A New View of Predictive State Methods for Dynamical System Learning

Ahmed Hefny

Abstract

Recently there has been substantial interest in predictive state methods for learning dynamical systems: these algorithms are popular since they often offer a good tradeoff between computational speed and statistical efficiency. Despite their desirable properties, though, predictive state methods can sometimes be difficult to use in practice. E.g., in contrast to the rich literature on supervised learning methods, which allows us to choose from an extensive menu of models and algorithms to suit the prior beliefs we have about properties of the function to be learned, predictive state dynamical system learning methods are comparatively inflexible: it is as if we were restricted to use only linear regression instead of being allowed to choose decision trees, nonparametric regression, or the lasso. To address this problem, we propose a new view of predictive state methods in terms of instrumental-variable regression. This view allows us to construct a wide variety of dynamical system learners simply by swapping in different supervised learning methods. We demonstrate that we can represent spectral learning algorithms for Hidden Markov Models and Kalman filters within this framework and that we can tweak the regression method or the feature representation to achieve a favorable outcome.

1. Introduction

Recently, there has been substantial interest in a new class of algorithms for learning dynamical systems. These algorithms combine several key intuitions, of which two are important to the current discussion.

First is the idea of **predictive state**: we can replace a belief about a latent variable S by the prediction of some observable variables X that depend on S. That is, on observing some evidence E about S, we could calculate the belief $\mathbb{P}(S \mid E)$. But, as long as the function $f(S) = \mathbb{E}(X \mid S)$ is sufficiently rich (e.g., for discrete X and S represented by one-hot encodings, invertible), it is equivalent to calculate $\mathbb{E}(X \mid E)$ directly: we could recover $\mathbb{P}(S \mid E)$ by inverting f, but in many cases we don't need to do so.

Second is the **method of moments**: for a dynamical system model with parameters θ, it is often intractable to solve the maximum likelihood problem $\max_\theta \ln P(\text{observations} \mid \theta)$. But, we can sometimes find a statistic T such that the expectation of T is a simple, invertible function of θ: $\mathbb{E}_\theta(T) = g(\theta)$. In this case, we can replace the expectation $\mathbb{E}_\theta(T)$ with the empirical average $\hat{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$. (Here T_i is the value of our statistic for the ith of a set of N observations.) We can then define an estimator $\hat{\theta}$ as

$$\hat{\theta} = g^{-1}(\hat{T}) \tag{1}$$

and seek an efficient algorithm for solving the inverse problem (1). The trick to designing a good method of moments algorithm is to discover a statistic T such that problem (1) is well-conditioned and efficiently solvable. One of the main tools that algorithm designers use for this purpose is to expand the class of models considered, thereby removing difficult constraints from (1): for example, instead of learning a hidden Markov model (HMM), we can expand the model class to include all observable operator models (OOMs) [14].

For brevity, we will call methods that use the above intuitions **predictive state methods**. Predictive state methods are popular for dynamical system learning because they often offer a good tradeoff between computational speed and statistical efficiency.

However, there are also some important difficulties with these methods. One is that it can be hard for predictive state methods to take advantage of prior knowledge about the structure or parameters of a dynamical system: for example, expanding from HMMs to OOMs removes our ability to directly refer to the conditional probability distribution of observations given states, a common place to incorporate structure. Another is that deriving, analyzing, and implementing new predictive state methods can require substantial expertise: it can be difficult to discover an appropriate statistic T, accumulate its empirical average $\hat{T}$ efficiently, and track how estimation errors in $\hat{T}$ propagate through the inverse problem (1) to affect $\hat{\theta}$.

We address both of these problems with a new view of predictive state methods for dynamical system learning. In this view, a dynamical system learning problem is reduced to a sequence of supervised learning problems. So, we can directly apply the rich literature on supervised learning methods to incorporate many types of prior knowledge about problem structure. We give a general convergence rate analysis that allows a high degree of flexibility in designing estimators. And finally, implementing a new estimator becomes as simple as rearranging our data and calling the appropriate supervised learning subroutines.

Our new view is based on **instrumental-variable regression** [18, 22]. Instrumental-variable regression is a well-known technique to compensate for certain types of observation noise in a linear regression problem; it can let us recover regression coefficients accurately where ordinary regression would yield biased estimates. The connection between predictive state learning and linear instrumental variable regression has been noted before, e.g., in [4]. We propose a generalization of the linear two stage ordinary least squares procedure [22], give error bounds for this generalization, and formulate dynamical systems learning as an instance of this regression technique.

More specifically, our contribution is to show that we can use much-more-general supervised learning algorithms in place of linear regression, and still get a meaningful theoretical analysis. In more detail: (1) we point out that we can equally well use any well-behaved supervised learning algorithm in place of linear regression in the first stage of instrumental-variable regression; (2) for the second stage of instrumental-variable regression, we generalize ordinary linear regression to its RKHS counterpart; (3) we analyze the resulting combination, and show that we get convergence to the correct answer, with a rate that depends on how quickly the individual supervised learners converge.

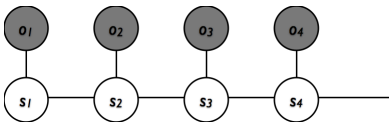

Figure 1: A dynamical system.

In the remainder of the paper, we first describe how to use instrumental-variable regression to learn a dynamical system (Sec. 2). We then provide theoretical guarantees for the two-stage instrumental-variable regression technique with non-linearity (Sec. 4). Finally, we give two examples of learning dynamical systems within our proposed framework; the first example is an HMM model for knowledge tracing (Sec. 5) and the second example is a Kalman filter model for neural activity in the motor cortex (Sec. 6).

2. Instrumental Regression for Dynamical Systems

We consider a dynamical system of the form in Fig. 1: a sequence of observations $o_t \in \mathcal{O}$ explained by latent states $s_t \in \mathcal{S}$ connected in a chain. A key question we need to solve in order be able to perform inference in the dynamical system is how to recursively update our belief about state: given a belief about s_t and a new observation o_{t+1}, compute a belief about s_{t+1}. This is referred to as *filtering*. Another inference task is *prediction*: predicting an observation o_{t+k} given our belief about the current state s_t. This involves computing a belief about future state state s_{t+1} given our belief about s_t.

If s_t and o_t have small, discrete ranges, the predictive state algorithm for learning a dynamical system is well known: see, e.g., [6, 3]. In fact, it is also known that we can interpret this algorithm as linear instrumental-variable regression [4]. Our proposed framework generalizes that direction by reducing dynamical system learning to solving three supervised learning problems, with the additional ability to incorporate arbitrary non-linear regression models in two of these problems.

The first step to formulate dynamical system learning as a supervised learning problem is to use an observable (predictive state) representation by replacing our belief about s_t with a predictive state: pick a statistic $\psi_t = \psi(o_{t:t+k-1})$ of a window of *future* observations $o_{t+1:t+k}$, and instead of tracking our belief $\mathbb{P}(s_t \mid o_{1:t-1})$, track the predictive state $\mathbb{E}[\psi_t \mid o_{1:t-1}]$. (The dimension of ψ must be at least as high as the number of discrete latent states.) We will use $Q_{t|t-k}$ to denote $\mathbb{E}[\psi_t \mid o_{1:t-k}]$ and hence our predictive predictive state is denoted by $Q_t \equiv Q_{t|t-1}$. We assume that the system is k-observable and hence latent states are distinguishable by the distribution of a window of k future observations[1] [25].

Second, we formulate a statistic $\xi_t = \xi(o_{t:t+k})$ over *extended future observations* $o_{t:t+k}$ with conditional expectation $P_t = \mathbb{E}[\xi_t \mid o_{1:t-1}]$ such that $Q_{t+1|t-1}$ and Q_{t+1} (i.e. our belief about the *shifted future* $o_{t+1:t+k}$) can be inferred from P_t and o_t.

1. In principle, the statistics can depend on the entire future. The restriction to a window of observations simplifies the notation and is commonly used in practice.

Learning a dynamical system then amounts to learning an operator W that satisfies the moment condition:[2]

$$P_t = WQ_t \quad \forall t \tag{2}$$

Filtering and prediction then correspond to inferring Q_{t+1} and $Q_{t+1|t-1}$ respectively from P_t. We assume that W is a linear operator. Unfortunately, we do not observe Q_t or P_t but noisy versions thereof. Moreover, due to the overlap between observation windows, the noise terms on ψ_t and ξ_t are correlated. This noise correlation means that naïve linear regression (using samples of ψ_t and ξ_t) will give a biased estimate of the dependence between Q_t and P_t.

To counteract this bias, we employ instrumental regression [18, 22]. Instrumental regression uses *instrumental variables* that are correlated with the input Q_t but not with the noise $\epsilon_{t:t+k}$. This property provides a criterion to denoise the inputs and outputs of the original regression problem: we remove that part of the input/output that is not correlated with the instrumental variables. Since past observations $o_{1:t-1}$ do not overlap with future or extended future windows, they are not correlated with the noise $\epsilon_{t:t+k+1}$. Therefore, we can use *history features* $h_t \equiv h(o_{1:t-1})$ as instrumental variables.

In more detail, by taking the expectation of (2) over h_t we obtain an instrument-based moment condition

$$\begin{aligned} \mathbb{E}[P_t \mid h_t] &= \mathbb{E}[WQ_t \mid h_t] \\ \mathbb{E}[\mathbb{E}[\xi_t \mid o_{1:t-1}] \mid h_t] &= W\mathbb{E}[\mathbb{E}[\psi_t \mid o_{1:t-1}] \mid h_t] \\ \mathbb{E}[\xi_t \mid h_t] &= W\mathbb{E}[\psi_t \mid h_t] \end{aligned} \tag{3}$$

Assuming that there are enough independent dimensions in h_t that are correlated with Q_t, we maintain the rank of the moment condition when moving from (2) to (3), and we can recover W by least squares if we can compute $\mathbb{E}[\psi_t \mid h_t]$ and $\mathbb{E}[\xi_t \mid h_t]$ for sufficiently many examples t.

In summary, learning and inference of a dynamical system through instrumental regression can be described as follows:

- **Model Specifcation:** Pick features of history $h_t = h(o_{1:t-1})$, future $\psi_t = \psi(o_{t:t+k-1})$ and extended future $\xi_t = \xi(o_{t:t+k})$. ψ_t must be a sufficient statistic for $\mathbb{P}(o_{t:t+k-1} \mid o_{1:t-1})$. ξ_t must satisfy
  - $\mathbb{E}[\psi_{t+1} \mid o_{1:t-1}] = f_{\text{predict}}(\mathbb{E}[\xi_t \mid o_{1:t-1}])$ for a known function f_{predict}.
  - $\mathbb{E}[\psi_{t+1} \mid o_{1:t}] = f_{\text{filter}}(\mathbb{E}[\xi_t \mid o_{1:t-1}], o_t)$ for a known function f_{filter}.
- **S1A (Stage 1A) Regression:** Learn a (possibly non-linear) regression model to estimate $\bar{\psi}_t \equiv \mathbb{E}[\psi_t \mid h_t]$. The training data for this model are (h_t, ψ_t) across time steps t.[3]

2. Note that, similar to [16], P_t is a deterministic function of Q_t and hence this condition has a unique solution if we observe sufficient examples of P_t and Q_t.
3. Our bounds assume that the training time steps t are sufficiently spaced for the underlying process to mix, but in practice, the error will only get smaller if we consider all time steps t.

