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. Speci cally, our algorithm includes a tuning parameter, call it μ, 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 L, and the amount of unlabelled data U that is necessary before the metric of interest can be non-trivial. We prove that each possible setting of μ parametrizes all possible trade-offs between unlabeled and labeled data.

Our work is perhaps closest to the streaming setting for agnostic active classi cation [8, 15] where each x_s is drawn i.i.d. from an underlying distribution 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 coef cient [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 speci c parametric form to their label distribution unlike our setting which is agnostic. Other works have characterized selective sampling for classi cation 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 , but we will evaluate the excess risk of our resulting classi er on a different stream [22, 23, 26].

Best-Arm Identi cation and Online Experimental Design. Our techniques are based on experimental design methods for best-arm identi cation in linear bandits, see [4, 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. Speci cally, 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 Identi cation

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 $x_t \overset{iid}{\sim} \mathcal{D}$ with support $\mathcal{D}(\mathcal{X}) = \mathcal{X}$
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 τ and output some $z \in \mathcal{Z}$

If the player stops at time U after observing $L = \sum_{t=1}^{U} Q_t$ 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 U, L.

This paper studies the relationship between U and L in the context of necessary and suf cient conditions to identify z^* with probability at least $1 - \delta$. Clearly U must be "large enough" for z^* to

be identi able even if all labels are requested (i.e., $L = U$). But if 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 L alone. Intuitively, L should decrease as U increases, but how?

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

De nition 1. For any $\in (0;1)$ we say a selective sampling algorithm is -PAC for $\in \mathcal{A}_X$ if for all $\in \mathbb{R}^d$ the algorithm terminates at time U which is nite almost surely and outputs $\arg\max_{z \in Z} \langle z; \rangle$ with probability at least 1.

2.1 Optimal design

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

$$(\ ) := \max_{z \in Z \setminus \{z\}} \frac{\|z - z\|^2_{E_X[XX^\top]^{-1}}}{\langle z - z; \rangle^2}. \tag{1}$$

Intuitively, $(\ )$ captures the number of labeled examples drawn from distribution to identify z. Speci cally, for any $\geq (\ )\log(|Z|/)$, if $x_1, \ldots, x_{\ } \sim$ and $y_i = \langle x_i; \rangle + \ _i$ where $\ _i$ is iid 1 sub-Gaussian noise, then there exists an estimator $\hat{} = \hat{}(\{(x_i; y_i)\}_{i=1}^{\ })$ such that $\langle \hat{}; z \rangle > \max_{z \in Z \setminus \{z\}} \langle \hat{}; z \rangle$ with probability at least 1 [11]. In particular, $\geq (\ )\log(|Z|/)$ samples suf ce to guarantee that $\arg\max_{z \in Z} \langle \hat{}; z \rangle = \arg\max_{z \in Z} \langle ; z \rangle =: z$.

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

$$P = \arg\min_{P : X \to [0;1]} E_X[P(X)] \quad \text{subject to} \quad \max_{z \in Z \setminus \{z\}} \frac{\|z - z\|^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z - z; \rangle^2} \leq 1 \tag{2}$$

for $= \log(|Z|/)$ 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 $\in \mathcal{A}_X$ instead of $P : X \to [0;1]$ as

$$E_X[P(X)] = \min_{\in \mathcal{A}_X} (\ ) \quad \text{subject to} \quad k = k_1 (\ )$$

where $k = k_1 = \max_{x \in X} (x) = (x)$, as shown in Proposition 2. Note that as $\to 1$ the constraint becomes inconsequential. Also notice that $(\ )$ 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 $\in (0;1)$, $X; Z \subset \mathbb{R}^d$, and $\in \mathbb{R}^d$. Any selective sampling algorithm that is -PAC for $\in \mathcal{A}_X$ and terminates after drawing U unlabelled examples from and requests the labels of just L of them satis es

- $E[U] \geq (\ )\log(1/)$, and
- $E[L] \geq \min_{\in \mathcal{A}_X} (\ )\log(1/)$ subject to $E[U]\ k = k_1 (\ )\log(1/)$.

The rst part of the theorem quanti es the number of rounds or unlabelled data that any algorithm must observe before it could hope to stop and output z correctly. The second part describes a

trade-off between U and L. One extreme is if $E[U] \to 1$, which effectively removes the constraint so that the number of observed labels must scale like $\frac{24}{\ } \ _x \ (\ ) \log(1/ \ )$. Note that this is precisely the number of labels required in the pool-based setting where the agent can choose $x \in X$ that she desires at each time (e.g. [11]). In the other extreme, $E[U] = \ (\ ) \log(1/ \ )$ so that the constraint in the label complexity $E[L]$ is equivalent to $(\ ) \ k \ = \ k_1 \ (\ )$. This implies that the minimizing $\ $ must either stay very close to, or must obtain a substantially smaller value of $(\ )$ relative to $(\ )$ to account for the in ation factor $k = \ k_1$. 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 know $\ $ exactly. The proof of Theorem 1 relies on standard techniques from best arm identi cation 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 $\ \in (0;1)$, $X ; Z \subseteq R^d$, and $\ \in R^d$. Let $\ = \min_{z \in Z \setminus \{ \ z \ \}} h \ , z \ - \ z; \ i$ and $\ / \log(\log(\frac{1}{\ })|Z| / \ )$ where the precise constant is given in the appendix. For any $\ (\ )$ there exists a $\ $-PAC selective sampling algorithm that observes U unlabeled examples and requests just L labels that satis es with probability at least $1 \ $*

- $U \ \ \log_2(\frac{4}{\ })$, and
- $L \ \ 3 \log_2(\frac{4}{\ }) \min_{\frac{24}{\ } \ _x} \ (\ ) \quad$ subject to $\quad k \ = \ k_1 \ (\ )$.

