Selective Sampling for Online Best-arm Identification

Romain Camilleri*, **Zhihan Xiong***, **Maryam Fazel, Lalit Jain, Kevin Jamieson**
University of Washington, Seattle, WA
{camilr,zhihanx,mfazel,lalitj,jamieson}@uw.edu

Abstract

This work considers the problem of *selective-sampling for best-arm identification*. Given a set of potential options $\mathcal{Z} \subset \mathbb{R}^d$, a learner aims to compute with probability greater than $1-\delta$, $\arg\max_{z\in\mathcal{Z}} z^\top \theta_*$ where θ_* is unknown. At each time step, a potential measurement $x_t \in \mathcal{X} \subset \mathbb{R}^d$ is drawn IID and the learner can either choose to take the measurement, in which case they observe a noisy measurement of $x^\top \theta_*$, or to abstain from taking the measurement and wait for a potentially more informative point to arrive in the stream. Hence the learner faces a fundamental trade-off between the number of labeled samples they take and when they have collected enough evidence to declare the best arm and stop sampling. The main results of this work precisely characterize this trade-off between labeled samples and stopping time and provide an algorithm that nearly-optimally achieves the minimal label complexity given a desired stopping time. In addition, we show that the optimal decision rule has a simple geometric form based on deciding whether a point is in an ellipse or not. Finally, our framework is general enough to capture binary classification improving upon previous works.

1 Introduction

In this work we consider *selective sampling for online best-arm identification*. In this setting, at every time step $t = 1, 2, \ldots,$ Nature reveals a potential measurement $x_t \in \mathcal{X} \subset \mathbb{R}^d$ to the learner. The learner can choose to either *query* x_t ($\xi_t = 1$) or abstain ($\xi_t = 0$) and immediately move on to the next time. If the learner chooses to take a query ($\xi_t = 1$), then Nature reveals a noisy linear measurement of an unknown $\theta_* \in \mathbb{R}^d$, i.e. $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ where ϵ_t is mean zero sub-Gaussian noise. Before the start of the game, the learner has knowledge of a set $\mathcal{Z} \subset \mathbb{R}^d$. The objective of the learner is to identify $z_* := \arg\max_{z\in\mathcal{Z}} \langle z, \theta_* \rangle$ with probability at least $1-\delta$ at a learner specified stopping time $\mathcal{U}$. It is desirable to minimize both the stopping time $\mathcal{U}$ which counts the total number of unlabeled or labeled queries and the number of labeled queries requested $\mathcal{L} := \sum_{t=1}^{\mathcal{U}} \mathbf{1}\{\xi_t = 1\}$. In this setting, at each time t the learner must make the decision of whether to accept the available measurement x_t, or abstain and wait for an even more informative measurement. While abstention may result in a smaller total labeled sample complexity $\mathcal{L}$, the stopping time $\mathcal{U}$ may be very large. This paper characterizes the set of feasible pairs $(\mathcal{U}, \mathcal{L})$ that are necessary and sufficient to identify z_* with probability at least $1-\delta$ when x_t are drawn IID at each time t from a distribution ν. Moreover, we propose an algorithm that nearly obtains the minimal information theoretic label sample complexity $\mathcal{L}$ for any desired unlabeled sample complexity $\mathcal{U}$.

While characterizing the sample complexity of selective sampling for online best arm identification is the primary theoretical goal of this work, the study was initially motivated by fundamental questions about how to optimally trade-off the value of information versus time. Even for this idealized linear setting, it is far from obvious a priori what an optimal decision rule ξ_t looks like and if it can even be succinctly described, or if it is simply the solution to an opaque optimization problem. Remarkably,

*Equal contribution. Alphabetical order.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

we show that for every feasible, optimal operating pair $(\mathcal{U}, \mathcal{L})$ there exists a matrix $A \in \mathbb{R}^{d\times d}$ such that the optimal decision rule takes on the form $\xi_t = \mathbf{1}\{x^\top Ax \geq 1\}$ when $x_t \sim \nu$ iid. The fact that for any smooth distribution ν the decision rule is a hard decision equivalent to x_t falling outside a fixed ellipse or not, and not a stochastic rule that varies complementarily with the density of ν over space is perhaps unexpected.

To motivate the problem description, suppose on each day $t = 1, 2, \ldots$ a food blogger posts the *Cocktail of the Day* with a recipe described by a feature vector $x_t \in \mathbb{R}^d$. You have the ingredients (and skills) to make any possible cocktail in the space of all cocktails $\mathcal{Z}$, but you don't know which one you'd like the most, i.e., $z_* := \arg\max_{z\in\mathcal{Z}} \langle z, \theta_* \rangle$, where θ_* captures your preferences over cocktail recipes. You decide to use the *Cocktail of the Day* to inform your search. That is, each day you are presented with the cocktail recipe $x_t \in \mathbb{R}^d$, and if you choose to make it ($\xi_t = 1$) you observe your preference for the cocktail y_t with $\mathbb{E}[y_t] = \langle x_t, \theta_* \rangle$. Of course, making cocktails can get costly, so you don't want to make each day's cocktail, but rather you will only make the cocktail if x_t is informative about θ_* (e.g., uses a new combination of ingredients). At the same time, waiting too many days before making the next cocktail of the day may mean that you never get to learn (and hence drink) the cocktail z_* you like best. The setting above is not limited to cocktails, but rather naturally generalizes to discovering the efficacy of drugs and other therapeutics where blood and tissue samples come to the clinic in a stream and the researcher has to choose whether to take a potentially costly measurement.

Our results hold for arbitrary $\theta_* \in \mathbb{R}^d$, sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Z} \subset \mathbb{R}^d$, and measures $\nu \in \triangle_{\mathcal{X}}$[1] for which we assume $x_t \sim \nu$ is drawn IID. The assumption that each x_t is IID allows us to make very strong statements about optimality. To summarize, our contributions are as follows:

- We present fundamental limits on the trade-off between the amount of unlabelled data and labelled data in the form of (the first) information theoretic lower bounds for selective sampling problems that we are aware of. Naturally, they say that there is an absolute minimum amount of unlabelled data that is necessary to solve the problem, but then for any amount of unlabelled data beyond this critical value, the bounds say that the amount of labelled data must exceed some value as a function of the unlabelled data used.
- We propose an algorithm that nearly matches the lower bound at all feasible trade-off points in the sense that given any unlabelled data budget that exceeds the critical threshold, the algorithm takes no more labels than the lower bound suggests. Thus, the upper and lower bounds sketch out a curve of all possible operating points, and the algorithm achieves any point on this curve.
- We characterize the optimal decision rule of whether to take a sample or not, based on any critical point is a simple test: Accept $x_t \in \mathbb{R}^d$ if $x_t^\top Ax_t \geq 1$ for some matrix A that depends on the desired operating point and geometry of the task. Geometrically, this is equivalent to x_t falling inside or outside an ellipsoid.
- Our framework is also general enough to capture binary classification, and consequently, we prove results there that improve upon state of the art.

1.1 Related Work

Selective Sampling in the Streaming Setting: Online prediction, the setting in which the selective sampling framework was introduced, is a closely related problem to the one studied in this paper and enjoys a much more developed literature [6, 9, 1, 7]. In the linear online prediction setting, for $t = 1, 2, \ldots$ Nature reveals $x_t \in \mathbb{R}^d$, the learner predicts $\widehat{y}_t$ and incurs a loss $\ell(\widehat{y}_t, y_t)$, and then the learner decides whether to observe y_t (i.e., $\xi_t = 1$) or not ($\xi_t = 0$), where y_t is a label generated by a composition of a known link function with a linear function of x_t. For example, in the classification setting [1, 6, 9], one setting assumes $y_t \in \{-1, 1\}$ with $\mathbb{E}[y_t|x_t] = \langle x_t, \theta_* \rangle$ for some unknown $\theta_* \in \mathbb{R}^d$, and $\ell(\widehat{y}_t, y_t) = \mathbf{1}\{\widehat{y}_t \neq y_t\}$. In the regression setting [7], one observes $y_t \in [-1, 1]$ with $\mathbb{E}[y_t|x_t] = \langle x_t, \theta_* \rangle$ again, and $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$. After any amount of time $\mathcal{U}$, the learner is incentivized to minimize both the amount of requested labels $\sum_{t=1}^{\mathcal{U}} \mathbf{1}\{\xi_t = 1\}$ and the cumulative loss $\sum_{t=1}^{\mathcal{U}} \ell(y_t, \widehat{y}_t)$ (or some measure of regret which compares to predictions using the unknown θ_*). If every label y_t is requested then $\mathcal{L} = \mathcal{U}$ and this is just the classical online learning setting.

[1] We denote the set of probability measures over $\mathcal{X}$ as $\triangle_{\mathcal{X}}$.

These works give a guarantee on the regret and labeled points taken in terms of the hardness of the stream relative to a learner which would see the label at every time. Most do not give the learner the ability to select an operating point that provides a trade-off between the amount of unlabeled versus labeled data taken. Those few works that propose algorithms that do provide this functionality do not provide lower bounds that match their given upper bounds, leaving it unclear whether their algorithm optimally negotiates this trade-off. In contrast, our work fully characterizes the trade-off between the amount of unlabeled and labeled data through an information-theoretic lower bound and a matching upper bound. Specifically, our algorithm includes a tuning parameter, call it τ, that controls the trade-off between the evaluation metric of interest (for us, the quality of the recommended $z \in \mathcal{Z}$), the label complexity $\mathcal{L}$, and the amount of unlabelled data $\mathcal{U}$ that is necessary before the metric of interest can be non-trivial. We prove that each possible setting of τ parametrizes *all* possible trade-offs between unlabeled and labeled data.

Our work is perhaps closest to the streaming setting for agnostic active classification [8, 15] where each x_s is drawn i.i.d. from an underlying distribution ν on $\mathcal{X}$, and indeed our results can be specialized to this setting as we discuss in Section 3. These papers also evaluate themselves at a single point on the tradeoff curve, namely the number of samples needed in passive supervised learning to obtain a learner with excess risk at most ϵ. They provide minimax guarantees on the amount of labeled data needed in terms of the disagreement coefficient [12]. In contrast, again, our results characterize the full trade-off between the amount of unlabeled data seen, and the amount of labeled data needed to achieve the target excess risk ϵ. We note that using online-to-batch conversion methods, [9, 1, 6] also provide results on the amount of labeled data needed but they assume a very specific parametric form to their label distribution unlike our setting which is agnostic. Other works have characterized selective sampling for classification in the realizable setting that assumes there exists a classifer among the set under consideration that perfectly labels every y_t [13]–our work addresses the agnostic setting where no such assumption is made. Finally, our results apply under the more general setting of *domain adaptation under covariate shift* where we are observing data drawn from the stream ν, but we will evaluate the excess risk of our resulting classifier on a different stream π [22, 23, 26].

Best-Arm Identification and Online Experimental Design. Our techniques are based on experimental design methods for best-arm identification in linear bandits, see [24, 11, 5]. In the setting of these works, there exists a pool of examples $\mathcal{X}$ and at each time any $x \in \mathcal{X}$ can be selected with replacement. The goal is to identify the best arm using as few total selections (labels) as possible. Their algorithms are based on arm-elimination. Specifically, they select examples with probability proportional to an approximate G-optimal design with respect to the current remaining arms. Then, during each round after taking measurements, those arms with high probability of being suboptimal will be eliminated. Remarkably, near-optimal sample complexity has been achieved under this setting. While we apply these techniques of arm-elimination and sampling through G-optimal design, the major difference is that we are facing a stream instead of a pool of examples. Finally, [10] considers a different online experiment design setup where (adversarially chosen) experiments arrive sequentially and a primal-dual algorithm decides whether to choose each, subject to a total budget. [10] studies the competitive ratio of such algorithms (in the manner of online packing algorithms) for problems such as D-optimal experiment design.

2 Selective Sampling for Best Arm Identification

Consider the following game: Given known $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ and unknown $\theta_* \in \mathbb{R}^d$ at each time $t = 1, 2, \ldots$:

1. Nature reveals $x_t \overset{iid}{\sim} \nu$ with support$(\nu) = \mathcal{X}$
2. Player chooses $Q_t \in \{0, 1\}$. If $Q_t = 1$ then nature reveals y_t with $\mathbb{E}[y_t] = \langle x_t, \theta_* \rangle$
3. Player optionally decides to stop at time t and output some $\widehat{z} \in \mathcal{Z}$

If the player stops at time $\mathcal{U}$ after observing $\mathcal{L} = \sum_{t=1}^{\mathcal{U}} Q_t$ labels, the objective is to identify $z_* = \arg\max_{z \in \mathcal{Z}} \langle z, \theta_* \rangle$ with probability at least $1 - \delta$ while minimizing a trade-off of $\mathcal{U}, \mathcal{L}$.

This paper studies the relationship between $\mathcal{U}$ and $\mathcal{L}$ in the context of necessary and sufficient conditions to identify z_* with probability at least $1 - \delta$. Clearly $\mathcal{U}$ must be "large enough" for z_* to

be identifiable even if all labels are requested (i.e., $\mathcal{L} = \mathcal{U}$). But if $\mathcal{U}$ is very large, the player can start to become more picky with their decision to observe the label or not. Indeed, one can easily imagine scenarios in which it is advantageous for a player to forgo requesting the label of the current example in favor of waiting for a more informative example to arrive later if they wished to minimize $\mathcal{L}$ alone. Intuitively, $\mathcal{L}$ should decrease as $\mathcal{U}$ increases, but how?

Any selective sampling algorithm for the above protocol at time t is defined by 1) a selection rule $P_t : \mathcal{X} \to [0, 1]$ where $Q_t \sim \text{Bernoulli}(P_t(x_t))$, 2) a stopping rule $\mathcal{U}$, and 3) a recommendation rule $\widehat{z} \in \mathcal{Z}$. The algorithm's behavior at time t can use all information collected up to time t

Definition 1. *For any $\delta \in (0, 1)$ we say a selective sampling algorithm is δ-PAC for $\nu \in \triangle_{\mathcal{X}}$ if for all $\theta \in \mathbb{R}^d$ the algorithm terminates at time $\mathcal{U}$ which is finite almost surely and outputs $\arg\max_{z\in\mathcal{Z}} \langle z, \theta\rangle$ with probability at least $1 - \delta$.*

2.1 Optimal design

Before introducing our own algorithm, let us consider a seemingly optimal procedure. For any $\lambda \in \triangle_{\mathcal{X}} = \{p : \sum_{x\in\mathcal{X}} p_x = 1,\, p_x \geq 0\ \forall x \in \mathcal{X}\}$ define

$$\rho(\lambda) := \max_{z\in\mathcal{Z}\setminus\{z_*\}} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\langle \theta_*, z_* - z\rangle^2}. \tag{1}$$

Intuitively, $\rho(\lambda)$ captures the number of labeled examples drawn from distribution λ to identify z_*. Specifically, for any $\tau \geq \rho(\lambda)\log(|\mathcal{Z}|/\delta)$, if $x_1, \ldots, x_\tau \sim \lambda$ and $y_i = \langle x_i, \theta_*\rangle + \epsilon_i$ where ϵ_i is iid 1 sub-Gaussian noise, then there exists an estimator $\widehat{\theta} := \widehat{\theta}(\{(x_i, y_i)\}_{i=1}^\tau)$ such that $\langle \widehat{\theta}, z_*\rangle > \max_{z\in\mathcal{Z}\setminus z_*} \langle \widehat{\theta}, z\rangle$ with probability at least $1 - \delta$ [11]. In particular, $\tau \geq \rho(\lambda)\log(|\mathcal{Z}|/\delta)$ samples suffice to guarantee that $\arg\max_{z\in\mathcal{Z}} \langle \widehat{\theta}, z\rangle = \arg\max_{z\in\mathcal{Z}} \langle \theta_*, z\rangle =: z_*$.

Thus, if our τ samples are coming from ν, we would expect any reasonable algorithm to require at least $\rho(\nu)\log(|\mathcal{Z}|/\delta)$ examples and labels. However, since we only want to take informative examples, we instead choose to select the tth example $x_t = x$ according to a probability $P(x)$ so that our final labeled samples are coming from the distribution λ where $\lambda(x) \propto P(x)\nu(x)$. In particular, $P(x)$ should be chosen according to the following optimization problem

$$P^* = \operatorname*{argmin}_{P:\mathcal{X}\to[0,1]} \tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \max_{z\in\mathcal{Z}\setminus\{z_*\}} \frac{\|z_* - z\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z_* - z, \theta_*\rangle^2}\beta_\delta \leq 1 \tag{2}$$

for $\beta_\delta = \log(|\mathcal{Z}|/\delta)$ where the objective captures the number of samples we select using P^*, and the constraint captures the fact that we have solved the problem. Remarkably, we can reparametrize this result in terms of an optimization problem over $\lambda \in \Delta_{\mathcal{X}}$ instead of $P^* : \mathcal{X} \to [0, 1]$ as

$$\tau\mathbb{E}_{X\sim\nu}[P^*(X)] = \min_{\lambda\in\triangle_{\mathcal{X}}} \rho(\lambda)\beta_\delta \quad \text{subject to} \quad \tau \geq \|\lambda/\nu\|_\infty \rho(\lambda)\beta_\delta$$

where $\|\lambda/\nu\|_\infty = \max_{x\in\mathcal{X}} \lambda(x)/\nu(x)$, as shown in Proposition 2. Note that as $\tau \to \infty$ the constraint becomes inconsequential. Also notice that $\rho(\nu)\beta_\delta$ appears to be a necessary amount of labels to solve the problem even if $P(x) \equiv 1$ (albeit, by arguing about minimizing the upperbound of above).

2.2 Main results

In this section we formally justify the sketched argument of the previous section, showing nearly matching upper and lower bounds.

Theorem 1 (Lower bound)**.** *Fix any $\delta \in (0, 1)$, $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, and $\theta_* \in \mathbb{R}^d$. Any selective sampling algorithm that is δ-PAC for $\nu \in \triangle_{\mathcal{X}}$ and terminates after drawing $\mathcal{U}$ unlabelled examples from ν and requests the labels of just $\mathcal{L}$ of them satisfies*

- $\mathbb{E}[\mathcal{U}] \geq \rho(\nu)\log(1/\delta)$*, and*
- $\mathbb{E}[\mathcal{L}] \geq \min_{\lambda\in\triangle_{\mathcal{X}}} \rho(\lambda)\log(1/\delta)$ *subject to* $\mathbb{E}[\mathcal{U}] \geq \|\lambda/\nu\|_\infty \rho(\lambda)\log(1/\delta)$.

The first part of the theorem quantifies the number of rounds or unlabelled draws $\mathcal{U}$ that *any* algorithm must observe before it could hope to stop and output z_* correctly. The second part describes a

trade-off between $\mathcal{U}$ and $\mathcal{L}$. One extreme is if $\mathbb{E}[\mathcal{U}] \to \infty$, which effectively removes the constraint so that the number of observed labels must scale like $\min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda) \log(1/\delta)$. Note that this is precisely the number of labels required in the pool-based setting where the agent can choose *any* $x \in \mathcal{X}$ that she desires at each time t (e.g. [11]). In the other extreme, $\mathbb{E}[\mathcal{U}] = \rho(\nu) \log(1/\delta)$ so that the constraint in the label complexity $\mathbb{E}[\mathcal{L}]$ is equivalent to $\rho(\nu) \geq \|\lambda/\nu\|_\infty \rho(\lambda)$. This implies that the minimizing λ must either stay very close to ν, or must obtain a substantially smaller value of $\rho(\lambda)$ relative to $\rho(\nu)$ to account for the inflation factor $\|\lambda/\nu\|_\infty$. In some sense, this latter extreme is the most interesting point on the trade-off curve because its asking the algorithm to stop as quickly as the algorithm that observes all labels, but after requesting a minimal number of labels. Note that this lower bound holds even for algorithms that known ν exactly. The proof of Theorem 1 relies on standard techniques from best arm identification lower bounds (see e.g. [17, 11]).

Remarkably, every point on the trade-off suggested by the lower bound is nearly achievable.

Theorem 2 (Upper bound)**.** *Fix any $\delta \in (0,1)$, $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, and $\theta_* \in \mathbb{R}^d$. Let $\Delta = \min_{z \in \mathcal{Z} \setminus \{z_*\}} \langle z_* - z, \theta_* \rangle$ and $\beta_\delta \propto \log(\log(\frac{1}{\Delta})|\mathcal{Z}|/\delta)$ where the precise constant is given in the appendix. For any $\tau \geq \rho(\nu)\beta_\delta$ there exists a δ-PAC selective sampling algorithm that observes $\mathcal{U}$ unlabeled examples and requests just $\mathcal{L}$ labels that satisfies with probability at least $1-\delta$*

- $\mathcal{U} \leq \log_2(\frac{4}{\Delta})\,\tau$, *and*
- $\mathcal{L} \leq 3\log_2(\frac{4}{\Delta}) \min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\,\beta_\delta$ *subject to* $\tau \geq \|\lambda/\nu\|_\infty \rho(\lambda)\,\beta_\delta$.

Aside from the $\log(\frac{1}{\Delta})$ factor and the $\log(|\mathcal{Z}|)$ that appears in the β_δ term, this nearly matches the lower bound. Note that the parameter τ parameterizes the algorithm and makes the trade-off between $\mathcal{U}$ and $\mathcal{L}$ explicit. The next section describes the algorithm that achieves this theorem.

2.3 Selective Sampling Algorithm

Algorithm 1 contains the pseudo-code of our selective sampling algorithm for best-arm identification. Note that it takes a confidence level $\delta \in (0,1)$ and a parameter τ that controls the unlabeled-labeled budget trade-off as input. The algorithm is effectively an elimination style algorithm and closely mirrors the RAGE algorithm for the pool-based setting of best-arm identification problem [11]. The key difference, of course, is that instead of being able to plan over the pool of measurements, this algorithm must plan over the x's that the algorithm may *potentially* see and account for the case that it might not see the x's it wants.

Algorithm 1 Selective Sampling for Best-arm Identification

1: **Input** $\mathcal{Z} \subset \mathbb{R}^d$, $\delta \in (0,1), \tau$
2: **while** $|\mathcal{Z}_\ell| \geq 1$ **do**
3: Let $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$OPTIMIZEDESIGN$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ // $\widehat{\Sigma}_{\widehat{P}_\ell}$ approximates $\mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)XX^\top]$
4: **for** $t = (\ell-1)\tau + 1, \ldots, \ell\tau$ **do**
5: Nature reveals x_t drawn iid from ν (with support $\mathbb{R}^d$)
6: Sample $Q_t(x_t) \sim$ Bernoulli$(\widehat{P}_\ell(x_t))$. If $Q_t = 1$ then observe y_t // $\mathbb{E}[y_t|x_t] = \langle \theta_*, x_t \rangle$
7: **end for**
8: Let $\widehat{\theta}_\ell \leftarrow$RIPS$(\{\widehat{\Sigma}_{\widehat{P}_\ell}^{-1} Q_s(x_s) x_s y_s\}_{s=(\ell-1)\tau+1}^{\ell\tau}, \mathcal{Z} \times \mathcal{Z})$ // $\widehat{\theta}_\ell$ approximates θ_*
9: $\mathcal{Z}_{\ell+1} = \mathcal{Z}_\ell \setminus \{z \in \mathcal{Z}_\ell : \max_{z' \in \mathcal{Z}_\ell} \langle z' - z, \widehat{\theta}_\ell \rangle \geq 2^{-\ell}\}$
10: **end while**

In round ℓ, the algorithm maintains an active set $\mathcal{Z}_\ell \subseteq \mathcal{Z}$ with the guarantee that each remaining $z \in \mathcal{Z}_\ell$ satisfies, $\langle z_* - z, \theta_* \rangle \leq 8 \cdot 2^{-\ell}$. In each round, on Line 3 of the algorithm, it calls out to a sub-routine OPTIMIZEDESIGN$(\mathcal{Z}, \epsilon, \tau)$ that is trying to approximate the ideal optimal design of (2). In particular, the ideal response to OPTIMIZEDESIGN$(\mathcal{Z}, \epsilon, \tau)$ would return a P_ϵ^* and $\Sigma_{P_\epsilon^*} = \mathbb{E}_{X\sim\nu}[P_\epsilon^*(X)XX^\top]$ where P_ϵ^* is the solution to Equation 2 with the one exception that the denominator of the constraint is replaced with $\max\{\epsilon^2, \langle \theta_*, z_* - z \rangle^2\}$. Of course, θ_* is unknown so we cannot solve Equation 2 (as well as other outstanding issues that we will address shortly). Consequently, our implementation will aim to *approximate* the optimization problem of Equation 2.