- **S1B Regression:** Learn a (possibly non-linear) regression model to estimate $\bar{\xi}_t \equiv \mathbb{E}[\xi_t \mid h_t]$. The training data for this model are (h_t, ξ_t) across time steps t.
- **S2 Regression:** Use the feature expectations estimated in the previous two steps to train a model to predict $\bar{\xi}_t = W\bar{\psi}_t$, where W is a linear operator. The training data for this model are estimates of $(\bar{\psi}_t, \bar{\xi}_t)$ across time steps t obtained from S1 steps.
- **Initial State Estimation:** Estimate an initial state $Q_1 = \mathbb{E}[\psi_1]$ by averaging ψ_1 across several example realizations of our time series.[4]
- **Inference:** Starting from the initial state Q_1, we can maintain the belief state $Q_t \equiv \mathbb{E}[\psi_t \mid o_{1:t-1}]$ through filtering: given Q_t we compute $P_t \equiv \mathbb{E}[\xi_t \mid o_{1:t-1}] = WQ_t$. Then, given the observation o_t, we can compute $Q_{t+1} = f_{\text{filter}}(P_t, o_t)$. Or, in the absence of o_t, we can predict the next state $Q_{t+1|t-1} = f_{\text{predict}}(P_t)$. Finally, by definition, the belief state Q_t is sufficient to predict $\mathbb{P}(o_{t:t+k-1} : o_{1:t-1})$.

The process of learning and inference is depicted in Figure 2. Modeling assumptions are reflected in the choice of the statistics ψ, ξ and h as well as the regression models in stages S1A and S1B. In the supplementary material we show that, with linear S1 models and certain choices of statistics, we can recover existing spectral algorithms for dynamical systems learning. The two stage framework not only provides a unifying view of some of the successful dynamical systems learning algorithms but also paves the way for extending them in a theoretically justified manner, as we demonstrate in the experiments.

3. Related Work

This work extends predictive state learning algorithms for dynamical systems, which include spectral algorithms for Kalman filters [2], Hidden Markov Models [10, 19] and Predictive State Representations (PSRs) [6, 3] as well as infinite-dimensional variants such as the Hilbert space embedding of hidden Markov models (HSE-HMM) [20] and predictive state representations (HSE-PSR) [5].

One common aspect in all these models is that they exploit the covariance structure between future and past observation sequences to obtain an unbiased observable state representation. Indeed, many of these algorithms can be reformulated as a two-stage linear instrumental regression. Boots and Gordon [4] note the connection between the HSE-HMM and instrumental variables, which is manifested in the use of kernel SVD of a future-past covariance operator to identify the latent state space. We use this connection to build a general framework for dynamical system learning where the state-space can be identified using supervised learning methods, including non-linear ones.

Reducing dynamical systems learning to supervised learning dates back to auto-regressive models [17], where the state of the system is assumed to be fully determined by the previous

4. This is the only step that needs multiple realizations of our time series. If only a single long realization is available, we need additional assumptions to be able to estimate an initial state; for example, if we assume stationarity, we can set the initial state to be the empirical average vector of future features, $\frac{1}{T}\sum_{t=1}^{T}\psi_t$.

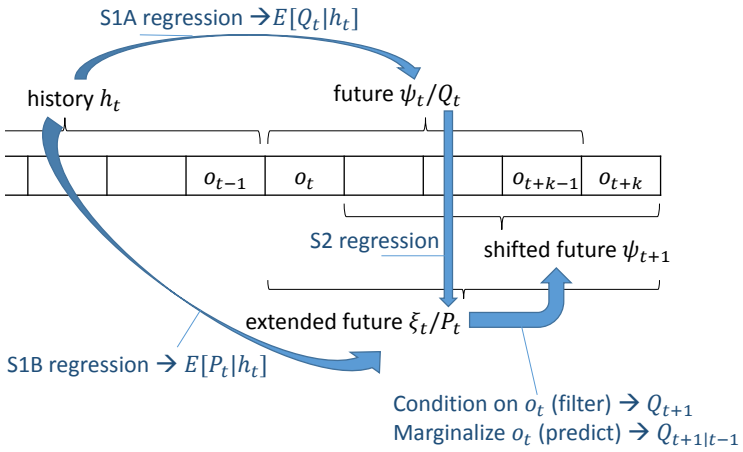

Figure 2: Learning and applying a dynamical system using instrumental regression. S1 regression is trained to provide data to train S2 regression. At test time, starting from an initial belief state Q_0, we alternate between S2 regression and filtering/prediction

k observations. Our aim is to use supervised learning methods to learn latent state models from observation sequences. This bears similarity to Langford et al.'s sufficient posterior representation (SPR) [16], which encodes the state by the sufficient statistics of the conditional distribution of the next observation and represents system dynamics by three vector-valued functions that are estimated using supervised learning approaches. While SPR allows all of these functions to be non-linear, there are some advantages that distinguish our work. First, while SPR is limited to 1-step observable systems (where the distribution over the next observation uniquely determines the state), our framework can seamlessly handle k-step observable systems by choosing a large enough (or even unbounded) window size. The use of instrumental variables ensures that correlated noise on overlapping windows does not bias our estimates of the system parameters. Secondly, SPR involves a rather complicated training procedure, involving multiple iterations of model refinement and model averaging, whereas our framework only requires solving three regression problems in sequence. Finally, the theoretical analysis of [16] only establishes the consistency of SPR learning assuming that all regression steps are solved perfectly. Our work, on the other hand, establishes convergence rates based on the performance of S1 regression.

4. Theoretical Analysis

In this section we present our main theoretical result: consistency and a convergence rate bound for two-stage instrumental regression, under the assumption that S1 predictions con-

verge to the true conditional expectations at an appropriate rate, regardless of the functional form of the S1 regressors.

We assume we are given i.i.d. triplets (x_t, y_t, z_t), where $x_t \in \mathcal{X}$, $y_t \in \mathcal{Y}$ and $z_t \in \mathcal{Z}$ denote input, output and instrumental variables respectively. (As mentioned above, we can equally well use correlated samples, as would result from successive time steps of a time series; our convergence rates will then include a factor that depends on the mixing rate of the underlying dynamical system.)

For generality, we assume that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are reproducing kernel Hilbert spaces (RKHS) of possibly infinite dimension, and that the operator W is estimated through (kernel) ridge regression

$$\hat{W}_\lambda = \left(\sum_{t=1}^{T} \hat{y}_t \otimes \hat{x}_t \right) \left(\sum_{t=1}^{T} \hat{x}_t \otimes \hat{x}_t + \lambda I_{\mathcal{X}} \right)^{-1} \tag{4}$$

where $\otimes$ denotes tensor product and $\lambda > 0$ is a regularization parameter that ensures the invertibility of the estimated covariance. λ can be 0 in finite dimensional cases where we have an invertible covariance matrix. The RKHS view is useful when the future statistics are represented in terms of kernels—for example, if they are kernel mean maps of the distribution of future observations, a case that is closely related to the HSE-HMM [20] and HSE-PSR [5] models.

Let $\bar{x}_t$ and $\bar{y}_t$ denote $\mathbb{E}[x_t|z_t]$ and $\mathbb{E}[y_t|z_t]$. Also let $\hat{x}_t$ and $\hat{y}_t$ denote $\hat{E}[x_t|z_t]$ and $\hat{E}[y_t|z_t]$, as estimated by the S1A and S1B regression steps. We assume that $\bar{x}_t, \hat{x}_t \in \mathcal{X}$ and $\bar{y}_t, \hat{y}_t \in \mathcal{Y}$. Let $\Sigma_{\bar{x}\bar{x}} \in \mathcal{X} \otimes \mathcal{X}$ and $\Sigma_{\bar{y}\bar{y}} \in \mathcal{Y} \otimes \mathcal{Y}$ denote the (uncentered) covariance operators of the distributions of $\bar{x}$ and $\bar{y}$ respectively: that is,

$$\Sigma_{\bar{x}\bar{x}} = \mathbb{E}[\bar{x} \otimes \bar{x}] \qquad \Sigma_{\bar{y}\bar{y}} = \mathbb{E}[\bar{y} \otimes \bar{y}]$$

Before we state our main theorem we need to quantify the quality of S1 regressions in a way that is independent of the functional form that we assume in S1.

Definition 1 (S1 Regression Bound) *For a given $\delta > 0$ and $N \in \mathbb{N}^+$, we define the S1 regression bound $\eta_{\delta,N} > 0$ to be a number satisfying the condition that, with probability at least $(1 - \delta/2)$, the following holds for all $1 \leq t \leq N$:*

$$\|\hat{x}_t - \bar{x}_t\|_{\mathcal{X}} < \eta_{\delta,N}$$
$$\|\hat{y}_t - \bar{y}_t\|_{\mathcal{Y}} < \eta_{\delta,N}$$

As long as, for each fixed δ,

$$\lim_{N \to \infty} \eta_{\delta,N} = 0, \tag{5}$$

our results show that the two stage estimator is consistent:

Theorem 2 *Assume that* $\|\bar{x}\|_{\mathcal{X}}, \|\bar{x}\|_{\mathcal{Y}} < c < \infty$ *almost surely. Also, assume that* $\mathrm{tr}(\Sigma_{\bar{x}\bar{x}}), \mathrm{tr}(\Sigma_{\bar{y}\bar{y}}) < \infty$. *Let* $\eta_{\delta,N}$ *be as defined in Definition 1 and assume it satisfies* (5). *Assume* W *is a Hilbert-Schmidt operator, let* $\hat{W}_\lambda$ *be as defined in* (4)*, and let* $\overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$ *denote the closure of the range of* $\Sigma_{\bar{x}\bar{x}}$. *Then the following statement holds with probability at least* $1-\delta$ *for each* $x_{\mathrm{test}} \in \overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$ *s.t.* $\|x_{\mathrm{test}}\|_{\mathcal{X}} \leq 1$.

$$\begin{aligned}\gamma_{\delta,N} \equiv \|\hat{W}_\lambda x_{\mathrm{test}} - W x_{\mathrm{test}}\|_{\mathcal{Y}} = \\ O\left(\eta_{\delta,N}\left(\frac{1}{\lambda} + \frac{\sqrt{1+\sqrt{\frac{\log(1/\delta)}{N}}}}{\lambda^{\frac{3}{2}}}\right)\right) \\ + O\left(\frac{\log(1/\delta)}{\sqrt{N}}\left(\frac{1}{\lambda} + \frac{1}{\lambda^{\frac{3}{2}}}\right)\right) \\ + O\left(\sqrt{\lambda}\right)\end{aligned}$$

Theorem 2 gives a generic error bound on S2 regression in terms of S1 regression performance. We defer the proof, as well as finite sample analysis, to the supplementary material. The main insight from the theorem is that the error in estimating the parameter W is the sum of three contributions: the first term captures the error in the S1 regressions. The second term captures the effect of estimating the covariance operators from finite data, assuming S1 regression is exact. Finally, the third term captures the effect of regularization assuming covariance estimates are exact. It can be shown that, if $\mathcal{X}$ and $\mathcal{Y}$ are finite dimensional, $\Sigma_{\bar{x}\bar{x}}$ spans $\mathcal{X}$ and we use least squares to estimate W (i.e. $\lambda = 0$), then the first and last terms will vanish, and λ in the middle two terms will be replaced by $\lambda_{x,\min}$, the minimum eigenvalue of $\Sigma_{\bar{x}\bar{x}}$.