Aside from the $\log(\frac{1}{\ })$ factor and the $\log(|Z|)$ that appears in the $\ $ term, this nearly matches the lower bound. Note that the parameter $\ $ parameterizes the algorithm and makes the trade-off between U and 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 identi cation. Note that it takes a con dence level $\ \in (0;1)$ and a parameter $\ $ 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 identi cation 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 Identi cation

1: Input $Z \subseteq R^d$, $\ \in (0;1)$;
2: while $|Z_\ell| > 1$ do
3: Let $P_\ell, \hat{\ }_\ell \leftarrow$ OPTIMIZEDESIGN$(Z_\ell; 2^{-\ell}; \ )$ // $\hat{\ }_\ell$ approximates $E_{X \sim P_\ell}[P_\ell(X)XX^\top]$
4: for $t = (\ell - 1) \ + 1; \ldots; \ell$ do
5: Nature reveals x_t drawn iid from $\ $ (with support R^d)
6: Sample $Q_t(x_t) \sim$ Bernoulli$(P_\ell(x_t))$. If $Q_t = 1$ then observe y_t // $E[y_t | x_t] = \langle \ , x_t \rangle$
7: end for
8: Let $\hat{\ }_\ell \leftarrow$ RIPS$(\{ \hat{\ }_\ell^{-1} Q_s(x_s) x_s y_s \}_{s=(\ell-1)\ +1}^{\ell}, Z, Z, \ )$ // $\hat{\ }_\ell$ approximates
9: $Z_{\ell+1} = Z_\ell \setminus \{ z \in Z_\ell : \max_{z' \in Z_\ell} \langle z' - z; \hat{\ }_\ell \rangle > 2^{-\ell} \}$
10: end while

In round ℓ, the algorithm maintains an active set $Z_\ell \subseteq Z$ with the guarantee that each remaining $z \in Z_\ell$ satis es, $\langle z - z; \ \rangle \ \ \

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 . Thus, the total number of samples should be bounded by the ideal sample complexity times the number of rounds, which is $O(\log(\ ^{1}))$. We will return to implementation issues in the next section.

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

$$|\langle z \quad z; \hat{b}\rangle \quad | \quad O\left(\|z \quad z\|_{E_X[\hat{P}(X)XX^\top]^{-1}} \sqrt{\log(2`^2|Z|^2/\ )} \quad 2^{`}\right)$$

for all $z \in Z_`$ simultaneously with probability at least 1 . Thus, the nal line of the algorithm eliminates any $z \in Z_`$ such that there exists another $z^0 \in Z_`$ (think $z$ ) that satis esh $\hat{b}; z^0 \quad z\rangle > 2^{`}$. The process continues until $Z_{`+1} = \{z\}$.

2.4 Implementation of OPTIMIZE DESIGN

For the subroutine OPTIMIZEDESIGN passed $(Z_`; \ ; \ )$ the next best thing to computing Equation 2 with the denominator of the constraint replaced with $\max\{\ ^{2}; \langle h \ ; z \quad z_i\rangle^2\}$, is to compute

$$P = \underset{P:X\to[0;1]}{\operatorname{argmin}} \ E_X \quad [P(X)] \text{ subject to } \max_{z;z^0\in Z_`} \frac{\|z \quad z^0\|^2_{E_X[P(X)XX^\top]^{-1}}}{2} \quad 1 \qquad (3)$$

and $_P = E_X \quad [P(X)XX^\top]$ for an appropriate choice of $= (\log(\ |Z|/\ ))$. To see this, rstly, any $z \in Z$ with gap $\langle h \ ; z \quad z_i\rangle$ that we could accurately estimate would not be included in $Z_`$, 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 $\max_{z;z^0}\|z \quad z^0\| \quad \max_z 2\|z \quad z\|$ $\max_{z;z^0} 2\|z \quad z^0\|$. Assuming we could solve this directly and compute $_P = E_X \quad [P(X)XX^\top]$, we can obtain the result of Theorem 2 (proven in the Appendix).

However, even if we knew exactly, the optimization problem of Equation 3 is quite daunting as it is a potentially in nite dimensional optimization problem over X. Fortunately, after forming the Lagrangian with dual variables for each $z^0 \in Z \quad Z$, optimizing the dual amounts to a nite dimensional optimization problem over the nite number of dual variables. Moreover, this optimization problem is maximizing a simple expectation with respect to x 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 x each iteration, it suf ces to "replay" a sequence of x's from historical data. Summing up, this construction allows us to compute a satisfactory P and avoid both an in nite-dimensional optimization problem and requiring knowledge of (as long as historical data is available).

Meanwhile, with historical data, we can also empirically compute $E_X \quad [P(X)XX^\top]$. Historical data could mean of ine samples from 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 of ine data have been used in other works in the streaming setting e.g. [15].

