Data Analysis Project: Σ-Optimality for Active Learning on Gaussian Random Fields

Yifei Ma
Machine Learning Department, Carnegie Mellon University
yifeim@cs.cmu.edu

Committee: Jeff Schneider, Barnabas Poczos, Roy Maxion.
Machine Learning Department, Carnegie Mellon University
{schneide,poczos,maxion}@cs.cmu.edu

Abstract

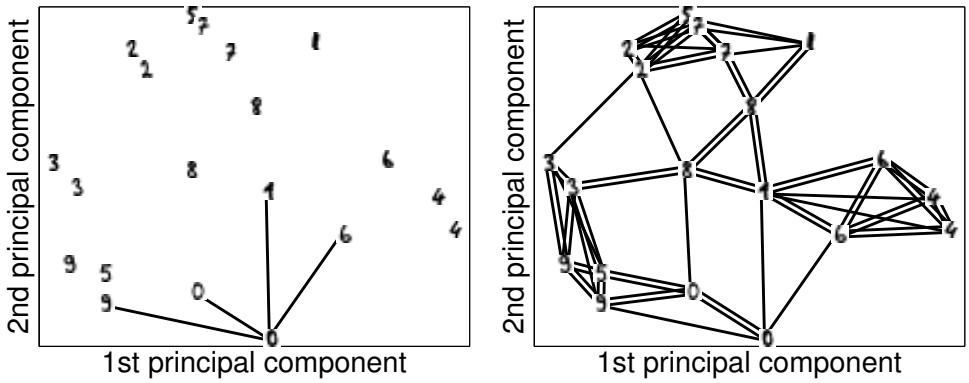
A common classifier for unlabeled nodes on undirected graphs uses label propagation from the labeled nodes, equivalent to the harmonic predictor on Gaussian random fields (GRFs). For active learning on GRFs, the commonly used V-optimality criterion queries nodes that reduce the L^2 (regression) loss. V-optimality satisfies a submodularity property showing that greedy reduction produces a $(1 - 1/e)$ globally optimal solution. However, L^2 loss may not characterise the true nature of 0/1 loss in classification problems and thus may not be the best choice for active learning.

We consider a new criterion we call Σ-optimality, which queries the node that minimizes the sum of the elements in the predictive covariance. Σ-optimality directly optimizes the risk of the surveying problem, which is to determine the proportion of nodes belonging to one class. In this paper we extend submodularity guarantees from V-optimality to Σ-optimality using properties specific to GRFs. We further show that GRFs satisfy the *suppressor-free condition* in addition to the conditional independence inherited from Markov random fields. We test Σ-optimality on real-world graphs with both synthetic and real data and show that it outperforms V-optimality and other related methods on classification.

1 Introduction

Real-world data are often presented as a graph where the nodes in the graph bear labels that vary smoothly along edges. For example, for scientific publications, the content of one paper is highly correlated with the content of papers that it references or is referenced by, the field of interest of a scholar is highly correlated with other scholars s/he coauthors with, etc. Many of these networks can be described using an undirected graph with nonnegative edge weights set to be the strengths of the connections between nodes.

The main character of graph-based representation of data is that all features of a node are implicitly characterized by its edges. Despite that many datasets are naturally represented by graphs, a feature-based database can also be easily turned into a graph by considering the k-nearest-neighbors among input feature vector pairs. In this way, the similarity of input features between different instances (i.e. nodes) is preserved. Figure 1 shows how to construct a graph on a toy dataset where input features are images of 8-by-8-pixel hand-written digits. To visualize the process, we used the scores of the first two principal components as the coordinate of an instance. The distance between two instances is designed to be the Euclidean distance of their corresponding 2-dimensional coordinates. Figure 1(a) demonstrate the 4 nearest neighbors on this feature space of the node at the bottom. Figure 1(b) is the final k-nn graph, where edge directions are removed. Using label propagation, which is a graph inference technique described below, the constructed k-nn graph can be used to predict the actual number indicated by each image in the test set, just like a feature-base database. In general, the more interesting results come from network graphs.



(a) The 4 nearest neighbors of one node (b) Final 4-nn graph

Figure 1: The 4-nn graph constructed from input images using the Euclidean distance on the first 2 PCA projection of concatenated pixel values.

The model for label prediction in this paper is the harmonic function on the Gaussian random field (GRF) by Zhu et al. (2003). It can generalize two popular and intuitive algorithms: label propagation (Zhu & Ghahramani, 2002), and random walk with absorptions (Wu et al., 2012). GRFs can be seen as a Gaussian process (GP) (Rasmussen & Williams, 2006) with its (maybe improper) prior covariance matrix whose (pseudo)inverse is set to be the graph Laplacian.

Specifically, the label propagation / random walk prediction used in our paper works as follows. To predict the label of a test node, start such a random walk from this node that if it arrives at node v^t at time t, during the next time step, it randomly traverses one outbound edge with probability proportional to the corresponding edge weight. This random walk terminates when it hits any labeled node. Given the above random experiment, the prediction model in this paper assigns the probability of this node having exactly the label A by the chance that this random walk hits a labeled node of class A before any other labeled classes. Figure 2 illustrates the probability that every unlabeled node belongs to the positive class, given the three labeled nodes (one positive and two negatives).

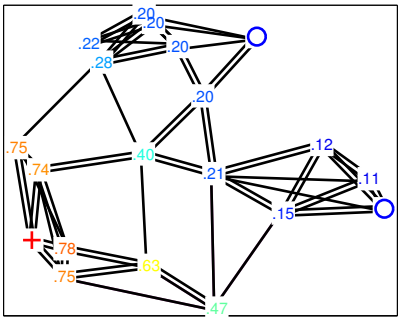
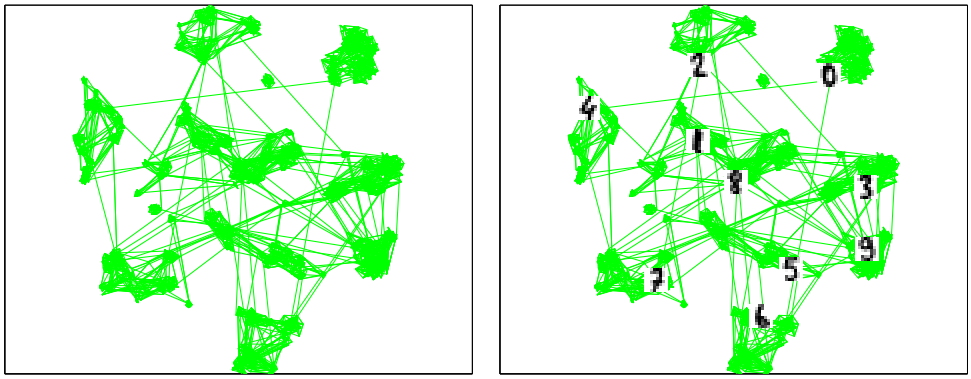


Figure 2: The number on every node is the chance that it belongs to class "+", predicted by the random walk model. Red "+" and blue "◯" are labeled nodes (i.e. training set) of both classes.

Like other learning problems, labels may be insufficient and expensive to gather, especially if one wants to discover a new phenomenon on the network. Active learning addresses these issues by making automated decisions on which nodes to query for labels from experts or the crowd. A visualization is shown in Figure 3.



(a) The actual knn graph visualized in 2D. (b) The first 10 labels picked by Σ-optimality.

Figure 3: Problem being solved: On actual graphs like (a) or network graphs, we want to find (b) a desirable training set labeling which helps classification the most, without any labels to start with.

1.1 Problem Being Solved

We consider the problem of designing a good active learning strategy that, under labeling budget constraints, selects which instances to query for labels that are most helpful for classification on a graph-represented database. We assume that the graph structure provides reasonable information for node classification in that the node class distribution is modeled by a Gaussian random field model with known hyper-parameters. The performance of a specific active learning strategy is measured by the classification accuracy using harmonic prediction that is based on label propagation.

Figure 3(b) shows an example of the first 10 labels picked by a reasonable active learning strategy, which is the Σ-optimality that we advocate.

1.2 Main Contributions and Related Work

We proposed a new strategy for the active learning problem by considering a criterion we call Σ-optimality, which is a variant of a recent variance minimization/Bayes risk minimization criterion. We compared its performance with other popular criteria including empirical risk minimization (Settles, 2010), mutual information gain (Krause et al., 2008), and V-optimality (Ji & Han, 2012). In our experiments, we show that Σ-optimality outperforms other approaches for active learning with GRFs for classification and surveying. Insights were also provided.

We also established several related theoretical results. Namely, we show that greedy reduction of Σ-optimality provides a $(1 - 1/e)$ approximation bound to the global optimum. We also show that Gaussian random fields satisfy the suppressor-free condition, described below.

1.2.1 V-optimality on Gaussian Random Fields

Ji & Han (2012) proposed greedy variance minimization as a cheap and high profile surrogate active classification criterion. To decide which node to query next, the active learning algorithm finds the unlabeled node which leads to the smallest average predictive variance on all other unlabeled nodes. It corresponds to standard V-optimality in optimal experiment design.

We will discuss several aspects of V-optimality on GRFs below: 1. The motivation behind V-optimality can be paraphrased as the expected risk minimization with the L^2-surrogate loss (Section 2.1). 2. The greedy solution to the set optimization problem in V-optimality is comparable to the global solution up to a constant (Theorem 1). 3. The greedy application of V-optimality can also be interpreted as a heuristic which selects nodes that have high correlation to nodes with high variances (Observation 4).

Some previous work is related to point 2 above. Nemhauser et al. (1978) shows that any *submodular*, monotone and normalized set function yields a $(1 - 1/e)$ global optimality guarantee for greedy

solutions. Our proof techniques coincides with Friedland & Gaubert (2011) in principle, but we are not restricted to spectral functions. Krause et al. (2008) showed a counter example where the V-optimality objective function with GP models does not satisfy submodularity.