For completeness, the following propositions provide concrete examples of S1 regression bounds $\eta_{\delta,N}$ for practical regression models.

Proposition 3 *Assume* $\mathcal{X} \equiv \mathbb{R}^{d_x}, \mathbb{R}^{d_y}, \mathbb{R}^{d_z}$ *for some* $d_x, d_y, d_z < \infty$ *and that* $\bar{x}$ *and* $\bar{y}$ *are linear vector functions of* z *where the parameters are estimated using ordinary least squares. Assume that* $\|\bar{x}\|_{\mathcal{X}}, \|\bar{y}\|_{\mathcal{Y}} < c < \infty$ *almost surely. Let* $\eta_{\delta,N}$ *be as defined in Definition 1. Then*

$$\eta_{\delta,N} = O\left(\sqrt{\frac{d_z}{N}}\log((d_x+d_y)/\delta)\right)$$

Proof (sketch) This is based on results that bound parameter estimation error in linear regression with univariate response (e.g. [11]). Note that if $\bar{x}_{ti} = U_i^\top z_t$ for some $U_i \in \mathcal{Z}$, then a bound on the error norm $\|\hat{U}_i - U_i\|$ implies a uniform bound of the same rate on

$\hat{x}_i - \bar{x}$. The probability of exceeding the bound is scaled by $1/(d_x + d_y)$ to correct for multiple regressions. ■

Variants of Proposition 3 can also be developed using bounds on non-linear regression models (e.g., generalized linear models).

The next proposition addresses a scenario where $\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional.

Proposition 4 *Assume that x and y are kernel evaluation functionals, $\bar{x}$ and $\bar{y}$ are linear vector functions of z where the linear operator is estimated using conditional mean embedding [21] with regularization parameter $\lambda_0 > 0$ and that $\|\bar{x}\|_{\mathcal{X}}, \|\bar{y}\|_{\mathcal{Y}} < c < \infty$ almost surely. Let $\eta_{\delta,N}$ be as defined in Definition 1. It follows that*

$$\eta_{\delta,N} = O\left(\sqrt{\lambda_0} + \sqrt{\frac{\log(N/\delta)}{\lambda_0 N}}\right)$$

Proof (sketch) This bound is based on [21], which gives a bound on the error in estimating the conditional mean embedding. The error probability is adjusted by $\delta/4N$ to accommodate the requirement that the bound holds for all training data. ■

In the following, we apply theorem 2 to the setting of learning dynamical systems, where $Q_t \in \mathcal{X}$, $P_t \in \mathcal{Y}$ and $h_t \in \mathcal{Z}$ (Q_t, P_t and h_t are as defined in Section 2). One issue to note is that theorem 2 assumes that the test input lies within $\overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$. In dynamical systems context, however, the test input is an *estimated* predictive state $\hat{Q}_t$. Since S1 regression can fail to identify the subspace of *true* states given finite data, $\hat{Q}_t$ can have a non-zero component ϵ_t in $\mathcal{R}^{\top}(\Sigma_{\bar{x}\bar{x}})$, the orthogonal complement of $\overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$. The following lemma states that, in a stable system, this component gets smaller as S1 regression performs better.

Lemma 5 *For a test sequence $o_{1:T}$, let $\hat{Q}_t$ denote the estimated state given $o_{1:t-1}$. Let $\tilde{Q}_t$ denote the projection of $\hat{Q}_t$ onto $\overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$. Assume that f_{filter} is L-Lipchitz continuous on P_t and that $f_{\text{filter}}(P_t, o_t) \in \overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$ for any $P_t \in \overline{\mathcal{R}(\Sigma_{\bar{y}\bar{y}})}$. Given the assumptions in theorem 2 and assuming that $\|\hat{Q}_t\|_{\mathcal{X}} \leq R$ for all $1 \leq t \leq T$, the following holds for all $1 \leq t \leq T$ with probability at least $1 - \delta/2$.*

$$\|\epsilon_t\|_{\mathcal{X}} = \|\hat{Q}_t - \tilde{Q}_t\|_{\mathcal{X}} = O\left(\frac{\eta_{\delta,N}}{\sqrt{\lambda}}\right)$$

Since $\hat{W}_\lambda$ is bounded. The prediction error due to adding ϵ_t to the input diminishes at the same rate of $\|\epsilon_t\|_{\mathcal{X}}$.

5. Case Study I: Learning A Knowledge Tracing Model

In this section we demonstrate that we can learn a hidden Markov model using the two stage regression framework. We also demonstrate that we can change the regression methods to gain advantage. Specifically, we consider a limited data scenario, where we have a conflict between using many history features (picking a long history window to reduce noise in our predictions, and rich features of that window to achieve a linear relationship between history and future) or using few history features (reducing the number of parameters we have to learn from limited data). We show that we can use non-linear S1 regression models to reduce the number of parameters we need to learn, resulting in better empirical prediction accuracy compared to linear models while still maintaining consistency.

In this experiment we attempt to model and predict the performance of students learning from an interactive computer-based tutor. We use the Bayesian knowledge tracing (BKT) model [8], which is essentially a 2-state HMM: the state s_t represents whether a student has learned a knowledge component (KC), and the observation o_t represents the success/failure of solving the t^{th} question in a sequence of question that cover the said KC. With high probability, the student remains in the same state (learned or unlearned) and with smaller probability, the student may transition from unlearned to learned (learning) or learned to unlearned (forgetting). In the learned state, the student is more likely to answer a question correctly than in the unlearned state. It is also possible for the student to answer a question correctly while in the unlearned state (guessing) or incorrectly while in the learned state (slipping). The possible transitions and observations are summarized in figure 3.

5.1 Data Description

The data set we used to evaluate the model is a publicly available data set from DataShop [15] called "Geometry Area (1996-97)." This data was generated by students learning introductory geometry, and contains attempts by 59 students in 12 knowledge components. As is typical for BKT, we consider a student's attempt at a question to be correct iff the student entered the correct answer on the first try, without requesting any hints from the help system. The sequence of first attempts for a student/KC pair constitutes a training sequence. We discard sequences of length less than 5, resulting in a total of 325 sequences. We pad each observation sequence at the beginning with dummy observations, to handle the case where the history window extends before the beginning of the sequence. (This procedure allows us to use more data in our regressions, which is important because of our limited sample size.) Therefore a history observation which is used as training input for S1 regression can be in one of three states: "correct", "incorrect" or "before beginning of time." We restrict the regression output however to be binary ("correct" or "incorrect").

5.2 Model Description

Under the (reasonable) assumption that the two states have distinct observation probabilities, this model is 1-observable. It is reasonable then to choose the predictive state to be the

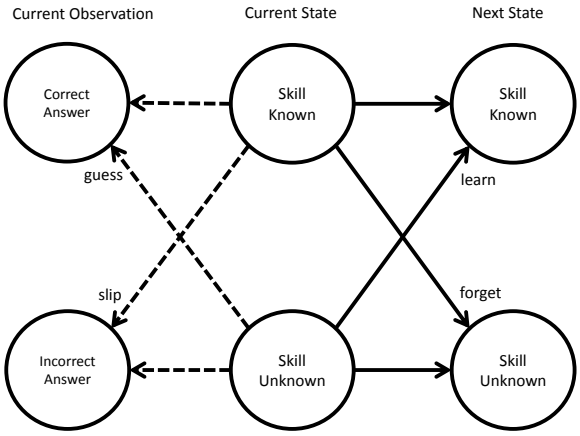

Figure 3: Transitions and observation emissions of the BKT model. (Each node represents a possible *value* of the state/observation). Solid arrows represent transitions while dashed arrows represent emissions. Horizontal arrows represent "ordinary" transitions and emissions, where skill level is maintained and correctly represented by the answer. Diagonal arrows represent emissions where the student guessed or slipped, or transitions where the student learned or forgot the skill.

expected next observation, which results in the following statistics:

$$\psi_t = o_t$$
$$\xi_t = o_t \otimes_k o_{t+1},$$

where o_t is represented by a 2 dimensional indicator vector and $\otimes_k$ denotes the Kronecker product. Given these statistics, $P_t = \mathbb{E}[\xi_t | o_{1:t-1}]$ is a joint probability table of $o_{t:t+1}$ from which conditioning on o_t (filtering) and marginalizing over o_t (prediction) are simple operations. It thus remains to choose the history features h_t and the S1 regression model. In the appendix, we show that if use $h_t = o_{t-1}$ and linear regression as S1 regression model, the resulting algorithm is equivalent to spectral HMM method of [10] and thus we use it as a baseline. In fact, if we had access to sufficient data, we could learn the HMM using this base line model. Not counting dummy observations, the model has to learn 7 parameters (7 free covariance entries). Under limited data, however, we can achieve faster learning by incorporating prior knowledge. Here, we will take advantage of the intuition that switching states (learning or forgetting) is a relatively unlikely event. Hence, aggregating observations over multiple previous time steps is a better predictor of the state, since aggregation will mitigate the effects of guessing and slipping. So we would like to use $h_t = o_{t-b:t-1}$ for some $b > 1$. We then have a choice: if we represent h_t by an indicator vector of dimension 2^b, then the optimal predictor of o_t from h_t will be linear, but the number of parameters we must learn will increase exponentially with b. On the other hand, if we represent h_t by a

Model	S1 Regression	History Features
model 1	Linear	o_{t-1}
model 2	Linear	$o_{t-4:t-1}$ (Indicator)
model 3	Logistic	$o_{t-4:t-1}$ (Separate)

Table 1: Evaluated models: "indicator" means we have one feature for each distinct sequence of length b, while "separate" means that we have separate discrete features for each observation in the history window.

binary vector of length b, then we will only need to learn $b + 1$ parameters, but the optimal predictor of o_t from h_t will no longer be linear leading to poor performance of linear regression. It is not obvious a priori which choice will result in better learning performance.[5]

Our formulation makes the choice much easier: we can use a history window of any length, pick the more-concise length-b representation, and train a nonlinear predictor such as a logistic regression. By doing so we combine the advantages of both of the previous paragraph's approaches: we only need to learn $O(b)$ parameters, but our class of predictors still includes a near-optimal choice. (Logistic regression becomes exactly optimal as the probabilities of learning and forgetting approach zero. Since these probabilities are typically small in practice, logistic regression will be close to optimal in practice.) As we will see below, the result is better learning from limited data.