But assuming our sample complexity is not too far off from this ideal, each round should not request more labels than the number of labels requested by the ideal program with $\epsilon = 0$. Thus, the total number of samples should be bounded by the ideal sample complexity times the number of rounds, which is $O(\log(\Delta^{-1}))$. We will return to implementation issues in the next section.

Assuming we are returned $(\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell})$ that approximate their ideals as just described, the algorithm then proceeds to process the incoming stream of $x_t \sim \nu$. As described above, the decision to request the label of x_t is determined by a coin flip coming up heads with probability $\widehat{P}_\ell(x_t)$–otherwise we do not request the label. Given the collected dataset $\{(x_t, y_t, Q_t, \widehat{P}_\ell(x_t))\}_t$, line 8 then computes an estimate $\widehat{\theta}_\ell$ of θ_* using the RIPS estimator of [5] which will satisfy

$$|\langle z_* - z, \widehat{\theta}_\ell - \theta_* \rangle| \leq O\left(\|z_* - z\|_{\mathbb{E}_{X\sim\nu}[\tau \widehat{P}_\ell(X) X X^\top]^{-1}} \sqrt{\log(2\ell^2 |\mathcal{Z}|^2/\delta)} \right) \leq 2^{-\ell}$$

for all $z \in \mathcal{Z}_\ell$ simultaneously with probability at least $1 - \delta$. Thus, the final line of the algorithm eliminates any $z \in \mathcal{Z}_\ell$ such that there exists another $z' \in \mathcal{Z}_\ell$ (think z_*) that satisfies $\langle \widehat{\theta}_\ell, z' - z \rangle > 2^{-\ell}$. The process continues until $\mathcal{Z}_\ell = \{z_*\}$.

2.4 Implementation of OPTIMIZEDESIGN

For the subroutine OPTIMIZEDESIGN passed $(\mathcal{Z}_\ell, \epsilon, \tau)$ the next best thing to computing Equation 2 with the denominator of the constraint replaced with $\max\{\epsilon^2, \langle \theta_*, z_* - z \rangle^2\}$, is to compute

$$P_\epsilon = \operatorname*{argmin}_{P:\mathcal{X}\to[0,1]} \mathbb{E}_{X\sim\nu}[P(X)] \text{ subject to } \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z - z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X) X X^\top]^{-1}}}{\epsilon^2} \beta_\delta \leq 1 \qquad (3)$$

and $\Sigma_{P_\epsilon} = \mathbb{E}_{X\sim\nu}[P_\epsilon(X) X X^\top]$ for an appropriate choice of $\beta_\delta = \Theta(\log(|\mathcal{Z}|/\delta))$. To see this, firstly, any $z \in \mathcal{Z}$ with gap $\langle \theta_*, z_* - z \rangle$ that we could accurately estimate would not be included in $\mathcal{Z}_\ell$, thus we don't need it in the $\max$ of the denominator. Secondly, to get rid of z_* in the numerator (which is unknown, of course), we note that for any norm $\max_{z,z'} \|z - z'\| \leq \max_z 2\|z - z_*\| \leq \max_{z,z'} 2\|z - z'\|$. Assuming we could solve this directly and compute $\Sigma_{P_\epsilon} = \mathbb{E}_{X\sim\nu}[P_\epsilon(X) X X^\top]$, we can obtain the result of Theorem 2 (proven in the Appendix).

However, even if we knew ν exactly, the optimization problem of Equation 3 is quite daunting as it is a potentially infinite dimensional optimization problem over $\mathcal{X}$. Fortunately, after forming the Lagrangian with dual variables for each $z - z' \in \mathcal{Z} \times \mathcal{Z}$, optimizing the dual amounts to a finite dimensional optimization problem over the finite number of dual variables. Moreover, this optimization problem is maximizing a simple expectation with respect to ν and thus we can apply standard stochastic gradient ascent and results from stochastic approximation [20]. Given the connection to stochastic approximation, instead of sampling a fresh $\widetilde{x} \sim \nu$ each iteration, it suffices to "replay" a sequence of $\widetilde{x}$'s from historical data. Summing up, this construction allows us to compute a satisfactory P_ϵ and avoid both an infinite-dimensional optimization problem and requiring knowledge of ν (as long as historical data is available).

Meanwhile, with historical data, we can also empirically compute $\mathbb{E}_{X\sim\nu}[P_\epsilon(X) X X^\top]$. Historical data could mean offline samples from ν or just samples from previous rounds. In this setting, Theorem 2 still holds albeit with larger constants. Theorem 7 in the appendix characterizes the necessary amount of historical data needed. Unfortunately (in full disclosure) the theoretical guarantees on the amount of historical data needed is absurdly large, though we suspect this arises from a looseness in our analysis. Similar assumptions and approaches to historical or offline data have been used in other works in the streaming setting e.g. [15].

3 Selective Sampling for Binary Classification

We now review streaming Binary Classification in the agnostic setting [8, 12, 15] and show that our approach can be adapted to this setting. Consider a binary classification problem where $\mathcal{X}$ is the example space and $\mathcal{Y} = \{-1, 1\}$ is the label space. Fix a hypothesis class $\mathcal{H}$ such that each $h \in \mathcal{H}$ is a classifier $h : \mathcal{X} \to \mathcal{Y}$. Assume there exists a fixed regression function $\eta : \mathcal{X} \to [0, 1]$ such that the label of x is Bernoulli with probability $\eta(x) = \mathbb{P}(Y = 1 | X = x)$. Being in the agnostic setting, we make no assumption on the relationship between $\mathcal{H}$ and η. Finally, fix any $\nu \in \triangle_\mathcal{X}$ and $\pi \in \triangle_\mathcal{X}$. Given known $\mathcal{X}, \mathcal{H}$ and unknown regression function η, at each time $t = 1, 2, \ldots$:

1. Nature reveals $x_t \sim \nu$
2. Player chooses $Q_t \in \{0, 1\}$. If $Q_t = 1$ then nature reveals $y_t \sim \text{Bernoulli}(\eta(x_t)) \in \{-1, 1\}$
3. Player optionally decides to stop at time t and output some $\widehat{h} \in \mathcal{H}$.

Define the *risk* of any $h \in \mathcal{H}$ as $R_\pi(h) := \mathbb{P}_{X\sim\pi, Y\sim\eta(X)}(Y \neq h(X))$. If the player stops at time $\mathcal{U}$ after observing $\mathcal{L} = \sum_{t=1}^{\mathcal{U}} Q_t$ labels, the objective is to identify $h_* = \arg\min_{h\in\mathcal{H}} R_\pi(h)$ with probability at least $1 - \delta$ while minimizing a trade-off of $\mathcal{U}, \mathcal{L}$. Note that h_* is the true risk minimizer with respect to distribution π but we observe samples $x_t \sim \nu$; π is not necessarily equal to ν. While we have posed the problem as identifying the potentially unique h^*, our setting naturally generalizes to identifying an ϵ-good h such that $R_\pi(h) - R_\pi(h_*) \leq \epsilon$.

We will now reduce selective sampling for binary classification problem to selective sampling for best arm identification, and thus immediately obtain a result on the sample complexity. For simplicity, assume that $\mathcal{X}$ and $\mathcal{H}$ are finite. Enumerate $\mathcal{X}$ and for each $h \in \mathcal{H}$ define a vector $z^{(h)} \in [0, 1]^{|\mathcal{X}|}$ such that $z_x^{(h)} := \pi(x)\mathbf{1}\{h(x) = 1\}$ for $z^{(h)} = [z_x^{(h)}]_{x\in\mathcal{X}}$. Moreover, define $\theta^* := [\theta_x^*]_{x\in\mathcal{X}}$ where $\theta_x^* := 2\eta(x) - 1$. Then

$$R_\pi(h) = \mathbb{E}_{X\sim\pi, Y\sim\eta(X)}[\mathbf{1}\{Y \neq h(X)\}] = \sum_{x\in\mathcal{X}} \pi(x)(\eta(x)\mathbf{1}\{h(x) \neq 1\} + (1 - \eta(x))\mathbf{1}\{h(x) \neq 0\})$$
$$= \sum_{x\in\mathcal{X}} \pi(x)\eta(x) + \sum_{x\in\mathcal{X}} \pi(x)(1 - 2\eta(x))\mathbf{1}\{h(x) = 1\} = c - \langle z^{(h)}, \theta^*\rangle$$

where $c = \sum_{x\in\mathcal{X}} \pi(x)\eta(x)$ does not depend on h. Thus, if $\mathcal{Z} := \{z^{(h)}\}_{h\in\mathcal{H}}$ then identifying $h_* = \arg\min_{h\in\mathcal{H}} R_\pi(h)$ is equivalent to identifying $z_* = \arg\max_{z\in\mathcal{Z}} \langle z, \theta^*\rangle$. We can now apply Theorem 2 to obtain a result describing the sample complexity trade-off. First define,

$$\rho_\pi(\lambda, \varepsilon) := \max_{z\in\mathcal{Z}\setminus\{z_*\}} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\max\{\langle\theta_*, z_* - z\rangle^2, \varepsilon^2\}} = \max_{h\in\mathcal{H}\setminus\{h_*\}} \frac{\mathbb{E}_{X\sim\pi}\left[\mathbf{1}\{h(X) \neq h'(X)\}\frac{\pi(X)}{\lambda(X)}\right]}{\max\{(R_\pi(h) - R_\pi(h^*))^2, \varepsilon^2\}}$$

An important case of the above setting is when $X \sim \nu$ and $\pi = \nu$, i.e. we are evaluating the performance of a classifier relative to the same distribution our samples are drawn from. This is the setting of [8, 15, 12]. The following theorem shows that the sample complexity obtained by our algorithm is at least as good as the results they present.

Theorem 3. *Fix any $\delta \in (0, 1)$, domain $\mathcal{X}$ with distribution ν, finite hypothesis class $\mathcal{H}$, regression function $\eta : \mathcal{X} \to [0, 1]$. Set $\epsilon \geq 0$ and $\beta_\delta = 2048 \log(4 \log_2^2(4/\epsilon)|\mathcal{H}|/\delta)$. Then for $\tau \geq \rho_\pi(\nu, \epsilon)\beta_\delta$ there exists a selective sampling algorithm that returns $h \in \mathcal{H}$ satisfying $R_\pi(h) - R_\pi(h^*) \leq \epsilon$ by observing $\mathcal{U}$ unlabeled examples and requesting just $\mathcal{L}$ labels such that*

- $\mathcal{U} \leq \log_2(4/\epsilon)\tau$
- $\mathcal{L} \leq 3 \log_2(\frac{4}{\varepsilon}) \min_{\lambda\in\Delta_\mathcal{X}} \rho_\pi(\lambda, \varepsilon)\beta_\delta \quad$ *s.t.* $\quad \tau \geq \|\lambda/\nu\|_\infty \rho_\pi(\lambda, \varepsilon)\beta_\delta$

with probability at least $1 - \delta$. Furthermore when $\nu = \pi$ and if $\tau \geq 16\rho(\nu, \epsilon)\beta_\delta$ we have that

$$\mathcal{L} \leq 36 \log_2(4/\epsilon) \left(\frac{R_\nu(h^*)^2}{\epsilon^2} + 4\right) \sup_{\xi\geq\epsilon} \theta^*(2R_\nu(h^*) + \xi, \nu)\beta_\delta$$

where $\theta^*(u, \nu)$ is the disagreement coefficient, defined in Appendix E.

Note that if τ is sufficiently large then the labeled sample complexity we obtain $\min_{\lambda\in\Delta_\mathcal{X}} \rho(\lambda, \epsilon)$ could be significantly smaller than previous results in the streaming setting, e.g. see [16]. The proof of Theorem 3 can be found in Appendix E.

4 Solving the Optimization Problem

Recall that in Algorithm 1, during round ℓ, we need to solve optimization problem (3). Solving this optimization problem is not trivial because the number of variables can potentially be infinite if $\mathcal{X}$ is

an infinite set. In this section, we will demonstrate how to reduce it to a finite-dimensional problem by considering its dual problem. To simplify the notation, let $\mathcal{Y}_\ell = \{z - z' : z, z' \in \mathcal{Z}_\ell, z \neq z'\}$, and rewrite the problem as follows, where $c_\ell > 0$ is a constant that may depend on round ℓ.

$$\begin{array}{rl} \min_P & \mathbb{E}_{X\sim\nu}\left[P(X)\right] \\ \text{subject to} & y^\top \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq P(x) \leq 1, \quad \forall x \in \mathcal{X}. \end{array} \tag{4}$$

Using the Schur complement technique, we show in Lemma 13 (Appendix C) the following equivalence: $y^\top \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1} y \leq c_\ell^2 \iff \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{1}{c_\ell^2} yy^\top$. This transforms a constraint involving matrix inversion into one with ordering between PSD matrices. Then, we remove the bound constraints $0 \leq P(x) \leq 1$, $\forall x \in \mathcal{X}$ by introducing the barrier function $-\log(1-x) - \log(x)$. That is, instead of working with the objective $\mathbb{E}_{X\sim\nu}\left[P(X)\right]$ directly, we consider the following problem.

$$\begin{array}{rl} \min_P & \mathbb{E}_{X\sim\nu}[P(X) - \mu_b(\log(1-P(X)) + \log(P(X)))] \\ \text{subject to} & \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{1}{c_\ell^2} yy^\top, \quad \forall y \in \mathcal{Y}_\ell. \end{array} \tag{5}$$

Here, $\mu_b \in (0,1)$ is some small constant that controls how strong the barrier is. Intuitively, a smaller μ_b will make problem (5) closer to the original problem. We now show that unlike the primal, the dual problem is indeed finite-dimensional. For each constraint of $y \in \mathcal{Y}_\ell$, let the matrix $\Lambda_y \succeq \mathbf{0}$ be its dual variable. Further, let $\Lambda = \sum_{y\in\mathcal{Y}_\ell} \Lambda_y$ and $\mathbf{\Lambda} = (\Lambda_y)_{y\in\mathcal{Y}_\ell}$. The corresponding Lagrangian is

$$\mathcal{L}\left(\mathbf{\Lambda}, P\right) = \mathbb{E}_{X\sim\nu}\left[P(X) - \mu_b\left(\log(1-P(X)) + \log(P(X))\right) - P(X)X^\top \Lambda X\right] + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y.$$

The dual problem is $\max_{\Lambda_y \succeq \mathbf{0}, \forall y\in\mathcal{Y}_\ell} \min_P \mathcal{L}\left(\mathbf{\Lambda}, P\right)$. Notice that minimization over $P : \mathcal{X} \mapsto [0,1]$ can be done via minimizing $P(x)$ point-wise for each $x \in \mathcal{X}$. To do this, we take the gradient with respect to each $P(x)$ and set it to zero to get

$$1 + \frac{\mu_b}{1-P(x)} - \frac{\mu_b}{P(x)} - x^\top \Lambda x = 0. \tag{6}$$

Solving this equation and defining $q_\Lambda(x) = x^\top \Lambda x - 1$, we get

$$P_\Lambda(x) = \frac{1}{2} - \frac{\mu_b}{q_\Lambda(x)} + \frac{\sqrt{\left(2\mu_b - q_\Lambda(x)\right)^2 + 4\mu_b q_\Lambda(x)}}{2q_\Lambda(x)}. \tag{7}$$

Note that if $\mu_b = 0$ (no barrier), the above reduces to the "threshold" decision rule $P_\Lambda(x) = \frac{1}{2} + \frac{|q_\Lambda(x)|}{2q_\Lambda(x)}$, which gives 0 when $q_\Lambda(x) < 0$ and 1 when $q_\Lambda(x) > 0$.[2] This is exactly the hard elliptical threshold rule mentioned before, in which whether to query the label for x depends on whether it falls inside ($x^\top \Lambda x < 1$) or outside ($x^\top \Lambda x > 1$) of the ellipsoid defined by the positive semidefinite matrix Λ. A visualization of the decision rule P_Λ is given in Figure 2 in the Appendix.

Now, by plugging in $P_\Lambda(x)$, our dual problem becomes $\max_{\Lambda_y \succeq \mathbf{0}, \forall y} D(\mathbf{\Lambda}) := \mathcal{L}\left(\mathbf{\Lambda}, P_\Lambda\right)$. This is a finite-dimensional optimization problem, and can be solved by projected gradient ascent (or projected stochastic gradient ascent when we have only samples from ν). The gradient of $D(\mathbf{\Lambda})$ is

$$\begin{aligned} \nabla_{\Lambda_y} D(\mathbf{\Lambda}) &= \mathbb{E}_{X\sim\nu}\left[\left(1 + \frac{\mu_b}{1-P_\Lambda(x)} - \frac{\mu_b}{P_\Lambda(X)} - X^\top \Lambda X\right)\nabla_{\Lambda_y} P_\Lambda(X) - P_\Lambda(X)XX^\top\right] + \frac{yy^\top}{c_\ell^2} \\ &= \frac{yy^\top}{c_\ell^2} - \mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)XX^\top\right]. \qquad \text{(Since } P_\Lambda(X) \text{ solves Eq. (6))} \end{aligned}$$

The algorithm to solve the problem has been summarized in Algorithm 2, in which the gradient during kth iteration is replaced by its unbiased estimator $\frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_k x_k^\top$. The adaptive learning rate is chosen by following the discussion in chapter 4 of [21]. Optimizing the assignment of $\hat{\Lambda}_y$ to each y in line 10 ensures that the re-scaling step in line 11 increases the function value in an optimized way. Finally, the re-scaling step is used to ensure that the output primal objective value $\mathbb{E}_{X\sim\nu}\left[P(X)\right]$ is bounded well, which will be explained in more details in Appendix C.

[2] When $q_\Lambda(x) = 0$, $P_\Lambda(x)$ is undetermined from the dual.

Algorithm 2 Projected Stochastic Gradient Ascent to Solve OPTIMIZEDESIGN

Input: Number of iterations K; number of samples u; barrier weight $\mu_b \in (0,1)$
Initialize $\hat{\Lambda}_y^{(0)} = \mathbf{0}$ for each $y \in \mathcal{Y}_\ell$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Sample $x_k \sim \nu$
  Set $g_{k,y} = \frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k) x_k x_k^\top$, where P_Λ is defined in Eq. (7)
  Set $\hat{\Lambda}_y^{(k+1)} \leftarrow \hat{\Lambda}_y^{(k)} + \eta_k g_{k,y}$ for each $y \in \mathcal{Y}_\ell$, where $\eta_k = \frac{1}{\sqrt{2\sum_{s=1}^k \sum_{y\in\mathcal{Y}_\ell} \|g_{s,y}\|_2^2}}$
  Update $\hat{\Lambda}_y^{(k+1)} \leftarrow \Pi_{\mathbb{S}_+^d}(\hat{\Lambda}_y^{(k+1)})$ for each $y \in \mathcal{Y}_\ell$, a projection to the set of $d \times d$ PSD matrices
end for
Let $\hat{\Lambda}_y = \frac{1}{K}\sum_{k=1}^K \hat{\Lambda}_y^{(k)}$ for each $y \in \mathcal{Y}_\ell$ and $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell} \hat{\Lambda}_y$
Update $(\hat{\Lambda}_y)_{y\in\mathcal{Y}_\ell} \leftarrow \operatorname{argmax}_{\mathbf{\Lambda}} \sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y$, subject to $\sum_{y\in\mathcal{Y}_\ell} \Lambda_y = \hat{\Lambda}, \Lambda_y \succeq \mathbf{0}, \forall y \in \mathcal{Y}_\ell$.
Find $s^* \leftarrow \operatorname{argmax}_{s\in[0,1]} D_E(s \cdot \hat{\mathbf{\Lambda}})$, where D_E empirically evaluates D using u i.i.d. samples
return $\widetilde{\Lambda} = s^* \cdot \sum_{y\in\mathcal{Y}_\ell} \hat{\Lambda}_y$

Let $\mathbf{\Lambda}^*$ be an optimal solution for $D(\mathbf{\Lambda})$. Intuitively, as long as we run this algorithm with sufficiently large number of iterations K and number of samples u, we can guarantee that $D(\widetilde{\mathbf{\Lambda}})$ and $D(\mathbf{\Lambda}^*)$ are close enough with high probability, which in turn guarantees that the primal constraints are violated by only a tiny amount and $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$ is close enough to the optimal value. Specifically, we can prove the following theorem.

Theorem 4. *Suppose* $\|x\|_2 \le M$ *for any* $x \in \operatorname{supp}(\nu)$ *and* $\Sigma = \mathbb{E}_{X\sim\nu}\left[XX^\top\right]$ *is invertible. Let* $\mathbf{\Lambda}^* \in \operatorname{argmax}_{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell} D(\mathbf{\Lambda})$ *and* $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *be its condition number. Assume* $\|\Lambda^*\|_F > 0$ *and define* $\omega = \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]$, *where* $\mathbb{S}^d$ *is the set of* $d\times d$ *symmetric matrices. Then,* $\Lambda^* = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y^*$ *is unique. Further, for any* $\epsilon > 0$ *and* $\delta > 0$, *if it holds that* $\mu_b \le O\left(\sqrt{\|\Lambda^*\|_F \kappa(\Sigma) M}\right) \cdot \sqrt{(1+\epsilon)/\epsilon}$ *and*

$$K \ge O\left(\frac{|\mathcal{Y}_\ell|^3 \kappa(\Sigma)^2 \|\Lambda^*\|_F^8 M^{16} \log(1/\delta)}{\omega^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2, u \ge O\left(\frac{\kappa(\Sigma)^2 \|\Lambda^*\|_F^6 M^{16} \log(1/\delta)}{\omega^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

then, with probability at least $1-\delta$, *Algorithm 2 will output* $\widetilde{\Lambda}$ *that satisfies*

- $y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X) XX^\top\right]^{-1} y \le (1+\epsilon)c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu_b}$, *where* $\widetilde{P}$ *is the optimal solution to problem* (4) *with barrier constraint repaced by* $0 \le P(x) \le 1-\mu_b, \forall x \in \mathcal{X}$.

The proof is in Appendix C. Although $\widetilde{P}$ is not exactly the same as the optimal solution of the original problem (4), when μ_b is sufficiently small, they will be very close. Meanwhile, it should be noted that Theorem 4 mainly reveals that with sufficiently large number of iterations and number of samples, Algorithm 2 can output sufficiently good solution. In future work, we plan to examine how much this bound can be improved via a tighter analysis.