1.2.2 Σ-optimality on Gaussian Random Fields

We define Σ-optimality on GRFs to be another variance minimization criterion that minimizes the sum of all entries in the predictive covariance matrix. As we will show in Lemma 7, the predictive covariance matrix is nonnegative entry-wise and thus the definition is proper. Σ-optimality was originally proposed by Garnett et al. (2012) in the context of *active surveying,* which is to determine the proportion of nodes belonging to one class. However, we focus on its performance as a criterion in active classification heuristics. The survey-risk of Σ-optimality replaces the L^2-risk of V-optimality as an alternative surrogate risk for the 0/1-risk.

We also prove that the greedy application of Σ-optimality has a similar theoretical bound as V-optimality. We will show that greedily minimizing Σ-optimality empirically outperforms greedily minimizing V-optimality on classification problems. The exact reason explaining the superiority of Σ-optimality as a surrogate loss in the GRF model is still an open question, but we observe that Σ-optimality tends to select cluster centers whereas V-optimality goes after outliers (Section 5.1). Finally, greedy application of both Σ-optimality and V-optimality need $\mathcal{O}(N)$ time per query candidate evaluation after one-time inverse of a $N \times N$ matrix.

1.2.3 GRFs Are Suppressor Free

In linear regression, an explanatory variable is called a suppressor if adding it as a new variable enhances correlations between the old variables and the dependent variable (Walker, 2003; Das & Kempe, 2008). Suppressors are persistent in real-world data. We show GRFs to be *suppressor-free*. Intuitively, this means that with more labels acquired, the conditional correlation between unlabeled nodes decreases even when their Markov blanket has not formed. That GRFs present natural examples for the otherwise obscure suppressor-free condition is interesting.

2 Approach: Learning Model & Active Learning Objectives

We use the *Gaussian random field/belief propagation* (GRF/BP) as our learning model. Suppose the dataset can be represented in the form of a connected undirected graph $G = (V, E)$ where each node has an (either known or unknown) label and each edge e_{ij} has a fixed nonnegative weight $w_{ij} (= w_{ji})$ that reflects the proximity, similarity, etc. between nodes v_i and v_j. Define the graph Laplacian of G to be $L = \text{diag}\,(W\mathbf{1}) - W$, i.e., $l_{ii} = \sum_j w_{ij}$ and $l_{ij} = -w_{ij}$ when $i \neq j$. Let $L_\delta = L + \delta I$ be the regularized Laplacian obtained by adding self-loops. In the following, we will write L to also encompass βL_δ for the set of hyper-parameters $\beta > 0$ and $\delta \geq 0$. The *binary* GRF is a Bayesian model to generate $y_i \in \{0, +1\}$ for every node v_i according to,

$$p(\boldsymbol{y}) \propto \exp\Big\{ -\frac{\beta}{2}\Big(\sum_{i,j} w_{ij}(y_i - y_j)^2 + \delta \sum_i y_i^2 \Big) \Big\} = \exp\Big(-\frac{1}{2} \boldsymbol{y}^T L \boldsymbol{y} \Big). \tag{2.1}$$

Suppose nodes $\boldsymbol{\ell} = \{v_{\ell_1}, \ldots, v_{\ell_{|\boldsymbol{\ell}|}}\}$ are labeled as $\boldsymbol{y_\ell} = (y_{\ell_1}, \ldots, y_{\ell_{|\boldsymbol{\ell}|}})^T$; A GRF infers the output distribution on unlabeled nodes, $\boldsymbol{y_u} = (y_{u_1}, \ldots, y_{u_{|\boldsymbol{u}|}})^T$ by the conditional distribution given $\boldsymbol{y}_\ell$, as

$$\Pr(\boldsymbol{y_u} | \boldsymbol{y_\ell}) \propto \mathcal{N}(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, L_{\boldsymbol{u}}^{-1}) = \mathcal{N}(\hat{\boldsymbol{y}}_{\boldsymbol{u}}, L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}), \tag{2.2}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{u}} = (-L_{\boldsymbol{u}}^{-1} L_{\boldsymbol{u}\boldsymbol{\ell}} \boldsymbol{y}_{\boldsymbol{\ell}})$ is the vector of predictive means on unlabeled nodes and $L_{\boldsymbol{u}}$ is the principal submatrix consisting of the unlabeled row and column indices in L, that is, the lower-right block of $L = \begin{pmatrix} L_{\boldsymbol{\ell}} & L_{\boldsymbol{\ell}\boldsymbol{u}} \\ L_{\boldsymbol{u}\boldsymbol{\ell}} & L_{\boldsymbol{u}} \end{pmatrix}$. By convention, $L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}$ means the inverse of the principal submatrix. We use $L_{(\boldsymbol{v}-\boldsymbol{\ell})}$ and $L_{\boldsymbol{u}}$ interchangeably because $\boldsymbol{\ell}$ and $\boldsymbol{u}$ partition the set of all nodes $\boldsymbol{v}$.

Finally, GRF, or GRF/LP, is a relaxation of the *binary* GRF to continuous outputs, because the latter is computationally intractable even for *a-priori* generations. LP stands for label propagation, because the predictive mean on a node is the probability of a random walk leaving that node hitting a positive

label before hitting a zero label. For multi-class problems, Zhu et al. (2003) proposed the *harmonic predictor* which looks at predictive means in one-versus-all comparisons.

Remark: An alternative approximation to the *binary* GRF is the GRF-sigmoid model, which draws the binary outputs from Bernoulli distributions with means set to be the sigmoid function of the GRF (latent) variables. However, this alternative is very slow to compute and may not be compatible with the theoretical results in this paper.

2.1 Active Learning Objective 1: L^2 Risk Minimization (V-Optimality)

Since in GRFs, regression responses are taken directly as probability predictions, it is computationally and analytically more convenient to apply the regression loss directly in the GRF as in Ji & Han (2012). Assume the L^2 loss to be our classification loss. The risk function, whose input variable is the labeled subset $\boldsymbol{\ell}$, is:

$$\begin{aligned} R_V(\boldsymbol{\ell}) &= \mathbb{E}^{\boldsymbol{y_\ell} \boldsymbol{y_u}} \textstyle\sum_{v_{u_i} \in \boldsymbol{u}} (y_{u_i} - \hat{y}_{u_i})^2 \\ &= \mathbb{E}\left[\mathbb{E}\left[\textstyle\sum_i \left(\boldsymbol{y}_{u_i} - (-L_{\boldsymbol{u}}^{-1} L_{\boldsymbol{u}\boldsymbol{\ell}} \boldsymbol{y}_{\boldsymbol{\ell}})_i\right)^2 \middle| \boldsymbol{y}_{\boldsymbol{\ell}}\right]\right] = \operatorname{tr}(L_{\boldsymbol{u}}^{-1}). \end{aligned} \tag{2.3}$$

This risk is written with a subscript V because minimizing (2.3) is also the V-optimality criterion, which minimizes mean prediction variance in active learning.

In active learning, we strive to select a subset $\boldsymbol{\ell}$ of nodes to query for labels, constrained by a given budget C, such that the risk is minimized. Formally,

$$\underset{\boldsymbol{\ell}:\ |\boldsymbol{\ell}| \le C}{\arg\min} \quad R(\boldsymbol{\ell}) = R_V(\boldsymbol{\ell}) = \operatorname{tr}(L_{(\boldsymbol{v}-\boldsymbol{\ell})}^{-1}). \tag{2.4}$$

2.2 Active Learning Objective 2: Survey Risk Minimization (Σ-Optimality)

Another objective building on the GRF model (2.2) is to determine the proportion of nodes belonging to class 1, as would happen when performing a survey. For active surveying, the risk would be:

$$R_\Sigma(\boldsymbol{\ell}) = \mathbb{E}^{y_{\boldsymbol{\ell}} y_{\boldsymbol{u}}}\Big(\sum_{u_i \in \boldsymbol{u}} y_{u_i} - \sum_{u_i \in \boldsymbol{u}} \hat{y}_{u_i}\Big)^2 = \mathbb{E}\big[\mathbb{E}\big[\big(\mathbf{1}^T \boldsymbol{y_u} - \mathbf{1}^T \hat{\boldsymbol{y}}_{\boldsymbol{u}}\big)^2 | y_{\boldsymbol{\ell}}\big]\big] = \mathbf{1}^T L_{\boldsymbol{u}}^{-1} \mathbf{1}, \tag{2.5}$$

which could substitute the risk $R(\boldsymbol{\ell})$ in (2.4) and yield another heuristic for selecting nodes in batch active learning. We will refer to this modified optimization objective as the Σ-optimality heuristic:

$$\underset{\boldsymbol{\ell}:\ |\boldsymbol{\ell}| \le C}{\arg\min} \quad R(\boldsymbol{\ell}) = R_\Sigma(\boldsymbol{\ell}) = \mathbf{1}^T L_{(\boldsymbol{v}-\boldsymbol{\ell})}^{-1} \mathbf{1}. \tag{2.6}$$

Further, we will also consider the application of Σ-optimality in active classification because (2.6) is another metric of the predictive variance. Surprisingly, although both (2.3) and (2.5) are approximations of the real objective (the 0/1 risk), greedy reduction of the Σ-optimality criterion outperforms greedy reduction of the V-optimality criterion in active classification (Section 5.1), as well as several other methods including expected error reduction.

3 Methods

Our method that directly solves the active learning problem is the greedy application of Σ-optimality. Because it is a variant of the greedy application of V-optimality, both are described below, followed by algorithmic guarantees with proofs in the appendix. In the end is a summary of our method and other baseline comparison methods.

3.1 Algorithm for Greedy Application of Σ- and V-Optimality

Both (2.4) and (2.6) are subset optimization problems. Calculating the global optimum may be intractable. As will be shown later in the theoretical results, both objectives are submodular set functions and the greedy sequential update algorithm (Algorithm 1) yields a solution that has guaranteed approximation ratio to the optimum (Theorem 1).

Algorithm 1 Greedy subset selection.