5.3 Evaluation Procedure and Results

We evaluated three variants of HMM learning via two-stage regression. They are summarized in Table 1. We evaluated the models using 1000 random splits of the 325 sequences into 200 training and 125 testing. For each split, we trained each model on the training sequences. Then for each test sequence, we filter through the first 3 observations then predict the rest of the sequence, reporting the root mean square error for each split. The results are depicted in figure 4. The results show that, in terms of accuracy, model 3 outperforms model 2, which in turn outperforms model 1. In other words, feature expansion does increase predictive accuracy. However, even more gain is achieved using non-linear S1 models that require fewer parameters.

6. Case Study II: Neural Spike Data

In this section we demonstrate that we can use two stage regression to learn a Kalman filter and that state of that Kalman filter has good predictive power. Specifically, we learn a Kalman filter on neural trajectories of a reach task, where a monkey is expected to acquire by hand a target in one of 16 directions. We show that the state of the Kalman filter condi-

5. The numbers above ignore the effect of padding observation sequences, but the conclusions are similar in either case.

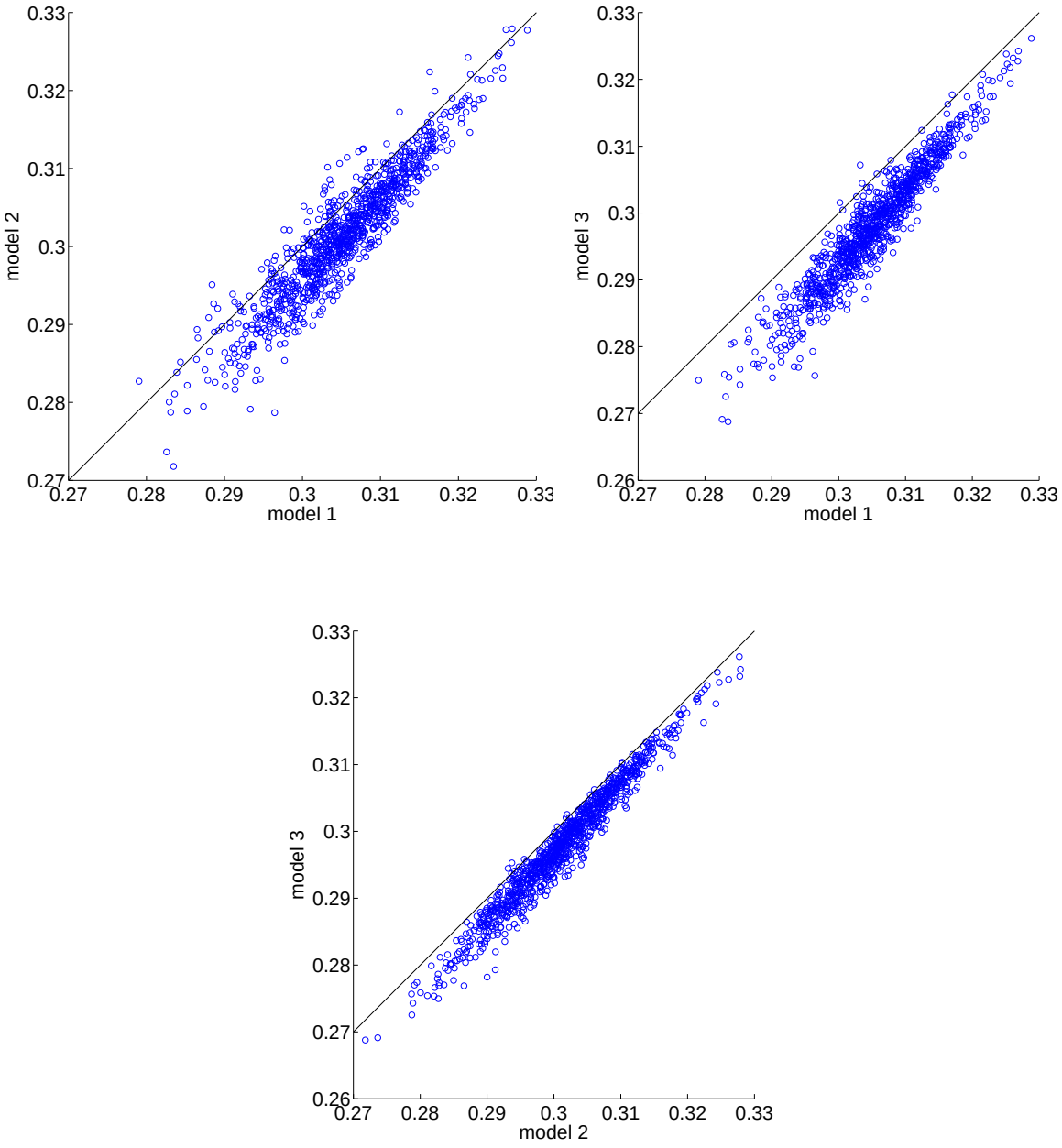

Figure 4: Experimental results: each graph depicts the performance of two models (measured by RMSE) on 1000 train/test splits. The black line represents the $x = y$ lines. More points below the line indicates that model y is better than model x.

tioned on the neural activity in the planning phase (the time interval before movement) is a good predictor of the direction.

6.1 Data Description

We give a summary the data collection procedure but more details can be found in [23]. All procedures were performed in accordance with the guidelines of the Institutional Care and Use Committee of the University of Pittsburgh.

One male Rhesus monkey (Macaca mulatta) was trained to perform a hand-controlled two-dimensional center out task in a virtual reality setup. An infrared marker was used to continuously track the hand position (Optotrak 3020 motion tracking system), which was then presented as visual feedback in the form of a cursor on the screen. Sixteen targets were radially located at the edges of an imaginary circle in the virtual setup. In each trial the monkey began by holding the cursor at a central-start position, then one of sixteen targets was randomly presented, the monkey reached to acquire the target and was required to hold for about 200ms at the end. If the monkey fails to acquire the right target the trial is discarded. Figure 5 displays hand trajectories for the recorded trials.

The monkey was implanted with a 96-channel array (Blackrock Microsystems, Salt Lake City, UT). The implant was visually placed in the proximal arm area of primary motor cortex (M1). A 96-channel Plexon MAP system (Plexon, Dallas, TX) was used to amplify, filter, and record the data. Spike sorting was used to isolate single unit activity as described in [23], which resulted in 93 identified units (neurons).

The dataset contains 47 trials for each direction. For each trial, we obtained spike counts for each of the 93 units binned in 20ms intervals. We aligned *movement onset* as the 20ms bin in which the movement speed of the hand reached 15% of the maximum speed in the trial. We trimmed all trials so that they contain exactly 6 bins (120 ms) before movement onset and 13 bins (260 ms) after movement onset. The 120ms and 260ms are the shortest pre-onset and post-onset intervals across all trials.

6.2 Model Description

A Kalman filter is given by

$$
\begin{aligned}
s_t &= O s_{t-1} + \nu_t \\
o_t &= T s_t + \epsilon_t \\
\nu_t &\sim \mathcal{N}(0, \Sigma_s) \\
\epsilon_t &\sim \mathcal{N}(0, \Sigma_o)
\end{aligned}
$$

In our case, o_t is the square root of the spike count in time bin t for each unit (i.e., a 93 dimensional vector). The square root transform is known to stabilize the variance of Poisson-distributed counts [7]. We assume a *stationary* filter where $\Sigma_t \equiv \mathbb{E}[s_t s_t^\top]$ is independent of t. We also assume a k-observable system where $k = 3$. We start by identifying a low dimensional subspace that contains the predictive state.

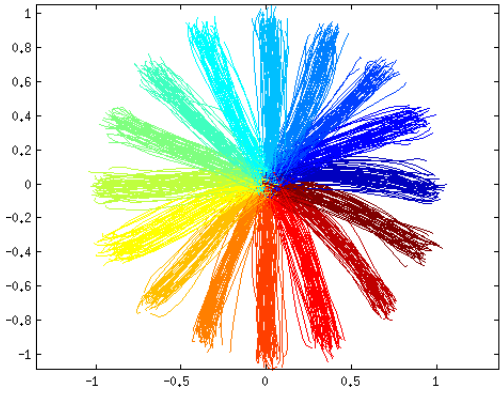

Figure 5: Hand motion traces of the recorded trials experiment in x, y coordinates. Each color indicates one of the 16 targets. All trials start from the center.

We choose our statistics

$$h_t = o_{t-k:t-1}$$
$$\psi_t = o_{t:t+k-1},$$

where a sequence of observations $o_{t_1:t_2}$ is represented by stacking their corresponding vectors into a single long vector. We then use reduced rank regression [13] as our S1A regression model. Reduced rank regression, for a regression problem from $\mathbb{R}^n$ to $\mathbb{R}^m$, finds regression weight matrix $W_{m\times n} = U_{m\times r}V_{n\times r}^\top$, where U has orthogonal columns and $r < m, n$. In our case $m = n = 93 \times 3$ and we set r to 20.

The matrix U then acts as a basis for the predictive state. Therefore we use the same basis for the shifted future

$$\xi_t = \begin{pmatrix} o_t \\ U^\top o_{t+1:t+k} \end{pmatrix}$$

In other words, ξ_t is the result of stacking the next observation and shifted future expressed in the U basis. Given ξ_t, we use linear regression as the S1B model. In the appendix, we show that the resulting algorithm is indeed a spectral learning algorithm for Kalman filters.

It remains to specify filtering and prediction functions. Prediction is trivially done by reading off from ξ_t the coordinates corresponding to o_t and $U^\top\psi_{t+1}$ as desired. Recognizing that ξ_t is normally distributed, filtering can be done under a steady-state approximation, where the covariance of ξ_t is assumed to be constant and hence can be estimated from the training data.

6.3 Evaluation Procedure and Results

We would like to evaluate whether the state of the Kalman filter can be used to predict movement direction given the neural activity before movement onset (i.e., during the first 120ms in the trial).

We use 5-fold cross validation. In each fold, we train the Kalman filter on the whole duration of the training trials and then perform filtering on the same trials for 6 time steps (i.e., up to movement onset). The predictive state at $t = 6$ is treated as a feature vector in a classification problem where the target class is the direction. We construct a nearest neighbor (NN) classifier from the training trials using these feature vectors.

Afterwards, for each test trial, we filter up to $t = 6$ given the Kalman filter parameters learned in the training phase and use the resulting predictive state as an input to the NN classifier. We measure the prediction error as the average angular difference between predicated and actual directions (in degrees). A random classifier would achieve an error of 90 degrees. We compare the Kalman filter based classifier to two baselines.

1. **Overall activity:** This is an NN classifier where the feature vector is the average spike count up to movement onset.