3 Selective Sampling for Binary Classi cation

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

1. Nature reveals x_t
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 τ and output some $\hat{h} \in \mathcal{H}$.

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

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 $z_x^{(h)} := \eta(x) \mathbb{1}\{h(x) = 1\}$ for $z^{(h)} = [z_x^{(h)}]_{x \in \mathcal{X}}$. Moreover, define $\mu := [\mu_x]_{x \in \mathcal{X}}$ where $\mu_x := 2\eta(x) - 1$. Then

$$R(h) = \mathbb{E}_{X \sim \mathcal{D}_X, Y \sim \eta(X)}[\mathbb{1}\{Y \neq h(X)\}] = \sum_{x \in \mathcal{X}} \eta(x)(\eta(x)\mathbb{1}\{h(x) \neq 1\} + (1 - \eta(x))\mathbb{1}\{h(x) \neq 0\})$$

$$= \sum_{x \in \mathcal{X}} \eta(x)\eta(x) + \sum_{x \in \mathcal{X}} \eta(x)(1 - 2\eta(x))\mathbb{1}\{h(x) = 1\} = c - \langle \mu, z^{(h)} \rangle,$$

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

$$\Delta(\gamma; \varepsilon) := \max_{z \in \mathcal{Z} \setminus \{z^*\}} \frac{\|z^* - z\|_{\mathbb{E}_X[\mathcal{X}\mathcal{X}^\top]^{-1}}^2}{\max\{\gamma, \langle \mu, z^* - z \rangle^2, \varepsilon^2\}} = \max_{h \in \mathcal{H} \setminus \{h^*\}} \frac{\mathbb{E}_X\left[\mathbb{1}\{h(X) \neq h^*(X)\} \frac{\eta(X)}{\eta(X)}\right]}{\max\{(R(h) - R(h^*))^2, \varepsilon^2\}}$$

An important case of the above setting is when $\mathcal{D}_X$ and $\mathcal{D}$ = $\mathcal{D}$, 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 $\mathcal{D}$, finite hypothesis class $\mathcal{H}$, regression function $\eta : \mathcal{X} \to [0, 1]$. Set $\gamma = 0$ and $\varepsilon = 2048 \log(4 \log_2^2(4/\delta)|\mathcal{H}|/\delta)$. Then for $\Delta(\gamma; \varepsilon)$ there exists a selective sampling algorithm that returns $h \in \mathcal{H}$ satisfying $R(h) - R(h^*) \leq \varepsilon$ by observing U unlabeled examples and requesting just L labels such that

- $U \geq \log_2(4/\delta)$
- $L \geq 3\log_2(\frac{4}{\delta}) \min_{\mu \in \mathcal{M}_{\mathcal{X}}} \Delta(\gamma; \varepsilon)$ s.t. $\|\mu - \mu_1\| \leq \Delta(\gamma; \varepsilon)$

with probability at least $1 - \delta$. Furthermore when $\gamma = 0$ and if $\varepsilon \geq 16\Delta(\gamma; \varepsilon)$ we have that

$$L \leq 36\log_2(4/\delta)\frac{R(h^*)^2}{\varepsilon^2} + 4\sup_{h \in \mathcal{H}}(2R(h^*) + \varepsilon; \varepsilon)$$

where $\theta(u; \varepsilon)$ is the disagreement coefficient, defined in Appendix E.

Note that if ε is sufficiently large then the labeled sample complexity we obtain $\min_{\mu \in \mathcal{M}_{\mathcal{X}}} \Delta(\gamma; \varepsilon)$ could be significantly smaller than previous results in the streaming setting, e.g. [16, 8, 15]. 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 in $\mathcal{X}$ if

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

$$\begin{aligned} \min_P \quad & E_X[P(X)] \\ \text{subject to} \quad & y^\top E_X[P(X)XX^\top]^{-1} y \le c^2, \quad \forall y \in Y, \\ & 0 \le P(x) \le 1, \quad \forall x \in X. \end{aligned} \tag{4}$$

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

$$\begin{aligned} \min_P \quad & E_X[P(X) - \lambda_b(\log(1 - P(X)) + \log(P(X)))] \\ \text{subject to} \quad & E_X[P(X)XX^\top] \succeq \frac{1}{c^2} yy^\top, \quad \forall y \in Y. \end{aligned} \tag{5}$$

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

$$L(\Lambda; P) = E_X[P(X) - \lambda_b(\log(1 - P(X)) + \log(P(X)))] - P(X)X^\top \Lambda X + \frac{1}{c^2} \sum_{y \in Y} y^\top \Lambda_y y.$$

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

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

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

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

Note that if $\lambda_b = 0$ (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 de ned by the positive semide nite 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 $\max_{\Lambda_y \succeq 0, \forall y} D(\Lambda) := L(\Lambda; P_\Lambda)$. This is a nite-dimensional optimization problem, and can be solved by projected gradient ascent (or projected stochastic gradient ascent when we have only samples from P). The gradient of $D(\Lambda)$ is

$$\begin{aligned} \nabla_{\Lambda_y} D(\Lambda) &= E_X\left[\left(1 + \frac{\lambda_b}{1 - P_\Lambda(x)} - \frac{\lambda_b}{P_\Lambda(X)}\right) X^\top \Lambda X \nabla_{\Lambda_y} P_\Lambda(X) - P_\Lambda(X)XX^\top\right] + \frac{yy^\top}{c^2} \\ &= \frac{yy^\top}{c^2} - E_X[P_\Lambda(X)XX^\top]. \qquad \text{(Since } P_\Lambda(X) \text{ solves Eq. (6))} \end{aligned}$$

