenstern, and Harold William Kuhn. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[VV11] G. Valiant and P. Valiant. "The Power of Linear Estimators". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. Oct. 2011, pp. 403–412. DOI: 10.1109/FOCS.2011.81.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[Wal49] Abraham Wald. "Statistical decision functions". In: *The Annals of Mathematical Statistics* (1949), pp. 165–205.

[Wel15] Jon A Wellner. "Maximum Likelihood in modern times: the ugly, the bad, and the good". 2015. URL: https://www.stat.washington.edu/jaw/RESEARCH/TALKS/LeCam-v2.pdf.

[Wij90] Robert A Wijsman. "Invariant Measures on Groups and Their Use in Statistics". In: *Lecture Notes-Monograph Series* 14 (1990), pp. i–218.

[WY16] Yihong Wu and Pengkun Yang. "Minimax rates of entropy estimation on large alphabets via best polynomial approximation". In: *IEEE Transactions on Information Theory* 62.6 (2016), pp. 3702–3720.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: cs.LG/1708.07747 [cs.LG].

[Yan74] EB Yanovskaya. "Infinite zero-sum two-person games". In: *Journal of Soviet Mathematics* 2.5 (1974), pp. 520–541.

[YB99] Yuhong Yang and Andrew Barron. "Information-theoretic determination of minimax rates of convergence". In: *Annals of Statistics* (1999), pp. 1564–1599.

[Zah+17] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. "Deep sets". In: *Advances in neural information processing systems*. 2017, pp. 3391–3401.

[Zar73] Eduardo H Zarantonello. "Dense single-valuedness of monotone operators". In: *Israel Journal of Mathematics* 15.2 (1973), pp. 158–166.

Part V

Appendix

Appendix A

Supplementary Material for Chapter 2

A.1 Proof of Proposition 1

For any deterministic algorithm, we show that there exists a sequence of loss functions over which the algorithm has $\Omega(T)$ regret. We work in the 1-dimensional setting and assume that the domain $\mathcal{X}$ is equal to $[-D, D]$. Suppose the adversary chooses the loss functions from the following class of 1-Lipschitz functions $\mathcal{F} = \{g_a(\mathbf{x}) : a \in [-D, D]\}$, where g_a is given by

$$g_a(\mathbf{x}) = \max\left\{0, \frac{D}{2} - |\mathbf{x} - a|\right\}.$$

We now describe our construction of the sequence of losses that cause the deterministic algorithm to fail. Let $f_{<t} = \{f_1, \dots f_{t-1}\}$ be the sequence of loss functions chosen until iteration $t-1$. Let $\mathbf{x}_t$ be the prediction of the deterministic learner at iteration t. Then we choose the loss at iteration t as $f_t(\mathbf{x}) = g_{\mathbf{x}_t}(\mathbf{x})$. It is easy to see that, after T iterations, the loss suffered by the learner is equal to $\frac{DT}{2}$. Whereas, the loss of the best action in hindsight can be upper bounded as

$$\inf_{\mathbf{x}\in[-D,D]} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{DT}{4}.$$

This shows that the regret of any deterministic algorithm is $\Omega(1)$.

A.2 Non-oblivious to Oblivious Adversary Model

In the oblivious adversary model, the actions $\{f_t\}_{t=1}^T$ of the adversary are assumed to be independent of the predictions $\{\mathbf{x}_t\}_{t=1}^T$ of the FTPL/OFTPL algorithm. In this model, we assume that the sequence of losses $\{f_t\}_{t=1}^T$ is fixed ahead of time. Whereas in the non-oblivious adversary model, the actions of the adversary are allowed to depend on the past predictions of the algorithm, *i.e.*, each f_t is given by $f_t \coloneqq F_t[\mathbf{x}_{<t}]$ for some function $F_t : \mathcal{X}^{t-1} \rightarrow \mathcal{F}$, where $\mathcal{F}$ is the set of all possible actions of the adversary and $\mathbf{x}_{<t}$ is a

shorthand for $\{\mathbf{x}_1 \dots \mathbf{x}_{t-1}\}$ and F_1 is a constant function. Note that the functions $F_1 \dots F_T$ uniquely determine a non-oblivious adversary.

Let P_t be the conditional distribution of the prediction $\mathbf{x}_t$ of the FTPL/OFTPL algorithm, conditioned on the past predictions $\mathbf{x}_{<t}$. Note that when the adversary is oblivious, P_t is independent of $\mathbf{x}_{<t}$. Moreover, in both oblivious and non-oblivious models, P_t is fully determined by the past actions $f_{<t}$ of the adversary. Let $f_t(P_t)$ denote the expected loss $\mathbb{E}_{\mathbf{x}\sim P_t}\left[f_t(\mathbf{x})|\mathbf{x}_{<t}\right]$.

The following Theorem shows that any algorithm which is guaranteed to work against an oblivious adversary also works against a non-oblivious adversary. This is an adaptation of Lemma 4.1 of [CL06] to the setting studied in this paper.

Theorem 22. *Let B be a positive constant. Suppose the FTPL, OFTPL algorithms satisfy the following regret bound against an oblivious adversary*

$$\mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - \inf_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})\right] \leq B, \quad \forall f_1 \dots f_T \in \mathcal{F}. \tag{A.1}$$

Then these algorithms satisfy the following regret bound against a non-oblivious adversary

$$\sum_{t=1}^T f_t(P_t) - \inf_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x}) \leq B.$$

Proof. Consider the non-oblivious adversary model. For any $\mathbf{x} \in \mathcal{X}$ we have

$$\begin{aligned}
&\sum_{t=1}^T f_t(P_t) - \sum_{t=1}^T f_t(\mathbf{x}) \\
&\quad = \sum_{t=1}^T F_t[\mathbf{x}_{<t}](P_t) - \sum_{t=1}^T F_t[\mathbf{x}_{<t}](\mathbf{x}) \\
&\quad \overset{(a)}{\leq} \sup_{F_1,\dots F_T} \left(\sum_{t=1}^T F_t[\mathbf{x}_{<t}](P_t) - \sum_{t=1}^T F_t[\mathbf{x}_{<t}](\mathbf{x})\right) \\
&\quad \overset{(b)}{=} \sup_{g_1\in\mathcal{F}} \left(g_1(P_1) - g_1(\mathbf{x}) + \sup_{g_2\in\mathcal{F}} \left(g_2(P_2) - g_2(\mathbf{x}) + \sup_{g_3\in\mathcal{F}} \left(\dots + \sup_{g_T\in\mathcal{F}} g_T(P_T) - g_T(\mathbf{x})\right)\right)\right),
\end{aligned}$$

where the supremum in (a) is over all possible non-oblivious adversaries. To see why (b) holds, consider $T = 2$. Then

$$\begin{aligned}
&\sup_{F_1,F_2} \left(F_1[\mathbf{x}_{<1}](P_1) - F_1[\mathbf{x}_{<1}](\mathbf{x}) + F_2[\mathbf{x}_{<2}](P_2) - F_2[\mathbf{x}_{<2}](\mathbf{x})\right) \\
&\quad = \sup_{g_1\in\mathcal{F},F_2} \left(g_1(P_1) - g_1(\mathbf{x}) + F_2[\mathbf{x}_{<2}](P_2) - F_2[\mathbf{x}_{<2}](\mathbf{x})\right) \\
&\quad = \sup_{g_1} \left(g_1(P_1) - g_1(\mathbf{x}) + \sup_{g_2\in\mathcal{F}} g_2(P_2) - g_2(\mathbf{x})\right).
\end{aligned}$$

This shows that a good strategy for the adversary is to set $F_2[\mathbf{x}_{<2}]$ to be a maximizer of $g_2(P_2) - g_2(\mathbf{x})$. Using a similar argument we can show that (b) holds for $T > 2$.

Next, we show that

$$\sup_{g_1\in\mathcal{F}}\left(g_1(P_1)-g_1(\mathbf{x})+\sup_{g_2\in\mathcal{F}}\left(g_2(P_2)-g_2(\mathbf{x})+\sup_{g_3\in\mathcal{F}}\left(\cdots+\sup_{g_T\in\mathcal{F}}g_T(P_T)-g_T(\mathbf{x})\right)\right)\right)$$
$$=\sup_{g_1\dots g_T\in\mathcal{F}}\left(\sum_{t=1}^T g_t(P_t)-g_t(\mathbf{x})\right).$$

Moreover, we show that the maximizers of the RHS objective are independent of the predictions $\{\mathbf{x}_t\}_{t=1}^T$ of the algorithm. This would then imply that the RHS is exactly equal to the regret of the algorithm under the oblivious adversary model, which is upper bounded by B. To see why the above statements are true, again consider the case of $T=2$. First note that $g_1(P_1)-g_1(\mathbf{x})$ is independent of g_2. So $g_1(P_1)-g_1(\mathbf{x})$ can be pushed inside the inner supermum. So we have

$$\sup_{g_1\in\mathcal{F}}\left(g_1(P_1)-g_1(\mathbf{x})+\sup_{g_2\in\mathcal{F}}(g_2(P_2)-g_2(\mathbf{x}))\right)$$
$$=\sup_{g_1,g_2\in\mathcal{F}}(g_1(P_1)-g_1(\mathbf{x})+g_2(P_2)-g_2(\mathbf{x}))$$

To see why the maximizers of the RHS are independent of $\mathbf{x}_1,\mathbf{x}_2$, note that P_1 is independent of $\mathbf{x}_1,\mathbf{x}_2$. Moreover, P_2 is fully determinimed by g_1. So the objective is independent of $\mathbf{x}_1,\mathbf{x}_2$. This shows that the maximizers are independent of $\mathbf{x}_1,\mathbf{x}_2$. Using a similar argument we can show that the above claim holds for $T>2$. Finally, from the regret bound against an oblivious adversary in Equation (A.1), we have

$$\sup_{g_1\dots g_T\in\mathcal{F}}\left(\sum_{t=1}^T g_t(P_t)-g_t(\mathbf{x})\right)=\sup_{g_1\dots g_T\in\mathcal{F}}\mathbb{E}\left[\sum_{t=1}^T g_t(\mathbf{x}_t)-\sum_{t=1}^T g_t(\mathbf{x})\right]\le B.$$

This shows that for any $\mathbf{x}\in\mathcal{X}$, $\sum_{t=1}^T f_t(P_t)-\sum_{t=1}^T f_t(\mathbf{x})\le B$. □

A.3 Proof of Lemma 2

Let $\gamma(\sigma)=\alpha+\beta\|\sigma\|_1$. For any $\mathbf{x}^*\in\mathcal{X}$ we have

$$\sum_{t=1}^T[f_t(\mathbf{x}_t)-f_t(\mathbf{x}^*)]$$
$$=\sum_{t=1}^T[f_t(\mathbf{x}_t)-f_t(\mathbf{x}_{t+1})]+\sum_{t=1}^T[f_t(\mathbf{x}_{t+1})-f_t(\mathbf{x}^*)]$$
$$\le\sum_{t=1}^T L\|\mathbf{x}_t-\mathbf{x}_{t+1}\|_1+\sum_{t=1}^T[f_t(\mathbf{x}_{t+1})-f_t(\mathbf{x}^*)].$$

We now use induction to show that $\sum_{t=1}^T[f_t(\mathbf{x}_{t+1})-f_t(\mathbf{x}^*)]\le\gamma(\sigma)T+\langle\sigma,\mathbf{x}_2-\mathbf{x}^*\rangle$.

Base Case ($T = 1$). Since $\mathbf{x}_2$ is an approximate minimizer of $f_1(\mathbf{x}) - \langle \sigma, \mathbf{x} \rangle$, we have

$$f_1(\mathbf{x}_2) - \langle \sigma, \mathbf{x}_2 \rangle \leq \min_{\mathbf{x} \in \mathcal{X}} f_1(\mathbf{x}) - \langle \sigma, \mathbf{x} \rangle + \gamma(\sigma) \leq f_1(\mathbf{x}^*) - \langle \sigma, \mathbf{x}^* \rangle + \gamma(\sigma),$$

where the last inequality holds for any $\mathbf{x}^* \in \mathcal{X}$. This shows that $f_1(\mathbf{x}_2) - f_1(\mathbf{x}^*) \leq \gamma(\sigma) + \langle \sigma, \mathbf{x}_2 - \mathbf{x}^* \rangle$.

Induction Step. Suppose the claim holds for all $T \leq T_0 - 1$. We now show that it also holds for T_0.

$$\begin{aligned}
&\sum_{t=1}^{T_0} f_t(\mathbf{x}_{t+1}) \\
&\quad \overset{(a)}{\leq} \left[\sum_{t=1}^{T_0 - 1} f_t(\mathbf{x}_{T_0+1}) + \langle \sigma, \mathbf{x}_2 - \mathbf{x}_{T_0+1} \rangle + \gamma(\sigma)(T_0 - 1) \right] + f_{T_0}(\mathbf{x}_{T_0+1}) \\
&\quad = \left[\sum_{t=1}^{T_0} f_t(\mathbf{x}_{T_0+1}) - \langle \sigma, \mathbf{x}_{T_0+1} \rangle \right] + \langle \sigma, \mathbf{x}_2 \rangle + \gamma(\sigma)(T_0 - 1) \\
&\quad \overset{(b)}{\leq} \sum_{t=1}^{T_0} f_t(\mathbf{x}^*) + \langle \sigma, \mathbf{x}_2 - \mathbf{x}^* \rangle + \gamma(\sigma) T_0, \quad \forall \mathbf{x}^* \in \mathcal{X},
\end{aligned}$$

where (a) follows since the claim holds for any $T \leq T_0 - 1$, and (b) follows from the approximate optimality of $\mathbf{x}_{T_0+1}$.

Using this result, we get the following upper bound on the expected regret of FTPL

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \inf_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \right] &\leq L \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_1 \right] + \mathbb{E}\left[\gamma(\sigma) T + \langle \sigma, \mathbf{x}_2 - \mathbf{x}^* \rangle \right] \\
&\leq L \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_1 \right] + (\beta T + D) \left(\sum_{i=1}^{d} \mathbb{E}\left[\sigma_i \right] \right) + \alpha T
\end{aligned}$$

The proof of the Lemma now follows from the following property of exponential distribution

$$\mathbb{E}\left[\sigma_i \right] = \frac{1}{\eta_i}.$$

Appendix B

Supplementary Material for Chapter 3

B.1 Dual view of Perturbations as Regularization

B.1.1 Proof of Theorem 2

We first define a convex function $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$\Psi(f) = \mathbb{E}_\sigma \left[\sup_{\mathbf{x} \in \mathcal{X}} \langle f + \sigma, \mathbf{x} \rangle \right] = \mathbb{E}_\sigma \left[\sup_{\mathbf{x} \in \mathcal{X}} \langle f + \sigma, \mathbf{x} \rangle \right],$$

where perturbation σ follows probability distribution P_{PRTB} which is absolutely continuous w.r.t the Lebesgue measure. For our choice of P_{PRTB}, we now show that Ψ is differentiable. Consider the function $\psi(g) = \sup_{\mathbf{x} \in \mathcal{X}} \langle g, \mathbf{x} \rangle$. Since $\psi(g)$ is a proper convex function, we know that it is differentiable almost everywhere, except on a set of Lebesgue measure 0 [see Theorem 25.5 of Roc70]. Moreover, it is easy to verify that $\text{argmax}_{\mathbf{x} \in \mathcal{X}} \langle g, \mathbf{x} \rangle \in \partial \psi(g)$. These two observations, together with the fact that P_{PRTB} is absolutely continuous, show that the sup expression inside the expectation of Ψ has a unique maximizer with probability one.

Since the sup expression inside the expectation has a unique maximizer with probability 1, we can swap the expectation and gradient to obtain [see Proposition 2.2 of Ber73]

$$\nabla \Psi(f) = \mathbb{E}_\sigma \left[\underset{\mathbf{x} \in \mathcal{X}}{\text{argmax}} \langle f + \sigma, \mathbf{x} \rangle \right]. \tag{B.1}$$

Note that $\nabla\Psi$ is related to the prediction of deterministic version of FTPL. Specifically, $\nabla\Psi(-\nabla_{1:t-1})$ is the prediction of deterministic FTPL in the t^{th} iteration. We now show that $\nabla\Psi(f) = \text{argmin}_{\mathbf{x} \in \mathcal{X}} \langle -f, \mathbf{x} \rangle + R(\mathbf{x})$, for some convex function R.

Since all differentiable functions are closed, $\Psi(f)$ is a proper, closed and differentiable convex function over $\mathbb{R}^d$. Let $R(\mathbf{x})$ denote the Fenchel conjugate of $\Psi(f)$

$$R(\mathbf{x}) = \sup_{f \in \text{dom}(\Phi)} \langle \mathbf{x}, f \rangle - \Psi(f),$$

where $\text{dom}(\Psi)$ denotes the domain of Ψ. Following Theorem 32 (see Appendix B.8), $\Psi(f)$ is the Fenchel conjugate of $R(\mathbf{x})$

$$\Psi(f) = \sup_{\mathbf{x}\in \text{dom}(R)} \langle f, \mathbf{x}\rangle - R(\mathbf{x}).$$

Furthermore, from Theorem 33 we have

$$\nabla\Psi(f) = \operatorname*{argmax}_{\mathbf{x}\in \text{dom}(R)} \langle f, \mathbf{x}\rangle - R(\mathbf{x}).$$

We now show that the domain of R is a subset of $\mathcal{X}$. This, together with the previous two equations, would then immediately imply

$$\Psi(f) = \sup_{\mathbf{x}\in\mathcal{X}} \langle f, \mathbf{x}\rangle - R(\mathbf{x}), \tag{B.2}$$

$$\nabla\Psi(f) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \langle f, \mathbf{x}\rangle - R(\mathbf{x}). \tag{B.3}$$

From Theorem 35, we know that the domain of R satisfies

$$\text{ri}(\text{dom}(R)) \subseteq \text{range}\nabla\Psi \subseteq \text{dom}(R),$$

where $\text{ri}(A)$ denotes the relative interior of a set A. Moreover, from the definition of $\nabla\Psi(f)$ in Equation (B.1), we have $\text{range}\nabla\Psi \subseteq \mathcal{X}$. Combining these two properties, we can show that one of the following statements is true

$$\text{ri}(\text{dom}(R)) \subseteq \text{range}\nabla\Psi \subseteq \mathcal{X} \subseteq \text{dom}(R),$$
$$\text{ri}(\text{dom}(R)) \subseteq \text{range}\nabla\Psi \subseteq \text{dom}(R) \subseteq \mathcal{X}.$$

Suppose the first statement is true. Since $\mathcal{X}$ is a compact set, it is easy to see that $\mathcal{X} = \text{dom}(R)$. If the second statement is true, then $\text{dom}(R) \subseteq \mathcal{X}$. Together, these two statements imply $\text{dom}(R) \subseteq \mathcal{X}$.

Connecting back to FTPL. We now connect the above results to FTPL. From Equation (B.1), we know that the prediction at iteration t of deterministic FTPL is equal to $\nabla\Psi(-\nabla_{1:t-1})$. From Equation (B.3), $\nabla\Psi(-\nabla_{1:t-1})$ is defined as

$$\mathbf{x}_t = \nabla\Psi(-\nabla_{1:t-1}) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \langle -\nabla_{1:t-1}, \mathbf{x}\rangle - R(\mathbf{x}).$$

This shows that

$$\mathbf{x}_t = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \langle \nabla_{1:t-1}, \mathbf{x}\rangle + R(\mathbf{x}).$$

So the prediction of FTPL can also be obtained using FTRL for some convex regularizer $R(\mathbf{x})$. Finally, to show that $-\nabla_{1:t-1} \in \partial R(\mathbf{x}_t), \mathbf{x}_t = \partial R^{-1}\left(-\nabla_{1:t-1}\right)$, we rely on Theorem 34. Since $\mathbf{x}_t = \nabla\Psi(-\nabla_{1:t-1})$, from Theorem 34, we have

$$-\nabla_{1:t-1} \in \partial R(\mathbf{x}_t), \quad \mathbf{x}_t = \nabla\Psi(-\nabla_{1:t-1}) = \partial R^{-1}\left(-\nabla_{1:t-1}\right),$$

where ∂R^{-1} is the inverse of ∂R in the sense of multivalued mappings. Note that, even though ∂R can be a multivalued mapping, its inverse $\partial R^{-1} = \nabla\Psi$ is a singlevalued mapping (this follows form differentiability of Ψ). This finishes the proof of the Theorem.

B.2 Online Convex Learning

B.2.1 Proof of Theorem 5

Before presenting the proof of the Theorem, we introduce some notation.

Notation

We define functions $\Phi : \mathbb{R}^d \to \mathbb{R}$, $R : \mathbb{R}^d \to \mathbb{R}$ as follows

$$\Phi(f) = \mathbb{E}_\sigma \left[\inf_{\mathbf{x}\in\mathcal{X}} \langle f - \sigma, \mathbf{x}\rangle \right], \quad R(\mathbf{x}) = \sup_{f\in\mathbb{R}^d} \langle f, \mathbf{x}\rangle + \Phi(-f).$$

Note that Φ is related to the function Ψ defined in the proof of Proposition 2. To be precise, $\Psi(f) = -\Phi(-f)$. Moreover, $R(\mathbf{x})$ is the Fenchel conjugate of Ψ. For our choice of perturbation distribution, Ψ is differentiable (see proof of Proposition 2). This implies Φ is also differentiable with gradient $\nabla\Phi$ defined as

$$\nabla\Phi\left(f\right) = \mathbb{E}_\sigma \left[\operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \langle f - \sigma, \mathbf{x}\rangle \right].$$

Note that $\nabla\Phi$ is the prediction of deterministic version of FTPL. In Proposition 2 we showed that

$$\nabla\Phi\left(f\right) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \langle f, \mathbf{x}\rangle + R(\mathbf{x}).$$

Main Argument

Since $\mathbf{x}_t^\infty$ is the prediction of deterministic version of FTPL, following FTPL-FTRL duality proved in Proposition 2, $\mathbf{x}_t^\infty$ can equivalently be written as

$$\mathbf{x}_t^\infty = \nabla\Phi\left(\nabla_{1:t-1} + g_t\right) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \langle \nabla_{1:t-1} + g_t, \mathbf{x}\rangle + R(\mathbf{x}).$$

Similarly, $\tilde{\mathbf{x}}_t^\infty$ can be written as

$$\tilde{\mathbf{x}}_t^\infty = \nabla\Phi\left(\nabla_{1:t}\right) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \langle \nabla_{1:t}, \mathbf{x}\rangle + R(\mathbf{x}).$$

We use the notation $\nabla_{1:0} = 0$. So $\tilde{\mathbf{x}}_0^\infty, \mathbf{x}_1^\infty$ are equal to $\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} R(\mathbf{x})$. From the first order optimality conditions, we have

$$-\nabla_{1:t-1} - g_t \in \partial R\left(\mathbf{x}_t^\infty\right), \quad -\nabla_{1:t} \in \partial R\left(\tilde{\mathbf{x}}_t^\infty\right).$$

Define functions $B(\cdot, \mathbf{x}_t^\infty), B(\cdot, \tilde{\mathbf{x}}_t^\infty)$ for any $t \in [T]$ as

$$\begin{aligned} B(\mathbf{x}, \mathbf{x}_t^\infty) &= R(\mathbf{x}) - R(\mathbf{x}_t^\infty) + \langle \nabla_{1:t-1} + g_t, \mathbf{x} - \mathbf{x}_t^\infty\rangle, \\ B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) &= R(\mathbf{x}) - R(\tilde{\mathbf{x}}_t^\infty) + \langle \nabla_{1:t}, \mathbf{x} - \tilde{\mathbf{x}}_t^\infty\rangle. \end{aligned}$$

From the stability of predictions of OFTPL we know that: $\|\nabla\Phi\left(g_1\right) - \nabla\Phi\left(g_2\right)\| \leq C\eta^{-1}\|g_1 - g_2\|_*$. Following our connection between Ψ, Φ, this implies $\|\nabla\Psi(g_1) - \nabla\Psi(g_2)\| \leq C\eta^{-1}\|g_1 - g_2\|_*$. This implies the following smoothness condition on Ψ [see Lemma 15 of Sha07]

$$\Psi(g_2) \leq \Psi(g_1) + \langle\nabla\Psi(g_1), g_2 - g_1\rangle + \frac{C\eta^{-1}}{2}\|g_1 - g_2\|_*^2.$$

Since Ψ is $C\eta^{-1}$-smooth w.r.t $\|\cdot\|_*$, following duality between strong convexity and strong smoothness properties (see Theorem 36), we can infer that R is $C^{-1}\eta$- strongly convex w.r.t $\|\cdot\|$ norm and satisfies

$$B(\mathbf{x}, \mathbf{x}_t^\infty) \geq \frac{\eta}{2C}\|\mathbf{x} - \mathbf{x}_t^\infty\|^2, \quad B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) \geq \frac{\eta}{2C}\|\mathbf{x} - \tilde{\mathbf{x}}_t^\infty\|^2.$$

We now go ahead and bound the regret of the learner. For any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned}
f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \overset{(a)}{\leq} \langle\mathbf{x}_t - \mathbf{x}, \nabla_t\rangle &= \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \langle\mathbf{x}_t^\infty - \mathbf{x}, \nabla_t\rangle \\
&= \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \langle\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty, \nabla_t - g_t\rangle + \langle\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty, g_t\rangle \\
&\quad + \langle\tilde{\mathbf{x}}_t^\infty - \mathbf{x}, \nabla_t\rangle \\
&\leq \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_* + \langle\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty, g_t\rangle \\
&\quad + \langle\tilde{\mathbf{x}}_t^\infty - \mathbf{x}, \nabla_t\rangle,
\end{aligned}$$

where (a) follows from convexity of f. Next, a simple calculation shows that

$$\begin{aligned}
\langle\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty, g_t\rangle &= B(\tilde{\mathbf{x}}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\tilde{\mathbf{x}}_t^\infty, \mathbf{x}_t^\infty) - B(\mathbf{x}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) \\
\langle\tilde{\mathbf{x}}_t^\infty - \mathbf{x}, \nabla_t\rangle &= B(\mathbf{x}, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) - B(\tilde{\mathbf{x}}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty).
\end{aligned}$$

Substituting this in the previous inequality gives us

$$\begin{aligned}
f_t(\mathbf{x}_t) - f_t(\mathbf{x}) &\leq \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_* \\
&\quad + B(\tilde{\mathbf{x}}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\tilde{\mathbf{x}}_t^\infty, \mathbf{x}_t^\infty) - B(\mathbf{x}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) \\
&\quad + B(\mathbf{x}, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) - B(\tilde{\mathbf{x}}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) \\
&= \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_* \\
&\quad + B(\mathbf{x}, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) - B(\tilde{\mathbf{x}}_t^\infty, \mathbf{x}_t^\infty) - B(\mathbf{x}_t^\infty, \tilde{\mathbf{x}}_{t-1}^\infty) \\
&\overset{(a)}{\leq} \langle\mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_* \\
&\quad + B(\mathbf{x}, \tilde{\mathbf{x}}_{t-1}^\infty) - B(\mathbf{x}, \tilde{\mathbf{x}}_t^\infty) - \frac{\eta\|\tilde{\mathbf{x}}_t^\infty - \mathbf{x}_t^\infty\|^2}{2C} - \frac{\eta\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2}{2C},
\end{aligned}$$

where (a) follows from strongly convexity of R. Summing over $t = 1, \dots T$, gives us

$$\begin{aligned}\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \leq \sum_{t=1}^{T} \langle \mathbf{x}_t - \mathbf{x}_t^{\infty}, \nabla_t \rangle + \underbrace{B(\mathbf{x}, \tilde{\mathbf{x}}_0^{\infty}) - B(\mathbf{x}, \tilde{\mathbf{x}}_T^{\infty})}_{S_1} \\ + \sum_{t=1}^{T} \|\mathbf{x}_t^{\infty} - \tilde{\mathbf{x}}_t^{\infty}\| \|\nabla_t - g_t\|_* \\ - \frac{\eta}{2C} \sum_{t=1}^{T} \left(\|\tilde{\mathbf{x}}_t^{\infty} - \mathbf{x}_t^{\infty}\|^2 + \|\mathbf{x}_t^{\infty} - \tilde{\mathbf{x}}_{t-1}^{\infty}\|^2 \right).\end{aligned}$$

Bounding S_1. We now bound $B(\mathbf{x}, \tilde{\mathbf{x}}_0^{\infty}) - B(\mathbf{x}, \tilde{\mathbf{x}}_T^{\infty})$. From the definition of B, we have

$$B(\mathbf{x}, \tilde{\mathbf{x}}_0^{\infty}) - B(\mathbf{x}, \tilde{\mathbf{x}}_T^{\infty}) = R(\tilde{\mathbf{x}}_T^{\infty}) - \langle \nabla_{1:T}, \mathbf{x} - \tilde{\mathbf{x}}_T^{\infty} \rangle - R(\tilde{\mathbf{x}}_0^{\infty}) + \langle \nabla_{1:0}, \mathbf{x} - \tilde{\mathbf{x}}_T^{\infty} \rangle.$$

Note that $\nabla_{1:0} = 0$. This gives us

$$B(\mathbf{x}, \tilde{\mathbf{x}}_0^{\infty}) - B(\mathbf{x}, \tilde{\mathbf{x}}_T^{\infty}) = R(\tilde{\mathbf{x}}_T^{\infty}) - \langle \nabla_{1:T}, \mathbf{x} - \tilde{\mathbf{x}}_T^{\infty} \rangle - R(\tilde{\mathbf{x}}_0^{\infty}).$$

We now use duality to convert the RHS of the above equation, which is currently in terms of R, into a quantity which depends on Φ. From Proposition 2 we have

$$\Phi(g) = -\Psi(-g) = \inf_{\mathbf{x} \in \mathcal{X}} \langle g, \mathbf{x} \rangle + R(\mathbf{x}).$$

Since $\tilde{\mathbf{x}}_T^{\infty}$ is the minimizer of $\langle \nabla_{1:T}, \mathbf{x} \rangle + R(\mathbf{x})$, we have $\Phi(\nabla_{1:T}) = \langle \nabla_{1:T}, \tilde{\mathbf{x}}_T^{\infty} \rangle + R(\tilde{\mathbf{x}}_T^{\infty})$. Similarly, $\Phi(0) = R(\tilde{\mathbf{x}}_0^{\infty})$. Substituting these in the previous equation gives us

$$\begin{aligned}B(\mathbf{x}, \tilde{\mathbf{x}}_0^{\infty}) - B(\mathbf{x}, \tilde{\mathbf{x}}_T^{\infty}) &= \Phi(\nabla_{1:T}) - \langle \nabla_{1:T}, \mathbf{x} \rangle - \Phi(0) \\ &= \mathbb{E}_{\sigma}\left[\inf_{\mathbf{x}' \in \mathcal{X}} \langle \nabla_{1:T} - \sigma, \mathbf{x}' \rangle \right] - \langle \nabla_{1:T}, \mathbf{x} \rangle - \mathbb{E}_{\sigma}\left[\inf_{\mathbf{x}' \in \mathcal{X}} \langle -\sigma, \mathbf{x}' \rangle \right] \\ &\leq \mathbb{E}_{\sigma}\left[\langle \nabla_{1:T} - \sigma, \mathbf{x} \rangle \right] - \langle \nabla_{1:T}, \mathbf{x} \rangle - \mathbb{E}_{\sigma}\left[\inf_{\mathbf{x}' \in \mathcal{X}} \langle -\sigma, \mathbf{x}' \rangle \right] \\ &= \mathbb{E}_{\sigma}\left[\inf_{\mathbf{x}' \in \mathcal{X}} \langle \sigma, \mathbf{x}' \rangle \right] - \mathbb{E}_{\sigma}\left[\langle \sigma, \mathbf{x} \rangle \right] \\ &\leq D\mathbb{E}_{\sigma}\left[\|\sigma\|_* \right] = \eta D\end{aligned}$$

Bounding Regret. Substituting this in our regret bound and taking expectation on both sides gives us

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right] &\leq \sum_{t=1}^T \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle\right] + \eta D + \sum_{t=1}^T \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_*\right]\\ &\quad - \frac{\eta}{2C}\sum_{t=1}^T \left(\mathbb{E}\left[\|\tilde{\mathbf{x}}_t^\infty - \mathbf{x}_t^\infty\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right]\right)\\ &\leq \sum_{t=1}^T \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle\right] + \eta D + \sum_{t=1}^T \frac{C}{2\eta}\mathbb{E}\left[\|\nabla_t - g_t\|_*^2\right]\\ &\quad - \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right]\end{aligned}$$

To finish the proof, we make use of the Holder's smoothness assumption on f_t to bound the first term in the RHS above. From Holder's smoothness assumption, we have

$$\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t - \nabla f_t(\mathbf{x}_t^\infty)\rangle \leq L\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha}.$$

Using this, we get

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right] &\leq \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla f_t(\mathbf{x}_t^\infty)\rangle + L\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha} | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right]\\ &\overset{(a)}{=} L\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha} | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right]\\ &\overset{(b)}{\leq} \Psi_1^{1+\alpha} L\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|_2^{1+\alpha} | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right]\\ &\overset{(c)}{\leq} \Psi_1^{1+\alpha} L\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|_2^{2} | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right]^{(1+\alpha)/2}\\ &\overset{(d)}{\leq} L\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha},\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla f_t(\mathbf{x}_t^\infty)\rangle | g_t, \mathbf{x}_{1:t-1}, f_{1:t}\right] = 0$, (b) follows from the definition of norm compatibility constant Ψ_1, (c) follows from Holders inequality and (d) uses the fact that conditioned on $\{g_t, \mathbf{x}_{1:t-1}, f_{1:t}\}$, $\mathbf{x}_t - \mathbf{x}_t^\infty$ is the average of m i.i.d bounded mean 0 random variables, the variance of which scales as $O(D^2/m)$. Substituting this in the above regret bound gives us the required result.

B.2.2 Proof of Corollary 1

We first bound $\mathbb{E}_\sigma\left[\|\sigma\|_2\right]$. Relying on spherical symmetry of the perturbation distribution and the fact that the density of P_{PRTB} on the spherical shell of radius r is proportional to r^{d-1}, we get

$$\mathbb{E}_\sigma\left[\|\sigma\|_2\right] = \frac{\int_{r=0}^{(1+d^{-1})\eta} r \times r^{d-1} dr}{\int_{r=0}^{(1+d^{-1})\eta} r^{d-1} dr} = \eta.$$

We now bound the stability of predictions of OFTPL. Our technique for bounding the stability uses similar arguments as Hazan and Minasyan [HM20] (see Lemma 4.2 of [HM20]). Recall, to bound stability, we need to show that $\Phi(g) = \mathbb{E}_\sigma\left[\inf_{\mathbf{x}\in\mathcal{X}} \langle g-\sigma, \mathbf{x}\rangle\right]$ is smooth. Let $\phi_0(g) = \inf_{\mathbf{x}\in\mathcal{X}}\langle g, \mathbf{x}-\mathbf{x}_{00}\rangle$, where $\mathbf{x}_{00}$ is an arbitrary point in $\mathcal{X}$. We can rewrite $\Phi(g)$ as

$$\Phi(g) = \mathbb{E}_\sigma\left[\phi_0(g-\sigma)\right] + \langle g, \mathbf{x}_{00}\rangle.$$

Since the second term in the RHS above is linear in g, any upper bound on the smoothness of $\mathbb{E}_\sigma\left[\phi_0(g-\sigma)\right]$ is also a bound on the smoothness of $\Phi(g)$. So we focus on bounding the smoothness of $\mathbb{E}_\sigma\left[\phi_0(g-\sigma)\right]$.

First note that $\phi_0(g)$ is D Lipschitz and satisfies the following for any $g_1, g_2 \in \mathbb{R}^d$

$$\begin{aligned}
\phi_0(g_1) - \phi_0(g_2) &= \inf_{\mathbf{x}\in\mathcal{X}}\langle -g_2, \mathbf{x}-\mathbf{x}_{00}\rangle - \inf_{\mathbf{x}\in\mathcal{X}}\langle -g_1, \mathbf{x}-\mathbf{x}_{00}\rangle\\
&\leq \sup_{\mathbf{x}\in\mathcal{X}}\langle g_1-g_2, \mathbf{x}-\mathbf{x}_{00}\rangle\\
&\leq D\|g_1-g_2\|_2.
\end{aligned}$$

Letting $\Phi_0(g) = \mathbb{E}_\sigma\left[\phi_0(g-\sigma)\right]$, Lemma 4.2 of Hazan and Minasyan [HM20] shows that $\Phi_0(g)$ is smooth and satisfies

$$\|\nabla\Phi_0(g_1) - \nabla\Phi_0(g_2)\|_2 \leq dD\eta^{-1}\|g_1-g_2\|_2.$$

This shows that the predictions of OFTPL are $dD\eta^{-1}$ stable. The rest of the proof involves substituting $C = dD$ in the regret bound of Theorem 5 and setting $g_t = 0$ and using the fact that $\|\nabla_t\|_2 \leq G$.

B.3 Online Nonconvex Learning

B.3.1 Proof of Theorem 6

Before we present the proof of the Theorem, we introduce some notation and present some useful intermediate results. We note that unlike the convex case, there are no know Fenchel duality theorems for infinite dimensional setting. So more careful arguments are need to obtain tight regret bounds. Our proof mimics the proof of Theorem 5.

Notation

Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{X}$. We define functions $\Phi : \mathcal{F} \to \mathbb{R}$, $R : \mathcal{P} \to \mathbb{R}$ as follows

$$\begin{aligned}
\Phi(f) &= \mathbb{E}_\sigma\left[\inf_{P\in\mathcal{P}} \mathbb{E}_{\mathbf{x}\sim P}\left[f(\mathbf{x}) - \sigma(\mathbf{x})\right]\right],\\
R(P) &= \sup_{f\in\mathcal{F}} -\mathbb{E}_{\mathbf{x}\sim P}\left[f(\mathbf{x})\right] + \Phi(f).
\end{aligned}$$

Also, note that the function $\nabla\Phi : \mathcal{F} \to \mathcal{P}$ defined in Section 3.3.2 can be written as

$$\nabla\Phi\left(f\right) = \mathbb{E}_\sigma\left[\operatorname*{argmin}_{P\in\mathcal{P}} \mathbb{E}_{\mathbf{x}\sim P}\left[f(\mathbf{x}) - \sigma(\mathbf{x})\right]\right].$$

Note that, $\nabla\Phi\left(f\right)$ is well defined because from our assumption on the perturbation distribution, the minimization problem inside the expectation has a unique minimizer with probability one. To simplify the notation, in the sequel, we use the shorthand notation $\langle P, f\rangle$ to denote $\mathbb{E}_{\mathbf{x}\sim P}\left[f(\mathbf{x})\right]$, for any $P \in \mathcal{P}$ and $f \in \mathcal{F}$. Similarly, for any $P_1, P_2 \in \mathcal{P}$ and $f \in \mathcal{F}$, we use the notation $\langle P_1 - P_2, f\rangle$ to denote $\mathbb{E}_{\mathbf{x}\sim P_1}\left[f(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim P_2}\left[f(\mathbf{x})\right]$.

Intermediate Results

Lemma 23. *For any $g \in \mathcal{F}$, $R(\nabla\Phi\left(g\right)) = -\langle\nabla\Phi\left(g\right), g\rangle + \Phi(g)$.*

Proof. Define $P_{g,\sigma}$ as

$$P_{g,\sigma} = \operatorname*{argmin}_{P\in\mathcal{P}} \mathbb{E}_{\mathbf{x}\sim P}\left[g(\mathbf{x}) - \sigma(\mathbf{x})\right].$$

Note that $\nabla\Phi\left(g\right) = \mathbb{E}_\sigma\left[P_{g,\sigma}\right]$. For any $g, h \in \mathcal{F}$, we have

$$\begin{aligned}\Phi(h) &= \mathbb{E}_\sigma\left[\inf_{P\in\mathcal{P}} \langle P, h - \sigma\rangle\right]\\ &\leq \mathbb{E}_\sigma\left[\langle P_{g,\sigma}, h - \sigma\rangle\right]\\ &= \mathbb{E}_\sigma\left[\langle P_{g,\sigma}, g - \sigma\rangle\right] + \mathbb{E}_\sigma\left[\langle P_{g,\sigma}, h - g\rangle\right]\\ &= \Phi(g) + \langle\nabla\Phi\left(g\right), h - g\rangle.\end{aligned}$$

This shows that for any $g, h \in \mathcal{F}$

$$\Phi(h) - \langle\nabla\Phi\left(g\right), h\rangle \leq \Phi(g) - \langle\nabla\Phi\left(g\right), g\rangle. \tag{B.4}$$

Taking supremum over h of the LHS quantity gives us

$$R(\nabla\Phi\left(g\right)) = \sup_{h\in\mathcal{F}} \Phi(h) - \langle\nabla\Phi\left(g\right), h\rangle = \Phi(g) - \langle\nabla\Phi\left(g\right), g\rangle.$$

□

Lemma 24 (Strong Smoothness)**.** *The function $-\Phi$ is convex and strongly smooth and satisfies the following inequality for any $g_1, g_2 \in \mathcal{F}$*

$$-\Phi(g_2) \leq -\Phi(g_1) - \langle\nabla\Phi\left(g_1\right), g_2 - g_1\rangle + \frac{C}{2\eta}\|g_2 - g_1\|_{\mathcal{F}}^2.$$

Proof. Let $g_1, g_2 \in \mathcal{F}$ and $\alpha \in [0, 1]$. Then

$$\begin{aligned}\Phi(\alpha g_1 + (1-\alpha)g_2) &= \mathbb{E}_\sigma\left[\inf_{P\in\mathcal{P}} \langle P, \alpha g_1 + (1-\alpha)g_2 - \sigma\rangle\right]\\ &\geq \alpha\mathbb{E}_\sigma\left[\inf_{P\in\mathcal{P}} \langle P, g_1 - \sigma\rangle\right] + (1-\alpha)\mathbb{E}_\sigma\left[\inf_{P\in\mathcal{P}} \langle P, g_2 - \sigma\rangle\right]\\ &= \alpha\Phi(g_1) + (1-\alpha)\Phi(g_2).\end{aligned}$$

This shows that $-\Phi$ is convex. To show smoothness, we rely on the following stability property

$$\forall g_1, g_2 \in \mathcal{F} \quad \gamma_{\mathcal{F}}(\nabla\Phi\left(g_1\right), \nabla\Phi\left(g_2\right)) \leq \frac{C}{\eta}\|g_1 - g_2\|_{\mathcal{F}}.$$

Let T be an arbitrary positive integer and for $t \in \{0, 1, \dots T\}$, define $\alpha_t = t/T$. Let $h = g_2 - g_1$. We have

$$\begin{aligned}\Phi(g_1) - \Phi(g_2) &= \Phi(g_1 + \alpha_0 h) - \Phi(g_1 + \alpha_T h)\\ &= \sum_{t=0}^{T-1}\left(\Phi(g_1 + \alpha_t h) - \Phi(g_1 + \alpha_{t+1} h)\right)\end{aligned}$$

Since $-\Phi$ is convex and satisfies Equation (B.4), we have

$$\begin{aligned}\Phi(g_1) - \Phi(g_2) &= \sum_{t=0}^{T-1}\left(\Phi(g_1 + \alpha_t h) - \Phi(g_1 + \alpha_{t+1} h)\right)\\ &\leq -\sum_{t=0}^{T-1}\frac{1}{T}\langle\nabla\Phi\left(g_1 + \alpha_{t+1}h\right), h\rangle\end{aligned}$$

Using stability, we get

$$\begin{aligned}\Phi(g_1) - \Phi(g_2) &\leq -\sum_{t=0}^{T-1}\frac{1}{T}\langle\nabla\Phi\left(g_1 + \alpha_{t+1}h\right), h\rangle\\ &= \sum_{t=0}^{T-1}\frac{1}{T}\left(\langle\nabla\Phi\left(g_1\right) - \nabla\Phi\left(g_1 + \alpha_{t+1}h\right), h\rangle - \langle\nabla\Phi\left(g_1\right), h\rangle\right)\\ &\overset{(a)}{\leq} -\langle\nabla\Phi\left(g_1\right), h\rangle + \sum_{t=0}^{T-1}\frac{1}{T}\gamma_{\mathcal{F}}(\nabla\Phi\left(g_1\right), \nabla\Phi\left(g_1 + \alpha_{t+1}h\right))\|h\|_{\mathcal{F}}\\ &\overset{(b)}{\leq} -\langle\nabla\Phi\left(g_1\right), h\rangle + \sum_{t=0}^{T-1}\frac{C}{T\eta}\|\alpha_{t+1}h\|_{\mathcal{F}}\|h\|_{\mathcal{F}}\\ &= -\langle\nabla\Phi\left(g_1\right), h\rangle + \sum_{t=0}^{T-1}\frac{C\alpha_{t+1}}{T\eta}\|h\|_{\mathcal{F}}^2\\ &= -\langle\nabla\Phi\left(g_1\right), h\rangle + \frac{C}{\eta}\frac{T+1}{2T}\|h\|_{\mathcal{F}}^2,\end{aligned}$$

where (a) follows from the definition of $\gamma_{\mathcal{F}}$ and (b) follows from the stability assumption. Taking $T \to \infty$, we get

$$-\Phi(g_2) \leq -\Phi(g_1) - \langle\nabla\Phi\left(g_1\right), g_2 - g_1\rangle + \frac{C}{2\eta}\|g_2 - g_1\|_{\mathcal{F}}^2.$$

□

Lemma 25 (Strong Convexity)**.** *For any $P \in \mathcal{P}$ and $g \in \mathcal{F}$, R satisfies the following inequality*

$$R(P) \geq R(\nabla\Phi\left(g\right)) + \langle \nabla\Phi\left(g\right) - P, g\rangle + \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(g\right))^2.$$

Proof. From Lemma 24 we know that the following holds for any $g, h \in \mathcal{F}$

$$\Phi(g) \geq \underbrace{\Phi(h) + \langle \nabla\Phi\left(h\right), g - h\rangle - \frac{C}{2\eta}\|g - h\|_{\mathcal{F}}^2}_{\Phi_{\text{lb},h}(g)}.$$

Define $R_{\text{lb},h}(P)$ as

$$R_{\text{lb}}(P) = \sup_{g\in\mathcal{F}} -\langle P, g\rangle + \Phi_{\text{lb},h}(g).$$

Since $\Phi(g) \geq \Phi_{\text{lb},h}(g)$ for all $g \in \mathcal{F}$, $R(P) \geq R_{\text{lb},h}(P)$ for all P. We now derive an expression for $R_{\text{lb},h}(P)$. Note that from Lemma 23 we have $R(\nabla\Phi\left(h\right)) = -\langle\nabla\Phi\left(h\right), h\rangle + \Phi(h)$. Using this, we get

$$\begin{aligned}
R_{\text{lb},h}(P) &= \sup_{g\in\mathcal{F}} -\langle P, g\rangle + \Phi_{\text{lb},h}(g)\\
&\overset{(a)}{=} \sup_{g\in\mathcal{F}}\left(-\langle P, g\rangle + \Phi(h) + \langle\nabla\Phi\left(h\right), g - h\rangle - \frac{C}{2\eta}\|g - h\|_{\mathcal{F}}^2\right)\\
&\overset{(b)}{=} R(\nabla\Phi\left(h\right)) + \sup_{g\in\mathcal{F}}\left(\langle\nabla\Phi\left(h\right) - P, g\rangle - \frac{C}{2\eta}\|g - h\|_{\mathcal{F}}^2\right),
\end{aligned}$$

where (a) follows from the definition of $\Phi_{\text{lb},h}(g)$ and (b) follows from Lemma 23. We now do a change of variables in the supremum of the above expression. Substituting $g' = g - h$, we get

$$R_{\text{lb},h}(P) = R(\nabla\Phi\left(h\right)) + \langle\nabla\Phi\left(h\right) - P, h\rangle + \sup_{g'\in\mathcal{F}}\left(\langle\nabla\Phi\left(h\right) - P, g'\rangle - \frac{C}{2\eta}\|g'\|_{\mathcal{F}}^2\right).$$

We now show that

$$\sup_{g'\in\mathcal{F}}\left(\langle\nabla\Phi\left(h\right) - P, g'\rangle - \frac{C}{2\eta}\|g'\|_{\mathcal{F}}^2\right) \geq \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2.$$

To this end, we choose a $g'' \in \mathcal{F}$ such that

$$\|g''\|_{\mathcal{F}} = \frac{\eta}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right)), \quad \langle\nabla\Phi\left(h\right) - P, g''\rangle = \frac{\eta}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2. \tag{B.5}$$

If such a g'' can be found, we have

$$\begin{aligned}
\sup_{g'\in\mathcal{F}}\left(\langle\nabla\Phi\left(h\right) - P, g'\rangle - \frac{C}{2\eta}\|g'\|_{\mathcal{F}}^2\right) &\geq \langle\nabla\Phi\left(h\right) - P, g''\rangle - \frac{C}{2\eta}\|g''\|_{\mathcal{F}}^2\\
&= \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2.
\end{aligned}$$

This would then imply the main claim of the Lemma.

$$R(P) \geq R_{\text{lb},h}(P) \geq R(\nabla\Phi\left(h\right)) + \langle\nabla\Phi\left(h\right) - P, h\rangle + \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2.$$

Finding g''. We now construct a g'' which satisfies Equation (B.5). From the definition of $\gamma_{\mathcal{F}}$ we know that

$$\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right)) = \sup_{\|g'\|_{\mathcal{F}}\leq 1} |\langle \nabla\Phi\left(h\right) - P, g'\rangle|$$

Suppose the supremum is achieved at g^*. Define g'' as $\frac{\eta s}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))g^*$, where $s = \text{sign}(\langle \nabla\Phi\left(h\right) - P, g^*\rangle)$. It can be easily verified that g'' satifies Equation (B.5).

If the supremum is never achieved, the same argument as above can still be made using a sequence of functions $\{g_n\}_{n=1}^{\infty}$ such that

$$\|g_n\|_{\mathcal{F}} \leq 1, \quad \lim_{n\to\infty} |\langle \nabla\Phi\left(h\right) - P, g_n\rangle| = \gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right)).$$

Define g''_n as $\frac{\eta s_n}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))g_n$, where $s_n = \text{sign}(\langle \nabla\Phi\left(h\right) - P, g_n\rangle)$. Since $\lim_{n\to\infty}\|g_n\|_{\mathcal{F}} = 1$, we have $\lim_{n\to\infty}\|g''_n\|_{\mathcal{F}} = \frac{\eta}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))$. Moreover,

$$\lim_{n\to\infty}\langle \nabla\Phi\left(h\right) - P, g''_n\rangle = \lim_{n\to\infty}\frac{\eta}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))\Big|\langle \nabla\Phi\left(h\right) - P, g_n\rangle\Big| = \frac{\eta}{C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2.$$

This shows that

$$\begin{aligned}\sup_{g'\in\mathcal{F}}\left(\langle \nabla\Phi\left(h\right) - P, g'\rangle - \frac{C}{2\eta}\|g'\|_{\mathcal{F}}^2\right) &\geq \lim_{n\to\infty}\langle \nabla\Phi\left(h\right) - P, g''_n\rangle - \frac{C}{2\eta}\|g''_n\|_{\mathcal{F}}^2 \\ &= \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \nabla\Phi\left(h\right))^2.\end{aligned}$$

This finishes the proof of the Lemma. □

Main Argument

We are now ready to prove Theorem 6. Our proof relies on Lemma 25 and uses similar arguments as used in the proof of Theorem 5. We first rewrite $P_t, \tilde{P}_t$ as

$$\begin{aligned}P_t &= \frac{1}{m}\sum_{j=1}^{m}\mathop{\text{argmin}}_{P\in\mathcal{P}}\mathbb{E}_{\mathbf{x}\sim P}\left[\sum_{i=1}^{t-1} f_i(\mathbf{x}) + g_t(\mathbf{x}) - \sigma_{t,j}(\mathbf{x})\right], \\ \tilde{P}_t &= \frac{1}{m}\sum_{j=1}^{m}\mathop{\text{argmin}}_{P\in\mathcal{P}}\mathbb{E}_{\mathbf{x}\sim P}\left[\sum_{i=1}^{t} f_i(\mathbf{x}) - \sigma'_{t,j}(\mathbf{x})\right].\end{aligned}$$

Note that

$$\begin{aligned}P_t^{\infty} &= \mathbb{E}\left[P_t | g_t, f_{1:t-1}, P_{1:t-1}\right] = \nabla\Phi\left(f_{1:t-1} + g_t\right), \\ \tilde{P}_t^{\infty} &= \mathbb{E}\left[\tilde{P}_t | f_{1:t-1}, P_{1:t-1}\right] = \nabla\Phi\left(f_{1:t}\right),\end{aligned}$$

with $P_1^{\infty} = \tilde{P}_0^{\infty} = \nabla\Phi\left(0\right)$. Define functions $B(\cdot, P_t^{\infty}), B(\cdot, \tilde{P}_t^{\infty})$ as

$$\begin{aligned}B(P, P_t^{\infty}) &= R(P) - R(P_t^{\infty}) + \langle P - P_t^{\infty}, f_{1:t-1} + g_t\rangle, \\ B(P, \tilde{P}_t^{\infty}) &= R(P) - R(\tilde{P}_t^{\infty}) + \langle P - \tilde{P}_t^{\infty}, f_{1:t}\rangle.\end{aligned}$$

From Lemma 25, we have

$$B(P, P_t^\infty) \geq \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, P_t^\infty)^2, \quad B(P, \tilde{P}_t^\infty) \geq \frac{\eta}{2C}\gamma_{\mathcal{F}}(P, \tilde{P}_t^\infty)^2.$$

For any $P \in \mathcal{P}$, we have

$$\begin{aligned}
\mathbb{E}\left[f_t(\mathbf{x}_t) - f_t(P)\right] &= \mathbb{E}\left[f_t(P_t) - f_t(P)\right] \\
&= \mathbb{E}\left[\langle P_t - P, f_t\rangle\right] \\
&= \mathbb{E}\left[\langle P_t - P_t^\infty, f_t\rangle\right] + \mathbb{E}\left[\langle P_t^\infty - P, f_t\rangle\right] \\
&= \mathbb{E}\left[\langle P_t - P_t^\infty, f_t\rangle\right] + \mathbb{E}\left[\langle P_t^\infty - \tilde{P}_t^\infty, f_t - g_t\rangle\right] \\
&\quad + \mathbb{E}\left[\langle P_t^\infty - \tilde{P}_t^\infty, g_t\rangle\right] + \mathbb{E}\left[\langle \tilde{P}_t^\infty - P, f_t\rangle\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_t^\infty)\|f_t - g_t\|_{\mathcal{F}}\right] + \mathbb{E}\left[\langle P_t^\infty - \tilde{P}_t^\infty, g_t\rangle\right] \\
&\quad + \mathbb{E}\left[\langle \tilde{P}_t^\infty - P, f_t\rangle\right],
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}\left[\langle P_t - P_t^\infty, f_t\rangle | g_t, f_{1:t-1}, P_{1:t-1}\right] = 0$ and as a result $\mathbb{E}\left[\langle P_t - P_t^\infty, f_t\rangle\right] = 0$. Next, a simple calculation shows that

$$\begin{aligned}
\langle P_t^\infty - \tilde{P}_t^\infty, g_t\rangle &= B(\tilde{P}_t^\infty, \tilde{P}_{t-1}^\infty) - B(\tilde{P}_t^\infty, P_t^\infty) - B(P_t^\infty, \tilde{P}_{t-1}^\infty) \\
\langle \tilde{P}_t^\infty - P, f_t\rangle &= B(P, \tilde{P}_{t-1}^\infty) - B(P, \tilde{P}_t^\infty) - B(\tilde{P}_t^\infty, \tilde{P}_{t-1}^\infty).
\end{aligned}$$

Substituting this in the previous regret bound gives us

$$\begin{aligned}
\mathbb{E}\left[f_t(\mathbf{x}_t) - f_t(P)\right] &\leq \mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_t^\infty)\|f_t - g_t\|_{\mathcal{F}}\right] + \mathbb{E}\left[B(\tilde{P}_t^\infty, \tilde{P}_{t-1}^\infty) - B(\tilde{P}_t^\infty, P_t^\infty) - B(P_t^\infty, \tilde{P}_{t-1}^\infty)\right] \\
&\quad + \mathbb{E}\left[B(P, \tilde{P}_{t-1}^\infty) - B(P, \tilde{P}_t^\infty) - B(\tilde{P}_t^\infty, \tilde{P}_{t-1}^\infty)\right] \\
&= \mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_t^\infty)\|f_t - g_t\|_{\mathcal{F}}\right] \\
&\quad + \mathbb{E}\left[B(P, \tilde{P}_{t-1}^\infty) - B(P, \tilde{P}_t^\infty) - B(\tilde{P}_t^\infty, P_t^\infty) - B(P_t^\infty, \tilde{P}_{t-1}^\infty)\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_t^\infty)\|f_t - g_t\|_{\mathcal{F}}\right] \\
&\quad + \mathbb{E}\left[B(P, \tilde{P}_{t-1}^\infty) - B(P, \tilde{P}_t^\infty)\right] - \mathbb{E}\left[\frac{\eta}{2C}\gamma_{\mathcal{F}}(\tilde{P}_t^\infty, P_t^\infty)^2 + \frac{\eta}{2C}\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\right] \\
&\overset{(b)}{\leq} \frac{C}{2\eta}\mathbb{E}\left[\|f_t - g_t\|_{\mathcal{F}}^2\right] + \mathbb{E}\left[B(P, \tilde{P}_{t-1}^\infty) - B(P, \tilde{P}_t^\infty)\right] - \mathbb{E}\left[\frac{\eta}{2C}\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\right]
\end{aligned}$$

where (a) follows from Lemma 25, and (b) uses the fact that $|xy| \leq \frac{1}{2c}|x|^2 + \frac{c}{2}|y|^2$, for any $x, y, c > 0$. Summing over $t = 1, \dots T$ gives us

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[f_t(\mathbf{x}_t) - f_t(P)\right] \leq \underbrace{\mathbb{E}\left[B(P, \tilde{P}_0^\infty) - B(P, \tilde{P}_T^\infty)\right]}_{S_1} + \sum_{t=1}^{T}\frac{C}{2\eta}\mathbb{E}\left[\|f_t - g_t\|_{\mathcal{F}}^2\right] \\
- \sum_{t=1}^{T}\frac{\eta}{2C}\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\right]
\end{aligned}$$

To finish the proof of the Theorem, we need to bound S_1.