Finally, notice that Algorithm 2 needs to maintain $|\mathcal{Y}_\ell|\, d^2 = O(|\mathcal{Z}_\ell|^2\, d^2)$ variables, which can be large when we have a large set $\mathcal{Z}_\ell$. Therefore, as an alternative, we also propose Algorithm 3 that only needs to maintain d^2 variables but requires more computational power in each iteration. The details are given in Appendix C.

5 Empirical results

In this section we present a benchmark experiment validating the fundamental trade-offs that are theoretically characterized in Theorem 1 and Theorem 2. We take inspiration from [24] to define our experimental protocol:

- $d = 2$, a two-dimensional problem.
- $\mathcal{Z} = [\mathbf{e}_1, \mathbf{e}_2, (\cos(\omega), \sin(\omega))]$ for $\omega = 0.3$, where $\mathbf{e}_1, \mathbf{e}_2$ are canonical vectors.
- $\theta_* = 2\mathbf{e}_1$ and $y = x^\top \theta_* + \eta$, where $\eta \sim \mathcal{N}(0, 1)$.
- The distribution ν for streaming measurements $x_t \overset{i.i.d.}{\sim} \nu$ is such that $x_t = (\cos(2I_t\pi/N), \sin(2I_t\pi/N))$ where $I_t \in \{0, \ldots, N-1\}$, $\mathbb{P}(I_t = i) \propto \cos(2i\pi/N)^2$, and $N = 30$.

In this problem, the angle ω is small enough that the item $(\cos(\omega), \sin(\omega))$ is hard to discriminate from the best item $\mathbf{e}_1$. As argued in [24], an efficient sampling strategy for this problem instance would be to pull arms in the direction of $\pm\mathbf{e}_2$ in order to reduce the uncertainty in the direction of interest, $\mathbf{e}_1 - (\cos(\omega), \sin(\omega))$. However, the distribution ν is defined such that it is more likely to receive a vector x_t in the direction of $\pm\mathbf{e}_1$ rather than $\pm\mathbf{e}_2$. Thus, if one seeks a small label complexity, then P should be taken to reject measurements in the direction of $\pm e_1$.

In the benchmark experiment, we compare the following three algorithms which all use Algorithm 1 as a meta-algorithm and just swap out the definition of $\widehat{P}_\ell$. `Naive Algorithm` uses no selective sampling so that $\widehat{P}_\ell(x) = 1$ for all x; the `Oracle Algorithm` uses $\widehat{P}_\ell = P_*$ where P_* is the ideal solution to (2), and `Our Algorithm` uses the solution to (5) for $\widehat{P}_\ell$, where we take $\mu_b = 2 \times 10^{-5}$. We swept over the values of τ and plotted on the y-axis the amount of labeled data needed before termination, as shown in Figure 1.

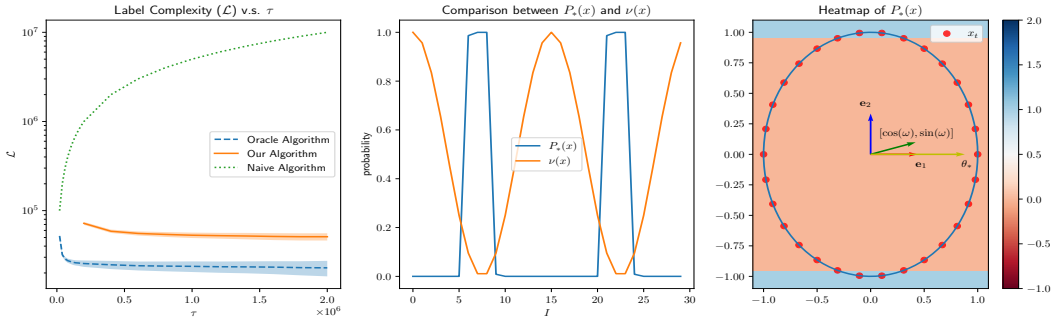

Figure 1: (left) For each value of τ, we plot the average label complexity over 50 repeated trials. (middle) Visualization of $P_*(x)$ and $\nu(x)$ v.s. x, where x is indexed by I such that $x_I = (\cos(2I\pi/N), \sin(2I\pi/N))$. Here, P_* is solved with $\tau = 4 \times 10^5$ and distribution ν is not normalized. (right) A heat map of $P_*(x)$ along with the setting of experimental protocol.

We observe in Figure 1 that the algorithms using non-naive selection rules require far less label complexity than the naive algorithm for all τ. This reflects the intuition that selection strategies that focus on requesting the more informative streaming measurements are much more efficient than naively observing every streaming measurement. Meanwhile, the trade-off between label complexity $\mathcal{L}$ and sample complexity $\mathcal{U}$ characterized in Theorem 1 and Theorem 2 is precisely illustrated in Figure 1. Indeed, we see the number of labels queried by the two selective sampling algorithms decrease as the number of unlabeled data seen in each round increases.

6 Conclusion

In this paper, we proposed a new approach for the important problem of *selective sampling for best arm identification*. We provide a lower bound that quantifies the trade-off between labeled samples and stopping time and also presented an algorithm that nearly achieves the minimal label complexity given a desired stopping time.

One of the main limitations of this work is that our approach depends on a well-specified model following stationary stochastic assumptions. In practice, dependencies over time and model mismatch are common. Utilizing the proposed algorithm outside of our assumptions may lead to poor performance and unexpected behavior with adverse consequences. While negative results justify some of the most critical assumptions we make (e.g., allowing the stream x_t to be arbitrary, rather than iid, can lead to trivial algorithms, see Theorem 7 of [7]), exploring what theoretical guarantees are possible under relaxed assumptions is an important topic of future work.

Acknowledgements

We sincerely thank Chunlin Sun for the insightful discussion on the alternative approach to the optimal design. This work was supported in part by the NSF TRIPODS II grant DMS 2023166, NSF TRIPODS CCF 1740551, NSF CCF 2007036 and NSF TRIPODS+X DMS 1839371.

References

[1] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228. PMLR, 2013.

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

[3] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[4] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[5] Romain Camilleri, Julian Katz-Samuels, and Kevin Jamieson. High-dimensional experimental design and kernel bandits, 2021.

[6] Nicolo Cesa-Bianchi, Claudio Gentile, and Francesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the 26th annual international conference on machine learning*, pages 121–128, 2009.

[7] Yining Chen, Haipeng Luo, Tengyu Ma, and Chicheng Zhang. Active online learning with hidden shifting domains. In *International Conference on Artificial Intelligence and Statistics*, pages 2053–2061. PMLR, 2021.

[8] S. Dasgupta, D. J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 2008.

[9] Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

[10] Reza Eghbali, James Saunderson, and Maryam Fazel. Competitive online algorithms for resource allocation over the positive semidefinite cone. *Mathematical Programming*, 170(1):267–292, 2018.

[11] Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *arXiv preprint arXiv:1906.08399*, 2019.

[12] Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[13] Steve Hanneke and Liu Yang. Toward a general theory of online selective sampling: Trading off mistakes and queries. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2021.

[14] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[15] Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *arXiv preprint arXiv:1506.08669*, 2015.

[16] Julian Katz-Samuels, Jifan Zhang, Lalit Jain, and Kevin Jamieson. Improved algorithms for agnostic pool-based active classification, 2021.

[17] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research, 17:1–42*, 2016.

[18] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[19] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[20] A Nemirovski, A Juditsky, G Lan, and A Shapiro. Stochastic approximation approach to stochastic programming. In *SIAM J. Optim*. Citeseer.

[21] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[22] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.

[23] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[24] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *arXiv preprint arXiv:1409.6110*, 2014.

[25] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011.

[26] Min Xiao and Yuhong Guo. Online active learning for cost sensitive domain adaptation, 2013.

Contents

A Selective Sampling Lower Bound

First, we review the standard argument for best-arm identification lower bounds applied to linear bandits. Fix $\theta_* \in \mathbb{R}^d$ and let $z_* = \arg\max_{z\in\mathcal{Z}} \langle z, \theta_*\rangle$. Define the set $\mathcal{C} = \{\theta \in \mathbb{R}^d : \exists z \in \mathcal{Z} \text{ s.t. } \langle \theta, z - z_*\rangle \geq 0\}$ as those θ in which z_* is note the best arm under θ. We now recall the transportation lemma of [17]. Under a δ-PAC strategy for finding the best arm for the bandit instance $(\mathcal{X}, \mathcal{Z}, \theta_*)$, let T_x denote the random variable which is the number of times arm x is pulled. In addition let $\mathcal{N}_{\theta,x}$ denote the reward distribution of the arm x of $\mathcal{X}$, i.e. $\mathcal{N}_{\theta,x} = \mathcal{N}(x^\top\theta, 1)$. Then for any δ-PAC algorithm

$$\begin{aligned}
\log(1/2.4\delta) &\leq \min_{\theta\in\mathcal{C}} \sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\mathrm{KL}(\mathcal{N}_{\theta_*,x}, \mathcal{N}_{\theta,x}) \\
&= \min_{\theta\in\mathcal{C}} \sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, \tfrac{1}{2}\|\theta_* - \theta\|^2_{xx^\top} \\
&= \min_{\theta\in\mathcal{C}} \tfrac{1}{2}\|\theta_* - \theta\|^2_{(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)} \\
&\leq \min_{z\in\mathcal{Z}\setminus z_*} \tfrac{1}{2}\|\theta_* - \theta_z(\epsilon)\|^2_{(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)}
\end{aligned}$$

where

$$\theta_z(\varepsilon) = \theta_* - \frac{((z_* - z)^\top\theta_* + \varepsilon)(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)^{-1}(z_* - z)^\top}{(z_* - z)^\top(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)^{-1}(z_* - z)}$$

for some small ϵ. This is a valid choice since for all $z \in \mathcal{Z} \setminus z_*$ we have $(z_* - z)^\top\theta_z(\varepsilon) = -\varepsilon < 0$ and thus $\theta_z(\varepsilon) \in \mathcal{C}$. A straightforward calculation shows that

$$\|\theta_* - \theta_z(\epsilon)\|^2_{(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)} = \frac{(\langle z_* - z, \theta_*\rangle + \varepsilon)^2}{\|z_* - z\|^2_{(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)^{-1}}}$$

so that after rearranging and lettering $\epsilon \to 0$ we have that any δ-PAC algorithm satisfies

$$\max_{z\in\mathcal{Z}\setminus z_*} \frac{2\|z_* - z\|^2_{(\sum_{x\in\mathcal{X}} \mathbb{E}[T_x]\, xx^\top)^{-1}}}{\langle z_* - z, \theta_*\rangle^2} \log(1/2.4\delta) \leq 1. \tag{8}$$

This series of steps will be applied for each bullet point of the theorem.

A.1 Proof of Theorem 1, part I

We use the consequence of Lemma 19 of [17]. Consider a δ-PAC algorithm that sets $P(x) = 1$ for all $x \in \mathcal{X}$ for all time until it exits at time $\mathcal{U}$ after this many unlabelled examples have been observed. If T_x denotes the number of times $x \in \mathcal{X}$ was observed before stopping time $\mathcal{U}$, then by Wald's identity we have that

$$\mathbb{E}[T_x] = \mathbb{E}\left[\sum_{t=1}^{\mathcal{U}} \mathbf{1}\{x_t = x\}\right] = \nu(x)\mathbb{E}[\mathcal{U}].$$

Plugging this into Equation 8 and rearranging we conclude that

$$\mathbb{E}[\mathcal{U}] \geq \max_{z\in\mathcal{Z}\setminus z_*} \frac{2\|z_* - z\|^2_{(\sum_{x\in\mathcal{X}} \nu(x)\, xx^\top)^{-1}}}{\langle z_* - z, \theta_*\rangle^2} \log(1/2.4\delta) =: \rho(\nu)\log(1/2.4\delta)$$

which concludes the proof of the first bullet.

A.2 Proof of Theorem 1, part II

By definition, the (random) number of times we measure x is

$$\mathcal{L}_x = \sum_{s=1}^{\mathcal{U}} \mathbf{1}\{x_s = x, Q_s(x) = 1\}$$

and we want to show that $\mathbb{E}[\mathcal{L}_x] = \nu(x)\mathbb{E}\left[\sum_{\ell=1}^{\mathcal{U}} P_\ell(x)\right]$. To do so, we define

$$M_t = \sum_{s=1}^{t} \left(\mathbf{1}\{x_s = x, Q_s(x) = 1\} - \nu(x)P_s(x)\right)$$

It is easy to check that $P_{t+1} \in \mathcal{F}_t := \{(x_s, y_s, Q_s)\}_{s=1}^t$ and that

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = M_t + \mathbb{E}[\mathbf{1}\{x_s = x, Q_s(x) = 1\} - \nu(x)P_s(x)|\mathcal{F}_t] = M_t$$

Applying Doob's equality $\mathbb{E}[M_{\mathcal{U}}] = \mathbb{E}[M_0] = 0$. Consequence:

$$\mathbb{E}[\mathcal{L}_x] = \mathbb{E}\left[\sum_{s=1}^{\mathcal{U}} \mathbf{1}\{x_s = x, Q_s(x) = 1\}\right] = \nu(v)\mathbb{E}\left[\sum_{s=1}^{\mathcal{U}} P_s(x)\right]$$

Define $\alpha(x) := \frac{\mathbb{E}\left[\sum_{s=1}^{\mathcal{U}} P_s(x)\right]}{\mathbb{E}[\mathcal{U}]}$ and note that each $\alpha_x \in [0, 1]$. Then $\mathbb{E}[\mathcal{L}_x] = \mathbb{E}[\mathcal{U}]\alpha(x)\nu(x)$ so applying equation (18) of [17] again, we have

$$\begin{aligned}
\log(1/2.4\delta) &\leq \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] \mathrm{KL}(\mathcal{N}_{\theta_*, x}, \mathcal{N}_{\theta, x}) \\
&= \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] \, \|\theta - \theta_*\|^2_{xx^\top}/2 \\
&= \min_{z \in \mathcal{Z} \setminus z_*} \frac{\langle \theta_*, z_* - z\rangle^2}{2\|z - z_*\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] xx^\top)^{-1}}} \\
&= \min_{z \in \mathcal{Z} \setminus z_*} \frac{\langle \theta_*, z_* - z\rangle^2}{2\|z - z_*\|^2_{(\sum_{x \in \mathcal{X}} \nu(x)\alpha(x) xx^\top)^{-1}}} \mathbb{E}[\mathcal{U}].
\end{aligned}$$

Rearranging, and applying the identity $\mathbb{E}_{X \sim \nu}[\alpha(X)XX^\top] = \sum_{x \in \mathcal{X}} \nu(x)\alpha(x)xx^\top$, the above implies that

$$\mathbb{E}[\mathcal{U}] \geq \max_{z \in \mathcal{Z} \setminus z_*} \frac{2\|z - z_*\|^2_{\mathbb{E}_{X \sim \nu}[\alpha(X)XX^\top]^{-1}}}{\langle \theta_*, z_* - z\rangle^2} \log(1/2.4\delta).$$

Noting that the total expected number of labels is equal to

$$\mathbb{E}[\mathcal{L}] = \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] = \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{U}]\alpha(x)\nu(x) = \mathbb{E}[\mathcal{U}] \, \mathbb{E}_{X \sim \nu}[\alpha(X)]$$

we conclude that

$$\begin{aligned}
\mathbb{E}[\mathcal{L}] \geq \min_{\alpha : \mathcal{X} \to [0,1]} \quad & \mathbb{E}[\mathcal{U}] \, \mathbb{E}_{X \sim \nu}[\alpha(X)] \\
\text{subject to} \quad & \mathbb{E}[\mathcal{U}] \geq \max_{z \in \mathcal{Z} \setminus \{z_*\}} \frac{2\|z - z_*\|^2_{\mathbb{E}_{X \sim \nu}[\alpha(X)XX^\top]^{-1}}}{\langle \theta_*, z_* - z\rangle^2} \log(1/2.4\delta).
\end{aligned}$$

The second bullet point result follows by denoting α as P and applying Proposition 2.

B Selective Sampling Algorithm for Known Distribution ν

B.1 Proof of Theorem 2, upper bound

At each round ℓ we assume an implementation such that $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$ OPTIMIZEDESIGN$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ returns the solution of Equation 3 with $\epsilon = 2^{-\ell}$, essentially. More explicitly, let $\epsilon_\ell := 2^{-\ell}$, $B < \infty$ such that $\max_{x \in \mathcal{X}} |\langle x, \theta_*\rangle| \leq B$, and $\sigma < \infty$ such that $\mathbb{E}[(y_s - \langle \theta_*, x_s\rangle)^2 | x_s] \leq \sigma^2$. If

$$\beta_{\delta,\ell} := 16(B^2 + \sigma^2) \log(2\ell^2|\mathcal{Z}|^2/\delta)$$

then $\widehat{P}_\ell = P_\ell$ where

$$P_\ell := \operatorname*{argmin}_{P : \mathcal{X} \to [0,1]} \mathbb{E}_{X \sim \nu}[P(X)] \text{ subject to } \max_{z, z' \in \mathcal{Z}_\ell} \frac{\|z - z'\|^2_{\mathbb{E}_{X \sim \nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \beta_{\delta,\ell} \leq 1$$

and $\widehat{\Sigma}_{\widehat{P}_\ell} := \mathbb{E}_{X\sim\nu}[P_\ell(X)XX^\top]$

We first provide an intermediate lemma on the correctness of Algorithm 1 that relies on the feasibility of P_ℓ which we will show shortly.

Lemma 1. *With probability at least $1-\delta$ we have for all stages $\ell \in \mathbb{N}$ such that P_ℓ is feasible, that $z_* \in \mathcal{Z}_\ell$ and $\max_{z\in\mathcal{Z}_\ell}\langle z_* - z, \theta_*\rangle \le 4\epsilon_\ell$.*

Proof. Define the event $\mathcal{E}$ as

$$\mathcal{E} := \bigcap_{\ell=1}^{\infty}\bigcap_{z,z'\in\mathcal{Z}_\ell}\left\{|\langle z-z', \widehat{\theta}_\ell - \theta_*\rangle| \le \epsilon_\ell\right\}$$

By Lemma 2, we know that $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. Then, the rest of the proof is the same as the one in [11], but we include it here for completeness. Assume that $\mathcal{E}$ holds. Then for any $z' \in \mathcal{Z}_\ell$

$$\begin{aligned}\langle z'-z^*, \widehat{\theta}_\ell\rangle &= \langle z'-z^*, \widehat{\theta}_\ell - \theta^*\rangle + \langle z'-z^*, \theta^*\rangle \\ &= \langle z'-z^*, \widehat{\theta}_\ell - \theta^*\rangle \\ &\le \epsilon_\ell\end{aligned}$$

so that z^* would survive to round $\mathcal{Z}_{\ell+1}$. And for any $z \in \mathcal{Z}_\ell$ such that $\langle z^* - z, \theta^*\rangle > 2\epsilon_\ell$, we have

$$\begin{aligned}\max_{z'\in\mathcal{Z}_\ell}\langle z'-z, \widehat{\theta}_\ell\rangle &\ge \langle z^*-z, \widehat{\theta}_\ell\rangle \\ &= \langle z^*-z, \widehat{\theta}_\ell - \theta^*\rangle + \langle z^*-z, \theta^*\rangle \\ &> -\epsilon_\ell + 2\epsilon_\ell \\ &= \epsilon_\ell\end{aligned}$$

which implies this z would be kicked out. Note that this implies that $\max_{z\in\mathcal{Z}_{\ell+1}}\langle z^*-z,\theta^*\rangle \le 2\epsilon_\ell = 4\epsilon_{\ell+1}$. □

We can now prove Theorem 2. After $L := \lceil \log_2(\frac{4}{\Delta})\rceil$ rounds $\mathcal{Z}_\ell = \{z_*\}$ by the above lemma. Thus, the total number of labels requested after L rounds is equal to $\mathcal{L} := \sum_{\ell=1}^{L}\sum_{t=(\ell-1)\tau+1}^{\ell\tau} Q_\ell(x_t)$. By Freedman's inequality (c.f., Theorem 1 of [4]) we have that

$$\sum_{\ell=1}^{L}\sum_{t=(\ell-1)\tau+1}^{\ell\tau} Q_\ell(x_t) \le 2\sum_{\ell=1}^{L}\tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell] + \log(1/\delta)$$

We can now bound the expected sample complexity of this algorithm.

$$\begin{aligned}&\sum_{\ell=1}^{L}\tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell] \\ &= \sum_{\ell=1}^{L}\left[\min_{P:\mathcal{X}\to[0,1]}\tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z,z'\in\mathcal{Z}_\ell}\frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1\right].\end{aligned}$$

Using Lemma 3, we have

$$\begin{aligned}\max_{z,z'\in\mathcal{Z}_\ell}\frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} &\le \beta_{\delta,L}\max_{z,z'\in\mathcal{Z}_\ell}\frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \\ &\le 64\beta_{\delta,L}\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2} \\ &=: \max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}\beta_\delta\end{aligned}$$

Note that the last line also describes a condition for which a P_ℓ is feasible. Indeed, at round ℓ, a sufficient condition for a feasible P_ℓ (i.e., the RHS ≤ 1) is if τ exceeds $\rho(\nu)\beta_\delta$ with $\beta_\delta := 1024(B^2 +$

$\sigma^2)\log(2L^2|\mathcal{Z}|^2/\delta)$ and $\rho(\nu) = \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}$, which holds by assumption in the theorem.

Plugging this constraint back into above we have

$$\begin{aligned}
&\sum_{\ell=1}^{L} \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell] \\
&\le \sum_{\ell=1}^{L}\left[\min_{P:\mathcal{X}\to[0,1]} \tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}\beta_\delta \le 1\right] \\
&\le L \min_{\lambda\in\triangle_{\mathcal{X}}} \rho(\lambda)\beta_\delta \quad \text{subject to} \quad \|\lambda/\nu\|_\infty\rho(\lambda)\beta_\delta \le \tau
\end{aligned}$$

where the last line follows by applying the reparameterization of Proposition 2.

B.1.1 High-probability Events

Lemma 2. *We have* $\mathbb{P}(\mathcal{E}) \ge 1-\delta$.

Proof. For any $\mathcal{V}\subseteq\mathcal{Z}$ and $z,z'\in\mathcal{V}$ define

$$\mathcal{E}_{z,z',\ell}(\mathcal{V}) = \{|\langle z-z', \widehat{\theta}_\ell(\mathcal{V})-\theta_*\rangle| \le \epsilon_\ell\}$$

where $\widehat{\theta}_\ell(\mathcal{V})$ is the estimator that would be constructed by the algorithm at stage ℓ with $\mathcal{Z}_\ell = \mathcal{V}$. For fixed $\mathcal{V}\subset\mathcal{Z}$ and $\ell\in\mathbb{N}$ we apply Proposition 1 so that with probability at least $1-\frac{\delta}{\ell^2|\mathcal{Z}|^2}$ we have that for any $z,z'\in\mathcal{V}$

$$\begin{aligned}
|\langle z-z', \widehat{\theta}_\ell(\mathcal{V})-\theta_*\rangle| &\le \|z-z'\|_{\mathbb{E}_{X\sim\nu}[\tau P_\ell(X)XX^\top]^{-1}}\sqrt{16(B^2+\sigma^2)\log(2\ell^2|\mathcal{Z}|^2/\delta)} \\
&\le \epsilon_\ell
\end{aligned}$$

Noting that $\mathcal{E} := \bigcap_{\ell=1}^\infty \bigcap_{z,z'\in\mathcal{Z}_\ell} \mathcal{E}_{z,z',\ell}(\mathcal{Z}_\ell)$ we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\ell=1}^\infty \bigcup_{z,z'\in\mathcal{Z}_\ell} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{Z}_\ell)\}\right) &\le \sum_{\ell=1}^\infty \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{Z}_\ell} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{Z}_\ell)\}\right) \\
&= \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{Z}} \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{V}} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{V})\}, \mathcal{Z}_\ell = \mathcal{V}\right) \\
&= \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{Z}} \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{V}} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{V})\}\right)\mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \\
&\le \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{Z}} \tfrac{\delta}{\ell^2|\mathcal{Z}|^2}\binom{|\mathcal{V}|}{2}\mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \\
&\le \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{Z}} \tfrac{\delta}{2\ell^2}\mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \le \delta
\end{aligned}$$

□

B.2 Technical Lemmas

The following definition characterizes the RIPS estimator we used in Algorithm 1.