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 $\frac{yy^\top}{c^2} - P_{\hat{\Lambda}(k)}(x_k)x_k x_k^\top$. The adaptive learning rate is chosen by following the discussion in chapter 4.2.1 [1]. 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 $E_X[P(X)]$ 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 n; barrier weight $\beta_b \in (0,1)$
Initialize $\hat{\lambda}_y^{(0)} = 0$ for each $y \in \mathcal{Y}$
for $k = 0, 1, 2, \ldots, K-1$ do
  Sample x_k
  Set $g_{k,y} = \frac{yy^\top}{c^2} - P_{\hat{\lambda}^{(k)}}(x_k) x_k x_k^\top$, where P is defined in Eq. (7)
  Set $\hat{\lambda}_y^{(k+1)} \leftarrow \hat{\lambda}_y^{(k)} + \eta_k g_{k,y}$ for each $y \in \mathcal{Y}$, where $\eta_k = \frac{1}{\sqrt{\frac{1}{2}\sum_{s=1}^{k}\sum_{y \in \mathcal{Y}} \|g_{s,y}\|_2^2}}$
  Update $\hat{\lambda}_y^{(k+1)} \leftarrow \mathcal{S}_+^d(\hat{\lambda}_y^{(k+1)})$ for each $y \in \mathcal{Y}$, 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 $y \in \mathcal{Y}$ and $\hat{\lambda} = \sum_{y \in \mathcal{Y}} \hat{\lambda}_y$
Update $(\hat{\lambda}_y)_{y \in \mathcal{Y}} \leftarrow \operatorname{argmax} \sum_{y \in \mathcal{Y}} y^\top \lambda_y y$, subject to $\sum_{y \in \mathcal{Y}} \lambda_y = \hat{\lambda}$; $\lambda_y \succeq 0, \forall y \in \mathcal{Y}$:
Find $s \leftarrow \operatorname{argmax}_{s \in [0,1]} D_E(s\hat{\lambda})$, where D_E empirically evaluates D using u i.i.d. samples
return $\hat{\pi} = s \sum_{y \in \mathcal{Y}} \hat{\lambda}_y$

Let π^* be an optimal solution for $D(\cdot)$. Intuitively, as long as we run this algorithm with sufficiently large number of iterations K and number of samples n, we can guarantee that $D(\hat{\pi})$ and $D(\pi^*)$ are close enough with high probability, which in turn guarantees that the primal constraints are violated by only a tiny amount and $\mathbb{E}_X P_{\hat{\pi}}(X)$ is close enough to the optimal value. Specifically, we can prove the following theorem.

Theorem 4. Suppose $\|x\|_2 \le M$ for any $x \in \operatorname{supp}(\mathcal{X})$ and $\Lambda = \mathbb{E}_X XX^\top$ is invertible. Let $\lambda^* \in \operatorname{argmax}_{\lambda_y \succeq 0, \forall y \in \mathcal{Y}} D(\lambda)$ and $\kappa(\Lambda) = \frac{\max(\Lambda)}{\min(\Lambda)}$ be its condition number. Assume $\|\Lambda\|_F > 0$ and define $\Delta = \min_{\Delta \in \mathcal{S}^d : \|\Delta\|_F = 1} \mathbb{E}_X \left[(X^\top \Delta X)^2 \right]$, where $\mathcal{S}^d$ is the set of $d \times d$ symmetric matrices.

Then, $\pi^* = \sum_{y \in \mathcal{Y}} \lambda_y^*$ is unique. Further, for any $\epsilon > 0$ and $\delta > 0$, if it holds that $\beta_b = O\left(\sqrt{\frac{\|\Lambda\|_F \, \Delta}{\kappa(\Lambda) M}}\right)\sqrt{(1+\epsilon)\epsilon}$ and

$$K \ge O\left(\frac{|\mathcal{Y}|^3 \, \kappa(\Lambda)^2 \|\Lambda\|_F^8 M^{16} \log(1/\delta)}{\Delta^2 \beta_b^6}\left(\frac{1+\epsilon}{\epsilon}\right)^2\right); \quad n \ge O\left(\frac{\kappa(\Lambda)^2 \|\Lambda\|_F^6 M^{16} \log(1/\delta)}{\Delta^2 \beta_b^6}\left(\frac{1+\epsilon}{\epsilon}\right)^2\right);$$

then, with probability at least $1-\delta$, Algorithm 2 will output $\hat{\pi}$ that satisfies

- $y^\top \mathbb{E}_X \left[P_{\hat{\pi}}(X) XX^\top \right]^{-1} y \le (1+\epsilon)c^2, \quad \forall y \in \mathcal{Y}$.
- $\mathbb{E}_X \left[P_{\hat{\pi}}(X) \right] \le \mathbb{E}_X \left[P_{\pi^*}(X) \right] + 4\sqrt{\epsilon}$, where π^* is the optimal solution to problem (4) with barrier constraint repaced by $\sum_{y \in \mathcal{Y}} P(x) \le 1 - \beta_b, \forall x \in \mathcal{X}$.