Bounding S_1. From the definition of B, we have

$$B(P, \tilde{P}_0^\infty) - B(P, \tilde{P}_T^\infty) = R(\tilde{P}_T^\infty) - \langle P - \tilde{P}_T^\infty, f_{1:T}\rangle - R(\tilde{\mathbf{x}}_0^\infty),$$

where we used the fact that $f_{1:0} = 0$. We now rely on Lemma 23 to convert the above equation, which is currently in terms of R, into a quantity which depends on Φ. Using Lemma 23, we get

$$B(P, \tilde{P}_0^\infty) - B(P, \tilde{P}_T^\infty) = \Phi(f_{1:T}) - \langle P, f_{1:T}\rangle - \Phi(0).$$

From the definition of Φ we have

$$\begin{aligned}
B(P, \tilde{P}_0^\infty) - B(P, \tilde{P}_T^\infty) &= \Phi(f_{1:T}) - \langle P, f_{1:T}\rangle - \Phi(0)\\
&= \mathbb{E}_\sigma\left[\inf_{P'\in\mathcal{P}} \langle P', f_{1:T} - \sigma\rangle\right] - \langle P, f_{1:T}\rangle - \mathbb{E}_\sigma\left[\inf_{P'\in\mathcal{P}} \langle P', -\sigma\rangle\right]\\
&\leq \mathbb{E}_\sigma\left[\langle P, f_{1:T} - \sigma\rangle\right] - \langle P, f_{1:T}\rangle - \mathbb{E}_\sigma\left[\inf_{P'\in\mathcal{P}} \langle P', -\sigma\rangle\right]\\
&= \mathbb{E}_\sigma\left[\sup_{P'\in\mathcal{P}} \langle P', \sigma\rangle\right] - \mathbb{E}_\sigma\left[\langle P, \sigma\rangle\right]\\
&\leq D\mathbb{E}_\sigma\left[\|\sigma\|_{\mathcal{F}}\right] = \eta D,
\end{aligned}$$

where the last inequality follows from our bound on the diameter of $\mathcal{P}$. Substituting this in the above regret bound gives us the required result.

B.3.2 Proof of Corollary 2

To prove the corollary we first show that for our choice of perturbation distribution, $\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x})$ has a unique minimizer with probability one, for any $f \in \mathcal{F}$. Next, we show that the predictions of OFTPL are stable.

Intermediate Results

Lemma 26 (Unique Minimizer)**.** *Suppose the perturbation function is such that $\sigma(\mathbf{x}) = \langle\bar{\sigma}, \mathbf{x}\rangle$, where $\bar{\sigma} \in \mathbb{R}^d$ is a random vector whose entries are sampled independently from Exp(η). Then, for any $f \in \mathcal{F}$, $\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x})$ has a unique minimizer with probability one.*

Proof. Define $\mathbf{x}_f(\sigma)$ as

$$\mathbf{x}_f(\bar{\sigma}) \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \langle\bar{\sigma}, \mathbf{x}\rangle.$$

For any $\bar{\sigma}_1, \bar{\sigma}_2$ we now show that $\mathbf{x}_f(\bar{\sigma})$ satisfies the following monotonicity property

$$\langle \mathbf{x}_f(\bar{\sigma}_1) - \mathbf{x}_f(\bar{\sigma}_2), \bar{\sigma}_1 - \bar{\sigma}_2\rangle \geq 0.$$

From the optimality of $\mathbf{x}_f(\bar{\sigma}_1), \mathbf{x}_f(\bar{\sigma}_2)$ we have

$$\begin{aligned} f(\mathbf{x}_f(\bar{\sigma}_1)) - \langle \bar{\sigma}_1, \mathbf{x}_f(\bar{\sigma}_1)\rangle &\leq f(\mathbf{x}_f(\bar{\sigma}_2)) - \langle \bar{\sigma}_1, \mathbf{x}_f(\bar{\sigma}_2)\rangle \\ &= f(\mathbf{x}_f(\bar{\sigma}_2)) - \langle \bar{\sigma}_2, \mathbf{x}_f(\bar{\sigma}_2)\rangle + \langle \bar{\sigma}_2 - \bar{\sigma}_1, \mathbf{x}_f(\bar{\sigma}_2)\rangle \\ &\leq f(\mathbf{x}_f(\bar{\sigma}_1)) - \langle \bar{\sigma}_2, \mathbf{x}_f(\bar{\sigma}_1)\rangle + \langle \bar{\sigma}_2 - \bar{\sigma}_1, \mathbf{x}_f(\bar{\sigma}_2)\rangle. \end{aligned}$$

This shows that $\langle \bar{\sigma}_2 - \bar{\sigma}_1, \mathbf{x}_f(\bar{\sigma}_2) - \mathbf{x}_f(\bar{\sigma}_1)\rangle \geq 0$. To finish the proof of Lemma, we rely on Theorem 1 of Zarantonello [Zar73], which shows that the set of points for which a monotone operator is not single-valued has Lebesgue measure zero. Since the distribution of $\bar{\sigma}$ is absolutely continuous w.r.t Lebesgue measure, this shows that $\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x})$ has a unique minimizer with probability one. □

Main Argument

For our choice of perturbation distribution, $\mathbb{E}_{\sigma}\left[\|\sigma\|_{\mathcal{F}}\right] = \mathbb{E}_{\bar{\sigma}}\left[\|\bar{\sigma}\|_{\infty}\right] = \eta \log d$. We now bound the stability of predictions of OFTPL. First note that for our choice of primal space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$, $\gamma_{\mathcal{F}}$ is the Wasserstein-1 metric, which is defined as

$$\gamma_{\mathcal{F}}(P_1, P_2) = \sup_{f\in\mathcal{F}, \|f\|_{\mathcal{F}}\leq 1} \left| \mathbb{E}_{\mathbf{x}\sim P_1}\left[f(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim P_2}\left[f(\mathbf{x})\right] \right| = \inf_{Q\in\Gamma(P_1,P_2)} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim Q}\left[\|\mathbf{x}_1 - \mathbf{x}_2\|_1\right],$$

where $\Gamma(P_1, P_2)$ is the set of all probability measures on $\mathcal{X}\times\mathcal{X}$ with marginals P_1, P_2 on the first and second factors respectively. Define $\mathbf{x}_f(\bar{\sigma})$ as

$$\mathbf{x}_f(\bar{\sigma}) \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \langle \bar{\sigma}, \mathbf{x}\rangle.$$

Note that $\nabla\Phi\left(f\right)$ is the distribution of random variable $\mathbf{x}_f(\bar{\sigma})$. In Chapter 2 we showed that for any $f, g\in\mathcal{F}$

$$\mathbb{E}_{\bar{\sigma}}\left[\|\mathbf{x}_f(\bar{\sigma}) - \mathbf{x}_g(\bar{\sigma})\|_1\right] \leq \frac{125 d^2 D}{\eta}\|f-g\|_{\mathcal{F}}.$$

Since $\gamma_{\mathcal{F}}(\nabla\Phi\left(f\right), \nabla\Phi\left(g\right)) \leq \mathbb{E}_{\bar{\sigma}}\left[\|\mathbf{x}_f(\bar{\sigma}) - \mathbf{x}_g(\bar{\sigma})\|_1\right]$, this shows that OFTPL is $O\left(d^2 D\eta^{-1}\right)$ stable w.r.t $\|\cdot\|_{\mathcal{F}}$. Substituting the stability bound in the regret bound of Theorem 6 shows that

$$\begin{aligned} \sup_{P\in\mathcal{P}} \mathbb{E}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(P)\right] &= \eta D \log d \\ &\quad + O\left(\sum_{t=1}^T \frac{d^2 D}{\eta}\mathbb{E}\left[\|f_t - g_t\|_{\mathcal{F}}^2\right] - \sum_{t=1}^T \frac{\eta}{d^2 D}\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^{\infty}, \tilde{P}_{t-1}^{\infty})^2\right]\right). \end{aligned}$$

B.4 Convex-Concave Games

Our algorithm for convex-concave games is presented in Algorithm 13. Before presenting the proof of Theorem 7, we first present a more general result in Section B.4.1. Theorem 7 immediately follows from our general result by instantiating it for the uniform noise distribution.

Algorithm 13 OFTPL for convex-concave games

```
Input: Perturbation Distributions $P^1_{\text{PRTB}}, P^2_{\text{PRTB}}$ of $\mathbf{x}, \mathbf{y}$ players, number of samples $m$, iterations $T$
for $t = 1 \dots T$ do
    if $t = 1$ then
        Sample $\{\sigma^1_{1,j}\}_{j=1}^m$, $\{\sigma^2_{1,j}\}_{j=1}^m$ from $P^1_{\text{PRTB}}, P^2_{\text{PRTB}}$
        $\mathbf{x}_1 = \frac{1}{m}\sum_{j=1}^m \left[\text{argmin}_{\mathbf{x}\in\mathcal{X}}\langle -\sigma^1_{1,j}, \mathbf{x}\rangle\right], \mathbf{y}_1 = \frac{1}{m}\left[\sum_{j=1}^m \text{argmax}_{\mathbf{y}\in\mathcal{Y}}\langle \sigma^2_{1,j}, \mathbf{y}\rangle\right]$
        continue
    end if
    //Compute guesses
    for $j = 1 \dots m$ do
        Sample $\sigma^1_{t,j} \sim P^1_{\text{PRTB}}, \sigma^2_{t,j} \sim P^2_{\text{PRTB}}$
        $\tilde{\mathbf{x}}_{t-1,j} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\langle \sum_{i=1}^{t-1} \nabla_{\mathbf{x}} f(\mathbf{x}_i, \mathbf{y}_i) - \sigma^1_{t,j}, \mathbf{x}\rangle$
        $\tilde{\mathbf{y}}_{t-1,j} = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\langle \sum_{i=1}^{t-1} \nabla_{\mathbf{y}} f(\mathbf{x}_i, \mathbf{y}_i) + \sigma^2_{t,j}, \mathbf{y}\rangle$
    end for
    $\tilde{\mathbf{x}}_{t-1} = \frac{1}{m}\sum_{j=1}^m \tilde{\mathbf{x}}_{t-1,j}$, $\tilde{\mathbf{y}}_{t-1} = \frac{1}{m}\sum_{j=1}^m \tilde{\mathbf{y}}_{t-1,j}$
    //Use the guesses to compute the next action
    for $j = 1 \dots m$ do
        Sample $\sigma^1_{t,j} \sim P^1_{\text{PRTB}}, \sigma^2_{t,j} \sim P^2_{\text{PRTB}}$
        $\mathbf{x}_{t,j} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\langle \sum_{i=1}^{t-1} \nabla_{\mathbf{x}} f(\mathbf{x}_i, \mathbf{y}_i) + \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}) - \sigma^1_{t,j}, \mathbf{x}\rangle$
        $\mathbf{y}_{t,j} = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\langle \sum_{i=1}^{t-1} \nabla_{\mathbf{y}} f(\mathbf{x}_i, \mathbf{y}_i) + \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}) + \sigma^2_{t,j}, \mathbf{y}\rangle$
    end for
    $\mathbf{x}_t = \frac{1}{m}\sum_{j=1}^m \mathbf{x}_{t,j}, \mathbf{y}_t = \frac{1}{m}\sum_{j=1}^m \mathbf{y}_{t,j}$
end for
return $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$
```

B.4.1 General Result

Theorem 27. *Consider the minimax game in Equation (3.1). Suppose f is convex in $\mathbf{x}$, concave in $\mathbf{y}$ and is Holder smooth w.r.t some norm $\|\cdot\|$*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}')\|_* \leq L_1\|\mathbf{x} - \mathbf{x}'\|^\alpha + L_2\|\mathbf{y} - \mathbf{y}'\|^\alpha,$$
$$\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}')\|_* \leq L_2\|\mathbf{x} - \mathbf{x}'\|^\alpha + L_1\|\mathbf{y} - \mathbf{y}'\|^\alpha.$$

Define diameter of sets $\mathcal{X}, \mathcal{Y}$ as $D = \max\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}}\|\mathbf{x}_1 - \mathbf{x}_2\|, \sup_{\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}}\|\mathbf{y}_1 - \mathbf{y}_2\|\}$. Let $L = \{L_1, L_2\}$. Suppose both $\mathbf{x}$ and $\mathbf{y}$ players use Algorithm 2 to solve the minimax game. Suppose the perturbation distributions P^1_{PRTB}, P^2_{PRTB}, used by $\mathbf{x}$, $\mathbf{y}$ players are absolutely continuous and satisfy $\mathbb{E}_{\sigma\sim P^1_{PRTB}}[\|\sigma\|_*] = \mathbb{E}_{\sigma\sim P^2_{PRTB}}[\|\sigma\|_*] = \eta$. Suppose the predictions of both the players are $C\eta^{-1}$-stable w.r.t $\|\cdot\|_*$. Suppose the guesses used by $\mathbf{x}, \mathbf{y}$ players in the t^{th} iteration are $\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}), \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})$, where $\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}$ denote the predictions of $\mathbf{x}, \mathbf{y}$ players in the t^{th} iteration, if guess $g_t = 0$ was used in that iteration. Then the *iterates* $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ *generated by the OFTPL based algorithm satisfy*

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^T \mathbf{y}_t\right)\right] \leq 2L_1\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + \frac{2\eta D}{T} + \frac{20CL^2}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} + 10L\left(\frac{5CL}{\eta}\right)^{\frac{1+\alpha}{1-\alpha}}$$

Proof. Since both the players are responding to each others actions using OFTPL, using Theorem 5, we get the following regret bounds for the players

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}, \mathbf{y}_t)\right] \leq{}& L_1 T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + \eta D \\ &+ \frac{C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})\|_*^2\right] \\ &- \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right].\end{aligned}$$

$$\begin{aligned}\sup_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}_t, \mathbf{y}_t)\right] \leq{}& L_1 T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + \eta D \\ &+ \frac{C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})\|_*^2\right] \\ &- \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^2\right].\end{aligned}$$

First, consider the regret of the $\mathbf{x}$ player. Since $\|a_1 + \cdots + a_5\|^2 \leq 5(\|a_1\|^2 \cdots + \|a_5\|^2)$, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})\|_*^2 \leq{}& 5\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t)\|_*^2 \\ &+ 5\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty)\|_*^2 \\ &+ 5\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2 \\ &+ 5\|\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1})\|_*^2 \\ &+ 5\|\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})\|_*^2 \\ \overset{(a)}{\leq}{}& 5L_1^2\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{2\alpha} + 5L_1^2\|\tilde{\mathbf{x}}_{t-1} - \tilde{\mathbf{x}}_{t-1}^\infty\|^{2\alpha} \\ &+ 5L_2^2\|\mathbf{y}_t - \mathbf{y}_t^\infty\|^{2\alpha} + 5L_2^2\|\tilde{\mathbf{y}}_{t-1} - \tilde{\mathbf{y}}_{t-1}^\infty\|^{2\alpha} \\ &+ 5\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2.\end{aligned}$$

where (a) follows from the Holder's smoothness of f. Using a similar technique as in the proof of Theorem 5, relying on Holders inequality, we get

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{2\alpha} | \tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}, \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}\right] &\leq \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{2} | \tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}, \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}\right]^\alpha \\
&\leq \Psi_1^{2\alpha}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_t^\infty\|_2^{2} | \tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}, \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}\right]^\alpha \\
&\overset{(a)}{\leq} \left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha},
\end{aligned}$$

where (a) follows from the fact that conditioned on past randomness, $\mathbf{x}_t - \mathbf{x}_t^\infty$ is the average of m i.i.d bounded mean 0 random variables, the variance of which scales as $O(D^2/m)$. A similar bound holds for the expectation of other quantities appearing in the RHS of the above equation. Using this, the regret of $\mathbf{x}$ player can be upper bounded as

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}, \mathbf{y}_t)\right] &\leq L_1 T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + \eta D + \frac{10CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} \\
&\quad + \frac{5C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\
&\quad - \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right].
\end{aligned}$$

Similarly, the regret of $\mathbf{y}$ player can be bounded as

$$\begin{aligned}
\sup_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}_t, \mathbf{y}_t)\right] &\leq L_1 T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + \eta D + \frac{10CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} \\
&\quad + \frac{5C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\
&\quad - \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^2\right].
\end{aligned}$$

Summing the above two inequalities, we get

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_t)\right] &\leq 2L_1 T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + 2\eta D + \frac{20CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} \\
&\quad + \frac{5C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\
&\quad + \frac{5C}{2\eta}\sum_{t=1}^T \mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\
&\quad - \frac{\eta}{2C}\sum_{t=1}^T \left(\mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right]\right).
\end{aligned}$$

From Holder's smoothness assumption on f, we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] &\leq 2\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\ &\quad + 2\mathbb{E}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty, \tilde{\mathbf{y}}_{t-1}^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \\ &\overset{(a)}{\leq} 2L^2\mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^{2\alpha}\right] + 2L^2\mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^{2\alpha}\right],\end{aligned}$$

Using a similar argument, we get

$$\mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty, \mathbf{y}_t^\infty) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty, \tilde{\mathbf{y}}_{t-1}^\infty)\|_*^2\right] \leq 2L^2\mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^{2\alpha}\right] + 2L^2\mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^{2\alpha}\right].$$

Plugging this in the previous bound, we get

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_t)\right] &\leq 2L_1T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + 2\eta D + \frac{20CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} \\ &\quad + \frac{10CL^2}{\eta}\sum_{t=1}^T\left(\mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^{2\alpha}\right] + \mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^{2\alpha}\right]\right) \\ &\quad - \frac{\eta}{2C}\sum_{t=1}^T\left(\mathbb{E}\left[\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right]\right).\end{aligned}$$

Case $\alpha = 1$. We first consider the case of $\alpha = 1$. In this case, choosing $\eta > \sqrt{20}CL$, we get

$$\sup_{\mathbf{x}\in\mathcal{X}\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_t)\right] \leq 2L_1T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + 2\eta D + \frac{20CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha}.$$

General α. The more general case relies on AM-GM inequality. Consider the following

$$\begin{aligned}\frac{10CL^2}{\eta}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^{2\alpha} &= \left((2\alpha C)^{\frac{\alpha}{1-\alpha}}\eta^{-\frac{1+\alpha}{1-\alpha}}(10CL^2)^{\frac{1}{1-\alpha}}\right)^{1-\alpha}\left(\frac{\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2}{2\alpha C\eta^{-1}}\right)^\alpha \\ &\overset{(a)}{\leq} (1-\alpha)\left((2\alpha C)^{\frac{\alpha}{1-\alpha}}\eta^{-\frac{1+\alpha}{1-\alpha}}(10CL^2)^{\frac{1}{1-\alpha}}\right) + \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2 \\ &= \sqrt{20}L\left(\frac{\sqrt{20}CL}{\eta}\right)^{\frac{1+\alpha}{1-\alpha}} + \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\end{aligned}$$

where (a) follows from AM-GM inequality. Plugging this in the previous bound, we get

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_t)\right] \leq& 2L_1T\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{1+\alpha} + 2\eta D \\ &+ \frac{20CL^2T}{\eta}\left(\frac{\Psi_1\Psi_2 D}{\sqrt{m}}\right)^{2\alpha} + 4\sqrt{5}LT\left(\frac{\sqrt{20}CL}{\eta}\right)^{\frac{1+\alpha}{1-\alpha}}.\end{aligned}$$

The claim of the theorem then follows from the observation that

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y}\right)-f\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\right]\leq\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}f(\mathbf{x}_t,\mathbf{y})-f(\mathbf{x},\mathbf{y}_t)\right].$$

□

B.4.2 Proof of Theorem 7

To prove the Theorem, we instantiate Theorem 27 for the uniform noise distribution. As shown in Corollary 1, the predictions of OFTPL are $dD\eta^{-1}$-stable in this case. Plugging this in the bound of Theorem 27 and using the fact that $\Psi_1=\Psi_2=1$ and $\alpha=1$ gives us

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y}\right)-f\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\right]\leq&2L\left(\frac{D}{\sqrt{m}}\right)^2+\frac{2\eta D}{T}\\&+\frac{20dDL^2}{\eta}\left(\frac{D}{\sqrt{m}}\right)^2+10L\left(\frac{5dDL}{\eta}\right)^{\infty}.\end{aligned}$$

Plugging in $\eta=6dD(L+1)$, $m=T$ in the above bound gives us

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y}\right)-f\left(\mathbf{x},\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\right]\leq O\left(\frac{dD^2(L+1)}{T}\right).$$

B.5 Nonconvex-Nonconcave Games

Our algorithm for nonconvex-nonconcave games is presented in Algorithm 14. Note that in each iteration of this game, both the players play empirical distributions (P_t,Q_t). Before presenting the proof of Theorem 8, we first present a more general result in Section B.5.1. Theorem 8 immediately follows from our general result by instantiating it for exponential noise distribution.

B.5.1 General Result

Theorem 28. *Consider the minimax game in Equation (3.1). Suppose the domains $\mathcal{X},\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$. Let $\mathcal{F},\mathcal{F}'$ be the set of Lipschitz functions over $\mathcal{X},\mathcal{Y}$, and $\|g_1\|_{\mathcal{F}},\|g_2\|_{\mathcal{F}'}$ be the Lipschitz constants of functions $g_1:\mathcal{X}\to\mathbb{R}$, $g_2:\mathcal{Y}\to\mathbb{R}$ w.r.t some norm $\|\cdot\|$. Suppose f is such that $\max\{\sup_{\mathbf{x}\in\mathcal{X}}\|f(\cdot,\mathbf{y})\|_{\mathcal{F}},\sup_{\mathbf{y}\in\mathcal{Y}}\|f(\mathbf{x},\cdot)\|_{\mathcal{F}'}\}\leq G$ and satisfies the following smoothness property*

$$\begin{aligned}\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{x}}f(\mathbf{x}',\mathbf{y}')\|_*&\leq L\|\mathbf{x}-\mathbf{x}'\|+L\|\mathbf{y}-\mathbf{y}'\|,\\\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f(\mathbf{x}',\mathbf{y}')\|_*&\leq L\|\mathbf{x}-\mathbf{x}'\|+L\|\mathbf{y}-\mathbf{y}'\|.\end{aligned}$$

Let $\mathcal{P},\mathcal{Q}$ be the set of probability distributions over $\mathcal{X},\mathcal{Y}$. Define diameter of $\mathcal{P},\mathcal{Q}$ as $D=\max\{\sup_{P_1,P_2\in\mathcal{P}}\gamma_{\mathcal{F}}(P_1,P_2),\sup_{Q_1,Q_2\in\mathcal{Q}}\gamma_{\mathcal{F}'}(Q_1,Q_2)\}$. Suppose both $\mathbf{x},\mathbf{y}$ players use Algorithm 3 to solve the game. Suppose the perturbation distributions P^1_{PRTB},P^2_{PRTB}, used by $\mathbf{x}$,

Algorithm 14 OFTPL for nonconvex-nonconcave games

```
Input: Perturbation Distributions $P^1_{\text{PRTB}}, P^2_{\text{PRTB}}$ of $\mathbf{x}, \mathbf{y}$ players, number of samples $m$, iterations $T$
for $t = 1 \dots T$ do
    if $t = 1$ then
        for $j = 1 \dots m$ do
            Sample $\sigma^1_{t,j} \sim P^1_{\text{PRTB}}, \sigma^2_{t,j} \sim P^2_{\text{PRTB}}$
            $\mathbf{x}_{1,j} = \text{argmin}_{\mathbf{x}\in\mathcal{X}} -\sigma^1_{1,j}(\mathbf{x})$
            $\mathbf{y}_{1,j} = \text{argmax}_{\mathbf{y}\in\mathcal{Y}} \sigma^2_{1,j}(\mathbf{y})$
        end for
        Let $P_1, Q_1$ be the empirical distributions over $\{\mathbf{x}_{1,j}\}_{j=1}^m, \{\mathbf{y}_{1,j}\}_{j=1}^m$
        continue
    end if
    //Compute guesses
    for $j = 1 \dots m$ do
        Sample $\sigma^1_{t,j} \sim P^1_{\text{PRTB}}, \sigma^2_{t,j} \sim P^2_{\text{PRTB}}$
        $\tilde{\mathbf{x}}_{t-1,j} = \text{argmin}_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{t-1} f(\mathbf{x}, Q_i) - \sigma^1_{t,j}(\mathbf{x})$
        $\tilde{\mathbf{y}}_{t-1,j} = \text{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{i=1}^{t-1} f(P_i, \mathbf{y}) + \sigma^2_{t,j}(\mathbf{y})$
    end for
    Let $\tilde{P}_{t-1}$, $\tilde{Q}_{t-1}$ be the empirical distributions over $\{\tilde{\mathbf{x}}_{t-1,j}\}_{j=1}^m, \{\tilde{\mathbf{y}}_{t-1,j}\}_{j=1}^m$
    //Use the guesses to compute the next action
    for $j = 1 \dots m$ do
        Sample $\sigma^1_{t,j} \sim P^1_{\text{PRTB}}, \sigma^2_{t,j} \sim P^2_{\text{PRTB}}$
        $\mathbf{x}_{t,j} = \text{argmin}_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{t-1} f(\mathbf{x}, Q_i) + f(\mathbf{x}, \tilde{Q}_{t-1}) - \sigma^1_{t,j}(\mathbf{x})$
        $\mathbf{y}_{t,j} = \text{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{i=1}^{t-1} f(P_i, \mathbf{y}) + f(\tilde{P}_{t-1}, \mathbf{y}) + \sigma^2_{t,j}(\mathbf{y})$
    end for
    Let $P_t, Q_t$ be the empirical distributions over $\{\mathbf{x}_{t,j}\}_{j=1}^m, \{\mathbf{y}_{t,j}\}_{j=1}^m$
end for
return $\{(P_t, Q_t)\}_{t=1}^T$
```

$\mathbf{y}$ *players are such that* $\text{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - \sigma(\mathbf{x}), \text{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{y}) + \sigma(\mathbf{y})$ *have unique optimizers with probability one, for any* f *in* $\mathcal{F}, \mathcal{F}'$ *respectively. Moreover, suppose* $\mathbb{E}_{\sigma\sim P^1_{PRTB}}[\|\sigma\|_{\mathcal{F}}] = \mathbb{E}_{\sigma\sim P^2_{PRTB}}[\|\sigma\|_{\mathcal{F}'}] = \eta$ *and predictions of both the players are* $C\eta^{-1}$*-stable w.r.t norms* $\|\cdot\|_{\mathcal{F}}, \|\cdot\|_{\mathcal{F}'}$*. Suppose the guesses used by* $\mathbf{x}, \mathbf{y}$ *players in the* t^{th} *iteration are* $f(\cdot, \tilde{Q}_{t-1}), f(\tilde{P}_{t-1}, \cdot)$*, where* $\tilde{P}_{t-1}, \tilde{Q}_{t-1}$ *denote the predictions of* $\mathbf{x}, \mathbf{y}$ *players in the* t^{th} *iteration, if guess* $g_t = 0$ *was used. Then the iterates* $\{(P_t, Q_t)\}_{t=1}^T$ *generated by the Algorithm 13 satisfy the following, for* $\eta > \sqrt{3}CL$

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T P_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^T Q_t\right)\right] = O\left(\frac{\eta D}{T} + \frac{CD^2L^2}{\eta m}\right)$$
$$+ O\left(\min\left\{\frac{dC\Psi_1^2\Psi_2^2G^2\log(2m)}{\eta m}, \frac{CD^2L^2}{\eta}\right\}\right).$$

Proof. The proof of this Theorem uses similar arguments as Theorem 27. Since both the players are responding to each others actions using OFTPL, using Theorem 6, we get the following regret bounds for the players

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} f(P_t, Q_t) - f(\mathbf{x}, Q_t)\right] \leq \eta D + \sum_{t=1}^{T} \frac{C}{2\eta}\mathbb{E}\left[\|f(\cdot, Q_t) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2\right] - \frac{\eta}{2C}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^{\infty}, \tilde{P}_{t-1}^{\infty})^2\right],$$

$$\sup_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^{T} f(P_t, \mathbf{y}) - f(P_t, Q_t)\right] \leq \eta D + \sum_{t=1}^{T} \frac{C}{2\eta}\mathbb{E}\left[\|f(P_t, \cdot) - f(\tilde{P}_{t-1}, \cdot)\|_{\mathcal{F}'}^2\right] - \frac{\eta}{2C}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_{\mathcal{F}'}(Q_t^{\infty}, \tilde{Q}_{t-1}^{\infty})^2\right],$$

where $P_t^{\infty}, \tilde{P}_{t-1}^{\infty}, Q_t^{\infty}, \tilde{Q}_{t-1}^{\infty}$ are as defined in Theorem 6. First, consider the regret of the $\mathbf{x}$ player. We upper bound $\|f(\cdot, Q_t) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2$ as

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2 \leq\ & 3\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}^2 \\ & + 3\|f(\cdot, Q_t^{\infty}) - f(\cdot, \tilde{Q}_{t-1}^{\infty})\|_{\mathcal{F}}^2 \\ & + 3\|f(\cdot, \tilde{Q}_{t-1}^{\infty}) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2.\end{aligned}$$

We now show that $\mathbb{E}\left[\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}^2 | \tilde{P}_{t-1}, \tilde{Q}_{t-1}, P_{1:t-1}, Q_{1:t-1}\right]$ is $O(1/m)$. To simplify the notation, we let $\zeta_t = \{\tilde{P}_{t-1}, \tilde{Q}_{t-1}, P_{1:t-1}, Q_{1:t-1}\}$. Let $\mathcal{N}_\epsilon$ be the ϵ-net of $\mathcal{X}$ w.r.t $\|\cdot\|$. Then

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}} &\overset{(a)}{=} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_* \\ &\overset{(b)}{\leq} \sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_* + 2L\epsilon,\end{aligned}$$

where (a) follows from the definition of Lipschitz constant and (b) follows from our smoothness assumption on f. Using this, we get

$$\mathbb{E}\left[\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}^2 | \zeta_t\right] \leq 2\mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_*^2 \Big| \zeta_t\right] + 8L^2\epsilon^2,$$

Since f is Lipschitz, $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|_*$ is bounded by G. So $\|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_*$ is bounded by $2G$ and $\|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_2$ is bounded by $2\Psi_1 G$. Moreover, conditioned on past randomness (ζ_t), $\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})$ is a sub-Gaussian random vector and satisfies the following bound

$$\mathbb{E}\left[\langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\rangle | \zeta_t\right] \leq \exp\left(2\Psi_1^2 G^2 \|\mathbf{u}\|_2^2 / m\right).$$

From tail bounds of sub-Gaussian random vectors [HKZ+12], we have

$$\mathbb{P}\left(\|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_2^2 > \frac{4\Psi_1^2 G^2}{m}(d + 2\sqrt{ds} + 2s)\Big|\zeta_t\right) \leq e^{-s},$$

for any $s > 0$. Using union bound, and the fact that $\log|\mathcal{N}_\epsilon|$ is upper bounded by $d\log(1 + 2D/\epsilon)$, we get

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_2^2 > \frac{4\Psi_1^2 G^2}{m}(d + 2\sqrt{ds} + 2s)\Big|\zeta_t\right) \leq e^{-s + d\log(1+2D/\epsilon)}.$$

Let $Z = \sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_2^2$. The expectation of Z can be bounded as follows

$$\begin{aligned}\mathbb{E}\left[Z|\zeta_t\right] &= \mathbb{P}(Z \leq a|\zeta_t)\mathbb{E}\left[Z|\zeta_t, Z \leq a\right] + \mathbb{P}(Z > a|\zeta_t)\mathbb{E}\left[Z|\zeta_t, Z > a\right]\\ &\leq a + 4\Psi_1^2 G^2 \mathbb{P}(Z > a|\zeta_t).\end{aligned}$$

Choosing $\epsilon = Dm^{-1/2}$, $s = 3d\log(1 + 2m^{1/2})$, and $a = \frac{44d\Psi_1^2 G^2 \log(1+2m^{1/2})}{m}$, we get

$$\mathbb{E}\left[Z|\zeta_t\right] \leq \frac{48d\Psi_1^2 G^2 \log(1 + 2m^{1/2})}{m}.$$

This shows that $\mathbb{E}\left[\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}^2|\zeta_t\right] \leq \frac{96d\Psi_1^2\Psi_2^2 G^2 \log(1+2m^{1/2})}{m} + \frac{8D^2L^2}{m}$. Note that another trivial upper bound for $\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}$ is DL, which can obtained as follows

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}} &= \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^{\infty})\|_*\\ &= \|\mathbb{E}_{\mathbf{y}_1\sim Q_t, \mathbf{y}_2\sim Q_t^{\infty}}\left[\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_2)\right]\|_*\\ &\overset{(a)}{\leq} LD,\end{aligned}$$

where (a) follows from the smoothness assumption on f and the fact that the diameter of $\mathcal{X}$ is D. When L is close to 0, this bound can be much better than the above bound. So we have

$$\mathbb{E}\left[\|f(\cdot, Q_t) - f(\cdot, Q_t^{\infty})\|_{\mathcal{F}}^2|\zeta_t\right] \leq \min\left(\frac{96d\Psi_1^2\Psi_2^2 G^2 \log(1 + 2m^{1/2})}{m} + \frac{8D^2L^2}{m}, L^2D^2\right).$$

Using this, the regret of the $\mathbf{x}$ player can be bounded as follows

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T f(P_t, Q_t) - f(\mathbf{x}, Q_t)\right] &\leq \eta D + \frac{24CD^2L^2T}{\eta m}\\ &\quad + \min\left(\frac{288dC\Psi_1^2\Psi_2^2 G^2 T\log(1 + 2m^{1/2})}{\eta m}, \frac{3CD^2L^2T}{\eta}\right)\\ &\quad + \sum_{t=1}^T \frac{3C}{2\eta}\mathbb{E}\left[\|f(\cdot, Q_t^{\infty}) - f(\cdot, \tilde{Q}_{t-1}^{\infty})\|_{\mathcal{F}}^2\right]\\ &\quad - \frac{\eta}{2C}\sum_{t=1}^T \mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^{\infty}, \tilde{P}_{t-1}^{\infty})^2\right].\end{aligned}$$

A similar analysis shows that the regret of $\mathbf{y}$ player can be bounded as

$$\begin{aligned}\sup_{\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[\sum_{t=1}^{T} f(P_t,\mathbf{y})-f(P_t,Q_t)\right] \leq \eta D + \frac{24CD^2L^2T}{\eta m}\\
+\min\left(\frac{288dC\Psi_1^2\Psi_2^2G^2T\log(1+2m^{1/2})}{\eta m},\frac{3CD^2L^2T}{\eta}\right)\\
+\sum_{t=1}^{T}\frac{3C}{2\eta}\mathbb{E}\left[\|f(P_t^{\infty},\cdot)-f(\tilde{P}_{t-1}^{\infty},\cdot)\|_{\mathcal{F}'}^2\right]\\
-\frac{\eta}{2C}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_{\mathcal{F}'}(Q_t^{\infty},\tilde{Q}_{t-1}^{\infty})^2\right],\end{aligned}$$

Summing the above two inequalities, we get

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\mathbb{E}\left[\sum_{t=1}^{T} f(P_t,\mathbf{y})-f(P,Q_t)\right] \leq 2\eta D + \frac{48CD^2L^2T}{\eta m}\\
+\min\left(\frac{576dC\Psi_1^2\Psi_2^2G^2T\log(1+2m^{1/2})}{\eta m},\frac{6CD^2L^2T}{\eta}\right)\\
+\sum_{t=1}^{T}\frac{3C}{2\eta}\mathbb{E}\left[\|f(\cdot,Q_t^{\infty})-f(\cdot,\tilde{Q}_{t-1}^{\infty})\|_{\mathcal{F}}^2\right]\\
+\sum_{t=1}^{T}\frac{3C}{2\eta}\mathbb{E}\left[\|f(P_t^{\infty},\cdot)-f(\tilde{P}_{t-1}^{\infty},\cdot)\|_{\mathcal{F}'}^2\right]\\
-\frac{\eta}{2C}\sum_{t=1}^{T}\left(\mathbb{E}\left[\gamma_{\mathcal{F}}(P_t^{\infty},\tilde{P}_{t-1}^{\infty})^2\right]+\mathbb{E}\left[\gamma_{\mathcal{F}'}(Q_t^{\infty},\tilde{Q}_{t-1}^{\infty})^2\right]\right).\end{aligned}$$

From our assumption on smoothness of f, we have

$$\|f(\cdot,Q_t^{\infty})-f(\cdot,\tilde{Q}_{t-1}^{\infty})\|_{\mathcal{F}}\leq L\gamma_{\mathcal{F}'}(Q_t^{\infty},\tilde{Q}_{t-1}^{\infty}),\quad \|f(P_t^{\infty},\cdot)-f(\tilde{P}_{t-1}^{\infty},\cdot)\|_{\mathcal{F}'}\leq L\gamma_{\mathcal{F}}(P_t^{\infty},\tilde{P}_{t-1}^{\infty}).$$

To see this, consider the following

$$
\begin{aligned}
\|f(\cdot, Q_t^\infty) - f(\cdot, \tilde{Q}_{t-1}^\infty)\|_{\mathcal{F}} &= \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty) - \nabla_{\mathbf{x}} f(\mathbf{x}, \tilde{Q}_{t-1}^\infty)\|_* \\
&= \sup_{\mathbf{x}\in\mathcal{X}, \|\mathbf{u}\|\leq 1} \langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty) - \nabla_{\mathbf{x}} f(\mathbf{x}, \tilde{Q}_{t-1}^\infty)\rangle \\
&= \sup_{\mathbf{x}\in\mathcal{X}, \|\mathbf{u}\|\leq 1} \mathbb{E}_{\mathbf{y}\sim Q_t^\infty}\left[\langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\rangle\right] - \mathbb{E}_{\mathbf{y}\sim \tilde{Q}_{t-1}^\infty}\left[\langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\rangle\right] \\
&\leq \gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty) \sup_{\mathbf{x}\in\mathcal{X}, \|\mathbf{u}\|\leq 1} \|\langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, \cdot)\rangle\|_{\mathcal{F}'} \\
&= \gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty) \sup_{\mathbf{x}\in\mathcal{X}, \|\mathbf{u}\|\leq 1} \left(\sup_{\mathbf{y}_1\neq\mathbf{y}_2\in\mathcal{Y}} \frac{|\langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_1)\rangle - \langle \mathbf{u}, \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_2)\rangle|}{\|\mathbf{y}_1 - \mathbf{y}_2\|} \right) \\
&\leq \gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty) \sup_{\mathbf{x}\in\mathcal{X}} \left(\sup_{\mathbf{y}_1\neq\mathbf{y}_2\in\mathcal{Y}} \frac{\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_2)\|_*}{\|\mathbf{y}_1 - \mathbf{y}_2\|} \right) \\
&\overset{(a)}{\leq} L\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty),
\end{aligned}
$$

where (a) follows from smoothness of f. Substituting this in the previous equation, and choosing $\eta > \sqrt{3}CL$, we get

$$
\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[\sum_{t=1}^T f(P_t, \mathbf{y}) - f(P, Q_t)\right] &\leq 2\eta D + \frac{48CD^2L^2T}{\eta m} \\
&\quad + \min\left(\frac{576dC\Psi_1^2\Psi_2^2G^2T\log(1+2m^{1/2})}{\eta m}, \frac{6CD^2L^2T}{\eta}\right)
\end{aligned}
$$

This finishes the proof of the Theorem. □

Remark B.5.1. *We note that a similar result can be obtained for other choice of function classes such as the set of all bounded and Lipschitz functions. The only difference between proving such a result vs. proving Theorem 28 is in bounding $\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}$.*

B.5.2 Proof of Theorem 8

To prove the Theorem, we instantiate Theorem 28 for exponential noise distribution. Recall, in Corollary 2, we showed that $\mathbb{E}_\sigma\left[\|\sigma\|_{\mathcal{F}}\right] = \eta \log d$ and OFTPL is $O\left(d^2D\eta^{-1}\right)$ stable w.r.t $\|\cdot\|_{\mathcal{F}}$, for this choice of perturbation distribution (similar results hold for $(\mathcal{F}', \|\cdot\|_{\mathcal{F}'})$). Substituting this in the bounds of Theorem 28 and using the fact that $\Psi_1 = \sqrt{d}, \Psi_2 = 1$, we get

$$
\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^T P_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^T Q_t\right)\right] &= O\left(\frac{\eta D\log d}{T} + \frac{d^2D^3L^2}{\eta m}\right) \\
&\quad + O\left(\min\left\{\frac{d^4DG^2\log(2m)}{\eta m}, \frac{d^2D^3L^2}{\eta}\right\}\right).
\end{aligned}
$$

Choosing $\eta = 10d^2D(L+1), m = T$, we get

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}P_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^{T}Q_t\right)\right] = O\left(\frac{d^2D^2(L+1)\log d}{T}\right)$$
$$+ O\left(\min\left\{\frac{d^2G^2\log(T)}{LT}, D^2L\right\}\right).$$

B.6 Choice of Perturbation Distributions

Regularization of some Perturbation Distributions. We first study the regularization effect of various perturbation distributions. Table B.1 presents the regularizer R corresponding to some commonly used perturbation distributions, when the action space $\mathcal{X}$ is ℓ_∞ ball of radius 1 centered at origin.

Perturbation Distribution P_{PRTB}	Regularizer
Uniform over $[0,\eta]^d$	$\eta\|\mathbf{x}-1\|_2^2$
Exponential $P(\sigma > t) = \exp(-t/\eta)$	$\sum_i \eta(\mathbf{x}_i+1)\left[\log(\mathbf{x}_i+1) - (1+\log 2)\right]$
Gaussian $P(\sigma = t) \propto e^{-t^2/2\eta^2}$	$\sum_i \sup_{u\in\mathbb{R}} u\left[\mathbf{x}_i - 1 + 2F(-u/\eta)\right]$

Table B.1: Regularizers corresponding to various perturbation distributions used in FTPL when the action space $\mathcal{X}$ is ℓ_∞ ball of radius 1 centered at origin. Here, F is the CDF of a standard normal random variable.

Dimension independent rates. Recall, the OFTPL algorithm described in Algorithm 13 converges at $O(d/T)$ rate to a Nash equilibrium of smooth convex-concave games (see Theorem 7). We now show that for certain constraint sets $\mathcal{X}, \mathcal{Y}$, by choosing the perturbation distributions appropriately, the dimension dependence in the rates can *potentially* be removed.

Suppose the action set is $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$. Suppose the perturbation distribution P_{PRTB} is the multivariate Gaussian distribution with mean 0 and covariance $\eta^2 I_{d\times d}$, where $I_{d\times d}$ is the identity matrix. We now try to explicitly compute the reguralizer corresponding to this perturbation distribution and action set. Define function Ψ as

$$\Psi(f) = \mathbb{E}_\sigma\left[\max_{\mathbf{x}\in\mathcal{X}}\langle f+\sigma, \mathbf{x}\rangle\right] = \mathbb{E}_\sigma\left[\|f+\sigma\|_2\right].$$

As shown in Proposition 2, the regularizer R corresponding to any perturbation distribution is given by the Fenchel conjugate of Ψ

$$R(\mathbf{x}) = \sup_f \langle f, \mathbf{x}\rangle - \Psi(f).$$

Since getting an exact expression for R is a non-trivial task, we only compute an *approximate expression* for R. Consider the high dimensional setting (*i.e.,* very large d). In this setting, $\|f+\sigma\|_2$, for σ drawn from $\mathcal{N}(0, \eta^2 I_{d\times d})$, can be approximated as follows

$$\begin{aligned} \|f+\sigma\|_2 &= \sqrt{\|f\|_2^2 + \|\sigma\|_2^2 + 2\langle f, \sigma\rangle} \\ &\overset{(a)}{\approx} \sqrt{\|f\|_2^2 + \eta^2 d + 2\langle f, \sigma\rangle} \\ &\overset{(b)}{\approx} \sqrt{\|f\|_2^2 + \eta^2 d} \end{aligned}$$

where (a) follows from the fact that $\|\sigma\|_2^2$ is highly concentrated around $\eta^2 d$ [HKZ+12]. To be precise

$$\mathbb{P}(\|\sigma\|_2^2 \geq \eta^2(d + 2\sqrt{dt} + 2t)) \leq e^{-t}.$$

A similar bound holds for the lower tail. Approximation (b) follows from the fact that $\langle f, \sigma\rangle$ is a Gaussian random variable with mean 0 and variance $\eta^2\|f\|_2^2$, and with high probability its magnitude is upper bounded by $\tilde{O}(\eta\|f\|_2)$. Since $\eta\|f\|_2 \ll \sqrt{d}\eta\|f\|_2 \leq \|f\|_2^2 + \eta^2 d$, approximation (b) holds. This shows that $\Psi(f)$ can be approximated as

$$\Psi(f) \approx \sqrt{\|f\|_2^2 + \eta^2 d}.$$

Using this approximation, we now compute the reguralizer corresponding to the perturbation distribution

$$R(\mathbf{x}) = \sup_f \langle f, \mathbf{x}\rangle - \Psi(f) \approx \sup_f \langle f, \mathbf{x}\rangle - \sqrt{\|f\|_2^2 + \eta^2 d} = -\eta\sqrt{d}\sqrt{1 - \|\mathbf{x}\|_2^2}.$$

This shows that R is $\eta\sqrt{d}$-strongly convex w.r.t $\|\cdot\|_2$ norm. Following duality between strong convexity and strong smoothness, $\Psi(f)$ is $(\eta^2 d)^{-1/2}$ strongly smooth w.r.t $\|\cdot\|_2$ norm and satisfies

$$\|\nabla\Psi(f_1) - \nabla\Psi(f_2)\|_2 \leq (\eta^2 d)^{-1/2}\|f_1 - f_2\|_2.$$

This shows that the predictions of OFTPL are $(\eta^2 d)^{-1/2}$ stable w.r.t $\|\cdot\|_2$ norm. We now instantiate Theorem 27 for this perturbation distribution and for constraint sets which are unit balls centered at origin, and use the above stability bound, together with the fact that $\mathbb{E}_\sigma\left[\|\sigma\|_2\right] \approx \eta\sqrt{d}$. Suppose f is smooth w.r.t $\|\cdot\|_2$ norm and satisfies

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}')\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}')\|_2 \leq L\|\mathbf{x} - \mathbf{x}'\|_2 + L\|\mathbf{y} - \mathbf{y}'\|_2.$$

Then Theorem 27 gives us the following rates of convergence to a NE

$$\begin{aligned} \sup_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\right)\right] \leq & \frac{2L_1}{m} + \frac{2\eta\sqrt{d}}{T} \\ & + \frac{20L^2}{\eta\sqrt{d}}\left(\frac{1}{m}\right) + 10L\left(\frac{5L}{\eta\sqrt{d}}\right)^{\infty} \end{aligned}$$

Choosing $\eta = 6L/\sqrt{d}, m = T$, we get $O\left(\frac{L}{T}\right)$ rate of convergence. Although, these rates are dimension independent, we note that our stability bound is only approximate. More accurate analysis is needed to actually claim that Algorithm 13 achieves dimension independent rates in this setting. That being said, for general constraints sets, we believe one can get dimension independent rates by choosing the perturbation distribution appropriately.

B.7 High Probability Bounds

In this section, we provide high probability bounds for Theorems 5, 7. Our results rely on the following concentration inequalities.

Proposition 14 (Jin, Netrapalli, Ge, Kakade, and Jordan [Jin+19])**.** *Let $X_1, \dots X_K$ be K independent mean 0 vector-valued random variables such that $\|X_i\|_2 \leq B_i$. Then*

$$\mathbb{P}\left(\|\sum_{i=1}^K X_i\|_2 \geq t\right) \leq 2\exp\left(-c\frac{t^2}{\sum_{i=1}^K B_i^2}\right),$$

where $c > 0$ is a universal constant.

We also need the following concentration inequality for martingales.

Proposition 15 (Wainwright [Wai19])**.** *Let $X_1, \dots X_K \in \mathbb{R}$ be a martingale difference sequence, where $\mathbb{E}\left[X_i|\mathcal{F}_{i-1}\right] = 0$. Assume that X_i satisfy the following tail condition, for some scalar $B_i > 0$*

$$\mathbb{P}\left(\left|\frac{X_i}{B_i}\right| \geq z\middle|\mathcal{F}_{i-1}\right) \leq 2\exp(-z^2).$$

Then

$$\mathbb{P}\left(\left|\sum_{i=1}^K X_i\right| \geq z\right) \leq 2\exp\left(-c\frac{z^2}{\sum_{i=1}^K B_i^2}\right),$$

where $c > 0$ is a universal constant.

B.7.1 Online Convex Learning

In this section, we present a high probability version of Theorem 5.

Theorem 29. *Suppose the perturbation distribution P_{PRTB} is absolutely continuous w.r.t Lebesgue measure. Let D be the diameter of $\mathcal{X}$ w.r.t $\|\cdot\|$, which is defined as $D = \sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}}\|\mathbf{x}_1 - \mathbf{x}_2\|$. Let $\eta = \mathbb{E}_\sigma\left[\|\sigma\|_*\right]$, and suppose the predictions of OFTPL are $C\eta^{-1}$-stable w.r.t $\|\cdot\|_*$, where C is a constant that depends on the set $\mathcal{X}$. Suppose, the sequence of loss functions $\{f_t\}_{t=1}^T$ are G-Lipschitz w.r.t $\|\cdot\|$ and satisfy $\sup_{\mathbf{x}\in\mathcal{X}}\|\nabla f_t(\mathbf{x})\|_* \leq G$. Moreover, suppose $\{f_t\}_{t=1}^T$ are Holder smooth and satisfy*

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \quad \|\nabla f_t(\mathbf{x}_1) - \nabla f_t(\mathbf{x}_2)\|_* \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|^\alpha,$$

for some constant $\alpha \in [0,1]$. Then the regret of Algorithm 2 satisfies the following with probability at least $1-\delta$

$$\sup_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \leq \eta D + \sum_{t=1}^T \frac{C}{2\eta}\|\nabla_t - g_t\|_*^2 - \sum_{t=1}^T \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2$$
$$+ cGD\sqrt{\frac{T\log 2/\delta}{m}} + cLT\left(\frac{\Psi_1^2\Psi_2^2D^2\log 4T/\delta}{m}\right)^{\frac{1+\alpha}{2}},$$

where c is a universal constant, $\mathbf{x}_t^\infty = \mathbb{E}\left[\mathbf{x}_t | g_t, f_{1:t-1}, \mathbf{x}_{1:t-1}\right]$ and $\tilde{\mathbf{x}}_{t-1}^\infty = \mathbb{E}\left[\tilde{\mathbf{x}}_{t-1} | f_{1:t-1}, \mathbf{x}_{1:t-1}\right]$ and $\tilde{\mathbf{x}}_{t-1}$ denotes the prediction in the t^{th} iteration of Algorithm 2, if guess $g_t = 0$ was used. Here, Ψ_1, Ψ_2 denote the norm compatibility constants of $\|\cdot\|$.

Proof. Our proof uses the same notation and similar arguments as in the proof Theorem 5. Recall, in Theorem 5 we showed that the regret of OFTPL is upper bounded by

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \leq \sum_{t=1}^T \langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \eta D + \sum_{t=1}^T \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_t^\infty\|\|\nabla_t - g_t\|_*$$
$$- \frac{\eta}{2C}\sum_{t=1}^T \left(\|\tilde{\mathbf{x}}_t^\infty - \mathbf{x}_t^\infty\|^2 + \|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2\right)$$
$$\leq \sum_{t=1}^T \langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t\rangle + \eta D + \sum_{t=1}^T \frac{C}{2\eta}\|\nabla_t - g_t\|_*^2 - \sum_{t=1}^T \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2.$$

From Holder's smoothness assumption, we have

$$\langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla_t - \nabla f_t(\mathbf{x}_t^\infty)\rangle \leq L\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha}.$$

Substituting this in the previous bound gives us

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \leq \underbrace{\sum_{t=1}^T \langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla f_t(\mathbf{x}_t^\infty)\rangle}_{S_1} + \sum_{t=1}^T L\underbrace{\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha}}_{S_2} + \eta D$$
$$+ \sum_{t=1}^T \frac{C}{2\eta}\|\nabla_t - g_t\|_*^2 - \sum_{t=1}^T \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2.$$

We now provide high probability bounds for S_1 and S_2.

Bounding S_1. Let $\xi_i = \{g_{i+1}, f_{i+1}, \mathbf{x}_i\}$ and let $\xi_{0:t}$ denote the union of sets $\xi_0, \xi_1, \ldots, \xi_t$. Let $\zeta_t = \langle \mathbf{x}_t - \mathbf{x}_t^\infty, \nabla f_t(\mathbf{x}_t^\infty)\rangle$ with $\zeta_0 = 0$. Note that $\{\zeta_t\}_{t=0}^T$ is a martingale difference sequence w.r.t $\xi_{0:T}$. This is because $\mathbb{E}\left[\mathbf{x}_t|\xi_{0:t-1}\right] = \mathbf{x}_t^\infty$ and $\nabla f_t(\mathbf{x}_t^\infty)$ is a deterministic quantity conditioned on $\xi_{0:t-1}$. As a result $\mathbb{E}\left[\zeta_t|\xi_{0:t-1}\right] = 0$. Moreover, conditioned on

$\xi_{0:t-1}$, ζ_t is the average of m independent mean 0 random variables, each of which is bounded by GD. Using Proposition 14, we get

$$\mathbb{P}\left(|\zeta_t| \geq s \middle| \xi_{0:t-1}\right) \leq 2\exp\left(-\frac{ms^2}{G^2D^2}\right).$$

Using Proposition 42 on the martingale difference sequence $\{\zeta_t\}_{t=0}^T$, we get

$$\mathbb{P}\left(\left|\sum_{t=1}^T \zeta_t\right| \geq s\right) \leq 2\exp\left(-c\frac{ms^2}{G^2D^2T}\right),$$

where $c > 0$ is a universal constant. This shows that with probability at least $1-\delta/2$, S_1 is upper bounded by $O\left(\sqrt{\frac{G^2D^2T\log\frac{2}{\delta}}{m}}\right)$.

Bounding S_2. Conditioned on $\{g_t, f_{1:t-1}, \mathbf{x}_{1:t-1}\}$, $\mathbf{x}_t - \mathbf{x}_t^\infty$ is the average of m independent mean 0 random variables which are bounded by D in $\|\cdot\|$ norm. From our definition of norm compatibility constant Ψ_2, this implies the random variables are bounded by $\Psi_2 D$ in $\|\cdot\|_2$. Using Proposition 14, we get

$$\mathbb{P}\left(\|\mathbf{x}_t - \mathbf{x}_t^\infty\|_2 \geq \Psi_2 D\sqrt{\frac{c\log 4T/\delta}{m}} \middle| g_t, f_{1:t-1}, \mathbf{x}_{1:t-1}\right) \leq \frac{\delta}{2T}.$$

Since the above bound holds for any set of $\{g_t, f_{1:t}, \mathbf{x}_{1:t-1}\}$, the same tail bound also holds without the conditioning. This shows that

$$\mathbb{P}\left(\|\mathbf{x}_t - \mathbf{x}_t^\infty\|^{1+\alpha} \geq \left(\frac{c\Psi_1^2\Psi_2^2D^2\log 4T/\delta}{m}\right)^{\frac{1+\alpha}{2}}\right) \leq \frac{\delta}{2T},$$

where we converted back to $\|\cdot\|$ by introducing the norm compatibility constant Ψ_1.

Bounding the regret. Plugging the above high probability bounds for S_1, S_2 in the previous regret bound and using union bound, we get the following regret bound which holds with probability at least $1-\delta$

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \leq cGD\sqrt{\frac{T\log 2/\delta}{m}} + cLT\left(\frac{\Psi_1^2\Psi_2^2D^2\log 4T/\delta}{m}\right)^{\frac{1+\alpha}{2}} + \eta D$$
$$+ \sum_{t=1}^T \frac{C}{2\eta}\|\nabla_t - g_t\|_*^2 - \sum_{t=1}^T \frac{\eta}{2C}\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|^2,$$

where $c > 0$ is a universal constant. □

B.7.2 Convex-Concave Games

In this section, we present a high probability version of Theorem 7.