Definition 2. *Let $X_1,\dots,X_n$ be i.i.d. random variables with mean $\bar{x}$ and variance ν^2. Let $\delta\in(0,1)$. We say that $\widehat{\mu}(X_1,\dots,X_n)$ is a* δ-robust estimator *if there exist universal constants $c_1, c_0 > 0$ such that if $n \ge c_1\log(1/\delta)$, then with probability at least $1-\delta$*

$$|\widehat{\mu}(\{X_t\}_{t=1}^n) - \bar{x}| \le c_0\sqrt{\frac{\nu^2\log(1/\delta)}{n}}.$$

Examples of δ-robust estimators include the median-of-means estimator and Catoni's estimator [18]. This work employs the use of the Catoni estimator which satisfies $|\widehat{\mu}(\{X_t\}_{t=1}^n) - \bar{x}| \leq \sqrt{\frac{2\nu^2 \log(1/\delta)}{n - 2\log(1/\delta)}}$ for $n > 2\log(1/\delta)$ which leads to an optimal leading constant as $n \to \infty$. See [5] or [18] for more details.

Proposition 1. *Let $x_1, \ldots, x_n$ be drawn IID from a distribution ν. Assume that $|\langle \theta, x_s \rangle| \leq B$ and $\mathbb{E}[|\langle \theta, x_s \rangle - y_s|^2] \leq \sigma^2$. Let $P : \mathcal{X} \to [0,1]$ be arbitrary. Let $Q(x_s) \sim$ Bernoulli$(P(x_s))$ independently for all $s \in [n]$. For a given finite set $\mathcal{V} \subset \mathbb{R}^d$ define for any $v \in \mathcal{V}$*

$$w_v = \mathrm{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s \rangle\}_{s=1}^n).$$

If $\widehat{\theta} = \arg\min_\theta \max_v \frac{|w_v - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}$ and $n \geq 4\log(2|\mathcal{V}|/\delta)$, then with probability at least $1-\delta$, it holds that

$$|\langle v, \widehat{\theta} - \theta \rangle| \leq \|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}} \sqrt{16(B^2+\sigma^2)\log(2|\mathcal{V}|/\delta)}$$

Proof. Inspired by [5], we note that

$$\begin{aligned}
\max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - \langle \theta, v\rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} &= \max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - w_v + w_v - \langle \theta, v\rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&\leq \max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} + \max_{v\in\mathcal{V}} \frac{|w_v - \langle \theta, v\rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&= \min_\theta \max_{v\in\mathcal{V}} \frac{|\langle \theta, v\rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} + \max_{v\in\mathcal{V}} \frac{|w_v - \langle \theta, v\rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&\leq 2\max_{v\in\mathcal{V}} \frac{|\langle \theta, v\rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}}
\end{aligned}$$

So it suffices to show that each $|\langle \theta, v\rangle - w_v|$ is small. We begin by fixing some $v \in \mathcal{V}$ and bounding the variance of $v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s$ for any $s \leq n$ which is necessary to use the robust estimator. For readability purposes, we shorten $\mathbb{E}_{x_s\sim\nu, Q(x_s)\sim P(x_s)}$ as $\mathbb{E}_{x_s,Q}$ in the rest of this proof. Note that

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}_{x_s\sim\nu, Q(x_s)\sim P(x_s)}(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s) \\
=&\mathbb{E}_{x_s,Q}[(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s)^2] \\
&- \mathbb{E}_{x_s,Q}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s]^2
\end{aligned}$$

which means we can drop the second term to bound the variance by

$$\begin{aligned}
&\mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s\right)^2] \\
&= \mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s (x_s^\top \theta + \xi_s)\right)^2] \\
&= \mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s (x_s^\top \theta)\right)^2] \\
&\quad + \mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s\right)^2 \xi_t^2] \\
&\leq B^2 \mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s\right)^2] \\
&\quad + \sigma^2 \mathbb{E}_{x_s,Q}[\left(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s\right)^2] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q(x_s)\sim P(x_s)}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s x_s^\top Q(x_s) \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} v]\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q(x_s)\sim P(x_s)}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s x_s^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} v]\right] \\
&\leq \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2) v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} P(x_s) x_s x_s^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} v]\right],
\end{aligned}$$

where we used that $Q(x_s)^2 = Q(x_s)$ in equality (i) above. Thus, we have

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s) \\
\leq&(B^2+\sigma^2) v^\top (\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} \mathbb{E}_{x_s\sim\nu}[P(x_s) x_s x_s^\top] (\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}) v \\
=&(B^2+\sigma^2)\|v\|^2_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}
\end{aligned}$$

By using the property of Catoni estimator stated in Definition 2, we have $c_0 = \sqrt{2}$ and

$$
\begin{aligned}
&|\langle \theta_*, v\rangle - w_v| \\
=&|\text{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle\}_{s=1}^n) - \mathbb{E}[\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle]| \\
\leq&\sqrt{2}\sqrt{(\mathbb{V}\text{ar}(\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle))\,\frac{\log(\frac{2}{\delta})}{n/2}} \\
&\hspace{6cm}\text{(with probability at least } 1-\delta \text{ if } n \geq 4\log(2/\delta)) \\
\leq&\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}\sqrt{\frac{4(B^2+\sigma^2)\log(\frac{2}{\delta})}{n}} \\
=&\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}\sqrt{4(B^2+\sigma^2)\log(2/\delta)}.
\end{aligned}
$$

Finally, the proof is complete by taking union bounding over all $v \in \mathcal{V}$. □

Lemma 3. *Holds*

$$
\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \leq 64 \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*, \theta_*\rangle^2}
$$

Proof. Let $\mathcal{S}_\ell = \{z \in \mathcal{Z} : \langle z_* - z, \theta_*\rangle \leq 4\epsilon_\ell\}$. We have

$$
\begin{aligned}
\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} &\leq \max_{z,z'\in\mathcal{S}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \\
&= 16 \max_{z,z'\in\mathcal{S}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{(4\epsilon_\ell)^2} \\
&\leq 64 \max_{z\in\mathcal{S}_\ell} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{(4\epsilon_\ell)^2} \\
&= 64 \max_{z\in\mathcal{S}_\ell\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\max\{(4\epsilon_\ell)^2, \langle z-z_*, \theta_*\rangle^2\}} \\
&\leq 64 \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*, \theta_*\rangle^2}.
\end{aligned}
$$

□

B.2.1 Reparameterization

Proposition 2. *Fix $\nu \in \triangle_{\mathcal{X}}$ and any $\lambda \in \triangle_{\mathcal{X}}$. Define $\|\lambda/\nu\|_\infty = \sup_{x\in\mathcal{X}} \lambda(x)/\nu(x)$ and $\rho(\lambda) = \max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\langle z_*-z, \theta_*\rangle^2}$. For any $t, \beta \in \mathbb{R}_+$ the following optimization problems achieve the same value*

- $\min_{P:\mathcal{X}\to[0,1]} t\,\mathbb{E}_{X\sim\nu}[P(X)]$ *subject to* $\max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}{\langle z_*-z, \theta_*\rangle^2}\beta \leq t$
- $\min_{\lambda\in\triangle_{\mathcal{X}}} \rho(\lambda)\beta$ *subject to* $\|\lambda/\nu\|_\infty \rho(\lambda)\beta \leq t$

Let us first prove a simple lemma.

Lemma 4. *Let $\mathcal{P}$ denote the set of all functions $P : \mathcal{X} \to [0,1]$. And for any $\nu \in \triangle_{\mathcal{X}}$ with support $\mathcal{X}$ let $\mathcal{P}' = \{\kappa\lambda_x/\nu_x : \lambda \in \triangle_{\mathcal{X}}, \kappa \geq 0 : \kappa\lambda_x/\nu_x \in [0,1]\}$. Then $\mathcal{P} = \mathcal{P}'$.*

Proof. Fix any $P \in \mathcal{P}$. If $\lambda_x = P_x\nu_x/\|P \circ \nu\|_1$ and $\kappa = \|P \circ \nu\|_1$ then $\kappa\lambda/\nu \in \mathcal{P}'$ and is equal to P. This implies $\mathcal{P} \subseteq \mathcal{P}'$.

For the other direction, fix any $\lambda \in \triangle_{\mathcal{X}}$ and $\kappa \geq 0$ such that $\kappa\lambda_x/\nu_x \in [0,1]$ for all x. If $P = \kappa\lambda/\nu$ then $P \in \mathcal{P}$ which implies $\mathcal{P}' \subseteq \mathcal{P}$ and concludes the proof. □

Proof of Proposition 2. Using the above lemma we have that

$$\min_{P:\mathcal{X}\to[0,1]} t\,\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}{\langle z_*-z,\theta_*\rangle^2}\beta \le t$$

is equivalent to

$$\min_{\kappa\ge 0,\lambda\in\triangle_{\mathcal{X}}} t\,\mathbb{E}_{X\sim\nu}[\kappa\lambda(X)/\nu(X)] \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\kappa\lambda(X)/\nu(X)XX^\top]^{-1}}}{\langle z_*-z,\theta_*\rangle^2}\beta \le t$$
$$\kappa\lambda(x)/\nu(x)\le 1 \quad \forall x\in\mathcal{X}$$

which is equal to, after simplifying,

$$\min_{\kappa\ge 0,\lambda\in\triangle_{\mathcal{X}}} t\,\kappa \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\langle z_*-z,\theta_*\rangle^2}\beta \le t\kappa$$
$$\kappa\lambda(x)/\nu(x)\le 1 \quad \forall x\in\mathcal{X}$$

which is equal to

$$\min_{u\ge 0,\lambda\in\triangle_{\mathcal{X}}} u \quad \text{subject to} \quad \rho(\lambda)\beta \le u$$
$$\|\lambda/\nu\|_\infty \le \frac{t}{u}.$$

Note, there exists a feasible (λ,u) precisely when there exists a $\lambda\in\triangle_{\mathcal{X}}$ such that $\|\lambda/\nu\|_\infty\rho(\lambda)\le t$, in which case the optimization problem is equal to

$$\min_{\lambda\in\triangle_{\mathcal{X}}} \rho(\lambda)\beta \quad \text{subject to} \quad \|\lambda/\nu\|_\infty\rho(\lambda)\beta \le t$$

□

C Analysis of the Optimization Problem

C.1 Proof of Theorem 4

For simplicity, we will use μ instead of μ_b to denote the number that controls the intensity of barrier function.

The proof relies on analyzing another function $\overline{D}:\mathbb{R}^{d\times d}_{\succeq\mathbf{0}}\mapsto\mathbb{R}$. For simplicity, first, we define

$$h_\Lambda(x) = P_\Lambda(x) - \mu\left(\log(1-P_\Lambda(x)) + \log(P_\Lambda(x))\right) - P_\Lambda(x)x^\top\Lambda x. \tag{9}$$

Recall that our dual objective is $D(\mathbf{\Lambda}) = \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top\Lambda_y y$. Since the first term in $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ only depends on $\Lambda=\sum_{y\in\mathcal{Y}_\ell}\Lambda_y$, we can consider the following optimization problem.

$$\begin{array}{rl} f(\Lambda) = \max_{\Lambda_y} & \sum_{y\in\mathcal{Y}_\ell} y^\top\Lambda_y y \\ \text{subject to} & \sum_{y\in\mathcal{Y}_\ell}\Lambda_y = \Lambda \\ & \Lambda_y\succeq\mathbf{0}, \quad \forall y\in\mathcal{Y}_\ell. \end{array} \tag{10}$$

Then, the alternative dual objective $\overline{D}(\Lambda)$ is defined as $\overline{D}(\Lambda) = \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] + \frac{1}{c_\ell^2}f(\Lambda)$. We can immediately see that maximizing $\overline{D}(\cdot)$ is equivalent to maximizing $D(\cdot)$. In particular, let $\Lambda^*\in\operatorname{argmax}_{\Lambda\succeq\mathbf{0}}\overline{D}(\Lambda)$ and $\left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$ be the set of PSD matrices that solve problem (10) and evaluate $f(\Lambda^*)$. We can see that $\left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$ also maximizes $D(\cdot)$. Conversely, for $\mathbf{\Lambda}^*=\left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}\in\operatorname{argmax}_{\Lambda_y\succeq\mathbf{0},\forall y} D(\mathbf{\Lambda})$, we also have $\sum_{y\in\mathcal{Y}_\ell}\Lambda_y^*\in\operatorname{argmax}_{\Lambda\succeq\mathbf{0}}\overline{D}(\Lambda)$.

Further, we also define their empirical version D_E and $\overline{D}_E$ with extra i.i.d. samples $x_1,\ldots,x_u$ as

$$D_E(\mathbf{\Lambda}) = \frac{1}{u}\sum_{i=1}^u h_\Lambda(x_i) + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top\Lambda_y y \quad \text{and} \quad \overline{D}_E(\Lambda) = \frac{1}{u}\sum_{i=1}^u h_\Lambda(x_i) + \frac{1}{c_\ell^2}f(\Lambda). \tag{11}$$

Recall that the problem Algorithm 2 tries to solve is

$$\begin{array}{rl} \min_P & \mathbb{E}_{X\sim\nu}[P(X) - \mu(\log(1-P(X)) + \log(P(X)))] \\ \text{subject to} & \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{1}{c_\ell^2} yy^\top, \quad \forall y \in \mathcal{Y}_\ell. \end{array} \tag{12}$$

We will restate a more precise version of Theorem 4 and then prove it.

Theorem 5. *Suppose $\|x\|_2 \le M$ for any $x \in \operatorname{supp}(\nu)$ and $\Sigma = \mathbb{E}_{X\sim\nu}\left[XX^\top\right]$ is invertible. Let $\mathbf{\Lambda}^* \in \operatorname{argmax}_{\Lambda_y \succeq \mathbf{0}, \forall y} D(\mathbf{\Lambda})$ and $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ be condition number. Assume $\|\Lambda^*\|_F > 0$ and define $\omega = \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top \Gamma X\right)^2\right]$, where $\mathbb{S}^d$ is the set of $d \times d$ symmetric matrices. Let $|\mathcal{Y}_\ell|\, C_\ell^2 = \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\|y\|_2^4$.*

Then, $\Lambda^* = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y^*$ is unique. Further, for any $\epsilon > 0$ and $\delta > 0$, suppose it holds that

$$\begin{aligned} \mu &\le \min\left\{\sqrt{\frac{3\kappa(\Sigma)\,\|\Lambda^*\|_F\, M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}}, \frac{4}{9}\|\Lambda^*\|_F^2\, M^4, \frac{1}{2\sqrt{3}}\right\} \\ K &\ge \frac{288\kappa(\Sigma)^2\,|\mathcal{Y}_\ell|^3\,\|\Lambda^*\|_F^4\, M^4(M^4+C_\ell^2)\cdot\left(2\,\|\Lambda^*\|_F\, M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2 \\ u &\ge \frac{576\kappa(\Sigma)^2\,\|\Lambda^*\|_F^2\, M^8\cdot\left(2\,\|\Lambda^*\|_F\, M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2. \end{aligned}$$

Then, with probability at least $1-\delta$, Algorithm 2 will output $\widetilde{\Lambda}$ that satisfies

- $y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1} y \le (1+\epsilon)c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, *where $\widetilde{P}$ is the optimal solution to problem* (20).

Proof. **First Bullet Point.** Fix some $\epsilon > 0$. Let $\hat{\mathbf{\Lambda}}$ and corresponding $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$ be the parameters obtained by Algorithm 2 just before the re-scaling step, which means that at line 10 of Algorithm 2, the assignment of $\hat{\Lambda}_y$ to each $y \in \mathcal{Y}_\ell$ has been optimized by solving problem (10). That is, we have $D(\hat{\mathbf{\Lambda}}) = \overline{D}(\hat{\Lambda})$ and $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$. Let $\widetilde{\mathbf{\Lambda}}$ and $\widetilde{\Lambda}$ be the ones after the re-scaling step. Then, by Theorem 3.13 of [21], with probability at least $1 - \frac{\delta}{3}$, it holds that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) = D(\mathbf{\Lambda}^*) - D(\hat{\mathbf{\Lambda}}) \le \frac{\operatorname{Reg}(K) + 2\sqrt{2K\log(6/\delta)}}{K},$$

where $\operatorname{Reg}(K)$ is the regret of running projected stochastic gradient ascent for K steps with η_k specified in Algorithm 2. Meanwhile, by Theorem 4.14 of [21] also, we have $\operatorname{Reg}(K) = \sqrt{2}B^2\sqrt{\sum_{k=1}^K\sum_{y\in\mathcal{Y}_\ell}\|g_{k,y}\|_2^2}$, where $B = \sqrt{|\mathcal{Y}_\ell|}\,\|\Lambda^*\|_F$ bound the norm of $\mathbf{\Lambda}^* = \left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$. Since $g_{k,y} = \frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_kx_k^\top$, we can easily get $\sum_{y\in\mathcal{Y}_\ell}\|g_{k,y}\|_2^2 \le 2\,|\mathcal{Y}_\ell|\,M^4 + \frac{2}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\|y\|_2^4 = 2\,|\mathcal{Y}_\ell|\,M^4 + 2\,|\mathcal{Y}_\ell|\,C_\ell^2$. Thus, we have

$$\operatorname{Reg}(K) \le 2\,|\mathcal{Y}_\ell|\,\|\Lambda^*\|_F^2\sqrt{|\mathcal{Y}_\ell|\,M^4 + |\mathcal{Y}_\ell|\,C_\ell^2}\cdot\sqrt{K} := C_{\operatorname{Reg}}\sqrt{K} \tag{13}$$

$$\implies \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{C_{\operatorname{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}}, \tag{14}$$

We now consider the effect of using u i.i.d. samples in the re-scaling step. First, since re-scaling always increases the function value, we must have $D_E(\hat{\mathbf{\Lambda}}) \le D_E(\widetilde{\mathbf{\Lambda}})$. Meanwhile, since $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$, by Lemma 10, we have $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$, which together implies $\overline{D}_E(\hat{\Lambda}) \le \overline{D}_E(\widetilde{\Lambda})$.

By Lemma 5, we know that Λ^* is unique and as long as $\mu \le \frac{1}{2\sqrt{3}}$, $\overline{D}(\Lambda)$ is G-strongly concave with respect to ℓ_2 norm over $\mathcal{S} = \{\Lambda \succeq \mathbf{0} : \|\Lambda\|_F \le 2\|\Lambda^*\|_F\}$, where G is defined in Eq. (21). Thus, by Lemma 11, if K is large enough such that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}} \le \frac{G\|\Lambda^*\|_F}{2},$$

then $\left\|\hat{\Lambda} - \Lambda^*\right\|_F \le \|\Lambda^*\|_F$, which implies $\left\|\hat{\Lambda}\right\|_F \le 2\|\Lambda^*\|_F$. That is, $\hat{\Lambda} \in \mathcal{S}$. Then, under this condition, by using Lemma 8, when $\mu \le \frac{4}{9}\|\Lambda^*\|_F M^4$ and

$$u \ge \left(\frac{6\kappa(\Sigma)\|\Lambda^*\|_F M^4\left(2+\sqrt{2\log(6/\delta)}\right)}{G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2, \tag{15}$$

for $\widetilde{\Lambda}$ after re-scaling, with probability at least $1-\frac{\delta}{3}$, it holds simultaneously that

$$\left|\overline{D}_E(\hat{\Lambda}) - \overline{D}(\hat{\Lambda})\right| \le \frac{G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \quad \text{and} \quad \left|\overline{D}_E(\widetilde{\Lambda}) - \overline{D}(\widetilde{\Lambda})\right| \le \frac{G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \tag{16}$$

$$\begin{aligned}
\implies \overline{D}(\Lambda^*) - \overline{D}(\widetilde{\Lambda}) &\le \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) + \overline{D}(\hat{\Lambda}) - \overline{D}(\widetilde{\Lambda}) \\
&\le \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) + \overline{D}(\hat{\Lambda}) - \overline{D}_E(\hat{\Lambda}) + \overline{D}_E(\widetilde{\Lambda}) - \overline{D}(\widetilde{\Lambda}) \\
&\qquad \text{(Since } \overline{D}_E(\hat{\Lambda}) \le \overline{D}_E(\widetilde{\Lambda})\text{)} \\
&\le \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}} + \frac{2G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}. \quad \text{(By Eq. (14) and (16))}
\end{aligned}$$

Since $\widetilde{\Lambda}$ is a smaller re-scaling of $\hat{\Lambda}$, we have $\widetilde{\Lambda} \in \mathcal{S}$, which implies $\frac{G}{2}\left\|\Lambda^* - \widetilde{\Lambda}\right\|_F \le \overline{D}(\Lambda^*) - \overline{D}(\widetilde{\Lambda})$ by property of strongly concave function [3]. Therefore, by Lemma 12, to guarantee an at most ϵ multiplicative constraint violation, it is sufficient to choose K such that

$$\begin{aligned}
\frac{G}{2}\left\|\Lambda^* - \widetilde{\Lambda}\right\|_F &\le \overline{D}(\Lambda^*) - \overline{D}(\widetilde{\Lambda}) \\
&\le \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}} + \frac{2G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \\
&\le \min\left\{\frac{4G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}, \frac{G\|\Lambda^*\|_F}{2}\right\} \\
&= \frac{4G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}. \qquad \text{(If } \mu \le \sqrt{\tfrac{3\kappa(\Sigma)\|\Lambda^*\|_F M^2}{8}\cdot\tfrac{1+\epsilon}{\epsilon}}\text{)}
\end{aligned}$$

An algebraic rearrangement gives us

$$K \ge \left(\frac{3\kappa(\Sigma)M^2\left(C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}\right)}{2G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2. \tag{17}$$

Second Bullet Point. We then prove the upper bound for primal objective value $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$, which explains the reason why an extra re-scaling step is needed. Define $g(s) = D_E(s\cdot\widetilde{\mathbf{\Lambda}})$. By construction, we know that $g(s)$ is maximized at $s=1$ because $\widetilde{\mathbf{\Lambda}} = s^*\cdot\hat{\mathbf{\Lambda}}$, where $s^* = \mathrm{argmax}_{s\in[0,1]} D_E(s\cdot\hat{\mathbf{\Lambda}})$. Therefore, we have $g'(1) \ge 0$, which in turn gives us

$$g'(1) = \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top\widetilde{\Lambda}_y y - \frac{1}{u}\sum_{i=1}^{u} P_{\widetilde{\Lambda}}(x_i)x_i^\top\widetilde{\Lambda}x_i \ge 0.$$