The proof is in Appendix C. Although π^* is not exactly the same as the optimal solution of the original problem (4), when β_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 $|\mathcal{Y}| d^2 = O(|\mathcal{Z}|^2 d^2)$ variables, which can be large when we have a large set $\mathcal{Z}$. 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.
- $Z = [e_1; e_2; (\cos(\omega), \sin(\omega))]$ for $\omega = 0.3$, where $e_1; e_2$ are canonical vectors.
- $\mu = 2e_1$ and $y = x^\top \mu + \eta$, where $\eta \sim N(0; 1)$.
- The distribution for streaming measurements $x_t \overset{i.i.d.}{\sim}$ is such that $x_t = (\cos(2\pi I_t/N), \sin(2\pi I_t/N))$ where $I_t \in \{0; \ldots; N-1\}$, $P(I_t = i) \propto \cos(2\pi i/N)^2$, 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 e_1. As argued in [24], an efficient sampling strategy for this problem instance would be to pull arms in the direction of e_2 in order to reduce the uncertainty in the direction of interest, $e_1 \approx (\cos(\omega), \sin(\omega))$. However, the distribution is defined such that it is more likely to receive a vector x_t in the direction of e_1 rather than e_2. Thus, if one seeks a small label complexity, then P should be taken to reject measurements in the direction of 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 P^b. Naive Algorithm uses no selective sampling so that $P^b(x) = 1$ for all x; the Oracle Algorithm uses $P^b = P$ where P is the ideal solution to (2), and Our Algorithm uses the solution to (5) for P^b, where we take $\beta_b = 2 \cdot 10^{-5}$. We swept over the values of ω 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 ω, we plot the average label complexity over 50 repeated trials. (middle) Visualization of $P(x)$ and $\tau(x)$ v.s. x, where x is indexed by I such that $x_I = (\cos(2\pi I/N), \sin(2\pi I/N))$. Here, P is solved with $\beta = 4 \cdot 10^{-5}$ and distribution τ is not normalized. (right) A heat map of $\tau(x)$ 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 ω. 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 L and sample complexity 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(b)), 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. 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. 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 identi cation 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

1 Introduction 1

1.1 Related Work 2

2 Selective Sampling for Best Arm Identi cation 3

2.1 Optimal design 4

2.2 Main results 4

2.3 Selective Sampling Algorithm 5

2.4 Implementation of OptimizeDesign 6

3 Selective Sampling for Binary Classi cation 6

4 Solving the Optimization Problem 7

5 Empirical results 9

6 Conclusion 10

A Selective Sampling Lower Bound 14

A.1 Proof of Theorem 1, part I 14

A.2 Proof of Theorem 1, part II 14

B Selective Sampling Algorithm for Known Distribution 15

B.1 Proof of Theorem 2, upper bound 15

B.1.1 High-probability Events 17

B.2 Technical Lemmas 17

B.2.1 Reparameterization 19

C Analysis of the Optimization Problem 20

C.1 Proof of Theorem 4 20

C.2 Relevant Lemmas 23

C.2.1 Strong Concavity of $\overline{D}(\ )$ 23

C.2.2 Concentration Inequalities 25

C.2.3 Other Lemmas 26

C.2.4 Properties of $\mathbb{P}$ 29

C.3 An Alternative Approach to OptimizeDesign 31

C.3.1 Technical Lemmas 33

D Selective Sampling Algorithm for Unknown Distribution 34

D.1 Statement and proof of Theorem 7 34

D.2 Lemmas for the correctness 35

E Classi cation 40

A Selective Sampling Lower Bound

First, we review the standard argument for best-arm identi cation lower bounds applied to linear bandits. Fix $\ 2\ R^d$ and let $z\ =\ \arg\max_{z2Z}\ hz;\ \ i$. De ne the set $C\ =\ f\ \ 2\ R^d\ :\ 9z\ 2\ Z\ \text{s.t.}\ h;z\ \ \ z\ i\ \ \ 0g$ as those in which z is note the best arm under We now recall the transportation lemma of [17]. Under a -PAC strategy for nding the best arm for the bandit instance $(X;Z;\ \ )$, let T_x denote the random variable which is the number of times x is pulled. In addition let $N_{;x}$ denote the reward distribution of the arm x of X, i.e. $N_{;x}\ =\ N(x^>\ ;1)$. Then for any -PAC algorithm

$$\begin{aligned}
\log(1=2.4\ ) \quad & \min_{2C}\ \sum_{x2X} E[T_x]KL(N\ \ _{;x};N_{;x}) \\
&= \min_{2C}\ \sum_{x2X} E[T_x]\ \tfrac{1}{2}k\ \ \ \ \ k^2_{xx^>} \\
&= \min_{2C}\ \tfrac{1}{2}k\ \ \ \ \ k^2_{(\ \ \sum_{x2X}\ E[T_x]xx^>)} \\
&\quad \min_{z2Zn\ z}\ \tfrac{1}{2}k\ \ \ \ \ {}_z(\ )k^2_{(\ \ \sum_{x2X}\ E[T_x]xx^>)}
\end{aligned}$$

where

$$_z(") = \frac{((z\ \ \ z)^> \ \ + ")(\sum_{x2X}\ E[T_x]xx^>)\ \ ^1(z\ \ \ z)^>}{(z\ \ \ z)^>(\sum_{x2X}\ E[T_x]xx^>)\ \ ^1(z\ \ \ z)}$$