Theorem 30. *Consider the minimax game in Equation (3.1). Suppose both the domains $\mathcal{X}, \mathcal{Y}$ are compact subsets of $\mathbb{R}^d$, with diameter $D = \max\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}} \|\mathbf{x}_1 - \mathbf{x}_2\|_2, \sup_{\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}} \|\mathbf{y}_1 - \mathbf{y}_2\|_2\}$. Suppose f is convex in $\mathbf{x}$, concave in $\mathbf{y}$ and is Lipschitz w.r.t $\|\cdot\|_2$ and satisfies*

$$\max\left\{\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|_2, \sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|_2\right\} \leq G.$$

Moreover, suppose f is smooth w.r.t $\|\cdot\|_2$

$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}',\mathbf{y}')\|_2 + \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}',\mathbf{y}')\|_2 \leq L\|\mathbf{x}-\mathbf{x}'\|_2 + L\|\mathbf{y}-\mathbf{y}'\|_2.$$

Suppose Algorithm 13 is used to solve the minimax game. Suppose the perturbation distributions used by both the players are the same and equal to the uniform distribution over $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq (1+d^{-1})\eta\}$. Suppose the guesses used by $\mathbf{x},\mathbf{y}$ players in the t^{th} iteration are $\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}), \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1})$, where $\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{y}}_{t-1}$ denote the predictions of $\mathbf{x},\mathbf{y}$ players in the t^{th} iteration, if guess $g_t = 0$ was used. If Algorithm 13 is run with $\eta = 6dD(L+1), m = T$, then the iterates $\{(\mathbf{x}_t,\mathbf{y}_t)\}_{t=1}^T$ satisfy the following bound with probability at least $1-\delta$

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \left[f\left(\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t, \mathbf{y}\right) - f\left(\mathbf{x}, \frac{1}{T}\sum_{t=1}^T \mathbf{y}_t\right)\right] = O\left(\frac{GD\sqrt{\log\frac{8}{\delta}}}{T} + \frac{D^2(L+1)\left(d + \log\frac{16T}{\delta}\right)}{T}\right).$$

Proof. We use the same notation and proof technique as Theorems 27, 7. From Theorem 1 we know that the predictions of OFTPL are $dD\eta^{-1}$ stable w.r.t $\|\cdot\|_2$, for the particular perturbation distribution we consider here. We use this stability bound in our proof. From Theorem 29, we have the following regret bound for both the players, which holds with probability at least $1-\delta/2$

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}} \left[\sum_{t=1}^T f(\mathbf{x}_t,\mathbf{y}_t) - f(\mathbf{x},\mathbf{y}_t)\right] &\leq cGD\sqrt{\frac{T\log 8/\delta}{m}} + cLT\left(\frac{D^2\log 16T/\delta}{m}\right) + \eta D \\ &\quad + \frac{dD}{2\eta}\sum_{t=1}^T \left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})\|_2^2\right] \\ &\quad - \frac{\eta}{2dD}\sum_{t=1}^T \left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|_2^2\right].\end{aligned}$$

$$\begin{aligned}\sup_{\mathbf{y}\in\mathcal{Y}}\left[\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y})-f(\mathbf{x}_t,\mathbf{y}_t)\right] \leq cGD\sqrt{\frac{T\log 8/\delta}{m}} + cLT\left(\frac{D^2\log 16T/\delta}{m}\right) + \eta D\\ + \frac{dD}{2\eta}\sum_{t=1}^{T}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})\|_2^2\right]\\ - \frac{\eta}{2dD}\sum_{t=1}^{T}\left[\|\mathbf{y}_t^\infty-\tilde{\mathbf{y}}_{t-1}^\infty\|_2^2\right].\end{aligned}$$

First, consider the regret of the $\mathbf{x}$ player. From the proof of Theorem 27, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1},\tilde{\mathbf{y}}_{t-1})\|_2^2 \leq 5L^2\|\mathbf{x}_t-\mathbf{x}_t^\infty\|_2^2 + 5L^2\|\tilde{\mathbf{x}}_{t-1}-\tilde{\mathbf{x}}_{t-1}^\infty\|_2^2\\ + 5L^2\|\mathbf{y}_t-\mathbf{y}_t^\infty\|_2^2 + 5L^2\|\tilde{\mathbf{y}}_{t-1}-\tilde{\mathbf{y}}_{t-1}^\infty\|_2^2\\ + 5\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty)-\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2.\end{aligned}$$

Moreover, from the proof of Theorem 29, we know that $\|\mathbf{x}_t-\mathbf{x}_t^\infty\|_2^2$ satisfies the following tail bound

$$\mathbb{P}\left(\|\mathbf{x}_t-\mathbf{x}_t^\infty\|_2^2 \geq \frac{cD^2\log 16T/\delta}{m}\right) \leq \frac{\delta}{8T}.$$

Similar bounds hold for the quantities appearing in the regret bound of $\mathbf{y}$ player. Plugging this in the previous regret bounds, we get the following which hold with probability at least $1-\delta$

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}}\left[\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}_t)-f(\mathbf{x},\mathbf{y}_t)\right] \leq cGD\sqrt{\frac{T\log 8/\delta}{m}} + \left(L+\frac{10dDL^2}{\eta}\right)\left(\frac{cD^2\log 16T/\delta}{m}\right)T\\ + \eta D + \frac{5dD}{2\eta}\sum_{t=1}^{T}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty)-\nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\right]\\ - \frac{\eta}{2dD}\sum_{t=1}^{T}\left[\|\mathbf{x}_t^\infty-\tilde{\mathbf{x}}_{t-1}^\infty\|_2^2\right].\end{aligned}$$

$$\begin{aligned}\sup_{\mathbf{y}\in\mathcal{Y}}\left[\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y})-f(\mathbf{x}_t,\mathbf{y}_t)\right] \leq cGD\sqrt{\frac{T\log 8/\delta}{m}} + \left(L+\frac{10dDL^2}{\eta}\right)\left(\frac{cD^2\log 16T/\delta}{m}\right)T\\ + \eta D + \frac{5dD}{2\eta}\sum_{t=1}^{T}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty)-\nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\right]\\ - \frac{\eta}{2dD}\sum_{t=1}^{T}\left[\|\mathbf{y}_t^\infty-\tilde{\mathbf{y}}_{t-1}^\infty\|_2^2\right].\end{aligned}$$

Summing these two regret bounds, we get

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\left[\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}) - f(\mathbf{x},\mathbf{y}_t)\right] &\leq 2cGD\sqrt{\frac{T\log 8/\delta}{m}} + \left(L + \frac{10dDL^2}{\eta}\right)\left(\frac{2cD^2\log 16T/\delta}{m}\right)T + 2\eta D\\
&+ \frac{10dD}{2\eta}\sum_{t=1}^{T}\left[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\right]\\
&+ \frac{10dD}{2\eta}\sum_{t=1}^{T}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\right]\\
&- \frac{\eta}{2dD}\sum_{t=1}^{T}\left[\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|_2^2 + \|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|_2^2\right].
\end{aligned}$$

From Holder's smoothness assumption on f, we have

$$\begin{aligned}
\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2 &\leq 2\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\\
&\quad + 2\|\nabla_{\mathbf{x}} f(\mathbf{x}_t^\infty,\tilde{\mathbf{y}}_{t-1}^\infty) - \nabla_{\mathbf{x}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2\\
&\leq 2L^2\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|_2^2 + 2L^2\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|_2^2,
\end{aligned}$$

Using a similar argument, we get

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_t^\infty,\mathbf{y}_t^\infty) - \nabla_{\mathbf{y}} f(\tilde{\mathbf{x}}_{t-1}^\infty,\tilde{\mathbf{y}}_{t-1}^\infty)\|_2^2 \leq 2L^2\|\mathbf{x}_t^\infty - \tilde{\mathbf{x}}_{t-1}^\infty\|_2^2 + 2L^2\|\mathbf{y}_t^\infty - \tilde{\mathbf{y}}_{t-1}^\infty\|_2^2.$$

Plugging this in the previous bound, and setting $\eta = 6dD(L+1), m = T$, we get the following bound which holds with probability at least $1-\delta$

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\left[\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}) - f(\mathbf{x},\mathbf{y}_t)\right] \leq O\left(GD\sqrt{\log\frac{8}{\delta}} + D^2(L+1)\left(d + \log\frac{16T}{\delta}\right)\right).$$

□

B.7.3 Nonconvex-Nonconcave Games

In this section, we present a high probability version of Theorem 8.

Theorem 31. *Consider the minimax game in Equation (3.1). Suppose the domains $\mathcal{X},\mathcal{Y}$ are compact subsets of $\mathbb{R}^d$ with diameter $D = \max\{\sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}}\|\mathbf{x}_1-\mathbf{x}_2\|_1, \sup_{\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}}\|\mathbf{y}_1-\mathbf{y}_2\|_1\}$. Suppose f is Lipschitz w.r.t $\|\cdot\|_1$ and satisfies*

$$\max\left\{\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|_\infty, \sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|_\infty\right\} \leq G.$$

Moreover, suppose f satisfies the following smoothness property

$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}',\mathbf{y}')\|_\infty + \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}',\mathbf{y}')\|_\infty \leq L\|\mathbf{x}-\mathbf{x}'\|_1 + L\|\mathbf{y}-\mathbf{y}'\|_1.$$

Suppose both $\mathbf{x}$ *and* $\mathbf{y}$ *players use Algorithm 14 to solve the game with linear perturbation functions* $\sigma(\mathbf{z}) = \langle \bar{\sigma}, \mathbf{z} \rangle$*, where* $\bar{\sigma} \in \mathbb{R}^d$ *is such that each of its entries is sampled independently from Exp*(η)*. Suppose the guesses used by* $\mathbf{x}$ *and* $\mathbf{y}$ *players in the* t^{th} *iteration are* $f(\cdot, \tilde{Q}_{t-1}), f(\tilde{P}_{t-1}, \cdot)$*, where* $\tilde{P}_{t-1}, \tilde{Q}_{t-1}$ *denote the predictions of* $\mathbf{x}, \mathbf{y}$ *players in the* t^{th} *iteration, if guess* $g_t = 0$ *was used. If Algorithm 14 is run with* $\eta = 10d^2D(L+1), m = T$*, then the iterates* $\{(P_t, Q_t)\}_{t=1}^T$ *satisfy the following with probability at least* $1-\delta$

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \sum_{t=1}^T f(P_t,\mathbf{y}) - f(\mathbf{x},Q_t) = O\left(\frac{d^2D^2(L+1)\log d}{T} + \frac{GD}{T}\sqrt{\log\frac{8}{\delta}}\right) + O\left(\min\left\{D^2L, \frac{d^2G^2\log T + dG^2\log\frac{8}{\delta}}{LT}\right\}\right).$$

Proof. We use the same notation used in the proofs of Theorems 6, 28. Let $\mathcal{F}, \mathcal{F}'$ be the set of Lipschitz functions over $\mathcal{X}, \mathcal{Y}$, and $\|g_1\|_{\mathcal{F}}, \|g_2\|_{\mathcal{F}'}$ be the Lipschitz constants of functions $g_1 : \mathcal{X} \to \mathbb{R}$, $g_2 : \mathcal{Y} \to \mathbb{R}$ w.r.t $\|\cdot\|_1$. Recall, in Corollary 2 we showed that for our choice of perturbation distribution, $\mathbb{E}_\sigma\left[\|\sigma\|_{\mathcal{F}}\right] = \eta \log d$ and OFTPL is $O\left(d^2D\eta^{-1}\right)$ stable. We use this in our proof.

From Theorem 6, we know that the regret of $\mathbf{x}, \mathbf{y}$ players satisfy

$$\begin{aligned}\sum_{t=1}^T f(P_t,Q_t) - f(\mathbf{x},Q_t) \leq \eta D\log d &+ \underbrace{\sum_{t=1}^T \langle P_t - P_t^\infty, f(\cdot,Q_t)\rangle}_{S_1} \\ &+ \sum_{t=1}^T \frac{cd^2D}{2\eta} \underbrace{\|f(\cdot,Q_t) - f(\cdot,\tilde{Q}_{t-1})\|_{\mathcal{F}}^2}_{S_2} \\ &- \sum_{t=1}^T \frac{\eta}{2cd^2D}\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^T f(P_t,\mathbf{y}) - f(P_t,Q_t) \leq \eta D\log d &+ \sum_{t=1}^T \langle Q_t - Q_t^\infty, f(P_t,\cdot)\rangle \\ &+ \sum_{t=1}^T \frac{cd^2D}{2\eta}\|f(P_t,\cdot) - f(\tilde{P}_{t-1},\cdot)\|_{\mathcal{F}'}^2 \\ &- \sum_{t=1}^T \frac{\eta}{2cd^2D}\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty)^2,\end{aligned}$$

where $c > 0$ is a positive constant. We now provide high probability bounds for S_1, S_2.

Bounding S_1**.** Let $\xi_i = \{\tilde{P}_i, \tilde{Q}_i, P_i, Q_{i+1}\}$ with $\xi_0 = \{Q_1\}$ and let $\xi_{0:t}$ denote the union of sets $\xi_0, \ldots, \xi_t$. Let $\zeta_t = \langle P_t - P_t^\infty, f(\cdot, Q_t)\rangle$ with $\zeta_0 = 0$. Note that $\{\zeta_t\}_{t=0}^T$ is a martingale difference sequence w.r.t $\xi_{0:T}$. This is because $\mathbb{E}\left[P_t|\xi_{0:t-1}\right] = P_t^\infty$ and $f(\cdot, Q_t)$ is a deterministic quantity conditioned on $\xi_{0:t-1}$. As a result $\mathbb{E}\left[\zeta_t|\xi_{0:t-1}\right] = 0$. Moreover, conditioned on $\xi_{0:t-1}$, ζ_t is the average of m independent mean 0 random variables, each of which is bounded by $2GD$. Using Proposition 14, we get

$$\mathbb{P}\left(|\zeta_t| \geq s\middle|\xi_{0:t-1}\right) \leq 2\exp\left(-\frac{ms^2}{4G^2D^2}\right).$$

Using Proposition 42 on the martingale difference sequence $\{\zeta_t\}_{t=0}^T$, we get

$$\mathbb{P}\left(\left|\sum_{t=1}^T \zeta_t\right| \geq s\right) \leq 2\exp\left(-c\frac{ms^2}{G^2D^2T}\right),$$

where $c > 0$ is a universal constant. This shows that with probability at least $1 - \delta/8$, S_1 is upper bounded by $O\left(\sqrt{\frac{G^2D^2T\log\frac{8}{\delta}}{m}}\right)$.

Bounding S_2. We upper bound S_2 as

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2 \leq\ & 3\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}^2\\ &+ 3\|f(\cdot, Q_t^\infty) - f(\cdot, \tilde{Q}_{t-1}^\infty)\|_{\mathcal{F}}^2\\ &+ 3\|f(\cdot, \tilde{Q}_{t-1}^\infty) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2.\end{aligned}$$

We first provide a high probability bound for $\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}^2$. A trivial bound for this quantity is L^2D^2, which can be obtained as follows

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}} &= \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty)\|_\infty\\ &= \|\mathbb{E}_{\mathbf{y}_1\sim Q_t, \mathbf{y}_2\sim Q_t^\infty}\left[\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_2)\right]\|_\infty\\ &\overset{(a)}{\leq} LD,\end{aligned}$$

where (a) follows from the smoothness assumption on f and the fact that the diameter of $\mathcal{X}$ is D. A better bound for this quantity can be obtained as follows. From proof of Theorem 28, we have

$$\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}^2 \leq 2\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty)\|_\infty^2 + 8L^2\epsilon^2.$$

where $\mathcal{N}_\epsilon$ be the ϵ-net of $\mathcal{X}$ w.r.t $\|\cdot\|$. Recall, in the proof of Theorem 28, we showed the following high probability bound for the RHS quantity

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty)\|_2^2 > \frac{4dG^2}{m}(d + 2\sqrt{ds} + 2s)\right) \leq e^{-s+d\log(1+2D/\epsilon)}.$$

Choosing $\epsilon = Dm^{-1/2}, s = \log\frac{8}{\delta} + d\log(1+2m^{1/2})$, we get the following bound for $\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty)\|_2^2$ which holds with probability at least $1-\delta/8$

$$\sup_{\mathbf{x}\in\mathcal{N}_\epsilon} \|\nabla_{\mathbf{x}} f(\mathbf{x}, Q_t) - \nabla_{\mathbf{x}} f(\mathbf{x}, Q_t^\infty)\|_2^2 \leq \frac{20dG^2}{m}\left(\log\frac{8}{\delta} + d\log(1+2m^{1/2})\right).$$

Together with our trivial bound of D^2L^2, this gives us the following bound for $\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}^2$, which holds with probability at least $1-\delta/8$

$$\|f(\cdot, Q_t) - f(\cdot, Q_t^\infty)\|_{\mathcal{F}}^2 \leq \min\left(\frac{20dG^2}{m}\left(\log\frac{8}{\delta} + d\log(1+2m^{1/2})\right), D^2L^2\right) + \frac{8D^2L^2}{m}.$$

Next, we bound $\|f(\cdot, Q_t^\infty) - f(\cdot, \tilde{Q}_{t-1}^\infty)\|_{\mathcal{F}}^2$. From our smoothness assumption on f, we have

$$\|f(\cdot, Q_t^\infty) - f(\cdot, \tilde{Q}_{t-1}^\infty)\|_{\mathcal{F}} \leq L\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty).$$

Combining the previous two results, we get the following upper bound for S_2 which holds with probability at least $1-\delta/8$

$$\begin{aligned}\|f(\cdot, Q_t) - f(\cdot, \tilde{Q}_{t-1})\|_{\mathcal{F}}^2 \leq 3L^2\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty)^2 + \frac{48D^2L^2}{m} \\ + \min\left(\frac{120dG^2}{m}\left(\log\frac{8}{\delta} + d\log(1+2m^{1/2})\right), 6D^2L^2\right).\end{aligned}$$

Regret bound. Substituting the above bounds for S_1, S_2 in the regret bound for $\mathbf{x}$ player gives us the following bound, which holds with probability at least $1-\delta/2$

$$\begin{aligned}\sum_{t=1}^T f(P_t, Q_t) - f(\mathbf{x}, Q_t) \leq \eta D\log d + O\left(GD\sqrt{\frac{T\log\frac{8}{\delta}}{m}} + \frac{d^2D^3L^2T}{\eta m}\right) \\ + O\left(\min\left(\frac{d^3DG^2T}{\eta m}\left(\log\frac{8}{\delta} + d\log(2m)\right), \frac{d^2D^3L^2T}{\eta}\right)\right) \\ + \sum_{t=1}^T \frac{3cd^2DL^2}{2\eta}\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty)^2 - \sum_{t=1}^T \frac{\eta}{2cd^2D}\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2\end{aligned}$$

Using a similar analysis, we get the following regret bound for the $\mathbf{y}$ player

$$\begin{aligned}\sum_{t=1}^T f(P_t, Q_t) - f(\mathbf{x}, Q_t) \leq \eta D\log d + O\left(GD\sqrt{\frac{T\log\frac{8}{\delta}}{m}} + \frac{d^2D^3L^2T}{\eta m}\right) \\ + O\left(\min\left(\frac{d^3DG^2T}{\eta m}\left(\log\frac{8}{\delta} + d\log(2m)\right), \frac{d^2D^3L^2T}{\eta}\right)\right) \\ + \sum_{t=1}^T \frac{3cd^2DL^2}{2\eta}\gamma_{\mathcal{F}}(P_t^\infty, \tilde{P}_{t-1}^\infty)^2 - \sum_{t=1}^T \frac{\eta}{2cd^2D}\gamma_{\mathcal{F}'}(Q_t^\infty, \tilde{Q}_{t-1}^\infty)^2\end{aligned}$$

Choosing, $\eta = 10d^2D(L+1), m = T$, and adding the above two regret bounds, we get

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \sum_{t=1}^{T} f(P_t,\mathbf{y}) - f(\mathbf{x},Q_t) = O\left(d^2D^2(L+1)\log d + GD\sqrt{\log\frac{8}{\delta}}\right) + O\left(\min\left\{D^2LT, \frac{d^2G^2\log T}{L} + \frac{dG^2\log\frac{8}{\delta}}{L}\right\}\right).$$

□

B.8 Background on Convex Analysis

Fenchel Conjugate. The Fenchel conjugate of a function f is defined as

$$f^*(x^*) = \sup_x \langle x, x^*\rangle - f(x).$$

We now state some useful properties of Fenchel conjugates. These properties can be found in Rockafellar [Roc70].

Theorem 32. *Let f be a proper convex function. The conjugate function f^* is then a closed and proper convex function. Moreover, if f is lower semi-continuous then $f^{**} = f$.*

Theorem 33. *For any proper convex function f and any vector x, the following conditions on a vector x^* are equivalent to each other*

- $x^* \in \partial f(x)$
- $\langle z, x^*\rangle - f(z)$ *achieves its supremum in z at $z = x$*
- $f(x) + f^*(x^*) = \langle x, x^*\rangle$

If $(clf)(x) = f(x)$, the following condition can be added to the list

- $x \in \partial f^*(x^*)$

Theorem 34. *If f is a closed proper convex function, ∂f^* is the inverse of ∂f in the sense of multivalued mappings,* i.e., *$x \in \partial f^*(x^*)$ iff $x^* \in \partial f(x)$.*

Theorem 35. *Let f be a closed proper convex function. Let ∂f be the subdifferential mapping. The effective domain of ∂f, which is the set $dom(\partial f) = \{x|\partial f \neq 0\}$, satisfies*

$$ri(dom(f)) \subseteq dom(\partial f) \subseteq dom(f).$$

The range of ∂f is defined as $range\partial f = \cup\{\partial f(x)|x \in \mathbb{R}^d\}$. The range of ∂f is the effective domain of ∂f^*, so

$$ri(dom(f^*)) \subseteq range\partial f \subseteq dom(f^*).$$

Strong Convexity and Smoothness. We now define strong convexity and strong smoothness and show that these two properties are duals of each other.

Definition B.8.1 (Strong Convexity)**.** A function $f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is β-strongly convex w.r.t a norm $\|\cdot\|$ if for all $x, y \in \mathrm{ri}(\mathrm{dom}(f))$ and $\alpha \in (0,1)$ we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{1}{2}\beta\alpha(1-\alpha)\|x-y\|^2.$$

This definition of strong convexity is equivalent to the following condition on f [see Lemma 13 of Sha07]

$$f(y) \geq f(x) + \langle g, y - x\rangle + \frac{1}{2}\beta\|y - x\|^2, \quad \text{for any } x, y \in \mathrm{ri}(\mathrm{dom}(f)), g \in \partial f(x)$$

Definition B.8.2 (Strong Smoothness)**.** A function $f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is β-strongly smooth w.r.t a norm $\|\cdot\|$ if f is everywhere differentiable and if for all x, y we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\beta\|y - x\|^2.$$

Theorem 36 (Kakade, Shalev-Shwartz, and Tewari [KST09])**.** *Assume that f is a proper closed and convex function. Suppose f is β-strongly smooth w.r.t a norm $\|\cdot\|$. Then its conjugate f^* satisfies the following for all a, x with $u = \nabla f(x)$*

$$f^*(a + u) \geq f^*(u) + \langle x, a\rangle + \frac{1}{2\beta}\|a\|_*^2.$$

Theorem 37 (Kakade, Shalev-Shwartz, and Tewari [KST09])**.** *Assume that f is a closed and convex function. Then f is β-strongly convex w.r.t a norm $\|\cdot\|$ iff f^* is $\frac{1}{\beta}$-strongly smooth w.r.t the dual norm $\|\cdot\|_*$.*

Appendix C

Supplementary Material for Chapter 4

C.1 Proof of Proposition 3

Let $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \langle \mathbf{b}, \mathbf{x}\rangle + c$, for some $A \in \mathbb{R}^{d\times d}, \mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$. The gradient and Hessian of f at $\mathbf{x}$ are given by

$$\nabla f(\mathbf{x}) = \frac{1}{2}(A + A^T)\mathbf{x} + \mathbf{b}, \quad \nabla^2 f(\mathbf{x}) = \frac{1}{2}(A + A^T).$$

Gradient. From the definition of f, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1 f(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)\right] = \frac{1}{2}&\underbrace{\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)^T A(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)\right]}_{T_1} \\ &+ \underbrace{\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\langle \mathbf{b}, \mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2\rangle\right]}_{T_2}.\end{aligned}$$

First consider T_1

$$\begin{aligned}&\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)^T A(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)\right] \\ &= \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\right]\mathbf{x}^T A\mathbf{x} + \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1(\mathbf{v}_1 + \mathbf{v}_2)^T CAC(\mathbf{v}_1 + \mathbf{v}_2)\right] \\ &\quad + \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\mathbf{x}^T AC(\mathbf{v}_1 + \mathbf{v}_2) + C^{-1}\mathbf{v}_1(\mathbf{v}_1 + \mathbf{v}_2)^T CA\mathbf{x}\right].\end{aligned}$$

Since $\mathbf{v}_1, \mathbf{v}_2$ are independent random variables whose distributions are symmetric around origin, it is easy to see that the first two terms in the RHS are 0. So we get

$$\begin{aligned}&\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)^T A(\mathbf{x} + C\mathbf{v}_1 + C\mathbf{v}_2)\right] \\ &\quad = \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\mathbf{x}^T AC(\mathbf{v}_1 + \mathbf{v}_2) + C^{-1}\mathbf{v}_1(\mathbf{v}_1 + \mathbf{v}_2)^T CA\mathbf{x}\right] \\ &\quad = \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\mathbf{x}^T AC\mathbf{v}_1 + C^{-1}\mathbf{v}_1\mathbf{v}_1^T CA\mathbf{x}\right] \\ &\quad = C^{-1}\mathbb{E}_{\mathbf{v}_1\sim\mathbb{S}^{d-1}}\left[\mathbf{v}_1\mathbf{v}_1^T\right] C(A\mathbf{x} + A^T\mathbf{x}) = \frac{1}{d}(A + A^T)\mathbf{x},\end{aligned}$$

where we used the fact that $\mathbb{E}_{\mathbf{v}_1\sim\mathbb{S}^{d-1}}\left[\mathbf{v}_1\mathbf{v}_1^T\right] = \frac{1}{d}I_{d\times d}$. Now consider T_2

$$\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1\langle \mathbf{b}, \mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2\rangle\right] = \mathbb{E}_{\mathbf{v}_1}\left[C^{-1}\mathbf{v}_1\langle \mathbf{b}, C\mathbf{v}_1\rangle\right] = \frac{1}{d}\mathbf{b}.$$

Substituting the above expressions for T_1, T_2 in the first display gives us

$$\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}\mathbf{v}_1 f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right] = \frac{1}{d}\nabla f(\mathbf{x}).$$

Hessian. From the definition of f, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right]\\
&\quad= \frac{1}{2}\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)^TA(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right]\\
&\quad\quad+\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}\langle\mathbf{b},\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2\rangle\right]
\end{aligned}$$

Since $\mathbf{v}_1, \mathbf{v}_2$ are independent random variables whose distributions are symmetric around origin, it is easy to see that the second term in the RHS above is 0. So, consider the first term

$$\begin{aligned}
&\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)^TA(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right]\\
&\quad= \mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}(\mathbf{v}_1+\mathbf{v}_2)^TCAC(\mathbf{v}_1+\mathbf{v}_2)\right]\\
&\quad= 2\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_1^T)CAC(\mathbf{v}_2\mathbf{v}_2^T)C^{-1}+C^{-1}(\mathbf{v}_1\mathbf{v}_2^T)CAC(\mathbf{v}_1\mathbf{v}_2^T)C^{-1}\right],
\end{aligned}$$

where we relied on the fact that odd moments of $\mathbf{v}_1, \mathbf{v}_2$ are zero. Continuing, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)^TA(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right]\\
&\quad= \frac{2}{d^2}(A+A^T),
\end{aligned}$$

where we used the fact that $\mathbb{E}_{\mathbf{v}_1\sim\mathbb{S}^{d-1}}\left[\mathbf{v}_1\mathbf{v}_1^T\right] = \frac{1}{d}I$ and $\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[(\mathbf{v}_1\mathbf{v}_2^T)W(\mathbf{v}_1\mathbf{v}_2^T)\right] = \frac{1}{d^2}W^T$. Substituting this in the first display gives us

$$\mathbb{E}_{\mathbf{v}_1,\mathbf{v}_2\sim\mathbb{S}^{d-1}}\left[C^{-1}(\mathbf{v}_1\mathbf{v}_2^T+\mathbf{v}_2\mathbf{v}_1^T)C^{-1}f(\mathbf{x}+C\mathbf{v}_1+C\mathbf{v}_2)\right] = \frac{2}{d^2}\nabla^2 f(\mathbf{x}).$$

C.2 Proof of Proposition 4

This proposition was proved in Nemirovski [Nem04]. For the sake of completeness, we reproduce the proof here. Let $\mathbf{h} = \mathbf{y}-\mathbf{x}$ and $r = \|\mathbf{h}\|_{\nabla^2 R(\mathbf{x})}$. Let $\phi(t) = \nabla^2 R(\mathbf{x}+t\mathbf{h})[\mathbf{h},\mathbf{h}]$. The function ϕ satisfies the following properties

$$0 \leq \phi(t),\ r^2 = \phi(0),\ |\phi'(t)| = |\nabla^3 R(\mathbf{x}+t\mathbf{h})[\mathbf{h},\mathbf{h},\mathbf{h}]| \leq 2\phi^{3/2}(t).$$

So, for all positive ϵ, we have

$$0 < \phi_\epsilon(t) = \epsilon + \phi(t),\ |\phi_\epsilon'(t)| \leq 2\phi_\epsilon^{3/2}(t).$$

Continuing,

$$\left|\frac{d}{dt}\phi_\epsilon^{-1/2}(t)\right| \le 1.$$

It follows that

$$\phi_\epsilon^{-1/2}(t) \le \phi_\epsilon^{-1/2}(0) + t.$$

This gives us

$$\frac{\phi_\epsilon(0)}{(1 + t\phi_\epsilon^{1/2}(0))^2} \le \phi_\epsilon(t).$$

The above inequality holds for any $t \in [0, 1]$ and any $\epsilon > 0$. Passing to limit as $\epsilon \to 0+$, we get

$$\frac{r^2}{(1 + rt)^2} \le \phi(t) = \nabla^2 R(\mathbf{x} + t\mathbf{h})[\mathbf{h}, \mathbf{h}]$$

Setting $t = 1$, we get $\nabla^2 R(\mathbf{y}) \ge \frac{1}{(1+r)^2}\nabla^2 R(\mathbf{x})$. Using the fact that $r \le \lambda$ gives us the required result.

C.3 Proof of Proposition 5

Let $\mathcal{X} = \{\mathbf{x} : \langle \mathbf{a}_i, \mathbf{x}\rangle \ge b_i, \text{ for } i = 1, \dots m\}$. Consider the logarithmic barrier for $\mathcal{X}$

$$R(\mathbf{x}) = -\sum_i \log(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i).$$

It is well know that $R(\mathbf{x})$ is a m-self concordant barrier for $\mathcal{X}$ [Nem04]. The Hessian of R is given by

$$\nabla^2 R(\mathbf{x}) = \sum_i \frac{\mathbf{a}_i\mathbf{a}_i^T}{(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i)^2}.$$

Since $\|\mathbf{y} - \mathbf{x}\|_{\nabla^2 R(\mathbf{x})} \le \lambda$, we have

$$\begin{aligned}
&\sum_i \frac{\langle \mathbf{a}_i, \mathbf{y} - \mathbf{x}\rangle^2}{(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i)^2} \le \lambda^2 \\
&\implies \forall i, \frac{\langle \mathbf{a}_i, \mathbf{y} - \mathbf{x}\rangle^2}{(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i)^2} \le \lambda^2 \\
&\implies \forall i, \langle \mathbf{a}_i, \mathbf{y} - \mathbf{x}\rangle \le \lambda(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i),
\end{aligned}$$

where we used the fact that $\mathbf{x} \in \mathcal{X}$ and hence $\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \ge 0$ in the last step. This then implies that

$$\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i \le (1 + \lambda)(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i).$$

Since $\mathbf{y} \in \mathcal{X}$ and hence $\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i \ge 0$, we have $(\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i)^2 \le (1+\lambda)^2(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i)^2$. So, we have

$$\nabla^2 R(\mathbf{y}) = \sum_i \frac{\mathbf{a}_i\mathbf{a}_i^T}{(\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i)^2} \succeq \frac{1}{(1+\lambda)^2}\sum_i \frac{\mathbf{a}_i\mathbf{a}_i^T}{(\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i)^2} = \frac{1}{(1+\lambda)^2}\nabla^2 R(\mathbf{x}).$$

This finishes the proof of the Proposition.

C.4 Warm up: Hypothetical case of known Hessians

In this section, we consider a hypothetical scenario where we are given access to the Hessian H_t of loss function f_t at the beginning of iteration t. In such a scenario, instead of estimating the Hessian from single point feedback (as done in Algorithm 4), one can rely on H_t. In this section, we study such an algorithm; that is, we study a variant of Algorithm 4 where we replace the Hessian estimate $\hat{H}_t$ with H_t.

Studying this hypothetical scenario helps the readers understand the intuition behind Algorithm 4. Moreover, it greatly simplifies our proofs and makes it easier to understand the key ideas in the proof of Theorem 9. Finally, this hypothetical scenario encompasses the important special case of linear loss functions (*i.e.*, $H_t = 0$) that is often studied in the literature of bandit optimization [AHR09].

The following Theorem bounds the regret of this hypothetical algorithm. To further simplify the analysis, we assume the loss functions are exactly quadratic (*i.e.*, $\epsilon = 0$).

Theorem 38 (Approximately quadratic losses)**.** *Suppose f_t is a convex, quadratic function $f_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A_t \mathbf{x} + \langle \mathbf{b}_t, \mathbf{x}\rangle + c_t$. Let R be a ν-self-concordant barrier of $\mathcal{X}$ that satisfies Assumption 1. Suppose the diameter of $\mathcal{X}$ is bounded by T, and the Lipschitz constants of $\{f_t\}_{t=1}^T$ are bounded by T. Suppose Algorithm 4 is run for T iterations with $\hat{H}_t = \frac{1}{2}(A_t + A_t^T)$ and the following hyper-parameters*

$$\lambda = \frac{1}{4}, \;\; \alpha = c_1(\nu + d)d\log^2 dT, \;\; \beta = 4d\log dT, \;\; \gamma = \frac{c_2}{d\log T}, \;\; \eta_1 = \frac{c_3}{d^{2.5}B\alpha\sqrt{T}\log T},$$

for some universal constants $c_1, c_2, c_3 > 0$. Let $\mathcal{T}$ be the minimum between T and the first time at which the algorithm restarts. Then with probability at least $1-\delta$

$$\sum_{t=1}^{\mathcal{T}} f_t(\mathbf{y}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{\mathcal{T}} f_t(\mathbf{x}) \le \begin{cases} \tilde{O}\left(d^{3.5}(d+\nu)^2\sqrt{T}\right) & \textit{if } \mathcal{T} = T \\ 0 \quad \textit{otherwise} \end{cases}.$$

Remark C.4.1 (Linear losses)**.** *The above regret bound can be improved to $\tilde{O}\left(d^{3.5}\nu^2\sqrt{T}\right)$ for linear loss functions. This is because for linear losses, we can obtain a tighter bound for $\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - f_s(\mathbf{x}_s)$ than the one we obtained for general quadratic functions in the proof of Theorem 38 (see Equation C.4 below).*

Remark C.4.2 (Convex losses)**.** *The above Theorem can be generalized in a straightforward way to general convex loss functions. Suppose f_t's are general convex loss functions and suppose we have access to a lower bound for of $\nabla^2 f_t$'s. In particular, suppose at the beginning iteration t, we have access to H_t which satisfies: $\forall \mathbf{x} \in \mathcal{X}, H_t \preceq \nabla^2 f_t(\mathbf{x})$. Suppose we run Algorithm 4 with $\hat{H}_t = H_t$. Then we can use similar proof techniques as in Theorem 38 to obtain regret bounds. There are two special cases of particular interest here.*

1. *(**Strongly convex and smooth**) Suppose f_t's are strongly convex and smooth and we have access to the strong convex parameter of f_t (say κ_t) at each iteration t. Suppose Algorithm 4 is run with $\hat{H}_t = \kappa_t I$. Then its regret is $\tilde{O}\left(d^{3.5}\nu^2\sqrt{T}\right)$.*
2. *(**Smooth**) Suppose f_t's are smooth and we run Algorithm 4 with $\hat{H}_t = 0$. Then its regret can be bounded by $\tilde{O}\left(d^{7/3}T^{2/3}\right)$.*

Before we present a proof of this Theorem, we present some useful intermediate results.

C.4.1 Intermediate Results

Lemma 39 (Initial focus region). *For any $\alpha \geq \nu + 2\sqrt{\nu}$,*

$$F_1 \subseteq \mathcal{X} \subseteq B_{\alpha,\nabla^2 R(\mathbf{x}_1)}(\mathbf{x}_1).$$

Proof. Consider property (P4) of self-concordant barriers stated in Equation (C.19) of Appendix C.7. It says that for any $\mathbf{x} \in \text{int}(\mathcal{X})$

$$\mathcal{X} \cap \{\mathbf{y} : \langle \nabla R(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0\} \subseteq B_{\nu+2\sqrt{\nu},\nabla^2 R(\mathbf{x})}(\mathbf{x}).$$

Since $\mathbf{x}_1$ is the minimizer of $R(\mathbf{x})$ over $\mathcal{X}$, and since it is in the interior of $\mathcal{X}$, we have $\nabla R(\mathbf{x}_1) = 0$. So, from property (P4) we have $\mathcal{X} \subseteq B_{\nu+2\sqrt{\nu},\nabla^2 R(\mathbf{x}_1)}(\mathbf{x}_1)$. The lemma then immediately follows from the definition of F_1 (recall $F_1 = \mathcal{X}_\xi \subseteq \mathcal{X}$). □

Lemma 40 (Lemma 5 of Bubeck, Lee, and Eldan [BLE17]). *Let $\mathcal{K}$ be a convex body and $\mathcal{E}$ be an ellipsoid centered at the origin. Suppose that Vol$(\mathcal{K} \cap \mathcal{E}) \geq \frac{1}{2}Vol(\mathcal{K})$. Then $\mathcal{K} \subset 4d\mathcal{E}$.*

Lemma 41 (Lemma 4.6 of Hazan [Haz16]). *Let B_0 be a symmetric positive definite matrix and let $\{B_t\}_{t=1}^T$ be symmetric positive semi-definite matrices. Let $A_t = \sum_{s=0}^t B_s$. Then*

$$\sum_{t=1}^{T} tr(A_t(A_t - A_{t-1})) \leq \log_2 \frac{\det A_T}{\det A_0}.$$

Lemma 42 (Wainwright [Wai19]). *Let $X_1, \ldots X_K \in \mathbb{R}$ be a martingale difference sequence, where $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$. Assume that X_i satisfy the following tail condition, for some scalar $B_i > 0$*

$$\mathbb{P}\left(\left|\frac{X_i}{B_i}\right| \geq z \middle| \mathcal{F}_{i-1}\right) \leq 2\exp(-z^2).$$

Then

$$\mathbb{P}\left(\left|\sum_{i=1}^{K} X_i\right| \geq z\right) \leq 2\exp\left(-c\frac{z^2}{\sum_{i=1}^K B_i^2}\right),$$

where $c > 0$ is a universal constant.

Lemma 43 (Matrix Azuma; Tropp [Tro12]). *Consider a finite adapted sequence $\{X_i\}$ of symmetric matrices in dimension d, and fixed sequence $\{A_i\}$ of symmetric matrices that satisfy*

$$\mathbb{E}_i[X_i] = 0 \textit{ and } X_i^2 \preceq A_i^2 \textit{ almost surely.}$$

Compute the variance parameter $\sigma^2 := \|\sum_i A_i^2\|_2$. Then, for all $t \geq 0$,

$$\mathbb{P}\left(\lambda_{max}\left(\sum_i X_i\right) \geq t\right) \leq de^{-t^2/8\sigma^2}.$$

C.4.2 Proof of Theorem 38

To prove Theorem 38, we work with a slightly modified algorithm and show that with high probability, the iterates of the modified algorithm are exactly same as the actual algorithm. Consequently, proving the proposition for the modified algorithm entails that the Theorem also holds for the actual algorithm. In the modified algorithm, we slightly change $\hat{g}_t, \hat{H}_t$ and work with the following sequence of random variables

$$\hat{g}_t = \lambda^{-1} d\iota_t f_t(\mathbf{y}_t) M_t^{1/2}\mathbf{v}_{1,t}, \quad \hat{H}_t = \frac{\iota_t}{2}\left(A_t + A_t^T\right).$$

where ι_t is an indicator random variable which is equal to 1 if and only if the following event happen

$$\sup_{\mathbf{x}\in F_t}\left|\sum_{s=1}^{t-1}(\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - \iota_s f_s(\mathbf{x}) + \iota_s f_s(\mathbf{x}_s))\right| \leq \frac{1}{\eta_1}.$$

This event happens when the cumulative loss estimate $\sum_{s=1}^{t-1}\hat{f}_s(\mathbf{x})$ is close to the true cumulative loss $\sum_{s=1}^{t-1} f_s(\mathbf{x})$ over the focus region F_t. We assume the algorithm is run with these modified estimates of gradients and Hessians. The main benefit of working with the modified gradient and Hessian estimates is that they are more amenable to analysis. Our proof shows that with high probability, the modified random variables $\hat{g}_t, \hat{H}_t$ are exactly equal to the original definitions of $\hat{g}_t, \hat{H}_t$. In particular, we show that in every iteration before the algorithm restarts, $\iota_t = 1$ with high probability. This entails that the actions output by the modified algorithm are exactly same as the actual algorithm, with high probability. As a result, it suffices to prove Theorem 38 for the modified algorithm.

We now derive some useful properties of the iterates produced by the modified algorithm. Some of these properties are very basic and pertain to the well-behavedness of the iterates of the algorithm. For example, the first property ensures that $\mathbf{y}_t$ always lies in $\mathcal{X}$.

Lemma 44 (Properties of iterates)**.** *Consider the setting of Theorem 38. Let $\mathcal{T}$ be the minimum between T and the first iteration at which the modified algorithm restarts. For any $t < \mathcal{T}$ such that $\eta_t \leq 10\eta_1$, the iterates of the algorithm satisfy the following stability properties*

1. *M_t is positive definite and $\mathbf{y}_t \in \mathcal{X}$.*
2. *$R_t(\mathbf{x})$ is a strictly convex function over F_t.*
3. *For all $\mathbf{x} \in F_t$, $\|\mathbf{x} - \mathbf{x}_t\|_{M_t} \leq 4d\alpha$ and $\nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+4d\alpha)^2}\nabla^2 R(\mathbf{x}_t)$.*
4. *$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} \leq 2\lambda^{-1} dB\eta_t$ and $\|I - M_t^{-1/2} M_{t+1} M_t^{-1/2}\|_2 \leq 12\lambda^{-2} d^2 B\eta_t$.*
5. *if $\iota_t = 0$, then $\iota_t = \iota_{t+1} = \cdots = \iota_{\mathcal{T}}$, $\mathbf{x}_t = \mathbf{x}_{t+1} \cdots = \mathbf{x}_{\mathcal{T}}$ and $F_t = F_{t+1} \cdots = F_{\mathcal{T}}$.*

Proof. We use induction to prove the lemma.

Base Case (t=1).

1. First note that $M_1 = \nabla^2 R(\mathbf{x}_1)$. From property P3 of SCB stated in Appendix C.7, we know that $R(\mathbf{x})$ is strictly convex over $\text{int}(\mathcal{X})$. So M_1 is positive definite and invertible. Moreover, from the Dikin ellipsoid property (P1) of SCB stated in Section 4.1, and from our choice of λ, it is easy to see that $\mathbf{y}_1 \in \mathcal{X}$.

2. The strict convexity property of $R(\mathbf{x})$ over F_1 follows from property P3 of SCB stated in Appendix C.7.
3. To show that for all $\mathbf{x} \in F_1$, $\nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+4d\alpha)^2}\nabla^2 R(\mathbf{x}_1)$, we rely on Assumption 1 and Lemma 39. In particular, from Assumption 1 we know that if $\|\mathbf{x}-\mathbf{x}_1\|_{\nabla^2 R(\mathbf{x}_1)} \leq \lambda$, then $\nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+\lambda)^2}\nabla^2 R(\mathbf{x}_1)$. Moreover, from Lemma 39 we know that any $\mathbf{x} \in \mathcal{X}$ satisfies

$$\|\mathbf{x}-\mathbf{x}_1\|_{\nabla^2 R(\mathbf{x}_1)} \leq \nu + 2\sqrt{\nu} \leq \alpha.$$

   Combining these two facts gives us the required result.
4. We now show that $\mathbf{x}_2$ and $\mathbf{x}_1$ are close to each other. Note that $\mathbf{x}_2$ is the minimizer of the following objective

$$\mathbf{x}_2 \in \operatorname*{argmin}_{\mathbf{x}\in F_1} \eta_1\langle \hat{g}_1, \mathbf{x}\rangle + \Phi_{R_2}(\mathbf{x}, \mathbf{x}_1). \tag{C.1}$$

   From first order optimality conditions we have

$$\forall \mathbf{x} \in F_1, \quad \langle \nabla R_2(\mathbf{x}_2) - \nabla R_2(\mathbf{x}_1) + \eta_1\hat{g}_1, \mathbf{x}-\mathbf{x}_2\rangle \geq 0.$$

   Substituting $\mathbf{x}_1$ in the above equation gives us

$$\langle \nabla R_2(\mathbf{x}_2) - \nabla R_2(\mathbf{x}_1) + \eta_1\hat{g}_1, \mathbf{x}_1-\mathbf{x}_2\rangle \geq 0.$$

   This can equivalently be written as

$$\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1\hat{g}_1 + \eta_1\hat{H}_1(\mathbf{x}_2-\mathbf{x}_1), \mathbf{x}_2-\mathbf{x}_1\rangle \leq 0. \tag{C.2}$$

   Now suppose $\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1} > 2\lambda^{-1}dB\eta_1$. Then we have

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1\hat{g}_1 + \eta_1\hat{H}_1(\mathbf{x}_2-\mathbf{x}_1), \mathbf{x}_2-\mathbf{x}_1\rangle \\
&\overset{(a)}{\geq} \frac{\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}^2}{1+\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}} + \langle \eta_1\hat{g}_1 + \eta_1\hat{H}_1(\mathbf{x}_2-\mathbf{x}_1), \mathbf{x}_2-\mathbf{x}_1\rangle \\
&\overset{(b)}{\geq} \frac{\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}^2}{1+\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}} - \eta_1\|\hat{g}_1\|_{M_1}^*\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1} \\
&= \|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}\left(\frac{\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}}{1+\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}} - \eta_1\|\hat{g}_1\|_{M_1}^*\right),
\end{aligned}$$

   where (a) follows from property P7 of SCBs stated in Appendix C.7 and (b) follows from the fact that $\hat{H}_1$ is a positive semi-definite matrix. Next, consider the following

$$(\|\hat{g}_1\|_{M_1}^*)^2 = \hat{g}_1^T M_1^{-1}\hat{g}_1 = \lambda^{-2}d^2 f_1^2(\mathbf{y}_1)\mathbf{v}_{1,1}^T\mathbf{v}_{1,1} \leq \lambda^{-2}d^2B^2.$$

   Substituting this in the previous inequality and using the fact that $\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1} > 2\lambda^{-1}dB\eta_1$ gives us

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1\hat{g}_1 + \eta_1\hat{H}_1(\mathbf{x}_2-\mathbf{x}_1), \mathbf{x}_2-\mathbf{x}_1\rangle \\
&\geq \lambda^{-1}dB\eta_1\|\mathbf{x}_2-\mathbf{x}_1\|_{M_1}\left(\frac{2}{1+2\lambda^{-1}dB\eta_1} - 1\right) \\
&\overset{(a)}{>} 0,
\end{aligned}$$

where (a) follows from the fact that $\lambda^{-1}dB\eta_1 < 1/2$. This contradicts the first order optimality condition in Equation (C.2). This shows that $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \leq 2\lambda^{-1}dB\eta_1$. Next, we show that $M_1^{-1/2}M_2M_1^{-1/2}$ is close to identity. From the definitions of M_1, M_2, we have

$$M_1^{-1/2}M_2M_1^{-1/2} - I = M_1^{-1/2}(\nabla^2 R(\mathbf{x}_2) - \nabla^2 R(\mathbf{x}_1))M_1^{-1/2} + \eta_1 M_1^{-1/2}\hat{H}_1 M_1^{-1/2}.$$

Since $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \leq 2\lambda^{-1}dB\eta_1$, we can rely on property P2 of SCB stated in Section 4.1 to infer that

$$\nabla^2 R(\mathbf{x}_2) \preceq \frac{1}{(1-2\lambda^{-1}dB\eta_1)^2}\nabla^2 R(\mathbf{x}_1) \preceq (1+6\lambda^{-1}dB\eta_1)\nabla^2 R(\mathbf{x}_1),$$

where the last inequality follows since $\lambda^{-1}dB\eta_1 < 1/10$. Next, note that $\hat{H}_1$ can be written as

$$\hat{H}_1 = \mathbb{E}\left[\frac{\lambda^{-2}}{2}d^2 f_1(\mathbf{y}_1)M_1^{1/2}\left(\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T + \mathbf{v}_{2,1}\mathbf{v}_{1,1}^T\right)M_1^{1/2}\right].$$

So we have $M_1^{-1/2}\hat{H}_1 M_1^{-1/2} = \mathbb{E}\left[\frac{\lambda^{-2}}{2}d^2 f_1(\mathbf{y}_1)\left(\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T + \mathbf{v}_{2,1}\mathbf{v}_{1,1}^T\right)\right]$ which is a bounded quantity. Substituting the previous two bounds in our expression for $M_1^{-1/2}M_2M_1^{-1/2} - I$ we get

$$\|M_1^{-1/2}M_2M_1^{-1/2} - I\|_2 \leq 6\lambda^{-1}dB\eta_1 + \lambda^{-2}d^2B\eta_1$$

5. Note that ι_1 is always equal to 1. So the last property trivially holds. This finishes the proof of the base case.

Induction Step. Suppose the proposition holds for the first $t-1$ iterations. We now show that it also holds for the t^{th} iteration.

1. The first part on positive definiteness of M_t and $\mathbf{y}_t \in \mathcal{X}$ follows from the same arguments as in the base case.
2. Note that $R_t(\mathbf{x}) = R(\mathbf{x}) + \sum_{s=0}^{t-1}\frac{\eta_s}{2}(\mathbf{x}-\mathbf{x}_s)^T\hat{H}_s(\mathbf{x}-\mathbf{x}_s)$. Since $\hat{H}_s$ is positive semi-definite, we have $\nabla^2 R_t(\mathbf{x}) \succeq \nabla^2 R(\mathbf{x})$. The strict convexity of $R_t(\mathbf{x})$ then follows from the fact that $R(\mathbf{x})$ is strictly convex over $\text{int}(\mathcal{X})$.
3. The focus region update condition of our algorithm (lines 21-25 of Algorithm 4) always ensures that

$$\text{Vol}(F_t \cap B_{\alpha,M_t}(\mathbf{x}_t)) \geq \frac{1}{2}\text{Vol}(F_t).$$

   So, from Lemma 40 we know that for any $\mathbf{x} \in F_t$, $\|\mathbf{x} - \mathbf{x}_t\|_{M_t} \leq 4d\alpha$. By relying on Assumption 1 on SCB, we then get

$$\forall \mathbf{x} \in F_t,\ \nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+4d\alpha)^2}\nabla^2 R(\mathbf{x}_t).$$

4. We now prove stability of the iterates. In particular, we show that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} \leq 2\lambda^{-1}dB\eta_t$. If $\iota_{t-1} = 0$, then this trivially holds (because $\mathbf{x}_{t+1} = \mathbf{x}_t$). So lets consider the case where $\iota_{t-1} = 1$. From the first order optimality conditions, we have

$$\forall \mathbf{x} \in F_t, \quad \langle \nabla R_{t+1}(\mathbf{x}_{t+1}) - \nabla R_{t+1}(\mathbf{x}_t) + \eta_t \hat{g}_t, \mathbf{x} - \mathbf{x}_{t+1}\rangle \geq 0. \tag{C.3}$$

Note that from our definition of F_t, F_{t-1} we always have $F_t \subseteq F_{t-1}$ and $\mathbf{x}_t \in F_t$. So substituting $\mathbf{x}_t$ in the above equation and rearranging terms gives us

$$\langle \nabla R(\mathbf{x}_{t+1}) - \nabla R(\mathbf{x}_t) + \eta_t \hat{g}_t + \sum_{s=1}^{t} \eta_s \hat{H}_s(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \leq 0.$$

Now suppose $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} > 2\lambda^{-1}dB\eta_t$. Then we have

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_{t+1}) - \nabla R(\mathbf{x}_t) + \eta_t \hat{g}_t + \sum_{s=1}^{t} \eta_s \hat{H}_s(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \\
&\overset{(a)}{\geq} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{\nabla^2 R(\mathbf{x}_t)}}{1 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 R(\mathbf{x}_t)}} + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{\eta_{1:t}\hat{H}_{1:t}} + \langle \eta_t \hat{g}_t, \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \\
&\overset{(b)}{\geq} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{\nabla^2 R(\mathbf{x}_t)}}{1 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 R(\mathbf{x}_t)}} + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{\eta_{1:t-1}\hat{H}_{1:t-1}} - \eta_t \|\hat{g}_t\|^*_{M_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t},
\end{aligned}$$

where (a) follows from property P7 of SCBs stated in Appendix C.7 and (b) follows from the fact that $\hat{H}_t$ is a positive semi-definite matrix. Here $\eta_{1:t}\hat{H}_{1:t} = \sum_{s=1}^{t} \eta_s \hat{H}_s$. Continuing

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_{t+1}) - \nabla R(\mathbf{x}_t) + \eta_t \hat{g}_t + \sum_{s=1}^{t} \eta_s \hat{H}_s(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \\
&\overset{(b)}{\geq} \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2_{M_t}}{1 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t}} - \eta_t \|\hat{g}_t\|^*_{M_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t},
\end{aligned}$$

Next, consider the following

$$(\|\hat{g}_t\|^*_{M_t})^2 = \hat{g}_t^T M_t^{-1} \hat{g}_t = \lambda^{-2} d^2 f_t^2(\mathbf{y}_t) \mathbf{v}_{1,t}^T \mathbf{v}_{1,t} \leq \lambda^{-2} d^2 B^2.$$

Substituting this in the previous inequality and using the fact that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} > 2\lambda^{-1}dB\eta_t$ gives us

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_{t+1}) - \nabla R(\mathbf{x}_t) + \eta_t \hat{g}_t + \sum_{s=1}^{t} \eta_s \hat{H}_s(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle \\
&\geq \lambda^{-1} dB\eta_t \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} \left(\frac{2}{1 + 2\lambda^{-1}dB\eta_t} - 1 \right) \\
&\overset{(a)}{>} 0,
\end{aligned}$$

where (a) follows from the fact that $\lambda^{-1}dB\eta_t < 1/2$. This contradicts the first order optimality condition in Equation (C.3). This shows that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{M_t} \leq 2\lambda^{-1}dB\eta_t$.
Next, we show that $M_t^{-1/2}M_{t+1}M_t^{-1/2}$ is close to identity. From the definitions of M_t, M_{t+1}, we have

$$M_t^{-1/2}M_{t+1}M_t^{-1/2} - I = M_t^{-1/2}(\nabla^2 R(\mathbf{x}_{t+1}) - \nabla^2 R(\mathbf{x}_t))M_t^{-1/2} + \eta_t M_t^{-1/2}\hat{H}_t M_t^{-1/2}.$$

Using similar arguments as in the base case, we get

$$\nabla^2 R(\mathbf{x}_{t+1}) \preceq (1 + 6\lambda^{-1}dB\eta_t)\nabla^2 R(\mathbf{x}_t),\ M_t^{-1/2}\hat{H}_t M_t^{-1/2} = \mathbb{E}_t\left[\frac{\lambda^{-2}}{2}d^2 f_t(\mathbf{y}_t)\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right)\right].$$

Substituting these quantities in our expression for $M_t^{-1/2}M_{t+1}M_t^{-1/2} - I$ we get

$$\|M_t^{-1/2}M_{t+1}M_t^{-1/2}\|_2 \leq 12\lambda^{-2}d^2B\eta_t.$$

5. The last property that remains to be shown is that if $\iota_t = 0$, then $\iota_t = \iota_{t+1} = \cdots = \iota_{\mathcal{T}}$, $\mathbf{x}_t = \mathbf{x}_{t+1} \cdots = \mathbf{x}_{\mathcal{T}}$ and $F_t = F_{t+1} \cdots = F_{\mathcal{T}}$. We assume $\iota_{t-1} = 1$, since otherwise the property is trivially true. Also note that $R_t(\mathbf{x})$ is strictly convex over F_t and so the OMD update in line 19 of Algorithm 4 has a unique minimizer.
When $\iota_t = 0$, we have $\hat{g}_t = 0, \hat{H}_t = 0$. So the OMD update in line 19 of Algorithm 4 is given by $\mathbf{x}_{t+1} = \text{argmin}_{\mathbf{x}\in F_t} \Phi_{R_{t+1}}(\mathbf{x}, \mathbf{x}_t)$. Since $R_{t+1}(\mathbf{x}) = R_t(\mathbf{x})$ and $\mathbf{x}_t \in F_t$, it is easy to see that $\mathbf{x}_{t+1} = \mathbf{x}_t$. So the algorithm wouldn't make any progress in further rounds.

This finishes the proof of the lemma. □

We now show that the focus region doesn't get updated more than $12d\log T$ times. This helps us show that the learning η_t doesn't gets too large.

Lemma 45 (Focus region updates)**.** *Consider the setting of Theorem 38. Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then the focus region gets updated no more than $12d\log T$ times before $\mathcal{T}$. Moreover, $\eta_s \leq 10\eta_1$ for any $s \leq \mathcal{T}$.*

Proof. We prove the proposition using contradiction. Assume that the focus region gets updated more than $12d\log T$ times before the algorithm restarts. Let $\tau < \mathcal{T}$ be the iteration where the focus region update happens for $12d\log T^{th}$ time. We now show that the restart condition should have triggered in iteration τ.

We have the following upper bound on the volume of $F_{\tau+1}$:

$$\text{Vol}(F_{\tau+1}) \leq \text{Vol}(F_\tau) \leq \frac{1}{T^{6d}}\text{Vol}(\mathcal{X}_\xi).$$

This follows from the fact that the volume of the focus region reduces by a factor of 1/2 whenever the focus region update condition triggers. In the rest of the proof, we show that if the volume of focus region is less than $\frac{1}{T^{6d}}\text{Vol}(\mathcal{X}_\xi)$, then the restart condition should have triggered.

Step 1. First of all, for our choice of γ, we have $(1+\gamma)^{12d\log T} \leq 10$. Consequently, $\eta_\tau \leq 10\eta_1$. So the properties of the iterates we proved in Lemma 44 apply to our setting here. From this Lemma, we can infer that $\iota_\tau = 1$. Otherwise, we know that the focus region shouldn't have changed in the τ^{th} iteration (recall, in Lemma 44 we showed that if $\iota_\tau = 0$, then $F_\tau = F_{\tau+1}$). Moreoever, from this Lemma we can infer that $\forall t \leq \tau, \iota_t = 1$. So the cumulative loss estimate is close to the true cumulative loss and satisfies

$$\sup_{\mathbf{x}\in F_\tau} \Big| \sum_{s=1}^{\tau-1} (\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - f_s(\mathbf{x}) + f_s(\mathbf{x}_s)) \Big| \leq \frac{1}{\eta_1}.$$

Step 2. Let $\mathbf{u}_{\tau+1}$ be the minimizer of $\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x})$ over F_τ. Suppose $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right) \cap \mathcal{X}_\xi \subset F_\tau$. Then

$$\text{Vol}(F_\tau) \geq \text{Vol}\left(B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right) \cap \mathcal{X}_\xi\right).$$

Next, from our assumption that $\mathcal{X}$ contains a euclidean ball of radius 1, we can infer that $\mathcal{X}_\xi = \xi\mathbf{x}_1 + (1-\xi)\mathcal{X}$ contains a ball of radius $(1-\xi)$ in it. Let $\tilde{B}$ be the ball of radius $(1-\xi)$ that lies in $\mathcal{X}_\xi$. By convexity of $\mathcal{X}$ and the fact that the diameter of $\mathcal{X}$ is less than or equal to T, we have

$$\left(1 - \frac{1}{T^3}\right)\mathbf{u}_{\tau+1} + \frac{1}{T^3}\tilde{B} \subseteq B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right) \cap \mathcal{X}_\xi.$$

This shows that $\text{Vol}(F_\tau) \geq T^{-4d}\omega_d$, where ω_d is the volume of unit sphere in $\mathbb{R}^d$. Combining this with the previous upper bound on $\text{Vol}(F_\tau)$, we get

$$T^{-4d}\omega_d, \leq \text{Vol}(F_\tau) \leq T^{-6d}\text{Vol}(\mathcal{X}) \overset{(a)}{\leq} T^{-5d}\omega_d,$$

where (a) follows from the fact that the diameter of $\mathcal{X}$ is upper bounded by T. We arrived at a contradiction. This shows that $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right) \cap \mathcal{X}_\xi \not\subset F_\tau$.