Input: Graph Laplacian L, objective function $R(\boldsymbol{\ell})$, budget C.
Output: A subset $\boldsymbol{\ell} \subset \boldsymbol{v}$ by greedy selection.
Define $\boldsymbol{\ell}^{(0)} \leftarrow \emptyset$.
for $k = 1, 2, \ldots, C$ **do**
  Find $v_*^{(k)} \leftarrow \arg\min_v \Big(R(\boldsymbol{\ell}^{(k-1)} \cup \{v\}) - R(\boldsymbol{\ell}^{(k-1)}) \Big)$.
  Update $\boldsymbol{\ell}^{(k)} \leftarrow \boldsymbol{\ell}^{(k-1)} \cup \{v_*^{(k)}\}$.
end for

The following applies Algorithm 1 to our specific objective functions. At the k-th query decision, denote the covariance matrix conditioned on the previous $(k-1)$ queries as $\mathrm{C} = (L_{(\boldsymbol{v}-\boldsymbol{\ell}^{(k-1)})})^{-1}$. By Shur's Lemma (or the GP-regression update rule), the one-step look-ahead covariance matrix conditioned on $\boldsymbol{\ell}^{(k-1)} \cup \{v\}$, denoted as $\mathrm{C}' = (L_{(\boldsymbol{v}-(\boldsymbol{\ell}^{(k-1)} \cup \{v\}))})^{-1}$, has the following update formula:

$$\begin{pmatrix} \mathrm{C}' & 0 \\ 0 & 0 \end{pmatrix} = \mathrm{C} - \frac{1}{\mathrm{C}_{vv}} \cdot \mathrm{C}_{:v}\mathrm{C}_{v:}, \tag{3.1}$$

where without loss of generality v was positioned as the last node. Further denoting $\mathrm{C}_{ij} = \rho_{ij}\sigma_i\sigma_j$, we can put (3.1) inside $R_\Sigma(\cdot)$ and $R_V(\cdot)$ to get the following equivalent criteria:

$$\text{V-optimality}: \ v_*^{(k)} = \arg\max_{v \in \boldsymbol{u}} \ \frac{\sum_{t \in \boldsymbol{u}} (\mathrm{C}_{vt})^2}{\mathrm{C}_{vv}} = \sum_{t \in \boldsymbol{u}} \rho_{vt}^2 \sigma_t^2, \tag{3.2}$$

$$\Sigma\text{-optimality}: \ v_*^{(k)} = \arg\max_{v \in \boldsymbol{u}} \ \frac{(\sum_{t \in \boldsymbol{u}} \mathrm{C}_{vt})^2}{\mathrm{C}_{vv}} = (\sum_{t \in \boldsymbol{u}} \rho_{vt}\sigma_t)^2. \tag{3.3}$$

where the second equalities in both (3.2) and (3.3) come from the observation that

$$\Big((\rho_{vu_1}\sigma_{u_1}), \ldots, (\rho_{vu_{|\boldsymbol{u}|}}\sigma_{u_{|\boldsymbol{u}|}}) \Big)^T = \frac{1}{\sqrt{\mathrm{C}_{vv}}} \cdot \Big(\mathrm{C}_{vu_1}, \cdots, \mathrm{C}_{vu_{|\boldsymbol{u}|}} \Big)^T. \tag{3.4}$$

Remark: We may generalize the two optimalities to a broader class of λ_p-optimalities: [1]

$$\lambda_p\text{-optimality}: \ v_*^{(k)} = \arg\max_{v \in \boldsymbol{u}} \sum_{t \in \boldsymbol{u}} (\rho_{vt}\sigma_t)^p = \arg\max_{v \in \boldsymbol{u}} \sum_{t \in \boldsymbol{u}} \left(\frac{\mathrm{C}_{vt}}{\sqrt{\mathrm{C}_{vv}}} \right)^p \tag{3.5}$$

where V-optimality corresponds to $p = 2$ and Σ-optimality $p = 1$ (up to the same optimizer).

3.2 Theoretical Guarantee for the Greedy Applications

For the general GP model, greedy optimization of the L^2 risk has no guarantee that the solution can be comparable to the brute-force global optimum (taking exponential time to compute), because the objective function, the trace of the predictive covariance matrix, fails to satisfy submodularity in all cases (Krause et al., 2008). However, in the special case of GPs with kernel matrix equal to the inverse of a graph Laplacian (with $\boldsymbol{\ell} \neq \emptyset$ or $\delta > 0$), the GRF does provide such theoretical guarantees, both for V-optimality and Σ-optimality. The latter is a novel result.

The following theoretical results concern greedy maximization of the risk reduction function (which is shown to be submodular): $R_\Delta(\boldsymbol{\ell}) = R(\emptyset) - R(\boldsymbol{\ell})$ for either $R(\cdot) = R_V(\cdot)$ or $R_\Sigma(\cdot)$.

Theorem 1 (Near-optimal guarantee for greedy applications of V/Σ-optimality)**.** *In risk reduction,*

$$R_\Delta(\boldsymbol{\ell}_g) \geq (1 - 1/e) \cdot R_\Delta(\boldsymbol{\ell}_*), \tag{3.6}$$

where $R_\Delta(\ell) = R(\emptyset) - R(\ell)$ for either $R(\cdot) = R_V(\cdot)$ or $R_\Sigma(\cdot)$, e is Euler's number, $\boldsymbol{\ell}_g$ is the greedy optimizer, and $\boldsymbol{\ell}_*$ is the true global optimizer under the constraint $|\boldsymbol{\ell}_*| \leq |\boldsymbol{\ell}_g|$.[2]

[1] The base is never negative as Lemma 9 shows that in any conditional distribution of GRFs, $\rho_{vt} \geq 0, \forall v, t$.

[2] The results (3.7)–(3.6) can be extended to nonuniform node costs. Denote c_v as the node cost of $v \in \boldsymbol{v}$. In this case, a corresponding greedy algorithm maximizes the marginal risk reduction divided by the marginal cost and the constraint in (3.6) becomes $\sum_{v \in \boldsymbol{\ell}_*} c_v \leq \sum_{v \in \boldsymbol{\ell}_g} c_v$

According to Nemhauser et al. (1978), it suffices to show the following properties of $R_\Delta(\boldsymbol{\ell})$:

Lemma 2 (Normalization, Monotonicity, and Submodularity)**.** $\forall \boldsymbol{\ell}_1 \subset \boldsymbol{\ell}_2 \subset \boldsymbol{v}, v \in \boldsymbol{v}$,

$$R_\Delta(\emptyset) = 0, \tag{3.7}$$

$$R_\Delta(\boldsymbol{\ell}_2) \geq R_\Delta(\boldsymbol{\ell}_1), \tag{3.8}$$

$$R_\Delta\big(\boldsymbol{\ell}_1 \cup \{v\}\big) - R_\Delta(\boldsymbol{\ell}_1) \geq R_\Delta\big(\boldsymbol{\ell}_2 \cup \{v\}\big) - R_\Delta(\boldsymbol{\ell}_2). \tag{3.9}$$

3.3 Corollary on GRF Model Class

Another sufficient condition for Theorem 1, which is itself an interesting observation, is the *suppressor-free* condition. Walker (2003) describes a *suppressor* as a variable, knowing which will suddenly suppress a strong correlation between the predictors. An example is $y_i + y_j = y_k$. Knowing any one of these will suppress correlations between the others. Walker further states that suppressors are common in regression problems. Das & Kempe (2008) extend the suppressor-free condition to sets and showed that this condition is sufficient to prove (2.3). Formally, the condition is:

$$\big|\mathrm{corr}(y_i, y_j \mid \boldsymbol{\ell}_1 \cup \boldsymbol{\ell}_2)\big| \leq \big|\mathrm{corr}(y_i, y_j \mid \boldsymbol{\ell}_1)\big| \quad \forall v_i, v_j \in \boldsymbol{v}, \forall \boldsymbol{\ell}_1, \boldsymbol{\ell}_2 \subset \boldsymbol{v}. \tag{3.10}$$

In fact, it may be easier to understand (3.10) as a decreasing correlation property. It is well known for Markov random fields that the labels of two nodes on a graph become independent if conditioned on their Markov blanket. Here we establish that GRF boasts more than that: the correlation between any two nodes decreases as more nodes get labeled, even before a Markov blanket is formed. To summarize, we have:

Theorem 3 (Suppressor-Free Condition)**.** (3.10) *holds for pairs of nodes in the* GRF *model. Note that since the conditional covariance of the* GRF *model is* $L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}$*, we can properly define the corresponding conditional correlation to be*

$$\mathrm{corr}(\boldsymbol{y_u}|\boldsymbol{\ell}) = D^{-\frac{1}{2}} L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})} D^{-\frac{1}{2}}, \text{ with } D = diag\left(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}\right). \tag{3.11}$$

3.4 Summary of Our Method and Other Baseline Methods

All of the active learning strategies to be compared are:[3]

1. The new Σ**-optimality** with greedy sequential updates: $\min_{v'} \left(\mathbf{1}^\top (L_{u^k \setminus \{v'\}})^{-1} \mathbf{1}\right)$.
2. Greedy **V-optimality** (Ji & Han, 2012): $\min_{v'} \mathrm{tr}\left((L_{u^k \setminus \{v'\}})^{-1}\right)$.
3. Greedy **information gain (IG)**, which is the same as determinant-optimality (Krause et al., 2008):[4] $\max_{v'} \left(L^{-1}_{u^k}\right)_{v',v'}$.
4. **Mutual information gain (MIG)** (Krause et al., 2008): $\max_{v'} \left(L^{-1}_{u^k}\right)_{v',v'} / \left((L_{\ell^k \cup \{v'\}})^{-1}\right)_{v',v'}$
5. **Uncertainty sampling (Unc)** picking the largest prediction margin: $\max_{v'} \hat{y}^{(1)}_{v'} - \hat{y}^{(2)}_{v'}$.
6. **Expected error reduction (EER)** (Settles, 2010; Zhu et al., 2003). Selected nodes maximize the average prediction confidence in expectation: $\max_{v'} \mathbb{E}_{y_{v'}} \left[\left(\sum_{u_i \in u} \hat{y}^{(1)}_{u_i} \middle| y_{v'}\right) \middle| y_{\ell^k}\right]$.
7. **Random selection** with 12 repetitions.

We use GRF/BP model with $\delta = 0$ and $\beta = 1$ as our learning model. In such a setting, the connectivity between different nodes on a graph is the strongest and the effect of the outliers is at its minimum. We feel that these parameters generally yields to better baseline results.

4 Data

Comparisons are made on the following real-world network graphs or manifold graph embeddings.