By the concentration inequality in Lemma 8, we know that when

$$u \ge \left(\frac{2\|\Lambda^*\|_F M^2\left(\|\Lambda^*\|_F M^2 + \mu\sqrt{2\log(6/\delta)}\right)}{\mu^{3/2}}\right)^2, \tag{18}$$

with probability at least $1-\frac{\delta}{3}$, it holds that

$$\left|\mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)X^\top\Lambda X\right]-\frac{1}{u}\sum_{i=1}^{u}P_\Lambda(x_i)x_i^\top\Lambda x_i\right|\le\sqrt{\mu}$$
$$\implies \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}y^\top\widetilde{\Lambda}_y y-\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)X^\top\widetilde{\Lambda}X\right]+\sqrt{\mu}\ge 0. \tag{19}$$

Now, let $\widetilde{P}$ be the optimal solution of problem (20) and $\hat{P}$ be the optimal solution of the same problem with bound constraint $\mu\le P(x)\le 1-\mu$.

$$\begin{aligned}\min_P\quad & \mathbb{E}_{X\sim\nu}\left[P(X)\right]\\ \text{subject to}\quad & y^\top\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1}y\le c_\ell^2,\quad \forall y\in\mathcal{Y}_\ell,\\ & 0\le P(x)\le 1-\mu,\quad \forall x\in\mathcal{X}.\end{aligned} \tag{20}$$

Then, we can notice that

$$\begin{aligned}&\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]\\ \le&\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)-\mu(\log(1-P_{\widetilde{\Lambda}}(X))+\log(P_{\widetilde{\Lambda}}(X)))\right]\\ \le&\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)-\mu(\log(1-P_{\widetilde{\Lambda}}(X))+\log(P_{\widetilde{\Lambda}}(X)))\right]\\ &\quad+\frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}y^\top\widetilde{\Lambda}_y y-\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)X^\top\widetilde{\Lambda}X\right]+\sqrt{\mu} && \text{(By Eq. (19))}\\ =&\inf_P\mathcal{L}(P,\widetilde{\mathbf{\Lambda}})+\sqrt{\mu} && \text{(By definition of Lagrangian function and how we solve for } P_\Lambda)\\ \le&\max_{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell}\inf_P\mathcal{L}\left(P,\mathbf{\Lambda}\right)+\sqrt{\mu}\\ =&\mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)-\mu(\log(1-P_{\Lambda^*}(X))+\log(P_{\Lambda^*}(X)))\right]+\sqrt{\mu}\\ \le&\mathbb{E}_{X\sim\nu}\left[\hat{P}(X)-\mu\log(1-\hat{P}(X))\right]-\mu\log\left(\hat{P}(X)\right)+\sqrt{\mu}\\ &&\text{(Since } \hat{P} \text{ is feasible to problem (12))}\\ \le&\mathbb{E}_{X\sim\nu}\left[\hat{P}(X)\right]+3\sqrt{\mu}, && \text{(Since } -a\log(a)\le\sqrt{a} \text{ for } a\in(0,1))\\ \le&\mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right]+4\sqrt{\mu}. && \text{(Since } \hat{P}(x) \text{ can have at most } \mu \text{ more contribution than } \widetilde{P})\end{aligned}$$

Therefore, in summary, Suppose K and u satisfy conditions specified in Eq. (17), (15) and (18) and $\mu\le\min\left\{\sqrt{\frac{3\kappa(\Sigma)\|\Lambda^*\|_F M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}},\frac{4}{9}\|\Lambda^*\|_F^2 M^4,\frac{1}{2\sqrt{3}}\right\}$, where C_{Reg} and G are defined in Eq. (13) and (21), respectively. Then. by applying a simple union bound, with probability at least $1-\delta$, the output of Algorithm 2 $\widetilde{\Lambda}$ satisfies $y^\top\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1}y\le(1+\epsilon)c_\ell^2,\forall y\in\mathcal{Y}_\ell$ and $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]\le\mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right]+4\sqrt{\mu}$. □

C.2 Relevant Lemmas

C.2.1 Strong Concavity of $\overline{D}(\Lambda)$

Lemma 5. *As long as $\mu\le\frac{1}{2\sqrt{3}}$, $\overline{D}(\Lambda)$ is G-strongly concave with respect to ℓ_2-norm on the bounded region $\mathcal{S}=\{\Lambda\succeq\mathbf{0}:\|\Lambda\|_F\le 2\|\Lambda^*\|_F\}$ with coefficient*

$$G=\frac{\mu}{2\left(2\|\Lambda^*\|_F M^2+1\right)^2}\cdot\min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1}\mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]. \tag{21}$$

Because of this, as a corollary, Λ^* will be unique.

Proof. By Lemma 6, since $f(\Lambda)$ is concave in Λ, it is sufficient to prove that $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ is G-strongly concave on $\mathcal{S}$, where $h_\Lambda(x)$ is defined in Eq. (9). Then, we have

$$-\nabla_\Lambda^2\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]=\mathbb{E}_{X\sim\nu}\left[\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}(X)\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right].$$

Since $\|x\|_2 \le M$, for any $\Lambda \in \mathcal{S}$, we have $q_\Lambda(x) = x^\top \Lambda x - 1 \le 2\|\Lambda^*\|_F M^2 + 1$. By Lemma, 14, we know that if $12\mu^2 \le \left(2\|\Lambda^*\|_F M^2 + 1\right)^2$, which can be done by choosing $\mu \le \frac{1}{2\sqrt{3}}$, we have $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}(x) \ge \frac{\mu}{2\left(2\|\Lambda^*\|_F M^2+1\right)^2}$ for any $x \in \mathcal{X}$ and $\Lambda \in \mathcal{S}$. Therefore, we have

$$-\nabla^2_\Lambda \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] \succeq \gamma \cdot \mathbb{E}_{X\sim\nu}\left[\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right]$$

Now, let $\mathbb{S}$ be the set of all $d \times d$ symmetric matrices. It is obvious that $\mathbb{S}$ is a subspace of the vector space of all $d \times d$ matrices and $\mathcal{S} \subseteq \mathbb{S}$. Thus, by applying Lemma 7, we can conclude that $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ is G-strongly concave on $\mathcal{S}$ with respect to ℓ_2 norm and

$$\begin{aligned} G &= \frac{\mu}{2\left(2\|\Lambda^*\|_F M^2 + 1\right)^2} \cdot \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathrm{vec}(\Gamma)^\top \mathbb{E}_{X\sim\nu}\left[\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right]\mathrm{vec}(\Gamma) \\ &= \frac{\mu}{2\left(2\|\Lambda^*\|_F M^2 + 1\right)^2} \cdot \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]. \end{aligned}$$

Thus the proof is complete. □

Lemma 6. *$f(\Lambda)$ defined in Eq. (10) is concave in Λ.*

Proof. To show its concavity, consider $\Lambda^{(1)} \succeq \mathbf{0}$, $\Lambda^{(2)} \succeq \mathbf{0}$ and some $\gamma \in (0,1)$. Let $(\Lambda_y^{(i)})_{y\in\mathcal{Y}_\ell}$ be the optimal solution obtained by evaluating $f(\Lambda^{(i)})$ for $i \in \{1,2\}$. Then, we can notice that

$$\begin{aligned} \gamma f(\Lambda^{(1)}) + (1-\gamma) f(\Lambda^{(2)}) &= \gamma \sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y^{(1)} y + (1-\gamma)\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y^{(2)} y \\ &= \sum_{y\in\mathcal{Y}_\ell} y^\top(\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)})y \\ &\le f(\gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}). \end{aligned}$$

The last inequality above holds because $\sum_{y\in\mathcal{Y}_\ell}\Lambda_y^{(i)} = \Lambda^{(i)}$ for $i \in \{1,2\}$ and thus $\sum_{y\in\mathcal{Y}_\ell}\left(\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)}\right) = \gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}$, which means that $(\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)})_{y\in\mathcal{Y}_\ell}$ is a feasible solution for problem (10) with parameter $\gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}$. Therefore, we can conclude that $f(\Lambda)$ is concave in Λ. □

Lemma 7. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex and twice differentiable function in $\mathbb{R}^d$. If for some subspace $S \subseteq \mathbb{R}^d$, we have $\min_{w\in S:\|w\|_2=1} w^\top \nabla^2 f(x) w \ge \sigma > 0$, $\forall x \in S$, then f is σ-strongly convex with respect to ℓ_2-norm on S.*

Proof. Suppose S has dimension m and let $v_1, \ldots, v_m$ be a set of orthonormal basis that span S. Then, for each $x \in S$, there exists unique $z \in \mathbb{R}^m$ such that $x = Vz$, where $V = [v_1 \quad \ldots \quad v_m]$. That is, there is one-to-one correspondence between S and $\mathbb{R}^m$.

Now, we define $g : \mathbb{R}^m \mapsto \mathbb{R}$ as $g(z) = f(Vz)$. It is easy to compute $\nabla^2 g(z) = V^\top \nabla^2 f(Vz) V$. Then, notice that for any $w' \in \mathbb{R}^m$ such that $\|w'\|_2 = 1$, we have $Vw' \in S$ and $\|Vw'\|_2 = \sqrt{w'^\top V^\top V w'} = \sqrt{w'^\top w'} = 1$. Thus, we have

$$\begin{aligned} \min_{w'\in\mathbb{R}^m:\|w'\|_2=1} w'^\top \nabla^2 g(z) w' &= \min_{w'\in\mathbb{R}^m:\|w'\|_2=1} w'^\top V^\top \nabla^2 f(Vz) V w' \\ &= \min_{w\in S:\|w\|_2=1} w^\top \nabla^2 f(Vz) w \ge \sigma. \end{aligned}$$

Therefore, g is σ-strongly convex with respect to ℓ_2 norm. Then, for any $x_1, x_2 \in S$, there exists unique $z_1, z_2 \in \mathbb{R}^m$ such that $x_1 = Vz_1$ and $x_2 = Vz_2$. Notice that $\|z_1 - z_2\|_2 = \|x_1 - x_2\|_2$ since V preserves the norm. Further, by definition of strong convexity, for any $\alpha \in [0,1]$, we have

$$\begin{aligned} & g(\alpha z_1 + (1-\alpha)z_2) + \frac{\sigma}{2}\alpha(1-\alpha)\|z_1 - z_2\|_2^2 \le \alpha g(z_1) + (1-\alpha) g(z_2) \\ \implies & f(\alpha V z_1 + (1-\alpha) V z_2) + \frac{\sigma}{2}\alpha(1-\alpha)\|x_1 - x_2\|_2^2 \le \alpha f(Vz_1) + (1-\alpha) f(Vz_2) \\ \implies & f(\alpha x_1 + (1-\alpha)x_2) + \frac{\sigma}{2}\alpha(1-\alpha)\|x_1 - x_2\|_2^2 \le \alpha f(x_1) + (1-\alpha) f(x_2). \end{aligned}$$

Thus, f is also σ-strongly convex with respect to ℓ_2 norm on S. □

C.2.2 Concentration Inequalities

Lemma 8. *Let $x_1, \ldots, x_u \sim \nu$ be i.i.d. samples. If $\left\|\hat{\Lambda}\right\|_F \leq 2\left\|\Lambda^*\right\|_F$, $\|x\|_2 \leq M$ for any $x \in \mathcal{X}$ and $\mu \leq \frac{4}{9}\left\|\Lambda^*\right\|_F^2 M^4$, then with probability at least $1-\frac{2\delta}{3}$, it holds for any $\Lambda \in \Theta = \left\{s \cdot \hat{\Lambda} : s \in [0,1]\right\}$ simultaneously that*

$$\left|\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] - \frac{1}{u}\sum_{i=1}^{u} h_\Lambda(x_i)\right| \leq \frac{2\left\|\Lambda^*\right\|_F M^2\left(2+\sqrt{2\log(6/\delta)}\right)}{\sqrt{u}}$$

$$\left|\mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)X^\top \Lambda X\right] - \frac{1}{u}\sum_{i=1}^{u} P_\Lambda(x_i)x_i^\top \Lambda x_i\right| \leq \frac{2\left\|\Lambda^*\right\|_F M^2\left(\left\|\Lambda^*\right\|_F M^2 + \mu\sqrt{2\log(6/\delta)}\right)}{\mu\sqrt{u}}.$$

Proof. To prove the first inequality, first, notice that we have $h_\Lambda(x) = -P_\Lambda(x)q_\Lambda(x) - \mu\left(\log(1-P_\Lambda(x)) + \log(P_\Lambda(x))\right)$, where $q_\Lambda(x) = x^\top\Lambda x - 1$. Since $P_\Lambda(x)$, defined in Eq. (7), explicitly only depends on $q_\Lambda(x)$ instead of x directly, we can treat h_Λ as a function of q_Λ and define a function class $\mathcal{F} = \left\{x \mapsto x^\top(s\cdot\hat{\Lambda})x : s \in [0,1]\right\}$. It is well-known that if h_Λ is L_1-Lipschitz in q_Λ and $|h_\Lambda(x)| \leq R_1$ for any $\Lambda \in \Theta$ and $x \sim \nu$, then, with probability at least $1-\frac{\delta}{3}$, it holds simultaneously for all $\Lambda \in \Theta$ that [2, 19]

$$\left|\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] - \frac{1}{u}\sum_{i=1}^{u} h_\Lambda(x_i)\right| \leq 2L_1\cdot\mathcal{R}_u(\mathcal{F}) + R_1\sqrt{\frac{2\log(6/\delta)}{u}}, \tag{22}$$

where $\mathcal{R}_u(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

To find L_1, we can compute

$$\begin{aligned}
\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda} &= -\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}q_\Lambda - P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\left(\frac{\mu}{1-P_\Lambda} - \frac{\mu}{P_\Lambda}\right) \\
&= -\frac{\mathrm{d}P_\Lambda}{\mathrm{d}\cdot\, q_\Lambda}q_\Lambda - P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\cdot q_\Lambda \qquad \text{(Since } P_\Lambda \text{ satisfies Eq. (6))} \\
&= -P_\Lambda
\end{aligned}$$

Therefore, we have $\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda} \in \left[-1, -\frac{\mu}{3}\right]$ by Lemma 14. Therefore, we can set $L_1 = 1$.

Let h_0 be the value of h_Λ when $q_\Lambda = -1$, which means $x^\top\Lambda x = 0$. To find R_1, notice that since $\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda} \in \left[-1, -\frac{\mu}{3}\right]$, we must have $-q_\Lambda + h_0 \leq h_\Lambda \leq -\frac{\mu}{3}q_\Lambda + h_0$. By Lemma 14, we know that $h_0 \in \left[0, 2\sqrt{\mu}\right]$. Therefore, we have $-x^\top\Lambda x \leq h_\Lambda(x) \leq -\frac{\mu}{3}x^\top\Lambda x + 3\sqrt{\mu}$ for any $x \in \mathcal{X}$ and $\Lambda \in \Theta$. Since $\|\Lambda\|_F \leq \left\|\hat{\Lambda}\right\|_F \leq 2\left\|\Lambda^*\right\|_F$, we have $|h_\Lambda(x)| \leq 2\left\|\Lambda^*\right\|_F M^2 := R_1$, which holds when $\mu \leq \frac{4}{9}\left\|\Lambda^*\right\|_F^2 M^4$. Then, by Lemma 9, we know that $\mathcal{R}_u(\mathcal{F}) \leq \frac{2\|\Lambda^*\|_F M^2}{\sqrt{u}}$. Thus, plugging in values of L_1, R_1 and $\mathcal{R}_u(\mathcal{F})$ into Eq. (22) gives our first concentration inequality.

We can basically follow exactly the same strategy to prove the second concentration inequality. In particular, define $\tilde{h}_\Lambda(x) = P_\Lambda(x)x^\top\Lambda x = P_\Lambda(x)q_\Lambda(x) + P_\Lambda(x)$. Then, with probability at least $1-\frac{\delta}{3}$, it holds simultaneously for any $\Lambda \in \Theta$ that

$$\left|\mathbb{E}_{X\sim\nu}\left[\tilde{h}_\Lambda(X)\right] - \frac{1}{u}\sum_{i=1}^{u}\tilde{h}_\Lambda(x_i)\right| \leq 2L_2\cdot\mathcal{R}_u(\mathcal{F}) + R_2\sqrt{\frac{2\log(6/\delta)}{u}}, \tag{23}$$

where $\left|\tilde{h}_\Lambda(x)\right| \leq R_2$ for any $x \in \mathcal{X}$, $\Lambda \in \Theta$ and $\tilde{h}_\Lambda$ is L_2-Lipschitz in q_Λ.

To find L_2, we can compute

$$\frac{\mathrm{d}\tilde{h}_\Lambda}{\mathrm{d}q_\Lambda} = P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\cdot x^\top\Lambda x.$$

By Lemma 14, we know that $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \in \left[0, \frac{1}{8\mu}\right]$. Thus, we have $\left|\frac{\mathrm{d}\tilde{h}_\Lambda}{\mathrm{d}q_\Lambda}\right| \le 1 + \frac{\|\Lambda^*\|_F M^2}{4\mu} := L_2$. It is obvious that $\tilde{h}_\Lambda(x) \le 2\|\Lambda^*\|_F M^2 := R_2$. Thus, by plugging the values of L_2, R_2 and $\mathcal{R}_u(\mathcal{F})$ into Eq. (23), we can obtain the second concentration inequality.

Finally, both concentration inequalities hold simultaneously with probability at least $1 - \frac{2\delta}{3}$ by a simple union bound. □

Lemma 9. *If* $\left\|\hat{\Lambda}\right\|_F \le 2\|\Lambda^*\|_F$*, then, we have* $\mathcal{R}_u(\mathcal{F}) \le \sqrt{\frac{\mathbb{E}_{X\sim\nu}\left[(X^\top\hat{\Lambda}X)^2\right]}{u}} \le \frac{2\|\Lambda^*\|_F M^2}{\sqrt{u}}$*, where* $\mathcal{F} = \left\{x \mapsto x^\top(s\cdot\hat{\Lambda})x : s \in [0,1]\right\}$.

Proof. Let $\sigma_1, \ldots, \sigma_u$ be i.i.d. Rademacher random variables, which are uniform over $\{-1, +1\}$. Let $x_1, \ldots, x_u \sim \nu$ be i.i.d. samples. Then, by definition of Rademacher complexity, we have

$$
\begin{aligned}
\mathcal{R}_u(\mathcal{F}) &= \mathbb{E}\left[\sup_{q\in\mathcal{F}} \frac{1}{u}\sum_{i=1}^u \sigma_i q(x_i)\right] \\
&= \mathbb{E}\left[\sup_{s\in[0,1]} \frac{1}{u}\sum_{i=1}^u \sigma_i x_i^\top(s\hat{\Lambda})x_i\right] && \text{(By definition of } \mathcal{F}\text{)} \\
&\overset{\text{(i)}}{=} \frac{1}{u}\mathbb{E}\left[\mathbb{1}\left\{\sum_{i=1}^n \sigma_i x_i^\top\hat{\Lambda}x_i \ge 0\right\}\sum_{i=1}^n \sigma_i x_i^\top\hat{\Lambda}x_i\right]. \\
&\le \frac{1}{u}\mathbb{E}\left[\left|\sum_{i=1}^u \sigma_i x_i^\top\hat{\Lambda}x_i\right|\right] \\
&\le \frac{1}{u}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^u \sigma_i x_i^\top\hat{\Lambda}x_i\right)^2\right]} && \text{(By Jensen's inequality)} \\
&= \frac{1}{u}\sqrt{\mathbb{E}\left[\sum_{i=1}^u \left(x_i^\top\hat{\Lambda}x_i\right)^2\right]} && \text{(Since } \sigma_i\text{'s are i.i.d. and } \mathbb{E}[\sigma_i] = 0\text{)} \\
&= \sqrt{\frac{\mathbb{E}_{X\sim\nu}\left[\left(X^\top\hat{\Lambda}X\right)^2\right]}{u}} \le \frac{2\|\Lambda^*\|_F M^2}{\sqrt{u}}.
\end{aligned}
$$

Here, the equality (i) holds because when $\sum_{i=1}^n \sigma_i x_i^\top\hat{\Lambda}x_i < 0$, the supremum over $s \in [0,1]$ will be obtained by taking $s = 0$; otherwise, it will be obtained by taking $s = 1$. □

C.2.3 Other Lemmas

The following lemma basically shows that $f(\Lambda)$ is linear in scalar multiplication.

Lemma 10. *If* $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$*, with* $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$*, then, for any* $s \ge 0$*, it holds that* $D_E(s\cdot\hat{\mathbf{\Lambda}}) = \overline{D}_E(s\cdot\hat{\Lambda})$*, where* D_E *and* $\overline{D}_E$ *are defined in Eq.* (11)*.*

Proof. It suffices to show that if $\sum_{y\in\mathcal{Y}_\ell} y^\top\hat{\Lambda}_y y = f(\hat{\Lambda})$, then $\sum_{y\in\mathcal{Y}_\ell} y^\top(s\cdot\hat{\Lambda}_y)y = f(s\cdot\hat{\Lambda})$ for any $s > 0$. By definition, we have

$$
\begin{aligned}
f(s\cdot\hat{\Lambda}) = \max_{\Lambda_y} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} y^\top\Lambda_y y \\
\text{subject to} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell}\Lambda_y = s\cdot\hat{\Lambda} \\
& \Lambda_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell.
\end{aligned}
$$

For the above optimization problem, we can do a change of variable by setting $\Lambda'_y = \frac{1}{s} \cdot \Lambda_y \implies \Lambda_y = s \cdot \Lambda'_y$. Then, we have

$$\begin{aligned} f(s \cdot \hat{\Lambda}) = \max_{\Lambda_y} \quad & \textstyle\sum_{y \in \mathcal{Y}_\ell} y^\top (s \cdot \Lambda'_y) y \\ \text{subject to} \quad & \textstyle\sum_{y \in \mathcal{Y}_\ell} s \cdot \Lambda'_y = s \cdot \hat{\Lambda} \\ & s \cdot \Lambda'_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned}$$

$$\begin{aligned} \implies f(s \cdot \hat{\Lambda}) = \max_{\Lambda_y} \quad & s \textstyle\sum_{y \in \mathcal{Y}_\ell} y^\top \Lambda'_y y \\ \text{subject to} \quad & \textstyle\sum_{y \in \mathcal{Y}_\ell} \Lambda'_y = \hat{\Lambda} \\ & \Lambda'_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned}$$

$$\implies f(s \cdot \hat{\Lambda}) = s \cdot f(\hat{\Lambda}) = s \cdot \sum_{y \in \mathcal{Y}_\ell} y^\top \Lambda_y y = \sum_{y \in \mathcal{Y}_\ell} y^\top (s \cdot \hat{\Lambda}_y) y.$$

Thus, the proof is complete. □

Lemma 11. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a concave function with maximizer x^* over the convex set $\mathcal{C}$. Further, assume that f is G-strongly concave with respect to ℓ_2 norm in region $\mathcal{S} \cap \mathcal{C}$, where $\mathcal{S} = \{x : \|x - x^*\|_2 \leq A\}$. If $f(x^*) - f(x) \leq \frac{AG}{2}$ and $c \in \mathcal{C}$, then $x \in \mathcal{S}$.*

Proof. By property of strong concavity, we know that, $f(x^*) - f(x) \geq \frac{G}{2} \|x - x^*\|_2$ for any $x \in \mathcal{S} \cap \mathcal{C}$. Now, suppose x' satisfies $f(x^*) - f(x') \leq \frac{AG}{2}$, $x' \in \mathcal{C}$ and $x' \notin \mathcal{S}$. Then, we must have $\|x' - x^*\|_2 > A$.