Step 3. Since $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right) \cap \mathcal{X}_\xi \not\subset F_\tau$, the following holds: $\exists \mathbf{x} \in \partial F_\tau \cap \text{int}(\mathcal{X}_\xi)$ such that $\|\mathbf{x} - \mathbf{u}_{\tau+1}\|_2 \leq \frac{1}{T^2}$. Now, consider the following for such an $\mathbf{x}$

$$\begin{aligned}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) = &\sum_{s=1}^{\tau} f_s(\mathbf{x}) - f_s(\mathbf{u}_{\tau+1})\\ &+ \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) - f_s(\mathbf{x}) + f_s(\mathbf{u}_{\tau+1}).\end{aligned}$$

Since each f_s is T-Lipschitz, the first term in the RHS above is upper bounded by 1. Since the cumulative loss estimate is close to the true cumulative loss, the second term can be bounded as

$$\begin{aligned}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) - f_s(\mathbf{x}) + f_s(\mathbf{u}_{\tau+1}) &\leq \frac{1}{\eta_1} + \hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1}) - f_\tau(\mathbf{x}) + f_\tau(\mathbf{u}_{\tau+1})\\ &\overset{(a)}{=} \frac{1}{\eta_1} + \langle \hat{g}_\tau - \mathbb{E}_\tau\left[\hat{g}_\tau\right], \mathbf{x} - \mathbf{u}_{\tau+1}\rangle,\end{aligned}$$

where (a) follows from the definitions of $f_\tau, \hat{f}_\tau$. Next, from Lemma 44 we know that for any $\mathbf{x} \in F_\tau$, $\|\mathbf{x} - \mathbf{x}_\tau\|_{M_\tau} \leq 4d\alpha$. Since $\mathbf{x}, \mathbf{u}_{\tau+1}$ are points in F_τ, we have $\|\mathbf{x} - \mathbf{u}_{\tau+1}\|_{M_\tau} \leq 8d\alpha$. Using, this we get

$$\begin{aligned}\langle \hat{g}_\tau - \mathbb{E}_\tau\left[\hat{g}_\tau\right], \mathbf{x} - \mathbf{u}_{\tau+1}\rangle &\leq \|\hat{g}_\tau - \mathbb{E}_\tau\left[\hat{g}_\tau\right]\|_{M_\tau}^* \|\mathbf{x} - \mathbf{u}_{\tau+1}\|_{M_\tau}\\ &\leq 16\lambda^{-1}d^2\alpha B,\end{aligned}$$

where the last inequality follows from the fact that $\|\hat{g}_\tau\|_{M_\tau}^*$ is a bounded random variable which satisfies $\|\hat{g}_\tau\|_{M_\tau}^* \leq \lambda^{-1}dB$. Since $16\lambda^{-1}d^2B\eta_1 \leq 1$, we have

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) - f_s(\mathbf{x}) + f_s(\mathbf{u}_{\tau+1}) \leq \frac{2}{\eta_1}.$$

This shows that $\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) \leq \frac{4}{\eta_1}$. We now show that this implies the restart condition should have triggered. Consider the following

$$\begin{aligned}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) &= \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{u}_{\tau+1})\\ &\leq \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\\ &= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}\rangle - \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T\hat{H}_s(\mathbf{x} - \mathbf{x}_s)\\ &= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}_{s+1}\rangle + \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle\\ &\quad - \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T\hat{H}_s(\mathbf{x} - \mathbf{x}_s)\\ &\overset{(a)}{\leq} \frac{4}{\eta_1} + 2\lambda^{-2}d^2B^2\sum_{s=1}^{\tau}\eta_s + \sum_{s=1}^{\tau}\langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle\\ &\quad - \sum_{s=1}^{\tau}\frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T\hat{H}_s(\mathbf{x} - \mathbf{x}_s),\end{aligned}$$

where (a) follows from the stability of the iterates we proved in Lemma 44. Since $\mathbf{x}_{s+1}$ is the minimizer of $\min_{\mathbf{y}\in F_s} \eta_s\langle \hat{g}_s, \mathbf{y}\rangle + \Phi_{R_{s+1}}(\mathbf{y}, \mathbf{x}_s)$, we have the following from the first order optimality conditions

$$\langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle \leq \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1}) - \Phi_{R_{s+1}}(\mathbf{x}_{s+1}, \mathbf{x}_s)}{\eta_s}.$$

Using this in the previous display, we get

$$\begin{aligned}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau}\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2}d^2B^2\sum_{s=1}^{\tau}\eta_s + \sum_{s=1}^{\tau}\frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1})}{\eta_s}\\ - \sum_{s=1}^{\tau}\frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T\hat{H}_s(\mathbf{x} - \mathbf{x}_s).\end{aligned}$$

Rearranging the terms in the RHS above, we get

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2}d^2B^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \frac{\Phi_{R_{\tau+1}}(\mathbf{x}, \mathbf{x}_{\tau+1})}{\eta_\tau}$$
$$+ \sum_{s=2}^{\tau} \left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right) \Phi_{R_s}(\mathbf{x}, \mathbf{x}_s).$$

Recall, $\mathbf{x} \in \partial F_\tau \cap \text{int}(\mathcal{X}_\xi)$. Let τ' be such that $\mathbf{x} \in \partial B_{\alpha, M_{\tau'}}(\mathbf{x}_{\tau'})$. Then

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2}d^2B^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \gamma\frac{\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})}{\eta_{\tau'}}.$$

Since $\|\mathbf{x} - \mathbf{x}_{\tau'}\|_{M_{\tau'}} = \alpha$, we have the following lower bound on $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$ which follows from property (P6) of SCB stated in Appendix C.7

$$\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'}) \geq \alpha - \log\left(1 + \alpha\right).$$

For our choice of α, $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$ can be lower bounded by $\alpha/2$. We now upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$. Since $\mathbf{x} \in \mathcal{X}_\xi$, using property P8 of SCB stated in Appendix C.7, we can upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$ as

$$\Phi_R(\mathbf{x}, \mathbf{x}_1) = R(\mathbf{x}) \leq 4\nu \log T.$$

Substituting the above two bounds in the previous display and using the fact that $\eta_\tau \leq 10\eta_1$, we get

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 20\lambda^{-2}d^2B^2\eta_1 T + \frac{4\nu\log T}{\eta_1} - \frac{\alpha\gamma}{20\eta_1} \leq -\frac{\beta}{\eta_1}.$$

This implies, the restart condition should have triggered. This shows that the focus region doesn't get updated more than $12d\log T$ times. □

Lemma 46. *Consider the setting of Theorem 38. Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then for any $t \leq \mathcal{T}$,*

$$M_t \preceq T^8(\nu + 2\sqrt{\nu})^2(\nabla^2 R(\mathbf{x}_1) + I).$$

Proof. First note that the iterates generated by the algorithm lie in $\mathcal{X}_\xi$, where $\xi = T^{-4}$. So using property P8 of SCB stated in Appendix C.7, we have

$$\forall t \leq \mathcal{T}, \quad \nabla^2 R(\mathbf{x}_t) \preceq \left(\frac{\nu + 2\sqrt{\nu}}{\xi}\right)^2 \nabla^2 R(\mathbf{x}_1) = T^8(\nu + 2\sqrt{\nu})^2 \nabla^2 R(\mathbf{x}_1).$$

Next, since f_t is T Lipschitz and since $\mathcal{X}$ contains a euclidean ball of radius 1 in it, we have $\nabla^2 f_t(\mathbf{x}) \preceq TI$. We now use the above two inequalities to bound M_t

$$M_t = \nabla^2 R(\mathbf{x}_t) + \sum_{s=1}^{t-1} \eta_s \hat{H}_s \preceq T^8(\nu + 2\sqrt{\nu})^2 \nabla^2 R(\mathbf{x}_1) + \sum_{s=1}^{t-1} \eta_s TI$$
$$\overset{(a)}{\preceq} T^8(\nu + 2\sqrt{\nu})^2(\nabla^2 R(\mathbf{x}_1) + I),$$

where (a) relied on the fact that $\eta_s \leq 10\eta_1$ for any $s \leq \mathcal{T}$ which we proved in Lemma 45. □

The following Lemma is concerned about concentration of loss estimates $\{\hat{f}_t\}_{t=1}^T$ computed by the modified algorithm. This Lemma helps us show that with high probability, the iterates of the modified and the original algorithms are exactly the same. Before we proceed, note that the focus region gets updated at most $12d\log T$ times before the algorithm restarts. So, for our choice of γ, we have $(1+\gamma)^{12d\log T} \leq 10$. Consequently, for all $t \leq \mathcal{T}$, $\eta_t \leq 10\eta_1$. So the results of Lemma 44 apply to all the iterates in the first $\mathcal{T}$ iterations of the modified algorithm.

Lemma 47 (Concentration of loss estimates)**.** *Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then for any $t \leq \mathcal{T}$, the following statement holds with probability at least $1 - T^{-2}$*

$$\sup_{\mathbf{x}\in F_t}\left|\sum_{s=1}^{t-1}\eta_1(\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - \iota_s f_s(\mathbf{x}) + \iota_s f_s(\mathbf{x}_s))\right| \leq \tilde{O}\left(\lambda^{-1}d^{5/2}\alpha B\eta_1\sqrt{T}\right).$$

Proof. First, note that

$$\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T\hat{H}_s(\mathbf{x}-\mathbf{x}_s) + \langle\hat{g}_s, \mathbf{x}-\mathbf{x}_s\rangle$$
$$\iota_s f_s(\mathbf{x}) - \iota_s f_s(\mathbf{x}_s) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T\hat{H}_s(\mathbf{x}-\mathbf{x}_s) + \iota_s\langle\nabla f_s(\mathbf{x}_s), \mathbf{x}-\mathbf{x}_s\rangle.$$

So $\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - \iota_s f_s(\mathbf{x}) + \iota_s f_s(\mathbf{x}_s) = \langle\hat{g}_s - \iota_s\nabla f_s(\mathbf{x}_s), \mathbf{x}-\mathbf{x}_s\rangle$. For any $\mathbf{x}\in F_t$, define random variables $Z_{\mathbf{x},s}$ as

$$Z_{\mathbf{x},s} = \begin{cases}\eta_1\langle\hat{g}_s - \iota_s\nabla f_s(\mathbf{x}_s), \mathbf{x}-\mathbf{x}_s\rangle & \text{if } s\leq\mathcal{T}\\ 0 \quad \text{otherwise}\end{cases}.$$

Since $\mathbb{E}_s[\hat{g}_s] = \iota_s\nabla f_s(\mathbf{x}_s)$, it is easy to see that $\{Z_{\mathbf{x},s}\}_{s=1}^T$ is a martingale difference sequence. Moreover, $Z_{\mathbf{x},s}$ is a bounded random variable. This follows from the fact that $\|\hat{g}_s\|_{M_s}^*$ is bounded and satisfies $\|\hat{g}_s\|_{M_s}^* \leq \lambda^{-1}dB$. Moreover, for any $\mathbf{x}\in F_s$, $\|\mathbf{x}-\mathbf{x}_s\|_{M_s} \leq 4d\alpha$ (see Lemma 44). So we have

$$|Z_{\mathbf{x},s}| \leq \eta_1|\langle\hat{g}_s - \iota_s\nabla f_s(\mathbf{x}_s), \mathbf{x}-\mathbf{x}_s\rangle| \leq 8\lambda^{-1}d^2\alpha B\eta_1.$$

By relying on standard concentration bounds for martingale difference sequences (see Lemma 42), we get that with probability at least $1-\delta$,

$$\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}| = O\left(\lambda^{-1}d^2\alpha B\eta_1\sqrt{T\log T/\delta}\right).$$

Next, we bound $\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}|$ using ϵ-net arguments. Let $\mathcal{N}_\epsilon$ be an ϵ-net over F_t which satisfies the following: for every $\mathbf{x}$, there exists a $\mathbf{x}_\epsilon \in \mathcal{N}_\epsilon$ such that $\|\mathbf{x}-\mathbf{x}_\epsilon\|_{M_t} \leq \epsilon$. Then

$$\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}| \leq \underbrace{\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\epsilon,s}|}_{T_1} + \underbrace{\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\epsilon,s} - Z_{\mathbf{x},s}|}_{T_2}.$$

Using a simple union bound, T_1 can be bounded as

$$T_1 \leq O\left(\lambda^{-1}d^2\alpha B\eta_1\sqrt{T\log T|\mathcal{N}_\epsilon|/\delta}\right) \overset{(a)}{\leq} O\left(\lambda^{-1}d^{5/2}\alpha B\eta_1\sqrt{T\log\frac{\alpha dT}{\epsilon\delta}}\right),$$

where the bound holds with probability at least $1-\delta$ and (a) holds since $\forall \mathbf{x} \in F_t$, $\|\mathbf{x}-\mathbf{x}_t\| \leq 4d\alpha$ and as a result $|\mathcal{N}_\epsilon| \leq \left(\frac{4d\alpha}{\epsilon}\right)^d$. T_2 can be bounded as follows

$$\begin{aligned}
\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\epsilon,s} - Z_{\mathbf{x},s}| &= \sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1}\eta_1\langle\hat{g}_s - \iota_s\nabla f_s(\mathbf{x}_s), \mathbf{x}-\mathbf{x}_\epsilon\rangle| \\
&\overset{(a)}{\leq} 2\eta_1\lambda^{-1}dB\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left(\sum_{s=0}^{t-1}\|\mathbf{x}-\mathbf{x}_\epsilon\|_{M_s}\right) \\
&\overset{(b)}{\leq} 2(1+4d\alpha)^2\eta_1\lambda^{-1}dB\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left(\sum_{s=0}^{t-1}\|\mathbf{x}-\mathbf{x}_\epsilon\|_{M_t}\right) = O\left(\lambda^{-1}d^3\alpha^2 B\eta_1\epsilon T\right),
\end{aligned}$$

where (a) follows from the fact that $\|\hat{g}_s\|_{M_s}^* \leq \lambda^{-1}dB$ and (b) follows from Lemma 44 where we showed that $M_s \preceq (1+4d\alpha)^2 M_t$. Choosing $\epsilon = \frac{1}{\alpha\sqrt{dT}}$, and plugging the above bounds for T_1, T_2 in the upper bound for $\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}|$ gives us the required result. □

Proof of Theorem 38. From Lemma 47, we know that with high probability, the iterates of the modified algorithm which relies on indicator variables ι_t are exactly same as the original algorithm. So it suffices to prove the regret bound for the modified algorithm. In the sequel, we work with the modified algorithm. Throughout the proof, we let $\mathcal{T}$ be the minimum between T and the first time step at which the algorithm restarts. Let τ be the minimum between $\mathcal{T}$ and the last time step where $\iota_\tau = 1$. Our goal is to bound the following quantity

$$\sum_{s=1}^{\mathcal{T}}\iota_s f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{s=1}^{\mathcal{T}}\iota_s f_s(\mathbf{x}) = \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{s=1}^{\tau} f_s(\mathbf{x}).$$

Case 1 ($\mathcal{T} = T$). We first consider the case where the restart condition didn't trigger in the first T iterations (i.e., $\mathcal{T} = T$). In this case, we show that the regret is $\tilde{O}\left(T^{1/2}\right)$. Since the restart condition hasn't triggered, we know that

$$\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau}\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{y}) \geq -\frac{\beta}{\eta_1}.$$

From the proof of Lemma 45, this implies $\forall \mathbf{x} \in \partial F_\tau \cap \operatorname{int}(\mathcal{X}_\xi)$

$$\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x}) - \min_{\mathbf{y}\in F_\tau}\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{y}) \geq \frac{4}{\eta_1}.$$

For the sake of clarity, we reproduce the argument we used in Lemma 45. To show this, we prove the contrapositive statement. Suppose $\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1}$ for some $\mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X}_\xi)$. We now show that this implies the restart condition should have triggered. Consider the following

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) &\leq \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x}) \\
&= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x} \rangle - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s) \\
&= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}_{s+1} \rangle + \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle \\
&\quad - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s) \\
&\overset{(a)}{\leq} \frac{4}{\eta_1} + 2\lambda^{-2} d^2 B^2 \sum_{s=1}^{\tau} \eta_s + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle \\
&\quad - \sum_{s=1}^{\tau} \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s),
\end{aligned}$$

where (a) follows from the stability of the iterates we proved in Lemma 44. Since $\mathbf{x}_{s+1}$ is the minimizer of $\min_{\mathbf{y}\in F_s} \eta_s \langle \hat{g}_s, \mathbf{y} \rangle + \Phi_{R_{s+1}}(\mathbf{y}, \mathbf{x}_s)$, we have the following from the first order optimality conditions

$$\langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle \leq \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1}) - \Phi_{R_{s+1}}(\mathbf{x}_{s+1}, \mathbf{x}_s)}{\eta_s}.$$

Using this in the previous display, we get

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2} d^2 B^2 \sum_{s=1}^{\tau} \eta_s + \sum_{s=1}^{\tau} \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1})}{\eta_s} \\
- \sum_{s=1}^{\tau} \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s).
\end{aligned}$$

Rearranging the terms in the RHS above, we get

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2} d^2 B^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \frac{\Phi_{R_{\tau+1}}(\mathbf{x}, \mathbf{x}_{\tau+1})}{\eta_\tau} \\
+ \sum_{s=2}^{\tau} \left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}} \right) \Phi_{R_s}(\mathbf{x}, \mathbf{x}_s).
\end{aligned}$$

Recall, $\mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X})$. Let τ' be such that $\mathbf{x} \in \partial B_{\alpha, M_{\tau'}}(\mathbf{x}_{\tau'})$. Then

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 2\lambda^{-2} d^2 B^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \gamma \frac{\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})}{\eta_{\tau'}}.$$

Since $\|\mathbf{x} - \mathbf{x}_{\tau'}\|_{M_{\tau'}} = \alpha$, we have the following lower bound on $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$ which follows from property (P6) of SCB stated in Appendix C.7

$$\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'}) \geq \alpha - \log(1+\alpha).$$

For our choice of α, $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$ can be lower bounded by $\alpha/2$. We now upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$. Since $\mathbf{x} \in \mathcal{X}_\xi$, using property P8 of SCB stated in Appendix C.7, we can upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$ as

$$\Phi_R(\mathbf{x}, \mathbf{x}_1) = R(\mathbf{x}) \leq 4\nu \log T.$$

Substituting the above two bounds in the previous display and using the fact that $\eta_\tau \leq 10\eta_1$, we get

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y} \in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 20\lambda^{-2} d^2 B^2 \eta_1 T + \frac{4\nu \log T}{\eta_1} - \frac{\alpha\gamma}{20\eta_1} \leq -\frac{\beta}{\eta_1}.$$

This implies, the restart condition should have triggered. But since the restart condition hasn't triggered, this result shows that $\forall \mathbf{x} \in \partial F_\tau \cap \text{int}(\mathcal{X}_\xi), \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \min_{\mathbf{y} \in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \geq \frac{4}{\eta_1}$. Next, since our cumulative loss estimate concentrates well around the true cumulative loss (i.e., $\iota_\tau = 1$), this implies

$$\forall \mathbf{x} \in \partial F_\tau \cap \text{int}(\mathcal{X}_\xi), \quad \sum_{s=1}^{\tau} f_s(\mathbf{x}) - \min_{\mathbf{y} \in F_\tau} \sum_{s=1}^{\tau} f_s(\mathbf{y}) \geq \frac{2}{\eta_1}.$$

Since f_s's are convex, this implies the minimizer of $\min_{\mathbf{x} \in \mathcal{X}_\xi} \sum_{s=0}^{T} f_s(\mathbf{x})$ is in F_τ. So, the regret of the algorithm can be bounded as follows

$$\begin{aligned}\text{Reg}_T = \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{s=1}^{\tau} f_s(\mathbf{x}) &\overset{(a)}{\leq} 1 + \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x} \in \mathcal{X}_\xi} \sum_{s=1}^{\tau} f_s(\mathbf{x}) \\ &= 1 + \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x} \in F_\tau} \sum_{s=1}^{\tau} f_s(\mathbf{x}),\end{aligned}$$

where (a) follows from the definition of $\mathcal{X}_\xi = (1-\xi)\mathcal{X} + \xi \mathbf{x}_1$ and the fact that the loss functions are Lipschitz and the diameter of $\mathcal{X}$ is bounded. Next, consider the following for any $\mathbf{x} \in F_\tau$

$$\begin{aligned}\sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \sum_{s=1}^{\tau} f_s(\mathbf{x}) = \underbrace{\sum_{s=1}^{\tau} [f_s(\mathbf{y}_s) - f_s(\mathbf{x}_s)]}_{T_1} + \underbrace{\sum_{s=1}^{\tau} \left[f_s(\mathbf{x}_s) - f_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) + \hat{f}_s(\mathbf{x})\right]}_{T_2} \\ + \underbrace{\sum_{s=1}^{\tau} \left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right]}_{T_3}.\end{aligned}$$

Bounding T_1. We first bound T_1. Since f_s is a quadratic function with Hessian $\hat{H}_s$, we have

$$\sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - f_s(\mathbf{x}_s) = \sum_{s=1}^{\tau} \lambda\langle \nabla f_s(\mathbf{x}_s), M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})\rangle + \frac{\lambda^2}{2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2}\hat{H}_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})$$

Let $Z_s = \lambda\langle \nabla f_s(\mathbf{x}_s), M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})\rangle$ if $s \leq \tau$ and 0 if $s > \tau$. Note that $\{Z_s\}_{s=1}^T$ is a martingale difference sequence with each Z_s being bounded: $|Z_s| \leq 2dB$. This follows from the observation that $\nabla f_s(\mathbf{x}_s) = \mathbb{E}_s\left[\hat{g}_s\right]$ and the fact that $M_s^{-1/2}\hat{g}_s$ is a bounded random variable. By relying on standard concentration bounds for martingale difference sequences (see Lemma 42), we get that with probability at least $1-\delta$, $\sum_{s=1}^T Z_s = O\left(dB\sqrt{T\log 1/\delta}\right)$. We now bound the last term in the RHS above. Consider the following

$$\begin{aligned}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2}\hat{H}_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}) &\leq 4\|M_s^{-1/2}\hat{H}_s M_s^{-1/2}\|_2 \\ &\leq 4\|M_{s+1}^{-1/2}\hat{H}_s M_{s+1}^{-1/2}\|_2\|M_s^{-1/2}M_{s+1}M_s^{-1/2}\|_2\end{aligned}$$

From Lemma 44 we know that $\|M_s^{-1/2}M_{s+1}M_s^{-1/2}\|_2 \leq 1 + 12\lambda^{-2}d^2B\eta_t \leq 2$. So we have

$$(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2}\hat{H}_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}) \leq 8\|M_{s+1}^{-1/2}\hat{H}_s M_{s+1}^{-1/2}\|_2 = 8\|\hat{H}_s^{1/2}M_{s+1}^{-1}\hat{H}_s^{1/2}\|_2.$$

Define $N_t = (1+4d\alpha)^{-2}\nabla^2 R(\mathbf{x}_1) + \sum_{s=1}^{t-1}\eta_s\hat{H}_s$. From Lemma 44 we know that $\nabla^2 R(\mathbf{x}_t) \succeq (1+4d\alpha)^{-2}\nabla^2 R(\mathbf{x}_1)$. So $N_t \preceq M_t$ for all t. Using this in the previous inequality we get

$$\begin{aligned}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2}\hat{H}_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}) &\leq 8\|\hat{H}_s^{1/2}N_{s+1}^{-1}\hat{H}_s^{1/2}\|_2 \\ &\leq 8\mathrm{tr}\left(N_{s+1}^{-1}\hat{H}_s\right) \\ &= \frac{8}{\eta_s}\mathrm{tr}\left(N_{s+1}^{-1}(N_{s+1} - N_s)\right)\end{aligned}$$

By relying on Lemma 41 we can upper bound $\sum_{s=1}^{\tau}\frac{8}{\eta_s}\mathrm{tr}\left(N_{s+1}^{-1}(N_{s+1} - N_s)\right)$ as

$$\sum_{s=1}^{\tau}\frac{8}{\eta_s}\mathrm{tr}\left(N_{s+1}^{-1}(N_{s+1} - N_s)\right) \leq \frac{8}{\eta_1}\sum_{s=1}^{\tau}\mathrm{tr}\left(N_{s+1}^{-1}(N_{s+1} - N_s)\right) \leq \frac{8}{\eta_1}\log\frac{\det N_T}{\det N_1} \tag{C.4}$$

From Lemma 46 we know that $N_T \preceq \mathrm{poly}(dT)$. Assuming $\nabla^2 R(\mathbf{x}_1) \succeq \frac{1}{\mathrm{poly}(dT)}I$, the RHS above can be upper bounded as $O\left(\frac{d\log dT}{\eta_1}\right)$. To summarize, we have the following upper bound T_1: $O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right)$

Bounding T_2. Since $\iota_\tau = 1$, T_2 can be upper bounded as

$$\begin{aligned}T_2 &\leq \frac{1}{\eta_1} + \left[f_\tau(\mathbf{x}_\tau) - f_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{x}_\tau) + \hat{f}_\tau(\mathbf{x})\right] \\ &= \frac{1}{\eta_1} + \langle \hat{g}_\tau - \mathbb{E}_\tau\left[\hat{g}_\tau\right], \mathbf{x} - \mathbf{x}_\tau\rangle \leq \frac{2}{\eta_1},\end{aligned}$$

where the last inequality follows from the fact that $\|\mathbf{x}-\mathbf{x}_\tau\|_{M_\tau} \leq 4d\alpha$ and $\|\hat{g}_\tau\|_{M_\tau}^* \leq \lambda^{-1}dB$.

Bounding T_3. To bound T_3, we consider the following

$$\begin{aligned}\sum_{s=1}^{\mathcal{T}}\left[\hat{f}_s(\mathbf{x}_s)-\hat{f}_s(\mathbf{x})\right] &= \sum_{s=1}^{\mathcal{T}}\langle \hat{g}_s, \mathbf{x}_s-\mathbf{x}\rangle - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T\hat{H}_s(\mathbf{x}-\mathbf{x}_s)\\ &= \sum_{s=1}^{\mathcal{T}}\langle \hat{g}_s, \mathbf{x}_s-\mathbf{x}_{s+1}\rangle + \langle \hat{g}_s, \mathbf{x}_{s+1}-\mathbf{x}\rangle - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T\hat{H}_s(\mathbf{x}-\mathbf{x}_s)\end{aligned}$$

Using similar arguments as at the beginning of Case 1, this can be bounded as

$$\sum_{s=1}^{\mathcal{T}}\left[\hat{f}_s(\mathbf{x}_s)-\hat{f}_s(\mathbf{x})\right] \leq 2\lambda^{-2}d^2B^2\sum_{s=1}^{\mathcal{T}}\eta_s + \frac{\Phi_R(\mathbf{x},\mathbf{x}_1)}{\eta_1},$$

Since $\mathbf{x}\in\mathcal{X}_\xi$, using property P8 of SCB stated in Appendix C.7, we can upper bound $\Phi_R(\mathbf{x},\mathbf{x}_1)$ as

$$\Phi_R(\mathbf{x},\mathbf{x}_1) = R(\mathbf{x}) \leq 4\nu\log T.$$

Combining the bounds for T_1, T_2, T_3 shows that with probability at least $1-T^{-2}$ the regret is upper bounded by

$$\tilde{O}\left(dB\sqrt{T} + \frac{(\nu+d)}{\eta_1} + \lambda^{-2}d^2B^2\eta_1 T\right) = \tilde{O}\left(d^{3.5}(d+\nu)^2\sqrt{T}\right).$$

Case 2 ($\mathcal{T}<T$). We now consider the case where the restart condition triggered at some iteration $\mathcal{T}<T$. Using the fact that the restart condition hasn't triggered in iteration $\mathcal{T}-1$ and using similar arguments as in the beginning of Case 1, we can again show that the minimizer of the cumulative loss over the entire domain lies in the focus region $F_{\mathcal{T}}$, and $\iota_{\mathcal{T}}=1$. So regret until $\mathcal{T}$ is given by

$$\begin{aligned}\mathrm{Reg}_{\mathcal{T}} = \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}) &\overset{(a)}{\leq} 1 + \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}_\xi}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x})\\ &= 1 + \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in F_{\mathcal{T}}}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}),\end{aligned}$$

where (a) follows from the definition of $\mathcal{X}_\xi$. Using the same regret decomposition as in Case 1, for any $\mathbf{x}\in F_{\mathcal{T}}$

$$\begin{aligned}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}) = \underbrace{\sum_{s=1}^{\mathcal{T}}\left[f_s(\mathbf{y}_s)-f_s(\mathbf{x}_s)\right]}_{T_1} + \underbrace{\sum_{s=1}^{\mathcal{T}}\left[f_s(\mathbf{x}_s)-f_s(\mathbf{x})-\hat{f}_s(\mathbf{x}_s)+\hat{f}_s(\mathbf{x})\right]}_{T_2}\\ + \underbrace{\sum_{s=1}^{\mathcal{T}}\left[\hat{f}_s(\mathbf{x}_s)-\hat{f}_s(\mathbf{x})\right]}_{T_3}.\end{aligned}$$

We use the same arguments as in Case 1 to bound T_1, T_2 as

$$T_1 = O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right), \quad T_2 = \frac{2}{\eta_1}.$$

Since the restart condition triggered in round $\mathcal{T}$, T_3 is bounded by $-\frac{\beta}{\eta_1}$. Combining all these bounds, we get the following bound on regret

$$\text{Reg}_{\mathcal{T}} \leq O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right) + \frac{2}{\eta_1} - \frac{\beta}{\eta_1}.$$

For our choice of hyper-parameters, the above bound is less than 0.

C.5 Proof of Theorem 10

The proof of this Theorem uses similar arguments as the proof of "known Hessian" case in Appendix C.4. The additional complexity in proving Theorem 10 comes from dealing with Hessian estimates instead of exact Hessians used in Appendix C.4. In particular, in Theorem 10, we need to prove one additional result regarding the concentration of cumulative Hessian estimates.

We first introduce some notation we use in the proof. We let $r_t(\mathbf{x}) = f_t(\mathbf{x}) - q_t(\mathbf{x})$, where $q_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A_t\mathbf{x} + \langle \mathbf{b}_t, \mathbf{x}\rangle + c_t$. Recall, $\sup_{\mathbf{x}\in\mathcal{X}} |r_t(\mathbf{x})| \leq \epsilon$. We let $H_t = \frac{1}{2}(A_t + A_t^T)$ denote the Hessian of $q_t(\mathbf{x})$. Define random variable Z_t as

$$Z_t = 2^{-1}\lambda^{-2}d^2 f_t(\mathbf{y}_t)\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right).$$

Since f_t is bounded, it is easy to see that Z_t is a bounded random variable (assuming M_t is positive definite and $\mathbf{y}_t \in \mathcal{X}$). In particular, Z_t can be bounded as

$$\|Z_t\|_2 \leq \lambda^{-2}d^2(B+\epsilon). \tag{C.5}$$

Another important thing to note here is that

$$\mathbb{E}_t\left[2^{-1}\lambda^{-2}d^2 q_t(\mathbf{y}_t)\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right)\right] = M_t^{-1/2}H_t M_t^{-1/2}.$$

Consequently, $\|Z_t - M_t^{-1/2}H_tM_t^{-1/2}\|_2 \leq 2\lambda^{-2}d^2(B+\epsilon)$.

Similar to Appendix C.4, to prove Theorem 10, we work with a slightly modified algorithm and show that with high probability, the iterates of the modified algorithm are exactly same as the original algorithm. Consequently, proving the Theorem for the modified algorithm entails that the Theorem also holds for the actual algorithm. In the modified algorithm, we slightly change the random variables $\hat{g}_t, \hat{H}_t$ and work with the following sequence of random variables

$$\hat{g}_t = \lambda^{-1}d\iota_t f_t(\mathbf{y}_t)M_t^{1/2}\mathbf{v}_{1,t}, \quad \hat{H}_t = \frac{\lambda^{-2}}{2}d^2\iota_t f_t(\mathbf{y}_t)M_t^{1/2}\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right)M_t^{1/2}.$$

where ι_t is an indicator random variable which is equal to 1 if and only if the following two events happen

$$\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \leq \frac{1}{10(1+8d\alpha)^2},$$

$$\sup_{\mathbf{x}\in F_t} \Big| \sum_{s=1}^{t-1} (\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - \iota_s q_s(\mathbf{x}) + \iota_s q_s(\mathbf{x}_s)) \Big| \leq \frac{1}{\eta_1}.$$

Here, we define $\tilde{M}_t$ as $\tilde{M}_t = \nabla^2 R(\mathbf{x}_t) + \sum_{s=0}^{t-1} \eta_s \iota_s H_s$. Intuitively, the first event happens when M_t is spectrally close to $\tilde{M}_t$, and the second event happens when the cumulative loss estimate $\sum_{s=1}^{t-1} \hat{f}_s(\mathbf{x})$ is close to the true cumulative loss $\sum_{s=1}^{t-1} q_s(\mathbf{x})$. We assume the algorithm is run with these modified estimates of gradients and Hessians[1]. The main benefit of working with the modified gradient and Hessian estimates is that they are bounded and are more amenable to analysis. Our proof shows that with high probability, the modified random variables $\hat{g}_t, \hat{H}_t$ are exactly equal to the original random variables. As a result, it suffices to prove Theorem 10 for the hypothetical algorithm.

We now derive some useful properties of the iterates produced by the modified algorithm.

Lemma 48 (Properties of iterates)**.** *Consider the setting of Theorem 10. Let $\mathcal{T}$ be the minimum between T and the first iteration at which the modified algorithm restarts. For any $t < \mathcal{T}$ such that $\eta_t \leq 10\eta_1$, the iterates of the algorithm satisfy the following stability properties*

1. *M_t is positive definite and $\mathbf{y}_t \in \mathcal{X}$.*
2. *$R_t(\mathbf{x})$ is a strictly convex function over F_t.*
3. *For all $\mathbf{x} \in F_t$, $\|\mathbf{x} - \mathbf{x}_t\|_{M_t} \leq 4d\alpha$ and $\nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+8d\alpha)^2} \nabla^2 R(\mathbf{x}_t)$.*
4. *$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\tilde{M}_t} \leq c\eta_t$ and $\|I - \tilde{M}_t^{-1/2} \tilde{M}_{t+1} \tilde{M}_t^{-1/2}\|_2 \leq 4c\eta_t$. Here $c = 10(B+\epsilon)(\lambda^{-1}d + \lambda^{-2}d^3\alpha)$.*
5. *if $\iota_t = 0$, then $\iota_t = \iota_{t+1} = \cdots = \iota_{\mathcal{T}}$, $\mathbf{x}_t = \mathbf{x}_{t+1} \cdots = \mathbf{x}_{\mathcal{T}}$ and $F_t = F_{t+1} \cdots = F_{\mathcal{T}}$.*

Proof. The proof uses similar arguments as in the proof of Lemma 44. So to avoid redundancy, we often directly rely on some of the results proved in Lemma 44. We use induction to prove the lemma.

Base Case (t=1).

1. First note that $\tilde{M}_1 = M_1 = \nabla^2 R(\mathbf{x}_1)$. So the proof follows from the proof of corresponding part in Lemma 44.
2. The proof of this part follows from the proof of corresponding part in Lemma 44.
3. The proof of this part follows from the proof of corresponding part in Lemma 44.
4. We now show that $\mathbf{x}_2$ and $\mathbf{x}_1$ are close to each other. Note that $\mathbf{x}_2$ is the minimizer of the following objective

$$\mathbf{x}_2 \in \operatorname*{argmin}_{\mathbf{x}\in F_1} \eta_1 \langle \hat{g}_1, \mathbf{x} \rangle + \Phi_{R_2}(\mathbf{x}, \mathbf{x}_1). \tag{C.6}$$

[1] It should be noted that this is a hypothetical algorithm. We can not actually run this algorithm in practice as we can not compute ι_t

From first order optimality conditions we have

$$\forall \mathbf{x} \in F_1, \quad \langle \nabla R_2(\mathbf{x}_2) - \nabla R_2(\mathbf{x}_1) + \eta_1 \hat{g}_1, \mathbf{x} - \mathbf{x}_2 \rangle \geq 0.$$

Substituting $\mathbf{x}_1$ in the above equation gives us

$$\langle \nabla R_2(\mathbf{x}_2) - \nabla R_2(\mathbf{x}_1) + \eta_1 \hat{g}_1, \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq 0.$$

This can equivalently be written as

$$\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1 \hat{g}_1 + \eta_1 \hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle \leq 0. \tag{C.7}$$

Now suppose $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} > c\eta_1$, where $c = 10(B+\epsilon)(\lambda^{-1}d + \lambda^{-2}d^3\alpha)$. Then we have

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1 \hat{g}_1 + \eta_1 \hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle \\
&\overset{(a)}{\geq} \frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}^2}{1 + \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}} + \langle \eta_1 \hat{g}_1 + \eta_1 \hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle \\
&\geq \frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}^2}{1 + \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}} - \eta_1 \left(\|\hat{g}_1\|_{M_1}^* + \|\hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1)\|_{M_1}^* \right) \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \\
&= \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \left(\frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}}{1 + \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}} - \eta_1 \|\hat{g}_1\|_{M_1}^* - \eta_1 \|\hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1)\|_{M_1}^* \right),
\end{aligned}$$

where (a) follows from property P7 of SCBs stated in Appendix C.7. Next, consider the following

$$(\|\hat{g}_1\|_{M_1}^*)^2 = \hat{g}_1^T M_1^{-1} \hat{g}_1 = \lambda^{-2} d^2 f_1^2(\mathbf{y}_1) \mathbf{v}_{1,1}^T \mathbf{v}_{1,1} \leq \lambda^{-2} d^2 (B+\epsilon)^2.$$

$$\begin{aligned}
(\|\hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1)\|_{M_1}^*)^2 &= (\mathbf{x}_2 - \mathbf{x}_1)^T \hat{H}_1 M_1^{-1} \hat{H}_1 (\mathbf{x}_2 - \mathbf{x}_1)^T \\
&\leq \left(\frac{d^2 f_1(\mathbf{y}_1)}{2\lambda^2} \right)^2 \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1}^2 \|\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T + \mathbf{v}_{2,1}\mathbf{v}_{1,1}^T\|_2^2 \\
&\overset{(a)}{\leq} 16\lambda^{-4} d^6 (B+\epsilon)^2 \alpha^2,
\end{aligned}$$

where (a) follows from the fact that $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \leq 4d\alpha$ proved in point (3). Substituting this in the previous inequality and using the fact that $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} > c\eta_1$ gives us

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}_2) - \nabla R(\mathbf{x}_1) + \eta_1 \hat{g}_1 + \eta_1 \hat{H}_1(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle \\
&\geq \frac{c}{2} \eta_1 \|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \left(\frac{2}{1 + c\eta_1} - 1 \right) \overset{(a)}{>} 0,
\end{aligned}$$

where (a) follows from the fact that $c\eta_1 < 1/2$. This contradicts the first order optimality condition in Equation (C.7). This shows that $\|\mathbf{x}_2 - \mathbf{x}_1\|_{M_1} \leq c\eta_1$.

Next, we show that $\tilde{M}_1^{-1/2} \tilde{M}_2 \tilde{M}_1^{-1/2}$ is close to identity. From the definitions of $\tilde{M}_1, \tilde{M}_2$, we have

$$\tilde{M}_1^{-1/2} \tilde{M}_2 \tilde{M}_1^{-1/2} - I = \tilde{M}_1^{-1/2} (\nabla^2 R(\mathbf{x}_2) - \nabla^2 R(\mathbf{x}_1)) \tilde{M}_1^{-1/2} + \eta_1 \tilde{M}_1^{-1/2} H_1 \tilde{M}_1^{-1/2}.$$

Since $\|\mathbf{x}_2 - \mathbf{x}_1\|_{\tilde{M}_1} \leq c\eta_1 < 1$, we can rely on property P2 of SCB stated in Section 4.1 to infer that

$$\nabla^2 R(\mathbf{x}_2) \preceq \frac{1}{(1-c\eta_1)^2}\nabla^2 R(\mathbf{x}_1) \preceq (1+3c\eta_1)\nabla^2 R(\mathbf{x}_1),$$

where the last inequality follows since $c\eta_1 < 1/10$. Next, note that H_1 can be written as

$$H_1 = \mathbb{E}\left[\frac{\lambda^{-2}}{2}d^2 f_1(\mathbf{y}_1)\tilde{M}_1^{1/2}\left(\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T + \mathbf{v}_{2,1}\mathbf{v}_{1,1}^T\right)\tilde{M}_1^{1/2}\right].$$

So we have $\tilde{M}_1^{-1/2}H_1\tilde{M}_1^{-1/2} = \mathbb{E}\left[\frac{\lambda^{-2}}{2}d^2 f_1(\mathbf{y}_1)\left(\mathbf{v}_{1,1}\mathbf{v}_{2,1}^T + \mathbf{v}_{2,1}\mathbf{v}_{1,1}^T\right)\right]$ which is a bounded quantity. Substituting the previous two bounds in our expression for $\tilde{M}_1^{-1/2}\tilde{M}_2\tilde{M}_1^{-1/2} - I$, we get

$$\|\tilde{M}_1^{-1/2}\tilde{M}_2\tilde{M}_1^{-1/2} - I\|_2 \leq 4c\eta_1.$$

5. Since $M_1 = \tilde{M}_1$, ι_1 is always equal to 1. So the last property trivially holds. This finishes the proof of the base case.

Induction Step. Suppose the Lemma holds for the first $t-1$ iterations. We now show that it also holds for the t^{th} iteration.

1. **Invertibility.** We first show that M_t is positive definite. If $\iota_{t-1} = 0$, then it is easy to see that M_t is equal to M_{t-1}, which we know is positive definite. So lets consider the where $\iota_{t-1} = 1$. We know that $\iota_1 = \iota_2 = \dots = \iota_{t-1}$ and $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\tilde{M}_{t-1}} \leq c\eta_{t-1}$. Now, consider the following

$$\begin{aligned} M_t &= \nabla^2 R(\mathbf{x}_t) + \sum_{s=1}^{t-1}\eta_s\hat{H}_s \\ &= M_{t-1} + \eta_{t-1}\hat{H}_{t-1} + \nabla^2 R(\mathbf{x}_t) - \nabla^2 R(\mathbf{x}_{t-1}) \\ &\overset{(a)}{\succeq} M_{t-1} + \eta_{t-1}\hat{H}_{t-1} - 2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_{\nabla^2 R(\mathbf{x}_{t-1})}\nabla^2 R(\mathbf{x}_{t-1}), \end{aligned}$$

where (a) follows from the property of self-concordant functions stated in Equation (4.1). From stability, we have

$$\begin{aligned} M_t &\succeq M_{t-1} + \eta_{t-1}\hat{H}_{t-1} - 2c\eta_{t-1}\nabla^2 R(\mathbf{x}_{t-1}) \\ &= M_{t-1}^{1/2}\left[I + \eta_{t-1}Z_{t-1} - 2c\eta_{t-1}M_{t-1}^{-1/2}\nabla^2 R(\mathbf{x}_{t-1})M_{t-1}^{-1/2}\right]M_{t-1}^{1/2}. \end{aligned}$$

We now show that $\left[I + \eta_{t-1}Z_{t-1} - 2c\eta_{t-1}M_{t-1}^{-1/2}\nabla^2 R(\mathbf{x}_{t-1})M_{t-1}^{-1/2}\right] \succ 0$. To show this, we rely on the following argument

$$\begin{aligned} &\|Z_{t-1} - 2cM_{t-1}^{-1/2}\nabla^2 R(\mathbf{x}_{t-1})M_{t-1}^{-1/2}\|_2 \\ &\quad\leq \|Z_{t-1}\|_2 + 2c\|M_{t-1}^{-1/2}\tilde{M}_{t-1}M_{t-1}^{-1/2}\|_2\|\tilde{M}_{t-1}^{-1/2}\nabla^2 R(\mathbf{x}_{t-1})\tilde{M}_{t-1}^{-1/2}\|_2 \\ &\quad\leq \lambda^{-2}d^2(B+\epsilon) + 3c \leq 4c, \end{aligned}$$

where the last inequality follows from the fact that Z_t is a bounded random variable, and $\|I - \tilde{M}_{t-1}^{-1/2} M_{t-1} \tilde{M}_{t-1}^{-1/2}\|_2 \leq \frac{1}{10}$ (consequently, $\|M_{t-1}^{-1/2} \tilde{M}_{t-1} M_{t-1}^{-1/2}\|_2 \leq \frac{3}{2}$). This shows that for our choice of hyper-parameters, M_t is invertible.

Valid Iterates. Next, we show that $\mathbf{y}_t \in \mathcal{X}$. If $\iota_{t-1} = 0$, then it is easy to see that this is the case (because $M_t = M_{t-1}$). So we assume $\iota_{t-1} = 1$. In this case, we first bound $\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2$ (i.e., we show that M_t and $\tilde{M}_t$ are spectrally close). Consider the following

$$\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 = \|\tilde{M}_t^{-1/2}(\tilde{M}_{t-1} - M_{t-1} + \eta_{t-1}(\hat{H}_{t-1} - H_{t-1}))\tilde{M}_t^{-1/2}\|_2 \tag{C.8}$$

$$\leq \|\tilde{M}_t^{-1/2}(\tilde{M}_{t-1} - M_{t-1})\tilde{M}_t^{-1/2}\|_2 \tag{C.9}$$

$$+ \eta_{t-1}\|\tilde{M}_t^{-1/2}(\hat{H}_{t-1} - H_{t-1})\tilde{M}_t^{-1/2}\|_2 \tag{C.10}$$

Consider the first term in the RHS above

$$\begin{aligned}\|\tilde{M}_t^{-1/2}(\tilde{M}_{t-1} - M_{t-1})\tilde{M}_t^{-1/2}\|_2 &\leq \|\tilde{M}_{t-1}^{-1/2}(\tilde{M}_{t-1} - M_{t-1})\tilde{M}_{t-1}^{-1/2}\|\|\tilde{M}_t^{-1/2}\tilde{M}_{t-1}\tilde{M}_t^{-1/2}\| \\ &\leq \frac{1}{5(1+8d\alpha)^2},\end{aligned}$$

where the last inequality follows from the fact that $\|I - \tilde{M}_{t-1}^{-1/2} M_{t-1} \tilde{M}_{t-1}^{-1/2}\|_2 \leq \frac{1}{10(1+8d\alpha)^2}$ and the fact that $\tilde{M}_{t-1}$ is spectrally close to $\tilde{M}_t$. Now consider the second term in the RHS of Equation (C.8). Since $\hat{H}_{t-1} = M_{t-1}^{1/2} Z_{t-1} M_{t-1}^{1/2}$, we have

$$\begin{aligned}&\|\tilde{M}_t^{-1/2}(\hat{H}_{t-1} - H_{t-1})\tilde{M}_t^{-1/2}\|_2 \\ &\quad = \|\tilde{M}_t^{-1/2} M_{t-1}^{1/2}(Z_{t-1} - M_{t-1}^{-1/2} H_{t-1} M_{t-1}^{-1/2}) M_{t-1}^{1/2} \tilde{M}_t^{-1/2}\|_2 \\ &\quad \leq \|\tilde{M}_{t-1}^{-1/2} M_{t-1}^{1/2}(Z_{t-1} - M_{t-1}^{-1/2} H_{t-1} M_{t-1}^{-1/2}) M_{t-1}^{1/2} \tilde{M}_{t-1}^{-1/2}\|_2 \\ &\quad \overset{(a)}{\leq} 2\|Z_{t-1} - M_{t-1}^{-1/2} H_{t-1} M_{t-1}^{-1/2}\|_2 \\ &\quad \overset{(b)}{\leq} 2\,(B+\epsilon)\, d^2 \lambda^{-2}.\end{aligned}$$

where (a) follows from the fact that $\|I - \tilde{M}_{t-1}^{-1/2} M_{t-1} \tilde{M}_{t-1}^{-1/2}\|_2 \leq \frac{1}{10(1+8d\alpha)^2}$ and (b) follows from Equation (C.5). Combining the previous two displays shows that for our choice of η_1, $\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \leq \frac{1}{4(1+8d\alpha)^2}$. This shows that M_t is spectrally close to $\tilde{M}_t$ and

$$\|\mathbf{y}_t - \mathbf{x}_t\|_{\tilde{M}_t} \leq 2\|\mathbf{y}_t - \mathbf{x}_t\|_{M_t} \leq 4\lambda < 1.$$

Since $\|\mathbf{y}_t - \mathbf{x}_t\|_{\tilde{M}_t} \geq \|\mathbf{y}_t - \mathbf{x}_t\|_{\nabla^2 R(\mathbf{x}_t)}$, using the Dikin Ellipsoid property of SCB stated in Section 4.1, we have $\mathbf{y}_t \in \mathcal{X}$.

2. The focus region update condition of our algorithm always ensures that

$$\mathrm{Vol}(F_t \cap B_{\alpha, M_t}(\mathbf{x}_t)) \geq \frac{1}{2}\mathrm{Vol}(F_t).$$

So, from Lemma 40 we know that for any $\mathbf{x} \in F_t$, $\|\mathbf{x}-\mathbf{x}_t\|_{M_t} \leq 4d\alpha$. Using this, together with the fact that $\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \leq \frac{1}{4(1+8d\alpha)^2}$, we get $\|\mathbf{x} - \mathbf{x}_t\|_{\tilde{M}_t} \leq 8d\alpha$. By relying on Assumption 1 on SCB R, we then get

$$\forall \mathbf{x} \in F_t, \ \nabla^2 R(\mathbf{x}) \succeq \frac{1}{(1+8d\alpha)^2} \nabla^2 R(\mathbf{x}_t).$$

3. We now show that $R_t(\mathbf{x})$ is strictly convex over interior of F_t. Consider the following for any $\mathbf{x} \in \text{int}(F_t)$

$$\begin{aligned}
\nabla^2 R(\mathbf{x}) + \eta_{1:t-1}\hat{H}_{1:t-1} &\overset{(a)}{\succeq} \frac{1}{(1+8d\alpha)^2}\nabla^2 R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1}\\
&\succeq \frac{1}{(1+8d\alpha)^2}\nabla^2 R(\mathbf{x}_t) + \eta_{1:t-1}H_{1:t-1} + (M_t - \tilde{M}_t)\\
&\overset{(b)}{\succeq} \frac{1}{(1+8d\alpha)^2}\tilde{M}_t - \frac{1}{4(1+8d\alpha)^2}\tilde{M}_t\\
&\succ 0,
\end{aligned}$$

where (a) follows from the previous property and (b) follows from the fact that $\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \leq \frac{1}{4(1+8d\alpha)^2}$. This shows that R_t is strictly convex over F_t.

4. We now prove stability of the iterates. In particular, we show that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\tilde{M}_t} \leq c\eta_t$. If $\iota_{t-1} = 0$, then this trivially holds. So lets consider the case where $\iota_{t-1} = 1$. From the first order optimality conditions, we have

$$\forall \mathbf{x} \in F_t, \quad \langle \nabla R_{t+1}(\mathbf{x}_{t+1}) - \nabla R_{t+1}(\mathbf{x}_t) + \eta_t \hat{g}_t, \mathbf{x} - \mathbf{x}_{t+1}\rangle \geq 0.$$

Note that from our definition of F_t, F_{t-1} we always have $F_t \subseteq F_{t-1}$ and $\mathbf{x}_t \in F_t$. So substituting $\mathbf{x}_t$ in the first equation gives us

$$\langle \nabla R(\mathbf{x}_{t+1}) - \nabla R(\mathbf{x}_t) + \eta_t \hat{g}_t + \sum_{s=1}^{t} \eta_s \hat{H}_s(\mathbf{x}_{t+1} - \mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1}\rangle \geq 0.$$

To prove the required result, we show that for any $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_t\|_{\tilde{M}_t} > c\eta_t$, the following holds

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}) - \nabla R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1}(\mathbf{x} - \mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle\\
&\qquad > \eta_t \|\hat{g}_t\|_{\tilde{M}_t}^{*} \|\mathbf{x} - \mathbf{x}_t\|_{\tilde{M}_t} + \eta_t \|\hat{H}_t(\mathbf{x} - \mathbf{x}_t)\|_{\tilde{M}_t}^{*} \|\mathbf{x} - \mathbf{x}_t\|_{\tilde{M}_t}.
\end{aligned}$$

This would then imply that the above optimality condition doesn't hold. We first lower bound the LHS of the above equation. Consider the following for any $\mathbf{x} \in F_t$ such that

$\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t} > c\eta_t$

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}) - \nabla R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1}(\mathbf{x}-\mathbf{x}_t), \mathbf{x}-\mathbf{x}_t\rangle \\
&\quad = \int_{s=0}^{1} (\mathbf{x}-\mathbf{x}_t)^T \left[\nabla^2 R(\mathbf{x}_t + s(\mathbf{x}-\mathbf{x}_t)) + \eta_{1:t-1}\hat{H}_{1:t-1}\right](\mathbf{x}-\mathbf{x}_t)ds \\
&\quad \overset{(a)}{\geq} \int_{s=0}^{\frac{c\eta_t}{\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t}}} (\mathbf{x}-\mathbf{x}_t)^T \left[\nabla^2 R(\mathbf{x}_t + s(\mathbf{x}-\mathbf{x}_t)) + \eta_{1:t-1}\hat{H}_{1:t-1}\right](\mathbf{x}-\mathbf{x}_t)ds \\
&\quad \overset{(b)}{\geq} \frac{c\eta_t}{\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t}}(\mathbf{x}-\mathbf{x}_t)^T \left[(1-c\eta_t)^2\nabla^2 R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1}\right](\mathbf{x}-\mathbf{x}_t),
\end{aligned}$$

where (a) uses the fact that $\nabla^2 R(\mathbf{x}) + \eta_{1:t-1}\hat{H}_{1:t-1}$ is a PSD matrix for any $\mathbf{x} \in F_t$ and (b) relies on property P1 of SCB stated in Equation (4.1). We further lower bound the RHS of the above equation as follows

$$\begin{aligned}
&(1-c\eta_t)^2\nabla^2 R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1} \\
&\quad = (1-c\eta_t)^2\nabla^2 R(\mathbf{x}_t) + \eta_{1:t-1}H_{1:t-1} + M_t - \tilde{M}_t \\
&\quad \succeq (1-c\eta_t)^2\tilde{M}_t - \tilde{M}_t^{1/2}\left[I - \tilde{M}_t^{-1/2}M_t\tilde{M}_t^{-1/2}\right]\tilde{M}_t^{1/2} \\
&\quad = \tilde{M}_t^{1/2}\left[(1-c\eta_t)^2 I - \left(I - \tilde{M}_t^{-1/2}M_t\tilde{M}_t^{-1/2}\right)\right]\tilde{M}_t^{1/2}.
\end{aligned}$$

Substituting this in the previous equation gives us

$$\begin{aligned}
&\langle \nabla R(\mathbf{x}) - \nabla R(\mathbf{x}_t) + \eta_{1:t-1}\hat{H}_{1:t-1}(\mathbf{x}-\mathbf{x}_t), \mathbf{x}-\mathbf{x}_t\rangle \\
&\quad \geq c\eta_t\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t}\lambda_{min}\left((1-c\eta_t)^2 I - \left(I - \tilde{M}_t^{-1/2}M_t\tilde{M}_t^{-1/2}\right)\right) \\
&\quad > \frac{c\eta_t}{2}\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t},
\end{aligned}$$

where the last inequality follows from the fact that $\|(I - \tilde{M}_t^{-1/2}M_t\tilde{M}_t^{-1/2}\|_2 \leq \frac{1}{4(1+8d\alpha)^2}$, and our choice of hyper-parameters. Next, consider the following

$$\begin{aligned}
(\|\hat{g}_t\|^*_{\tilde{M}_t})^2 &= \hat{g}_t^T\tilde{M}_t^{-1}\hat{g}_t \\
&= \lambda^{-2}d^2 f_t^2(\mathbf{y}_t)\mathbf{v}_{1,t}^T M_t^{1/2}\tilde{M}_t^{-1}M_t^{1/2}\mathbf{v}_{1,t} \\
&\leq 2\lambda^{-2}d^2(B+\epsilon)^2.
\end{aligned}$$

$$\begin{aligned}
(\|\hat{H}_t(\mathbf{x}-\mathbf{x}_t)\|^*_{\tilde{M}_t})^2 &= (\mathbf{x}-\mathbf{x}_t)^T\hat{H}_t\tilde{M}_t^{-1}\hat{H}_t(\mathbf{x}-\mathbf{x}_t)^T \\
&\leq \left(\frac{d^2 f_t(\mathbf{y}_t)}{\lambda^2}\right)^2\|\mathbf{x}-\mathbf{x}_t\|^2_{M_t}\|M_t^{1/2}\tilde{M}_t^{-1}M_t^{1/2}\|_2 \\
&\overset{(a)}{\leq} 32\lambda^{-4}d^6(B+\epsilon)^2\alpha^2,
\end{aligned}$$

where (a) follows from the fact that for any $\mathbf{x} \in F_t$, $\|\mathbf{x}-\mathbf{x}_t\|_{M_t} \leq 4d\alpha$. This shows that

$$\eta_t\|\hat{g}_t\|^*_{\tilde{M}_t} + \eta_t\|\hat{H}_t(\mathbf{x}-\mathbf{x}_t)\|^*_{\tilde{M}_t} \leq \frac{c\eta_t}{2}$$

This shows that $\mathbf{x}_{t+1}$ should satisfy $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\tilde{M}_t} \leq c\eta_t$.

To shows that $\tilde{M}_t$ and $\tilde{M}_{t+1}$ are spectrally close to each other, we rely on the closeness of $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ and use the same arguments as in the base case.

5. The last property that remains to be shown is that if $\iota_t = 0$, then $\iota_t = \iota_{t+1} = \cdots = \iota_{\mathcal{T}}$, $\mathbf{x}_t = \mathbf{x}_{t+1} \cdots = \mathbf{x}_{\mathcal{T}}$ and $F_t = F_{t+1} \cdots = F_{\mathcal{T}}$. We assume $\iota_{t-1} = 1$, since otherwise the property is trivially true. In this case, we know that $R_t(\mathbf{x})$ is strictly convex over F_t and so the Newton update in line 19 of Algorithm 4 has a unique minimizer.