[3] Code available at `http://www.autonlab.org/autonweb/21763`

[4] Using the equivalence, $\left(L^{-1}_{u^k}\right)_{v',v'} = \det(L^{-1}_{u^k}) / \det\left((L_{u^k \setminus \{v'\}})^{-1}\right)$, when L is a generalized graph Laplacian matrix, we have $\arg\min_{v'} \det\left((L_{u^k \setminus \{v'\}})^{-1}\right) = \arg\max_{v'} \left(L^{-1}_{u^k}\right)_{v',v'}$.

1. **DBLP coauthorship network**.[5] The nodes represent scholars and the weighted edges are the number of papers bearing both scholars' names. The largest connected component has 1711 nodes and 2898 edges. The node labels were hand assigned in Ji & Han (2012) to one of the four expertise areas of the scholars: machine learning, data mining, information retrieval, and databases. Each class has around 400 nodes.
2. **Cora citation network**.[6] This is a citation graph of 2708 publications, each of which is classified into one of seven classes: case based, genetic algorithms, neural networks, probabilistic methods, reinforcement learning, rule learning, and theory. The network has 5429 links. We took its largest connected component, with 2485 nodes and 5069 undirected and unweighted edges.
3. **CiteSeer citation network**.[6] This is another citation graph of 3312 publications, each of which is classified into one of six classes: agents, artificial intelligence, databases, information retrieval, machine learning, human computer interaction. The network has 4732 links. We took its largest connected component, with 2109 nodes and 3665 undirected and unweighted edges.
4. **Scikit-learn handwritten digits (digits)**.[7] This is an image classification database published in the scikit-learn software. The database contains 1797 images of hand written digits (0-9) with 8×8 pixel resolution. Every digit class contains roughly 180 images. We created a 7-nearest neighbor **(7-nn)** graph using Euclidean distances of raw features and symmetrized the resulting graph.
5. **Isolated Letter Speech Recognition (ISOLETe / ISOLET4)**.[8] This is a UCI benchmark database of human pronunciations of the 26 English letters. For every letter pronunciation, 617 domain-specific features are created. We used the first 4 mini-batches which contain 120 human subjects **(ISOLET4)**. Further, we also looked at a harder problem that distinguishes letters containing "e" sound (B, C, D, E, G, P, T, V, Z) **(ISOLETe)**. For both problems, we constructed a 4-nearest neighbor **(4-nn)** graph using Euclidean distances of raw features and symmetrized the resulting graph.
6. **Face pose recognition (pose)**.[9] This is a database that regresses semantic information from images. 687 pictures of the same sculpture face were taken with different face poses and lighting conditions. The goal is to reconstruct the face poses (2-dimensional: left-right and up-down). To solve the problem, we constructed a 7-nearest neighbor **(7-nn)** graph using Euclidean distances of the first 240 principal components and symmetrized the resulting graph.

To summarize, our pool of databases aims to cover most of Table 1.

Table 1: Datasets and Experiments Overview

Model Type \ Task	Classification & Survey	Regression
Network graphs	**DBLP, Cora, CiteSeer**	N/A
Manifold graph embeddings of the Euclidean space	**digits, ISOLET4, ISOLETe**	**pose**

4.1 Visualization the Graphs via 2-Dimensional Embedding.

To gain insights of the of the graph databases or graphs generated from feature-represented databases, it is helpful to lay out the graphs on the 2D plane. We use the OpenOrd toolbox (Martin et al., 2011) in the Gephi software[10] for this purpose.

In Figure 4(a-f), it is clear that for classification, different clusters on the graphs, characterized by dense concentrations of nodes, connect to different classes which are shown by node colors. Among them, (a-c) are from network graphs and class boundaries are more unclear. On the contrary, (d-f) are from k-nearest-neighbors graphs for classification problems and classes are more separated. (f) is a noisier version of (e), containing only classes that are more difficult to classify.

[5] http://www.informatik.uni-trier.de/~ley/db/

[6] http://www.cs.umd.edu/projects/linqs/projects/lbc/index.html

[7] http://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html

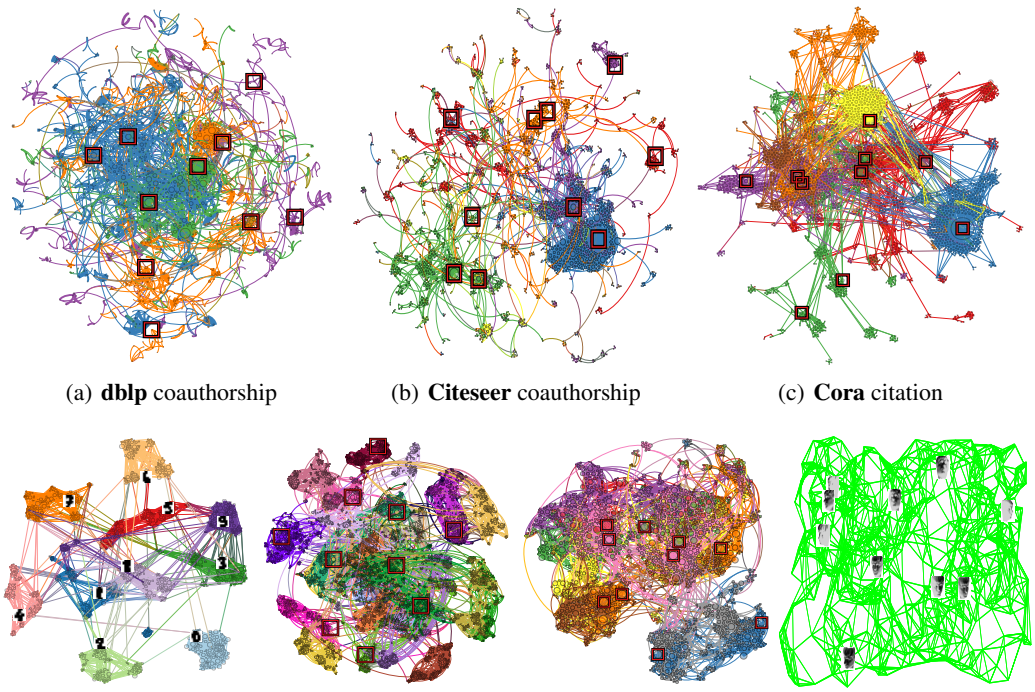
[8] http://archive.ics.uci.edu/ml/datasets/ISOLET

[9] http://isomap.stanford.edu/datasets.html

[10] https://gephi.org/

Figure 4(g) shows nodes more uniformly distributed throughout the embedded space in the pose regression problem. Validity of the GRF model can also be seen as the regression outputs, the yaw and nod of the face pose, vary smoothly along the 6-nn graph generated from Euclidean distance of the first 240 principal components of the face image pixels. For Figure 4, the graph is visualized via Isomap (Tenenbaum et al., 2000) with which the correspondence between the location of a node and its actual pose is more clear.

We also marked the first portion of nodes selected by Σ-optimality active learning criterion to gain insights about its behavior. They are marked with red squares or input snapshots if applicable. To the active learner, the only node labels visible are the marked ones.



(a) **dblp** coauthorship (b) **Citeseer** coauthorship (c) **Cora** citation

(d) **digits** 7-nn graph (e) **isolet4** 4-nn graph (f) **isolete** 4-nn graph (g) **pose** 6-nn graph

Figure 4: Visualization of the graphs we use via OpenOrd (a-f) or Isomap (g). Node colors in (a-f) indicate classes. (a-c) are network graphs. (d-f) are k-nn graphs for classification. (g) is a k-nn graph for regression. Squares or input snapshots mark the nodes queried decided by our Σ-optimality, starting from not knowing any class labels.

5 Analysis

At a first step, we analyzed our method and other baseline methods conceptually or under simple cases. The goal is to gain intuitions about the behavior of every method and to find out the types of problems that every method is more suitable for.

5.1 Insights From Comparing the Greedy Applications of Σ- and V-Optimality Criteria

First, we compare V- and Σ-optimality because the former is a recent successful method and the latter is a variant of the former. Both V/Σ-optimality criteria are approximations to the 0/1 risk minimization objective. Unfortunately, we cannot theoretically reason why Σ-optimality outperforms V-optimality in the experiments. However, we made two observations during our investigation that provide some insights.

Observation 4. *Eq.* (3.2) *and* (3.3) *suggest that both the greedy Σ/V-optimality selects nodes that (1) have high variance and (2) are highly correlated to high-variance nodes, conditioned on the labeled nodes.*

In order to contrast the Σ/V-optimality, rewrite (3.3) as:

$$(\Sigma\text{-optimality}): \ \arg\max_{v \in \boldsymbol{u}} \Big(\sum_{t \in \boldsymbol{u}} \rho_{vt}\sigma_t\Big)^2 = \sum_{t \in \boldsymbol{u}} \rho_{vt}^2 \sigma_t^2 + \sum_{t_1 \neq t_2 \in \boldsymbol{u}} \rho_{vt_1}\rho_{vt_2}\sigma_{t_1}\sigma_{t_2}. \tag{5.1}$$

Observation 5. *The Σ-optimality has one more term that involves cross products of $(\rho_{vt_1}\sigma_{t_1})$ and $(\rho_{vt_2}\sigma_{t_2})$ (which are nonnegative according to Lemma 9). By Cauchy–Schwartz Inequality, the sum of these cross products are maximized when they equal. So, the Σ-optimality additionally favors nodes that (3) have consistent global influence, i.e., that are more likely to be in cluster centers.*

To visualize the intuitions described above, Figure 5 shows the first few nodes selected by different optimality criteria. This graph is constructed by a breadth-first search from a random node in a larger **DBLP** coauthorship network graph that we will introduce in the next section. On this toy graph, both criteria pick the same center node to query first. However, for the second and third queries, V-optimality weighs the uncertainty of the candidate node more, choosing outliers, whereas Σ-optimality favors nodes with universal influence over the graph and goes to cluster centers.