Let $\gamma \in (0, 1)$ be some number such that $z = \gamma x' + (1-\gamma)x^*$ lies on the boundary of $\mathcal{S}$. By convexity, we also have $z \in \mathcal{C}$. Then, since f is concave, we have $f(z) \geq \gamma f(x') + (1 - \gamma) f(x^*) > f(x')$, where the second inequality is strict because f is strongly concave in a region around x^*. Since $f(x^*) - f(x') \leq \frac{AG}{2}$, f is G-strongly concave on $\mathcal{S}$ and z lies on the boundary of $\mathcal{S}$, we have

$$\frac{AG}{2} = \frac{G}{2} \|z - x^*\|_2 \leq f(x^*) - f(z) < f(x^*) - f(x') \leq \frac{AG}{2}.$$

This is a contradiction and thus we must have $x' \in \mathcal{S}$. □

The following lemma quantitatively describes how close $\widetilde{\Lambda}$ and Λ^* needs to be to ensure an at most ϵ multiplicative constraint violation.

Lemma 12. *Assume $\|x\|_2 \leq M$ for any $x \in \mathcal{X}$. Let $\Sigma = \mathbb{E}_{X \sim \nu}\left[XX^\top\right] \succ \mathbf{0}$ and $\Lambda^* = \operatorname{argmax}_{\Lambda \succeq \mathbf{0}} \overline{D}(\Lambda)$. Then, for any $\epsilon > 0$, if we have*

$$\left\|\widetilde{\Lambda} - \Lambda^*\right\|_F \leq \frac{8\mu^2 \lambda_{\min}(\Sigma)}{3M^2 \lambda_{\max}(\Sigma)} \cdot \frac{\epsilon}{1 + \epsilon},$$

then it holds that $y^\top \mathbb{E}_{X \sim \nu}\left[P_{\widetilde{\Lambda}}(X) XX^\top\right]^{-1} y \leq (1 + \epsilon) c_\ell^2$ for any $y \in \mathcal{Y}_\ell$.

Proof. Fix some $\epsilon > 0$. First, notice that if we regard P_Λ as a function of $q_\Lambda(x) = x^\top \Lambda x - 1$, it then holds that

$$\|\nabla_\Lambda P_\Lambda(x)\|_2 = \left\|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \nabla_\Lambda q_\Lambda(x)\right\|_2 \leq \left|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\right| \left\|xx^\top\right\|_2 \leq \left|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\right| M^2 \leq \frac{M^2}{8\mu},$$

where we obtain the last inequality by using Lemma 14. Therefore, for any $x \in \mathcal{X}$ and $\widetilde{\Lambda} \succeq \mathbf{0}$, we have $\left|P_{\widetilde{\Lambda}}(x) - P_{\Lambda^*}(x)\right| \leq \frac{M^2}{8\mu} \cdot \left\|\widetilde{\Lambda} - \Lambda^*\right\|_F$ by mean value theorem and Cauchy-Schwartz. inequality.

Therefore, if we have $\left\|\widetilde{\Lambda} - \Lambda^*\right\|_F \leq \delta$, then

$$\left|P_{\widetilde{\Lambda}}(x) - P_{\Lambda^*}(x)\right| \leq \frac{M^2 \delta}{8\mu} \implies P_{\widetilde{\Lambda}}(x) \geq P_{\Lambda^*}(x) - \frac{M^2 \delta}{8\mu}$$

$$\implies \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right] \succeq \mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)XX^\top\right] - \frac{M^2\delta}{8\mu}\mathbb{E}_{X\sim\nu}\left[XX^\top\right].$$

By Lemma 13, we know that

$$y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1} y \le c_\ell^2(1+\epsilon) \iff \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right] \succeq \frac{yy^\top}{(1+\epsilon)c_\ell^2}. \tag{24}$$

Let $\Sigma^* = \mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)XX^\top\right]$. Therefore, to guarantee the condition in Eq. (24), it is sufficient to guarantee that $\Sigma^* - \frac{M^2\delta}{8\mu}\Sigma \succeq \frac{yy^\top}{(1+\epsilon)c_\ell^2}$, which is equivalent to

$$w^\top\Sigma^* w - \frac{M^2\delta}{8\mu} w^\top\Sigma w \ge \frac{(w^\top y)^2}{c_\ell^2(1+\epsilon)}, \quad \forall\text{unit vector } w\in\mathbb{R}^d$$

$$\iff \frac{1}{w^\top\Sigma w}\cdot w^\top\left(\Sigma^* - \frac{yy^\top}{(1+\epsilon)c_\ell^2}\right)w \ge \frac{M^2\delta}{8\mu}, \quad \forall\text{unit vector } w\in\mathbb{R}^d.$$

Therefore, it is sufficient to choose δ such that

$$\frac{M^2\delta}{8\mu} \le \frac{1}{\lambda_{\max}(\Sigma)}\cdot\lambda_{\min}\left(\Sigma^* - \frac{yy^\top}{c_\ell^2(1+\epsilon)}\right) \le \min_{w:\|w\|_2=1}\frac{1}{w^\top\Sigma w}\cdot w^\top\left(\Sigma^* - \frac{yy^\top}{(1+\epsilon)c_\ell^2}\right)w.$$

Since P_{Λ^*} satisfies the constraint defined in problem (12), we have $\Sigma^* \succeq \frac{yy^\top}{c_\ell^2}$. Meanwhile, by Lemma 14, we know that $P_{\Lambda^*}(x) \ge \frac{\mu}{3}$ for any $x\in\mathcal{X}$, which means that $\Sigma^* \succeq \frac{\mu}{3}\cdot\Sigma$. That is, for any unit vector $w\in\mathbb{R}^d$, we have

$$w^\top\Sigma^* w \ge \frac{\left(w^\top y\right)^2}{c_\ell^2} \quad \text{and} \quad w^\top\Sigma^* w \ge \frac{\mu}{3}\lambda_{\min}\left(\Sigma\right),$$

which together implies $w^\top\Sigma^* w \ge \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma), \frac{\left(w^\top y\right)^2}{c_\ell^2}\right\}$. Therefore, it holds that

$$\begin{aligned}
w^\top\Sigma w - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2} &\ge \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma), \frac{\left(w^\top y\right)^2}{c_\ell^2}\right\} - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2} \\
&= \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma) - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2}, \frac{\epsilon\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2}\right\} \\
&\ge \frac{\epsilon\mu}{3(1+\epsilon)}\cdot\lambda_{\min}\left(\Sigma\right) \\
\implies \lambda_{\min}\left(\Sigma^* - \frac{yy^\top}{c_\ell^2(1+\epsilon)}\right) &\ge \frac{\epsilon\mu}{3(1+\epsilon)}\cdot\lambda_{\min}\left(\Sigma\right).
\end{aligned}$$

Therefore, to guarantee the condition in Eq. (24), it is sufficient to have

$$\frac{M^2\delta}{8\mu} = \frac{\epsilon\mu\lambda_{\min}(\Sigma)}{3(1+\epsilon)\lambda_{\max}(\Sigma)} \implies \mu = \frac{8\mu^2\lambda_{\min}(\Sigma)}{3M^2\lambda_{\max}(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon},$$

Thus, the proof is complete. □

The following lemma is a result of standard Schur complement technique.

Lemma 13. *If $\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]$ is invertible and $c_\ell > 0$, then*

$$y^\top\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1}y \le c_\ell^2 \iff \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{yy^\top}{c_\ell^2}.$$

Proof. For simplicity, let $A = E_{X\sim\nu}\left[P(X)XX^\top\right] \succ \mathbf{0}$. Then, we consider the block matrix $\begin{bmatrix} A & y \\ y^\top & c_\ell^2\end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$. Let $\begin{bmatrix} u & a\end{bmatrix}^\top \in \mathbb{R}^{d+1}$ with $u\in\mathbb{R}^d$ be some vector.

Now, for one direction, suppose $y^\top A^{-1} y \le c_\ell^2$ holds. Consider

$$\begin{bmatrix} u & a \end{bmatrix} \begin{bmatrix} A & y \\ y^\top & c_\ell^2 \end{bmatrix} \begin{bmatrix} u \\ a \end{bmatrix} = u^\top A u + 2 a u^\top y + 2 c_\ell^2 a^2 := r(u, a).$$

If we minimize $r(u, a)$ over u, which means to treat a as fixed, we can get (by taking gradient and setting it to zero)

$$u^* = -a A^{-1} y \implies r(u^*, a) = a^2 (c_\ell^2 - y^\top A^{-1} y).$$

Since $y^\top A^{-1} y \le c_\ell^2$, we know that $r(u^*, a) \ge 0$, which means $r(u, a) \ge 0$ for any $\begin{bmatrix} u & a \end{bmatrix}^\top \in \mathbb{R}^{d+1}$.

Then, if we minimize $r(u, a)$ over a, we can get

$$a^* = -\frac{u^\top y}{c_\ell^2} \implies r(u, a^*) = u^\top A u - \frac{\left(u^\top y\right)^2}{c_\ell^2}.$$

Since $r(u, a) \ge 0$ for any $\begin{bmatrix} u & a \end{bmatrix}^\top \in \mathbb{R}^{d+1}$, we know that $u^\top A u - \frac{\left(u^\top y\right)^2}{c_\ell^2} \ge 0$ for any $u \in \mathbb{R}^d$. That is, we have $A \succeq \frac{y y^\top}{c_\ell^2}$.

The other direction simply takes the above calculation in a reversed way and thus the proof is complete. □

C.2.4 Properties of P_Λ

A visualization of P_Λ is given in Figure 2.

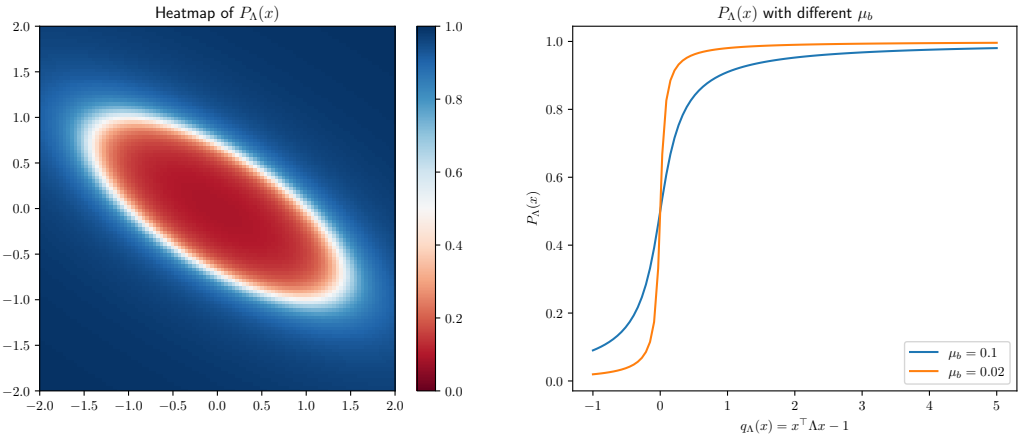

Figure 2: (left) A heatmap of some P_Λ when problem dimension is $d = 2$, which shows that P_Λ is approximately an 0-1 threshold rule characterized by an ellipsoid. (right) A plot of P_Λ as a function of $q_\Lambda(x) = x^\top \Lambda x - 1$, which shows that the change of P_Λ near the boundary of ellipsoid is sharper when the barrier weight μ is smaller.

Lemma 14. *The function $P_\Lambda(x)$ defined in* (7)*, if regarding as a function of $q_\Lambda(x) = x^\top \Lambda x - 1 \ge -1$, satisfies*

- $\lim_{q_\Lambda \to 0} P_\Lambda = \frac{1}{2}$ *for any* $\mu \in (0, 1)$
- *When* $q_\Lambda = -1$, $P_\Lambda = \frac{1}{2} + \mu - \frac{\sqrt{1+4\mu^2}}{2} \ge \frac{\mu}{3}$ *and* $P_\Lambda - \mu(\log(1 - P_\Lambda) + \log(P_\Lambda)) \le 2\sqrt{\mu}$ *for any* $\mu \in (0, 1)$.
- $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} = \frac{\mu\sqrt{q_\Lambda^2 + 4\mu^2} - 2\mu^2}{q_\Lambda^2 \sqrt{q_\Lambda^2 + 4\mu^2}}$ *decreases as* q_Λ^2 *increases. Further,* $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \in [0, \frac{1}{8\mu}]$. *Thus,* P_Λ *increases monotonically as* q_Λ *increases and* $P_\Lambda(x) \ge \frac{\mu}{3}$ *for any* $x \in \mathcal{X}$ *and* $\Lambda \succeq \mathbf{0}$.
- $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\big|_{q_\Lambda = \pm 1} \ge \frac{\mu}{10}$ *and* $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \ge \frac{\mu}{2q_\Lambda^2}$ *when* $q_\Lambda^2 \ge 12\mu^2$.

Proof. For simplicity, we will drop the subscript Λ and just treat P as a function of q. That is, we have

$$P(q) = \frac{1}{2} - \frac{\mu}{q} + \frac{\sqrt{(2\mu - q)^2 + 4\mu q}}{2q}.$$

We prove each bullet point separately.

- Since $P(q)$ also satisfies Eq. (6), which in simpler form is $\frac{\mu}{1-P(q)} - \frac{\mu}{P(q)} = q$, we can easily see that $P(q) = \frac{1}{2}$ satisfies this equation when $q = 0$.

- By direction computation, we can get $P(-1) = \frac{1}{2} + \mu - \frac{\sqrt{1+4\mu^2}}{2}$. To show this is greater than $\frac{\mu}{3}$ for any $\mu \in [0, 1]$, consider $\ell(\mu) = P(-1) - \frac{\mu}{3}$. It is easy to check that $\ell(0) = 0$ and $\ell(1) > 0$. Then, since $\ell'(\mu) = \frac{2}{3} - \frac{2\mu}{\sqrt{1+4\mu^4}}$ is initially greater than 0 and then smaller than 0, we know $\ell(\mu)$ first increases and then decreases on $[0, 1]$. Thus, $\ell(\mu) \geq 0$ on $[0, 1]$ and thus $P(-1) \geq \frac{\mu}{3}$ for any $\mu \in [0, 1]$.

  For the second part, define $\tilde{\ell}(\mu) = 2\sqrt{\mu} - P(-1) + \mu\left(\log(1 - P(-1)) + \log(P(-1))\right)$. Then, by utilizing the fact that P satisfies Eq. (6), we can compute its derivative and get $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu} = \frac{1}{\sqrt{\mu}} + \log(1 - P(-1)) + \log(P(-1))$. We can check that on the domain $(0, 1)$, we have $\frac{\mathrm{d}^2\tilde{\ell}}{\mathrm{d}\mu^2} = -\frac{1}{2\mu^{3/2}} + \frac{1}{\mu} - \frac{2}{\sqrt{1+4\mu^2}} \cdot \frac{2\sqrt{\mu(1+4\mu^2)} - 4\mu^{3/2} - \sqrt{1+4\mu^2}}{2\mu^{3/2}\sqrt{1+4\mu^2}} \leq 0$ on $(0, 1)$, which means that $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu}$ is monotonically decreasing. To see why the second derivative is smaller than 0, we can compute

  $$\left(4\mu^{3/2} + \sqrt{1 + 4\mu^2}\right) - 4\mu\left(1 + 4\mu^2\right) = (1 - 2\mu)^2 + 8\mu^{3/2}\sqrt{1 + 4\mu^2} \geq 0.$$

  Thus, $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu}$ is initially greater than 0 and then smaller than 0 on $(0, 1)$. It is easy to verify that $\lim_{\mu\to 0}\tilde{\ell} = 0$ and $\tilde{\ell}(1) > 0$. Therefore, we have $\tilde{\ell}(\mu) \geq 0$ for any $\mu \in (0, 1)$.

- We can get $\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{\mu\sqrt{q^2+4\mu^2} - 2\mu^2}{q^2\sqrt{q^2+4\mu^2}}$ by direct computation. To show it is decreasing as q^2 increasing, we consider $\tilde{f}(z) = \frac{\mu\sqrt{z+4\mu^2} - 2\mu^2}{z\sqrt{z+4\mu^2}}$ and it is sufficient to show that $\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} < 0$ for any $z > 0$. Again by direct computation, we have

  $$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} = \frac{\mu\left(8\mu^3 + 3\mu z - \left(z + 4\mu^2\right)^{3/2}\right)}{z^2\left(z + 4\mu^2\right)^{3/2}}.$$

  By direct computation, We can show that $\left(z + 4\mu^2\right)^3 - \left(8\mu^3 + 3\mu z\right)^2 = z^3 + 3z^2\mu^2 > 0$ for any $z > 0$ and $\mu \in [0, 1]$. Thus, $\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} < 0$ and thus $\frac{\mathrm{d}P}{\mathrm{d}q}$ is decreasing as q^2 increases.

  It is obvious that $\frac{\mathrm{d}P}{\mathrm{d}q} \geq 0$ for any $q^2 \geq 0$ and $\mu \in [0, 1]$ since we always have $\mu\sqrt{q^2 + 4\mu^2} \geq 2\mu^2$. Thus, the maximum value could potentially happen is when $q^2 \to 0$, which can be evaluated by using L'Hospital's rule. A direct computation gives us $\lim_{q^2\to 0}\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{1}{8\mu}$. Thus, we can conclude that $\frac{\mathrm{d}P}{\mathrm{d}q} \in \left[0, \frac{1}{8\mu}\right]$. Therefore, P increases monotonically as q increases, which implies that $P_\Lambda(x) \geq \frac{\mu}{3}$ for any $x \in \mathcal{X}$ and Λ.

- By direct computation, we have $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}|_{q_\Lambda=\pm 1} = \mu\left(1 - \frac{2\mu}{\sqrt{1+4\mu^2}}\right) \geq \mu\left(1 - \frac{2}{\sqrt{5}}\right) \geq \frac{\mu}{10}$ for any $\mu \in [0, 1]$. The reason is that we can easily see $\frac{2\mu}{\sqrt{1+4\mu^2}}$ is increasing in μ.

  Finally, notice that when $2\mu \leq \frac{1}{2}\sqrt{q^2 + 4\mu^2}$, which is equivalent to $q^2 \geq 12\mu^2$, we have

  $$\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{\mu\sqrt{q^2 + 4\mu^2} - 2\mu^2}{q^2\sqrt{q^2 + 4\mu^2}} \geq \frac{\mu\sqrt{q^2 + 4\mu^2} - \frac{\mu}{2}\sqrt{q^2 + 4\mu^2}}{q^2\sqrt{q^2 + 4\mu^2}} = \frac{\mu}{2q^2}.$$

Thus, the proof is complete. □

C.3 An Alternative Approach to OPTIMIZEDESIGN

Based on the analysis in Section C.1, we know that maximizing $\overline{D}(\cdot)$ is equivalent to maximizing $D(\cdot)$. Therefore, as an alternative to Algorithm 2, which maximizes $D(\cdot)$ through stochastic gradient ascent, it is natural to have an algorithm that directly maximizes $\overline{D}(\cdot)$. Here, we will consider subgradient ascent.

Recall that $\overline{D} : \mathbb{S}_+^d \mapsto \mathbb{R}$ is defined as

$$\overline{D}(\Lambda) = \mathbb{E}_{X\sim\nu}\left[P_\Lambda(X) - \mu\left(\log(1-P_\Lambda(X)) + \log(P_\Lambda(X))\right) - P_\Lambda(X)X^\top \Lambda X\right] + \frac{1}{c_\ell^2}\cdot f(\Lambda),$$

where $f(\Lambda)$ is defined in problem (10). The subgradient of $\overline{D}(\Lambda)$ is

$$\begin{aligned}\partial\overline{D}(\Lambda) &= \mathbb{E}_{X\sim\nu}\left[\left(1+\frac{\mu}{1-P_\Lambda(x)}-\frac{\mu}{P_\Lambda(X)}-X^\top\Lambda X\right)\nabla P_\Lambda(X) - P_\Lambda(X)XX^\top\right] + \frac{\partial f(\Lambda)}{c_\ell^2} \\ &\qquad\qquad \text{(The first term is differentiable)} \\ &= \frac{\partial f(\Lambda)}{c_\ell^2} - \mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)XX^\top\right]. \qquad \text{(Since } P_\Lambda(X) \text{ solves Eq. (6))}\end{aligned}$$

Therefore, to run subgradient ascent, we only need to find an element in $\partial f(\Lambda)$, which can be obtained by solving the following optimization problem as claimed by Lemma 15.

$$\begin{array}{rl} \min_\Gamma & \langle \Gamma, \Lambda\rangle \\ \text{subject to} & \Gamma \succeq yy^\top, \quad \forall y \in \mathcal{Y}_\ell, \\ & \Gamma \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top. \end{array} \tag{25}$$

As a result, we have Algorithm 3 as an alternative to solve OPTIMIZEDESIGN. Compared to Algorithm 2, which needs to maintain $|\mathcal{Y}_\ell|\, d^2$ number of objective variables, Algorithm 3 only has d^2 variables. However, each iteration of Algorithm 3 is computationally more intensive since finding a subgradient needs to solve the problem (25).

Algorithm 3 Projected Stochastic Subgradient Ascent to Solve OPTIMIZEDESIGN

Input: Number of iterations K; number of samples u; barrier weight $\mu_b \in (0,1)$
Initialize $\hat{\Lambda}^{(0)} = \mathbf{0}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Sample $x_k \sim \nu$
  Solve problem (25) with current $\hat{\Lambda}^{(k)}$ to get $\Gamma^{(k)}$
  Set $g_k = \frac{\Gamma^{(k)}}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_k x_k^\top$
  Set $\hat{\Lambda}^{(k+1)} \leftarrow \hat{\Lambda}^{(k)} + \eta_k g_k$, where $\eta_k = \frac{1}{\sqrt{2\sum_{s=1}^k \|g_s\|_2^2}}$
  Update $\hat{\Lambda}^{(k+1)} \leftarrow \Pi_{\mathbb{S}_+^d}(\hat{\Lambda}^{(k+1)})$, a projection to the set of $d\times d$ PSD matrices
end for
Let $\hat{\Lambda} = \frac{1}{K}\sum_{k=1}^K \hat{\Lambda}^{(k)}$
Find $s^* \leftarrow \operatorname{argmax}_{s\in[0,1]} \overline{D}_E(s\cdot\hat{\Lambda})$, where $\overline{D}_E$ is the empirical version of $\overline{D}$, evaluated using u i.i.d. samples
return $\widetilde{\Lambda} = s^*\cdot\hat{\Lambda}$

A result similar to Theorem 5 can also be obtained for Algorithm 3, which is given in Theorem 6. The bounds are almost identical except that the old lower bound for K depends on $|\mathcal{Y}_\ell|^3$ while the new one depends on $|\mathcal{Y}_\ell|$. Steps identical to the proof of Theorem 5 will be skipped in the proof of Theorem 6.