for some small . This is a valid choice since for all $2\ Z\ n\ z$ we have $(z\ \ \ z)^>\ \ _z(") = \ \ " < 0$ and thus $_z(")\ 2\ C$. A straightforward calculation shows that

$$k\ \ \ \ \ {}_z(\ )k^2_{(\ \ \sum_{x2X}\ E[T_x]xx^>)} = \frac{(hz\ \ \ z;\ \ i\ +\ ")^2}{kz\ \ \ \ zk^2_{(\ \ \sum_{x2X}\ E[T_x]xx^>)}\ \ ^1}$$

so that after rearranging and lettering $g\ \ 0$ we have that any-PAC algorithm satis es

$$\max_{z2Zn\ z}\ \frac{2kz\ \ \ \ zk^2_{(\ \ \sum_{x2X}\ E[T_x]xx^>)}\ \ ^1}{hz\ \ \ z;\ \ i^2}\log(1=2.4\ ) \quad 1: \tag{8}$$

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 -PAC algorithm that sets $B(x) = 1$ for all $x\ 2\ X$ for all time until it exits at time U after this many unlabelled examples have been observed. If T_x denotes the number of times $x\ 2\ X$ was observed before stopping time U, then by Wald's identity we have that

$$E[T_x] = E\left[\sum_{t=1}^{U} 1\{x_t\ =\ x\}\right] = \ \ (x)E[U]:$$

Plugging this into Equation 8 and rearranging we conclude that

$$E[U] \quad \max_{z2Zn\ z}\ \frac{2kz\ \ \ \ zk^2_{(\ \ \sum_{x2X}\ \ (x)xx^>)}\ \ ^1}{hz\ \ \ z;\ \ i^2}\log(1=2.4\ ) =: \ \ (\ )\log(1=2.4\ )$$

which concludes the proof of the rst bullet.

A.2 Proof of Theorem 1, part II

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

$$L_x = \sum_{s=1}^{U} 1\{x_s\ =\ x;Q_s(x) = 1\}$$

and we want to show that $E[L_x] = \quad (x) E\left[P_{\ell=1}^{U} P_\ell(x) \right]$. To do so, we define

$$M_t = \sum_{s=1}^{t} \left(1\{ x_s = x; Q_s(x) = 1 \} \quad (x) P_s(x) \right)$$

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

$$E[M_{t+1} | F_t] = M_t + E[1\{ x_s = x; Q_s(x) = 1 \} \quad (x) P_s(x) | F_t] = M_t$$

Applying Doob's equality $E[M_U] = E[M_0] = 0$. Consequence:

$$E[L_x] = E\left[\sum_{s=1}^{U} 1\{ x_s = x; Q_s(x) = 1 \} \right] = \quad (v) E\left[\sum_{s=1}^{U} P_s(x) \right]$$

Define $\quad (x) := \frac{E[\sum_{s=1}^{U} P_s(x)]}{E[U]}$ and note that each $_x \in [0;1]$. Then $E[L_x] = E[U] \quad (x) \quad (x)$ so applying equation (18) of [17] again, we have

$$\begin{aligned}
\log(1=2:4) &\quad \min_{\in C} \sum_{x \in X} E[L_x] KL(N_{;x}; N_{;x}) \\
&= \min_{\in C} \sum_{x \in X} E[L_x] \| \quad \|^2_{xx^\top} =2 \\
&= \min_{z \in Z \setminus z} \frac{\langle \quad ; z \quad \rangle^2}{2\| z \quad \|^2_{\left(\sum_{x \in X} E[L_x] xx^\top \right)^{-1}}} \\
&= \min_{z \in Z \setminus z} \frac{\langle \quad ; z \quad \rangle^2}{2\| z \quad \|^2_{\left(\sum_{x \in X} \quad (x) \quad (x) xx^\top \right)^{-1}}} E[U]:
\end{aligned}$$

Rearranging, and applying the identity $E_{X \sim \quad} [\quad (X) XX^\top] = \sum_{x \in X} \quad (x) \quad (x) xx^\top$, the above implies that

$$E[U] \quad \max_{z \in Z \setminus z} \frac{2\| z \quad \|^2_{E_X \quad [\quad (X) XX^\top]^{-1}}}{\langle \quad ; z \quad \rangle^2} \log(1=2:4):$$

Noting that the total expected number of labels is equal to

$$E[L] = \sum_{x \in X} E[L_x] = \sum_{x \in X} E[U] \quad (x) \quad (x) = E[U] E_{X \sim \quad} [\quad (X)]$$

we conclude that

$$\begin{aligned}
E[L] \quad & \min_{: X \to [0;1]} \quad E[U] E_{X \sim \quad} [\quad (X)] \\
& \text{subject to} \quad E[U] \quad \max_{z \in Z \setminus \{ \quad z \}} \frac{2\| z \quad \|^2_{E_X \quad [\quad (X) XX^\top]^{-1}}}{\langle \quad ; z \quad \rangle^2} \log(1=2:4):
\end{aligned}$$

The second bullet point result follows by denoting $\quad$ as P and applying Proposition 2.