   When $\iota_t = 0$, we have $\hat{g}_t = 0, \hat{H}_t = 0$. So the OMD update in line 19 of Algorithm 4 is given by $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in F_t} \Phi_{R_{t+1}}(\mathbf{x}, \mathbf{x}_t)$. Since $R_{t+1}(\mathbf{x}) = R_t(\mathbf{x})$ and $\mathbf{x}_t \in F_t$, it is easy to see that $\mathbf{x}_{t+1} = \mathbf{x}_t$. So the algorithm wouldn't make any progress in further rounds.

This finishes the proof of the lemma. □

We now show that the focus region doesn't get updated more than $12d \log T$ times.

Lemma 49 (Focus region updates)**.** *Consider the setting of Theorem 10. Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then the focus region gets updated no more than $12d \log T$ times before $\mathcal{T}$.*

Proof. The proof uses similar arguments as in Lemma 45. We prove the Lemma using contradiction. Assume that the focus region gets updated more than $12d \log T$ times before the algorithm restarts. Let $\tau < \mathcal{T}$ be the iteration where the focus region update happens for $12d \log T^{th}$ time. We now show that the restart condition should have triggered in iteration τ.

We have the following upper bound on the volume of $F_{\tau+1}$:

$$\mathrm{Vol}(F_{\tau+1}) \leq \mathrm{Vol}(F_\tau) \leq \frac{1}{T^{6d}} \mathrm{Vol}(\mathcal{X}_\xi).$$

This follows from the fact that the volume of the focus region reduces by a factor of $1/2$ whenever the focus region update condition triggers. In the rest of the proof, we show that if the volume of focus region is less than $\frac{1}{T^{6d}} \mathrm{Vol}(\mathcal{X}_\xi)$, then the restart condition should have triggered.

Step 1. First of all, for our choice of γ, we have $(1+\gamma)^{12d \log T} \leq 10$. Consequently, $\eta_\tau \leq 10\eta_1$. So the properties of the iterates we proved in Lemma 48 apply to our setting here. From this Lemma, we can infer that $\iota_\tau = 1$. Otherwise, we know that the focus region shouldn't have changed in the τ^{th} iteration (recall, in Lemma 48 we showed that if $\iota_\tau = 0$, then $F_\tau = F_{\tau+1}$). Moreoever, from this Lemma we can infer that $\forall t \leq \tau, \iota_t = 1$. So the cumulative loss estimate is close to the true cumulative loss and satisfies

$$\sup_{\mathbf{x} \in F_\tau} \left| \sum_{s=1}^{\tau-1} (\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - q_s(\mathbf{x}) + q_s(\mathbf{x}_s)) \right| \leq \frac{1}{\eta_1}.$$

Step 2. Let $\mathbf{u}_{\tau+1}$ be the minimizer of $\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x})$ over F_τ. Suppose $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right)\cap\mathcal{X}_\xi \subset F_\tau$. Then

$$\mathrm{Vol}(F_\tau) \geq \mathrm{Vol}\left(B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right)\cap\mathcal{X}_\xi\right).$$

Next, from our assumption that $\mathcal{X}$ contains a euclidean ball of radius 1, we can infer that $\mathcal{X}_\xi = \xi\mathbf{x}_1 + (1-\xi)\mathcal{X}$ contains a ball of radius $(1-\xi)$ in it. Let $\tilde{B}$ be the ball of radius $(1-\xi)$ that lies in $\mathcal{X}_\xi$. By convexity of $\mathcal{X}$ and the fact that the diameter of $\mathcal{X}$ is less than or equal to T, we have

$$\left(1-\frac{1}{T^3}\right)\mathbf{u}_{\tau+1} + \frac{1}{T^3}\tilde{B} \subseteq B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right)\cap\mathcal{X}_\xi.$$

This shows that $\mathrm{Vol}(F_\tau) \geq T^{-4d}\omega_d$, where ω_d is the volume of unit sphere in $\mathbb{R}^d$. Combining this with the previous upper bound on $\mathrm{Vol}(F_\tau)$, we get

$$T^{-4d}\omega_d, \leq \mathrm{Vol}(F_\tau) \leq T^{-6d}\mathrm{Vol}(\mathcal{X}) \overset{(a)}{\leq} T^{-5d}\omega_d,$$

where (a) follows from the fact that the diameter of $\mathcal{X}$ is upper bounded by T. We arrived at a contradiction. This shows that $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right)\cap\mathcal{X}_\xi \not\subset F_\tau$.

Step 3. Since $B\left(\mathbf{u}_{\tau+1}, \frac{1}{T^2}\right)\cap\mathcal{X}_\xi \not\subset F_\tau$, the following holds: $\exists \mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X}_\xi)$ such that $\|\mathbf{x}-\mathbf{u}_{\tau+1}\|_2 \leq \frac{1}{T^2}$. Now, consider the following for such an $\mathbf{x}$

$$\begin{aligned}\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) &= \sum_{s=1}^{\tau} q_s(\mathbf{x}) - q_s(\mathbf{u}_{\tau+1})\\ &\quad+\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) - q_s(\mathbf{x}) + q_s(\mathbf{u}_{\tau+1}).\end{aligned}$$

Since each q_s is T-Lipschitz, the first term in the RHS above is upper bounded by 1. Since the cumulative loss estimate is close to the true cumulative loss, the second term can be bounded as

$$\begin{aligned}\sum_{s=1}^{\tau}\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) - q_s(\mathbf{x}) + q_s(\mathbf{u}_{\tau+1}) &\leq \frac{2}{\eta_1} + \hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1}) - q_\tau(\mathbf{x}) + q_\tau(\mathbf{u}_{\tau+1})\\ &\overset{(a)}{\leq} \frac{2}{\eta_1} + 2B + \hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1}),\end{aligned}$$

where (a) follows from the fact that q_s is a bounded function. $\hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1})$ can be bounded as follows

$$\begin{aligned}\hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1}) &= \frac{1}{2}(\mathbf{x}-\mathbf{x}_\tau)^T\hat{H}_\tau(\mathbf{x}-\mathbf{x}_\tau) + \langle\hat{g}_\tau, \mathbf{x}-\mathbf{x}_\tau\rangle\\ &\quad-\frac{1}{2}(\mathbf{u}_{\tau+1}-\mathbf{x}_\tau)^T\hat{H}_\tau(\mathbf{u}_{\tau+1}-\mathbf{x}_\tau) - \langle\hat{g}_\tau, \mathbf{u}_{\tau+1}-\mathbf{x}_\tau\rangle\\ &\leq \frac{1}{2}\lambda^{-2}d^2(B+\epsilon)(\|\mathbf{x}-\mathbf{x}_\tau\|_{M_\tau}^2 + \|\mathbf{u}_{\tau+1}-\mathbf{x}_\tau\|_{M_\tau}^2)\\ &\quad+\lambda^{-1}d(B+\epsilon)(\|\mathbf{x}-\mathbf{x}_\tau\|_{M_\tau} + \|\mathbf{u}_{\tau+1}-\mathbf{x}_\tau\|_{M_\tau}).\end{aligned}$$

From Lemma 48, we know that for any $\mathbf{x} \in F_\tau$, $\|\mathbf{x} - \mathbf{x}_\tau\|_{M_\tau} \leq 4d\alpha$. Substituting this in the previous equation we get

$$\hat{f}_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{u}_{\tau+1}) \leq 16(B+\epsilon)\left(\lambda^{-2}d^4\alpha^2 + \lambda^{-1}d^2\alpha\right) \leq \frac{1}{\eta_1}.$$

This shows that $\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{u}_{\tau+1}) \leq \frac{4}{\eta_1}$. We now show that this implies the restart condition should have triggered. Consider the following

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) &= \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{u}_{\tau+1}) \\
&\leq \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x}) \\
&= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}\rangle - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s(\mathbf{x}-\mathbf{x}_s) \\
&= \frac{4}{\eta_1} + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}_{s+1}\rangle + \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle \\
&\quad - \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s(\mathbf{x}-\mathbf{x}_s) \\
&\overset{(a)}{\leq} \frac{4}{\eta_1} + 10\lambda^{-3}\alpha d^4 (B+\epsilon)^2 \sum_{s=1}^{\tau} \eta_s + \sum_{s=1}^{\tau} \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle \\
&\quad - \sum_{s=1}^{\tau} \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s(\mathbf{x}-\mathbf{x}_s),
\end{aligned}$$

where (a) follows from the stability of the iterates we proved in Lemma 48. Since $\mathbf{x}_{s+1}$ is the minimizer of $\min_{\mathbf{y}\in F_s} \eta_s\langle \hat{g}_s, \mathbf{y}\rangle + \Phi_{R_{s+1}}(\mathbf{y}, \mathbf{x}_s)$, we have the following from the first order optimality conditions

$$\langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x}\rangle \leq \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1}) - \Phi_{R_{s+1}}(\mathbf{x}_{s+1}, \mathbf{x}_s)}{\eta_s}.$$

Using this in the previous display, we get

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 10\lambda^{-3}\alpha d^4 (B+\epsilon)^2 \sum_{s=1}^{\tau} \eta_s + \sum_{s=1}^{\tau} \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1})}{\eta_s} \\
- \sum_{s=1}^{\tau} \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s(\mathbf{x}-\mathbf{x}_s).
\end{aligned}$$

Rearranging the terms in the RHS above, we get

$$\begin{aligned}
\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 10\lambda^{-3}\alpha d^4 (B+\epsilon)^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \frac{\Phi_{R_{\tau+1}}(\mathbf{x}, \mathbf{x}_{\tau+1})}{\eta_\tau} \\
+ \sum_{s=2}^{\tau} \left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right) \Phi_{R_s}(\mathbf{x}, \mathbf{x}_s).
\end{aligned}$$

Recall, $\mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X}_\xi)$. Let τ' be such that $\mathbf{x} \in \partial B_{\alpha, M_{\tau'}}(\mathbf{x}_{\tau'})$. Then

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 10\lambda^{-3}\alpha d^4 (B+\epsilon)^2 \sum_{s=1}^{\tau} \eta_s + \frac{\Phi_R(\mathbf{x}, \mathbf{x}_1)}{\eta_1} - \gamma \frac{\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})}{\eta_{\tau'}}.$$

Since $M_{\tau'}, \tilde{M}_{\tau'}$ are spectrally close to each other and since $\|\mathbf{x} - \mathbf{x}_{\tau'}\|_{M_{\tau'}} = \alpha$, we have $\|\mathbf{x} - \mathbf{x}_{\tau'}\|_{\tilde{M}_{\tau'}} \geq \alpha/2$. Using this, we now lower bound $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$

$$\begin{aligned}
\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'}) &= \Phi_R(\mathbf{x}, \mathbf{x}_{\tau'}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{\tau'})^T \left(\sum_{s=1}^{\tau'-1} \eta_s H_s \right) (\mathbf{x} - \mathbf{x}_{\tau'}) \\
&\quad + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{\tau'})^T \left(M_{\tau'} - \tilde{M}_{\tau'} \right) (\mathbf{x} - \mathbf{x}_{\tau'}) \\
&\overset{(a)}{\geq} \frac{\alpha}{2} - \log\left(1 + \frac{\alpha}{2}\right) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{\tau'})^T \left(M_{\tau'} - \tilde{M}_{\tau'} \right) (\mathbf{x} - \mathbf{x}_{\tau'}) \\
&\overset{(b)}{\geq} \frac{\alpha}{2} - \log\left(1 + \frac{\alpha}{2}\right) - \frac{\alpha}{20(1+8d\alpha)^2},
\end{aligned}$$

where (a) follows from property (P6) of SCB stated in Equation (C.20) and (b) follows from the fact that $M_{\tau'}, \tilde{M}_{\tau'}$ are spectrally close to each other. For our choice of α, $\Phi_{R_{\tau'}}(\mathbf{x}, \mathbf{x}_{\tau'})$ can be lower bounded by $\alpha/4$. We now upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$. Since $\mathbf{x} \in \mathcal{X}_\xi$, using property P8 of SCB stated in Appendix C.7, we can upper bound $\Phi_R(\mathbf{x}, \mathbf{x}_1)$ as

$$\Phi_R(\mathbf{x}, \mathbf{x}_1) = R(\mathbf{x}) \leq 4\nu \log T.$$

Substituting the above two bounds in the previous display and using the fact that $\eta_\tau \leq 10\eta_1$, we get

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1} + 100\lambda^{-3}\alpha d^4 (B+\epsilon)^2 \eta_1 T + \frac{4\nu \log T}{\eta_1} - \frac{\alpha\gamma}{20\eta_1} \leq -\frac{\beta}{\eta_1}.$$

This implies, the restart condition should have triggered. This shows that the focus region doesn't get updated more than $12d\log T$ times.

□

C.5.1 Proof of Proposition 6

In this section, we first show that the cumulative Hessian estimates and cumulative loss function estimates generated by the modified algorithm concentrate well around their expected values. In particular, Lemma 50 is concerned about concentration of the Hessian estimates $\{\hat{H}_t\}_{t=1}^T$, and Lemma 51 is concerned about loss estimates $\{\hat{f}_t\}_{t=1}^T$ of the modifed algorithm. These two Lemmas immediately imply that $\iota_t = 1$ for any $t \leq \mathcal{T}$ w.h.p, where $\mathcal{T}$ is the minimum between T and the first time at which the modified algorithm restarts. Consequently, with high probability, the iterates of the modified and the original algorithms are exactly the same. These two Lemmas together prove Proposition 6.

Before we proceed, note that the focus region gets updated at most $12d\log T$ times before the algorithm restarts. So, for our choice of γ, we have $(1+\gamma)^{12d\log T} \leq 10$. Consequently, for all $t \leq \mathcal{T}$, $\eta_t \leq 10\eta_1$. So the results of Lemma 48 apply to all the iterates in the first $\mathcal{T}$ iterations of the modified algorithm.

Lemma 50 (Concentration of Hessian estimates)**.** *Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then for any $t \leq \mathcal{T}$, the following statement holds with probability at least $1 - T^{-2}$*

$$\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 = O\left(\alpha^2\eta_1\lambda^{-2}d^5 B\sqrt{T\log(dT)}\right).$$

Proof. We first try to derive upper and lower bounds for $\tilde{M}_t$. From Lemma 48, we know that for all $s \leq \mathcal{T}$, and for all $\mathbf{x} \in F_s$, $\|\mathbf{x} - \mathbf{x}_s\|_{\tilde{M}_s} \leq 8d\alpha$. So, from Assumption 1 we have $\tilde{M}_t \succeq \frac{1}{(1+8d\alpha)^2}\tilde{M}_s$ for all $s \leq t$. This implies

$$\tilde{M}_t \succeq \frac{1}{(1+8d\alpha)^2}\tilde{M}_1 = \frac{1}{(1+8d\alpha)^2}\nabla^2 R(\mathbf{x}_1).$$

Moreover, from Lemma 46 we have $\tilde{M}_t \preceq T^8(\nu + 2\sqrt{\nu})^2(\nabla^2 R(\mathbf{x}_1) + I)$. Since $\nabla^2 R(\mathbf{x}_1)$ is a fixed quantity, for large enough T we have $\frac{1}{\text{poly}(T)}I \preceq \nabla^2 R(\mathbf{x}_1) \preceq \text{poly}(T)I$. This then shows that there exist positive constants c_l, c_u such that $T^{-c_l}I \preceq \tilde{M}_t \preceq T^{c_u}I$ for any $t \leq \mathcal{T}$.

Next consider the following

$$I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2} = \sum_{s=1}^{t-1} \eta_s \tilde{M}_t^{-1/2}\left(\iota_s H_s - \hat{H}_s\right)\tilde{M}_t^{-1/2}.$$

So we have

$$\begin{aligned}&\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \\ &\quad \leq \sup_{T^{-c_l}I \preceq A \preceq T^{c_u}I} \bar{\iota}_A \left\| \sum_{s=1}^{t-1} \eta_s A^{-1/2}\left(\hat{H}_s - \iota_s H_s\right) A^{-1/2}\mathbb{I}\left(A \succeq \frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\right\|_2,\end{aligned}$$

where $\bar{\iota}_A$ is an indicator random variable which is equal to 1 if and only if

$$\forall s \leq \mathcal{T}, \quad A \succeq \frac{1}{(1+8d\alpha)^2}\tilde{M}_s.$$

We now focus on bounding the RHS of the above equation. We write $\hat{H}_t$ as

$$\hat{H}_t = \hat{H}_{t,1} + \hat{H}_{t,2} = \frac{\lambda^{-2}}{2}d^2\iota_t(\underbrace{r_t(\mathbf{y}_t)}_{\hat{H}_{t,1}} + \underbrace{q_t(\mathbf{y}_t)}_{\hat{H}_{t,2}})M_t^{1/2}\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right)M_t^{1/2}$$

Now consider the RHS in the second-to-last display

$$\bar{\iota}_A\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,1}+\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$
$$\le\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\hat{H}_{s,1}A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$
$$+\bar{\iota}_A\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$

First consider the first term in the RHS above. We have

$$\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\hat{H}_{s,1}A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$
$$\le\sum_{s=1}^{t-1}\Big|\Big|\eta_s A^{-1/2}\hat{H}_{s,1}A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$

If $\iota_s=0$, then the s^{th} term in the RHs above is 0. On the other hand if $\iota_s=1$, then we know that $M_s,\tilde{M}_s$ are spectrally close to each other. In this case, the s^{th} term above is upper bounded by $20\epsilon\lambda^{-2}\eta_1 d^2(1+8d\alpha)^2$. This follows from the fact that $r_t(\mathbf{y}_t)$ is bounded by ϵ and $A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s$. So the RHS above is upper bounded by $20\epsilon\lambda^{-2}\eta_1 d^2(1+8d\alpha)^2T$. Now consider the second term

$$\bar{\iota}_A\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$
$$=\bar{\iota}_A\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\phi\left((1+8d\alpha)^2\lambda_{min}(\tilde{M}_s^{-1/2}A\tilde{M}_s^{-1/2})\right)\Big|\Big|_2,$$

where $\phi:\mathbb{R}_+\to\mathbb{R}_+$ is defined as

$$\phi(x)=\begin{cases}1 & \text{if } x\ge 1\\ 2x-1 & \text{if } 1>x>1/2\\ 0 & \text{if } \frac{1}{2}\ge x\ge 0\end{cases}.$$

Continuing, we get

$$\bar{\iota}_A\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\mathbb{I}\left(A\succeq\frac{1}{(1+8d\alpha)^2}\tilde{M}_s\right)\Big|\Big|_2$$
$$\le\Big|\Big|\sum_{s=1}^{t-1}\eta_s A^{-1/2}\left(\hat{H}_{s,2}-\iota_s H_s\right)A^{-1/2}\phi\left((1+8d\alpha)^2\lambda_{min}(\tilde{M}_s^{-1/2}A\tilde{M}_s^{-1/2})\right)\Big|\Big|_2.$$

So we have

$$\|I - \tilde{M}_t^{-1/2} M_t \tilde{M}_t^{-1/2}\|_2 \tag{C.11}$$

$$\leq \sup_{T^{-c_l} I \preceq A \preceq T^{c_u} I} \left\| \sum_{s=1}^{t-1} \eta_s A^{-1/2} \left(\hat{H}_{s,2} - \iota_s H_s \right) A^{-1/2} \phi \left((1+8d\alpha)^2 \lambda_{min}(\tilde{M}_s^{-1/2} A \tilde{M}_s^{-1/2}) \right) \right\|_2 \tag{C.12}$$

$$+ 20\epsilon\lambda^{-2}\eta_1 d^2 (1+8d\alpha)^2 T. \tag{C.13}$$

We now bound the first term in the RHS above using standard concentration results for matrix-valued martingales (see Lemma 43). Define random variable $Z_{A,s}$ as follows

$$Z_{A,s} = \begin{cases} \eta_s A^{-1/2} \left(\hat{H}_{s,2} - \iota_s H_s \right) A^{-1/2} \phi \left((1+8d\alpha)^2 \lambda_{min}(\tilde{M}_s^{-1/2} A \tilde{M}_s^{-1/2}) \right), & \text{if } s \leq \mathcal{T}, \\ 0 & \text{if } \mathcal{T} < s \leq T \end{cases}.$$

Note that $\{Z_{A,s}\}_{s=1}^{T}$ is a matrix-valued martingale difference sequence and satisfies $\mathbb{E}_t\left[Z_{A,t}\right] = 0$. Moreover, $Z_{A,s}$ is a bounded random variable which satisfies $\|Z_{A,s}\|_2 = O\left(\eta_1 \lambda^{-2} d^2 (1+8\alpha d)^2 B\right)$. This is easy to see when $\iota_s = 0$. When $\iota_s = 1$, it follows from the facts that $M_s, \tilde{M}_s$ are spectrally close to each other and $A \succeq \frac{1}{(1+8d\alpha)^2}\tilde{M}_s$ and $q_s(\mathbf{x})$ is bounded by B. By relying on standard concentration results for matrix martingale sequences, we get with probability at least $1-\delta$

$$\forall t \leq T, \quad \|\sum_{s=1}^{t} Z_{A,s}\|_2 \leq O\left(\alpha^2 \eta_1 \lambda^{-2} d^4 B \sqrt{T \log(2T/\delta)} \right). \tag{C.14}$$

We now do a union bound over all A such that $T^{-c_l} I \preceq A \preceq T^{c_u} I$. We first construct an Δ-net so that the following holds: for every A, there exists a A_Δ in the Δ-net such that $(1+(Td)^{-1})A_\Delta \succeq A \succeq (1-(Td)^{-1})A_\Delta$. We can show that the size of such an Δ-net is $\tilde{O}\left((Td)^{cd^2}\right)$, for some positive constant c. Moreover, we can show that for every A, there exists an A_Δ in the Δ-net such that

$$\|\sum_{s=1}^{t-1} Z_{A,s} - Z_{A_\Delta,s}\|_2 \leq \tilde{O}\left(\alpha^2 \eta_1 \lambda^{-2} d^4 B \sqrt{T} \right).$$

This follows from the fact that ϕ is bounded and Lipschitz. Now consider the following

$$\begin{aligned} \sup_{T^{-c_l} I \preceq A \preceq T^{c_u} I} \|\sum_{s=0}^{t-1} Z_{A,s}\|_2 &= \sup_{A} \|\sum_{s=0}^{t-1} Z_{A_\Delta,s}\|_2 + \sup_{A} \|\sum_{s=0}^{t-1} Z_{A_\Delta,s} - Z_{A,s}\|_2 \\ &\leq \sup_{A_\Delta \text{ in } \Delta\text{-net}} \|\sum_{s=0}^{t-1} Z_{A_\Delta,s}\|_2 + O\left(\alpha^2 \eta_1 \lambda^{-2} B d^4 \sqrt{T} \right), \end{aligned}$$

where A_Δ is the point in Δ-net which is closest to A. Finally, by relying on the bound in Equation (C.14) and performing a union bound over all the elements in the Δ-net gives us $\sup_A \|\sum_{s=0}^{t-1} Z_{A,s}\|_2 = \tilde{O}\left(\alpha^2 \eta_1 \lambda^{-2} d^5 B \sqrt{T} \right)$. Plugging this bound in Equation (C.11) and using the fact that $\epsilon = O\left(dBT^{-1/2}\right)$ gives us the required result. □

Remark C.5.1 (Convexifying the restart condition)**.** *We note that a similar argument as above can be used to show that the following two matrices are spectrally close to each other*

$$N_t = \nabla^2 R(\mathbf{x}_t) + \eta_1 (d\alpha)^2 \sum_{s=1}^{t-1} \hat{H}_s, \quad \tilde{N}_t = \nabla^2 R(\mathbf{x}_t) + \eta_1 (d\alpha)^2 \sum_{s=1}^{t-1} H_s.$$

In particular, we can show that $\|I - \tilde{N}_t^{-1/2} N_t \tilde{N}_t^{-1/2}\|_2 \leq \frac{1}{2}$. This would entail that N_t is invertible and positive definite. This in turn implies that the following objective is convex

$$\min_{\mathbf{y}\in F_t} \sum_{s=0}^{t} \hat{f}_s(\mathbf{y}) + (d^2\alpha^2\eta_1)^{-1}(\mathbf{y}-\mathbf{x}_t)^T \nabla^2 R(\mathbf{x}_t)(\mathbf{y}-\mathbf{x}_t).$$

Now consider the restart condition stated in line 16 of Algorithm 4. It involves solving $\min_{\mathbf{y}\in F_t} \sum_{s=0}^{t} \hat{f}_s(\mathbf{y})$. Note that this objective itself may not be convex. However, it is pointwise close to the above objective, which is convex. To see this, note that in Lemma 48 we showed that $\forall \mathbf{x} \in F_t$, $\|\mathbf{x}-\mathbf{x}_t\|_{\tilde{M}_t} \leq 8d\alpha$. As a result, $\forall \mathbf{x} \in F_t$, $(d^2\alpha^2\eta_1)^{-1}(\mathbf{y} - \mathbf{x}_t)^T \nabla^2 R(\mathbf{x}_t)(\mathbf{y}-\mathbf{x}_t) = O\left(\eta_1^{-1}\right)$. Consequently, the two objectives are $O\left(\frac{1}{\eta_1}\right)$ close to each other. So, one can efficiently check for an "approximate" restart condition by minimizing the above convex objective.

Lemma 51 (Concentration of loss estimates)**.** *Let $\mathcal{T}$ be the minimum between T and the first time at which the modified algorithm restarts. Then for any $t \leq \mathcal{T}$, the following statement holds with probability at least $1 - T^{-2}$*

$$\sup_{\mathbf{x}\in F_t} \left| \sum_{s=1}^{t-1} \eta_1 (\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) - \iota_s q_s(\mathbf{x}) + \iota_s q_s(\mathbf{x}_s)) \right| \leq \tilde{O}\left(\alpha^2 \eta_1 \lambda^{-2} B d^{9/2} \sqrt{T}\right).$$

Proof. First note that

$$\hat{f}_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_s (\mathbf{x}-\mathbf{x}_s) + \langle \hat{g}_s, \mathbf{x}-\mathbf{x}_s\rangle.$$

We split $\hat{H}_s, \hat{g}_s$ into two components, one corresponding to r_s and the other corresponding to q_s

$$\hat{H}_t = \frac{\lambda^{-2}}{2} d^2 \iota_t (\underbrace{r_t(\mathbf{y}_t)}_{\hat{H}_{t,1}} + \underbrace{q_t(\mathbf{y}_t)}_{\hat{H}_{t,2}}) M_t^{1/2}\left(\mathbf{v}_{1,t}\mathbf{v}_{2,t}^T + \mathbf{v}_{2,t}\mathbf{v}_{1,t}^T\right) M_t^{1/2}$$

$$\hat{g}_t = \lambda^{-1} d \iota_t (\underbrace{q_t(\mathbf{y}_t)}_{\hat{g}_{t,2}} + \underbrace{r_t(\mathbf{y}_t)}_{\hat{g}_{t,1}}) M_t^{1/2}\mathbf{v}_{1,t}.$$

Similarly, we define $\hat{r}_s(\mathbf{x})$ and $\hat{q}_s(\mathbf{x})$ as follows. These are obtained by splitting $\hat{f}_s(\mathbf{x})$ into two components based on r_s and q_s

$$\hat{r}_s(\mathbf{x}) - \hat{r}_s(\mathbf{x}_s) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_{s,1}(\mathbf{x}-\mathbf{x}_s) + \langle \hat{g}_{s,1}, \mathbf{x}-\mathbf{x}_s\rangle$$

$$\hat{q}_s(\mathbf{x}) - \hat{q}_s(\mathbf{x}_s) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_s)^T \hat{H}_{s,2}(\mathbf{x}-\mathbf{x}_s) + \langle \hat{g}_{s,2}, \mathbf{x}-\mathbf{x}_s\rangle.$$

We first upper bound $|\sum_{s=1}^{t-1} \hat{r}_s(\mathbf{x}) - \hat{r}_s(\mathbf{x}_s)|$. First note that from Lemma 48 we know that for any $\mathbf{x} \in F_t$, $\|\mathbf{x} - \mathbf{x}_t\|_{M_t} \leq 4d\alpha$. Using this, we have the following for any $\mathbf{x} \in F_t$.

$$\begin{aligned}
|\sum_{s=1}^{t-1} \hat{r}_s(\mathbf{x}) - \hat{r}_s(\mathbf{x}_s)| &\leq \sum_{s=1}^{t-1} |\hat{r}_s(\mathbf{x}) - \hat{r}_s(\mathbf{x}_s)| && \text{(C.15)}\\
&\leq 16\epsilon T(\lambda^{-2}d^4\alpha^2 + \lambda^{-1}d^2\alpha) && \text{(C.16)}\\
&\leq 32\alpha^2\epsilon\lambda^{-2}d^4T. && \text{(C.17)}
\end{aligned}$$

Next, we upper bound $|\sum_{s=1}^{t-1} \hat{q}_s(\mathbf{x}) - \hat{q}_s(\mathbf{x}_s) - \iota_s q_s(\mathbf{x}) + \iota_s q_s(\mathbf{x}_s)|$. Define random variables $Z_{\mathbf{x},s}$ as

$$Z_{\mathbf{x},s} = \begin{cases} \eta_1(\hat{q}_s(\mathbf{x}) - \hat{q}_s(\mathbf{x}_s) - \iota_s q_s(\mathbf{x}) + \iota_s q_s(\mathbf{x}_s)) & \text{if } s \leq \mathcal{T} \\ 0 \quad \text{otherwise} \end{cases}.$$

It is easy to see that $\{Z_{\mathbf{x},s}\}_{s=1}^{T}$ is a martingale difference sequence. Moreover, $Z_{\mathbf{x},s}$ is a bounded random variable which satisfies

$$|Z_{\mathbf{x},s}| \leq 32\alpha^2\eta_1\lambda^{-2}Bd^4.$$

This again follows from the fact that $\mathbf{x} \in F_t$, $\|\mathbf{x} - \mathbf{x}_t\|_{M_t} \leq 4d\alpha$ which we proved in Lemma 48. By relying on standard concentration bounds for martingale difference sequences (see Lemma 42), we get that with probability at least $1-\delta$,

$$\sup_{t \leq T} |\sum_{s=1}^{t-1} Z_{\mathbf{x},s}| = O\left(\lambda^{-2}d^4\alpha^2B\eta_1\sqrt{T\log T/\delta}\right).$$

Next, we bound $\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}|$ using Δ-net arguments. Let $\mathcal{N}_\Delta$ be an Δ-net over F_t which satisfies the following: for every $\mathbf{x}$, there exists a $\mathbf{x}_\Delta \in \mathcal{N}_\Delta$ such that $\|\mathbf{x} - \mathbf{x}_\Delta\|_{M_t} \leq \Delta$. Then

$$\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}| \leq \underbrace{\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\Delta,s}|}_{T_1} + \underbrace{\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\Delta,s} - Z_{\mathbf{x},s}|}_{T_2}. \tag{C.18}$$

Using a simple union bound, T_1 can be bounded as

$$T_1 \leq O\left(\lambda^{-2}d^4\alpha^2B\eta_1\sqrt{T\log T|\mathcal{N}_\Delta|/\delta}\right) \overset{(a)}{\leq} O\left(\lambda^{-2}d^{9/2}\alpha^2B\eta_1\sqrt{T\log\frac{\alpha dT}{\Delta\delta}}\right),$$

where the bound holds with probability at least $1-\delta$ and (a) holds since $\forall \mathbf{x} \in F_t, \|\mathbf{x} -$

$\mathbf{x}_t\|_{M_t} \leq 4d\alpha$ and as a result $|\mathcal{N}_\Delta| \leq \left(\frac{4d\alpha}{\Delta}\right)^d$. T_2 can be bounded as follows

$$\begin{aligned}
&\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1} Z_{\mathbf{x}_\Delta,s} - Z_{\mathbf{x},s}| \\
&\quad \overset{(a)}{\leq} \sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1}\eta_1\langle \hat{g}_{s,2} - \iota_s\nabla q_s(\mathbf{x}_s), \mathbf{x} - \mathbf{x}_\Delta\rangle| \\
&\qquad + \sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left|\sum_{s=0}^{t-1}\eta_1\langle \hat{H}_{s,2} - \iota_s H_s, (\mathbf{x}-\mathbf{x}_s)(\mathbf{x}-\mathbf{x}_s)^T - (\mathbf{x}_\Delta-\mathbf{x}_s)(\mathbf{x}_\Delta-\mathbf{x}_s)^T\rangle_F\right|
\end{aligned}$$

where (a) follows from the definitions of $Z_{\mathbf{x},s}$ and $q_s(\mathbf{x}), \hat{q}_s(\mathbf{x})$ and $\langle\cdot,\cdot\rangle_F$ is the frobenius inner product. The first term in the RHS above can be bounded as

$$\begin{aligned}
&\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=0}^{t-1}\eta_1\langle \hat{g}_{s,2} - \iota_s\nabla q_s(\mathbf{x}_s), \mathbf{x} - \mathbf{x}_\Delta\rangle| \\
&\quad \overset{(a)}{\leq} 2\eta_1\lambda^{-1}dB\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left(\sum_{s=0}^{t-1}\|\mathbf{x}-\mathbf{x}_\Delta\|_{M_s}\right) \\
&\quad \overset{(b)}{\leq} 2(1+8d\alpha)^2\eta_1\lambda^{-1}dB\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left(\sum_{s=0}^{t-1}\|\mathbf{x}-\mathbf{x}_\Delta\|_{M_t}\right) = O\left(\lambda^{-1}d^3\alpha^2B\eta_1\Delta T\right),
\end{aligned}$$

where (a) follows from the facts that $\|\hat{g}_{s,2}\|_{M_s}^* \leq \lambda^{-1}dB$, $\mathbb{E}_s[\hat{g}_{s,2}] = \iota_s\nabla q_s(\mathbf{x}_s)$, and (b) follows from Lemma 48 where we showed that $M_s \preceq (1+8d\alpha)^2 M_t$.

Using similar arguments and the fact that $\forall\mathbf{x}\in F_t, \|\mathbf{x}-\mathbf{x}_t\|_{M_t} \leq 4d\alpha$, the second term in the RHS of the second-to-last display can be bounded as

$$\begin{aligned}
&\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}\left|\sum_{s=0}^{t-1}\eta_1\langle \hat{H}_{s,2} - \iota_s H_s, (\mathbf{x}-\mathbf{x}_s)(\mathbf{x}-\mathbf{x}_s)^T - (\mathbf{x}_\Delta-\mathbf{x}_s)(\mathbf{x}_\Delta-\mathbf{x}_s)^T\rangle_F\right| \\
&\quad = O\left(\lambda^{-2}d^5\alpha^3B\eta_1\Delta T\right).
\end{aligned}$$

Choosing $\Delta = \frac{1}{\alpha\sqrt{dT}}$, and plugging the above bounds for T_1, T_2 in Equation (C.18) gives us $\sup_{\mathbf{x}\in F_t}\sup_{t\leq T}|\sum_{s=1}^{t-1} Z_{\mathbf{x},s}| = \tilde{O}\left(\alpha^2\eta_1\lambda^{-2}Bd^{9/2}\sqrt{T}\right)$. Finally, combining Equation (C.15) and Equation (C.18), and using the fact that $\epsilon = O\left(dBT^{-1/2}\right)$ gives us the requires result. □

Remark C.5.2. *For our choice of hyper-parameters, the concentration bounds in Lemmas 50, 51 show that the indicator random variables $\{\iota_t\}_{t=1}^T$ are equal to 1 with high probability. This entails that the iterates produced by the modified algorithm are exactly equal to the iterates produced by the actual algorithm with high probability. As a result all the properties we showed for the modified algorithm in Lemmas 48, 49, 50, 51 also hold for the original algorithm with high probability.*

C.5.2 Main argument for Theorem 10

We are now ready to prove Theorem 10. Since we know that with high probability, the iterates of the modified algorithm which relies on indicator variables ι_t are exactly same as the original algorithm, it suffices to prove the regret bound for the modified algorithm. In the sequel, we work with the modified algorithm. Throughout the proof, we let $\mathcal{T}$ be the minimum between T and the first time step at which the algorithm restarts. Let τ be the minimum between $\mathcal{T}$ and the last time step where $\iota_\tau = 1$. Our goal is to bound the following quantity

$$\sum_{s=1}^{\mathcal{T}} \iota_s f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{\mathcal{T}} \iota_s f_s(\mathbf{x}) = \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{\tau} f_s(\mathbf{x}).$$

Case 1 ($\mathcal{T} = T$). We first consider the case where the restart condition didn't trigger in the first T iterations (i.e., $\mathcal{T} = T$). In this case, we show that the regret is $\tilde{O}\left(T^{1/2}\right)$. Since the restart condition hasn't triggered, we know that

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \geq -\frac{\beta}{\eta_1}.$$

From the proof of Lemma 49, this implies $\forall \mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X}_\xi)$

$$\sum_{s=1}^{\tau} \hat{f}_s(\mathbf{x}) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} \hat{f}_s(\mathbf{y}) \geq \frac{4}{\eta_1}.$$

(In Lemma 49, we proved a contrapositive statement. We showed that if $\exists \mathbf{x} \in \partial F_T \cap \mathrm{int}(\mathcal{X}_\xi)$ such that $\sum_{s=0}^{T} \hat{f}_s(\mathbf{x}) - \min_{\mathbf{y}\in F_T} \sum_{s=0}^{T} \hat{f}_s(\mathbf{y}) \leq \frac{4}{\eta_1}$, then $\sum_{s=0}^{T} \hat{f}_s(\mathbf{x}_s) - \min_{\mathbf{y}\in F_T} \sum_{s=0}^{T} \hat{f}_s(\mathbf{y}) \leq -\frac{\beta}{\eta_1}$). Since our cumulative loss estimate concentrates well around the true cumulative loss (i.e., $\iota_\tau = 1$), this implies

$$\forall \mathbf{x} \in \partial F_\tau \cap \mathrm{int}(\mathcal{X}_\xi), \quad \sum_{s=1}^{\tau} q_s(\mathbf{x}) - \min_{\mathbf{y}\in F_\tau} \sum_{s=1}^{\tau} q_s(\mathbf{y}) \geq \frac{2}{\eta_1}.$$

Since q_s's are convex, this implies the minimizer of $\min_{\mathbf{x}\in\mathcal{X}_\xi} \sum_{s=1}^{\tau} q_s(\mathbf{x})$ is in F_τ. So, the regret of the algorithm can be bounded as follows

$$\begin{aligned}
\mathrm{Reg}_T = \sum_{s=1}^{\tau} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{\tau} f_s(\mathbf{x}) &\leq \epsilon T + \sum_{s=1}^{\tau} q_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{\tau} q_s(\mathbf{x}) \\
&\overset{(a)}{\leq} 1 + \epsilon T + \sum_{s=1}^{\tau} q_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}_\xi} \sum_{s=1}^{\tau} q_s(\mathbf{x}) \\
&= 1 + \epsilon T + \sum_{s=1}^{\tau} q_s(\mathbf{y}_s) - \min_{\mathbf{x}\in F_\tau} \sum_{s=1}^{\tau} q_s(\mathbf{x}),
\end{aligned}$$

where (a) follows from the definition of $\mathcal{X}_\xi = (1-\xi)\mathcal{X} + \xi\mathbf{x}_1$ and the fact that the loss functions are Lipschitz and the diameter of $\mathcal{X}$ is bounded. Next, consider the following for any $\mathbf{x} \in F_\tau$

$$\sum_{s=1}^{\tau} q_s(\mathbf{y}_s) - \sum_{s=1}^{\tau} q_s(\mathbf{x}) = \underbrace{\sum_{s=1}^{\tau} [q_s(\mathbf{y}_s) - q_s(\mathbf{x}_s)]}_{T_1} + \underbrace{\sum_{s=1}^{\tau} \left[q_s(\mathbf{x}_s) - q_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) + \hat{f}_s(\mathbf{x})\right]}_{T_2}$$
$$+ \underbrace{\sum_{s=1}^{\tau} \left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right]}_{T_3}.$$

Bounding T_1. Consider the following

$$\sum_{s=0}^{T} q_s(\mathbf{y}_s) - q_s(\mathbf{x}_s) \leq \sum_{s=0}^{T} \lambda\langle \nabla q_s(\mathbf{x}_s), M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})\rangle$$
$$+ \lambda^2 \frac{1}{2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2} H_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}).$$

Let $Z_s = \lambda\langle \nabla f_s(\mathbf{x}_s), M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})\rangle$ if $s \leq \tau$ and 0 if $s > \tau$. Note that $\{Z_s\}_{s=1}^T$ is a martingale difference sequence with each Z_s being bounded: $|Z_s| \leq 2dB$. This follows from the observation that $\nabla q_s(\mathbf{x}_s) = \mathbb{E}_s[\hat{g}_s]$ and the fact that $M_s^{-1/2}\hat{g}_s$ is a bounded random variable. By relying on standard concentration bounds for martingale difference sequences (see Lemma 42), we get that with probability at least $1-\delta$, $\sum_{s=1}^T Z_s = O\left(dB\sqrt{T\log 1/\delta}\right)$. Next, consider the last term in the RHS

$$\begin{aligned}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2} H_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}) &\leq 4\|M_s^{-1/2} H_s M_s^{-1/2}\|_2 \\ &\leq 4\|\tilde{M}_{s+1}^{-1/2} H_s \tilde{M}_{s+1}^{-1/2}\|_2 \|M_s^{-1/2}\tilde{M}_{s+1}M_s^{-1/2}\|_2 \\ &\leq 4\|\tilde{M}_{s+1}^{-1/2} H_s \tilde{M}_{s+1}^{-1/2}\|_2 \|M_s^{-1/2}\tilde{M}_s M_s^{-1/2}\|_2 \|\tilde{M}_s^{-1/2}\tilde{M}_{s+1}\tilde{M}_s^{-1/2}\|_2\end{aligned}$$

Since $\tilde{M}_s, M_s, M_{s+1}$ are spectrally close to each other, we can show that $\|M_s^{-1/2}\tilde{M}_s M_s^{-1/2}\|_2$, $\|\tilde{M}_s^{-1/2}\tilde{M}_{s+1}\tilde{M}_s^{-1/2}\|_2$ are close to 1. So we have

$$(\mathbf{v}_{1,s} + \mathbf{v}_{2,s})^T M_s^{-1/2} H_s M_s^{-1/2}(\mathbf{v}_{1,s} + \mathbf{v}_{2,s}) \leq 8\|\tilde{M}_{s+1}^{-1/2} H_s \tilde{M}_{s+1}^{-1/2}\|_2.$$

Using similar arguments as in the proof of Theorem 38 (see Equation (C.4)), we get the following upper bound for T_1: $O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right)$.

Bounding T_2. Since $\iota_\tau = 1$, T_2 can be upper bounded as

$$
\begin{aligned}
T_2 &\le \frac{1}{\eta_1} + \Big[q_\tau(\mathbf{x}_\tau) - q_\tau(\mathbf{x}) - \hat{f}_\tau(\mathbf{x}_\tau) + \hat{f}_\tau(\mathbf{x})\Big] \\
&\le \frac{1}{\eta_1} + + \langle \hat{g}_\tau - \nabla q_\tau(\mathbf{x}_\tau), \mathbf{x} - \mathbf{x}_\tau \rangle + \frac{1}{2}\langle \hat{H}_\tau - H_\tau, (\mathbf{x} - \mathbf{x}_\tau)(\mathbf{x} - \mathbf{x}_\tau)^T \rangle_F \\
&\le \frac{2}{\eta_1},
\end{aligned}
$$

where the last inequality follows from the facts that $\|\mathbf{x} - \mathbf{x}_\tau\|_{M_\tau} \le 4d\alpha$, $\|\hat{g}_\tau\|_{M_\tau}^* \le \lambda^{-1} d(B + \epsilon)$, $\|M_\tau^{-1/2} \hat{H}_\tau M_\tau^{-1/2}\|_2 \le \lambda^{-2} d^2 (B + \epsilon)$..

Bounding T_3. To bound T_3, we consider the following

$$
\begin{aligned}
\sum_{s=0}^{T} \Big[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\Big] &= \sum_{s=1}^{T} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x} \rangle - \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T \hat{H}_s (\mathbf{x} - \mathbf{x}_s) \\
&= \sum_{s=1}^{T} \langle \hat{g}_s, \mathbf{x}_s - \mathbf{x}_{s+1} \rangle + \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle - \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T \hat{H}_s (\mathbf{x} - \mathbf{x}_s) \\
&\overset{(a)}{\le} 10\lambda^{-3} \alpha d^4 (B + \epsilon)^2 \sum_{s=1}^{T} \eta_s + \sum_{s=1}^{T} \langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle \\
&\quad - \sum_{s=1}^{T} \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T \hat{H}_s (\mathbf{x} - \mathbf{x}_s),
\end{aligned}
$$

where (a) follows from the stability of the iterates we proved in Lemma 48. Since $\mathbf{x}_{s+1}$ is the minimizer of $\min_{\mathbf{y} \in F_s} \eta_s \langle \hat{g}_s, \mathbf{y} \rangle + \Phi_{R_{s+1}}(\mathbf{y}, \mathbf{x}_s)$, we have

$$
\langle \hat{g}_s, \mathbf{x}_{s+1} - \mathbf{x} \rangle \le \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1}) - \Phi_{R_{s+1}}(\mathbf{x}_{s+1}, \mathbf{x}_s)}{\eta_s}.
$$

Using this in the previous display, we get

$$
\begin{aligned}
\sum_{s=0}^{T} \Big[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\Big] \le 10\lambda^{-3} \alpha d^4 (B + \epsilon)^2 \sum_{s=1}^{T} \eta_s + \sum_{s=1}^{T} \frac{\Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_s) - \Phi_{R_{s+1}}(\mathbf{x}, \mathbf{x}_{s+1})}{\eta_s} \\
- \sum_{s=1}^{T} \frac{1}{2}(\mathbf{x} - \mathbf{x}_s)^T \hat{H}_s (\mathbf{x} - \mathbf{x}_s).
\end{aligned}
$$

Rearranging the terms in the RHS above, we get

$$\begin{aligned}\sum_{s=0}^{T}\left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right] &\leq 10\lambda^{-3}\alpha d^4(B+\epsilon)^2\sum_{s=1}^{T}\eta_s + \frac{\Phi_R(\mathbf{x},\mathbf{x}_1)}{\eta_1} - \frac{\Phi_{R_{T+1}}(\mathbf{x},\mathbf{x}_{T+1})}{\eta_T}\\ &\quad + \sum_{s=2}^{T}\left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right)\Phi_{R_s}(\mathbf{x},\mathbf{x}_s)\\ &\overset{(a)}{\leq} 10\lambda^{-3}\alpha d^4(B+\epsilon)^2\sum_{s=1}^{T}\eta_s + \frac{\Phi_R(\mathbf{x},\mathbf{x}_1)}{\eta_1},\end{aligned}$$

where (a) follows from the facts that R_s is convex, and $\eta_s \geq \eta_{s-1}$ for all s. Hence the last two terms are negatives and can be ignored. Since $\mathbf{x} \in \mathcal{X}_\xi$, using property P8 of SCB stated in Appendix C.7, we can upper bound $\Phi_R(\mathbf{x},\mathbf{x}_1)$ as

$$\Phi_R(\mathbf{x},\mathbf{x}_1) = R(\mathbf{x}) \leq 4\nu\log T.$$

Combining the bounds for T_1, T_2, T_3 shows that with probability at least $1-T^{-2}$ the regret is upper bounded by

$$\tilde{O}\left(\epsilon T + dB\sqrt{T} + \frac{(\nu+d)}{\eta_1} + \lambda^{-3}\alpha d^4(B+\epsilon)^2\eta_1 T\right) = \tilde{O}\left(d^{11}(d+\nu)^5\sqrt{T}\right).$$

Case 2 ($\mathcal{T} < T$). We now consider the case where the restart condition triggered at some iteration $\mathcal{T} < T$. Using the fact that the restart condition hasn't triggered in iteration $\mathcal{T}-1$ and using similar arguments as in the beginning of Case 1, we can again show that the minimizer of the cumulative loss over the entire domain lies in the focus region $F_{\mathcal{T}}$, and $\iota_{\mathcal{T}} = 1$. So regret until $\mathcal{T}$ is given by

$$\begin{aligned}\mathrm{Reg}_{\mathcal{T}} = \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}) &\overset{(a)}{\leq} 1 + \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in\mathcal{X}_\xi}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x})\\ &= 1 + \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \min_{\mathbf{x}\in F_{\mathcal{T}}}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}),\end{aligned}$$

where (a) follows from the definition of $\mathcal{X}_\xi$. Using the same regret decomposition as in Case 1, for any $\mathbf{x} \in F_{\mathcal{T}}$

$$\begin{aligned}\sum_{s=1}^{\mathcal{T}} f_s(\mathbf{y}_s) - \sum_{s=1}^{\mathcal{T}} f_s(\mathbf{x}) \leq \epsilon T + \underbrace{\sum_{s=1}^{\mathcal{T}}\left[q_s(\mathbf{y}_s) - q_s(\mathbf{x}_s)\right]}_{T_1} + \underbrace{\sum_{s=1}^{\mathcal{T}}\left[q_s(\mathbf{x}_s) - q_s(\mathbf{x}) - \hat{f}_s(\mathbf{x}_s) + \hat{f}_s(\mathbf{x})\right]}_{T_2}\\ + \underbrace{\sum_{s=1}^{\mathcal{T}}\left[\hat{f}_s(\mathbf{x}_s) - \hat{f}_s(\mathbf{x})\right]}_{T_3}.\end{aligned}$$

We use the same arguments as in Case 1 to bound T_1, T_2 as

$$T_1 = O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right), \quad T_2 = \frac{2}{\eta_1}.$$

Since the restart condition triggered in round $\mathcal{T}$, T_3 is bounded by $-\frac{\beta}{\eta_1}$. Combining all these bounds, we get the following bound on regret

$$\text{Reg}_{\mathcal{T}} \leq \epsilon T + O\left(dB\sqrt{T\log 1/\delta} + \frac{d\log dT}{\eta_1}\right) + \frac{2}{\eta_1} - \frac{\beta}{\eta_1}.$$

For our choice of hyper-parameters, the above bound is less than 0.

C.6 Additional Results

Proposition 16 (Gaussian Smoothing). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a potentially non-smooth function. Define the smoothed function $\hat{f}$ as $\hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,I)}\left[f(\mathbf{x}+C\mathbf{u})\right]$, for some symmetric positive definite matrix C. Then $\hat{f}$ is twice differentiable with the following gradient and Hessian*

$$\nabla\hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,I)}\left[C^{-1}\mathbf{u}f(\mathbf{x}+C\mathbf{u})\right], \ \nabla^2\hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,I)}\left[C^{-1}(\mathbf{u}\mathbf{u}^T - I)C^{-1}f(\mathbf{x}+C\mathbf{u})\right].$$

Proof. **Gradient.** Using the expression for probability density function of a multivariate Gaussian, we get

$$\begin{aligned}
\nabla\hat{f}(\mathbf{x}) &= \frac{\partial}{\partial\mathbf{x}}\int\frac{1}{(2\pi)^{d/2}}f(\mathbf{x}+C\mathbf{u})e^{-\|\mathbf{u}\|^2/2}d\mathbf{u} \overset{(a)}{=} \frac{\partial}{\partial\mathbf{x}}\int\frac{1}{(2\pi|C|^2)^{d/2}}f(\mathbf{y})e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2}d\mathbf{y} \\
&= \int\frac{\partial}{\partial\mathbf{x}}\frac{1}{(2\pi|C|^2)^{d/2}}f(\mathbf{y})e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2}d\mathbf{y} = \int\frac{C^{-2}(\mathbf{y}-\mathbf{x})}{(2\pi|C|^2)^{d/2}}f(\mathbf{y})e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2}d\mathbf{y} \\
&\overset{(b)}{=} \int\frac{C^{-1}\mathbf{u}}{(2\pi)^{d/2}}f(\mathbf{x}+C\mathbf{u})e^{-\|\mathbf{u}\|^2/2}d\mathbf{u},
\end{aligned}$$

where we used change of variables in (a) and (b). This shows that

$$\nabla\hat{f}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,I)}\left[C^{-1}\mathbf{u}f(\mathbf{x}+C\mathbf{u})\right].$$

Hessian. We use a similar argument as above to compute the Hessian. From the first display above, we have

$$\nabla\hat{f}(\mathbf{x}) = \int\frac{C^{-2}(\mathbf{y}-\mathbf{x})}{(2\pi|C|^2)^{d/2}}f(\mathbf{y})e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2}d\mathbf{y}.$$

Using the definition of Hessian, we get

$$\begin{aligned}
\nabla^2 \hat{f}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \nabla \hat{f}(\mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}} \int \frac{C^{-2}(\mathbf{y}-\mathbf{x})}{(2\pi|C|^2)^{d/2}} f(\mathbf{y}) e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2} d\mathbf{y} \\
&= \int \frac{\partial}{\partial \mathbf{x}} \frac{C^{-2}(\mathbf{y}-\mathbf{x})}{(2\pi|C|^2)^{d/2}} f(\mathbf{y}) e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2} d\mathbf{y} \\
&= \int \frac{C^{-2}(\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^T C^{-2} - C^{-2}}{(2\pi|C|^2)^{d/2}} f(\mathbf{y}) e^{-\|\mathbf{y}-\mathbf{x}\|^2_{C^{-2}}/2} d\mathbf{y} \\
&\overset{(a)}{=} \int \frac{C^{-1}\mathbf{u}\mathbf{u}^T C^{-1} - C^{-2}}{(2\pi|C|^2)^{d/2}} f(\mathbf{x}+C\mathbf{u}) e^{-\|u\|^2/2} d\mathbf{u}
\end{aligned}$$

where we used change of variables in (a). This shows that

$$\mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,I)}\left[C^{-1}(\mathbf{u}\mathbf{u}^T - I)C^{-1} f(\mathbf{x}+C\mathbf{u})\right].$$

□

C.7 Review of Self Concordant Barriers

This section reviews some useful properties of Self Concordant (SC) functions and Self Concordant Barriers (SCBs). Most of the content in this section is from Nemirovski [Nem04] and Nesterov [Nes18].

- **(P3)** *Non-degeneracy*: If $\mathcal{X}$ doesn't contain straight lines, then the Hessian $\nabla^2 R(\mathbf{x})$ is nondegenerate (*i.e.,* $\nabla^2 R(\mathbf{x}) \succ 0$) at all points $\mathbf{x} \in \operatorname{int}(\mathcal{X})$.
- **(P4)** For any $\mathbf{x} \in \operatorname{int}(\mathcal{X})$, we have
$$\mathcal{X} \cap \{\mathbf{y} : \langle \nabla R(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq 0\} \subseteq B_{\nu+2\sqrt{\nu}, \nabla^2 R(\mathbf{x})}(\mathbf{x}). \tag{C.19}$$
- **(P5)** *Semiboundedness*: For any $\mathbf{x} \in \operatorname{int}(\mathcal{X}), \mathbf{y} \in \mathcal{X}$, $\langle \nabla R(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \nu$.
- **(P6)** For any $\mathbf{x}, \mathbf{y} \in \operatorname{int}(\mathcal{X})$,
$$R(\mathbf{y}) - R(\mathbf{x}) - \langle \nabla R(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq \|\mathbf{y}-\mathbf{x}\|_{\nabla^2 R(\mathbf{x})} - \log(1 + \|\mathbf{y}-\mathbf{x}\|_{\nabla^2 R(\mathbf{x})}). \tag{C.20}$$
- **(P7)** For any $\mathbf{x}, \mathbf{y} \in \operatorname{int}(\mathcal{X})$, we have
$$\langle \nabla R(\mathbf{y}) - \nabla R(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq \frac{\|\mathbf{y}-\mathbf{x}\|^2_{\nabla^2 R(\mathbf{x})}}{1 + \|\mathbf{y}-\mathbf{x}\|_{\nabla^2 R(\mathbf{x})}}. \tag{C.21}$$
- **(P8)** Define the Minkowsky function of $\mathcal{X}$ with the pole at $\mathbf{x}$ as
$$\pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t > 0 | \mathbf{x} + t^{-1}(\mathbf{y}-\mathbf{x}) \in \mathcal{X}\}.$$
Then for any $\mathbf{x}, \mathbf{y} \in \operatorname{int}(\mathcal{X})$
$$R(\mathbf{y}) \leq R(\mathbf{x}) + \nu \log \frac{1}{1 - \pi_{\mathbf{x}}(\mathbf{y})} \tag{C.22}$$
$$\nabla^2 R(\mathbf{y}) \preceq \left(\frac{\nu + 2\sqrt{\nu}}{1 - \pi_{\mathbf{x}}(\mathbf{y})}\right)^2 \nabla^2 R(\mathbf{x}). \tag{C.23}$$

Appendix D

Supplementary Material for Chapter 5

D.1 Measurability of Bayes Estimators

For any prior Π, define $p_\Pi(\mathbb{X}^n)$ as

$$\int_\theta \prod_{i=1}^n p(X_i;\theta) d\Pi(\theta).$$

For any prior Π, define estimator $\hat{\theta}_\Pi$ as follows

$$\hat{\theta}_\Pi(\mathbb{X}^n) \in \operatorname*{argmin}_{\tilde{\theta}\in\Theta} \mathbb{E}_{\theta\sim\Pi(\cdot|\mathbb{X}^n)}\left[M(\tilde{\theta},\theta)\right].$$

Certain regularity conditions need to hold for this to be a Bayes estimator of Π. $\hat{\theta}_\Pi$ defined this way need not be a measurable function of $\mathbb{X}^n$. We now provide sufficient conditions on the statistical problem which guarantee measurability of $\hat{\theta}_\Pi$. These conditions are from Brown and Purves [BP73].

Assumption 2. *The sample space $\mathcal{X}^n$ and the parameter set Θ are non-empty Borel sets.*

Assumption 3. *Let $\mathcal{B}(\mathcal{X}^n)$ be the Borel σ-algebra corresponding to the sample space $\mathcal{X}^n$ and $\mathcal{B}(\Theta)$ be the Borel σ-algebra corresponding to parameter space Θ. Let Π be a prior probability measure on Θ. Suppose, for each $\theta \in \Theta$, P_θ is such that, for each $B \in \mathcal{B}(\mathcal{X}^n)$, the function $\theta \to P_\theta(B)$ is measurable w.r.t $\mathcal{B}(\Theta)$.*

Assumption 4. *The loss function M defined on $\Theta\times\Theta$ and taking non-negative real values, is measurable w.r.t $\mathcal{B}(\Theta) \times \mathcal{B}(\Theta)$. Moreover, $M(\cdot,\theta)$ is lower semi-continuous on Θ, for each $\theta \in \Theta$.*

Under these assumptions, when Θ is compact, Brown and Purves [BP73] show that there exists a Borel measurable function $\hat{\theta}_\Pi$ such that

$$\hat{\theta}_\Pi(\mathbb{X}^n) \in \operatorname*{argmin}_{\tilde{\theta}\in\Theta} \mathbb{E}_{\theta\sim\Pi(\cdot|\mathbb{X}^n)}\left[M(\tilde{\theta},\theta)\right].$$

Moreover, $\hat{\theta}_\Pi$ is the Bayes estimator for Π.