2. **Cosine tuning:** This is based on the direction profile concept [1]. This baseline is based on the assumption that, for each neuronal activity unit i, the spiking rate s_i is a function of the movement angle θ, where the maximum of this function is the *preferred direction* of that unit. We used the cosine tuning model $s_i = f(\theta) = c_i + w_i \cos(\theta - \theta_i)$, where the parameters c_i, w_i and θ_i are estimated for each of the 93 units from the training trials by minimizing mean square error. In more detail, for each unit we solve a regression problem where an input example consists of a direction as the input and the spike count for that unit in all training trials for that direction ,averaged over time and trials. The value of θ_i is the preferred direction for unit i. At test time, the predicted direction is given by

$$\sum_{t=1}^{6} \sum_{i=1}^{93} s_{it} u_i \ ,$$

   where s_{it} is the spike count of the i^{th} unit in the t^{th} time bin and u_i is the unit vector corresponding to θ_i. We try two variations of cosine tuning; the first version estimates the direction profile based on the entire time in training trials while the second version estimates the profile based only on the time before movement onset.

The results are shown in Table 2.

Model	Prediction error (degrees)
Kalman state	22.9375
Overall activity	34.5469
Cosine tuning (all)	44.5745
Cosine tuning (preonset)	24.3182

Table 2: Experimental results for predicting movement direction based on pre-onset activity

7. Conclusion

In this work we developed a general framework for dynamical system learning using supervised learning methods. The proposed framework is based on two-stage regression: in the first stage we use history features to train regression models that denoise future observation windows into state estimates. In the second stage we use these state estimates to train a linear model that represents system dynamics.

This framework encompasses and provides a unified view of some successful dynamical system learning algorithms. We demonstrated the proposed framework in learning a Hidden Markov Model and a Kalman filter. We have shown in the HMM case that we can use non-linear regression to incorporate more history features in identifying the latent state without an exponential increase in the number of parameters.

As future work, we would like to apply this framework to more scenarios where we can leverage additional techniques such as manifold embedding, sparse learning and transfer learning in stage 1 regression. We would also like to extend the framework to controlled processes.

References

[1] Bagrat Amirikian and Apostolos P Georgopulos. Directional tuning profiles of motor cortical cells. *Neuroscience research*, 36(1):73–79, 2000.

[2] Byron Boots. *Spectral Approaches to Learning Predictive Representations*. PhD thesis, Carnegie Mellon University, December 2012.

[3] Byron Boots and Geoffrey Gordon. An online spectral learning algorithm for partially observable nonlinear dynamical systems. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI-2011)*, 2011.

[4] Byron Boots and Geoffrey Gordon. Two-manifold problems with applications to nonlinear system identification. In *Proc. 29th Intl. Conf. on Machine Learning (ICML)*, 2012.

[5] Byron Boots, Arthur Gretton, and Geoffrey J. Gordon. Hilbert Space Embeddings of Predictive State Representations. In *Proc. 29th Intl. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[6] Byron Boots, Sajid Siddiqi, and Geoffrey Gordon. Closing the learning planning loop with predictive state representations. volume 30, pages 954–956, 2011.

[7] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

[8] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modelling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.*, 4(4):253–278, 1995.

[9] Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

[10] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. In *COLT*, 2009.

[11] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 9.1–9.24, 2012.

[12] Daniel Hsu, Sham M Kakade, and Tong Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14):1–13, 2012.

[13] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[14] Herbert Jaeger. Observable Operator Models for Discrete Stochastic Time Series. *Neural Computation*, 12(6):1371–1398, June 2000.

[15] Kenneth R. Koedinger, R. S. J. Baker, K. Cunningham, A. Skogsholm, B. Leber, and John Stamper. A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, pages 43–55, 2010.

[16] John Langford, Ruslan Salakhutdinov, and Tong Zhang. Learning nonlinear dynamic models. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 593–600, 2009.

[17] S.M. Pandit and S.M. Wu. *Time series and system analysis, with applications*. Wiley, 1983.

[18] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

[19] Sajid Siddiqi, Byron Boots, and Geoffrey J. Gordon. Reduced-rank hidden Markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-2010)*, 2010.

[20] L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert space embeddings of hidden Markov models. In *Proc. 27th Intl. Conf. on Machine Learning (ICML)*, 2010.

[21] Le Song, Jonathan Huang, Alexander J. Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 961–968, 2009.

[22] J.H. Stock and M.W. Watson. *Introduction to Econometrics*. Addison-Wesley series in economics. Addison-Wesley, 2011.

[23] Dawn M. Taylor, Stephen I. Helms Tillery, and Andrew B. Schwartz. Direct cortical control of 3d neuroprosthetic devices. *Science*, pages 1829–1832, 2002.

[24] Joel A. Tropp. User-friendly tools for random matrices: An introduction. NIPS Tutorial, 2012.

[25] P. van Overschee and L.R. de Moor. *Subspace identification for linear systems: theory, implementation, applications*. Number Volume 1. Kluwer Academic Publishers, 1996.

Appendix A. Spectral and HSE Dynamical System Learning as Regression

In this section we provide examples of mapping some of the successful dynamical system learning algorithms to our framework.

A.1 HMM

In this section we show that we can use instrumental regression framework to reproduce the spectral learning algorithm for learning HMM [10]. We consider 1-observable models but the argument applies to k-observable models. In this case we use $\psi_t = e_{o_t}$ and $\xi_t = e_{o_{t:t+1}} = e_{o_t} \otimes_k e_{o_{t+1}}$, where $\otimes_k$ denotes the kronecker product. We start with the (very

restrictive) case where $P_{1,2}$ is invertible. Given samples of $h_1 = o_1$, $\psi_2 = o_2$ and $\xi_2 = o_{2:3}$, in S1 regression we learn two matrices:

$$\hat{W}_{2,1} = \hat{\Sigma}_{o_2 o_1} \hat{\Sigma}_{o_1}^{-1} = \hat{P}_{2,1} \hat{P}_{1,1}^{-1} \tag{A.1}$$

$$\hat{W}_{2:3,1} = \hat{\Sigma}_{o_{2:3} o_1} \hat{\Sigma}_{o_1}^{-1} = \hat{P}_{2:3,1} \hat{P}_{1,1}^{-1} \tag{A.2}$$

In S2 regression we learn the matrix

$$\begin{aligned}
\hat{W} &= \hat{W}_{2:3,1} \hat{\mathbb{E}}[e_{o_1} e_{o_1}^\top] \hat{W}_{2,1}^\top \left(\hat{W}_{2,1} \hat{\mathbb{E}}[e_{o_1} e_{o_1}^\top] \hat{W}_{2,1}^\top \right)^{-1} \\
&= \left(\hat{P}_{2:3,1} P_{1,1}^{-1} \hat{P}_{2,1}^\top \right) \left(\hat{P}_{2,1} P_{1,1}^{-1} \hat{P}_{2,1}^\top \right)^{-1} \\
&= \hat{P}_{2:3,1} \left(\hat{P}_{2,1} \right)^{-1}
\end{aligned} \tag{A.3}$$

For a given value x of o_2, define

$$B_x = u_x^\top \hat{W} = u_x^\top \hat{P}_{2:3,1} \left(\hat{P}_{2,1}^\top \right)^{-1}, \tag{A.4}$$

where u_x is an $|\mathcal{O}| \times |\mathcal{O}|^2$ matrix which selects a block of rows in $\hat{P}_{2:3,1}$ corresponding to $o_2 = x$. Specifically, $u_x = \delta_x \otimes_k I_{|O|}$. This gives

$$\begin{aligned}
Q_{t+1} = \hat{\mathbb{E}}[e_{o_{t+1}} | o_{1:t}] &\propto u_{o_t}^\top \hat{\mathbb{E}}[e_{o_{t:t+1}} | o_{1:t-1}] \\
&= u_{o_t}^\top \hat{\mathbb{E}}[\xi_t | o_{1:t-1}] = B_{o_t} Q_t
\end{aligned}$$

with a normalization constant given by

$$\frac{1}{1^\top B_{o_t} Q_t} \tag{A.5}$$

In a realistic setting, we have $\mathrm{rank}(P_{2,1}) = m < |\mathcal{O}|$. Therefore we project the predictive state using a matrix U that preserves the dynamics, by requiring that $U^\top O$ (i.e. U is an independent set of columns spanning the range of the observation matrix O).

It can be shown [10] that $\mathcal{R}(O) = \mathcal{R}(P_{2,1}) = \mathcal{R}(P_{2,1} P_{1,1}^{-1})$. Therefore, we can use the leading m left singular vectors of $\hat{W}_{2,1}$, which corresponds to replacing the linear regression in S1A with a reduced rank regression. However, for the sake of our discussion will use the singular vectors of $P_{2,1}$. In more detail, let $[U, S, V]$ be the rank-m SVD decomposition of $P_{2,1}$. We use $\psi_t = U^\top e_{o_t}$ and $\xi_t = e_{o_t} \otimes_k U^\top o_{t+1}$. S1 weights are then given by $\hat{W}_{2,1}^{rr} = U^\top \hat{W}_{2,1}$ and $\hat{W}_{2:3,1}^{rr} = U^\top \hat{W}_{2:3,1}$ and S2 weights are given by

$$\begin{aligned}
\hat{W}^{rr} &= U^\top \hat{W}_{2:3,1} \hat{\mathbb{E}}[e_{o_1} e_{o_1}^\top] \hat{W}_{2,1}^\top U \left(U^\top \hat{W}_{2,1} \hat{\mathbb{E}}[e_{o_1} e_{o_1}^\top] \hat{W}_{2,1}^\top U \right)^{-1} \\
&= U^\top \hat{P}_{2:3,1} \hat{P}_{1,1}^{-1} V S \left(S V^\top \hat{P}_{1,1}^{-1} V S \right)^{-1} \\
&= U^\top \hat{P}_{2:3,1} \hat{P}_{1,1}^{-1} V \left(V^\top \hat{P}_{1,1}^{-1} V \right)^{-1} S^{-1}
\end{aligned} \tag{A.6}$$

In the limit of infinite data, V spans $\text{range}(O) = \text{rowspace}(P_{2:3,1})$ and hence $P_{2:3,1} = P_{2:3,1}VV^\top$. Substituting in (A.6) gives

$$W^{rr} = U^\top P_{2:3,1} V S^{-1} = U^\top P_{2:3,1} \left(U^\top P_{2,1}\right)^+$$

Similar to the full-rank case we define, for each observation x an $m \times |\mathcal{O}|^2$ selector matrix $u_x = \delta_x \otimes_k I_m$ and an observation operator

$$B_x = u_x^\top \hat{W}^{rr} \to U^\top P_{3,x,1} \left(U^\top P_{2,1}\right)^+ \tag{A.7}$$

This is exactly the observation operator obtained in [10]. However, instead of using A.6, they use A.7 with $P_{3,x,1}$ and $P_{2,1}$ replaced by their empirical estimates.