B Selective Sampling Algorithm for Known Distribution

B.1 Proof of Theorem 2, upper bound

At each round ℓ we assume an implementation such that $(\hat{h}_\ell, \hat{p}_\ell) \leftarrow \textsc{OptimizeDesign}(Z_\ell; 2^\ell; \quad)$ returns the solution of Equation 3 with $= 2^\ell$, essentially. More explicitly, let $_\ell := 2^\ell$, $B < 1$ such that $\max_{x \in X} \langle x; \quad \rangle \quad B$, and < 1 such that $E[(y_s \quad \langle \quad ; x_s \rangle)^2 | x_s] \quad ^2$. If

$$_{;\ell} := 16(B^2 + \quad ^2) \log(2\ell^2 \| Z \|^2 =)$$

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

$$P_\ell := \operatorname*{argmin}_{P : X \to [0;1]} E_{X \sim \quad} [P(X)] \text{ subject to } \max_{z; z' \in Z_\ell} \frac{\| z \quad z' \|^2_{E_X \quad [P(X) XX^\top]^{-1}}}{\langle \quad ; z \quad \rangle^2} \quad _{;\ell} \quad 1$$

and $\hat{b}_{\hat{p}_\ell} := E_X[P_\ell(X)XX^\top]$

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 $\ell \geq 2$ such that P_ℓ is feasible, that $z^* \in Z_\ell$ and $\max_{z \in Z_\ell} \langle z - z^*, \hat{b}_\ell \rangle \leq 4\epsilon_\ell$.

Proof. Define the event E as

$$E := \bigcap_{\ell=1}^{L} \bigcap_{z, z' \in Z_\ell} \left\{ |\langle z - z', \hat{b}_\ell \rangle - \langle z - z', b_\ell \rangle| \leq \epsilon_\ell \right\}$$

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

$$\begin{aligned} \langle z' - z, \hat{b}_\ell \rangle &= \langle z' - z, b_\ell \rangle + \langle z' - z, \hat{b}_\ell - b_\ell \rangle \\ &\leq \langle z' - z, b_\ell \rangle + \epsilon_\ell \end{aligned}$$

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

$$\begin{aligned} \max_{z' \in Z_\ell} \langle z' - z, \hat{b}_\ell \rangle &\geq \langle z^* - z, \hat{b}_\ell \rangle \\ &= \langle z^* - z, b_\ell \rangle + \langle z^* - z, \hat{b}_\ell - b_\ell \rangle \\ &> \epsilon_\ell + 2\epsilon_\ell \\ &= \epsilon_\ell \end{aligned}$$

which implies this z would be kicked out. Note that this implies that $\max_{z \in Z_{\ell+1}} \langle z^* - z, b_\ell \rangle \leq 2\epsilon_\ell = 4\epsilon_{\ell+1}$. □

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

$$\sum_{\ell=1}^{L} \sum_{t=(\ell-1)+1}^{\ell} Q_\ell(x_t) \leq 2 \sum_{\ell=1}^{L} E_X[P_\ell(X) | Z_\ell] + \log(1/\delta)$$

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

$$\begin{aligned} &\sum_{\ell=1}^{L} E_X[P_\ell(X) | Z_\ell] \\ &= \sum_{\ell=1}^{L} \left\{ \min_{P: X \to [0,1]} E_X[P(X)] \quad \text{subject to} \quad \max_{z, z' \in Z_\ell} \frac{\|z - z'\|^2_{E_X[P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \leq 1 \right\} \end{aligned}$$

Using Lemma 3, we have

$$\begin{aligned} \max_{z, z' \in Z_\ell} \frac{\|z - z'\|^2_{E_X[P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} &\leq L \max_{z, z' \in Z_\ell} \frac{\|z - z'\|^2_{E_X[P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \\ &\leq 64 L \max_{z \in Z \setminus z^*} \frac{\|z - z^*\|^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z - z^*, b_\ell \rangle^2} \\ &=: \max_{z \in Z \setminus z^*} \frac{\|z - z^*\|^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z - z^*, b_\ell \rangle^2} \end{aligned}$$

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

$^2) \log(2L\,^2 jZj\,^2 = \ )$ and $(\ ) = \max_{z2Zn\ z} \frac{kz\ \ z\ k^2_{E_X\ \ [XX\ >]\ 1}}{hz\ \ z\ ;\ \ i^2}$, which holds by assumption in the theorem.

Plugging this constraint back into above we have

$$\begin{aligned}
&\sum_{`=1}^{L} E_X\ \ [P_`(X)jZ_`] \\
&\sum_{`=1}^{L} \min_{P:X!\ [0;1]} E_X\ \ [P(X)] \quad \text{subject to} \quad \max_{z2Zn\ z} \frac{kz\ \ z\ k^2_{E_X\ \ [P(X)XX\ >]\ 1}}{hz\ \ z\ ;\ \ i^2} \quad 1 \\
&L \min_{24\ x} (\ ) \quad \text{subject to} \quad k = \ k_1\ \ (\ )
\end{aligned}$$

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

B.1.1 High-probability Events

Lemma 2. We have $P(E) \quad 1 \quad$.