D.2 Minimax Estimators, LFPs and Nash Equilibirium

Proposition 17. *Consider the statistical game in Equation (5.1). If $(\hat{\theta}^*, P^*)$ is a mixed strategy NE of (5.1), then the minmax and maxmin values of the linearized game are equal to each other. Moreover, $\hat{\theta}^*$ is a minimax estimator and P^* is an LFP. Conversely, if $\hat{\theta}^*$ is a minimax estimator, and P^* is an LFP, and the minmax and maxmin values of the linearized game (5.4) are equal to each other, then $(\hat{\theta}^*, P^*)$ is a mixed strategy NE of (5.1). Moreover, θ^* is a Bayes estimator for P^*.*

Proof. Suppose $(\hat{\theta}^*, P^*)$ is a mixed strategy NE. Then, from the definition of mixed strategy NE, we have

$$\sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P) \leq R(\hat{\theta}^*, P^*) \leq \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*).$$

This further implies

$$\begin{aligned}\inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}, P) &\overset{(a)}{\leq} \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P) \leq R(\hat{\theta}^*, P^*) \\ &\overset{(b)}{\leq} \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*) \overset{(c)}{\leq} \sup_{P \in \mathcal{M}_\Theta} \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P).\end{aligned}$$

Since $\inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}, P) \geq \sup_{P \in \mathcal{M}_\Theta} \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P)$, the above set of inequalities all hold with an equality and imply that the minmax and maxmin values of the linearized game are equal to each other. Moreover, from (a), we have $\sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P) = \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}, P)$. This implies $\hat{\theta}^*$ is a minimax estimator. From (c), we have $\inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*) = \sup_{P \in \mathcal{M}_\Theta} \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P)$. This implies P^* is an LFP. Finally, from (b), we have $R(\hat{\theta}^*, P^*) = \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*)$. This implies $\hat{\theta}^*$ is a Bayes estimator for P^*.

We now prove the converse. Since $\hat{\theta}^*$ is a minimax estimator and P^* is an LFP, we have

$$\sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P) = \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}, P), \qquad \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*) = \sup_{P \in \mathcal{M}_\Theta} \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P).$$

Moreover, since minmax and maxmin values of the linearized game are equal to each other, all the above 4 quantities are equal to each other. Since $R(\hat{\theta}^*, P^*) \leq \sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P)$ and $R(\hat{\theta}^*, P^*) \geq \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*)$, we have

$$\sup_{P \in \mathcal{M}_\Theta} R(\hat{\theta}^*, P) = R(\hat{\theta}^*, P^*) = \inf_{\hat{\theta} \in \mathcal{M}_\mathcal{D}} R(\hat{\theta}, P^*).$$

This shows that $(\hat{\theta}^*, P^*)$ is a mixed strategy NE of the linear game in Equation (5.4). □

D.3 Minimax Estimation via Online Learning

D.3.1 Proof of Proposition 7

We have the following bounds on the regret of the minimization and maximization players

$$\sum_{t=1}^{T} R(\hat{\theta}_t, P_t) - \inf_{\hat{\theta}\in\mathcal{D}} \sum_{t=1}^{T} R(\hat{\theta}, P_t) \leq \epsilon_1(T),$$
$$\sup_{\theta\in\Theta} \sum_{t=1}^{T} R(\hat{\theta}_t, \theta) - \sum_{t=1}^{T} R(\hat{\theta}_t, P_t) \leq \epsilon_2(T).$$

Now consider the following

$$\begin{aligned}
&\inf_{\hat{\theta}\in\mathcal{D}} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}, P_t)\\
&\quad \geq \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}_t, P_t) - \frac{\epsilon_1(T)}{T} \qquad\qquad (\text{D.1})\\
&\quad \geq \sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}_t, \theta) - \frac{\epsilon_1(T)+\epsilon_2(T)}{T},
\end{aligned}$$

where the first and the second inequalities follow from the regret bounds of the minimization and maximization players. We further bound the LHS and RHS of the above inequality as follows

$$\inf_{\hat{\theta}\in\mathcal{D}} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}, P_t) \leq \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} R(\hat{\theta}_{t'}, P_t) = R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}),$$
$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}_t, \theta) \geq \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} R(\hat{\theta}_{t'}, P_t) = R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}).$$

Combining the previous two sets of inequalities gives us

$$R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}) \geq \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND}}, \theta) - \frac{\epsilon_1(T)+\epsilon_2(T)}{T},$$
$$R(\hat{\theta}_{\text{RND}}, P_{\text{AVG}}) \leq \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG}}) + \frac{\epsilon_1(T)+\epsilon_2(T)}{T}.$$

D.3.2 Proof of Theorem 11

To prove the Theorem we first bound the regret of each player and then rely on Proposition 7 to show that the iterates converge to a NE. Since the maximization player is responding using FTPL to the actions of minimization player, we rely on Theorem 1 to bound her regret. First note that the sequence of reward functions seen by the maximization player $R(\hat{\theta}_i, \cdot)$ are L-Lipschitz. Moreover, the domain Θ has ℓ_∞ diameter of D. So applying Theorem 1 gives us the following regret bound

$$\mathbb{E}_\sigma\left[\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}R(\hat{\theta}_t,\theta)-\frac{1}{T}\sum_{t=1}^{T}R(\hat{\theta}_t,\theta_t(\sigma))\right]\leq O\left(\eta d^2DL^2+\frac{d(\beta T+D)}{\eta T}+\alpha+\beta dL\right).$$

Taking the expectation inside, we get the following

$$\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}R(\hat{\theta}_t,\theta)-\frac{1}{T}\sum_{t=1}^{T}R(\hat{\theta}_t,P_t)\leq O\left(\eta d^2DL^2+\frac{d(\beta T+D)}{\eta T}+\alpha+\beta dL\right). \tag{D.2}$$

Since the minimization player is using BR, her regret is upper bounded by 0. Plugging in these two regret bounds in Proposition 7 gives us the required result.

D.3.3 Proof of Corollary 3

Note that this corollary is only concerned about existence of minimax estimators and LFPs, and showing that minmax and maxmin values of Equation (5.4) are equal to each other. So we can ignore the approximation errors introduced by the oracles and set $\alpha=\beta=\alpha'=0$ in the results of Theorem 11 (that is, we assume access to exact optimization oracles, as we are only concerned with existence of NE and not about computational tractability of the algorithm).

Minimax Theorem To prove the first part of the corollary, we set $\eta=\sqrt{\frac{1}{dL^2T}}$ in Theorem 11 and let $T\to\infty$. We get

$$\begin{aligned}&\sup_{\theta\in\Theta}R(\hat{\theta}_{\text{RND}},\theta)=\inf_{\hat{\theta}\in\mathcal{D}}R(\hat{\theta},P_{\text{AVG}})\\ \implies&\sup_{P\in\mathcal{M}_\Theta}R(\hat{\theta}_{\text{RND}},P)=\inf_{\hat{\theta}\in\mathcal{M}_\mathcal{D}}R(\hat{\theta},P_{\text{AVG}})\\ \implies&\inf_{\hat{\theta}\in\mathcal{M}_\mathcal{D}}\sup_{P\in\mathcal{M}_\Theta}R(\hat{\theta},P)\leq\sup_{P\in\mathcal{M}_\Theta}\inf_{\hat{\theta}\in\mathcal{M}_\mathcal{D}}R(\hat{\theta},P).\end{aligned}$$

Since minmax value of any game is always greater than or equal to maxmin value of the game, we get

$$\inf_{\hat{\theta}\in\mathcal{M}_\mathcal{D}}\sup_{P\in\mathcal{M}_\Theta}R(\hat{\theta},P)=\sup_{P\in\mathcal{M}_\Theta}\inf_{\hat{\theta}\in\mathcal{M}_\mathcal{D}}R(\hat{\theta},P)R^*.$$

Existence of LFP We now show that the statistical game has an LFP. To prove this result, we make use of the following result on the compactness of probability spaces. If Θ is a compact space, then $\mathcal{M}_\Theta$ is sequentially compact; that is, any sequence $P_n\in\mathcal{M}_\Theta$ has a convergent subsequence converging to a point in $\mathcal{M}_\Theta$ (the notion of convergence here is weak convergence). Let $P_{\text{AVG},t}=\frac{1}{t}\sum_{i=1}^{t}P_i$ be the mixture distribution obtained from the first t iterates of Algorithm 5 when run with $\eta=\sqrt{\frac{1}{dL^2T}}$ and exact optimization oracles.

Consider the sequence of probability measures $\{P_{\text{AVG},t}\}_{t=1}^{\infty}$. Since the parameter space Θ is compact, we know that there exists a converging subsequence $\{P_{\text{AVG},t_i}\}_{i=1}^{\infty}$. Let $P^* \in \mathcal{M}_\Theta$ be the limit of this sequence. In the rest of the proof, we show that P^* is an LFP; that is, $\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P^*) = R^*$. Since $R(\hat{\theta}, \theta)$ is bounded, and Lipschitz in its second argument, we have

$$\forall \hat{\theta} \in \mathcal{M}_{\mathcal{D}} \quad \lim_{i\to\infty} R(\hat{\theta}, P_{\text{AVG},t_i}) = R(\hat{\theta}, P^*). \tag{D.3}$$

This follows from the equivalent formulations of weak convergence of measures. We now make use of the following result from Corollary 4 (which we prove later in Appendix D.3.4)

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG},t}) \geq R^* - O(t^{-\frac{1}{2}}).$$

Combining this with the fact that $\sup_{P\in\mathcal{M}_\Theta} \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P) = R^*$, we get

$$\lim_{i\to\infty} \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG},t_i}) = R^*. \tag{D.4}$$

Equations (D.3), (D.4) show that $\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG},t_i})$, $R(\tilde{\theta}, P_{\text{AVG},t_i})$ are converging sequences as $i \to \infty$. Since $\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG},t_i}) \leq R(\tilde{\theta}, P_{\text{AVG},t_i})$ for all $i, \tilde{\theta} \in \mathcal{D}$, we have

$$\lim_{i\to\infty} \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P_{\text{AVG},t_i}) \leq \lim_{i\to\infty} R(\tilde{\theta}, P_{\text{AVG},t_i}), \quad \forall \tilde{\theta} \in \mathcal{D}.$$

From Equations (D.3), (D.4), we then have

$$\begin{aligned} R^* &\leq R(\tilde{\theta}, P^*), \quad \forall \tilde{\theta} \in \mathcal{D} \\ \implies R^* &\leq \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P^*), \end{aligned}$$

Combining this with the fact that $\sup_{P\in\mathcal{M}_\Theta} \inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P) = R^*$, we get

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, P^*) = R^*.$$

This shows that P^* is an LFP.

Existence of Minimax Estimator To show the existence of a minimax estimator, we make use of the following result from Wald [Wal49], which is concerned about the "compactness" of the space of estimators $\mathcal{M}_{\mathcal{D}}$.

Proposition 18. *Suppose Θ is compact w.r.t $\Delta_M(\theta_1, \theta_2) = \sup_{\theta\in\Theta} |M(\theta_1, \theta) - M(\theta_2, \theta)|$. Moreover, suppose the risk R is bounded. Then for any sequence of $\{\hat{\theta}_i\}_{i=1}^{\infty}$ of estimators there exists a subsequence $\{\hat{\theta}_{i_j}\}_{j=1}^{\infty}$ such that $\lim_{j\to\infty} \hat{\theta}_{i_j} = \hat{\theta}_0$ and for any $\theta \in \Theta$*

$$\lim_{i\to\infty} \inf R(\hat{\theta}_{i_j}, \theta) \geq R(\hat{\theta}_0, \theta).$$

Let $\hat{\theta}_{\text{RND},t}$ be the randomized estimator obtained by uniformly sampling an estimator from $\{\hat{\theta}_i\}_{i=1}^t$. Consider the sequence of estimators $\{\hat{\theta}_{\text{RND},t}\}_{t=1}^\infty$. From the above proposition, we know that there exists a subsequence $\{\hat{\theta}_{\text{RND},t_j}\}_{j=1}^\infty$ and an estimator $\hat{\theta}^*$ such that $\liminf_{j\to\infty} R(\hat{\theta}_{\text{RND},t_j}, \theta) \geq R(\hat{\theta}^*, \theta)$. We now show that $\hat{\theta}^*$ is a minimax estimator; that is, we show that $\sup_{\theta\in\Theta} R(\hat{\theta}^*, \theta) = R^*$. We make use of the following result from Corollary 4

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t}, \theta) \leq R^* + O(t^{-\frac{1}{2}}).$$

Combining this with the fact that $\inf_{\hat{\theta}\in\mathcal{D}} \sup_{P\in\mathcal{M}_\Theta} R(\hat{\theta}, P) = R^*$, we get

$$\lim_{j\to\infty} \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \theta) = R^*. \tag{D.5}$$

Since $\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \theta) \geq R(\hat{\theta}_{\text{RND},t_j}, \tilde{\theta})$ for any $j, \tilde{\theta} \in \Theta$, we have

$$\liminf_{j\to\infty} \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \theta) \geq \liminf_{j\to\infty} R(\hat{\theta}_{\text{RND},t_j}, \tilde{\theta}) \geq R(\hat{\theta}^*, \theta), \quad \forall\tilde{\theta} \in \Theta.$$

Since $\{R(\hat{\theta}_{\text{RND},t_j}, \theta)\}_{j=1}^\infty$ is a converging sequence, we have

$$\liminf_{j\to\infty} \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \theta) = \lim_{j\to\infty} \sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \theta) = R^*.$$

This together with the previous inequality gives us $\sup_{\tilde{\theta}\in\Theta} R(\hat{\theta}_{\text{RND},t_j}, \tilde{\theta}) \leq R^*$. This shows that θ^* is a minimax estimator.

D.3.4 Proof of Corollary 4

Minimax Estimator From Theorem 11 we have

$$\begin{aligned}
\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{RND}}, \theta) &= \sup_{\theta\in\Theta} \frac{1}{T} \sum_{i=1}^T R(\hat{\theta}_i, \theta) \\
&\leq \inf_{\hat{\theta}\in\mathcal{D}} \frac{1}{T} \sum_{i=1}^T R(\hat{\theta}, P_i) + O\left(\eta d^2 D L^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta d L\right) \\
&= \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \frac{1}{T} \sum_{i=1}^T R(\hat{\theta}, P_i) + O\left(\eta d^2 D L^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta d L\right) \\
&\overset{(a)}{\leq} \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \sup_{P\in\mathcal{M}_\Theta} R(\hat{\theta}, P) + O\left(\eta d^2 D L^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta d L\right),
\end{aligned}$$

where (a) follows from the fact that $\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) \geq \frac{1}{T}\sum_{i=1}^T R(\hat{\theta}, P_i)$. Substituting $\eta = \sqrt{\frac{1}{dL^2T}}$ in the above equation shows that the randomized estimator is approximately minimax. This completes the first part of the proof. If the metric M is convex in its first argument, then from Jensen's inequality we have

$$\forall\theta, \quad R(\hat{\theta}_{\text{AVG}}, \theta) \leq R(\hat{\theta}_{\text{RND}}, \theta).$$

This shows that the worst-case risk of $\hat{\theta}_{\text{AVG}}$ is upper bounded as

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{AVG}},\theta) \leq \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \sup_{P\in\mathcal{M}_{\Theta}} R(\hat{\theta},\theta) + O\left(\eta d^2 DL^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta dL\right). \tag{D.6}$$

Substituting $\eta = \sqrt{\frac{1}{dL^2T}}$ in Equation (D.6) gives us the required bound on the worst-case risk of $\hat{\theta}_{\text{AVG}}$.

LFP We now prove the results pertaining to LFP. From Theorem 11, we have

$$\begin{aligned}
\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} R(\hat{\theta}, P_{\text{AVG}}) &= \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \frac{1}{T}\sum_{i=1}^{T} R(\hat{\theta}, P_i)\\
&\geq \sup_{P\in\mathcal{M}_{\Theta}} \frac{1}{T}\sum_{i=1}^{T} R(\hat{\theta}_i, P) - O\left(\eta d^2 DL^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta dL\right)\\
&\geq \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \sup_{P\in\mathcal{M}_{\Theta}} R(\hat{\theta}, P) - O\left(\eta d^2 DL^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta dL\right).
\end{aligned}$$

Substituting $\eta = \sqrt{\frac{1}{dL^2T}}$ in the above equation shows that P_{AVG} is approximately least favourable. Now consider the case where M is convex in its first argument. To show that $\hat{\theta}_{\text{AVG}}$ is an approximate Bayes estimator for P_{AVG}, we again rely on Theorem 11 where we showed that

$$\sup_{P\in\mathcal{M}_{\Theta}} \frac{1}{T}\sum_{i=1}^{T} R(\hat{\theta}_i, P) \leq \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}, P_t) + O\left(\eta d^2 DL^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta dL\right).$$

Since $\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} R(\hat{\theta}_{t'}, P_t) \leq \sup_{P\in\mathcal{M}_{\Theta}} \frac{1}{T}\sum_{i=1}^{T} R(\hat{\theta}_i, P)$, we have

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} R(\hat{\theta}_{t'}, P_t) \leq \inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D}}} \frac{1}{T}\sum_{t=1}^{T} R(\hat{\theta}, P_t) + O\left(\eta d^2 DL^2 + \frac{d(\beta T + D)}{\eta T} + \alpha + \alpha' + \beta dL\right).$$

Since M is convex in its first argument, we have

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} R(\hat{\theta}_{t'}, P_t) \geq \frac{1}{T}\sum_{i=1}^{T} R(\hat{\theta}_{\text{AVG}}, P_i).$$

Combining the above two equations shows that $\hat{\theta}_{\text{AVG}}$ is an approximate Bayes estimator for P_{AVG}.

D.4 Invariance of Minimax Estimators

D.4.1 Proof of Theorem 12

In our proof, we rely on the following property of left Haar measure μ of a compact group G. For any real valued integrable function f on G and any $g \in G$ [see Chapter 7 of Wij90]

$$\int_G f(g^{-1}h)d\mu(h) = \int_G f(h)d\mu(h). \tag{D.7}$$

We now proceed to the proof of the Theorem. For any estimator $\hat{\theta} : \mathcal{X}^n \to \Theta$, define the following estimator $\hat{\theta}_G$

$$\hat{\theta}_G(\mathbb{X}^n) = \int_G g\hat{\theta}(g^{-1}\mathbb{X}^n)d\mu(g),$$

where μ is the left Haar measure on G and $g\mathbb{X}^n = \{gX_1, \dots gX_n\}$. The above integral is well defined because $\hat{\theta}$ is measurable, G is compact and the action of the group G is continuous. We first show that $\hat{\theta}_G$ is invariant under group transformations G. For any $h \in G$, consider the following

$$\begin{aligned}
\hat{\theta}_G(h\mathbb{X}^n) &= \int_G g\hat{\theta}((g^{-1}h)\mathbb{X}^n)d\mu(g)\\
&= \int_G h(h^{-1}g)\hat{\theta}((h^{-1}g)^{-1}\mathbb{X}^n)d\mu(g)\\
&= h\left[\int_G (h^{-1}g)\hat{\theta}((h^{-1}g)^{-1}\mathbb{X}^n)d\mu(g)\right]\\
&\overset{(a)}{=} h\left[\int_G g\hat{\theta}(g^{-1}\mathbb{X}^n)d\mu(g)\right]\\
&= h\hat{\theta}_G(\mathbb{X}^n),
\end{aligned}$$

where (a) follows from Equation (D.7). This shows that $\hat{\theta}_G$ is an invariant estimator. We now show that the worst case risk of $\hat{\theta}_G$ is less than or equal to the worst case risk of $\hat{\theta}$. Consider the following upper bound on the risk of $\hat{\theta}_G$ at any $\theta \in \Theta$

$$\begin{aligned}
R(\hat{\theta}_G, \theta) &= \mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[M(\hat{\theta}_G(\mathbb{X}^n), \theta)\right]\\
&\leq \mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[\int_G M(g\hat{\theta}(g^{-1}\mathbb{X}^n), \theta)d\mu(g)\right] \quad \text{(convexity of } M\text{)}\\
&= \mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[\mathbb{E}_{g\sim\mu}\left[M(g\hat{\theta}(g^{-1}\mathbb{X}^n), \theta)\right]\right]\\
&\overset{(a)}{=} \mathbb{E}_{g\sim\mu}\left[\mathbb{E}_{\mathbb{X}^n\sim P_{g^{-1}\theta}^n}\left[M(g\hat{\theta}(\mathbb{X}^n), \theta)\right]\right] \quad \text{(change of variables)}\\
&\overset{(b)}{=} \mathbb{E}_{g\sim\mu}\left[\mathbb{E}_{\mathbb{X}^n\sim P_{g^{-1}\theta}^n}\left[M(\hat{\theta}(\mathbb{X}^n), g^{-1}\theta)\right]\right] \quad \text{(invariance of } M\text{)}\\
&= \mathbb{E}_{g\sim\mu}\left[R(\hat{\theta}, g^{-1}\theta)\right]\\
&\leq \sup_{\theta'\in\Theta} R(\hat{\theta}, \theta'),
\end{aligned}$$

where (a) follows from Fubini's theorem and change of variables $X' = g^{-1}X$ and the fact that if $X \sim P_\theta$, then $g^{-1}X \sim P_{g^{-1}\theta}$. (b) follows from the invariance property of the metric M. This shows that $\sup_{\theta\in\Theta} R(\hat{\theta}_G, \theta) \leq \sup_{\theta\in\Theta} R(\hat{\theta}, \theta)$. This shows that we can always improve a given estimator by averaging over the group G and hence there should be a minimax estimator which is invariant under the action of G.

D.4.2 Proof of Theorem 13

We first prove some intermediate results which we require in the proof of the Theorem.

Intermediate Results

Lemma 52. *Suppose $\hat{\theta}$ is a deterministic estimator that is invariant to group transformations G. Then $R(\hat{\theta}, \theta_1) = R(\hat{\theta}, \theta_2)$, whenever $\theta_1 \sim \theta_2$.*

Proof. Suppose $\theta_2 = g\theta_1$ for some $g \in G$. From the definition of $R(\hat{\theta}, g\theta_1)$ we have

$$\begin{aligned} R(\hat{\theta}, \theta_2) = R(\hat{\theta}, g\theta_1) &= \mathbb{E}_{\mathbb{X}^n \sim P^n_{g\theta_1}} \left[M(\hat{\theta}(\mathbb{X}^n), g\theta_1)\right] \\ &= \mathbb{E}_{\mathbb{X}^n \sim P^n_{g\theta_1}} \left[M(g^{-1}\hat{\theta}(\mathbb{X}^n), \theta_1)\right] \quad \text{(invariance of loss metric)} \\ &= \mathbb{E}_{\mathbb{X}^n \sim P^n_{g\theta_1}} \left[M(\hat{\theta}(g^{-1}\mathbb{X}^n), \theta_1)\right] \quad \text{(invariance of estimator)} \\ &\overset{(a)}{=} \mathbb{E}_{\mathbb{X}^n \sim P^n_{\theta_1}} \left[M(\hat{\theta}(\mathbb{X}^n), \theta_1)\right] \\ &= R(\hat{\theta}, \theta_1), \end{aligned}$$

where (a) follows from the fact that $gX \sim P_{g\theta}$ whenever $X \sim P_\theta$. This shows that $R(\hat{\theta}, \theta_1) = R(\hat{\theta}, \theta_2)$. □

Lemma 53. *Suppose Π is a probability distribution which is invariant to group transformations G. For any deterministic estimator $\hat{\theta}$, there exists an invariant estimator $\hat{\theta}_G$ such that the Bayes risk of $\hat{\theta}_G$ is no larger than the Bayes risk of $\hat{\theta}$*

$$R(\hat{\theta}, \Pi) \geq R(\hat{\theta}_G, \Pi).$$

Proof. Define estimator $\hat{\theta}_G$ as follows

$$\hat{\theta}_G(\mathbb{X}^n) = \int_G g\hat{\theta}(g^{-1}\mathbb{X}^n) d\mu(g),$$

where μ is the left Haar measure on G. Note that, in the proof of Theorem 12 we showed that this estimator is invariance to the action of group G. We now show that the Bayes risk of $\hat{\theta}_G$ is less than equal to the Bayes risk of $\hat{\theta}$. Consider the following

$$\begin{aligned}
R(\hat{\theta}_G, \Pi) &= \mathbb{E}_{\theta\sim\Pi}[R(\hat{\theta}_G, \theta)] \\
&= \mathbb{E}_{\theta\sim\Pi}\left[\mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[M\left(\int_G g\hat{\theta}(g^{-1}\mathbb{X}^n)d\mu(g), \theta\right)\right]\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{\theta\sim\Pi}\left[\mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[\mathbb{E}_{g\sim\mu}\left[M\left(g\hat{\theta}(g^{-1}\mathbb{X}^n), \theta\right)\right]\right]\right] \\
&= \mathbb{E}_{g\sim\mu}\left[\mathbb{E}_{\theta\sim\Pi}\left[\mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[M\left(g\hat{\theta}(g^{-1}\mathbb{X}^n), \theta\right)\right]\right]\right] \\
&\overset{(b)}{=} \mathbb{E}_{g\sim\mu}\left[\mathbb{E}_{\theta\sim\Pi}\left[\mathbb{E}_{\mathbb{X}^n\sim P_\theta^n}\left[M\left(\hat{\theta}(g^{-1}\mathbb{X}^n), g^{-1}\theta\right)\right]\right]\right] \\
&= \mathbb{E}_{g\sim\mu}\left[\mathbb{E}_{\theta\sim\Pi}\left[R(\hat{\theta}, g^{-1}\theta)\right]\right] \\
&\overset{(c)}{=} \mathbb{E}_{\theta\sim\Pi}\left[R(\hat{\theta}, \theta)\right],
\end{aligned}$$

where (a) uses convexity of M and follows from Jensen's inequality, (b) follows from the invariance of M and (c) follows from the invariance of distribution Π to actions of group G. □

Main Argument

We now proceed to the proof of Theorem 13. We first prove the second part of the Theorem. The first part immediately follows from the proof of second part. Suppose $(\hat{\theta}_G^*, P_G^*)$ is an ϵ-approximate mixed strategy Nash equilibirium of the reduced statistical game in Equation (5.8). Our goal is to construct an approximate Nash equilibrium of the original statistical game in Equation (5.1), using $(\hat{\theta}_G^*, P_G^*)$.

Note that $\hat{\theta}_G^*$ is a randomized estimator over the set of deterministic invariant estimators $\mathcal{D}_G$ and P_G^* is a distribution on the quotient space Θ/G. To construct an approximate Nash equilibrium of the original statistical game (5.1), we extend P_G^* to the entire parameter space Θ. We rely on Bourbaki's approach to measure theory, which is equivalent to classical measure theory in the setting of locally compact spaces we consider in this work [Wij90]. In Bourbaki's approach, any measure ν on a set Θ is defined as a linear functional on the set of integrable functions (that is, a measure is defined by its action on integrable functions)

$$\nu[f] = \int_\Theta f(\theta)d\nu(\theta).$$

We define P^*, the extension of P_G^* to the entire parameter space Θ, as follows

$$P^*[f] = \int_{\Theta/G} f'(\Theta_\beta)dP_G^*(\Theta_\beta),$$

where $f' : \Theta/G \to \mathbb{R}$ is a function that depends on f, and is defined as follows. First define $f_I : \Theta \to \mathbb{R}$, an invariant function constructed using f, as $f_I(\theta) = \int_\Theta f(g\theta)d\mu(g)$, where μ is the left invariant Haar measure of G. From Equation (D.7), it is easy to see that $f_I(h\theta) = f_I(\theta)$, for all $h \in G$. So f_I is constant on the equivalence classes of Θ. So f_I can be written in terms of a function $f' : \Theta/G \to \mathbb{R}$, as follows

$$f_I = f' \circ \gamma,$$

where $\gamma : \Theta \to \Theta/G$ is the orbit projection function which projects $\theta \in \Theta$ onto the quotient space. We first show that P^* defined this way is an invariant measure. To this end, we use the following equivalent definition of an invariant measure.

Proposition 19. *A probability measure ν on Θ is invariant to transformations of group G iff for any ν-integrable function f and for any $h \in G$, $\int f(\theta)d\nu(\theta) = \int f(h\theta)d\nu(\theta)$.*

Since f_I is an invariant function, relying on the above proposition, it is easy to see that P^* is an invariant measure. We now show that $(\hat{\theta}^*_G, P^*)$ is an ϵ-approximate mixed strategy Nash equilibrium of Equation (5.1). Since $(\hat{\theta}^*_G, P^*_G)$ is an ϵ-approximate Nash equilibrium of Equation (5.8), we have

$$\sup_{\Theta_\beta \in \Theta/G} R_G(\hat{\theta}^*_G, \Theta_\beta) - \epsilon \leq \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}^*_G, \Theta_\beta)] \leq \inf_{\hat{\theta} \in \mathcal{D}_G} \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}, \Theta_\beta)] + \epsilon, \tag{D.8}$$

where $\mathcal{D}_G$ is the set of deterministic invariant estimators. Now consider the following

$$\begin{aligned}
\mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}^*_G, \theta)] &\overset{(a)}{=} \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}^*_G, \Theta_\beta)] \quad \text{(Lemma 52)}\\
&\leq \inf_{\hat{\theta} \in \mathcal{D}_G} \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}, \Theta_\beta)] + \epsilon \quad \text{(Equation (D.8))}\\
&= \inf_{\hat{\theta} \in \mathcal{D}_G} \mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}, \theta)] + \epsilon \quad \text{(definition of } P^*)\\
&\overset{(b)}{=} \inf_{\hat{\theta} \in \mathcal{D}} \mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}, \theta)] + \epsilon \quad \text{(Lemma 53)},
\end{aligned}$$

where (a) follows from the definition of P^* and Lemma 52. (b) follows from the fact that for any invariant prior, there exists a Bayes estimator which is invariant to group transformations (Lemma 53). Next, we provide a lower bound for $\mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}^*_G, \theta)]$

$$\begin{aligned}
\mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}^*_G, \theta)] &= \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}^*_G, \Theta_\beta)]\\
&\geq \sup_{\Theta_\beta \in \Theta/G} R_G(\hat{\theta}^*_G, \Theta_\beta) - \epsilon\\
&= \sup_{\theta \in \Theta} R(\hat{\theta}^*_G, \theta) - \epsilon \quad \text{(Lemma 52)}
\end{aligned}$$

The upper and lower bounds for $\mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}^*_G, \theta)]$ derived in the previous two equations shows that $(\hat{\theta}^*_G, P^*)$ is an ϵ-approximate mixed strategy Nash equilibrium of the original statistical game in Equation 5.1. The above inequalites also show that

$$\sup_{\theta \in \Theta} R(\hat{\theta}^*_G, \theta) - \epsilon \leq \mathbb{E}_{\Theta_\beta \sim P^*_G}[R_G(\hat{\theta}^*_G, \Theta_\beta)] \leq \inf_{\hat{\theta} \in \mathcal{D}} \mathbb{E}_{\theta \sim P^*}[R(\hat{\theta}, \theta)] + \epsilon.$$

This, together with Equation (D.8), shows that

$$\inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D}}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \inf_{\hat{\theta} \in \mathcal{M}_{\mathcal{D},G}} \sup_{\Theta_\beta \in \Theta/G} R_G(\hat{\theta}, \Theta_\beta).$$

D.4.3 Applications of Invariance Theorem

In our proofs, we establish homeomorphisms between the quotient spaces and another natural space over which we run our algorithm. Note that establishing a homeomorphism is sufficient since we are only dealing with Borel σ-algebras on our spaces and homeomorphism would imply that there is an isomorphism between the Borel σ-algebras of the two spaces. Hence, measures learnt on one space can be transferred to another.

Proof of Theorem 14

First note that for any $g \in \mathbb{O}(d)$ and $\theta \in \Theta$, we have $g\theta \in \Theta$ and the distribution of gX is $P_{g\theta}$. Moreover, for any orthogonal matrix $g \in \mathbb{O}(d)$ we have $\|g\theta - gX\|^2 = \|\theta - X\|^2$, which implies the statistical game is invariant to group transformations G.

For the second part, note that for any $\theta_1, \theta_2 \in \Theta$ such that $\|\theta_1\|_2 = \|\theta_2\|_2$, $\exists g \in \mathbb{O}(d)$ s.t. $g\theta_1 = \theta_2$. Mapping all elements to their norm gives us a bijection between the quotient space and the interval $[0, B]$. The continuity of this bijection and it's inverse can easily be checked using the standard basis for both the topologies.

Proof of Theorem 15

Note that for any $\theta \in \Theta$, $g\theta = [g_1\theta^{1:k},\ g_2\theta^{k+1:d}] \in \Theta$. Since g_1 is orthogonal, for any $\theta_1, \theta_2 \in \Theta$ we have $\|g_1\theta_1^{1:k} - g_1\theta_2^{1:k}\| = \|\theta_1^{1:k} - \theta_2^{1:k}\|$. Hence the invariance of the statistical game follows.

Now, for any $\theta_1, \theta_2 \in \Theta$ such that $\|\theta_1^{1:k}\| = \|\theta_2^{1:k}\|$ and $\|\theta_1^{k+1:d}\| = \|\theta_2^{k+1:d}\|$, $\exists g_1 \in \mathbb{O}(k)$ and $g_2 \in \mathbb{O}(d-k)$ such that $g_1\theta_1^{1:k} = \theta_2^{1:k}$ and $g_2\theta_1^{k+1:d} = \theta_2^{k+1:d}$. Hence $\exists g \in \mathbb{O}(k) \times \mathbb{O}(d-k)$ such that $g\theta_1 = \theta_2$. This means that in each equivalence class the parameters $B_1 = \|\theta_1^{1:k}\|^2$ and $B_2 = \|\theta_1^{k+1:d}\|^2$ are constant. Since $\|\theta\|^2 \leq B$ we have $B_1 + B_2 \leq B$, this gives us a bijection. The continuity of this bijection and it's inverse can easily be checked using the standard basis for both the topologies.

Proof of Theorem 16

We define the action of any $g \in \mathbb{O}(d)$ on the samples $\{(X_i, Y_i)\}_{i=1}^n$ as transforming them to $\{(gX_i, Y_i)\}_{i=1}^n$. Since $Y_i = X_i^T\theta + \epsilon_i = X_i^T g^T g\theta + \epsilon_i = (gX_i)^T g\theta + \epsilon_i$ and $\|g\theta_1 - g\theta_2\| = \|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2 \in \Theta$ we have the invariance of the statistical game. The rest of the proof uses similar arguments as in Theorem 14.

Proof of Theorem 17

First note that for any Σ such that $\|\Sigma\|_2 \leq B$, and any $g \in \mathbb{O}(d)$, we have $\|g\Sigma g^T\| \leq B$. If $X \sim N(0, \Sigma)$ then for any $g \in \mathbb{O}(d)$

$$\mathbb{E}[gXX^Tg^T] = g\mathbb{E}[XX^T]g^T = g\Sigma g^T.$$

Hence $gX \sim N(0, g\Sigma g^T)$. Moreover, we have

$$\begin{aligned}
&M(g\Sigma_1 g^T, g\Sigma_2 g^T) \\
&= \operatorname{tr}\left((g\Sigma_1 g^T)^{-1} g\Sigma_2 g^T\right) - \log|(g\Sigma_1 g^T)^{-1} g\Sigma_2 g^T| - d \\
&= \operatorname{tr}\left(g\Sigma_1^{-1} g^T g\Sigma_2^{-1} g^T\right) - \log|g\Sigma_1^{-1} g^T g\Sigma_2^{-1} g^T| - d \\
&= \operatorname{tr}(g\Sigma_1^{-1}\Sigma_2 g^T) - \log|g\Sigma_1^{-1}\Sigma_2 g^T| - d \\
&= M(\Sigma_1, \Sigma_2),
\end{aligned}$$

where the last equality follows from the invariance of trace to multiplication with orthogonal matrices and the property of the determinant to split over the multiplication of matrices. This shows the desired invariance of the statistical game.

Now, consider two covariance matrices Σ_1, Σ_2 with singular value decompositions (SVD) $\Sigma_1 = U_1\Delta_1 U_1^T$ and $\Sigma_2 = U_2\Delta_2 U_2^T$ respectively. Here all matrices are square and of full rank. In particular, Δ_1 and Δ_2 are diagonal matrices with decreasing entries from left to right and, U_1 and U_2 are orthogonal matrices. Since the orthogonal group is transitive $\exists g \in \mathbb{O}(d)$ such that $gU_1 = U_2$. If $\Delta_1 = \Delta_2$ we have $g\Sigma_1 g^T = \Sigma_2$. Hence under the action of $\mathbb{O}(d)$, all covariance matrices with the same singular values fall in the same equivalence class. It is easy to see that this is also a necessary condition. These equivalence classes naturally form a bijection with a sequence of d decreasing positive real numbers bounded above by B. The continuity of this bijection and it's inverse can easily be checked using the standard basis for both the topologies.

Proof of Theorem 18

Let P, Q be any two distributions on d elements $\{1, \ldots d\}$ such that $\exists g \in S_d$ s.t. $gP = Q$. They are indistinguishable from the samples they generate. Since the entropy is defined as

$$f(P) = -\sum_{i=1}^{d} p_i \log(p_i)$$

it doesn't depend upon the ordering of the individual probabilites. Hence the statistical game is invariant under the action of S_d.

Since using a permutation we can always order a given set of probabilities in decreasing order, there is a natural bijection between the quotient space and the given space. The continuity of this map and it's inverse can easily be checked using the standard basis for both the topologies.

Mixture of Gaussians

In the problem of mixture of Gaussians we are given n samples $X_1, \ldots, X_n \in \mathbb{R}^d$ which come from a mixture distribution of k Gaussians with different means

$$P_\theta = \sum_{i=1}^{k} p_i \mathcal{N}(\theta_i, \Sigma_i).$$

We assume that all k Gaussians have the same covariance, let's say identity, and we also assume that we know the mixture probabilities. Finally, we assume that the mean vectors θ_i are such that $\|\theta_i\| \leq B$. Under this setting we want to estimate the k different means while minimizing the sum of the L_2^2 losses of all the estimates of the mean parameters.

We will show the invariance of this statistical game under the action of the group $G = \mathbb{O}(d) \times \mathbb{O}(d-1) \times \ldots \times \mathbb{O}(d-k+1)$. But first we describe an element in the group and it's operation on the parameter and sample space.

An element of $g \in G$ is made up of a sequence of k orthonormal matrices $(g_1, \ldots, g_k)$ such that for a given set of parameters $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^{d \times k}$ (where each $\theta_i \in \mathbb{R}^d$) the matrix g_i leaves the first $(i-1)$ parameters unchanged, i.e. for $j = 1, \ldots, i-1$ $g_i\theta_j = \theta_j$. Hence the ith orthonormal matrix has $(d-i+1)$ degrees of freedom and can be viewed as an element in $\mathbb{O}(d-i+1)$.

The action of g on θ is defined as

$$\begin{aligned}
g\theta &= g(\theta_1, \ldots, \theta_k) \\
&= (g\theta_1, \ldots, g\theta_k) \\
&= (g_k \ldots g_1\theta_1, \ldots, g_k \ldots g_1\theta_1) \\
&= (g_1\theta_1, \ldots, g_i \ldots g_1\theta_i, \ldots, g_k \ldots g_1\theta_k)
\end{aligned}$$

where the last equality follows from the definition of our group. The group acts in a similar manner on the sample space, i.e., for an $X \in \mathcal{X}$ $gX = g_k \ldots g_1 X$.

Theorem 54. *The statistical game defined by mixture of k-Gaussians with identity covariance and known mixture probabilities under L_2^2 loss is invariant under the action of the group $\mathbb{O}(d) \times \mathbb{O}(d-1) \times \ldots \times \mathbb{O}(d-k+1)$. Moreover, the quotient space is homeomorphic to $(0, B]^k \times [0, \pi]^{\binom{k}{2}}$.*

Proof. First we show the invariance of the mixture distribution $P_\theta = \sum_i p_i \mathcal{N}(\theta_i, I)$, i.e., if $X \sim P_\theta$ then $gX \sim P_{g\theta}$. Note that from the proof of Theorem 14 it follows that for a given normal distribution $N(\tilde{\theta}, I)$ and an orthonormal matrix $h \in \mathbb{O}(d)$ s.t. $h\tilde{\theta} = \tilde{\theta}$ if $X \sim N(\tilde{\theta}, I)$ then $hX \sim N(h\tilde{\theta}, I) = N(\tilde{\theta}, I)$. The invariance of P follows directly from this by substituting each $\|X - \theta_i\|^2$ in the pdf with $\|g_k \ldots g_1 X - g_k \ldots g_1\theta_i\|^2$ and the definition of the group. The L_2^2 loss is trivially invariant and hence we establish the invariance of the statistical game.

Now, notice that for any two given parameters $\theta = (\theta_1, \ldots, \theta_k), \phi = (\phi_1, \ldots, \phi_k) \in \mathbb{R}^{dk}$ if we have the property that $\forall i$ $\|\theta_i\| = \|\phi_i\|$ and $\forall i, j$ $\theta_i^T\theta_j = \phi_i^T\phi_j$ then we can find orthonormal matrices $g_1, \ldots, g_k$ s.t. $\forall i$ $g_i \ldots g_1\theta_i = \phi_i$. This follows from the following inductive argument: Assume we have $g_1, \ldots, g_{i-1}$ which satisfy the given constraints. Consider $\theta' = g_{i-1} \ldots g_1\theta_i$. We have $\forall j = 1, \ldots, i-1$ $\theta'^T\phi_j = \theta_i^T\theta_j = \phi_i^T\phi_j$ because $g^T = g^{-1}$. Now if ϕ_i lies in the span of $\phi_1, \ldots, \phi_{i-1}$ then $\theta' = \phi_i$ and we can pick g_i to be any orthonormal matrix which doesn't transform this spanned space. Otherwise, we can pick an orthonormal matrix which rotates the axis orthogonal to the spanned subspace and in the direction of the high component of θ' to the corresponding axis for ϕ_i. This completes the desired construction.

It is easy to see that given θ, ϕ, g which satisfy $g\theta = \phi$, we have $\forall i$ $\|\theta_i\| = \|\phi_i\|$ and $\forall i, j$ $\theta_i^T\theta_j = \phi_i^T\phi_j$. Hence the equivalence classes are defined uniquely by the norms of the individual gaussians and the angles between them, since there are k different norms and $\binom{k}{2}$ many angles we can establish a bijection between the quotient space and $(0, B]^k \times [0, \pi]^{\binom{k}{2}}$. The continuity of this map and it's inverse can easily be checked using the standard basis for both the topologies. □

D.5 Finite Gaussian Sequence Model

D.5.1 Proof of Proposition 8

In this section we derive a closed-form expression for the minimizer $\hat{\theta}_t$ of the following objective

$$\mathop{\mathrm{argmin}}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right],$$

where $\mathcal{D}_G$ is the set of deterministic estimators which are invariant to transformations of orthogonal group $\mathbb{O}(d)$. From Lemma 52, we know that for any invariant estimator $\hat{\theta} \in \mathcal{D}_G$ and any $g \in \mathbb{O}(d)$, $R(\hat{\theta}, b\mathbf{e}_1) = R(\hat{\theta}, bg\mathbf{e}_1)$. So the above problem can be rewritten as follows

$$\mathop{\mathrm{argmin}}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[\mathbb{E}_{\theta\sim U_b}\left[R(\hat{\theta}, \theta)\right]\right],$$

where U_b is the uniform distribution over spherical shell of radius b, centered at origin; that is, its density $u_b(\theta)$ is defined as

$$u_b(\theta) \propto \begin{cases} 0, & \text{if } \|\theta\|_2 \neq b \\ b^{-d+1}, & \text{otherwise} \end{cases}.$$

The above optimization problem can be further rewritten as

$$\mathop{\mathrm{argmin}}_{\hat{\theta}\in\mathcal{D}_G} R(\hat{\theta}, \Pi_t),$$

where $R(\hat{\theta}, \Pi_t) \stackrel{\text{def}}{=} \mathbb{E}_{\theta\sim\Pi_t}\left[R(\hat{\theta}, \theta)\right]$, and Π_t is the distribution of a random variable θ which is generated by first sampling b from P_t and then generating a sample from U_b. Note that Π_t is a spherically symmetric distribution. From Lemma 53, we know that the Bayes estimator corresponding to any invariant prior is an invariant estimator. So the minimization over $\mathcal{D}_G$ in the above optimization problem can be replaced with minimization over the set of all estimators $\mathcal{D}$. This leads us to the following equivalent optimization problem

$$\mathop{\mathrm{argmin}}_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \Pi_t).$$

Let $\hat{\theta}_t$ be the minimizer of this equivalent problem. We now obtain an expression for $\hat{\theta}_t(X)$ in terms of modified Bessel functions. Let $\Pi_t(\cdot|X)$ be the posterior distribution of θ given the data X and let $p(X;\theta)$ be the probability density function for distribution P_θ. Since the risk is measured with respect to ℓ_2^2 metric, the Bayes estimator $\hat{\theta}_t(X)$ is given by the posterior mean

$$\begin{aligned}
\hat{\theta}_t(X) &= \mathbb{E}_{\theta\sim\Pi_t(\cdot|X)}\left[\theta\right]\\
&= \frac{\mathbb{E}_{\theta\sim\Pi_t}\left[\theta p(X;\theta)\right]}{\mathbb{E}_{\theta\sim\Pi_t}\left[p(X;\theta)\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[\int \theta u_b(\theta)p(X;\theta)d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[\int u_b(\theta)p(X;\theta)d\theta\right]} \quad \text{(definition of } \Pi_t)\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}\theta p(X;\theta)d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}p(X;\theta)d\theta\right]} \quad \text{(since } U_b \text{ is uniform on sphere)}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}e^{-b^2/2}\int_{\|\theta\|_2=b}\theta e^{\langle X,\theta\rangle}d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}e^{-b^2/2}\int_{\|\theta\|_2=b}e^{\langle X,\theta\rangle}d\theta\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{2}e^{-b^2/2}\int_{\|\theta\|_2=1}\theta e^{b\langle X,\theta\rangle}d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[be^{-b^2/2}\int_{\|\theta\|_2=1}e^{b\langle X,\theta\rangle}d\theta\right]} \quad \text{(change of variables).}
\end{aligned}$$

We now obtain a closed-form expression for the terms $\int_{\|\theta\|_2=1}\theta e^{b\langle X,\theta\rangle}d\theta$ and $\int_{\|\theta\|_2=1}e^{b\langle X,\theta\rangle}d\theta$ appearing in the RHS of the above equation. We do this by relating them to the mean and normalization constant of Von Mises-Fisher (vMF) distribution, which is a probability distribution on the unit sphere centered at origin in $\mathbb{R}^d$. This distribution is usually studied in directional statistics [MJ09]. The probability density function of a random unit vector $Z\in\mathbb{R}^d$ distributed according to vMF distribution is given by

$$p(Z;\mu,\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}\exp(\kappa\langle\mu, Z\rangle),$$

where $\kappa\geq 0$, $\|\mu\|_2=1$, I_ν is the modified Bessel function of the first kind of order ν. Using the fact that a probability density function integrates to 1, we get the following closed-form expression for $\int_{\|\theta\|_2=1}e^{b\langle X,\theta\rangle}d\theta$

$$\int_{\|\theta\|_2=1}e^{b\langle X,\theta\rangle}d\theta = \frac{(2\pi)^{d/2}I_{d/2-1}(b\|X\|_2)}{(b\|X\|_2)^{d/2-1}}. \tag{D.9}$$

To get a closed-form expression for $\int_{\|\theta\|_2=1}\theta e^{b\langle X,\theta\rangle}d\theta$, we relate it to mean of vMF distribution. We have the following expression for the mean of a random vector distributed according to vMF distribution [Ban+05]

$$\int_{\|Z\|=1}Zp(Z;\mu,\kappa)dZ = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}\mu.$$

Using the above equality, we get the following expression for $\int_{\|\theta\|_2=1}\theta e^{b\langle X,\theta\rangle}d\theta$

$$\int_{\|\theta\|_2=1}\theta e^{b\langle X,\theta\rangle}d\theta = \frac{(2\pi)^{d/2}I_{d/2}(b\|X\|_2)}{(b\|X\|_2)^{d/2-1}}\frac{X}{\|X\|_2}. \tag{D.10}$$

Substituting Equations (D.9), (D.10) in the expression for $\hat{\theta}_t(X)$ obtained above, we get an expression for $\hat{\theta}_t(X)$ which involves the modified Bessel function I_ν and integrals over variable b. We note that I_ν can be computed to very high accuracy and there exist accurate implementations of I_ν in a number of programming languages. So in our analysis of the approximation error of Algorithm 7, we assume the error from the computation of I_ν is 0.

D.5.2 Proof of Theorem 19

Before we present the proof of the Theorem we present useful intermediate results which we require in our proof.

Intermediate Results

Lemma 55 (Lipschitz Continuity). *Consider the problem of finite Gaussian sequence model. Let $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq B\}$ be the ball of radius B centered at origin in $\mathbb{R}^d$. Let $\hat{\theta}$ be any estimator which maps X to an element in Θ. Then the risk $R(\hat{\theta},\theta) = \mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[\|\hat{\theta}(X)-\theta\|_2^2\right]$ is Lipschitz continuous in its second argument w.r.t ℓ_2 norm over the domain Θ, with Lipschitz constant $4(B+\sqrt{d}B^2)$. Moreover, $R(\hat{\theta},b\mathbf{e}_1) = \mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[\|\hat{\theta}(X)-b\mathbf{e}_1\|_2^2\right]$ is Lipschitz continuous in b over the domain $[0,B]$, with Lipschitz constant $4(B+B^2)$.*

Proof. Let $R_{\hat{\theta}}(\theta) = R(\hat{\theta},\theta)$. The gradient of $R_{\hat{\theta}}(\theta)$ with respect to θ is given by

$$\nabla_\theta R_{\hat{\theta}}(\theta) = \mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[2(\theta-\hat{\theta}(X)) + (X-\theta)\|\hat{\theta}(X)-\theta\|_2^2\right].$$

The norm of $\nabla_\theta R_{\hat{\theta}}(\theta)$ can be upper bounded as follows

$$\begin{aligned}
\|\nabla_\theta R_{\hat{\theta}}(\theta)\|_2 &\leq \left\|\mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[2(\theta-\hat{\theta}(X))\right]\right\|_2 + \left\|\mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[(X-\theta)\|\hat{\theta}(X)-\theta\|_2^2\right]\right\|_2\\
&\overset{(a)}{\leq} 4B + \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\theta,I)}\left[\|X-\theta\|_2\|\hat{\theta}(X)-\theta\|_2^2\right]\\
&\overset{(b)}{\leq} 4B + 4B^2\mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[\|X-\theta\|_2\right]\\
&\leq 4B + 4\sqrt{d}B^2,
\end{aligned}$$

where the first term in (a) follows from the fact that $\theta, \hat{\theta}(X) \in \Theta$ and the second term follows from Jensen's inequality. This shows that $R_{\hat{\theta}}(\theta)$ is Lipschitz continuous over Θ. This finishes the first part of the proof. To show that $R(\hat{\theta},b\mathbf{e}_1)$ is Lipschitz continuous in

b, we use similar arguments. Let $R_{\hat{\theta}}(b) = R(\hat{\theta}, b\mathbf{e}_1)$. Then

$$\begin{aligned}
\left|R'_{\hat{\theta}}(b)\right| &= \left|\left\langle \mathbf{e}_1, \nabla_\theta R_{\hat{\theta}}(\theta)\Big|_{\theta=b\mathbf{e}_1}\right\rangle\right| \\
&\overset{(a)}{\leq} \left|\mathbb{E}_{X\sim\mathcal{N}(b\mathbf{e}_1,I)}\left[2(b-[\hat{\theta}(X)]_1)\right]\right| + \left\|\mathbb{E}_{X\sim\mathcal{N}(b\mathbf{e}_1,I)}\left[(X_1-b)\|\hat{\theta}(X)-b\mathbf{e}_1\|_2^2\right]\right\|_2 \\
&\leq 4B + \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(b\mathbf{e}_1,I)}\left[|X_1-b|\|\hat{\theta}(X)-b\mathbf{e}_1\|_2^2\right] \\
&\leq 4B + 4B^2\mathbb{E}_{X\sim\mathcal{N}(b\mathbf{e}_1,I)}\left[|X_1-b|\right] \\
&\leq 4B + 4B^2,
\end{aligned}$$

where (a) follows from the expression for $\nabla_\theta R_{\hat{\theta}}(\theta)$ obtained above. □

Lemma 56 (Approximation of risk)**.** *Consider the setting of Lemma 55. Let $\hat{\theta}$ be any estimator which maps X to an element in Θ. Let $\{X_i\}_{i=1}^N$ be N i.i.d samples from $\mathcal{N}(\theta, I)$. Then with probability at least $1-\delta$*

$$\left|\frac{1}{N}\sum_{i=1}^N \|\hat{\theta}(X_i)-\theta\|_2^2 - R_{\hat{\theta}}(\theta)\right| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N}}.$$

Proof. The proof of the Lemma relies on concentration properties of sub-Gaussian random variables. Let $Z(X) = \|\hat{\theta}(X)-\theta\|^2$. Note that $R_{\hat{\theta}}(\theta) = \mathbb{E}_{X\sim\mathcal{N}(\theta,I)}\left[Z(X)\right]$. Since $Z(X)$ is bounded by $4B^2$, it is a sub-Gaussian random variable. Using Hoeffding bound we get

$$\left|\frac{1}{N}\sum_{i=1}^N Z(X_i) - \mathbb{E}\left[Z(X)\right]\right| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N}}, \quad \text{w.p} \geq 1-\delta.$$

□

Main Argument

The proof relies on Corollary 4 to show that the averaged estimator $\hat{\theta}_{\text{AVG}}$ is approximately minimax and $\hat{P}_{\text{LFP}}$ is approximately least favorable. Here is a rough sketch of the proof. We first apply the corollaries on the following reduced statistical game that we are aiming to solve

$$\inf_{\hat{\theta}\in\mathcal{D}_G}\sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1).$$

To apply these corollaries, we need the risk $R(\hat{\theta}, b\mathbf{e}_1)$ to be Lipscthiz continuous in b. This holds for us because of Lemma 55. Next, we convert the guarantees for the reduced statistical game to the orginial statistical game to show that we learn a minimax estimator and LFP for finite Gaussian sequence model.

To use Corollary 4, we first need to bound α, β, α', the approximation errors of the optimization subroutines described in Algorithms 6, 7. A major part of the proof involves bounding these quantities.

Approximation error of Algorithm 6 There are two causes for error in the optimization oracle described in Algorithm 6: (a) grid search and (b) approximate computation of risk $R(\hat{\theta}, b\mathbf{e}_1)$. We now bound the error due to both (a) and (b). From Lemma 56 we know that for any estimator $\hat{\theta}_i$ and grid point b_j, the following holds with probability at least $1-\delta$

$$\Big|\frac{1}{N_1}\sum_{k=1}^{N_1}\|\hat{\theta}_i(X_k) - b_j\mathbf{e}_1\|_2^2 - R(\hat{\theta}_i, b_j\mathbf{e}_1)\Big| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N_1}}.$$

Taking a union bound over all estimators $\{\hat{\theta}_i\}_{i=1}^T$ and grid points $\{b_j\}_{j=1}^{B/w}$, we can show that with probability at least $1-\delta$, the following holds for all $i \in [T], j \in [B/w]$

$$\Big|\frac{1}{N_1}\sum_{k=1}^{N_1}\|\hat{\theta}_i(X_k) - b_j\mathbf{e}_1\|_2^2 - R(\hat{\theta}_i, b_j\mathbf{e}_1)\Big| \leq 4B^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}}. \tag{D.11}$$

Let $f_{t,\sigma}(b)$ be the actual objective we would like to optimize in iteration t of Algorithm 5, which is given by

$$f_{t,\sigma}(b) = \sum_{i=1}^{t-1} R(\hat{\theta}_i, b\mathbf{e}_1) + \sigma b.$$

Let $\hat{f}_{t,\sigma}(b)$ be the approximate objective we are optimizing by replacing $R(\hat{\theta}_i, b\mathbf{e}_1)$ with its approximate estimate. Let b_t^* be a maximizer of $f_{t,\sigma}(b)$ and $b_{t,\text{approx}}^*$ be the maximizer of $\hat{f}_{t,\sigma}(b)$ (which is also the output of Algorithm 6). Finally, let $b_{t,\text{NN}}^*$ be the point on the grid which is closest to b_t^*. Using Lemma 55 we first show that $f_{t,\sigma}(b)$ is Lipschitz continuous in b. The derivative of $f_{t,\sigma}(b)$ with respect to b is given by

$$f'_{t,\sigma}(b) = \sum_{i=1}^{t-1}\left\langle \mathbf{e}_1, \nabla_\theta R(\hat{\theta}_i, \theta)\Big|_{\theta=b\mathbf{e}_1}\right\rangle + \sigma$$

Using Lemma 55, the magnitude of $f'_{t,\sigma}(b)$ can be upper bounded as

$$|f'_{t,\sigma}(b)| \leq 4(t-1)(B+B^2) + \sigma.$$

This shows that $f_{t,\sigma}(b)$ is Lipschitz continuous in b. We now bound $f_{t,\sigma}(b_t^*) - f_{t,\sigma}(b_{t,\text{approx}}^*)$, the approximation error of the optimization oracle

$$\begin{aligned}
f_{t,\sigma}(b_t^*) &\overset{(a)}{\leq} f_{t,\sigma}(b_{t,\text{NN}}^*) + \big(4t(B+B^2)+\sigma\big)\,w\\
&\overset{(b)}{\leq} \hat{f}_{t,\sigma}(b_{t,\text{NN}}^*) + 4tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \big(4t(B+B^2)+\sigma\big)\,w\\
&\overset{(c)}{\leq} \hat{f}_{t,\sigma}(b_{t,\text{approx}}^*) + 4tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \big(4t(B+B^2)+\sigma\big)\,w\\
&\overset{(d)}{\leq} f_{t,\sigma}(b_{t,\text{approx}}^*) + 8tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \big(4t(B+B^2)+\sigma\big)\,w,
\end{aligned}$$

where (a) follows from Lipschitz property of the loss function and $(b), (d)$ follow from Equation (D.11) and hold with probability at least $1-\delta$ and (c) follows from the optimality of $b^*_{t,\text{approx}}$. This shows that Algorithm 6 is a $\left(O\left(TB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + TB(1+B)w\right), w\right)$-approximate maximization oracle; that is

$$\alpha = O\left(TB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + TB(1+B)w\right), \quad \beta = w.$$

Approximation error of Algorithm 7 There are two sources of approximation error in Algorithm 7: (a) computation of modified Bessel functions I_ν, and (b) approximation of P_t with its samples. In this analysis we assume that I_ν can be computed to very high accuracy. This is a reasonable assumption because many programming languages have accurate and efficient implementations of I_ν. So the main focus here is on bounding the error from approximation of P_t.