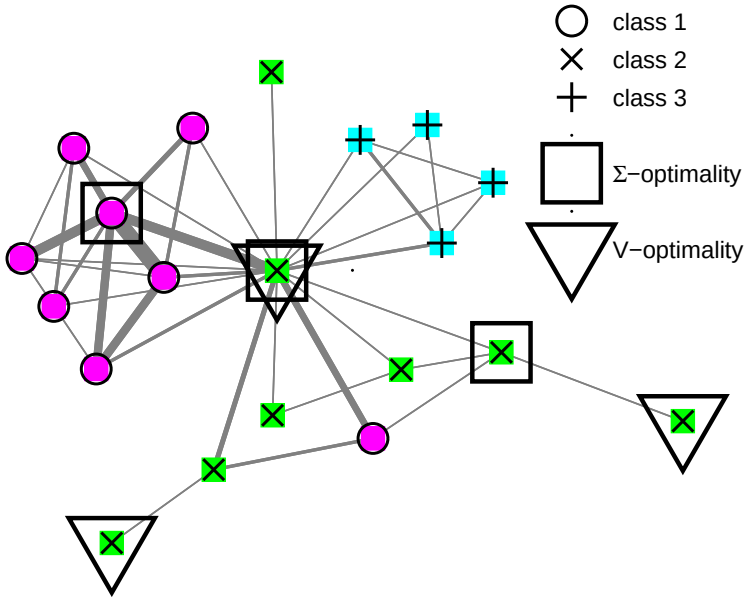


Figure 5: Toy graph demonstrating the behavior of Σ-optimality vs. V-optimality.

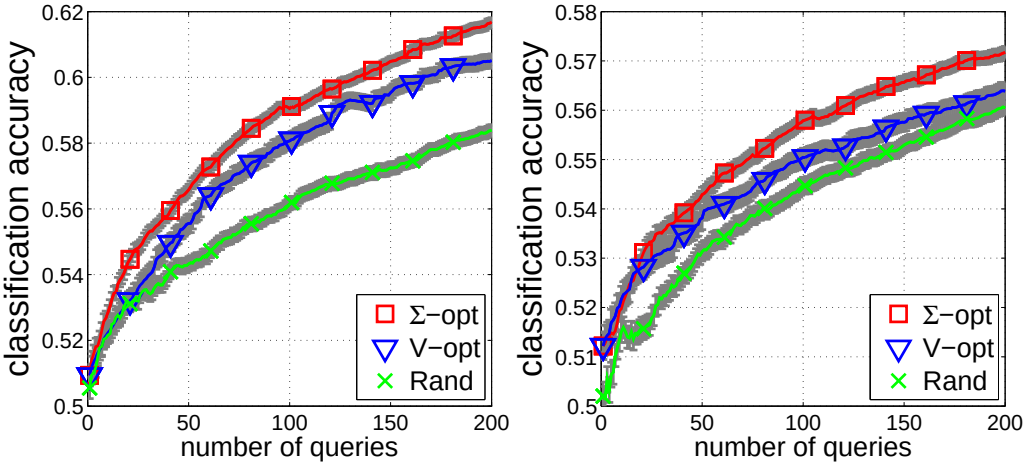
5.2 Simulating the Node Labels on a Graph

To further investigate the behavior of Σ- and V-optimality, we conducted experiments on synthetic labels generated on real-world network graphs. The node labels were first simulated using the model in order to compare the active learning criteria directly without raising questions of model fit. We carry out tests on the same graphs with real data in the next section.

For active learning, Σ-optimality outperforms V-optimality on various graphs.

We simulated the binary labels with the GRF-sigmoid model and performed active learning with the GRF/LP model for predictions. The parameters in the generation phase were $\beta = 0.01$ and $\delta = 0.05$, which maximizes the average classification accuracy increases from 50 random training nodes to 200 random training nodes using the GRF/LP model for predictions. Figure **??** shows the binary classification accuracy versus the number of queries on both the DBLP coauthorship graph and the CORA citation graph that we will describe below. The best possible classification results are indicated by the leave-one-out (LOO) accuracies given under each plot.

Figure 6 can be a surprise due to the reasoning behind the L^2 surrogate loss, especially when the predictive means are trapped between $[-1, 1]$, but we see here that our reasoning in Section 5.1 can lead to the survey loss actually making a better active learning objective.



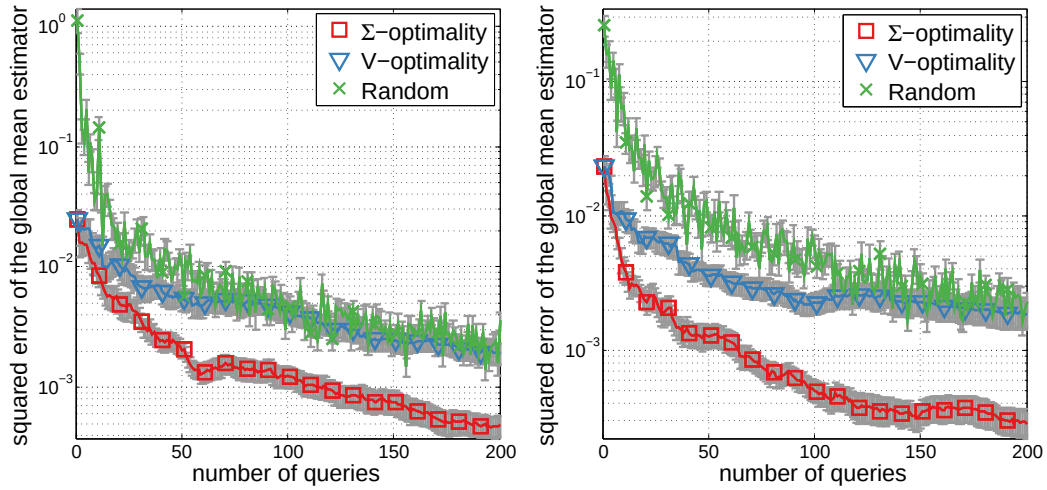
(a) DBLP coauthorship graph. 68.3% LOO accuracy. (b) Cora citation graph. 60.8% LOO accuracy

Figure 6: Simulating binary labels by the GRF-Sigmoid; learning with the GRF/BP, 250 repetitions.

We have also performed preliminary experiments with different values of β and δ. Despite that larger β and smaller δ increase label independence on the graph structure and undermine the effectiveness of both V/Σ-optimality heuristics, we have seen that whenever the V-optimality establishes a superiority over random selections, the Σ-optimality yields better performances. Particularly, a larger δ increases the performance gap between Σ- and V-optimality, because it decreases the influence from one node on the graph to another, yielding more significant outliers that hinder the performance of V-optimality more seriously than Σ-optimality.

For active surveying, Σ-optimality also outperforms V-optimality.

The active surveying problem strives to determine the mean of the (continuous) node labels. When the objective was set to active surveying, Σ-optimality also outperformed V-optimality (Figure 7). Here, for simplicity, we set $\delta = 0$ and $\beta = 1$. Notice that the random selection is actually a very competent baseline as the squared standard error of the mean decreases at the rate $\mathcal{O}(k^2)$, where k is the number of random query points.

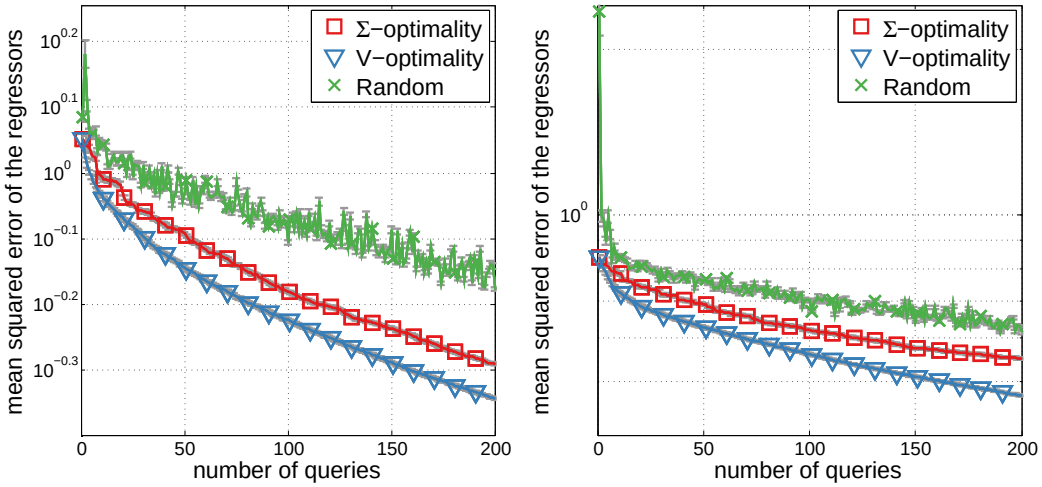


(a) The DBLP dataset (b) The CORA dataset

Figure 7: Active surveying risk comparison. Lower is better. 24 repetitions.

Fail case: when the goal is set to minimize the regression mean-squared-error.

Apart from classification and surveying problems, another broad active learning application is the regression problem. The superiority of Σ-optimality over V-optimality predicates on the fact that the L^2 surrogate loss does not reveal the true binary/survey risk in these previous objectives. Yet, for the regression problem, V-optimality directly minimizes its expected mean-squared-error (MSS).



(a) The DBLP dataset (b) The CORA dataset

Figure 8: Average regression risk comparison. Lower is better. 100 repetitions.

Figure 8 show simulation results in regression settings. Here, we simulate 100 independent draws of GRF/BP model with $\delta = 0$ and $\beta = 1$. The evaluation is the empirical MSE of the predictors among all node labels. Complying with our intuition, both V-optimality and Σ-optimality win over random selection, yet V-optimality reduces the MSE even more efficiently than Σ-optimality.

6 Results

6.1 Network Graphs

Classification. For active classification, Figure 9 shows the prediction accuracy of the unlabeled nodes using only the labels from the nodes that each active learning queries, except for the first common seed node which was assigned at random. Every curve shows the mean and its standard error after 12 runs.