Theorem 6. *Let* $\Lambda^* \in \operatorname{argmax}_{\Lambda\succeq\mathbf{0}} \overline{D}(\Lambda)$ *and take other settings the same as that in Theorem 5.*

Then, Λ^* is unique. Further, for any $\epsilon > 0$ and $\delta > 0$, suppose it holds that

$$\mu \le \min\left\{\sqrt{\frac{3\kappa(\Sigma)\left\|\Lambda^*\right\|_F M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}}, \frac{4}{9}\left\|\Lambda^*\right\|_F^2 M^4, \frac{1}{2\sqrt{3}}\right\}$$

$$K \ge \frac{288\kappa(\Sigma)^2\left\|\Lambda^*\right\|_F^4 M^4(M^4+4\left|\mathcal{Y}_\ell\right|C_\ell^2)\cdot\left(2\left\|\Lambda^*\right\|_F M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2$$

$$u \ge \frac{576\kappa(\Sigma)^2\left\|\Lambda^*\right\|_F^2 M^8\cdot\left(2\left\|\Lambda^*\right\|_F M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2.$$

Then, with probability at least $1-\delta$, Algorithm 2 will output $\widetilde{\Lambda}$ that satisfies

- $y^\top\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1}y \le (1+\epsilon)c_\ell^2, \quad \forall y\in\mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, *where $\widetilde{P}$ is the optimal solution to problem* (20).

Proof. **First Bullet Point.** Similar to the proof of Theorem 5, let $\hat{\Lambda}$ be the parameter obtained by Algorithm 3 just before the re-scaling step (line 11). Then, by Theorem 3.13 of [21], with probability at least $1-\frac{\delta}{3}$, it holds that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{\mathrm{Reg}(K) + 2\sqrt{2K\log(6/\delta)}}{K},$$

where $\mathrm{Reg}(K)$ is the regret of running projected stochastic subgradient ascent for K steps with η_k specified in Algorithm 3. Meanwhile, by Theorem 4.14 of [21] also, we have $\mathrm{Reg}(K) = \sqrt{2}B^2\sqrt{\sum_{k=1}^K\left\|g_k\right\|_2^2}$, where $B = \left\|\Lambda^*\right\|_F$. Since $g_k = \frac{\Gamma^{(k)}}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_kx_k^\top$ and $\left\|\Gamma^{(k)}\right\|_F \le 2\left\|\sum_{y\in\mathcal{Y}_\ell}yy^\top\right\|_F$, we can easily get $\left\|g_k\right\|_2^2 \le 2M^4 + \frac{8}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\left\|y\right\|_2^4 = 2M^4 + 8\left|\mathcal{Y}_\ell\right|C_\ell^2$. Thus, we have

$$\mathrm{Reg}(K) \le 2\left\|\Lambda^*\right\|_F^2\sqrt{M^4+4\left|\mathcal{Y}_\ell\right|C_\ell^2}\cdot\sqrt{K} := C_{\mathrm{Reg}}\sqrt{K} \tag{26}$$

$$\implies \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}}, \tag{27}$$

We now consider the effect of using u i.i.d. samples in the re-scaling step. Since re-scaling always increases the function value, we must have $\overline{D}_E(\hat{\Lambda}) \le \overline{D}_E(\widetilde{\Lambda})$.

Then, after **exactly the same** steps of analysis, we can get the following same lower bound for K,

$$K \ge \left(\frac{3\kappa(\Sigma)M^2\left(C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}\right)}{2G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2, \tag{28}$$

with a different value of C_{Reg}.

Second Bullet Point. We then prove the upper bound for primal objective value $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$, which explains the reason why an extra re-scaling step is needed. Let $\hat{\mathbf{\Lambda}} = (\hat{\Lambda}_y)_{y\in\mathcal{Y}_\ell}$ be a set of PSD matrices that solves problem (10) with parameter $\hat{\Lambda}$ and $\widetilde{\mathbf{\Lambda}} = s^*\cdot\hat{\mathbf{\Lambda}}$, where $s^* = \mathrm{argmax}_{s\in[0,1]}\,\overline{D}_E(s\cdot\hat{\Lambda})$. Since the constraint in problem (10) requires $\sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y = \hat{\Lambda}$, we have $\sum_{y\in\mathcal{Y}_\ell}\widetilde{\Lambda}_y = \widetilde{\Lambda}$, which is the output of Algorithm 3.

Define $g(s) = D_E(s\cdot\widetilde{\mathbf{\Lambda}})$. By construction, we know that $g(s)$ is maximized at $s=1$ because $\overline{D}_E(s\cdot\hat{\Lambda}) = D_E(s\cdot\hat{\mathbf{\Lambda}})$ for any $s\ge 0$ as shown in Lemma 10, which means that $s^* = \mathrm{argmax}_{s\in[0,1]}\,D_E(s\cdot\hat{\mathbf{\Lambda}})$. Therefore, we have $g'(1)\ge 0$, which in turn gives us

$$g'(1) = \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}y^\top\widetilde{\Lambda}_y y - \frac{1}{u}\sum_{i=1}^u P_{\widetilde{\Lambda}}(x_i)x_i^\top\widetilde{\Lambda}x_i \ge 0.$$

Then, after **exactly the same** steps of analysis, we can get $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \leq \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, where $\widetilde{P}$ is the optimal solution of the problem (20). □

C.3.1 Technical Lemmas

Lemma 15. *The optimal value of the optimization problem* (25) *with parameter $\Lambda \succeq \mathbf{0}$ is equal to $f(\Lambda)$. Further, let $\Gamma^*(\Lambda)$ be an optimal solution to* (25)*. Then, it holds that $\Gamma^*(\Lambda) \in \partial f(\Lambda)$ and $\|\Gamma^*(\Lambda)\| \leq 2\left\|\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\|_F$.*

Proof. Alternatively, we first consider the following optimization problem.

$$\begin{array}{ll} \max_{\Lambda_y,\Sigma} & \sum_{y\in\mathcal{Y}_\ell} y^\top\left(\Lambda_y - 2\Sigma\right) y \\ \text{subject to} & \Lambda = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y - \Sigma, \\ & \Sigma \succeq \mathbf{0}, \Lambda_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{array} \tag{29}$$

Since $y^\top \Sigma y \geq 0$ for any $y \in \mathcal{Y}_\ell$ and $\Sigma \succeq \mathbf{0}$, it is clear that problem (29) has the same optimal value as problem (10). Then, let $\Gamma \in \mathbb{R}^{d\times d}$ be the dual variable for the equality constraint in problem (29). We can have its dual problem to be

$$\begin{aligned} &\min_{\Gamma} \max_{\substack{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell,\\ \Sigma\succeq\mathbf{0}}} \sum_{y\in\mathcal{Y}_\ell}\left\langle yy^\top, \Lambda_y - 2\Sigma\right\rangle + \left\langle \Gamma, \Lambda + \Sigma - \sum_{y\in\mathcal{Y}_\ell}\Lambda_y\right\rangle \\ \Longrightarrow &\min_{\Gamma} \max_{\substack{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell,\\ \Sigma\succeq\mathbf{0}}} \langle\Gamma,\Lambda\rangle + \left\langle \Sigma, \Gamma - 2\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\rangle + \sum_{y\in\mathcal{Y}_\ell}\left\langle \Lambda_y, yy^\top - \Gamma\right\rangle. \end{aligned}$$

In order for the above optimization problem to have finite value, we must have $\Gamma \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top$ and $\Gamma \succeq yy^\top$ for any $y \in \mathcal{Y}_\ell$. Therefore, we obtain the following dual problem.

$$\begin{array}{ll} \min_{\Gamma} & \langle\Gamma,\Lambda\rangle \\ \text{subject to} & \Gamma \succeq yy^\top, \quad \forall y \in \mathcal{Y}_\ell, \ . \\ & \Gamma \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top. \end{array}$$

This is exactly the problem (25). Then, we can notice the Slater's condition is clearly satisfied by problem (25), which means the strong duality holds. Therefore, problem (25) has the same optimal value as (29), which is the same as (10).

Since $f(\Lambda)$ is concave in Λ as shown in Lemma 6, to show that $\Gamma^*(\Lambda) \in \partial f(\Lambda)$, consider arbitrary $\Lambda, \Lambda' \succeq \mathbf{0}$. Then, we have

$$f(\Lambda) + \langle\Gamma^*(\Lambda), \Lambda' - \Lambda\rangle = \langle\Gamma^*(\Lambda),\Lambda\rangle + \langle\Gamma^*(\Lambda),\Lambda'-\Lambda\rangle = \langle\Gamma^*(\Lambda),\Lambda'\rangle \geq f(\Lambda').$$

The first equality holds because the optimal value of problem (25) is $f(\Lambda)$ as just shown above. The last inequality holds because $\Gamma^*(\Lambda)$ is a feasible solution to the problem (25) with parameter Λ'. Therefore, we have $\Gamma^*(\Lambda) \in \partial f(\Lambda)$.

Finally, since the constraint of problem (25) requires $\Gamma^*(\Lambda) \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top$, we can obtain $\|\Gamma^*(\Lambda)\|_F \leq 2\left\|\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\|_F$ as a direct consequence of Lemma 16. □

Lemma 16. *For $A, B \in \mathbb{S}^{d\times d}$, if $A \succeq B \succeq \mathbf{0}$, then $\|A\|_F \geq \|B\|_F$.*

Proof. Let $\lambda_1,\ldots,\lambda_d$ and $\gamma_1,\ldots,\gamma_d$ be eigenvalues of A and B, respectively. Let $v_1,\ldots,v_d$ be a set of orthogonal unit eigenvectors of matrix A. Then, we have

$$\|A\|_F = \sqrt{\mathrm{tr}\left(AA\right)} = \sqrt{\mathrm{tr}\left(\left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right)\left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right)\right)} = \sqrt{\sum_{i=1}^d \lambda_i^2}.$$

Similarly, we have $\|B\|_F = \sqrt{\sum_{i=1}^d \gamma_i^2}$. By Corollary 7.7.4 in [14], since $A \succeq B \succeq \mathbf{0}$, we know that $\lambda_i \geq \gamma_i \geq 0$ for each i. Therefore, we have $\|A\|_F \geq \|B\|_F$. □

D Selective Sampling Algorithm for Unknown Distribution ν

D.1 Statement and proof of Theorem 7

Consider now the case where we do not know ν exactly, and are returned $(\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell})$ that only approximate their ideals. Algorithm 1 can still be employed to solve this case where ν is unknown, but at the cost of sampling some historical data. Note that compared to the case where ν is know, it assumes the knowledge of an upper bound on $\sup_{x\in\text{support}(\nu)} \|x\|$. It also relies on a multiplicative factor change in the constraint of the optimization problem, in order to account for the possible constraint violation of the output of the subroutine. The last difference is the use of an approximation of the covariance matrix to compute the estimator. The covariance matrix is empirically approximated by injecting additional unlabeled samples (historical data). With that, although we do not know ν but we can approximate the relevant quantities, such as the covariance matrix $\mathbb{E}_{X\sim\nu}[XX^\top]$.

Let us detail the properties of the implementation of $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$OPTIMIZEDESIGN$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ we use at each round ℓ.

First, $\widehat{P}_\ell$ has the properties described in Theorem 4 (by using Algorithm 2). More explicitly, let $\epsilon_\ell := 2^{-\ell}$, $B < \infty$ such that $\max_{x\in\mathcal{X}} |\langle x, \theta_*\rangle| \le B$, and $\sigma < \infty$ such that $\mathbb{E}[(y_s - \langle\theta_*, x_s\rangle)^2 | x_s] \le \sigma^2$. If

$$\beta_{\delta,\ell} := 4(1+\varepsilon)^2 \left(4\sqrt{B^2+\sigma^2}+1\right)^2 \log(4\ell^2|\mathcal{Z}|^2/\delta)$$

then $\widehat{P}_\ell$ is such that

- $\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1+\varepsilon.$
- $\mathbb{E}_{X\sim\nu}\left[\widehat{P}_\ell(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}_\ell(X)\right] + 4\sqrt{\mu_b}$, where $\widetilde{P}_\ell$ is the optimal solution to problem (30).

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[P(X)\right] \\ \text{subject to} \quad & \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1, \\ & 0 \le P(x) \le 1-\mu_b, \quad \forall x\in\mathcal{X}. \end{aligned} \tag{30}$$

where $\mu_b \ge 0$. The quantity $\mathbb{E}_{X\sim\nu}\left[\widetilde{P}_\ell(X)\right]$ that uses $\mu_b > 0$ is easily related to the value when $\mu_b = 0$ through a simple scaling factor of $\frac{1}{1-\mu_b}$ (see proof below).

$\widehat{\Sigma}_{\widehat{P}_\ell}$ is the empirical covariance matrix of $\Sigma_{\widehat{P}_\ell} := \mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)XX^\top]$ using historical data and is such that

$$(1-\gamma)\Sigma_{\widehat{P}_\ell} \preceq \widehat{\Sigma}_{\widehat{P}_\ell} \preceq (1+\gamma)\Sigma_{\widehat{P}_\ell}$$

where $\gamma \ge 0$.

Again, while we think of historical data as independent data collected offline before the start of the game, in practice this historical data could just come from previous rounds (which is not technically correct since its use may introduce some dependencies).

Theorem 7 (Upper bound)**.** *Fix any $\delta \in (0,1)$. Let $\Delta = \min_{z\in\mathcal{Z}\setminus z_*}\langle z_* - z, \theta_*\rangle$ and set*

$$\beta_\delta = 256(1+\varepsilon)^2\left(4\sqrt{B^2+\sigma^2}+1\right)^2 \log(4\log_2^2(\tfrac{4}{\Delta})|\mathcal{Z}|^2/\delta).$$

For any $\tau \ge \rho(\nu)\beta_\delta$ there exists a δ-PAC selective sampling algorithm that collects $\mathcal{T}$ historical data before the start of the game, observes $\mathcal{U}$ unlabeled examples, and requests just $\mathcal{L}$ labels that satisfies

- $\mathcal{U} \le \log_2(\frac{4}{\Delta})\tau$,
- $\mathcal{L} \le \frac{1}{1-\mu_b}\min_{\lambda\in\triangle_\mathcal{X}} \rho(\lambda)\beta_\delta + \frac{5\tau}{1-\mu_b}\sqrt{\mu_b}$ *subject to* $\tau \ge \|\lambda/\nu\|_\infty \rho(\lambda)\,\beta_\delta$, *and*
- $\mathcal{T} \le \log_2(\frac{4}{\Delta})(K + u + \kappa_\delta)$

with probability at least $1-\delta$.

Here, the sample complexity for estimating the covariance matrix is bounded by $\kappa_\delta = \lceil 2K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})\max\{1, 20\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}\rceil$ *(where the sub-gaussian norm* $K_{\psi_2} = \max_{s,P}\|\sqrt{P(\widetilde{x}_s)}\Sigma_P^{-1/2}\widetilde{x}_s\|_{\psi_2}$ *), and the contributions from the optimization problem to compute* $\{\widehat{P}_\ell\}_\ell$ *are*

$$K = \widetilde{O}\left(\frac{|\mathcal{Z}|^6\,\kappa(\Sigma)^2\,\|\Lambda^*\|_2^8\,M^{16}}{\omega^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2, \quad u = \widetilde{O}\left(\frac{\kappa(\Sigma)^2\,\|\Lambda^*\|_2^6\,M^{16}}{\omega^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

Naturally, we have a trade-off on the subroutine tolerance μ_b. In order to get a better solution of the optimization over the selection rule P (and thus get a smaller $\sum_{t=(\ell-1)\tau+1}^{\ell\tau} P(x_t)$ term), the subroutine needs more unlabeled samples. However, it suffices to take $\mu_b = \frac{1}{\tau^2}$ to make $\mathcal{U}$, and $\mathcal{L}$ roughly match those of the case when ν was known.

The proof of this theorem is established through several results, which we provide in Section D.2.

D.2 Lemmas for the correctness

We first state here the correctness of Algorithm 1 in the case where ν is unknown.

Lemma 17. *With probability at least* $1-\delta$ *we have for all stages* $\ell\in\mathbb{N}$*, we have that* $z_*\in\mathcal{Z}_\ell$ *and* $\max_{z\in\mathcal{Z}_\ell}\langle z_*-z,\theta_*\rangle \le 4\epsilon_\ell$.

The proof of the correctness lemma is established though several lemmas. First we provide Lemma 18 guaranteeing concentration of empirical covariance matrices, which is obtained by sampling κ additional measurements. Then we show in Proposition 3 that the RIPS estimator does not suffer from using that empirical covariance matrix.

Lemma 18. *For any* $P:\mathcal{X}\to[0,1]$*, let* $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$, $\widehat{\Sigma}_P = \frac{1}{\kappa}\sum_{s=1}^{\kappa} P(\widetilde{x}_s)\widetilde{x}_s\widetilde{x}_s^\top$. *Define* $K_{\psi_2} = \max_s\|\sqrt{P(\widetilde{x}_i)}\Sigma_P^{-1/2}\widetilde{x}_s\|_{\psi_2}$ *. With probability at least* $1-2\exp(-c_1t^2/K_{\psi_2}^4)$ *holds*

$$(1-c)x^\top\Sigma_P x \le x^\top\widehat{\Sigma}_P x \le (1+c)x^\top\Sigma_P x$$

where $c = \max\left\{\frac{C\sqrt{d}+t}{\sqrt{\kappa}}, \left(\frac{C\sqrt{d}+t}{\sqrt{\kappa}}\right)^2\right\}$, $C = K_{\psi_2}^2\sqrt{\ln 9/c_1}$ *and* c_1 *is an absolute constant.*

Consequently for $\kappa \ge c_\delta := K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})$, holds with probability at least $1-\delta$

$$\left(1-\frac{c_\delta}{\sqrt{\kappa}}\right)x^\top\Sigma_P x \le x^\top\widehat{\Sigma}_P x \le \left(1+\frac{c_\delta}{\sqrt{\kappa}}\right)x^\top\Sigma_P x.$$

Proof. Let $A\in\mathbb{R}^{\kappa\times d}$ whose rows A_i are independent sub-gaussian isotropic random vectors in R^d and define $K_{\psi_2} = \max_i\|A_i\|_{\psi_2}$. We can apply Theorem 5.39 of [25] on A to have that with probability at least $1-2\exp(-c_1t^2/K_{\psi_2}^4)$ holds

$$1-\frac{C\sqrt{d}+t}{\sqrt{\kappa}} \le \sigma_{\min}(A) \le \sigma_{\max}(A) \le 1+\frac{C\sqrt{d}+t}{\sqrt{\kappa}},$$

where $C = K_{\psi_2}^2\sqrt{\ln 9/c_1}$ and c_1 is an absolute constant.

With Lemma 5.36 of [25], this implies that with probability at least $1-2\exp(-c_0t^2)$ holds

$$\|A^\top A - I\| \le \max\left\{\frac{C\sqrt{d}+t}{\sqrt{\kappa}}, \left(\frac{C\sqrt{d}+t}{\sqrt{\kappa}}\right)^2\right\} =: c \tag{31}$$

Recall $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$, so $Y = \sqrt{P(X)}\Sigma_P^{-1/2}X$ satisfies $\mathbb{E}[YY^\top] = \mathbb{E}[\Sigma_P^{-1/2}P(X)XX^\top\Sigma_P^{-1/2}] = \Sigma_P^{-1/2}\Sigma_P\Sigma_P^{-1/2} = I$. So we can apply (31) to get $\|\Sigma_P^{-1/2}\widehat{\Sigma}_P\Sigma_P^{-1/2} - I\| \le c$. Thus for any $y \in \mathbb{R}^d$,

$$1 - c \le \frac{y^\top}{\|y\|}\Sigma_P^{-1/2}\widehat{\Sigma}_P\Sigma_P^{-1/2}\frac{y}{\|y\|} \le 1 + c$$

so setting $y = \Sigma_P^{1/2}x$

$$(1-c)x^\top\Sigma_P x \le x^\top\widehat{\Sigma}_P x \le (1+c)x^\top\Sigma_P x.$$

Also, the sub-gaussian bound becomes $K_{\psi_2} = \max_i \|\sqrt{P(\widetilde{x}_i)}\Sigma_P^{-1/2}\widetilde{x}_i\|_{\psi_2}$. □

Proposition 3 (RIPS guarantees on empirical covariance matrix)**.** *Let $x_1, \dots, x_n$ and $\widetilde{x}_1, \dots, \widetilde{x}_\kappa$ be drawn IID from a distribution ν. For $s = 1, \dots, n$, assume that $|\langle\theta, x_s\rangle| \le B$ and $\mathbb{E}[|\langle\theta, x_s\rangle - y_s|^2] \le \sigma^2_{noise}$. For $s = 1, \dots, \kappa$, assume that $\mathbb{E}[|\langle\theta, x_s\rangle - y_s|^2] \le \sigma^2_{noise}$. Let $P \in [0, 1]$ be arbitrary and let $Q_s(x_s) \sim$ Bernoulli(P) independently for all $s \in [n]$. Let $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$ and $\widehat{\Sigma}_P = \frac{1}{\kappa}\sum_{s=1}^{\kappa} P(\widetilde{x}_s)\widetilde{x}_s\widetilde{x}_s^\top$. Assume that Σ_P is invertible and that there exists $\gamma \ge 0$ such that $(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P \preceq (1+\gamma)\Sigma_P$. For a given finite set $\mathcal{V} \subset \mathbb{R}^d$ define*

$$w_v = \mathrm{Catoni}(\{\langle v, \widehat{\Sigma}_P^{-1}Q_s(x_s)x_s y_s\rangle\}_{s=1}^n),$$

If $\widehat{\theta} = \arg\min_\theta \max_v \frac{|w_v - \langle\theta, v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}}$ and $n \ge 4\log(2|\mathcal{V}|/\delta)$, then with probability at least $1-\delta$, it holds that

$$|\langle v, \widehat{\theta} - \theta\rangle| \le 4\left(\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}\sqrt{\log(2\,|\mathcal{V}|\,/\delta)}$$

We first state an intermediate matrix lemma before the proof of Proposition 3.