First, note that since we are using grid search to optimize the maximization problem, the true distribution P_t for which we are supposed to compute the Bayes estimator is a discrete distribution supported on grid points $\{b_1, \dots b_{B/w}\}$. Algorithm 7 does not compute the Bayes estimator for P_t. Instead, we generate samples from P_t and use them as a proxy for P_t. Let $\hat{P}_t$ be the empirical distribution obtained by sampling N_2 points from P_t. Let $p_{t,j}$ be the probability mass on grid point b_j. Using Bernstein inequality we can show that the following holds with probability at least $1-\delta$

$$\forall j \in [B/w] \quad |\hat{p}_{t,j} - p_{t,j}| \leq \sqrt{p_{t,j}\frac{\log\frac{B}{w\delta}}{N_2}}. \tag{D.12}$$

Define estimators $\hat{\theta}'_t, \hat{\theta}_t$ as

$$\hat{\theta}'_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right], \quad \hat{\theta}_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right].$$

$\hat{\theta}'_t$ is what we ideally want to compute. $\hat{\theta}_t$ is what we end up computing using Algorithm 7. We now show that $\hat{\theta}_t$ is an approximate minimizer of the left hand side optimization problem above. To this end, we try to bound the following quantity

$$\mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}_t, b\mathbf{e}_1) - R(\hat{\theta}'_t, b\mathbf{e}_1)\right].$$

Let $f_t(\hat{\theta}) = \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right]$ and $\hat{f}_t(\hat{\theta}) = \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right]$. We would like to bound the quantity $f_t(\hat{\theta}_t) - f_t(\hat{\theta}'_t)$. Consider the following

$$\begin{aligned} f_t(\hat{\theta}_t) &\overset{(a)}{\leq} \hat{f}_t(\hat{\theta}_t) + \frac{4B^3}{w}\sqrt{\frac{\log\frac{B}{w\delta}}{N_2}} \\ &\overset{(b)}{\leq} \hat{f}_t(\hat{\theta}'_t) + \frac{4B^3}{w}\sqrt{\frac{\log\frac{B}{w\delta}}{N_2}} \\ &\overset{(c)}{\leq} f_t(\hat{\theta}'_t) + \frac{8B^3}{w}\sqrt{\frac{\log\frac{B}{w\delta}}{N_2}}, \end{aligned}$$

where (a) follows from Equation (D.12) and the fact that the risk $R(\hat{\theta}, \theta)$ of any estimator is bounded by $4B^2$, (b) follows since $\hat{\theta}_t$ is a minimizer of $\hat{f}_t$ and (c) follows from Equation (D.12). This shows that with probability at least $1-\delta$, Algorithm 7 is an $O\left(\frac{B^3}{w}\sqrt{\frac{\log\frac{B}{w\delta}}{N_2}}\right)$-approximate optimization oracle; that is,

$$\alpha' = O\left(\frac{B^3}{w}\sqrt{\frac{\log\frac{B}{w\delta}}{N_2}}\right).$$

Minimax Estimator We are now ready to show that $\hat{\theta}_{\text{AVG}}$ is an approximate minimax estimator. Instantiating Corollary 4 for the reduced statistical game gives us the following bound, which holds with probability at least $1-\delta$

$$\sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq \inf_{\hat{\theta}\in\mathcal{D}_G}\sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) + \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}} + \alpha + \alpha' + \beta B(B+1)\sqrt{T}\right),$$

where we used the fact that the risk $R(\hat{\theta}, b\mathbf{e}_1)$ is $4B(B+1)$-Lipschitz continuous w.r.t b. The $\tilde{O}$ notation in the above inequality hides logarithmic factors. Plugging in the values of α, α', β in the above equation gives us

$$\sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq \inf_{\hat{\theta}\in\mathcal{D}_G}\sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) + \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right).$$

We now convert this bound to a bound on the original statistical game. From Theorem 13 we know that $\inf_{\hat{\theta}\in\mathcal{D}_G}\sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) = \inf_{\hat{\theta}\in\mathcal{D}}\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) = R^*$. Since the estimator $\hat{\theta}_{\text{AVG}}$ is invariant to transformations of orthogonal group, we have $R(\hat{\theta}_{\text{AVG}}, \theta) = R(\hat{\theta}_{\text{AVG}}, \|\theta\|_2\mathbf{e}_1)$ for any $\theta \in \Theta$. Using these two results in the above inequality, we get

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{AVG}}, \theta) = \sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq R^* + \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right).$$

This shows that the worst-case risk of $\hat{\theta}_{\text{AVG}}$ is close to the minimax risk R^*. This finishes the first part of the proof.

LFP To prove the second part, we rely on Corollary 4. Instantiating it for the reduced statistical game gives us

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}} + \alpha + \alpha' + \beta B(B+1)\sqrt{T}\right).$$

Plugging in the values of α, α', β in the above equation gives us

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right).$$

From Equation (D.12) we know that P_t is close to $\hat{P}_t$ with high probability. Using this, we can replace P_t in the above bound with $\hat{P}_t$ and obtain the following bound, which holds with probability at least $1-\delta$

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B+1)}{\sqrt{T}}\right). \tag{D.13}$$

In the rest of the proof, we show that $\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] = \inf_{\hat{\theta}} R(\hat{\theta}, \hat{P}_{\text{LFP}})$. Recall, the density function of $\hat{P}_{\text{LFP}}$ is given by: $\hat{p}_{\text{LFP}}(\theta) \propto \|\theta\|_2^{1-d}\hat{P}_{\text{AVG}}(\|\theta\|_2)$, where $\hat{P}_{\text{AVG}}(\|\theta\|_2)$ is the probability mass placed by $\hat{P}_{\text{AVG}}$ at $\|\theta\|_2$. This distribution is equivalent to the distribution of a random variable which is generated by first sampling b from $\hat{P}_t$ and then sampling θ from the uniform distribution on $(d-1)$ dimensional sphere of radius b, centered at origin in $\mathbb{R}^d$. Using this equivalence, we can equivalently rewrite $R(\hat{\theta}, \hat{P}_{\text{LFP}})$ for any estimator $\hat{\theta}$ as

$$R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[\mathbb{E}_{\theta\sim U}\left[R(\hat{\theta}, b\theta)\right]\right],$$

where U is the uniform distribution on the $(d-1)$ dimensional unit sphere centered at origin, in $\mathbb{R}^d$. Next, from Lemma 53, we know that the Bayes estimator corresponding to any invariant prior is an invariant estimator. Since $\hat{P}_{\text{LFP}}$ is an invariant distribution, we have

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[\mathbb{E}_{\theta\sim U}\left[R(\hat{\theta}, b\theta)\right]\right].$$

From Lemma 52 we know that for any invariant estimator $\hat{\theta}$, we have $R(\hat{\theta}, \theta_1) = R(\hat{\theta}, \theta_2)$, whenever $\theta_1 \sim \theta_2$. Using this result in the above equation gives us

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right].$$

Combining the above result with Equation (D.13) shows that $\hat{P}_{\text{LFP}}$ is approximately least favorable.

D.5.3 Loss on few co-ordinates

In this section, we present the optimization oracles for the problem of finite Gaussian sequence model, when the loss is evaluated on a few co-ordinates. Recall, in Theorem 15 we showed that the original min-max statistical game can be reduced to the following simpler problem

$$\inf_{\hat{\theta}\in\mathcal{M}_{\mathcal{D},G}} \sup_{b:b[1]^2+b[2]^2\leq B^2} R(\hat{\theta}, [b[1]\mathbf{e}_{1,k}, b[2]\mathbf{e}_{1,d-k}]), \tag{D.14}$$

where $b[j]$ represents the j^{th} co-ordinate of b. We now provide efficient implementations of the optimization oracles required by Algorithm 5 for finding a Nash equilibrium of this game. The optimization problems corresponding to the two optimization oracles are as follows

$$\hat{\theta}_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, [b[1]\mathbf{e}_{1,k}, b[2]\mathbf{e}_{1,d-k}])\right],$$

$$b_t(\sigma) \leftarrow \operatorname*{argmax}_{b:b[1]^2+b[2]^2\leq B^2} \sum_{i=1}^{t-1} R(\hat{\theta}_i, [b[1]\mathbf{e}_{1,k}, b[2]\mathbf{e}_{1,d-k}]) + \langle\sigma, b\rangle,$$

where $\mathcal{D}_G$ is the set of deterministic invariant estimators and P_t is the distribution of random variable $b_t(\sigma)$. The maximization oracle can be efficiently implemented via a grid search over $\{b : b[1]^2 + b[2]^2 \leq B^2\}$ (see Algorithm 15). The minimization oracle can also be efficiently implemented. The minimizer has a closed form expression which depends on P_t and modified Bessel functions (see Algorithm 16).

Algorithm 15 Maximization Oracle

```
Input: Number of coordinates to evaluate loss on $k$, estimators $\{\hat{\theta}_i\}_{i=1}^{t-1}$, perturbation $\sigma$, grid
  width $w$, number of samples for computation of expected risk $R(\hat{\theta}, \theta)$: $N_1$
Let $\{b_1, b_2 \dots b_{N(w)}\}$ be the $w$-covering of $\{b : b[1]^2 + b[2]^2 \leq B^2\}$
for $j = 1 \dots N(w)$ do
    for $i = 1 \dots t-1$ do
        Generate $N_1$ independent samples $\{X_l\}_{l=1}^{N_1}$ from the following distribution
            $\mathcal{N}([b_j[1]\mathbf{e}_{1,k}, b_j[2]\mathbf{e}_{1,d-k}], I)$
        Estimate $R(\hat{\theta}_i, [b_j[1]\mathbf{e}_{1,k}, b_j[2]\mathbf{e}_{1,d-k}])$ as
            $\frac{1}{N_1}\sum_{l=1}^{N_1} \|\hat{\theta}_i(X_l)[1:k] - b_j[1]\mathbf{e}_{1,k}\|_2^2.$
    end for
    Evaluate the objective at $b_j$ using the above estimates
end for
Output: $b_j$ which maximizes the objective
```

Algorithm 16 Minimization Oracle

```
Input: Samples $\{b_i\}_{i=1}^{N_2}$ generated from distribution $P_t$, number of coordinates to evaluate loss on $k$.
For any $X$, compute $\hat{\theta}_t(X)$ as
    $\left(\frac{\sum_{i=1}^{N_2} w_i b_i[1] A_k(b_i[1]\|X[1:k]\|_2)}{\sum_{i=1}^{N_2} w_i}\right)\frac{X[1:k]}{\|X[1:k]\|_2},$
    where $A_k(\gamma) = \frac{I_{k/2}(\gamma)}{I_{k/2-1}(\gamma)}$,
    $w_i = b_i[1]^{2-\frac{k}{2}} b_i[2]^{2-\frac{d-k}{2}} e^{-\frac{\|b\|^2}{2}} I_{k/2-1}(b_i[1]\|X[1:k]\|_2) I_{(d-k)/2-1}(b_i[2]\|X[k+1:d]\|_2),$
    and $I_\nu$ is the modified Bessel function of the first kind of order $\nu$.
```

D.6 Linear Regression

D.6.1 Proof of Proposition 9

In this section we derive a closed-form expression for the minimizer $\hat{\theta}_t$ of the following objective

$$\operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right].$$

Using the same arguments as in proof of Proposition 8, we can show that the above optimization problem can be rewritten as the following equivalent optimization problem over the set of all deterministic estimators

$$\operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}} \mathbb{E}_{\theta\sim\Pi_t}\left[R(\hat{\theta}, \theta)\right],$$

where Π_t is the distribution of a random variable θ which is generated by first sampling a b from P_t and then drawing a random sample from U_b, the uniform distribution on a spherical shell of radius b. The density function of U_b is given by

$$u_b(\theta) \propto \begin{cases} 0, & \text{if } \|\theta\|_2 \neq b \\ b^{-d+1}, & \text{otherwise} \end{cases}.$$

Since the risk is measured with respect to ℓ_2^2 metric, the minimizer $\hat{\theta}_t(D_n)$ is given by the posterior mean

$$\begin{aligned}
\hat{\theta}_t(D_n) &= \mathbb{E}_{\theta\sim\Pi_t(\cdot|D_n)}\left[\theta\right]\\
&= \frac{\mathbb{E}_{\theta\sim\Pi_t}\left[\theta p(D_n;\theta)\right]}{\mathbb{E}_{\theta\sim\Pi_t}\left[p(D_n;\theta)\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[\int \theta u_b(\theta)p(D_n;\theta)d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[\int u_b(\theta)p(D_n;\theta)d\theta\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}\theta p(D_n;\theta)d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}p(D_n;\theta)d\theta\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}\theta e^{-\frac{\|\mathbf{Y}-\mathbf{X}\theta\|_2^2}{2}}d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[b^{-d+1}\int_{\|\theta\|_2=b}e^{-\frac{\|\mathbf{Y}-\mathbf{X}\theta\|_2^2}{2}}d\theta\right]}\\
&= \frac{\mathbb{E}_{b\sim P_t}\left[b^{2}\int_{\|\theta\|_2=1}\theta e^{-\frac{b^2\|\mathbf{X}\theta\|_2^2-2b\left\langle\theta,\mathbf{X}^T\mathbf{Y}\right\rangle}{2}}d\theta\right]}{\mathbb{E}_{b\sim P_t}\left[b\int_{\|\theta\|_2=1}e^{-\frac{b^2\|\mathbf{X}\theta\|_2^2-2b\langle\theta,\mathbf{X}^T\mathbf{Y}\rangle}{2}}d\theta\right]} \quad \text{(change of variables)}.
\end{aligned}$$

We now relate the terms appearing in the above expression to the mean and normalization constant of Fisher-Bingham (FB) distribution. As stated in Section 5.5, the probability density function of a random unit vector $Z \in \mathbb{R}^d$ distributed according to FB distribution is given by

$$p(Z;A,\gamma) = C(A,\gamma)^{-1}\exp\left(-Z^TAZ + \langle\gamma,Z\rangle\right),$$

where $Z \in \mathbb{S}^{d-1}$, and $\gamma \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$ are the parameters of the distribution with A being positive semi-definite and $C(A,\gamma)$ is the normalization constant which is given by

$$C(A,\gamma) = \int_{\|Z\|_2=1}\exp\left(-Z^TAZ + \langle\gamma,Z\rangle\right)dZ.$$

The mean of Z is given by

$$\begin{aligned}
\int_{\|Z\|_2=1} Zp(Z;A,\gamma)dZ &= C(A,\gamma)^{-1}\int_{\|Z\|_2=1} Z\exp\left(-Z^TAZ + \langle\gamma,Z\rangle\right)dZ\\
&= C(A,\gamma)^{-1}\frac{\partial}{\partial\gamma}C(A,\gamma).
\end{aligned}$$

Using these in the previously derived expression for $\hat{\theta}(D_n)$ gives us the required result.

D.6.2 Mean and normalization constant of Fisher-Bingham distribution

In this section, we present our technique for computation of $C\left(A,\gamma\right)$. Once we have an accurate technique for its computation, computing $\frac{\partial}{\partial\gamma}C(A,\gamma)$ should be straight forward as one can rely on efficient numerical differentiation techniques for its computation. Recall, to implement Algorithm 9 we need to compute $C\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, b\mathbf{X}^T\mathbf{Y}\right)$. Let $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and let $U\Lambda U^T$ be its eigen decomposition. Then it is easy to see that $C\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, b\mathbf{X}^T\mathbf{Y}\right)$ can be rewritten as

$$C\left(2^{-1}b^2\mathbf{X}^T\mathbf{X}, b\mathbf{X}^T\mathbf{Y}\right) = C(2^{-1}nb^2\Lambda, bU^T\mathbf{X}^T\mathbf{Y}).$$

So it suffices to compute $C(A,\gamma)$ for some positive semi-definite, diagonal matrix A and vector γ. Let a_i be the i^{th} diagonal entry of A and let γ_i be the i^{th} element of γ. Kume and Wood [KW05] derive the following expression for $C(A,\gamma)$

$$C(A,\gamma) = (2\pi)^{d/2}\left(\prod_{i=1}^d a_i^{-1/2}\right)\exp\left(\frac{1}{4}\sum_{i=1}^d \frac{\gamma_i^2}{a_i}\right) f_{A,\gamma}(1),$$

where $f_{A,\gamma}$ is the probability density of a non-central chi-squared random variable $\sum_{i=1}^d z_i^2$ with $z_i \sim \mathcal{N}(\frac{\gamma_i}{2a_i}, \frac{1}{2a_i})$. There are number of efficient techniques for computation of $f_{A,\gamma}(1)$ [Imh61; KW05]. We first present the technique of Imhof [Imh61] for exact computation of $f_{A,\gamma}(1)$. Imhof [Imh61] showed that $f_{A,\gamma}(1)$ can be written as the following integral

$$f_{A,\gamma}(1) = \pi^{-1}\int_0^\infty [\rho(u)]^{-1}\cos\zeta(u)du,$$

where $\rho : \mathbb{R} \to \mathbb{R}$ and $\zeta : \mathbb{R} \to \mathbb{R}$ are defined as

$$\begin{aligned}
\zeta(u) &= \frac{1}{2}\sum_{i=1}^d \left(\tan^{-1}\left(\frac{u}{2a_i}\right) + \frac{\gamma_i^2}{8a_i^3}\left(1+\frac{u^2}{4a_i^2}\right)^{-1} u\right) - \frac{1}{2}u,\\
\rho(u) &= \prod_{i=1}^d \left(1+\frac{u^2}{4a_i^2}\right)^{1/4}\exp\left(\frac{1}{32}\frac{(u\gamma_i/a_i^2)^2}{1+\frac{u^2}{4a_i^2}}\right).
\end{aligned}$$

One can rely on numerical integration techniques to compute the above integral to desired accuracy. In our analysis of the approximation error of Algorithm 9, we assume the error from the computation of $f_{A,\gamma}(1)$ is negligible.

Before we conclude this subsection, we present another technique for computation of $f_{A,\gamma}(1)$, which is typically faster than the above approach. This approach was proposed by Kume and Wood [KW05] and relies on the saddle point density approximation technique. While this approach is faster, the downside of it is that it only provides an approximate estimate of $f_{A,\gamma}(1)$. To explain this method, we first present some facts about non-central chi-squared random variables. The cumulant generating function of a non-central chi-squared random variable with density $f_{A,\gamma}$ is given by

$$K(t) = \sum_{i=1}^d \left(-\frac{1}{2}\log\left(1-\frac{t}{a_i}\right) + \frac{1}{4}\frac{\gamma_i^2}{a_i - t} - \frac{\gamma_i^2}{4a_i}\right) \quad (t < \min_i a_i).$$

The first derivative of $K(t)$ is given by

$$K^{(1)}(t) = \sum_{i=1}^{d} \left(\frac{1}{2} \frac{1}{a_i - t} + \frac{1}{4} \frac{\gamma_i^2}{(a_i - t)^2} \right),$$

and higher derivatives are given by

$$K^{(j)}(t) = \sum_{i=1}^{d} \left(\frac{(j-1)!}{2} \frac{1}{(a_i - t)^j} + \frac{j!}{4} \frac{\gamma_i^2}{(a_i - t)^{j+1}} \right), \quad (j \geq 2).$$

Let $\hat{t}$ be the unique solution in $(-\infty, \min_i a_i)$ to the saddle point equation $K^{(1)}(\hat{t}) = 1$. Kume and Wood [KW05] show that $\hat{t}$ has finite upper and lower bounds

$$\min_i a_i - \frac{d}{4} - \frac{1}{2} \left(\frac{d^2}{4} + d \max_i \gamma_i^2 \right)^{1/2} \leq \hat{t} \leq \min_i a_i - \frac{1}{4} - \frac{1}{2} \left(\frac{1}{4} + \gamma_{\min}^2 \right)^{1/2},$$

where $\gamma_{\min}$ is equal to γ_{i^*} for $i^* = \operatorname{argmin}_i a_i$. So, to find $\hat{t}$, one can perform grid search in the above range. Given $\hat{t}$, the first-order saddle point density approximation of $f_{A,\gamma}(1)$ is given by

$$\hat{f}_{A,\gamma,1}(1) = \left(2\pi K^{(2)}(\hat{t})\right)^{-1/2} \exp(K(\hat{t}) - \hat{t}).$$

The second-order saddle point density approximation of $Z_{g,h}(1)$ is given by

$$\hat{f}_{A,\gamma,2}(1) = \hat{f}_{A,\gamma,1}(1)(1 + T),$$

where $T = \frac{1}{8}\hat{\rho}_4 - \frac{5}{24}\hat{\rho}_3^2$, where $\hat{\rho}_j = K^{(j)}(\hat{t})/(K^{(2)}(\hat{t}))^{j/2}$.

D.6.3 Proof of Theorem 20

Before we present the proof of the Theorem we present useful intermediate results which we require in our proof.

Intermediate Results

Lemma 57 (Lipschitz Continuity)**.** *Consider the problem of linear regression described in Section 5.3.2. Let $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq B\}$ and let $\hat{\theta}$ be any estimator which maps the data $D_n = \{(X_i, Y_i)\}_{i=1}^n$ to an element in Θ. Then the risk $R(\hat{\theta}, \theta) = \mathbb{E}_{D_n}\left[\|\hat{\theta}(D_n) - \theta\|_2^2\right]$ is Lipschitz continuous in its second argument w.r.t ℓ_2 norm over the domain Θ, with Lipschitz constant $4(B + B^2\sqrt{nd})$. Moreover, the risk $R(\hat{\theta}, b\mathbf{e}_1) = \mathbb{E}_{D_n}\left[\|\hat{\theta}(D_n) - b\mathbf{e}_1\|_2^2\right]$ is Lipschitz continuous in b over the domain $[0, B]$, with Lipschitz constant $4(B + B^2\sqrt{n})$.*

Proof. Let $R_{\hat{\theta}}(\theta) = R(\hat{\theta}, \theta)$. The gradient of $R_{\hat{\theta}}(\theta)$ with respect to θ is given by

$$\nabla_\theta R_{\hat{\theta}}(\theta) = \mathbb{E}_{D_n}\left[2(\theta - \hat{\theta}(D_n))\right] + \mathbb{E}_{D_n}\left[\|\hat{\theta}(D_n) - \theta\|_2^2 \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)\right],$$

where $\mathbf{X} = [X_1, X_2, \dots X_n]^T, \mathbf{Y} = [Y_1, \dots Y_n]$. The norm of $\nabla_\theta R_{\hat{\theta}}(\theta)$ can be upper bounded as follows

$$\begin{aligned}
\|\nabla_\theta R_{\hat{\theta}}(\theta)\|_2 &\leq \left\|\mathbb{E}_{D_n}\left[2(\theta - \hat{\theta}(D_n))\right]\right\|_2 + \left\|\mathbb{E}_{D_n}\left[\|\hat{\theta}(D_n) - \theta\|_2^2 \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)\right]\right\|_2 \\
&\overset{(a)}{\leq} 4B + \mathbb{E}_{D_n}\left[\|\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)\|_2 \|\hat{\theta}(D_n) - \theta\|_2^2\right] \\
&\overset{(b)}{\leq} 4B + 4B^2 \mathbb{E}_{D_n}\left[\|\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)\|_2\right] \\
&\leq 4B + 4B^2\sqrt{nd},
\end{aligned}$$

where the first term in (a) follows from the fact that $\theta, \hat{\theta}(X) \in \Theta$ and the second term follows from Jensen's inequality. This shows that $R_{\hat{\theta}}(\theta)$ is Lipschitz continuous over Θ. This finishes the first part of the proof. To show that $R(\hat{\theta}, b\mathbf{e}_1)$ is Lipschitz continuous in b, we use similar arguments. Let $R_{\hat{\theta}}(b) = R(\hat{\theta}, b\mathbf{e}_1)$. Then

$$\begin{aligned}
\left|R'_{\hat{\theta}}(b)\right| &= \left|\left\langle \mathbf{e}_1, \nabla_\theta R_{\hat{\theta}}(\theta)\Big|_{\theta = b\mathbf{e}_1}\right\rangle\right| \\
&\overset{(a)}{\leq} \left|\mathbb{E}_{D_n}\left[2(b - [\hat{\theta}(D_n)]_1)\right]\right| + \left\|\mathbb{E}_{D_n}\left[\mathbf{e}_1^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)\|\hat{\theta}(D_n) - b\mathbf{e}_1\|_2^2\right]\right\|_2 \\
&\leq 4B + 4B^2\mathbb{E}_{D_n}\left[|\mathbf{e}_1^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\theta)|\right] \\
&\leq 4B + 4B^2\sqrt{n},
\end{aligned}$$

where (a) follows from our bound for $\|\nabla_\theta R_{\hat{\theta}}(\theta)\|_2$ obtained above. □

Lemma 58 (Approximation of risk). *Consider the setting of Lemma 57. Let $\hat{\theta}$ be any estimator which maps D_n to an element in Θ. Let $\{D_{n,k}\}_{k=1}^N$ be N independent datasets generated from the linear regression model with true parameter θ. Then with probability at least $1 - \delta$*

$$\left|\frac{1}{N}\sum_{i=1}^N \|\hat{\theta}(D_{n,i}) - \theta\|_2^2 - R_{\hat{\theta}}(\theta)\right| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N}}$$

Proof. The proof of the Lemma relies on concentration properties of sub-Gaussian random variables. Let $Z(D_n) = \|\hat{\theta}(D_n) - \theta\|^2$. Note that $R_{\hat{\theta}}(\theta) = \mathbb{E}_{D_n}[Z(D_n)]$. Since $Z(D_n)$ is bounded by $4B^2$, it is a sub-Gaussian random variable. Using Hoeffding bound we get

$$\left|\frac{1}{N}\sum_{i=1}^N Z(D_{n,i}) - \mathbb{E}[Z(D_n)]\right| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N}}, \quad \text{w.p} \geq 1 - \delta.$$

□

Main Argument

The proof uses exactly the same arguments as in the proof of Theorem 19. The only difference between the two proofs are the Lipschitz constants derived in Lemmas 55, 57. The Lipschitz constant in the case of regression is $O(B + B^2\sqrt{n})$, whereas in the case of finite Gaussian sequence model it is $O(B + B^2)$.

Approximation Error of Algorithm 8 There are two causes for error in the optimization oracle described in Algorithm 8: (a) grid search and (b) approximate computation of risk $R(\hat{\theta}, b\mathbf{e}_1)$. We now bound the error due to both (a) and (b). From Lemma 58 we know that for any estimator $\hat{\theta}_i$ and grid point b_j, the following holds with probability at least $1-\delta$

$$\left|\frac{1}{N_1}\sum_{k=1}^{N_1}\|\hat{\theta}_i(D_{n,k}) - b_j\mathbf{e}_1\|_2^2 - R(\hat{\theta}_i, b_j\mathbf{e}_1)\right| \leq 4B^2\sqrt{\frac{\log\frac{1}{\delta}}{N_1}}.$$

Taking a union bound over all estimators $\{\hat{\theta}_i\}_{i=1}^T$ and grid points $\{b_j\}_{j=1}^{B/w}$, we can show that with probability at least $1-\delta$, the following holds for all $i \in [T], j \in [B/w]$

$$\left|\frac{1}{N_1}\sum_{k=1}^{N_1}\|\hat{\theta}_i(D_{n,k}) - b_j\mathbf{e}_1\|_2^2 - R(\hat{\theta}_i, b_j\mathbf{e}_1)\right| \leq 4B^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}}. \tag{D.15}$$

Let $f_{t,\sigma}(b)$ be the actual objective we would like to optimize in iteration t of Algorithm 5, which is given by

$$f_{t,\sigma}(b) = \sum_{i=1}^{t-1} R(\hat{\theta}_i, b\mathbf{e}_1) + \sigma b.$$

Let $\hat{f}_{t,\sigma}(b)$ be the approximate objective we are optimizing by replacing $R(\hat{\theta}_i, b\mathbf{e}_1)$ with its approximate estimate. Let b_t^* be a maximizer of $f_{t,\sigma}(b)$ and $b_{t,\text{approx}}^*$ be the maximizer of $\hat{f}_{t,\sigma}(b)$ (which is also the output of Algorithm 8). Finally, let $b_{t,\text{NN}}^*$ be the point on the grid which is closest to b_t^*. Using Lemma 57 we first show that $f_{t,\sigma}(b)$ is Lipschitz continuous in b. The derivative of $f_{t,\sigma}(b)$ with respect to b is given by

$$f'_{t,\sigma}(b) = \sum_{i=1}^{t-1}\left\langle \mathbf{e}_1, \nabla_\theta R(\hat{\theta}_i, \theta)\Big|_{\theta = b\mathbf{e}_1}\right\rangle + \sigma$$

Using Lemma 57, the magnitude of $f'_{t,\sigma}(b)$ can be upper bounded as

$$|f'_{t,\sigma}(b)| \leq 4(t-1)(B + B^2\sqrt{n}) + \sigma.$$

This shows that $f_{t,\sigma}(b)$ is Lipschitz continuous in b. We now bound $f_{t,\sigma}(b_t^*) - f_{t,\sigma}(b_{t,\text{approx}}^*)$, the approximation error of the optimization oracle

$$\begin{aligned}
f_{t,\sigma}(b_t^*) &\overset{(a)}{\leq} f_{t,\sigma}(b_{t,\text{NN}}^*) + \left(4t(B + B^2\sqrt{n}) + \sigma\right) w \\
&\overset{(b)}{\leq} \hat{f}_{t,\sigma}(b_{t,\text{NN}}^*) + 4tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \left(4t(B + B^2\sqrt{n}) + \sigma\right) w \\
&\overset{(c)}{\leq} \hat{f}_{t,\sigma}(b_{t,\text{approx}}^*) + 4tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \left(4t(B + B^2\sqrt{n}) + \sigma\right) w \\
&\overset{(d)}{\leq} f_{t,\sigma}(b_{t,\text{approx}}^*) + 8tB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + \left(4t(B + B^2\sqrt{n}) + \sigma\right) w,
\end{aligned}$$

where (a) follows from Lipschitz property of the loss function and $(b), (d)$ follow from Equation (D.15) and hold with probability at least $1-\delta$ and (c) follows from the optimality of $b^*_{t,\text{approx}}$. This shows that Algorithm 8 is a $\left(O\left(TB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + TB(1+B\sqrt{n})w\right), w\right)$-approximate maximization oracle; that is

$$\alpha = O\left(TB^2\sqrt{\frac{\log\frac{BT}{w\delta}}{N_1}} + TB(1+B\sqrt{n})w\right), \quad \beta = w.$$

Approximation Error of Algorithm 9 There are two sources of approximation error in Algorithm 9: (a) computation of mean and normalization constant of FB distribution, and (b) approximation of P_t with its samples. In this analysis we assume that mean and normalization constant of FB distribution can be computed to very high accuracy. So the main focus here is on bounding the error from approximation of P_t.

First, note that since we are using grid search to optimize the maximization problem, the true distribution P_t for which we are supposed to compute the Bayes estimator is a discrete distribution supported on grid points $\{b_1, \dots b_{B/w}\}$. Algorithm 9 does not compute the Bayes estimator for P_t. Instead, we generate samples from P_t and use them as a proxy for P_t. Let $\hat{P}_t$ be the empirical distribution obtained by sampling N_2 points from P_t. Let $p_{t,j}$ be the probability mass on grid point b_j. Using Bernstein inequality we can show that the following holds with probability at least $1-\delta$

$$\forall j \in [B/w] \quad |\hat{p}_{t,j} - p_{t,j}| \leq \sqrt{p_{t,j}\frac{\log\frac{B}{w\delta}}{N_2}}. \tag{D.16}$$

Define estimators $\hat{\theta}'_t, \hat{\theta}_t$ as

$$\hat{\theta}'_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right], \quad \hat{\theta}_t \leftarrow \operatorname*{argmin}_{\hat{\theta}\in\mathcal{D}_G} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right].$$

$\hat{\theta}'_t$ is what we ideally want to compute. $\hat{\theta}_t$ is what we end up computing using Algorithm 9. We now show that $\hat{\theta}_t$ is an approximate minimizer of the left hand side optimization problem above. To this end, we try to bound the following quantity

$$\mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}_t, b\mathbf{e}_1) - R(\hat{\theta}'_t, b\mathbf{e}_1)\right].$$

Let $f_t(\hat{\theta}) = \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right]$ and $\hat{f}_t(\hat{\theta}) = \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right]$. We would like to bound the quantity $f_t(\hat{\theta}_t) - f_t(\hat{\theta}'_t)$. Consider the following

$$\begin{aligned} f_t(\hat{\theta}_t) &\overset{(a)}{\leq} \hat{f}_t(\hat{\theta}_t) + \frac{4B^3}{w}\sqrt{\frac{\log \frac{B}{w\delta}}{N_2}} \\ &\overset{(b)}{\leq} \hat{f}_t(\hat{\theta}'_t) + \frac{4B^3}{w}\sqrt{\frac{\log \frac{B}{w\delta}}{N_2}} \\ &\overset{(c)}{\leq} f_t(\hat{\theta}'_t) + \frac{8B^3}{w}\sqrt{\frac{\log \frac{B}{w\delta}}{N_2}}, \end{aligned}$$

where (a) follows from Equation (D.16) and the fact that the risk $R(\hat{\theta}, \theta)$ of any estimator is bounded by $4B^2$, (b) follows since $\hat{\theta}_t$ is a minimizer of $\hat{f}_t$ and (c) follows from Equation (D.16). This shows that with probability at least $1-\delta$, Algorithm 9 is an $O\left(\frac{B^3}{w}\sqrt{\frac{\log \frac{B}{w\delta}}{N_2}}\right)$-approximate optimization oracle; that is,

$$\alpha' = O\left(\frac{B^3}{w}\sqrt{\frac{\log \frac{B}{w\delta}}{N_2}}\right).$$

The rest of the proof is same as the proof of Theorem 19 and involves substituting the approximation errors computed above in Corollary 4.

Minimax Estimator We now show that $\hat{\theta}_{\text{AVG}}$ is an approximate minimax estimator. Instantiating Corollary 4 for the reduced statistical game gives us the following bound, which holds with probability at least $1-\delta$

$$\begin{aligned} \sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq \inf_{\hat{\theta}\in\mathcal{D}_G} \sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) \\ + \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}} + \alpha + \alpha' + \beta B(B\sqrt{n}+1)\sqrt{T}\right), \end{aligned}$$

where we used the fact that the risk $R(\hat{\theta}, b\mathbf{e}_1)$ is $4B(B\sqrt{n}+1)$-Lipschitz continuous w.r.t b. The $\tilde{O}$ notation in the above inequality hides logarithmic factors. Plugging in the values of α, α', β in the above equation gives us

$$\sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq \inf_{\hat{\theta}\in\mathcal{D}_G} \sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) + \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}}\right).$$

We now convert this bound to a bound on the original statistical game. From Theorem 13 we know that $\inf_{\hat{\theta}\in\mathcal{D}_G} \sup_{b\in[0,B]} R(\hat{\theta}, b\mathbf{e}_1) = \inf_{\hat{\theta}\in\mathcal{D}} \sup_{\theta\in\Theta} R(\hat{\theta}, \theta) = R^*$. Since the estimator $\hat{\theta}_{\text{AVG}}$ is invariant to transformations of orthogonal group, we have $R(\hat{\theta}_{\text{AVG}}, \theta) = R(\hat{\theta}_{\text{AVG}}, \|\theta\|_2\mathbf{e}_1)$ for any $\theta \in \Theta$. Using these two results in the above inequality, we get

$$\sup_{\theta\in\Theta} R(\hat{\theta}_{\text{AVG}}, \theta) = \sup_{b\in[0,B]} R(\hat{\theta}_{\text{AVG}}, b\mathbf{e}_1) \leq R^* + \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}}\right).$$

This shows that the worst-case risk of $\hat{\theta}_{\text{AVG}}$ is close to the minimax risk R^*. This finishes the first part of the proof.

LFP To prove the second part, we rely on Corollary 4. Instantiating it for the reduced statistical game gives us

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}} + \alpha + \alpha' + \beta B(B\sqrt{n}+1)\sqrt{T}\right).$$

Plugging in the values of α, α', β in the above equation gives us

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim P_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}}\right).$$

From Equation (D.12) we know that P_t is close to $\hat{P}_t$ with high probability. Using this, we can replace P_t in the above bound with $\hat{P}_t$ and obtain the following bound, which holds with probability at least $1-\delta$

$$\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] \geq R^* - \tilde{O}\left(\frac{B^2(B\sqrt{n}+1)}{\sqrt{T}}\right). \tag{D.17}$$

In the rest of the proof, we show that $\inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right] = \inf_{\hat{\theta}} R(\hat{\theta}, \hat{P}_{\text{LFP}})$. From the definition of $\hat{P}_{\text{LFP}}$, we can equivalently rewrite $R(\hat{\theta}, \hat{P}_{\text{LFP}})$ for any estimator $\hat{\theta}$ as

$$R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[\mathbb{E}_{\theta\sim U}\left[R(\hat{\theta}, b\theta)\right]\right],$$

where U is the uniform distribution on the $(d-1)$ dimensional unit sphere centered at origin, in $\mathbb{R}^d$. Next, from Lemma 53, we know that the Bayes estimator corresponding to any invariant prior is an invariant estimator. Since $\hat{P}_{\text{LFP}}$ is an invariant distribution, we have

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[\mathbb{E}_{\theta\sim U}\left[R(\hat{\theta}, b\theta)\right]\right].$$

From Lemma 52 we know that for any invariant estimator $\hat{\theta}$, we have $R(\hat{\theta}, \theta_1) = R(\hat{\theta}, \theta_2)$, whenever $\theta_1 \sim \theta_2$. Using this result in the above equation gives us

$$\inf_{\hat{\theta}\in\mathcal{D}} R(\hat{\theta}, \hat{P}_{\text{LFP}}) = \inf_{\hat{\theta}\in\mathcal{D}_G} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{b\sim \hat{P}_t}\left[R(\hat{\theta}, b\mathbf{e}_1)\right].$$

Combining the above result with Equation (D.17) shows that $\hat{P}_{\text{LFP}}$ is approximately least favorable.

D.7 Covariance Estimation

D.7.1 Proof of Proposition 10

In this proof, we rely on permutation invariant functions and a representer theorem for such functions. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called permutation invariant, if for any permutation π and any $X \in \mathbb{R}^d$

$$f(\pi(X)) = f(X).$$

The following proposition provides a representer theorem for such functions.

Proposition 20 (Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, and Smola [Zah+17])**.** *A function $f(X)$ from $\mathbb{R}^d$ to $\mathbb{R}$ is permutation invariant and continuous iff it can be decomposed in the form $\rho(\sum_{i=1}^d \phi(X_i))$, for some suitable transformations $\phi : \mathbb{R} \to \mathbb{R}^{d+1}$ and $\rho : \mathbb{R}^{d+1} \to \mathbb{R}$.*

We now prove Proposition 10. First note that from Blackwell's theorem we know that there exists a minimax estimator which is just a function of the sufficient statistic, which in this case is the empirical covariance $S_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ [see Theorem 2.1 of IH81]. So we restrict ourselves to estimators which are functions of S_n. This, together with Theorem 12, shows that there is a minimax estimator which is a function S_n and which is invariant under the action of the orthogonal group $\mathbb{O}(d)$. Let $\hat{\Sigma}$ be such an estimator. Since $\hat{\Sigma}$ is an invariant estimator, it satisfies the following equality for any orthogonal matrix V

$$\hat{\Sigma}(VS_nV^T) = V\hat{\Sigma}(S_n)V^T.$$

Setting $V = U^T$ in the above equation, we get $\hat{\Sigma}(S_n) = U\hat{\Sigma}(\Delta)U^T$. Hence, $\hat{\Sigma}$ is completely determined by it's action on diagonal matrices. So, in the rest of the proof we try to understand $\hat{\Sigma}(\Delta)$. Again relying on invariance of $\hat{\Sigma}$ and setting $V = \Delta' U^T$ for some diagonal matrix Δ' with diagonal elements ±1, we get

$$\hat{\Sigma}(\Delta'\Delta\Delta') = \Delta' U^T \hat{\Sigma}(S_n) U \Delta' \overset{(a)}{=} \Delta'\hat{\Sigma}(\Delta)\Delta',$$

where (a) follows from the fact that $\hat{\Sigma}(S_n) = U\hat{\Sigma}(\Delta)U^T$. Since $\Delta'\Delta\Delta' = \Delta$, the above equation shows that $\Delta'\hat{\Sigma}(\Delta)\Delta' = \hat{\Sigma}(\Delta)$ for any diagonal matrix Δ' with diagonal elements ± 1. This shows that $\hat{\Sigma}(\Delta)$ is a diagonal matrix. Next, we set $V = P_\pi U^T$, where P_π is the permutation matrix corresponding to some permutation π. This gives us

$$\hat{\Sigma}(P_\pi \Delta P_\pi^T) = P_\pi \hat{\Sigma}(\Delta) P_\pi^T.$$

This shows that for any permutation π, $\hat{\Sigma}(\pi(\Delta)) = \pi(\hat{\Sigma}(\Delta))$, where $\pi(\Delta)$ represents permutation of the diagonal elements of Δ. In the rest of the proof, we use the notation Δ_i to denote the i^{th} diagonal entry of Δ and $\hat{\Sigma}_i(\Delta)$ to denote the i^{th} diagonal entry of $\hat{\Sigma}(\Delta)$. The above property of $\hat{\Sigma}$ shows that $\hat{\Sigma}_i(\Delta)$ doesn't depend on the ordering of the elements in $\{\Delta_j\}_{j\neq i}$. This follows by choosing any permutation π which keeps the i^{th} element fixed. Next, by considering the permutation which only exchanges positions 1 and i, we get

$$\hat{\Sigma}_i(\Delta_1, \ldots \Delta_i, \ldots \Delta_d) = \hat{\Sigma}_1(\Delta_i, \ldots \Delta_1, \ldots \Delta_d).$$

Thus $\hat{\Sigma}_i$ can be expressed in terms of $\hat{\Sigma}_1$. Represent $\hat{\Sigma}_1$ by $\hat{\Sigma}_0$. Combining the above two properties, we have

$$\hat{\Sigma}_i(\Delta) = \hat{\Sigma}_0(\Delta_i, \{\Delta_j\}_{j\neq i}),$$

where $\{\Delta_j\}_{j\neq i}$ represents the independence of $\hat{\Sigma}_0$ on the ordering of elements $\{\Delta_j\}_{j\neq i}$. Now, consider the function $\hat{\Sigma}_0(\Delta_1, \{\Delta_j\}_{j=2}^d)$. For any fixed a, and $\Delta_1 = a$, $\hat{\Sigma}_0(a, \{\Delta_j\}_{j=2}^d)$ is a permutation invariant function. Using Proposition 20, $\hat{\Sigma}_0(a, \{\Delta_j\}_{j=2}^d)$ can be written as

$$\hat{\Sigma}_0(a, \{\Delta_j\}_{j=2}^d) = f_a\left(\sum_{j=2}^d g_a(\Delta_j)\right),$$

for some functions f_a, g_a. We overload the notation and define $f_a(x) = f(a,x)$ and $g_a(x) = g(a,x)$. Using this, we can represent $\hat{\Sigma}_i(\Delta)$ as

$$\hat{\Sigma}_i(\Delta) = f\left(\Delta_i, \sum_{j\neq i} g(\Delta_i, \Delta_j)\right),$$

for some functions f, g. There is a small technicality which we ignored while using Proposition 20 on $\hat{\Sigma}_0$. Proposition 20 only holds for continuous functions. Since $\hat{\Sigma}_0$ is not guaranteed to be continuous, the proposition can't be used on this function. However, this is not an issue because any measurable function is a limit of continuous functions. Since $\hat{\Sigma}_0$ is a measurable function, it can be approximated arbitrarily close in the form of $f_a\left(\sum_{j=2}^d g_a(\Delta_j)\right)$.

To conclude the proof of the proposition, we note that

$$\inf_{\hat{\Sigma}\in\mathcal{M}_{\mathcal{D},G}} \sup_{\lambda\in\Xi_G} R(\hat{\Sigma}, \mathrm{Diag}(\lambda)) = \inf_{\hat{\Sigma}\in\mathcal{M}_{f,g}} \sup_{\lambda\in\Xi_G} R(\hat{\Sigma}, \mathrm{Diag}(\lambda)).$$

This is because the minimax estimator can be approximated arbitrarily well using estimators of the form $\hat{\Sigma}_i(\Delta) = f\left(\Delta_i, \sum_{j\neq i} g(\Delta_i, \Delta_j)\right)$ and the fact that the model class has absolutely continuous distributions.

D.8 Entropy Estimation

D.8.1 Proof of Proposition 11

First note that any estimator of entropy is a function of $\hat{P}_n$, which is a sufficient statistic for the problem. This, together with Theorem 12, shows that there is a minimax estimator which is a function of $\hat{P}_n$ and which is invariant under the action of permutation group. Let $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ be such an estimator. Since $\hat{f}$ is invariant, it satisfies the following property for any permutation π

$$\hat{f}(\pi(\hat{P}_n)) = \hat{f}(\hat{P}_n).$$

If $\hat{f}(\hat{P}_n)$ is continuous, then Proposition 20 shows that it can written as $g\left(\sum_{j=1}^{d} h(\hat{p}_j)\right)$, for some functions $h : \mathbb{R} \to \mathbb{R}^{d+1}, g : \mathbb{R}^{d+1} \to \mathbb{R}$. Even if it is not continuous, since it is a measurable function, it is a limit of continuous functions. So $\hat{f}$ can be approximated arbitrarily close in the form of $g\left(\sum_{j=1}^{d} h(\hat{p}_j)\right)$. This also implies the statistical game in Equation (5.17) can reduced to the following problem

$$\inf_{\hat{f} \in \mathcal{M}_{\mathcal{D},G}} \sup_{P \in \mathcal{P}_G} R(\hat{f}, P) = \inf_{\hat{f} \in \mathcal{M}_{g,h}} \sup_{P \in \mathcal{P}_G} R(\hat{f}, P).$$

D.9 Experiments

D.9.1 Covariance Estimation

In this section, we compare the performance of various estimators at randomly generated Σ's. We use beta distribution to randomly generate Σ's with varying spectral decays and compute the average risks of all the estimators at these Σ's. Figure D.1 presents the results from this experiment. It can be seen that our estimator has better average case performance than empirical and James Stein estimators.

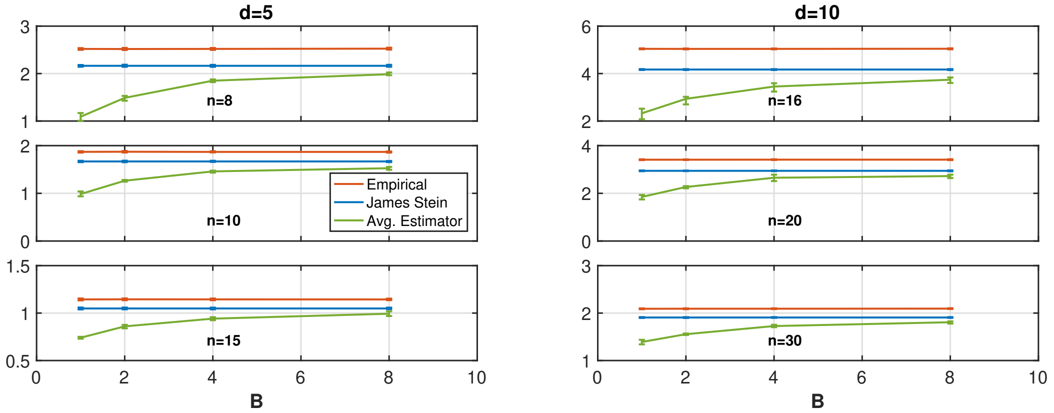


Figure D.1: Risk of various estimators for covariance estimation evaluated at randomly generated Σ's. We generated multiple Σ's whose eigenvalues are randomly sampled from a Beta distribution with various parameters and averaged the risks of estimators at these Σ's. Plots on the left correspond to $d = 5$ and the plots on the right correspond to $d = 10$.

D.9.2 Entropy Estimation

In this section, we compare the performance of various estimators at randomly generated P's. We use beta distribution to randomly generate P's and compute the average risks of all the estimators at these P's. Figure D.2 presents the results from this experiment.

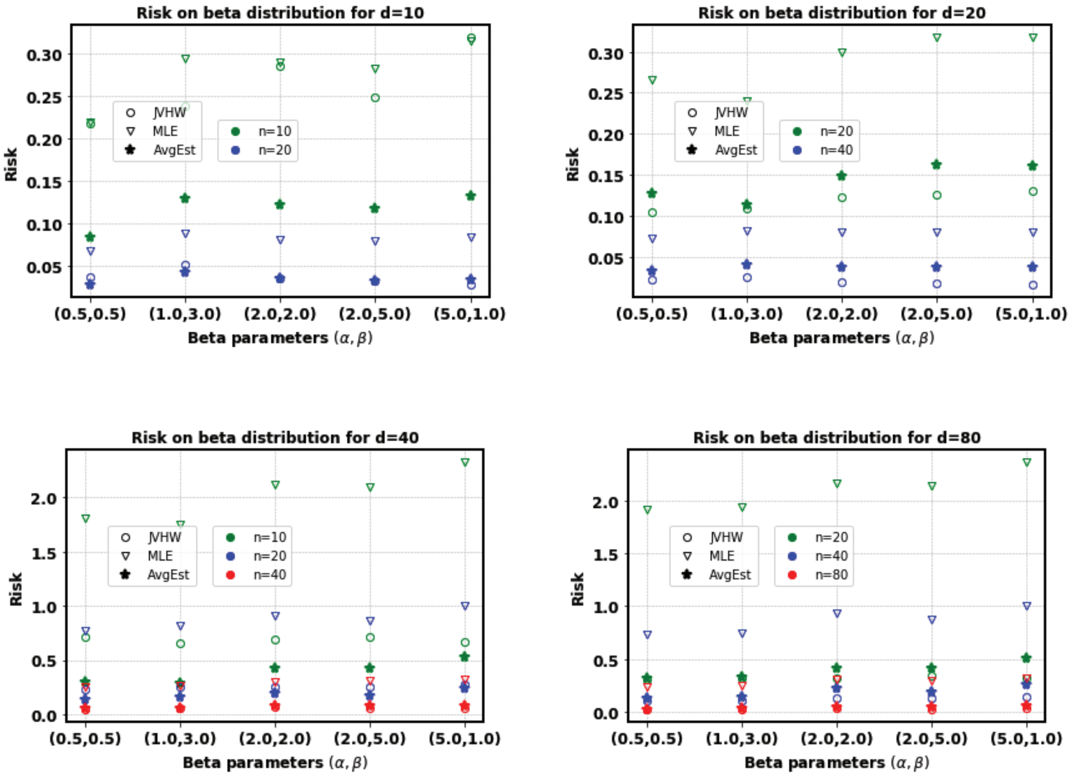


Figure D.2: Risk of various estimators for entropy estimation evaluated at randomly generated distributions. We generated multiple P's with p_i's sampled from a Beta distribution and averaged the risks of estimators at these P's.

Appendix E

Supplementary Material for Chapter 6

E.1 Notation and Terminology

Notation

Symbol	Description
X	feature vector
Y	label
$\mathcal{X}$	domain of feature vector
$\mathcal{Y}$	domain of the label
K	number of classes in multi-class classification problem
S	data set
P	true data distribution
P^X, P^Y	marginal distributions of X, Y
P_n	empirical distribution
P_n^X, P_n^Y	empirical marginal distributions of X, Y in data set S
$f : \mathcal{X} \to \mathbb{R}^K$	score based classifier
ϕ	feature transformer
W	linear classifier on top of feature transformer
ℓ_{0-1}	0/1 classification loss
ℓ	convex surrogate of ℓ_{0-1}
$R(f)$	population risk of classifier f, measured w.r.t ℓ
$\widehat{R}_S(f)$	empirical risk of classifier f, measured w.r.t ℓ
$R(W, \phi)$	population risk of classifier $f = W\phi$, measured w.r.t ℓ
$\widehat{R}_S(W, \phi)$	empirical risk of classifier $f = W\phi$, measured w.r.t ℓ
$L_2(P)$	set of square integrable functions w.r.t P
$f \circ g(\mathbf{x})$	denotes function composition $f(g(\mathbf{x}))$
$[\phi_0, \dots, \phi_t](\mathbf{x})$	denotes concatenation of vectors $\phi_0(\mathbf{x}) \dots \phi_t(\mathbf{x})$
$\mathcal{F}$	hypothesis class of weak classifiers
$\mathcal{G}$	hypothesis class of weak feature transformers
$\mathcal{G}_t$	hypothesis class of weak feature transformers used in the t^{th} iteration of greedy
$\mathcal{W}$	hypothesis class of linear classifiers on top of feature transformers

Terminology

Term	Description
Additive Boosting	Classical boosting framework which constructs a strong classifier using additive combinations of weak classifiers
Additive Feature Boosting	Feature boosting framework which constructs a strong classifier using additive combinations of weak feature transformers with a linear classifier on top of the feature transformer
Weak classifier	Any classifier which by itself doesn't achieve good performance on a given classification task and whose performance we wish to boost
Weak feature transformer	Any feature transformation which by itself doesn't provide good performance on a given classification task and whose performance we wish to boost

E.2 Proof of Proposition 12

Notation. We use the notation of Huang, Ash, Langford, and Schapire [Hua+17a] in this proof. We note that this notation will only be used in this section. Later sections use the notation introduced in Section 6.1. We let $g_t(\mathbf{x})$ be the output of the t^{th} residual block, which is given by the following recursion

$$g_t(\mathbf{x}) = f_{t-1} \circ g_{t-1}(\mathbf{x}) + g_{t-1}(\mathbf{x}) = \sum_{i=0}^{t-1} f_i \circ g_i(\mathbf{x}),$$

with g_0, f_0 equal to identity functions. The final output of a depth-T ResNet, given input $\mathbf{x}$, is rendered after a linear classifier $W \in \mathbb{R}^{K\times D}$ on representation $g_{T+1}(\mathbf{x})$. Let W_t be the auxiliary linear classifier on top of the residual block g_t. Define $o_t(\mathbf{x})$ as

$$o_t(\mathbf{x}) \overset{def}{=} W_t g_t(\mathbf{x}).$$

Note that $o_t(\mathbf{x}) = \sum_{i=0}^{t} W_t f_i \circ g_i(\mathbf{x})$. Define $h_t(\mathbf{x})$ as $h_t(\mathbf{x}) \overset{def}{=} \alpha_{t+1} o_{t+1}(\mathbf{x}) - \alpha_t o_t(\mathbf{x})$, where α_t is a scalar. Huang, Ash, Langford, and Schapire [Hua+17a] consider exponential loss in their work, which is defined as

$$\ell(o(\mathbf{x}), y) = \sum_{k\neq y} \exp\left([o(\mathbf{x})]_k - [o(\mathbf{x})]_y\right).$$

Algorithm of Bengio, Lamblin, Popovici, and Larochelle [Ben+07]. Using this notation, the greedy layer-by-layer training technique of Bengio, Lamblin, Popovici, and Larochelle [Ben+07] for learning ResNets is given by the following update rule

$$W_{t+1}, f_t \leftarrow \underset{W,f}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(W\left[f \circ g_t(\mathbf{x}_i) + g_t(\mathbf{x}_i)\right], y_i\right). \tag{E.1}$$

Algorithm of Huang, Ash, Langford, and Schapire [Hua+17a]. The algorithm of Huang, Ash, Langford, and Schapire [Hua+17a] for greedy learning of ResNets is given in Algorithm 17, which is a reproduction of Algorithm 3 of Huang, Ash, Langford, and Schapire [Hua+17a]. Note that the key update step is given in step 2 of Algorithm 18

$$f_t, \alpha_{t+1}, W_{t+1} \leftarrow \underset{f,\alpha,W}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(\alpha W[f \circ g_t(\mathbf{x}_i) + g_t(\mathbf{x}_i)], y_i). \tag{E.2}$$

Since α is a scalar, it can be consumed into the linear classifier W. This shows that the update step of Huang, Ash, Langford, and Schapire [Hua+17a] is equivalent to Equation (E.1).