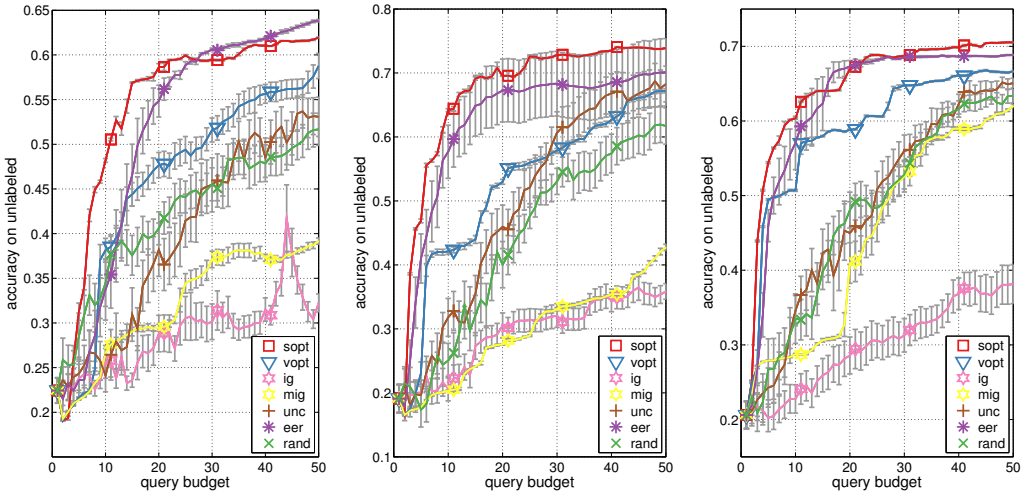
On all three datasets, Σ-optimality outperforms other methods by a large margin especially during the first five to ten queries. The runner-up, EER, catches up to Σ-optimality in some cases, but (1) it is an order slower to evaluate, (2) it requires query results immediately before the next query, whereas both V-optimality and Σ-optimality do not, and (3) it does not have theoretical guarantees.

The win of Σ-optimality over V-optimality has been intuitively explained in Section 5.1 as Σ-optimality having better exploration ability and robustness against outliers. That all three active learning algorithms win over random selection validates the effectiveness of the GRF model which assumes node labels cluster according to graph clusters.

We also noticed that IG, MIG, and Unc methods do not perform significantly better than random. This is because these heuristics tend to query mostly outliers on the graph.

Surveying. We also performed real-world experiments on the root-mean-square-error (RMSE) of the class proportion estimations, which is the survey risk that the Σ-optimality minimizes. The Σ-optimality beats the V-optimality (Figure 10).

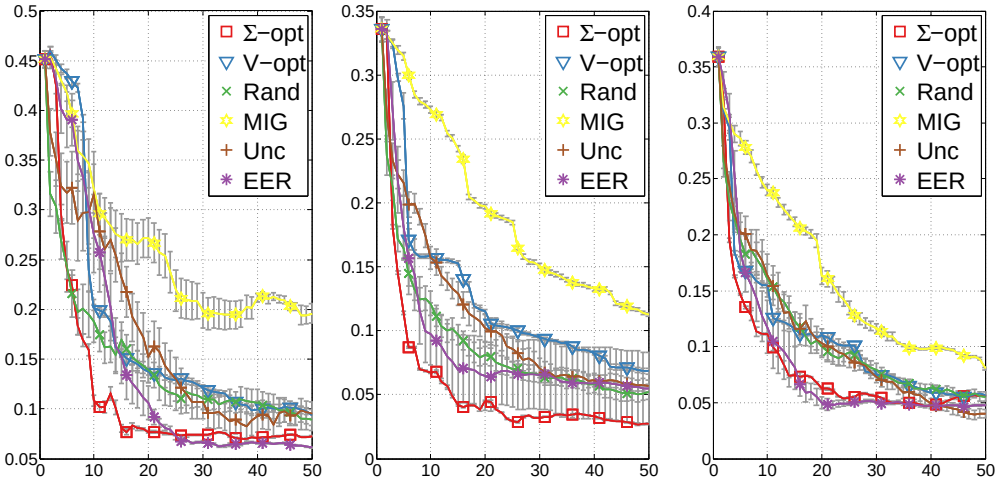
With the survey experiments, the objective is $\|\hat{\mathbb{E}}\hat{\boldsymbol{y}} - \boldsymbol{\pi}\|_2/\sqrt{C}$ on unlabeled set $\boldsymbol{u}$, where $\hat{\boldsymbol{y}}$ is the vector of prediction means in different one-vs-alls, C is the number of classes and $\boldsymbol{\pi}$ is the C-



(a) DBLP. 87% LOO accuracy. (b) CORA. 90% LOO accuracy. (c) CITESEER 80% LOO accuracy.

Figure 9: Classification accuracy vs the number of queries. $\beta = 1, \delta = 0$. Randomized first query.

dimensional true class distribution of unlabeled nodes. Every curve shows the mean and its standard error after 12 random initializations.



(a) **DBLP** coauthorship, 4 classes. (b) **Cora** citation, 7 classes. (c) **CiteSeer** citation, 6 classes.

Figure 10: Survey RMSE, $\|\hat{\mathbb{E}}\hat{\boldsymbol{y}} - \boldsymbol{\pi}\|_2/\sqrt{C}$, on unlabeled set $\boldsymbol{u}$. Model is GRF/BP with $\delta = 0$.

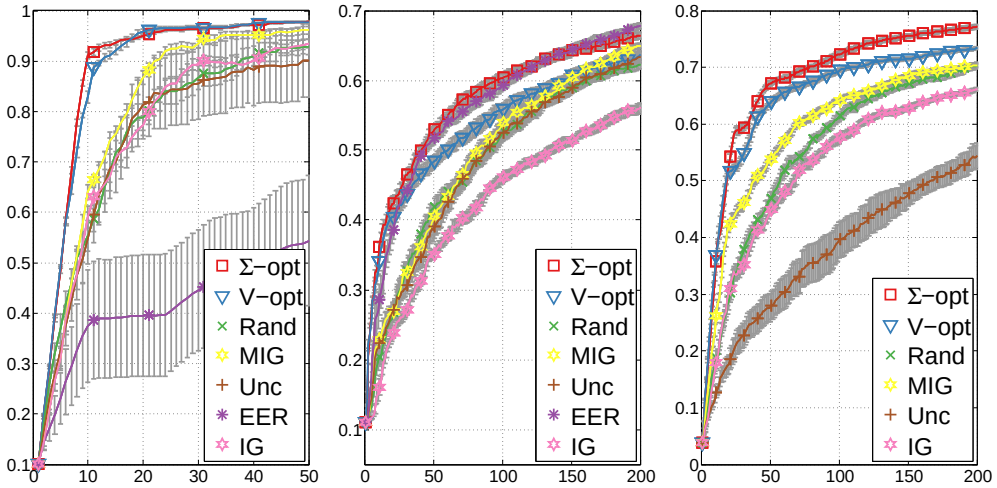
6.2 Manifold Graph Embeddings of the Euclidean Space

Detailed data preprocessing. To embed the Euclidean features from the databases **digits**, **ISOLETe**, **ISOLET4**, and **pose** in graphs, we used k-nearest neighbor graphs using the Euclidean distance. In **digits**, we created a 7-nearest neighbor graph based on the Euclidean distance of raw features, i.e. the concatenation of 64 image pixel gray values. The graph was further symmetrized by removing the direction information (and also doubling the edge weight if an edge was originally bi-directional). The resulting graph contain 1797 nodes and 8727 edges. Visual inspection shows that the resulting graph fits the labels well.

In both **ISOLETe** and **ISOLET4**, we found the 4-nearest neighbor graph based also on Euclidean distances of raw features, which is the 617 dimensional domain-specific features. The graphs were further symmetrized in the same manner. The resulting graph for **ISOLETe** contains 2160 nodes and 6337 edges and for **ISOLET4** 6238 nodes and 18662 edges. Visual inspection shows that the

resulting graphs are moderately difficult: while some classes are separated from other classes by sparse cuts, about half of the nodes are close to nodes of other classes in graph distances.

Classification results.



(a) **digits**. 99% LOO accuracy. (b) **ISOLETe**. 83% LOO accuracy. (c) **ISOLET4**. 91% LOO accuracy.

Figure 11: Classification accuracy vs the number of queries. Model is GRF/BP with $\delta = 0$.

Figure 11 shows the prediction accuracy of the unlabeled nodes using only the labels from the nodes that each active learning queries, except for the first common seed node which was assigned at random. Every curve shows the mean and its standard error after 12 runs.

On all three manifold graph embeddings of the Euclidean space, Σ-optimality again outperforms other methods by a large margin, while all baseline methods yield to acceptable classification accuracies. We reason that this result follows the spectral and cut similarity between manifold graph embeddings and the network graphs in previous experiments. Specifically, we observed that in the 2D layouts of these manifold graphs, graph clusters have purer labels and there are also smaller and less important clusters that distract the heuristics.

Regression. Finally, we performed a graph regression experiment on the **pose** database. To create a manifold graph embedding, we used the 7-nearest neighbor graph based on the 240 principal components of face images that come with the database we downloaded. Then we symmetrized the resulting graph. There are 698 nodes and 2562 edges on this graph. The validity of this graph is checked as we recover a 2-dimensional (2D) Euclidean space layout of our graph similar to the Isomap method (Tenenbaum et al., 2000). The relative positions of the recovered 2D coordinates agree with the relative yaws and pitches of the original face poses.

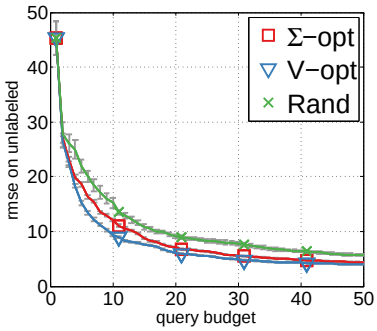


Figure 12: Regression RMSE vs the number of queries on the **pose** 7-nn graph. Lower is better.

Figure 12 show the RMSE of the 2D pose predictors of all unlabeled nodes based on the 2D pose labels queried by various active learning heuristics. The curves are averaged after 12 runs from

different randomly sampled starting nodes. The error bars show the standard error of the mean. V-optimality outperforms Σ-optimality and both outperformed random selection. The result is similar to what we have seen in the simulation. An explanation is that for active regression problems, V-optimality directly minimizes the corresponding risk and thus is the best-performing heuristic.