Lemma 19. *Assume that Σ_P is invertible and that there exists $\gamma \in [0, 1/2]$ such that $(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P \preceq (1+\gamma)\Sigma_P$. Then for any $v \in \mathcal{V}$*

$$\|v\|^2_{\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}} \le \frac{1}{(1-\gamma)^2}\|v\|^2_{\Sigma_P^{-1}}.$$

and

$$\|v\|_{(I-\Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})^2} \le \sqrt{1 - \frac{2}{1+\gamma} + \frac{1}{(1-\gamma)^2}}\|v\|_2 \le \sqrt{10\gamma}\|v\|_2.$$

Proof. We know that taking the inverse of two ordered positive definite matrices will flip the order, so here

$$\frac{1}{(1+\gamma)}\Sigma_P^{-1} \preceq \widehat{\Sigma}_P^{-1} \preceq \frac{1}{(1-\gamma)}\Sigma_P^{-1}.$$

$(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P$ implies that for all $u \in \mathbb{R}^d$ holds $u^\top\Sigma_P u \le 1/(1-\gamma)u^\top\widehat{\Sigma}_P u$. So taking $u = \widehat{\Sigma}_P^{-1}v$, we get $v^\top\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}v \le 1/(1-\gamma)v^\top\widehat{\Sigma}_P^{-1}v$. Conclusion

$$v^\top\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}v = \frac{1}{1-\gamma}v^\top\widehat{\Sigma}_P^{-1}v \le \frac{1}{(1-\gamma)^2}v^\top\Sigma_P^{-1}v$$

hence the first result of Lemma 19.
For the second one, we get

$$\begin{aligned}
\|v\|^2_{\left(I-\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}\right)^2} &= v^\top \left(I-\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}\right)^2 v \\
&= \|v\|_2^2 - 2v^\top\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}v + v^\top\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}v \\
&\overset{\text{(i)}}{\leq} \|v\|_2^2 - \frac{2}{1+\gamma}\|v\|_2^2 + \frac{1}{1-\gamma}v^\top\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}v \\
&\leq \|v\|_2^2 - \frac{2}{1+\gamma}\|v\|_2^2 + \frac{1}{(1-\gamma)^2}\|v\|_2^2 \qquad \text{(Since } \hat{\Sigma}_P \preceq \tfrac{1}{1-\gamma}\Sigma_P\text{)} \\
&\leq \left(1-\frac{2}{1+\gamma}+\frac{1}{(1-\gamma)^2}\right)\|v\|_2^2 \\
&\overset{\text{(ii)}}{\leq} 10\gamma\|v\|_2^2 .
\end{aligned}$$

The inequality (i) above holds because $\frac{1}{1+\gamma}\Sigma_P^{-1} \preceq \hat{\Sigma}_P^{-1}$ and $(1-\gamma)\Sigma_P \preceq \hat{\Sigma}_P \implies \Sigma_P \preceq \frac{1}{1-\gamma}\hat{\Sigma}_P$.
The inequality (ii) above holds because for $\gamma \in \left[0, \frac{1}{2}\right]$, we have

$$1-\frac{2}{1+\gamma}+\frac{1}{(1-\gamma)^2} \leq 1-2(1-\gamma)+(1+2\gamma)^2 \leq 10\gamma.$$

Taking square root on both sides gives us the results. □

Proof of Proposition 3. This proof is analogous to the proof of Proposition 1. We first note that

$$\begin{aligned}
\max_{v\in\mathcal{V}}\frac{|\langle\widehat{\theta},v\rangle-\langle\theta,v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} &= \max_{v\in\mathcal{V}}\frac{|\langle\widehat{\theta},v\rangle-w_v+w_v-\langle\theta,v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&\leq \max_{v\in\mathcal{V}}\frac{|\langle\widehat{\theta},v\rangle-w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} + \max_{v\in\mathcal{V}}\frac{|w_v-\langle\theta,v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&= \min_{\theta'}\max_{v\in\mathcal{V}}\frac{|\langle\theta',v\rangle-w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} + \max_{v\in\mathcal{V}}\frac{|w_v-\langle\theta',v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&\leq 2\max_{v\in\mathcal{V}}\frac{|\langle\theta,v\rangle-w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}}
\end{aligned}$$

So it suffices to show that each $|\langle\theta,v\rangle - w_v|$ is small. We begin by fixing some $v \in \mathcal{V}$ and bounding the variance of $v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s$ for any $s \leq n$ which is necessary to use the robust estimator. Note that

$$\begin{aligned}
\mathbb{V}\text{ar}_{x_s\sim\nu,Q_s(x_s)\sim P(x_s)}(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s) =& \mathbb{E}_{x_s\sim\nu,Q_s(x_s)\sim P(x_s)}[(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s)^2] \\
&- \mathbb{E}_{x_s\sim\nu,Q_s(x_s)\sim P(x_s)}[v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s]^2
\end{aligned}$$

which means we can drop the second term to bound the variance by

$$\begin{aligned}
&\mathbb{E}_{x_s\sim\nu,Q_s(x_s)\sim P(x_s)}[\left((v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s\right)^2] \\
&= \mathbb{E}_{x_s\sim\nu,Q_s(x_s)\sim P(x_s)}[\left(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_s(x_s^\top\theta+\xi_s)\right)^2] \\
&= \mathbb{E}_{x_s\sim\nu}\left[\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\left(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_s(x_s^\top\theta)\right)^2] + \mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\left(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_s\right)^2\xi_t^2]\right] \\
&\leq \mathbb{E}_{x_s\sim\nu}\left[B^2\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\left(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_s\right)^2] + \sigma^2\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\left(v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_s\right)^2]\right] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sx_s^\top Q_s(x_s)\widehat{\Sigma}_P^{-1}v]\right] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[v^\top\widehat{\Sigma}_P^{-1}Q_s(x_s)x_sx_s^\top\widehat{\Sigma}_P^{-1}v]\right] \\
&\leq \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)v^\top\widehat{\Sigma}_P^{-1}P(x_s)x_sx_s^\top\widehat{\Sigma}_P^{-1}v\right],
\end{aligned}$$

where we used that $Q_s^2(x_s) = Q_s(x_s)$. Thus, we have with Lemma 19

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s) &\le (B^2+\sigma^2) v^\top \widehat{\Sigma}_P^{-1} \mathbb{E}_{x_s\sim\nu}[P(x_s)x_s x_s^\top] \widehat{\Sigma}_P^{-1} v \\
&= (B^2+\sigma^2)\|v\|^2_{\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}} \\
&\le \frac{B^2+\sigma^2}{(1-\gamma)^2}\|v\|^2_{\Sigma_P^{-1}}.
\end{aligned}$$

We have

$$\begin{aligned}
|\langle\theta_*, v\rangle - w_v| &= |\langle\theta_*, v\rangle - \mathbb{E}[v^\top\widehat{\Sigma}_P^{-1}P(x_1)x_1y_1] + \mathbb{E}[v^\top\widehat{\Sigma}_P^{-1}P(x_1)x_1y_1] - w_v| \\
&\le |\langle\theta_*, v\rangle - \mathbb{E}[v^\top\widehat{\Sigma}_P^{-1}P(x_1)x_1y_1]| \\
&\quad + |\mathrm{Catoni}(\{\langle v, \widehat{\Sigma}_P^{-1}Q_s(x_s)x_sy_s\rangle\}_{s=1}^n) - \mathbb{E}_{X\sim\nu}[v^\top\widehat{\Sigma}_P^{-1}P(X)XY]|.
\end{aligned}$$

We now recall that we can write $y_t = x_t^\top\theta_* + \xi_t$ where ξ_t is a mean-zero, independent random variable with variance at most σ^2. Thus, using Cauchy-Schwarz and applying Lemma 19, we get

$$\begin{aligned}
|\langle\theta_*, v\rangle - \mathbb{E}[v^\top\widehat{\Sigma}_P^{-1}P(x_1)x_1y_1]| &= |v^\top\theta_* - v^\top\widehat{\Sigma}_P^{-1}\Sigma_P\theta_*| \\
&= |v^\top(I - \widehat{\Sigma}_P^{-1}\Sigma_P)\theta_*| \\
&= |v^\top\Sigma_P^{-1/2}(I - \Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})\Sigma_P^{1/2}\theta_*| \\
&\le \|\Sigma_P^{-1/2}v\|\ \|\Sigma_P^{1/2}\theta_*\|_{(I-\Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})^2} \\
&\le \sqrt{10\gamma}\|\Sigma_P^{-1/2}v\|\ \|\Sigma_P^{1/2}\theta_*\| \\
&= \sqrt{10\gamma}\|v\|_{\Sigma_P^{-1}}\|\theta_*\|_{\Sigma_P}.
\end{aligned}$$

By using the property of Catoni estimator stated in Definition 2, we have

$$\begin{aligned}
&|\langle\theta_*, v\rangle - w_v| \\
\le&|\mathrm{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_sy_s\rangle\}_{s=1}^n) - \mathbb{E}[\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_sy_s\rangle]| \\
&\quad + \sqrt{10\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}} \\
\le&\sqrt{2}\sqrt{(\mathbb{V}\mathrm{ar}(\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_sy_s\rangle))\,\frac{\log(\frac{2}{\delta})}{n/2}} \\
&\quad + \sqrt{10\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}} \\
&\qquad\qquad \text{(with probability at least } 1-\delta \text{ if } n \ge 4\log(2/\delta)) \\
\le&\left(\sqrt{4}\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{10n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}\sqrt{\frac{\log(\frac{2}{\delta})}{n}} \\
=&\left(\sqrt{4}\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{10n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}\sqrt{\log(2/\delta)}.
\end{aligned}$$

Finally, the proof is complete by taking union bounding over all $v \in \mathcal{V}$. □

Proof of Lemma 17. Most of this proof is exactly the one of Section B.1 and Section B.1.1 so we only state the concentration bound. For any $\mathcal{V} \subseteq \mathcal{Z}$ and $z, z' \in \mathcal{V}$ define

$$\mathcal{E}_{z,z',\ell}(\mathcal{V}) = \{|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_*\rangle| \le \epsilon_\ell\}$$

where $\widehat{\theta}_\ell(\mathcal{V})$ is the estimator that would be constructed by the algorithm at stage ℓ with $\mathcal{Z}_\ell = \mathcal{V}$. Naturally we want to apply Proposition 3 with τ labeled samples to obtain that $\mathcal{E}_{z,z',\ell}(\mathcal{V})$ holds with probability at least $1 - \frac{\delta}{2\ell^2|\mathcal{Z}|^2}$. Note that as Lemma 14 gives $P(x) \ge \mu/3$ so

$$\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top] \ge \frac{\mu}{3}\mathbb{E}_{X\sim\nu}[XX^\top]$$

Σ_P is invertible.

Defining $\delta_0 := \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ and setting $\kappa \geq 2c_{\delta_0} \max\{1, 20\|\theta_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}$ where we recall that was defined $c_\delta = K^2_{\psi_2}(\sqrt{d \ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})$, Lemma 18 leads to

$$\frac{c_{\delta_0}}{\kappa} \leq \frac{1}{2} \min\left\{1, \frac{1}{20\|\theta_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]}}\right\}$$

so that we can set $\gamma = c_{\delta_0}/(\tau\kappa)$ in the bound of Proposition 3 to get

$$\sqrt{10\tau\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]} \leq \frac{1}{2}$$

and

$$\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} \leq 2\sqrt{B^2+\sigma^2}$$

So for $\delta_0 = \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ the event $\widetilde{\mathcal{E}}_{\text{cov}}$ defined as

$$\widetilde{\mathcal{E}}_{\text{cov}} := \left\{\left(1 - \frac{c_{\delta_0}}{\sqrt{\kappa}}\right) x^\top \Sigma_P x \leq x^\top \widehat{\Sigma}_P x \leq \left(1 + \frac{c_{\delta_0}}{\sqrt{\kappa}}\right) x^\top \Sigma_P x\right\}.$$

happen with probability at least $1 - \delta_0$.

Now, let us for now condition on $\widetilde{\mathcal{E}}_{\text{cov}}$. For fixed $\mathcal{V} \subset \mathcal{Z}$ and $\ell \in \mathbb{N}$ we apply Proposition 3, instantiating the arbitrary P to $\widehat{P}_\ell$ (obtained with OPTIMIZEDESIGN, recall Section D.1) so that with probability at least $1 - \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ we have that for any $z, z' \in \mathcal{V}$ holds that the event $\widetilde{\mathcal{E}}_{\text{RIPS},z,z'}$ defined as

$$\widetilde{\mathcal{E}}_{\text{RIPS},z,z'} := \Bigg\{|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_*\rangle| \\ \leq 2\|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}} \left(4\sqrt{B^2+\sigma^2} + 1\right) \sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)}\Bigg\}$$

happen with probability at least $1 - \delta_0$.

So with probability at least $1 - \mathbb{P}(\widetilde{\mathcal{E}}^c_{\text{RIPS},z,z'}) - \mathbb{P}(\widetilde{\mathcal{E}}^c_{\text{cov}}) \geq 1 - \frac{\delta}{4\ell^2|\mathcal{Z}|^2} - \frac{\delta}{4\ell^2|\mathcal{Z}|^2} = 1 - \frac{\delta}{2\ell^2|\mathcal{Z}|^2}$, both events hold and we have that for any $z, z' \in \mathcal{V}$ holds

$$\begin{aligned}|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_*\rangle| &\leq 2\|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}} \left(4\sqrt{B^2+\sigma^2} + 1\right) \sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)} \\ &\leq 2(1+\varepsilon) \left(4\sqrt{B^2+\sigma^2} + 1\right) \|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}} \sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)} \\ &\leq \epsilon_\ell.\end{aligned}$$

where we used the property of $\widehat{P}_\ell$ as detailed in Section D.1 to conclude. □

Proof of Theorem 7. The total number of labels requested after L rounds is equal to $\sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} \widehat{P}_\ell(x_t)$. Again by Freedman's inequality we have that

$$\sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} \widehat{P}_\ell(x_t) \leq 2 \sum_{\ell=1}^{L} \tau \mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)|\mathcal{Z}_\ell] + \log(1/\delta)$$

From Theorem 4, it holds for any ℓ that $\mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)] \leq \mathbb{E}_{X\sim\nu}[\widetilde{P}_\ell(X)] + 4\sqrt{\mu}$ where $\widetilde{P}_\ell$ is the optimal solution to problem (20). So now, for some $\widetilde{\tau}$, we want to relate $\mathbb{E}_{X\sim\nu}[\widetilde{\tau}\widetilde{P}_\ell(X)]$ to $\mathbb{E}_{X\sim\nu}[\tau P_\ell(X)]$ where P_ℓ is the solution of problem (4). To do so, we rewrite problem (4) and problem (20) as

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[\tau P(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}\left[\tau P(X) XX^\top\right]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq \tau P(x) \leq \tau, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{32}$$

and

$$\begin{array}{rl} \min_P & \mathbb{E}_{X\sim\nu}\left[\widetilde{\tau}P(X)\right] \\ \text{subject to} & y^\top \mathbb{E}_{X\sim\nu}[\widetilde{\tau}P(X)XX^\top]^{-1}y \le c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \le \widetilde{\tau}P(x) \le \widetilde{\tau}(1-\mu_b), \quad \forall x \in \mathcal{X}. \end{array} \tag{33}$$

where problem (32) is equivalent to problem (4) and problem (33) is equivalent to problem (20). Thus taking $\widetilde{\tau} = \frac{\tau}{1-\mu_b}$, problem (33) becomes

$$\begin{array}{rl} \min_P & \mathbb{E}_{X\sim\nu}\left[\frac{\tau}{1-\mu_b}P(X)\right] \\ \text{subject to} & y^\top \mathbb{E}_{X\sim\nu}[\frac{\tau}{1-\mu_b}P(X)XX^\top]^{-1}y \le c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \le \frac{\tau}{1-\mu_b}P(x) \le \tau, \quad \forall x \in \mathcal{X}. \end{array}$$

which, using $Q = \frac{P}{1-\mu_b}$ is equivalent to

$$\begin{array}{rl} \min_Q & \mathbb{E}_{X\sim\nu}\left[\tau Q(X)\right] \\ \text{subject to} & y^\top \mathbb{E}_{X\sim\nu}[\tau Q(X)XX^\top]^{-1}y \le c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \le \tau Q(x) \le \tau, \quad \forall x \in \mathcal{X}. \end{array} \tag{34}$$

And we can now see that (34) and (32) are the same optimization problem. And Q_ℓ^* the solution of (34) is equal to $\frac{\widetilde{P}_\ell}{1-\mu_b}$. Thus the result $\mathbb{E}_{X\sim\nu}\left[\widetilde{\tau}\widetilde{P}_\ell(X)\right] = \mathbb{E}_{X\sim\nu}\left[\tau P_\ell(X)\right]$.

Remains to bound $\sum_{\ell=1}^L \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)]$ where

$$\begin{aligned} &\sum_{\ell=1}^L \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell] \\ &= \sum_{\ell=1}^L \left[\min_{P:\mathcal{X}\to[0,1]} \tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1\right], \end{aligned}$$

where $\beta_{\delta,\ell}$ is defined in Section D.1 as

$$\beta_{\delta,\ell} := 4(1+\varepsilon)^2\left(4\sqrt{B^2+\sigma^2}+1\right)^2\log(4\ell^2|\mathcal{Z}|^2/\delta).$$

As in the case where the distribution ν is known (Section B.1), we use Lemma 3 to bound $\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell}$ by $\max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}64\beta_{\delta,L}$. Last, the reparameterization of Proposition 2 also applies here.

In the unlabeled sample complexity, we get an additional $L\kappa = L\lceil 2K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})\max\{1, 20\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}\rceil$ term from the estimation of the covariance matrix. Last, we get an additional $L(K+u)$, where K and u are such that

$$K \ge \widetilde{O}\left(\frac{|\mathcal{Z}|^3\kappa(\Sigma)^2\|\Lambda^*\|_2^8 M^{16}}{\beta^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2, \quad u \ge \widetilde{O}\left(\frac{\kappa(\Sigma)^2\|\Lambda^*\|_2^6 M^{16}}{\beta^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

from the sample complexity of the subroutine. □

E Classification

In this section we adopt the implementation described in Section B.1. As described in the text, given a distribution $\pi \in \Delta_\mathcal{X}$, and a class of hypothesis $\mathcal{H}$, we can reduce classification to linear bandits by setting $\theta^* = [\theta_x^*]_{x\in\Delta_\mathcal{X}}$ where $\theta_x^* = 2\eta(x) - 1$, and $\mathcal{Z} := \{z^{(h)}\}_{h\in\mathcal{H}} \subset [0,1]^{|\mathcal{X}|}$ where $z_x^{(h)} = \pi(x)\mathbf{1}\{h(x) = 1\}$. With the quantities computed in Section 3, we now prove Theorem 3.

Proof of Theorem 3. We consider a slightly modified version of Algorithm 1 where we stop at round L where $L_\epsilon = \lceil\log_2(4/\epsilon)\rceil$ and return $\arg\max_{z^{(h)}\in\mathcal{Z}_\ell}\langle z^{(h)}, \widehat{\theta}_\ell\rangle$. By an identical analysis to that in the proof of Theorem 2, we are guaranteed that $h \in \mathcal{S}_\ell$, i.e. $R_\nu(h) - R_\nu(z^*) = \langle z^* - z, \theta_*\rangle \le 4\epsilon_\ell$.

In addition the analysis of the sample complexity given there immediately gives the first part of the theorem.

It remains to bound the sample complexity in terms of the disagreement coefficient. The total sample complexity is given by,

$$\sum_{\ell=1}^{L}\left[\min_{P:\mathcal{X}\to[0,1]}\tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z\in\mathcal{S}_\ell}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_\delta\le 1\right]$$

where we recall $\beta_\delta = 2048\log(2L^2|\mathcal{H}|/\delta)$ since we can take $B=1$ and $\sigma=1$.

We recall the proof of Theorem 2. From the proof, we see that with probability greater than $1-\delta$, our sample complexity is obtained by summing up to round L

$$\sum_{\ell=1}^{L}\left[\min_{P:\mathcal{X}\to[0,1]}\tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z\in\mathcal{S}_\ell}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_\delta\le 1\right]$$

By proposition 2 this is equivalent to

$$\sum_{\ell=1}^{L}\left[\min_{\lambda\in\Delta_X}\rho_\ell(\lambda)\beta_\delta \quad \text{subject to} \quad \left\|\frac{\lambda}{\nu}\right\|_\infty\rho_\ell(\lambda)\beta_\delta\le\tau\right], \text{ where } \rho_\ell(\lambda) := \max_{z\in\mathcal{S}_\ell}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\epsilon_\ell^2}.$$

Define

$$A_\ell = \{x\in\mathcal{X}:\exists h, h(x)\neq h^*(x), R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell\}, \ell\le L$$

and let $\lambda_\ell = \dfrac{\mathbf{1}\{x\in A_\ell\}\nu(x)}{\mathbb{E}[\mathbf{1}\{x\in A_\ell\}]}$, so $\left\|\dfrac{\lambda}{\nu}\right\|_\infty = \dfrac{1}{\mathbb{E}[\mathbf{1}\{x\in A_i\}]}$.

We first argue that λ_ℓ is feasible for the previous program. Note,

$$\begin{aligned}
\rho_\ell(\lambda_\ell) &= \max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{\mathbb{E}_{X\sim\nu}[\frac{\mathbf{1}\{h(x)\neq h^*(x)\}}{\lambda_\ell(x)/\nu(x)}\}]}{\epsilon_\ell^2}\\
&\overset{\text{(i)}}{=} \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\epsilon_\ell^2}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon_\ell^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{(4\epsilon_\ell)^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\overset{\text{(ii)}}{\le} \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h\in H}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le 16\mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\rho(\nu,\epsilon)
\end{aligned}$$

where the equality (i) holds because the following is true when we only consider h such that $R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell$

$$\frac{\mathbf{1}\{h(x)\neq h^*(x)\}}{\mathbf{1}\{x:\exists h, h(x)\neq h^*(x), (R_\nu(h)-R_\nu(h^*))\le 4\epsilon_\ell\}} = \mathbf{1}\{h(x)\neq h^*(x)\}.$$

The inequality (ii) above is true because $4\epsilon_\ell \ge \epsilon$. Thus we see that $\rho_\ell(\lambda_\ell)\|\lambda/\nu\|_\infty\beta_\delta \le 16\rho(\nu,\epsilon)\beta_\delta \le \tau$. It remains to argue about the disagreement coefficient. Firstly note that for any h such that $R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell$.

$$d_\nu(h,h^*) = \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(X)\neq h^*(X)\}] \le \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(X)\neq Y\}] + \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h^*(X)\neq Y\}] \tag{35}$$

$$\le R_\nu(h) + R_\nu(h^*) \tag{36}$$

$$\le 2R_\nu(h^*) + 4\epsilon_\ell \tag{37}$$

Using this we see that,

$$
\begin{aligned}
&\min_{\lambda\in\Delta} \rho_\ell(\lambda) \text{ subject to } \rho_\ell(\lambda)\|\lambda/\nu\|_\infty\beta_\delta \le \tau \\
&\qquad \le \rho_\ell(\lambda_\ell)\beta_\delta && \text{(since } \lambda_\ell \text{ is feasible.)} \\
&\qquad \le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}] \max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell} \frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\epsilon_\ell^2}\beta_\delta \\
&\qquad && \text{(imitating the above computation)} \\
&\qquad \le \frac{(2R(h^*)+4\epsilon_\ell)\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{\epsilon_\ell^2}\beta_\delta \\
&\qquad && \text{(Equation (35))} \\
&\qquad \le \beta_\delta \begin{cases} \frac{9R(h^*)^2}{\epsilon_\ell^2}\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell} & 4\epsilon_\ell \le R(h^*) \\ \frac{144\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell} & 4\epsilon_\ell > R(h^*) \end{cases} \\
&\qquad \le \left(\frac{9R(h^*)^2}{\epsilon_\ell^2}+144\right)\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{\ell=1}^{L}\left[\min_{\lambda\in\Delta_X}\rho_\ell(\lambda)\beta_\delta \quad \text{subject to} \quad \left\|\frac{\lambda}{\nu}\right\|_\infty \rho_\ell(\lambda)\beta_\delta \le \tau\right] \\
&\le \sum_{\ell=1}^{L}\rho_\ell(\lambda_\ell)\beta_\delta \\
&\le \sum_{\ell=1}^{L}\left(\frac{9R(h^*)^2}{\epsilon_\ell^2}+144\right)\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta \\
&\le \log_2\left(\frac{4}{\epsilon}\right)\sup_{\ell\le L}\left(\frac{9R(h^*)^2}{\epsilon_\ell^2}+144\right)\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta \\
&\le \log_2\left(\frac{4}{\epsilon}\right)\left(\frac{36R(h^*)^2}{\epsilon^2}+144\right)\sup_{\ell\le L}\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta \\
&\le 36\log_2\left(\frac{4}{\epsilon}\right)\left(\frac{R(h^*)^2}{\epsilon^2}+4\right)\sup_{\xi\ge\epsilon}\theta^*(2R(h^*)+\xi,\nu)\beta_\delta
\end{aligned}
$$

from which the result follows.

□