Note that for a state $b_t = \mathbb{E}[\xi_t | o_{1:t-1}]$, $B_x b_t = P(o_t | o_{1:t-1}) \mathbb{E}[\xi_{t+1} | o_{1:t}] = P(o_t | o_{1:t-1}) b_{t+1}$. To get b_{t+1}, the normalization constant becomes $\frac{1}{P(o_t|o_{1:t-1})} = \frac{1}{b_\infty^\top B_x b_t}$, where $b_\infty^\top b = 1$ for any valid predictive state b. To estimate b_∞ we solve the aforementioned condition for states estimated from all possible values of history features h_t. This gives,

$$b_\infty^\top \hat{W}_{2,1}^{rr} I_{|\mathcal{O}|} = b_\infty^\top U^\top \hat{P}_{2,1} \hat{P}_{1,1}^{-1} I_{|\mathcal{O}|} = 1_{|\mathcal{O}|}^\top,$$

where the columns of $I_{|\mathcal{O}|}$ represent all possible values of h_t. This in turn gives

$$\begin{aligned} b_\infty^\top &= 1_{|\mathcal{O}|}^\top \hat{P}_{1,1} (U^\top \hat{P}_{2,1})^+ \\ &= \hat{P}_1^\top (U^\top \hat{P}_{2,1})^+, \end{aligned}$$

the same estimator proposed in [10].

A.2 Stationary Kalman Filter

A Kalman filter is given by

$$\begin{aligned} s_t &= O s_{t-1} + \nu_t \\ o_t &= T s_t + \epsilon_t \\ \nu_t &\sim \mathcal{N}(0, \Sigma_s) \\ \epsilon_t &\sim \mathcal{N}(0, \Sigma_o) \end{aligned}$$

We consider the case of a *stationary* filter where $\Sigma_t \equiv \mathbb{E}[s_t s_t^\top]$ is independent of t. We choose our statistics

$$\begin{aligned} h_t &= o_{t-H:t-1} \\ \psi_t &= o_{t:t+F-1} \\ \xi_t &= o_{t:t+F} \end{aligned}$$

It can be shown [2, 25] that

$$\mathbb{E}[x_t|h_t] = \Sigma_{s,h}\Sigma_{h,h}^{-1}h_t$$

and it follows that

$$\mathbb{E}[\psi_t|h_t] = \Gamma\Sigma_{s,h}\Sigma_{h,h}^{-1}h_t = W_1 h_t$$
$$\mathbb{E}[\xi_t|h_t] = \Gamma_+\Sigma_{s,h}\Sigma_{h,h}^{-1}h_t = W_2 h_t$$

where Γ is the extended observation operator

$$\Gamma \equiv \begin{pmatrix} O \\ OT \\ \vdots \\ OT^F \end{pmatrix}, \Gamma_+ \equiv \begin{pmatrix} O \\ OT \\ \vdots \\ OT^{F+1} \end{pmatrix}$$

It follows that F and H must be large enough to have $\text{rank}(W) = n$. Let $U \in \mathbb{R}^{mF\times n}$ be the matrix of left singular values of W_1 corresponding to non-zero singular values. Then $U^\top\Gamma$ is invertible and we can write

$$\begin{aligned}\mathbb{E}[\psi_t|h_t] &= UU^\top\Gamma\Sigma_{s,h}\Sigma_{h,h}^{-1}h_t = W_1 h_t \\ \mathbb{E}[\xi_t|h_t] &= \Gamma_+\Sigma_{s,h}\Sigma_{h,h}^{-1}h_t = W_2 h_t \\ \mathbb{E}[\xi_t|h_t] &= \Gamma_+(U^\top\Gamma)^{-1}U^\top\left(UU^\top\Gamma\Sigma_{s,h}\Sigma_{h,h}^{-1}h_t\right) \\ &= W\mathbb{E}[\psi_t|h_t]\end{aligned}$$

which matches the instrumental regression framework. For the steady-state case (constant Kalman gain), one can estimate Σ_ξ from the data. $\mathbb{E}[\xi_{t+1}|h_t]$ and Σ_ξ then specify a joint Gaussian distribution where marginalization and conditioning can be easily performed.

A.3 HSE-PSR

We define a class of non-parametric two-stage instrumental regression models. By using conditional mean embedding [21] as S1 regression model, we recover a single-action variant of HSE-PSR [5]. Assume $\psi_t \in \mathcal{X}$ with a reproducing kernel $k_\mathcal{X}$ and that ξ_t is defined as the tuple $(o_t \otimes o_t, \psi_{t+1} \otimes o_t)$. Let $\mathbf{\Psi} \in \mathcal{X}\times\mathbb{R}^N$, $\mathbf{\Xi} \in \mathcal{Y}\times\mathbb{R}^N$ and $\mathbf{H} \in \mathcal{Z}\times\mathbb{R}^N$ be operators that represent training data. Specifically, ψ_s, ξ_s, h_t are the s^{th} "columns" in $\mathbf{\Psi}$ and $\mathbf{\Xi}$ and $\mathbf{H}$ respectively. It is possible to implement S1 using a non-parametric regression method that takes the form of a linear smoother. In such case the training data for S2 regression

take the form

$$\hat{\mathbb{E}}[\psi_t \mid h_t] = \sum_{s=1}^{N} \beta_{s|h_t} \psi_s$$
$$\hat{\mathbb{E}}[\xi_t \mid h_t] = \sum_{s=1}^{N} \gamma_{s|h_t} \xi_s,$$

where β_s and γ_s depend on h_t. This produces the following training operators for S2 regression:

$$\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}\mathbf{B}$$
$$\tilde{\boldsymbol{\Xi}} = \boldsymbol{\Xi}\boldsymbol{\Gamma},$$

where $\mathbf{B}_{st} = \beta_{s|h_t}$ and $\boldsymbol{\Gamma}_{st} = \gamma_{s|h_t}$. With this data, S2 regression uses a Gram matrix formulation to estimate the operator

$$W = \boldsymbol{\Xi}\boldsymbol{\Gamma}(\mathbf{B}^\top G_{\mathcal{X},\mathcal{X}}\mathbf{B} + \lambda I_N)^{-1}\mathbf{B}^\top \boldsymbol{\Psi}^* \tag{A.8}$$

Note that we can use an arbitrary method to estimate B. Using conditional mean maps, the weight matrix $\mathbf{B}$ is computed using kernel ridge regression

$$\mathbf{B} = (G_{\mathcal{Z},\mathcal{Z}} + \lambda I_N)^{-1} G_{\mathcal{Z},\mathcal{Z}} \tag{A.9}$$

HSE-PSR learning is similar to this setting, with ψ_t being a conditional expectation operator of test observations given test actions. For this reason, kernel ridge regression is replaced by application of kernel Bayes rule [9].

For each t, S1 regression will produce a denoised prediction $\hat{E}[\xi_t \mid h_t]$ as a linear combination of training feature maps

$$\hat{E}[\xi_t \mid h_t] = \boldsymbol{\Xi}\alpha_t = \sum_{s=1}^{N} \alpha_{t,s}\xi_s$$

This corresponds to the covariance operators

$$\hat{\Sigma}_{\psi_{t+1}o_t|h_t} = \sum_{s=1}^{N} \alpha_{t,s}\psi_{s+1} \otimes o_s = \boldsymbol{\Psi}'\mathrm{diag}(\alpha_t)\mathbf{O}^*$$
$$\hat{\Sigma}_{o_t o_t|h_t} = \sum_{s=1}^{N} \alpha_{t,s} o_s \otimes o_s = \mathbf{O}\mathrm{diag}(\alpha_t)\mathbf{O}^*$$

Where, $\boldsymbol{\Psi}'$ is the shifted future training operator satisfying $\boldsymbol{\Psi}' e_t = \psi_{t+1}$ Given these two covariance operators, we can use kernel Bayes rule [9] to condition on o_t which gives

$$Q_{t+1} = \hat{E}[\psi_{t+1} \mid h_t] = \hat{\Sigma}_{\psi_{t+1}o_t|h_t}(\hat{\Sigma}_{o_t o_t|h_t} + \lambda I)^{-1} o_t. \tag{A.10}$$

Replacing o_t in (A.10) with its conditional expectation $\sum_{s=1}^{N} \alpha_s o_s$ corresponds to marginalizing over o_t (i.e. prediction). A stable Gram matrix formulation for (A.10) is given by [9]

$$\begin{aligned} & Q_{t+1} \\ & \quad = \mathbf{\Psi}' \mathrm{diag}(\alpha_t) G_{\mathcal{O},\mathcal{O}} ((\mathrm{diag}(\alpha_t) G_{\mathcal{O},\mathcal{O}})^2 + \lambda N I)^{-1} \\ & \quad\quad .\mathrm{diag}(\alpha_t) \mathbf{O}^* o_{t+1} \\ & \quad = \mathbf{\Psi}' \tilde{\alpha}_{t+1}, \end{aligned} \tag{A.11}$$

which is the state update equation in HSE-PSR. Given $\tilde{\alpha}_{t+1}$ we perform S2 regression to estimate

$$\hat{P}_{t+1} = \hat{\mathbb{E}}[\xi_{t+1} \mid o_{1:t+1}] = \mathbf{\Xi} \alpha_{t+1} = W \mathbf{\Psi}' \tilde{\alpha}_{t+1},$$

where W is defined in (A.8).

Appendix B. Proofs

B.1 Proof of Main Theorem

In this section we provide a proof for theorem 2. We provide finite sample analysis of S1, covariance estimation and regularization effects. The asymptotic statement becomes a natural consequence.

We will make use of matrix Brenestein's inequality stated below:

Lemma B.1 (Matrix Bernestein's Inequality [12]) *Let A be a random square symmetric matrix, and $r > 0$, $v > 0$ and $k > 0$ be such that, almost surely,*

$$\begin{aligned} \mathbb{E}[A] = 0, \quad \lambda_{\max}[A] \leq r, \\ \lambda_{\max}[\mathbb{E}[A^2]] \leq v, \quad \mathrm{tr}(\mathbb{E}[A^2]) \leq k. \end{aligned}$$

If $A^{(1)}, A^{(2)}, \ldots, A_t$ are independent copies of A, then for any $t > 0$,

$$\begin{aligned} \Pr \left[\lambda_{\max} \left[\frac{1}{N} \sum_{t=1}^{N} A_n \right] > \sqrt{\frac{2vt}{N}} + \frac{rt}{3N} \right] \\ \leq \frac{kt}{v} (e^t - t - 1)^{-1}. \end{aligned} \tag{B.1}$$

If $t \geq 2.6$, then $t(E^t - t - 1)^{-1} \leq e^{-t/2}$.