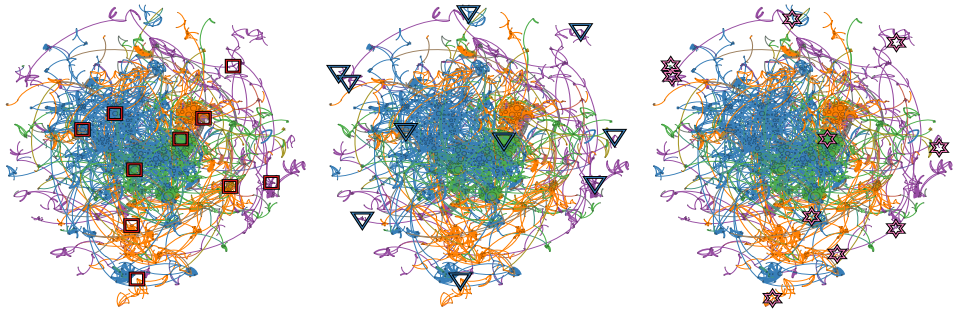
7 Discussions

For classification and surveying experiments, Σ-optimality reasonably outperformed all its competitors. The reason is explained below.

Notice that the randomness in node classifications on real network graphs is limited, because it is hard to subsample a graph when we are unclear what properties we should keep. As a result, the only randomness comes from the initiation of the first query node, which does not affect the behavior of most algorithms because they generally simply ignore the first query node and re-selects their own favorite nodes to start. The blips in many curves are for this reason. In k-nn graphs, randomness of the graph can be created by subsampling 70% of the examples. Thus, the curves appear much smoother.

7.1 Visualization of Node Selections

To gain insights of the empirical behavior of different active learning criteria, we visually inspected the choices of various criteria in the OpenOrd embeddings of the DBLP database.



(a) Σ-optimality (b) V-optimality (c) IG

(d) MIG (e) Unc (f) EER

Figure 13: **DBLP** coauthorship graph. First 10 queries decided in a greedy sequential manner.

We used the OpenOrd toolbox (Martin et al., 2011) in the Gephi software to lay out the graph in 2D. The node colors indicate the classes which the nodes belong to. They are not visible to active learners until the nodes have been queried. The query decisions were made in a greedy sequential manner. In these visualizations, the first query for all algorithms were fixed at the node of the largest degree.

In Figure 13, the central part corresponds to the majority of the nodes, which have strong connections with each other. The periphery contains many outlying nodes which provide little information to

classifications of the central nodes because their connections are weak. A desirable set of queries should explore denser regions in the central part, which correspond to clusters of reasonable size and interesting nodes. However, the presence of many sparse cuts on the periphery hinders the performance of many other active learning strategies. It is clear that Σ-optimality selected a set of very reasonable queries, because it exploits the cluster structure in the central part.

Figure 13(b) shows that V-optimality went after small clusters if not outliers in the periphery. It may be because that the 2-norm score used for greedy update in (3.2) gives a high score to a node which has fewer but stronger connections to very uncertain nodes. In contrast, the square of 1-norm used in (3.3) favors the number of links more than the quality of links and the uncertainty of the nodes that they connect to.

Figure 13(c) shows the information gain criterion, which is equivalent to a criterion that selects nodes with the highest variance. These nodes correspond to the edge of a graph and are not very helpful for predicting other node labels.

Figure 13(d) is an improvement over (c). The resulting sample set is similar to (b). It avoid querying the very edge of the graph, but still query the centers of very small yet sparse cuts at the edge of the graph. It is not as good as (a).

Figure 13(e) shows the decision of uncertainty sampling which is based solely on predictive means. An interesting observation in (e) is that all following query decisions are all made along the path from the first query point to the second query point. The reason is that in label propagation, only nodes that connect to more than one labeled nodes can possibly have a predictive mean between 0 and 1, because of the effect of Markove blankets. Generally, radial strings of nodes are common in network graphs and thus uncertainty sampling fail to explore the graph in most cases.

Figure 13(f) shows the decision of expected error reduction which is another criterion based solely on predictive means. EER fails because of its bias to avoid risks. In the very beginning, EER sees only one class and infers that the predictive means of all other labels have value 1 for the class EER has seen and 0 for any other class. Thus, any new query may cause surprises and lower the expected accuracy after it is made. In order to minimize risk, EER only queries nodes that have small variance, i.e. being very close to other queried nodes. The fact that EER remains low accuracy after many samples in k-nn graphs can be explained by the above observation. The result of EER in network graphs is much improved because in these graphs, the node labels contain much noise which helps resolve the bias in EER quickly. However, even so, the behavior of EER is unpredictable, because it is unclear which type of bias is resolved/accumulated.

A final remark is that for active regression, a set of query point at the periphery of a graph may be desirable, because they reolves regions with the highest variance. Examples have been shown in simulations (Figure 8 and with **pose** dataset (Figure 12). As for binary classification, the GRF makes its biggest relaxation error at the periphery, where the actual variance is always bounded. Thus, risk minimization methods that directly applies to the GRF model generally fails.

8 Limitations

The randomness of our experiments with network graphs is limited, because it is hard to subsample a network graph while arguably maintain its desirable properties. There is neither a convincingly principled way to simulate node labels because GRF is a relaxation of the actual binary model, which is computationally infeasible to deal with. We tried to combat this limitation with multiple attempts to increase randomness and using many different datasets of the same nature.

Another limitation comes from the fact that Σ-optimality does not include predictive means in its criterion. This limitation is alleviated if the relevancy of the graph structure and the node labels is high. However, it may raise concerns if the noise-level of node labels are not consistent with the model, in which case it could happen that one cluster requires more queries because the actual connections in that cluster are weaker than modeled by GRF, if not null. Nonetheless, while including predictive means in the criterion, e.g. in an EER fashion, increases robustness to GRF modeling error, there will also be more parameters to tune and the behavior of the EER part in the criterion may be hard to analyze.

9 Conclusion & Future Work

For active learning on GRFs, it is common to use variance minimization criteria with greedy one-step lookahead heuristics. V-optimality and Σ-optimality are two criteria based on statistics of the predictive covariance matrix. They both are also risk minimization criteria. V-optimality minimizes the L^2 risk (2.3) whereas Σ-optimality minimizes the survey risk (2.5).

Active learning with both criteria can be seen as subset optimization problems (2.4), (2.6). Both objective functions on the risk reduction are submodular set functions. Therefore, greedy one-step lookahead applications of these criteria can achieve a $(1 - 1/e)$ global optimality ratio for risk reduction. Moreover, GRFs may serve as a tangible example to the otherwise abstract suppressor-free condition.

While the V-optimality on GRFs inherits from label propagation (and random walk with absorptions) and have good empirical performance, it is not directly minimizing the 0/1 classification risk. We found that the Σ-optimality performs even better. The intuition is described in section 5.1. Our claim was also backed by extensive experiments on both synthetic data and real-world data.

It is unclear whether there exist more fundamental reasons that explain what types of graphs the greedy Σ- and V-optimality work better. Neither do we know the graph-theoretic motivation behind Σ-optimality. Future work aims to answer these questions.

Acknowledgements

I want to thank my collaborator Roman Garnett and my committee members, Jeff Schneider, Barnabas Poczos, and Roy Maxion, for their help and advice. I also received comments from Geoff Gordon, Peter Huggins, and Arthur Dubrawski.

I much appreciate supports from Ying Yang, my parents, and many peer students, particularly: Liang Xiong, Min Xu, Tzu-Kuo Huang, Yi Zhang, Madalina Fiterau, Haijie Gu, and Avinava Dubey. MLD program manager Diane Stidle is very considerate as well.

This work is funded in part by NSF grant IIS0911032 and DARPA grant FA87501220324.

References

Das, Abhimanyu and Kempe, David. Algorithms for subset selection in linear regression. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pp. 45–54. ACM, 2008.

Friedland, S and Gaubert, S. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 2011.

Garnett, Roman, Krishnamurthy, Yamuna, Xiong, Xuehan, Schneider, Jeff, and Mann, Richard. Bayesian optimal active search and surveying. In *ICML*, 2012.

Ji, Ming and Han, Jiawei. A variance minimization criterion to active learning on graphs. In *AISTAT*, 2012.

Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research (JMLR)*, 9:235–284, February 2008.

Martin, Shawn, Brown, W Michael, Klavans, Richard, and Boyack, Kevin W. Openord: an open-source toolbox for large graph layout. In *IS&T/SPIE Electronic Imaging*, pp. 786806–786806. International Society for Optics and Photonics, 2011.

Nemhauser, George L, Wolsey, Laurence A, and Fisher, Marshall L. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

Rasmussen, Carl Edward and Williams, Christopher KI. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

Settles, Burr. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

Tenenbaum, Joshua B, De Silva, Vin, and Langford, John C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Walker, David A. Suppressor variable (s) importance within a regression model: an example of salary compression from career services. *Journal of College Student Development*, 44(1):127–133, 2003.

Wu, Xiao-Ming, Li, Zhenguo, So, Anthony Man-Cho, Wright, John, and Chang, Shih-Fu. Learning with partially absorbing random walks. In *Advances in Neural Information Processing Systems 25*, pp. 3086–3094, 2012.

Zhu, Xiaojin and Ghahramani, Zoubin. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

Zhu, Xiaojin, Lafferty, John, and Ghahramani, Zoubin. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pp. 58–65, 2003.

Appendix

A Proofs

Our results predicate on and extend to GPs whose inverse covariance matrix meets Proposition 6.

Proposition 6. *L satisfies the following.* [11]

#	*Textual description*	*Mathematical expression*
p6.1	*L has proper signs.*	$l_{ij} \geq 0$ *if* $i = j$ *and* $l_{ij} \leq 0$ *if* $i \neq j$.
p6.2	*L is undirected and connected.*	$l_{ij} = l_{ji} \forall i, j$ *and* $\sum_{j \neq i}(-l_{ij}) > 0$.
p6.3	*Node degree no less than number of edges.*	$l_{ii} \geq \sum_{j \neq i}(-l_{ij}) = \sum_{j \neq i}(-l_{ji}) > 0, \forall i$.
p6.4	*L is nonsingular and positive-definite.*	$\exists i: \; l_{ii} > \sum_{j \neq i}(-l_{ij}) = \sum_{j \neq i}(-l_{ji}) > 0$.