Algorithm 17 Greedy algorithm of Huang, Ash, Langford, and Schapire [Hua+17a] for learning ResNets

```
**Input:** Training data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, iterations $T$, threshold $\gamma$
Initialize $t \leftarrow 0, \tilde{\gamma}_0 \leftarrow 0, \alpha_0 \leftarrow 0, o_0 \leftarrow \mathbf{0} \in \mathbb{R}^K, s_0(\mathbf{x}_i) = \mathbf{0} \in \mathbb{R}^K, \forall i \in [n]$
Initialize cost function $[C_0(i)]_k \leftarrow \begin{cases} 1 & \text{if } k \neq y_i \\ 1 - K & \text{if } k = y_i \end{cases}, \forall i \in [n], k \in [K]$
**while** $\gamma_t > \gamma$ **do**
    $f_t, \alpha_{t+1}, W_{t+1}, o_{t+1} \leftarrow$ Algorithm 18$(g_t)$
    Compute $\gamma_t \leftarrow \sqrt{\frac{\tilde{\gamma}_{t+1}^2 - \tilde{\gamma}_t^2}{1 - \tilde{\gamma}_t^2}}$, where $\tilde{\gamma}_{t+1} = \frac{-\sum_{i=1}^n C_t(i)^T o_{t+1}(\mathbf{x}_i)}{\sum_{i=1}^n \sum_{k \neq y_i} [C_t(i)]_k}$
    Update $s_{t+1}(\mathbf{x}_i) \leftarrow s_t(\mathbf{x}_i) + h_t(\mathbf{x}_i)$, where $h_t(\mathbf{x}_i) = \alpha_{t+1} o_{t+1}(\mathbf{x}_i) - \alpha_t o_t(\mathbf{x}_i)$
    Update cost function $[C_{t+1}(i)]_k \leftarrow \begin{cases} \exp\left([s_{t+1}(\mathbf{x}_i)]_k - [s_{t+1}(\mathbf{x}_i)]_{y_i}\right) & \text{if } k \neq y_i \\ -\sum_{k' \neq y_i} \exp\left([s_{t+1}(\mathbf{x}_i)]_{k'} - [s_{t+1}(\mathbf{x}_i)]_{y_i}\right) & \text{if } k = y_i \end{cases}, \forall i \in [n], k \in [K]$
    $t \leftarrow t + 1$
**end while**
$T \leftarrow t - 1$
**Return:** $W_{T+1}, \{f_t(\cdot), \forall t\}$
```

Algorithm 18 Training a ResNet module

```
**Input:** $g_t$
$(f_t, \alpha_{t+1}, W_{t+1}) \leftarrow \text{argmin}_{f,\alpha,W} \sum_{i=1}^n \ell(\alpha W[f \circ g_t(\mathbf{x}_i) + g_t(\mathbf{x}_i)], y_i)$
$o_{t+1}(\mathbf{x}) = W_{t+1}[f_t \circ g_t(\mathbf{x}) + g_t(\mathbf{x})]$
**Return:** $f_t, \alpha_{t+1}, W_{t+1}, o_{t+1}$
```

E.3 Proof of Proposition 13

Freund and Schapire [FS95] consider the problem of binary classification with $\mathcal{Y} = \{-1, +1\}$. Let $\mathcal{F}$ be a hypothesis space of weak classifiers mapping $\mathcal{X}$ to $\mathcal{Y}$. Freund and Schapire [FS95] consider the following weak learning condition. For any set of non-negative weights $\{w_i\}_{i=1}^n$ over points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that $\sum_i w_i = 1$, there is a classifier $f \in \mathcal{F}$ which achieves an error at most $\frac{1}{2} - \frac{\beta}{2}$, for some $\beta > 0$. That is, there exists $f \in \mathcal{F}$ such that

$$\sum_{i=1}^n w_i \mathbb{I}(y_i \neq f(\mathbf{x}_i)) \leq \frac{1}{2} - \frac{\beta}{2}.$$

This can equivalently be written as

$$\begin{aligned} \textstyle\sum_{i=1}^n w_i y_i f(\mathbf{x}_i) &= \textstyle\sum_{i:y_i=f(\mathbf{x}_i)} w_i y_i f(\mathbf{x}_i) - \sum_{i:y_i\neq f(\mathbf{x}_i)} w_i y_i f(\mathbf{x}_i) + 2\sum_{i:y_i\neq f(\mathbf{x}_i)} w_i y_i f(\mathbf{x}_i) \\ &= 1 + 2\textstyle\sum_{i:y_i\neq f(\mathbf{x}_i)} w_i y_i f(\mathbf{x}_i) \\ &\geq \beta \\ &= \beta\left(\textstyle\sum_{i=1}^n w_i\right) \end{aligned} \tag{E.3}$$

We now show that this condition implies Definition 6.3.1 in the label space. We first introduce the notion of inner product between functions mapping $\mathcal{X}$ to $\mathbb{R}$. For any f, g mapping $\mathcal{X}$ to $\mathbb{R}$, we define $\langle f, g\rangle_n$ as

$$\langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)g(\mathbf{x}_i).$$

Let the classification loss ℓ be such that $\ell(f(\mathbf{x}), y) = c(yf(\mathbf{x}))$ for some decreasing function $c : \mathbb{R} \to \mathbb{R}$. All the popular classification losses such as logistic, exponential, hinge losses satisfy this assumption. The functional gradient of $\widehat{R}_S$ w.r.t f in the above inner product space is defined as

$$\nabla_f \widehat{R}_S(f)(\mathbf{x}) = \begin{cases} y_i c'(y_i f(\mathbf{x}_i)), & \text{if } \mathbf{x} = \mathbf{x}_i \\ 0, & \text{otherwise} \end{cases},$$

where $c'(z)$ is the derivative of c at z. Note that since c is a decreasing function, $c'(z) < 0$ for any z. Using this notation, it is easy to see that any hypothesis class $\mathcal{F}$ satisfying Equation (E.3) satisfies the following condition for any function $h : \mathcal{X} \to \mathbb{R}$

$$\exists f \in \mathcal{F}, \quad \langle f, -\nabla_f \widehat{R}_S(h)\rangle_n \geq \beta\|\nabla_f \widehat{R}_S(h)\|_1 \geq \frac{\beta}{\sqrt{n}}\|\nabla_f \widehat{R}_S(h)\|_n,$$

where $\|\nabla_f \widehat{R}_S(h)\|_1 = n^{-1}\sum_{i=1}^n |\nabla_f \widehat{R}_S(h)(\mathbf{x}_i)|$. This can be shown by substituting w_i in Equation (E.3) with $-c'(y_i h(\mathbf{x}_i))$. This shows that the weak learning condition of Freund and Schapire [FS95] satisfies the weak learning condition in Definition 6.3.1, albeit in the label space.

E.4 Discussion of Theorem 21

In this section, we discuss the results of Theorem 21.

Remark E.4.1 (Reference Classifier)**.** *The reference classifier (W^*, ϕ^*) in the bound in Theorem 21 can be any classifier, as long as $\|W^*\|_2 < \infty, \|\phi^*\|_{P^X} < \infty$. In particular, if there exists a Bayes optimal classifier satisfying this condition, then the above Theorem provides an excess risk bound w.r.t the Bayes optimal classifier.*

Remark E.4.2 (Breakdown of Rates)**.** *The $T^{-\alpha}$ term in the bound corresponds to the* optimization error*. The $\eta_t\epsilon_t$ term corresponds to the* approximation error *and the rest of the terms correspond to the* generalization error*. As T increases, the optimization error goes down, and as $\tilde{n}$ increases, the generalization error goes down. If there is no approximation error, that is $\epsilon_t = 0$ for all t, then the excess risk goes down to 0 as $\tilde{n}, T \to \infty$ at appropriate rate.*

Remark E.4.3 (Optimization Error)**.** *If $\beta = 1$, then for appropriate choice of step size the optimization error goes down as $O\left(T^{-1/3+\gamma}\right)$, for some arbitrarily small $\gamma > 0$. This rate is slower than the $O(T^{-1})$ rates for inexact gradient descent obtained by Schmidt, Roux, and Bach [SRB11]* and *Devolder, Glineur, and Nesterov [DGN14]. However, we note that unlike our work, these works assume that the level sets of the objective are bounded. Under the assumption that the level sets of population risk are bounded, the optimization error in Theorem 21 can be improved to $O(T^{-1})$. However, such a condition need not hold in the our setting.*

Remark E.4.4 (Lipschitzness of loss)**.** *The assumptions of smoothness and Lipschitzness on ℓ are satisfied by popular loss functions such as logistic loss,* softmax + cross entropy loss*. Consider logistic loss for binary classification $\ell(z, y) = \log(1 + e^{-yz})$. It is easy to verify that $\ell(z, y)$ is 1-Lipschitz and 1-smooth w.r.t. z. Similarly, the softmax + cross entropy loss, which is given by, $\ell(\mathbf{z}, y) = -\mathbf{z}[y] + \log\left(\sum_{k=1}^{K} e^{\mathbf{z}[k]}\right)$ is 1-Lipschitz and 1-smooth w.r.t. $\mathbf{z}$.*

Remark E.4.5 (Bounded Feature Transformers)**.** *The boundedness assumption on the functions in $\mathcal{G}_t$ is satisfied by neural networks made up of bounded activation functions such as* sigmoid, tanh.

Remark E.4.6 (Modular Bounds)**.** *Note that the risk bounds are modular and only depend on the Rademacher complexity terms $\mathcal{R}(\mathcal{W}, \mathcal{G}_t), \mathcal{R}(\mathcal{G}_t)$ which capture the complexity of $\mathcal{G}_t$. To instantiate Theorem 21 for specific choices of $\mathcal{G}_t$, we need to bound these two complexity terms.*

Remark E.4.7 (Bounds on 0/1 risk)**.** *Since $0/1$ loss is upper bounded by surrogate losses such as exponential, logistic loss, our Theorem also provides generalization bounds for $0/1$ loss.*

Remark E.4.8 (Sample Splitting)**.** *A natural question that might arise regarding sample splitting is: "does this make our approach similar to bagging and random forests (RFs)?". We would like to note that even with sample splitting, our approach is not similar to bagging and RFs. Bagging and RFs create ensembles by independently training each base learner. Whereas, in boosting, the base learners are fit greedily and are not independent of each other. Another important distinction between RFs and boosting is that RFs work with complex base classifiers with good predictive power and aim to reduce the variance of these classifiers by averaging the predictions of multiple independently trained base classifiers. Whereas in boosting, one works with base classifiers with very little predictive power (i.e., high bias) and combines multiple such base classifiers to create a strong classifier with good predictive power (i.e., low bias). Viewed this way, our approach is very similar to boosting than RFs.*

E.5 Proof of Theorem 21

E.5.1 Intermediate Results

In this section we present some intermediate results which we use in the proof of Theorem 21. The proof of the Theorem can be found in Section E.5.2.

Lemma 59. *Consider the setting of Theorem 21. Let (W_t, ϕ_t) be the t^{th} iterate generated by Algorithm 10 with Algorithm 12 as update routine. Then for any t, the following holds with probability at least $1-\delta$ over datasets of size n*

$$R(W_t, \phi_t) \leq R(W_{t-1}, \phi_t) + 2\eta_t L \mathcal{R}(\mathcal{W}, \mathcal{G}_t) + \frac{4c\sigma_{max} B L t^{1-s}}{1-s}\left(\sqrt{\frac{\log 2/\delta}{\tilde{n}}} + \sqrt{\frac{K}{\tilde{n}}}\right),$$

where $\mathcal{R}(\mathcal{W}, \mathcal{G}_t)$ is the Rademacher complexity term, which is defined as

$$\mathcal{R}(\mathcal{W}, \mathcal{G}_t) = \mathbb{E}\left[\sup_{\substack{W\in\mathcal{W},\\ g\in\mathcal{G}_t}} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{k=1}^{K} \rho_{ik}[Wg(\mathbf{x}_{t,i})]_k\right],$$

and the expectation is over the randomness from S_t, ρ's.

Proof. Throughout the proof, we condition on the past datasets $S_1, \dots S_{t-1}$ and show that the Lemma holds for any choice of $S_1, \dots S_{t-1}$. Consider the following upper bound for $R(W_t, \phi_t)$

$$\begin{aligned} R(W_t, \phi_t) &\leq \widehat{R}_{S_t}(W_t, \phi_t) + \sup_{W\in\mathcal{W}, g\in\mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)| \\ &\overset{(a)}{\leq} \widehat{R}_{S_t}(W_{t-1}, \phi_t) + \sup_{W\in\mathcal{W}, g\in\mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)| \\ &\leq R(W_{t-1}, \phi_t) + 2\sup_{W\in\mathcal{W}, g\in\mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)|, \end{aligned}$$

where (a) follows from the definition of W_t. We now rely on Rademacher complexity bounds in Theorem 62 to bound the supremum in the RHS. To apply the bound, we first need to ensure $\ell(W\phi_{t-1}(\mathbf{x}) + \eta_t Wg(\mathbf{x}), y)$ is bounded. Since $\sup_X \|g(X)\|_2 \leq B$ and $\lambda_{\max}\left(WW^T\right) \leq \sigma_{\max}^2$, it is easy to see that

$$\sup_X \|W\phi_{t-1}(\mathbf{x}) + \eta_t Wg(\mathbf{x})\|_2 \leq \sigma_{\max} B \sum_{i=1}^{t} \eta_i \leq \frac{c\sigma_{\max} B t^{1-s}}{1-s},$$

where the last inequality follows from the definition of η_t. Since ℓ is L-Lipschitz in its first argument, we can show that $\ell(W\phi_{t-1}(\mathbf{x}) + \eta_t Wg(\mathbf{x}), y)$ lies in an interval of width $\frac{2c\sigma_{\max} B L t^{1-s}}{1-s}$. Applying Theorem 62, we get with probability at least $1-\delta$

$$\begin{aligned} R(W_t, \phi_t) \leq & R(W_{t-1}, \phi_t) + 2\mathbb{E}\left[\sup_{W\in\mathcal{W}, g\in\mathcal{G}_t} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \rho_i \ell(W\phi_{t-1}(\mathbf{x}_{t,i}) + \eta_t Wg(\mathbf{x}_{t,i}), y_{t,i})\right] \\ & + \frac{4c\sigma_{\max} B L t^{1-s}}{1-s}\sqrt{\frac{\log 2/\delta}{\tilde{n}}}. \end{aligned}$$

We now focus on bounding the Rademacher complexity term appearing above. To this end, we rely on the composition property of Rademacher complexity. Since ℓ is L-Lipscthiz in the first argument, applying Theorem 63 we get

$$\begin{aligned}
R(W_t, \phi_t) \leq & R(W_{t-1}, \phi_t) + 2L\mathbb{E}\left[\sup_{W\in\mathcal{W}, g\in\mathcal{G}_t} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \rho_{ik}[W\phi_{t-1}(\mathbf{x}_{t,i}) + \eta_t W g(\mathbf{x}_{t,i})]_k\right] \\
& + \frac{4c\sigma_{\max} BLt^{1-s}}{1-s}\sqrt{\frac{\log 2/\delta}{\tilde{n}}} \\
\leq & R(W_{t-1}, \phi_t) + 2\eta_t L\mathcal{R}(\mathcal{W}, \mathcal{G}_t) + 2L\,\underbrace{\mathbb{E}\left[\sup_{W\in\mathcal{W}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \rho_{ik}[W\phi_{t-1}(\mathbf{x}_{t,i})]_k\right]}_{T_1} \\
& + \frac{4c\sigma_{\max} BLt^{1-s}}{1-s}\sqrt{\frac{\log 2/\delta}{\tilde{n}}}
\end{aligned}$$

T_1 can be bounded as follows. Let $\rho \in \mathbb{R}^{K\times\tilde{n}}$ be the matrix whose $(k,i)^{th}$ entry is given by ρ_{ik} and $\phi_{t-1}(S_t) \in \mathbb{R}^{D\times\tilde{n}}$ be the matrix whose $(j,i)^{th}$ entry is given by $[\phi_{t-1}(\mathbf{x}_{t,i})]_j$. T_1 can be rewritten in terms of $\rho, \phi_{t-1}(S_t)$ as

$$\begin{aligned}
T_1 &= \mathbb{E}\left[\sup_{W\in\mathcal{W}} \frac{1}{\tilde{n}} \left\langle \rho\phi_{t-1}(S_t)^T, W\right\rangle_F\right] \\
&\leq \left[\sup_{W\in\mathcal{W}} \|W\|_2\right] \mathbb{E}\left[\frac{1}{\tilde{n}}\|\rho\phi_{t-1}(S_t)^T\|_F\right] \\
&\leq \sigma_{\max}\mathbb{E}\left[\frac{1}{\tilde{n}}\|\rho\phi_{t-1}(S_t)^T\|_F\right] \\
&\leq \frac{\sigma_{\max}}{\tilde{n}}\sqrt{\mathbb{E}\left[\|\rho\phi_{t-1}(S_t)^T\|_F^2\right]} \\
&= \frac{\sigma_{\max}}{\tilde{n}}\sqrt{K\mathbb{E}\left[\|\phi_{t-1}(S_t)\|_F^2\right]} \leq \sigma_{\max}\sqrt{\frac{K}{\tilde{n}}}\mathbb{E}\left[\sup_X \|\phi_{t-1}(X)\|_2\right] \\
&\leq \frac{c\sigma_{\max}Bt^{1-s}}{1-s}\sqrt{\frac{K}{\tilde{n}}},
\end{aligned}$$

where the last inequality follows from our choice of step size η_t and our assumption on the boundedness of the outputs of functions in $\mathcal{G}_t$. Substituting this upper bound on T_1 in the previous inequality gives us the required bound on $R(W_t, \phi_t)$. □

Lemma 60. *Consider the setting of Theorem 21. Let (W_t, ϕ_t) be the t^{th} iterate generated by Algorithm 10 with Algorithm 12 as update routine. Then for any t, the following holds with probability at least $1-2\delta$ over datasets of size n*

$$\langle g_t, -\nabla_\phi R(W_{t-1}, \phi_{t-1})\rangle_P \geq \beta B\|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P - \epsilon_t - 2\sigma_{max}L\mathcal{R}(\mathcal{G}_t) - 4\sigma_{max}BL\sqrt{\frac{\log 2/\delta}{\tilde{n}}}.$$

Proof. Let $\hat{P}_{\tilde{n},t}$ be the empirical distribution of dataset S_t. Since $\mathcal{G}_t$ satisfies the (β, ϵ_t)-weak learning condition w.r.t dataset S_t, we have

$$\langle g_t, -\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\rangle_{P^X_{\tilde{n},t}} \geq \beta B \|\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\|_{P^X_{\tilde{n},t}} - \epsilon_t.$$

Consider the following lower bound for $\langle g_t, -\nabla_\phi R(W_{t-1}, \phi_{t-1})\rangle_P$

$$\begin{aligned}\langle g_t, -\nabla_\phi R(W_{t-1}, \phi_{t-1})\rangle_P \geq & \underbrace{\langle g_t, -\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\rangle_{P^X_{\tilde{n},t}}}_{T_1} \\ & - \underbrace{\left|\langle g_t, -\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\rangle_{P^X_{\tilde{n},t}} - \langle g_t, -\nabla_\phi R(W_{t-1}, \phi_{t-1})\rangle_P\right|}_{T_2}\end{aligned}$$

We now lower bound each of the terms appearing the RHS of the above inequality. Similar to the proof of Lemma 59, throughout the proof we condition on the past datasets $S_1, \dots S_{t-1}$ and show that the Lemma holds for any choice of $S_1, \dots S_{t-1}$.

Bounding T_1. Using the weak learning condition, T_1 can be lower bounded as

$$T_1 \geq \beta B \|\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\|_{P^X_{\tilde{n},t}} - \epsilon_t.$$

Using triangle inequality, this can be further lower bounded as

$$T_1 \geq \beta B \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P - \beta B \left| \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P - \|\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\|_{P^X_{\tilde{n},t}} \right| - \epsilon_t.$$

We now bound the middle term in the RHS using standard concentration inequalities. Define random variable Z as

$$Z = W_{t-1}^T \nabla \ell(W_{t-1}\phi_{t-1}(X), Y),$$

for $(X, Y) \sim P$ and define $\mathbf{z}_{t,i}$ as

$$\mathbf{z}_{t,i} = W_{t-1}^T \nabla \ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i}),$$

where $\nabla \ell(u, y)$ denotes the gradient of ℓ w.r.t its first argument. Then from the definition of functional gradients $\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1}), \nabla_\phi R(W_{t-1}, \phi_{t-1})$, we have

$$\|\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\|^2_{P^X_{\tilde{n},t}} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \|\mathbf{z}_{t,i}\|^2, \quad \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|^2_P = \mathbb{E}\left[\|Z\|^2\right].$$

Since ℓ is L-Lipschitz, it is easy to see that $\|Z\|$ is a bounded random variable and always lies in the interval $[0, \sigma_{\max} L]$. So using Chernoff bounds in Theorem 61, we can show that the following holds with probability at least $1 - \delta$

$$\left| \sum_{i=1}^{\tilde{n}} \frac{1}{\tilde{n}} \|\mathbf{z}_{t,i}\|^2 - \mathbb{E}\left[\|Z\|^2\right] \right| \leq \sigma_{\max} L \sqrt{\frac{3\mathbb{E}\left[\|Z\|^2\right] \log 1/\delta}{\tilde{n}}}.$$

Now, consider the following

$$\left|\|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P - \|\nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})\|_{P^X_{\tilde{n},t}}\right| = \left|\sqrt{\sum_{i=1}^{\tilde{n}} \frac{1}{\tilde{n}}\|\mathbf{z}_{t,i}\|^2} - \sqrt{\mathbb{E}\left[\|Z\|^2\right]}\right|$$
$$\leq \frac{\left|\sum_{i=1}^{\tilde{n}} \frac{1}{\tilde{n}}\|\mathbf{z}_{t,i}\|^2 - \mathbb{E}\left[\|Z\|^2\right]\right|}{\sqrt{\mathbb{E}\left[\|Z\|^2\right]}},$$

where the last inequality follows from the fact that $|\sqrt{a} - \sqrt{b}| = \frac{|a-b|}{\sqrt{a}+\sqrt{b}} \leq \frac{|a-b|}{\sqrt{b}}$. This shows that, with probability at least $1-\delta$, T_1 can be lower bounded as

$$T_1 \geq \beta B\|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P - \beta\sigma_{\max}BL\sqrt{\frac{3\log 1/\delta}{\tilde{n}}} - \epsilon_t. \tag{E.4}$$

Bounding T_2. Using the definition of functional gradients, T_2 can be rewritten as follows

$$T_2 = \left|\mathbb{E}_X\left[\langle g_t(X), \nabla_\phi R(W_{t-1}, \phi_{t-1})(X)\rangle\right] - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\langle g_t(\mathbf{x}_{t,i}), \nabla_\phi \widehat{R}_{S_t}(W_{t-1}, \phi_{t-1})(\mathbf{x}_{t,i})\rangle\right|$$
$$= \left|\mathbb{E}_{X,Y}\left[\langle g_t(X), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(X), Y)\rangle\right] - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\langle g_t(\mathbf{x}_{t,i}), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle\right|$$
$$\leq \sup_{g\in\mathcal{G}_t}\left|\mathbb{E}_{X,Y}\left[\langle g(X), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(X), Y)\rangle\right] - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\langle g(\mathbf{x}_{t,i}), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle\right|.$$

We now rely on uniform convergence bounds and bound the RHS in terms of the Rademacher complexity term $\mathcal{R}(\mathcal{G}_t)$. First note that the random variable $\langle g(X), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(X), Y)\rangle$ is bounded and lies in the interval $[-\sigma_{\max}BL, \sigma_{\max}BL]$. This follows from the Lipschitz property of the loss ℓ and the boundedness of the functions in $\mathcal{G}_t$. Using Theorem 62, we get the following upper bound for T_2, which holds with probability at least $1-\delta$

$$T_2 \leq 2\mathbb{E}\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_i\langle g(\mathbf{x}_{t,i}), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle\right] + 2\sigma_{\max}BL\sqrt{\frac{\log 2/\delta}{\tilde{n}}}.$$

We now focus on bounding the Rademacher complexity term in the above inequality. Define function $h_i : \mathbb{R}^D \to \mathbb{R}$ as follows

$$h_i(\mathbf{u}) = \langle \mathbf{u}, W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle.$$

Note that, $h_i(\mathbf{u})$ is $\sigma_{\max}L$-Lipschitz in $\mathbf{u}$. The Rademacher complexity can be written in terms of h_i's as follows

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_i\langle g(\mathbf{x}_{t,i}), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle\right] = \mathbb{E}_{S_t}\left[\mathbb{E}_\rho\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_i h_i(g(\mathbf{x}_{t,i}))\Big|S_t\right]\right].$$

Using the composition property of Rademacher complexities stated in Theorem 63, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_i\langle g(\mathbf{x}_{t,i}), W_{t-1}^T\nabla\ell(W_{t-1}\phi_{t-1}(\mathbf{x}_{t,i}), y_{t,i})\rangle\right] &\leq \sigma_{\max}L\mathbb{E}_{S_t}\left[\mathbb{E}_\rho\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{D}\rho_{ij}[g(\mathbf{x}_{t,i})]_j\middle|S_t\right]\right]\\
&= \sigma_{\max}L\mathbb{E}\left[\sup_{g\in\mathcal{G}_t}\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{D}\rho_{ij}[g(\mathbf{x}_{t,i})]_j\right]\\
&= \sigma_{\max}L\mathcal{R}(\mathcal{G}_t).
\end{aligned}$$

So we have the following bound for T_2 which holds with probability at least $1-\delta$

$$T_2 \leq 2\sigma_{\max}L\mathcal{R}(\mathcal{G}_t) + 2\sigma_{\max}BL\sqrt{\frac{\log 2/\delta}{\tilde{n}}}. \tag{E.5}$$

Combining Equations (E.4), (E.5) gives us the required bound. □

E.5.2 Main Argument

Our analysis of inexact gradient descent uses similar arguments as in Temlyakov [Tem14]. Let $\phi_t = \phi_{t-1} + \eta_t g_t$ be the t^{th} iterate generated by the algorithm. We first derive an upper bound for the reduction in population risk in the t^{th} iteration of the algorithm. From Lemma 59 we know that with probability at least $1-\delta/3T$

$$R(W_t,\phi_t) \leq R(W_{t-1},\phi_t) + C_1(t), \tag{E.6}$$

where $C_1(t) = 2\eta_t L\mathcal{R}(\mathcal{W},\mathcal{G}_t) + \frac{4c\sigma_{\max}BLt^{1-s}}{1-s}\left(\sqrt{\frac{\log 6T/\delta}{\tilde{n}}} + \sqrt{\frac{K}{\tilde{n}}}\right)$. Since ℓ is M smooth, the following holds for any two vectors $\mathbf{u},\mathbf{v}\in\mathbb{R}^K$ and $y\in\mathcal{Y}$

$$\ell(\mathbf{u}+\mathbf{v},y) \leq \ell(\mathbf{u},y) + \langle\mathbf{v},\nabla\ell(\mathbf{u},y)\rangle + \frac{M\|\mathbf{v}\|_2^2}{2}.$$

Using this smoothness property, $R(W_{t-1},\phi_t) = \mathbb{E}\left[\ell(W_{t-1}\phi_{t-1}(\mathbf{x}) + \eta_t W_{t-1}g_t, y)\right]$ can be upper bounded as

$$R(W_{t-1},\phi_t) \leq R(W_{t-1},\phi_{t-1}) + \eta_t\langle g_t, \nabla_\phi R(W_{t-1},\phi_{t-1})\rangle_P + \frac{\eta_t^2 M\sigma_{\max}^2\|g_t\|_P^2}{2}. \tag{E.7}$$

Combining Equations (E.6), (E.7), we get the following bound on $R(W_t,\phi_t)$ which holds with probability at least $1-\delta/3T$

$$R(W_t,\phi_t) \leq R(W_{t-1},\phi_{t-1}) + \eta_t\langle g_t, \nabla_\phi R(W_{t-1},\phi_{t-1})\rangle_P + \frac{\eta_t^2 M\sigma_{\max}^2 B^2}{2} + C_1(t).$$

Next, from Lemma 60 we know that the g_t chosen by the algorithm satisfies the following with probability at least $1-2\delta/3T$

$$\langle g_t, -\nabla_\phi R(W_{t-1},\phi_{t-1})\rangle_P \geq \beta B\|\nabla_\phi R(W_{t-1},\phi_{t-1})\|_P - \epsilon_t - C_2(t),$$

where $C_2(t) = 2\sigma_{\max} L\mathcal{R}(\mathcal{G}_t) + 4\sigma_{\max} BL\sqrt{\frac{\log 6T/\delta}{\tilde{n}}}$. Substituting this in the previous equation, we get the following bound on $R(W_t, \phi_t)$ which holds with probability at least $1-\delta/T$

$$\begin{aligned} R(W_t, \phi_t) \leq & R(W_{t-1}, \phi_{t-1}) - \eta_t \beta B \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P + \frac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s} && \text{(E.8)} \\ & + \eta_t \epsilon_t + C_1(t) + \eta_t C_2(t). && \text{(E.9)} \end{aligned}$$

Let $r_t = R(W_t, \phi_t) - R(W^*, \phi^*) - \sum_{i=1}^t (\eta_i \epsilon_i + C_1(t) + \eta_t C_2(t))$. Then the above equation implies the following recurrence on r_t

$$r_t \leq r_{t-1} + \frac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s}. \tag{E.10}$$

We now try to tighten this recurrence. Let $W_{t-1}^\dagger$ be the pseudoinverse of W_{t-1}. From the convexity of ℓ we have

$$\begin{aligned} R(W_{t-1}, \phi_{t-1}) - R(W^*, \phi^*) & \overset{(a)}{=} R(W_{t-1}, \phi_{t-1}) - R(W_{t-1}, W_{t-1}^\dagger W^* \phi^*) \\ & \overset{(b)}{\leq} -\langle W_{t-1}^\dagger W^* \phi^* - \phi_{t-1}, \nabla_\phi R(W_{t-1}, \phi_{t-1}) \rangle_P \\ & \leq \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P \left(\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + \|\phi_{t-1}\|_P \right), \end{aligned}$$

where (a) follows from the definition of psuedoinverse and (b) follows from the convexity of ℓ. Letting $A_t = \sum_{i=1}^t \eta_i B$, we can lower bound $\|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|$ as

$$\|\nabla_\phi R(W_{t-1}, \phi_{t-1})\| \geq \frac{R(W_{t-1}, \phi_{t-1}) - R(W^*, \phi^*)}{\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + A_{t-1}}.$$

Substituting this in Equation (E.8), we get

$$\begin{aligned} R(W_t, \phi_t) \leq & R(W_{t-1}, \phi_{t-1}) - \eta_t \beta B \left(\frac{R(W_{t-1}, \phi_{t-1}) - R(W^*, \phi^*)}{\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + A_{t-1}} \right) + \frac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s} \\ & + \eta_t \epsilon_t + C_1(t) + \eta_t C_2(t). \end{aligned}$$

Rewriting the above equation in terms of r_t, we get

$$\begin{aligned} r_t & \leq r_{t-1} - \eta_t \beta B \left(\tfrac{r_{t-1}}{\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + A_{t-1}} \right) + \tfrac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s} \\ & = \left(1 - \tfrac{\eta_t \beta B}{\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + A_{t-1}} \right) r_{t-1} + \tfrac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s}. \end{aligned} \tag{E.11}$$

In the rest of the proof, we try to solve the above recurrence relation on r_t to obtain the required excess risk bound. First note that there exists t_0 such that for all $t \geq t_0$ [1]

$$\frac{\eta_t \beta B}{\sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P + A_{t-1}} \geq \frac{\alpha + 3\beta(1-s)}{4t}. \tag{E.12}$$

[1] To be precise, t_0 is such that $t_0^{1-s} = \frac{\alpha \sigma_{\min}^{-1} \|W^*\|_2 \|\phi^*\|_P}{c\beta B}$.

This follows from the observation that $\eta_t = ct^{-s}$ and $A_{t-1} \leq \frac{cBt^{1-s}}{1-s}$. We now make use of Theorem 65 for solving the recurrence in Equation (E.11). We first show that r_t satisfies the conditions for Theorem 65 with $a = \alpha, b = (\alpha + \beta(1-s))/2$ and $D = t_0$ and for some A which we specify later. From Equation (E.10) we have

$$r_{t+1} \leq r_t + \frac{c^2 MB^2 \sigma_{\max}^2}{2} t^{-2s} \leq r_t + A(t-1)^{-\alpha},$$

where the last inequality holds for any $A \geq \frac{c^2 MB^2 \sigma_{\max}^2}{2}$ and for our choice of α, s specified in the theorem statement. This shows that the first condition of Theorem 65 is satisfied by r_t. Next, suppose $r_t \geq At^{-\alpha}$, for some $t \geq t_0$. Then using Equations (E.11) and (E.12), r_{t+1} can be bounded as follows

$$\begin{aligned} r_{t+1} &\leq \left(1 - \frac{\alpha + 3\beta(1-s)}{4t}\right) r_t + \frac{c^2 MB^2 \sigma_{\max}^2}{2} t^{-2s} \\ &= \left(1 - \frac{\alpha + \beta(1-s)}{2t}\right) r_t \underbrace{- \left(\frac{\beta(1-s) - \alpha}{4t}\right) r_t + \frac{c^2 MB^2 \sigma_{\max}^2}{2} t^{-2s}}_{T_1}. \end{aligned}$$

Following our choices for α, s and using the fact that $r_t \geq At^{-\alpha}$, it is easy to verify that $T_1 \leq 0$ for sufficiently large A. This shows that for appropriately chosen A, we have

$$r_{t+1} \leq \left(1 - \frac{\alpha + \beta(1-s)}{2t}\right) r_t.$$

Since the conditions for Theorem 65 are satisfied, using it to solve our recurrence gives us the following bound on r_t which holds with probability at least $1 - \delta$

$$r_T \leq O\left(\frac{1}{T^\alpha}\right).$$

This finishes the proof of the Theorem.

E.6 Proof of Corollary 5

A simple intuition for why the exact greedy approach satisfies similar risk bounds as gradient greedy approach is that in exact greedy approach one solves the greedy step in Equation (6.2) exactly. Whereas, in gradient greedy approach, the greedy step is only solved approximately and so one would expect the objective value of exact greedy approach to be smaller than gradient greedy approach. We formalize this intuition in the proof. Let (W_t, ϕ_t), where $\phi_t = \phi_{t-1} + \eta_t g_t$, be the t^{th} iterate generated by the exact greedy algorithm. And let $(\tilde{W}, \tilde{\phi}_t)$, where $\tilde{\phi}_t = \phi_{t-1} + \eta_t \tilde{g}_t$, be the iterate obtained by running gradient greedy update in the t^{th} iteration of the algorithm. We now bound $R(W_t, \phi_t)$ in terms of $R(\tilde{W}, \tilde{\phi}_t)$

$$\begin{aligned} R(W_t, \phi_t) &\leq \widehat{R}_{S_t}(W_t, \phi_t) + \sup_{W \in \mathcal{W}, g \in \mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)| \\ &\overset{(a)}{\leq} \widehat{R}_{S_t}(\tilde{W}_t, \tilde{\phi}_t) + \sup_{W \in \mathcal{W}, g \in \mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)| \\ &\leq R((\tilde{W}_t, \tilde{\phi}_t)) + 2 \sup_{W \in \mathcal{W}, g \in \mathcal{G}_t} |R(W, \phi_{t-1} + \eta_t g) - \widehat{R}_{S_t}(W, \phi_{t-1} + \eta_t g)|, \end{aligned}$$

where (a) follows from the definition of W_t, ϕ_t which are obtained by minimizing Equation (6.2). Note that the supremum in the RHS above can be bounded using Lemma 59.

From the proof of Theorem 21, we know that $R((\tilde{W}_t, \tilde{\phi}_t))$ can be upper bounded in terms of $R((W_{t-1}, \phi_{t-1}))$. To be precise, from Equation (E.8) in the proof of Theorem 21, we know that with probability at least $1 - 2\delta/T$

$$\begin{aligned} R(\tilde{W}_t, \tilde{\phi}_t) \leq & R(W_{t-1}, \phi_{t-1}) - \eta_t \beta B \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P + \frac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s} \\ & + \eta_t \epsilon_t + C_1(t) + \eta_t C_2(t). \end{aligned}$$

This shows that

$$\begin{aligned} R(W_t, \phi_t) \leq & R(W_{t-1}, \phi_{t-1}) - \eta_t \beta B \|\nabla_\phi R(W_{t-1}, \phi_{t-1})\|_P + \frac{c^2 M B^2 \sigma_{\max}^2}{2} t^{-2s} \\ & + \eta_t \epsilon_t + C_1(t) + \eta_t C_2(t). \end{aligned}$$

Using the exact same techniques as in the proof of Theorem 21, we get the required risk bound on $R(W_T, \phi_T)$.

E.7 Proof of Corollary 6

The major part of the proof involves bounding the Rademacher complexity terms appearing in the risk bound of Theorem 21. We first bound $\mathcal{R}(\mathcal{G})$.

$$\begin{aligned}
\mathcal{R}\left(\mathcal{G}\right) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{D}\rho_{ij}[g(\mathbf{x}_{t,i})]_j\right]\\
&= \mathbb{E}\left[\sup_{C:\max_j \|C_{j,*}\|_1\leq\Lambda} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{D}\rho_{ij}\sigma(\langle C_{j,*},\mathbf{x}_{t,i}\rangle)\right]\\
&\leq \sum_{j=1}^{D}\mathbb{E}\left[\sup_{\|C_{j,*}\|_1\leq\Lambda} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_{ij}\sigma(\langle C_{j,*},\mathbf{x}_{t,i}\rangle)\right]\\
&\overset{(a)}{\leq} \sum_{j=1}^{D}\mathbb{E}\left[\sup_{\|C_{j,*}\|_1\leq\Lambda} \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\rho_{ij}\left\langle C_{j,*},\mathbf{x}_{t,i}\right\rangle\right]\\
&\leq \sum_{j=1}^{D}\Lambda\mathbb{E}\left[\frac{1}{\tilde{n}}\left\|\sum_{i=1}^{\tilde{n}}\rho_{ij}\mathbf{x}_{t,i}\right\|_\infty\right]\\
&= \frac{D\Lambda}{\tilde{n}}\mathbb{E}\left[\left\|\sum_{i=1}^{\tilde{n}}\rho_{i1}\mathbf{x}_{t,i}\right\|_\infty\right]\\
&\overset{(b)}{\leq} 2D\Lambda\sqrt{\frac{\log d}{\tilde{n}}},
\end{aligned}$$

where (a) follows from the Lipschitzness of sigmoid activation function and composition property of Rademacher complexities(see Theorem 63) and (b) follows from the following well known property of sub-Gaussian random variables. Let $Z_1, \dots Z_n$ be n random variables, not necessarily independent. Moreover, lets suppose each Z_i is sub-Gaussian with parameter σ. Then $\mathbb{E}\left[\max_i Z_i\right] \leq \sqrt{2\sigma^2\log n}$. Since $\mathcal{X} \subseteq [0,1]^d$, it is easy to see that conditioned on data S_t, each co-ordinate of $\sum_{i=1}^{\tilde{n}}\rho_{i1}\mathbf{x}_{t,i}$ is a sub-Gaussian random variable with parameter $\sqrt{\tilde{n}}$. So using the above stated property of sub-Gaussian random variables, we get

$$\begin{aligned}
\mathbb{E}_\rho\left[\left\|\sum_{i=1}^{\tilde{n}}\rho_{i1}\mathbf{x}_{t,i}\right\|_\infty\right] &= \mathbb{E}_\rho\left[\max_{j\in[d]}\max\left\{\sum_{i=1}^{\tilde{n}}\rho_{i1}[\mathbf{x}_{t,i}]_j, -\sum_{i=1}^{\tilde{n}}\rho_{i1}[\mathbf{x}_{t,i}]_j\right\}\right]\\
&\leq \sqrt{2\tilde{n}\log 2d}.
\end{aligned}$$

Next, we bound $\mathcal{R}(\mathcal{W}, \mathcal{G})$

$$
\begin{aligned}
\mathcal{R}(\mathcal{W}, \mathcal{G}) &= \mathbb{E}\left[\sup_{\substack{W \in \mathcal{W}, \\ g \in \mathcal{G}}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \rho_{ik} [W g(\mathbf{x}_{t,i})]_k\right] \\
&\leq \sum_{k=1}^{K} \mathbb{E}\left[\sup_{\substack{W \in \mathcal{W}, \\ g \in \mathcal{G}}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_{ik} \left\langle W_{k,*}, g(\mathbf{x}_{t,i}) \right\rangle\right] \\
&\overset{(a)}{\leq} 2\sigma_{\max} K \sum_{j=1}^{D} \mathbb{E}\left[\sup_{\|C_{j,*}\|_1 \leq \Lambda} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_i \left\langle C_{j,*}, \mathbf{x}_{t,i} \right\rangle\right] + O\left(\frac{\sigma_{\max} K \sqrt{D} \log D}{\sqrt{\tilde{n}}}\right) \\
&\overset{(b)}{\leq} 4\sigma_{\max} K D \Lambda \sqrt{\frac{\log d}{\tilde{n}}} + O\left(\frac{\sigma_{\max} K \sqrt{D} \log D}{\sqrt{\tilde{n}}}\right) \\
&\leq O\left(\frac{\sigma_{\max} K D \Lambda \log{(dD)}}{\sqrt{\tilde{n}}}\right).
\end{aligned}
$$

where (a) follows from the property of Rademacher complexity stated in Theorem 64 and (b) uses the arguments used to bound $\mathcal{R}(\mathcal{G})$ above. Substituting the above bounds for $\mathcal{R}(\mathcal{G})$ and $\mathcal{R}(\mathcal{W}, \mathcal{G})$ in Theorem 21 and using the fact that $\sup_X \|g(X)\|_2 \leq \sqrt{D}$, for all $g \in \mathcal{G}$, we get the required risk bound.

E.8 Proof of Corollary 7

Similar to the proof of Corollary 6, we focus on bounding the Radmacher complexity terms $\mathcal{R}(\mathcal{G}_t)$ and $\mathcal{R}(\mathcal{W}, \mathcal{G}_t)$. To bound $\mathcal{R}(\mathcal{G}_t)$, we use the same argument we used to bound $\mathcal{R}(\mathcal{G})$

in Corollary 6.

$$
\begin{aligned}
\mathcal{R}\left(\mathcal{G}_t\right) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}_t} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{D} \rho_{ij} [g(\mathbf{x}_{t,i})]_j\right] \\
&= \mathbb{E}\left[\sup_{C:\max_j \|C_{j,*}\|_1 \leq \Lambda} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{D} \rho_{ij} \sigma(\langle C_{j,*}, \phi_{t-1}(\mathbf{x}_{t,i})\rangle)\right] \\
&\leq \sum_{j=1}^{D} \mathbb{E}\left[\sup_{\|C_{j,*}\|_1 \leq \Lambda} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_{ij} \sigma(\langle C_{j,*}, \phi_{t-1}(\mathbf{x}_{t,i})\rangle)\right] \\
&\leq \sum_{j=1}^{D} \mathbb{E}\left[\sup_{\|C_{j,*}\|_1 \leq \Lambda} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_{ij} \left\langle C_{j,*}, \phi_{t-1}(\mathbf{x}_{t,i})\right\rangle\right] \\
&\leq \sum_{j=1}^{D} \Lambda \mathbb{E}\left[\frac{1}{\tilde{n}} \left\| \sum_{i=1}^{\tilde{n}} \rho_{ij} \phi_{t-1}(\mathbf{x}_{t,i}) \right\|_\infty\right] \\
&= \frac{D\Lambda}{\tilde{n}} \mathbb{E}\left[\left\| \sum_{i=1}^{\tilde{n}} \rho_{i} \phi_{t-1}(\mathbf{x}_{t,i}) \right\|_\infty\right] \\
&\overset{(a)}{\leq} \frac{2cD\Lambda t^{1-s}}{1-s}\sqrt{\frac{\log d}{\tilde{n}}},
\end{aligned}
$$

where (a) uses similar arguments as in the proof of Corollary 6 and relies on the fact that $\|\phi_{t-1}(\mathbf{x})\|_\infty \leq \sum_{i=1}^{t-1} \eta_i \leq \frac{ct^{1-s}}{1-s}$. Next, we bound $\mathcal{R}(\mathcal{W}, \mathcal{G}_t)$

$$
\begin{aligned}
\mathcal{R}(\mathcal{W}, \mathcal{G}_t) &= \mathbb{E}\left[\sup_{\substack{W\in\mathcal{W},\\ g\in\mathcal{G}_t}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{K} \rho_{ik} [Wg(\mathbf{x}_{t,i})]_k\right] \\
&\leq \sum_{k=1}^{K} \mathbb{E}\left[\sup_{\substack{W\in\mathcal{W},\\ g\in\mathcal{G}_t}} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_{ik} \left\langle W_{k,*}, g(\mathbf{x}_{t,i})\right\rangle\right] \\
&\overset{(a)}{\leq} 2\sigma_{\max} K \sum_{j=1}^{D} \mathbb{E}\left[\sup_{\|C_{j,*}\|_1 \leq \Lambda} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \rho_i \left\langle C_{j,*}, \phi_{t-1}(\mathbf{x}_{t,i})\right\rangle\right] + O\left(\frac{\sigma_{\max} K \sqrt{D} \log D}{\sqrt{\tilde{n}}}\right) \\
&\overset{(b)}{\leq} \frac{4\sigma_{\max} c D K \Lambda t^{1-s}}{1-s} \sqrt{\frac{\log d}{\tilde{n}}} + O\left(\frac{\sigma_{\max} K \sqrt{D} \log D}{\sqrt{\tilde{n}}}\right) \\
&\leq O\left(t^{1-s} \frac{\sigma_{\max} K D \Lambda \log dD}{\sqrt{\tilde{n}}}\right),
\end{aligned}
$$

where (a) follows from the property of Rademacher complexity stated in Theorem 64 and (b) relies on arguments used to bound $\mathcal{R}(\mathcal{G}_t)$. Substituting the above bounds for $\mathcal{R}(\mathcal{G}_t)$ and $\mathcal{R}(\mathcal{W}, \mathcal{G}_t)$ in Theorem 21, we get the required risk bound.

E.9 Some Useful Results

Theorem 61 (Chernoff Bounds). *Let $X = \sum_{i=1}^n X_i$, where X_i's are independently distributed in $[0,1]$. Then, for $\epsilon \in (0,1)$*

$$\mathbb{P}\left(X > (1+\epsilon)\mathbb{E}\left[X\right]\right) \leq \exp\left(-\frac{\epsilon^2}{3}\mathbb{E}\left[X\right]\right), \quad \mathbb{P}\left(X < (1-\epsilon)\mathbb{E}\left[X\right]\right) \leq \exp\left(-\frac{\epsilon^2}{2}\mathbb{E}\left[X\right]\right).$$

Theorem 62 (Bartlett and Mendelson [BM02]). *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $[a,b]$ and let $\{X_i\}_{i=1}^n$ be independently selected according to the probability measure P. Then for any integer n and any $0 < \delta < 1$, with probability at least $1-\delta$ over samples of length n, every f in $\mathcal{F}$ satisfies*

$$\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\left[f(X)\right]\right| \leq 2\mathcal{R}(\mathcal{F}) + (b-a)\sqrt{\frac{\log 2/\delta}{n}},$$

where $\mathcal{R}(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$ which is defined as

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \rho_i f(X_i)\right],$$

where the expectation is taken w.r.t the Rademacher random variables ρ's and data $\{X_i\}_{i=1}^n$.

We next present an important result on the composition property of Rademacher complexities.

Theorem 63 (Maurer [Mau16]). *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}^d$ and let $\{h_i\}_{i=1}^n$ be L-Lipschitz functions from $\mathbb{R}^d$ to $\mathbb{R}$. Then*

$$\mathbb{E}_\rho\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \rho_i h_i(f(X_i))\right] \leq L\mathbb{E}_\rho\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d \rho_{ij}[f(X_i)]_j\right].$$

Theorem 64 (Proposition A.12 of Allen-Zhu, Li, and Liang [ALL19]). *Let $u : \mathbb{R} \to \mathbb{R}$ be a fixed 1-Lipschitz function. Given $\mathcal{F}_1 \ldots \mathcal{F}_m$ classes of functions $\mathcal{X} \to \mathbb{R}$ and suppose for each $j \in [m]$ there exists a function $f_j^{(0)} \in \mathcal{F}_j$ satisfying $\sup_{\mathbf{x}\in\mathcal{X}} |u(f_j^{(0)}(\mathbf{x}))| \leq A$, then*

$$\mathcal{F}' = \left\{\mathbf{x} \to \sum_{j=1}^m v_j u(f_j(\mathbf{x}))\middle| f_j \in \mathcal{F}_j \wedge \|v\|_1 \leq B \wedge \|v\|_\infty \leq D\right\}$$

satisfies

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}'}\sum_{i=1}^n\frac{1}{n}\rho_i\sum_{j=1}^m v_j u(f_j(\mathbf{x}_i))\right] \leq 2D\sum_{j=1}^m\mathbb{E}\left[\sup_{f\in\mathcal{F}_j}\sum_{i=1}^n\frac{1}{n}\rho_i f(\mathbf{x}_i)\right] + O\left(\frac{AB\log m}{\sqrt{n}}\right).$$

Theorem 65 (Temlyakov [Tem14]). *Let four positive numbers $a < b \leq 1, A, D \in \mathbb{N}$ be given and let a sequence $\{r_t\}_{t=1}^\infty$ have the following properties: $r_1 \leq A$ and for any $t \geq 2$*

$$r_t \leq r_{t-1} + A(t-1)^{-a}.$$

Moreover, suppose the sequence is such that if $r_t \geq At^{-a}$ for some $t \geq D$, then $r_{t+1} \leq r_t(1-b/t)$. Then there exists a constant C such that for all $t \in \mathbb{N}$ we have

$$r_t \leq Ct^{-a}.$$

E.10 Experiments

This section provides experimental details, including the datasets, hyperparameter settings, and additional experimental evidence not presented in the main paper.

We first note that in our experiments for DenseCompBoost, we use a slight variant of $\mathcal{G}_t$ defined in Equation (6.3)

$$\mathcal{G}_t = \left\{ h + g \circ \left(\sum_{i=0}^{t-1} \alpha_i \phi_i \right), \text{ for } h \in \mathcal{H}, g \in \mathcal{G}, \alpha_i \in \mathbb{R} \right\},$$

where $\mathcal{H}, \mathcal{G}$ are weak feature transformer classes. We use this variant because the dimensions of the input feature space and the representation space need not be the same, and as a consequence $\mathcal{G}_t$ in Equation (6.3) can not always be used. Similar to StdCompBoost, we consider two choices for $\mathcal{H}, \mathcal{G}$: one based on fully connected blocks and the other based on convolution blocks.

E.10.1 Drawbacks of Layer-by-Layer fitting

In this section, we provide empirical evidence highlighting drawbacks of layer-by-layer fitting and how our proposed techniques address these drawbacks. Similar to Section 6.4, we use StdCompBoost to denote standard layer-by-layer fitting.

DenseCompBoost can recover from mistakes. We mentioned earlier that compared to StdCompBoost, one advantage of DenseCompBoost is that the dense connections allow it to more easily recover from mistakes made in earlier layers. We now provide empirical evidence to support this claim. We introduce mistakes in the weights of the first layers learned using StdCompBoost and DenseCompBoost. To be precise, we fix the weights of the first layer of both StdCompBoost and DenseCompBoost to (a) the same random matrix, (b) an all-0 matrix, and then continue the training of the later layers. Table E.1 shows the results: while StdCompBoost suffers a significant performance drop (from 82.49% when every layer is greedily trained, to 72.99% with a random first layer), the performance of dense greedy is barely affected (from 95.70% when every layer is trained, to 95.0% with a random first layer). Similar trend occurs when setting the first layer to 0: dense greedy still achieves a 93.69% test accuracy, while standard greedy would fail to train at all since any signal in the data has been cut off.

Narrow-to-Wide architecture of CmplxCompBoost. Note that in CmplxCompBoost, we increase the widths of layers over iterations. We now justify this choice of architecture. There are two possible ways to vary the complexity of the $\tilde{\mathcal{G}}_t$, increasing or decreasing. We tested both approaches on one tabular dataset CovType, and one image dataset SVHN. On CovType, we started with a layer width of 4096, then increase or decrease the width of subsequent layers by 512 at each layer. On SVHN, the starting layer width is 128, followed by 4 additional layers, each increasing or decreasing the width by 16. As can be seen in table E.2, increasing complexity gives slightly better results for both

		layer 1	layer 2	layer 3	layer 4	layer 5
StdCompBoost	Random	49.71	50.25	52.51	69.70	72.99
DenseCompBoost	Random	49.71	50.86	70.07	92.31	95.00
	Zero	50.06	61.76	89.19	93.17	93.69

Table E.1: Test accuracy at each layer, with the first layer being set to a random value or the all-0 matrix. Compared to the performance without corrupted first layer, StdCompBoost suffers a performance drop, while DenseCompBoost is almost unaffected, demonstrating its ability to recover from mistakes made in early layers.

the datasets, therefore we choose to increase the width for CmplxCompBoost in all other experiments.

	Decreasing width	Increasing width
CovType	95.58 ± 0.04	**95.64** ± 0.16
SVHN	88.30 ± 0.28	**89.05** ± 0.01

Table E.2: Test accuracy using CmplxCompBoost with decreasing or increasing layer widths.

E.10.2 Datasets and Hyperparameters

In this section, we present the details of datasets used in our experiments and describe our process for hyperparameter selection.

Simulated Datasets. We generated 3 synthetic binary classification datasets in $\mathbb{R}^{32}$. Simulation 1 is a concentric ellipsoids dataset, where a point $\mathbf{x}$ is classified based on $\mathbf{x}^T A\mathbf{x}$, for some randomly generated positive semi-definite matrix A. Simulations 2 and 3 are datasets whose classification boundaries are polynomials of degrees 8 and 9 respectively. For each of these datasets, we generated 10^6 samples for training and testing.

Hyper-parameters. We used hold-out set validation to pick the best hyper-parameters for all the methods. We used 20% of the training data as validation data and picked the best parameters using grid search, based on validation accuracy. After picking the best parameters, we train on the entire training data and report performance on the test data. For all the greedy techniques based on neural networks, we used fully connected blocks and tuned the following parameters: weight decay, width of weak feature transformers, number of iterations T. For CmplxCompBoost, we set $\Delta = D_0/5$. For end-to-end training, we tuned weight decay, width of layers, depth. We used SGD for optimization of all these techniques. The number of epochs and step size schedule of SGD are chosen to ensure convergence. For XGBoost, we tuned the number of trees, depth of each tree, learning rate.

Benchmark Datasets. We consider the following image datasets: CIFAR10, MNIST, FashionMNIST [XRV17], MNIST-rot-back-image [Lar+07], convex [XRV17], SVHN [Net+11], and the following tabular datasets from UCI repository [BM98]: letter recognition [FS91], forest cover type (covtype), connect4. The convex dataset involves classifying shapes in images as either convex or non-convex. MNIST-rot-back-image is generated from MNIST by rotating the images and adding random images in the background.

Hyper-parameters. For covtype dataset, which doesn't come with a test set, we randomly sample 20% of the original data and use it as the test set. We use a similar hyper-parameter selection technique as above and tune the same set of hyper-parameters as described above. We use convolution blocks for CIFAR10, SVHN, FashionMNIST, convex, MNIST-rot-back-image and fully connected blocks for the rest. We limit the width of fully connected blocks to 4096, and the number of output channels in convolution blocks to 128 while tuning the hyper-parameters for the composition boosting techniques and end-to-end training. For AdaBoost and additive representation boosting, we set these limits to 16000 and 350 respectively. For CmplxCompBoost with convolution blocks, we set $\Delta = D_0/8$. We *do not* use data augmentation in our experiments.

E.10.3 Further Experimental Details

Tables E.3, E.4 list the statistics of datasets used in our experiments. We now list the hyper-parameters tuned for each dataset and learning algorithm. Table E.5 presents the list of hyper-parameters tuned for XGBoost. All the other techniques we use in our experiments rely on neural networks. We use SGD with momentum to learn these models. In all our experiments, we set the initial learning rate of SGD to 0.01, momentum to 0.9, batch size to 64 and tune the following weight decay values: $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$. The number of epochs we used for SGD varied with the dataset and is chosen to be large enough to ensure convergence. Over the course of the SGD optimization, we reduce the learning rate by a factor of 0.5, if the training loss doesn't decrease for certain number of SGD iterations (we rely on scheduler-tolerance option in PyTorch to implement this). We run all the greedy techniques (*AdaBoost, additive feature boosting, StdCompBoost, DenseCompBoost, CmplxCompBoost*) for 10 iterations and use validation dataset to decide the best early stopping rule. For End-2-End training, we tune two values of depth: $5, 10$. Tables E.6, E.7 presents the list of all the other hyper-parameters tuned.

Table E.3: Details of simulated datasets used in our experiments. We use 20% of the training data as validation set for picking the best hyper-parameter

Dataset	Simulation 1	Simulation 2	Simulation 3
# Train samples	1000000	1000000	1000000
# Test samples	500000	500000	500000
# Classes	2	2	2

Table E.4: Details of benchmark datasets used in our experiments. We use 20% of the training data as validation set for picking the best hyper-parameter

Details	**Image Datasets**				
	SVHN	FashionMNIST	CIFAR10	Convex	MNIST-rot-back-image
# Train samples	73257	60000	50000	8000	12000
# Test samples	26032	10000	10000	50000	50000
# Classes	10	10	10	2	10

Details	**Tabular Datasets**			
	MNIST	Letter	CovType	Connect4
# Train samples	60000	15000	464809	54045
# Test samples	10000	5000	116203	13512
# Classes	10	26	7	3

Table E.5: List of hyper-parameters tuned for XGBoost, on all the datasets used in our experiments.

Parameter	Values Tuned
Tree Depth	$\{10, 15, 20\}$
Learning Rate	$\{0.1, 0.2\}$
Number of Trees	$\{400, 800, 1600\}$

Table E.6: List of hyper-parameters tuned for various compositional boosting techniques and end-2-end training.

Dataset	Hyper-parameters tuned
Simulation-1	width:$\{32, 64, 128\}$
Simulation-2	width:$\{64, 128, 256\}$
Simulation-3	width:$\{256, 512, 1024\}$
SVHN	output channels:$\{32, 64, 128\}$
FashionMNIST	output channels:$\{32, 64, 128\}$
Convex	output channels:$\{32, 64, 128\}$
MNIST-rot-back-image	output channels:$\{32, 64, 128\}$
CIFAR10	output channels:$\{32, 64, 128\}$
MNIST	width:$\{256, 512, 1024\}$
LETTER	width:$\{256, 512, 1024\}$
Covtype	width:$\{1024, 2048, 4096\}$
Connect4	width:$\{256, 512, 1024\}$

Table E.7: List of hyper-parameters tuned for AdaBoost and additive feature boosting. To be fair for additive boosting techniques, we considered wider weak learners than the ones used for compositional boosting and end-2-end training.

Dataset	Hyper-parameters tuned
Simulation-1	width:$\{256, 512, 1024\}$
Simulation-2	width:$\{256, 512, 1024\}$
Simulation-3	width:$\{4096, 8192, 16384\}$
SVHN	output channels:$\{128, 256, 350, 512\}$
FashionMNIST	output channels:$\{128, 256, 350, 512\}$
Convex	output channels:$\{128, 256, 350, 512\}$
MNIST-rot-back-image	output channels:$\{128, 256, 350, 512\}$
CIFAR10	output channels:$\{128, 256, 350, 512\}$
MNIST	width:$\{256, 512, 1024\}$
LETTER	width:$\{256, 512, 1024\}$
Covtype	width:$\{4096, 8192, 16384\}$
Connect4	width:$\{256, 512, 1024\}$