Recall that we have four sources of error: first, the error due to the input x_{test} not being in $\mathcal{R}(\Sigma_{\bar{x}\bar{x}})$; second, error in S1 regression causes the input to S2 regression procedure $(\hat{x}, \hat{y})$ to be perturbed version of the true $(\bar{x}, \bar{y})$; third, the covariance operators are estimated from a finite sample of size N; and fourth, there is the effect of regularization. In the

proof, we characterize the effect of each source of error. To do so, we define the following intermediate quantities:

$$W_\lambda = \Sigma_{\bar{y}\bar{x}} \left(\Sigma_{\bar{x}\bar{x}} + \lambda I\right)^{-1} \tag{B.2}$$

$$\bar{W}_\lambda = \hat{\Sigma}_{\bar{y}\bar{x}} \left(\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I\right)^{-1}, \tag{B.3}$$

where

$$\hat{\Sigma}_{\bar{y}\bar{x}} \equiv \frac{1}{N} \sum_{t=1}^{N} \bar{y}_t \otimes \bar{x}_t$$

and $\hat{\Sigma}_{\bar{x}\bar{x}}$ is defined similarly. Basically, W_λ captures only the effect of regularization and $\bar{W}_\lambda$ captures in addition the effect of finite sample estimate of the covariance. $\bar{W}_\lambda$ is the result of regression if $\bar{x}$ and $\bar{y}$ were observed instead of $\hat{x}$ and $\hat{y}$. It is important to note that $\hat{\Sigma}_{\bar{x}\bar{y}}$ and $\hat{\Sigma}_{\bar{x}\bar{x}}$ are *not* observable quantities since they depend on the true expectations $\bar{x}$ and $\bar{y}$. We will use λ_{xi} and λ_{yy} denote the eigenvalues of $\Sigma_{\bar{x}\bar{x}}$ and $\Sigma_{\bar{y}\bar{y}}$ respectively, in descending order and will use $\|.\|$ to denote the operator norm.

Before we prove the main theorem, we define the quantities $\zeta_{\delta,N}^{\bar{x}\bar{x}}$ and $\zeta_{\delta,N}^{\bar{x}\bar{y}}$ which we use to bound the effect of covariance estimation from finite data, as stated in the following lemma:

Lemma B.2 (Covariance error bound) *Let N be a positive integer and $\delta \in (0,1)$ and assume that $\|\bar{x}\|, \|\bar{y}\| < c < \infty$ almost surely. Let $\zeta_{\delta,N}^{\bar{x}\bar{y}}$ be defined as:*

$$\zeta_{\delta,N}^{\bar{x}\bar{y}} = \sqrt{\frac{2vt}{N}} + \frac{rt}{3N}, \tag{B.4}$$

where

$$\begin{aligned} t &= \max(2.6, 2\log(2k/\delta v)) \\ r &= c^2 + \|\Sigma_{\bar{x}\bar{y}}\| \\ v &= c^2 \max(\lambda_{y1}, \lambda_{x1}) + \|\Sigma_{\bar{x}\bar{y}}\|^2 \\ k &= 2c^2 \sqrt{\mathrm{tr}(\Sigma_{\bar{x}\bar{x}})\mathrm{tr}(\Sigma_{\bar{y}\bar{y}})} \end{aligned}$$

Similarly, let $\zeta_{\delta,N}^{\bar{x}\bar{x}}$ be defined as:

$$\zeta_{\delta,N}^{\bar{x}\bar{x}} = \sqrt{\frac{2v't'}{N}} + \frac{r't'}{3N}, \tag{B.5}$$

where

$$\begin{aligned} t' &= \max(2.6, 2\log(2k'/\delta v')) \\ r' &= c^2 + \lambda_{x1} \\ v' &= c^2\lambda_{x1} + \lambda_{x1}^2 \\ k' &= c^2 \mathrm{tr}(\Sigma_{\bar{x}\bar{x}}) \end{aligned}$$

It follows that, with probability at least $1-\delta/2$*,*

$$\|\hat{\Sigma}_{\bar{x}\bar{x}} - \Sigma_{\bar{x}\bar{x}}\| < \zeta_{\delta,N}^{\bar{x}\bar{x}}$$
$$\|\hat{\Sigma}_{\bar{y}\bar{x}} - \Sigma_{\bar{y}\bar{x}}\| < \zeta_{\delta,N}^{\bar{x}\bar{y}}.$$

Proof Lemma B.1 shows that in order to bound the probability of exceeding the bound to be below $\delta/2$, t can be set to $\max(2.6, 2k\log(2/\delta v))$. So, it remains to find suitable values for r, v and k.

We start with $\zeta_{\delta,N}^{\bar{x}\bar{x}}$. By setting $A_t = \bar{x}_t \otimes \bar{x}_t - \Sigma_{\bar{x}\bar{x}}$ we get

$$\begin{aligned}
&\lambda_{\max}[A] \leq \|\bar{x}\|^2 + \|\Sigma_{\bar{x}\bar{x}}\| \leq c^2 + \lambda_{x1} = r', \\
&\lambda_{\max}[\mathbb{E}[A^2]] = \lambda_{\max}[\mathbb{E}[\|\bar{x}\|^2(\bar{x}\otimes\bar{x}) \\
&\quad - (\bar{x}\otimes\bar{x})\Sigma_{\bar{x}\bar{x}} + \Sigma_{\bar{x}\bar{x}}(\bar{x}\otimes\bar{x}) - {\Sigma_{\bar{x}\bar{x}}}^2] \\
&\quad = \lambda_{\max}[\mathbb{E}[\|\bar{x}\|^2(\bar{x}\otimes\bar{x}) - {\Sigma_{\bar{x}\bar{x}}}^2] \\
&\quad \leq c^2\lambda_{x1} + \lambda_{x1}^2 = v' \\
&\mathrm{tr}[\mathbb{E}[A^2]] = \mathrm{tr}[\mathbb{E}[\|\bar{x}\|^2(\bar{x}\otimes\bar{x}) - {\Sigma_{\bar{x}\bar{x}}}^2] \\
&\quad \leq \mathrm{tr}[\mathbb{E}[\|\bar{x}\|^2(\bar{x}\otimes\bar{x})] \leq c^2\mathrm{tr}(\Sigma_{\bar{x}\bar{x}}) = k'
\end{aligned}$$

Now moving to $\zeta_{\delta,N}^{\bar{x}\bar{y}}$, we have $B_t = \bar{y}_t \otimes \bar{x}_t - \Sigma_{\bar{y}\bar{x}}$. Since B_t is not square, we use the Hermitian dilation similar to [24]:

$$A = \mathscr{H}(B) = \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$$

Note that

$$\lambda_{\max}[A] = \|B\|, \quad A^2 = \begin{bmatrix} BB^* & 0 \\ 0 & B^*B \end{bmatrix}$$

therefore suffices to bound $\lambda_{\max}[\frac{1}{N}\sum_{t=1}^{N} A_t]$ using an argument similar to that used in $\zeta_{\delta,N}xx$ case. ■

To prove theorem 2, we write

$$\begin{aligned}
\|\hat{W}_\lambda x_{\text{test}} - W x_{\text{test}}\|_{\mathcal{Y}} \leq &\|(\hat{W}_\lambda - \bar{W}_\lambda)\bar{x}_{\text{test}}\|_{\mathcal{Y}} \\
&+ \|(\bar{W}_\lambda - W_\lambda)\bar{x}_{\text{test}}\|_{\mathcal{Y}} \\
&+ \|(W_\lambda - W)\bar{x}_{\text{test}}\|_{\mathcal{Y}} \qquad \text{(B.6)}
\end{aligned}$$

We will now present bounds on each term. We consider the case where $\bar{x}_{\text{test}} \in \mathcal{R}(\Sigma_{\bar{x}\bar{x}})$. Extension to $\overline{\mathcal{R}(\Sigma_{\bar{x}\bar{x}})}$ is a result of the assumed boundedness of W, which implies the boundedness of $\hat{W}_\lambda - W$.

Lemma B.3 (Error due to S1 Regression) *Assume that $\|\bar{x}\|, \|\bar{y}\| < c < \infty$ almost surely, and let $\eta_{\delta,N}$ be as defined in Definition 1. The following holds with probability at least $1-\delta$*

$$\begin{aligned}\|\hat{W}_\lambda - \bar{W}_\lambda\| &\leq \sqrt{\lambda_{y1} + \zeta^{\bar{y}\bar{y}}_{\delta,N}}\frac{(2c\eta_{\delta,N} + {\eta_{\delta,N}}^2)}{\lambda^{\frac{3}{2}}} \\ &+ \frac{(2c\eta_{\delta,N} + {\eta_{\delta,N}}^2)}{\lambda} \\ &= O\left(\eta_{\delta,N}\left(\frac{1}{\lambda} + \frac{\sqrt{1+\frac{\log(1/\delta)}{\sqrt{N}}}}{\lambda^{\frac{3}{2}}}\right)\right).\end{aligned}$$

The asymptotic statement assumes $\eta_{\delta,N} \to 0$ as $N \to \infty$.

Proof Write $\hat{\Sigma}_{\hat{x}\hat{x}} = \Sigma_{\bar{x}\bar{x}} + \Delta_x$ and $\hat{\Sigma}_{\hat{y}\hat{x}} = \Sigma_{\bar{y}\bar{x}} + \Delta_{yx}$. We know that, with probability at least $1-\delta/2$, the following is satisfied for all unit vectors $\phi_x \in \mathcal{X}$ and $\phi_x \in \mathcal{Y}$

$$\begin{aligned}\langle \phi_y, \Delta_{yx}\phi_x\rangle_{\mathcal{Y}} &= \frac{1}{N}\sum_{t=1}^{N}\langle \phi_y, \hat{y}_t\rangle_{\mathcal{Y}}\langle \phi_x, \hat{x}_t\rangle_{\mathcal{X}} \\ &- \langle \phi_y, \hat{y}_t\rangle_{\mathcal{Y}}\langle \phi_x, \bar{x}_t\rangle_{\mathcal{X}} \\ &+ \langle \phi_y, \hat{y}_t\rangle_{\mathcal{Y}}\langle \phi_x, \bar{x}_t\rangle_{\mathcal{X}} - \langle \phi_y, \bar{y}_t\rangle_{\mathcal{Y}}\langle \phi_x, \bar{x}_t\rangle_{\mathcal{X}} \\ &= \frac{1}{N}\sum_{t}\langle \phi_y, \bar{y}_t + (\hat{y}_t - \bar{y}_t)\rangle_{\mathcal{Y}}\langle \phi_x, \hat{x}_t - \bar{x}_t\rangle_{\mathcal{X}} \\ &+ \langle \phi_y, \hat{y}_t - \bar{y}_t\rangle_{\mathcal{Y}}\langle \phi_x, \bar{x}_t\rangle_{\mathcal{X}} \\ &\leq 2c\eta_{\delta,N} + \eta^2_{\delta,N}\end{aligned}$$

Therefore,

$$\|\Delta_{yx}\| = \sup_{\|\phi_x\|_{\mathcal{X}}\leq 1, \|\phi_y\|_{\mathcal{Y}}\leq 1}\langle \phi_y, \Delta_{yx}\phi_x\rangle_{\mathcal{Y}} \leq 2c\eta_{\delta,N} + \eta^2_{\delta,N},$$

and similarly

$$\|\Delta_x\| \leq 2c\eta_{\delta,N} + {\eta_{\delta,N}}^2,$$

with probability $1-\delta/2$. We can write

$$\begin{aligned}\hat{W}_\lambda - \bar{W}_\lambda = \hat{\Sigma}_{\bar{y}\bar{x}}&\left((\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} - (\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I)^{-1}\right) \\ &+ \Delta_{yx}(\hat{\Sigma}_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}\end{aligned}$$