□

Although the properties of V-optimality fall into the more general class of *spectral functions* (Friedland & Gaubert (2011)), we have seen no proof of either the suppressor-free condition or the submodularity of Σ-optimality on GRFs.

Lemma 7. *For any L satisfying (p6.1-4), $L^{-1} \geq 0$ entry-wise.*[12]

Proof. Suppose $L = D - W = D(I - D^{-1}W)$, with $D = \text{diag}\,(L)$. According to (p6.1), $D \geq 0$, $W \geq 0$ and $D^{-1}W \geq 0$. Furthermore, by (p6.3),

$$0 \leq D^{-1}W \leq \big(\frac{w_{ij}}{\sum_k w_{ik}}\big)_{i,j=1}^{N}, \tag{.1}$$

and so the matrix norm $\|D^{-1}W\|_\infty \leq 1$. Thus, any eigenvalue λ_k and its corresponding eigenvector $\boldsymbol{v}_k$ of $D^{-1}W$ needs to satisfy $|\lambda_k| \|\boldsymbol{v}_k\|_\infty = \|D^{-1}W\boldsymbol{v}_k\|_\infty \leq \|\boldsymbol{v}_k\|_\infty$, i.e. $|\lambda_k| \leq 1, \forall k = 1, ..., N$.

When L is nonsingular, $(I - D^{-1}W)$ is invertible, i.e., has no zero eigenvalue. Hence, $|\lambda_k| < 1, \forall k = 1, ..., N$ and $\lim_{n\to\infty}(D^{-1}W)^n = 0$. The latter yields the convergence of Taylor expansion,

$$L^{-1} = [I + \textstyle\sum_{r=1}^{\infty}(D^{-1}W)^r]D^{-1}. \tag{.2}$$

It suffices to observe that every term on the right hand side (RHS) is nonnegative. □

Corollary 8. *The* GRF *prediction operator $L_{\boldsymbol{u}}^{-1}L_{ul}$ maps $\boldsymbol{y}_{\boldsymbol{\ell}} \in [0,1]^{|\boldsymbol{\ell}|}$ to $\hat{\boldsymbol{y}}_{\boldsymbol{u}} = -L_{\boldsymbol{u}}^{-1}L_{ul}\boldsymbol{y}_{\boldsymbol{\ell}} \in [0,1]^{|\boldsymbol{u}|}$. When L is singular, the mapping is onto.*

Proof. For $\boldsymbol{y}_{\boldsymbol{\ell}} = \mathbf{1}$, $(L_{\boldsymbol{u}}, L_{ul}) \cdot \mathbf{1} \geq 0$ and $L_{\boldsymbol{u}}^{-1} \geq 0$ imply $\big(I, L_{\boldsymbol{u}}^{-1}L_{ul}\big) \cdot \mathbf{1} \geq 0$, i.e. $\mathbf{1} \geq -L_{\boldsymbol{u}}^{-1}L_{ul}\mathbf{1} = \hat{\boldsymbol{y}}_{\boldsymbol{u}}$.

As both $L_{\boldsymbol{u}} \geq 0$ and $-L_{ul} \geq 0$, we have $\boldsymbol{y}_{\boldsymbol{\ell}} \geq 0 \Rightarrow \hat{\boldsymbol{y}}_{\boldsymbol{u}} \geq 0$ and $\boldsymbol{y}_{\boldsymbol{\ell}} \geq \boldsymbol{y}'_{\boldsymbol{\ell}} \Rightarrow \hat{\boldsymbol{y}}_{\boldsymbol{u}} \geq \hat{\boldsymbol{y}}'_{\boldsymbol{u}}$. □

Lemma 9. *Suppose $L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}$, then $L^{-1} - \begin{pmatrix} L_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \geq 0$ and is positive-semidefinite.*

Proof. When L is nonsingular, by the block matrix inversion theorem,

$$L^{-1} - \begin{pmatrix} L_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} L_{11}^{-1}(-L_{12}) \\ I \end{pmatrix} (L_{22} - L_{21}L_{11}^{-1}L_{12})^{-1} \big((-L_{21})L_{11}^{-1}, I\big) \tag{.3}$$

By assumption (p6.4), L^{-1} is positive-definite, so is its lower right principal submatrix $(L_{22} - L_{21}L_{11}^{-1}L_{12})^{-1}$. Thus, $L^{-1} - \begin{pmatrix} L_{11} & 0 \\ 0 & 0 \end{pmatrix}$ is positive-semidefinite.

By Lemma 7, $L^{-1} \geq 0$ and this implies that its lower right $(L_{22} - L_{21}L_{11}^{-1}L_{12})^{-1} \geq 0$. The submatrix L_{11} also satisfies (p6.1-4) and by Lemma 1, $L_{11}^{-1} \geq 0$. By the sign rule (p6.1), $(-L_{12}) =$

[11] Property p6.4 holds after the first query is done or when the regularizer $\delta > 0$ in (2.1).

[12] In the following, for any vector or matrix A, $A \geq 0$ always stands for A being (entry-wise) nonnegative.

$(-L_{21})^T \geq 0$. Now that every term on the right side of (.3) is nonnegative, the left side also has to be so. □

As a corollary, the **monotonicity in** (3.8) for both $R(\cdot) = R_V(\cdot)$ or $R_\Sigma(\cdot)$ can be shown. □

Both proofs for **submodularity in** (3.9) and **Theorem 3** result from more careful execution of matrix inversions. We first state the key property in these executions of matrix inversions and then prove both results.

Proposition 10. *Without loss of generality, let* $\boldsymbol{u} = \boldsymbol{v} - \boldsymbol{\ell} = \{1, \ldots, k\}$ *and* $v = v_k$. *Partition the matrix:*

$$L_{(\boldsymbol{v}-\boldsymbol{\ell})} = \begin{pmatrix} L_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\{v\})} & L_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\{v\}),\{v\}} \\ L_{\{v\},(\boldsymbol{v}-\boldsymbol{\ell}\cup\{v\})} & L_{\{v\}} \end{pmatrix} := \begin{pmatrix} A & b \\ b^T & c \end{pmatrix} \tag{.4}$$

By the block matrix inversion theorem,

$$\begin{pmatrix} C & d \\ d^T & e \end{pmatrix} := \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \frac{A^{-1}bb^TA^{-1}}{c-b^TA^{-1}b} & \frac{-A^{-1}b}{c-b^TA^{-1}b} \\ \frac{-b^TA^{-1}}{c-b^TA^{-1}b} & \frac{1}{c-b^TA^{-1}b} \end{pmatrix}. \tag{.5}$$

□

Proof. **submodularity in** (3.9) **for** $R_\Delta(\cdot)$**.** Adopting the notations in Proposition 10,

$$L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})} - L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}-\{v\})} = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}^{-1} - \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -A^{-1}b \\ 1 \end{pmatrix} \frac{1}{c - b^TA^{-1}b} \left(-b^TA^{-1}, 1\right) \tag{.6}$$

For V-optimality,

$$R_\Delta(\boldsymbol{\ell} \cup \{v\}) - R_\Delta(\boldsymbol{\ell}) = \operatorname{tr}\left(-L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}-\{v\})} + L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}\right) = \frac{((-b^T)A^{-1})(A^{-1}(-b)) + 1}{c - (-b)^TA^{-1}(-b)}.$$

As every term on the RHS has been written as nonnegative entry-wise, by taking submatrices/vectors of consistent rows/columns of A and $-b$, the values of $(-b^T)A^{-1}$ and $(-b^T)A^{-1}(-b)$ decrease.

Notice that both A and b correspond to $(\boldsymbol{v} - \boldsymbol{\ell} \cup \{v\})$. Thus, as $\boldsymbol{\ell}$ grows, A and b shrink in size, $R_\Delta(\boldsymbol{\ell} \cup \{v\}) - R_\Delta(\boldsymbol{\ell})$ diminishes.

For Σ-optimality,

$$R_\Delta(\boldsymbol{\ell} \cup \{v\}) - R_\Delta(\boldsymbol{\ell}) = \mathbf{1}^T \cdot \left(-L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}-\{v\})} + L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})}\right) \cdot \mathbf{1} = \frac{((-b^T) \cdot A^{-1} \cdot \mathbf{1})^2}{c - (-b)^TA^{-1}(-b)}.$$

Similar arguments hold. □

Proof. **Theorem 3.** Adopt the notations in Proposition 10. Dividing the vector d by the diagonal number e yields $\forall i \neq k$:

$$\frac{\operatorname{Cov}(y_i, y_k|\boldsymbol{\ell})}{\operatorname{Var}(y_k|\boldsymbol{\ell})} = \frac{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}_1)})_{ik}}{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}_1)})_{kk}} = \frac{1}{e} \cdot d_i = \frac{(-A^{-1}b)_i}{c - b^TA^{-1}b} \Big/ \frac{1}{c - b^TA^{-1}b} = (A^{-1}(-b))_i. \tag{.7}$$

That $-b \geq 0$ and $A^{-1} \geq 0$ leads to $A^{-1}(-b)^T \geq \tilde{A}^{-1}(-\tilde{b}) \geq 0$ if $\tilde{A}$ and $\tilde{b}$ are subsets of consistent columns/rows (Lemma 9), i.e.,

$$\frac{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})})_{ik}}{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})})_{kk}} \geq \frac{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\boldsymbol{\ell}_2)})_{ik}}{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\boldsymbol{\ell}_2)})_{kk}} \geq 0 \quad \forall i \neq k \notin \boldsymbol{\ell} \cup \boldsymbol{\ell}_2. \tag{.8}$$

Similarly, reordering the indices, $\dfrac{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})})_{ik}}{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell})})_{ii}} \geq \dfrac{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\boldsymbol{\ell}_2)})_{ik}}{(L^{-1}_{(\boldsymbol{v}-\boldsymbol{\ell}\cup\boldsymbol{\ell}_2)})_{ii}} \geq 0$. It suffices to multiply both sides of the above. □