Using the fact that $B^{-1} - A^{-1} = B^{-1}(A-B)A^{-1}$ for invertible operators A and B we get

$$\begin{aligned}\hat{W}_\lambda - \bar{W}_\lambda = &-\hat{\Sigma}_{\bar{y}\bar{x}}(\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I)^{-1}\Delta_x(\hat{\Sigma}_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} \\ &+ \Delta_{yx}(\hat{\Sigma}_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}\end{aligned}$$

we then use the decomposition $\hat{\Sigma}_{\bar{y}\bar{x}} = \hat{\Sigma}_{\bar{y}\bar{y}}^{\frac{1}{2}} V \hat{\Sigma}_{\bar{x}\bar{x}}^{\frac{1}{2}}$, where V is a correlation operator satisfying $\|V\| \leq 1$. This gives

$$\begin{aligned}
\hat{W}_\lambda - \bar{W}_\lambda = & \\
& - \hat{\Sigma}_{\bar{y}\bar{y}}^{\frac{1}{2}} V \hat{\Sigma}_{\bar{x}\bar{x}}^{\frac{1}{2}} (\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}} (\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}} \\
& . \Delta_x (\hat{\Sigma}_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} \\
& + \Delta_{yx} (\hat{\Sigma}_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}
\end{aligned}$$

Noting that $\|\hat{\Sigma}_{\bar{x}\bar{x}}^{\frac{1}{2}} (\hat{\Sigma}_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}}\| \leq 1$, the rest of the proof follows from triangular inequality and the fact that $\|AB\| \leq \|A\|\|B\|$ ■

Lemma B.4 (Error due to Covariance) *Assuming that $\|\bar{x}\|_{\mathcal{X}}, \|\bar{y}\|_{\mathcal{Y}} < c < \infty$ almost surely, the following holds with probability at least $1 - \frac{\delta}{2}$*

$$\|\bar{W}_\lambda - W_\lambda\| \leq \sqrt{\lambda_{y1}} \zeta_{\delta,N}^{\bar{x}\bar{x}} \lambda^{-\frac{3}{2}} + \frac{\zeta_{\delta,N}^{\bar{x}\bar{y}}}{\lambda}$$

, where $\zeta_{\delta,N}^{\bar{x}\bar{x}}$ and $\zeta_{\delta,N}^{\bar{x}\bar{y}}$ are as defined in Lemma B.2.

Proof Write $\hat{\Sigma}_{\bar{x}\bar{x}} = \Sigma_{\bar{x}\bar{x}} + \Delta_x$ and $\hat{\Sigma}_{\bar{y}\bar{x}} = \Sigma_{\bar{y}\bar{x}} + \Delta_{yx}$. Then we get

$$\begin{aligned}
\bar{W}_\lambda - W_\lambda = & \Sigma_{\bar{y}\bar{x}} \left((\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} - (\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-1} \right) \\
& + \Delta_{yx} (\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}
\end{aligned}$$

Using the fact that $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$ for invertible operators A and B we get

$$\begin{aligned}
\bar{W}_\lambda - W_\lambda = & -\Sigma_{\bar{y}\bar{x}} (\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-1} \Delta_x (\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} \\
& + \Delta_{yx} (\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}
\end{aligned}$$

we then use the decomposition $\Sigma_{\bar{y}\bar{x}} = \Sigma_{\bar{y}\bar{y}}{}^{\frac{1}{2}} V \Sigma_{\bar{x}\bar{x}}{}^{\frac{1}{2}}$, where V is a correlation operator satisfying $\|V\| \leq 1$. This gives

$$\begin{aligned}
\bar{W}_\lambda - W_\lambda = & \\
& - \Sigma_{\bar{y}\bar{y}}{}^{\frac{1}{2}} V \Sigma_{\bar{x}\bar{x}}{}^{\frac{1}{2}} (\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}} (\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}} \\
& . \Delta_x (\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1} \\
& + \Delta_{yx} (\Sigma_{\bar{x}\bar{x}} + \Delta_x + \lambda I)^{-1}
\end{aligned}$$

Noting that $\|\Sigma_{\bar{x}\bar{x}}{}^{\frac{1}{2}} (\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-\frac{1}{2}}\| \leq 1$, the rest of the proof follows from triangular inequality and the fact that $\|AB\| \leq \|A\|\|B\|$ ■

Lemma B.5 (Error due to Regularization on inputs within $\mathcal{R}(\Sigma_{\bar{x}\bar{x}})$) *For any $x \in \mathcal{R}(\Sigma_{\bar{x}\bar{x}})$ s.t. $\|x\|_{\mathcal{X}} \leq 1$ and $\|{\Sigma_{\bar{x}\bar{x}}}^{-\frac{1}{2}} x\|_{\mathcal{X}} \leq C$. The following holds*

$$\|(W_\lambda - W)x\|_{\mathcal{Y}} \leq \frac{1}{2}\sqrt{\lambda}\|W\|_{HS}C$$

Proof Since $x \in \mathcal{R}(\Sigma_{\bar{x}\bar{x}}) \subseteq \mathcal{R}(\Sigma_{\bar{x}\bar{x}}^{\frac{1}{2}})$, we can write $x = \Sigma_{\bar{x}\bar{x}}^{\frac{1}{2}} v$ for some $v \in \mathcal{X}$ s.t. $\|v\|_{\mathcal{X}} \leq C$. Then

$$(W_\lambda - W)x = \Sigma_{\bar{y}\bar{x}}((\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-1} - {\Sigma_{\bar{x}\bar{x}}}^{-1})\Sigma_{\bar{x}\bar{x}}^{\frac{1}{2}} v$$

Let $D = \Sigma_{\bar{y}\bar{x}}((\Sigma_{\bar{x}\bar{x}} + \lambda I)^{-1} - {\Sigma_{\bar{x}\bar{x}}}^{-1})\Sigma_{\bar{x}\bar{x}}^{\frac{1}{2}}$. We will bound the Hilbert-Schmidt norm of D. Let $\psi_{xi} \in \mathcal{X}$, $\psi_{yi} \in \mathcal{Y}$ denote the eigenvector corresponding to λ_{xi} and λ_{yi} respectively. Define $s_{ij} = |\langle \psi_{yj}, \Sigma_{\bar{x}\bar{y}}\psi_{xi}\rangle_{\mathcal{Y}}|$. Then we have

$$\begin{aligned}
|\langle \psi_{yj}, D\psi_{xi}\rangle_{\mathcal{Y}}| &= \langle \psi_{yj}, \Sigma_{\bar{y}\bar{x}} \frac{\lambda}{(\lambda_{xi} + \lambda)\sqrt{\lambda_{xi}}} \psi_{xi}\rangle_{\mathcal{Y}} \\
&= \frac{\lambda s_{ij}}{(\lambda_{xi} + \lambda)\sqrt{\lambda_{xi}}} = \frac{s_{ij}}{\sqrt{\lambda_{xi}}} \frac{1}{\frac{1}{\lambda/\lambda_{xi}} + 1} \\
&\leq \frac{s_{ij}}{\sqrt{\lambda_{xi}}} \cdot \frac{1}{2}\sqrt{\frac{\lambda}{\lambda_{xi}}} = \frac{1}{2}\sqrt{\lambda}\frac{s_{ij}}{\lambda_{xi}} \\
&= \frac{1}{2}\sqrt{\lambda}|\langle \psi_{yj}, W\psi_{xi}\rangle_{\mathcal{Y}}|,
\end{aligned}$$

where the inequality follows from the arithmetic-geometric-harmonic mean inequality. This gives the following bound

$$\|D\|_{HS}^2 = \sum_{i,j} \langle \psi_{yj}, D\psi_{xi}\rangle_{\mathcal{Y}}^2 \leq \frac{1}{2}\sqrt{\lambda}\|W\|_{HS}^2$$

and hence

$$\begin{aligned}
\|(W_\lambda - W)x\|_{\mathcal{Y}} &\leq \|D\|\|v\|_{\mathcal{X}} \leq \|D\|_{HS}\|v\|_{\mathcal{X}} \\
&\leq \frac{1}{2}\sqrt{\lambda}\|W\|_{HS}C
\end{aligned}$$

■

Note that the additional assumption that $\|{\Sigma_{\bar{x}\bar{x}}}^{-\frac{1}{2}} x\|_{\mathcal{X}} \leq C$ is not required to obtain an asymptotic $O(\sqrt{\lambda})$ rate for a given x. This assumption, however, allows us to uniformly bound the constant. Theorem 2 is simply the result of plugging the bounds in Lemmata B.3, B.4, and B.5 into (B.6) and using the union bound.

B.2 Proof of Lemma 5

for $t = 1$: Let $\mathcal{I}$ be an index set over training instances such that

$$\hat{Q}_1^{\text{test}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{Q}_i$$

Then

$$\|\hat{Q}_1^{\text{test}} - \tilde{Q}_1^{\text{test}}\|_{\mathcal{X}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\hat{Q}_i - \tilde{Q}_i\|_{\mathcal{X}} \leq \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\hat{Q}_i - Q_i\|_{\mathcal{X}} \leq \eta_{\delta,N}$$

for $t > 1$: Let A denote a projection operator on $\mathcal{R}^\top(\Sigma_{\bar{y}\bar{y}})$

$$\begin{aligned}
\|\hat{Q}_{t+1}^{\text{test}} - \tilde{Q}_{t+1}^{\text{test}}\|_{\mathcal{X}} &\leq L\|\hat{P}_t^{\text{test}} - \tilde{P}_t^{\text{test}}\|_{\mathcal{Y}} \leq L\|A\hat{W}_\lambda \hat{Q}_t^{\text{test}}\|_{\mathcal{Y}} \\
&\leq L \left\| \frac{1}{N} \left(\sum_{i=1}^{N} A\hat{P}_i \otimes \hat{Q}_i \right) \left(\frac{1}{N} \sum_{i=1}^{N} \hat{Q}_i \otimes \hat{Q}_i + \lambda I \right)^{-1} \right\| \left\| \hat{Q}_t^{\text{test}} \right\|_{\mathcal{X}} \\
&\leq L \left\| \frac{1}{N} \sum_{i=1}^{N} A\hat{P}_i \otimes A\hat{P}_i \right\|^{\frac{1}{2}} \frac{1}{\sqrt{\lambda}} \|\hat{Q}_t^{\text{test}}\|_{\mathcal{X}} \leq L \frac{\eta_{\delta,N}}{\sqrt{\lambda}} \|\hat{Q}_t^{\text{test}}\|_{\mathcal{X}},
\end{aligned}$$

where the second to last inequality follows from the decomposition similar to $\Sigma_{YX} = \Sigma_Y^{\frac{1}{2}} V \Sigma_X^{\frac{1}{2}}$, and the last inequality follows from the fact that $\|A\hat{P}_i\|_{\mathcal{Y}} \leq \|\hat{P}_i - \bar{P}_i\|_{\mathcal{Y}}$. □