Proof. For any $V \quad Z$ and $z; z^0 2\ V$ de ne

$$E_{z;z^0;`}(V) = \ fjhz \quad z^0; \hat{b}_`(V) \quad ij \quad `g$$

where $\hat{b}_`(V)$ is the estimator that would be constructed by the algorithm at stage $`$ with $Z_` = V$. For xed $V \quad Z$ and $` \ 2\ N$ we apply Proposition 1 so that with probability at least $1 \frac{}{`^2 jZj\,^2}$ we have that for any $z; z^0 2\ V$

$$jhz \quad z^0; \hat{b}_`(V) \quad ij \quad k\ z \quad z^0 k_{E_X\ \ [P_`(X)XX\ >]\ 1} \sqrt{\frac{}{16(B^2 + \ \ ^2) \log(2`^2 jZj\,^2 = \ )}}$$

Noting that $E := \bigcap_{`=1}^{1} \bigcap_{z;z^0 2Z_`} E_{z;z^0;`}(Z_`)$ we have

$$\begin{aligned}
P\Big(\bigcap_{`=1}^{1} \bigcap_{z;z^0 2Z_`} fE^c_{z;z^0;`}(Z_`)g\Big) &= \sum_{`=1}^{1} P\Big(\bigcap_{z;z^0 2Z_`} fE^c_{z;z^0;`}(Z_`)g\Big) \\
&= \sum_{`=1}^{1} \sum_{V\ Z} P\Big(\bigcap_{z;z^0 2V} fE^c_{z;z^0;`}(V)g; Z_` = V\Big) \\
&= \sum_{`=1}^{1} \sum_{V\ Z} P\Big(\bigcap_{z;z^0 2V} fE^c_{z;z^0;`}(V)g\Big) P(Z_` = V) \\
&\sum_{`=1}^{1} \sum_{V\ Z} \frac{}{`^2 jZj\,^2}\ \ \binom{jVj}{2} P(Z_` = V) \\
&\sum_{`=1}^{1} \sum_{V\ Z} \frac{}{2`^2} P(Z_` = V)
\end{aligned}$$

□

B.2 Technical Lemmas

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

De nition 2. Let $X_1; : : : ; X_n$ be i.i.d. random variables with mean x and variance ^{2}. Let $\ \ 2\ (0; 1)$. We say that $b(X_1; : : : ; X_n)$ is a $\ $-robust estimator if there exist universal constants $c_1; c_0 > 0$ such that if $n \quad c_1 \log(1= \ )$, then with probability at least 1

$$jb(f X_t g_{t=1}^n) \quad xj \quad c_0 \sqrt{\frac{^2 \log(1= \ )}{n}}:$$

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

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

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

If $\hat{b} = \arg\min \max_{v \in V} \frac{|\langle w_v - h, v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}}$ and $n \ge 4\log(2|V|/\delta)$, then with probability at least $1 - \delta$, it holds that

$$|\langle v, \hat{b} - h\rangle| \le \|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}} \sqrt{16(B^2 + \sigma^2)\log(2|V|/\delta)}$$

Proof. Inspired by [5], we note that

$$\begin{aligned}
\max_{v \in V} \frac{|\langle \hat{b} - h, v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} &= \max_{v \in V} \frac{|\langle \hat{b} - w_v + w_v - h, v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} \\
&\le \max_{v \in V} \frac{|\langle \hat{b} - w_v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} + \max_{v \in V} \frac{|\langle w_v - h, v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} \\
&= \min \max_{v \in V} \frac{|\langle h - v\rangle - \langle w_v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} + \max_{v \in V} \frac{|\langle w_v - h, v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}} \\
&\le 2 \max_{v \in V} \frac{|\langle h - v\rangle - \langle w_v\rangle|}{\|v\|_{\mathbb{E}_X [P(X)XX^\top]^{-1}}}
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{Var}_{x_s \sim Q(x_s) \sim P(x_s)}(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s y_s) \\
&= \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s y_s)^2] \\
&\quad - \mathbb{E}_{x_s;Q}[v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s y_s]^2
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s y_s)^2] \\
&= \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s (x_s^\top h + \epsilon_s))^2] \\
&= \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s (x_s^\top h))^2] \\
&\quad + \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s)^2 \epsilon_t^2] \\
&\quad B^2 \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s)^2] \\
&\quad + \sigma^2 \mathbb{E}_{x_s;Q}[(v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s)^2] \\
&= \mathbb{E}_{x_s}(B^2 + \sigma^2) \mathbb{E}_{Q(x_s) \sim P(x_s)}[v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s x_s^\top Q(x_s) \mathbb{E}_X [P(X)XX^\top]^{-1} v] \\
&\overset{(i)}{=} \mathbb{E}_{x_s}(B^2 + \sigma^2) \mathbb{E}_{Q(x_s) \sim P(x_s)}[v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} Q(x_s) x_s x_s^\top \mathbb{E}_X [P(X)XX^\top]^{-1} v] \\
&\quad \mathbb{E}_{x_s}(B^2 + \sigma^2) v^\top \mathbb{E}_X [P(X)XX^\top]^{-1} P(x_s) x_s x_s^\top \mathbb{E}_X [P(X)XX^\top]^{-1} v] \ ;
\end{aligned}$$

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

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

By using the property of Catoni estimator stated in De nition 2, we have $\alpha_0 = \frac{p}{2}$ and

$$
\begin{aligned}
& |\langle \; ; v\rangle - \langle w_v\rangle| \\
&= \left| \mathrm{Catoni}_s(\{\langle v; E_X[P(X)XX^\top]^{-1} Q(x_s)x_s y_s\rangle\}_{s=1}^n) - E[\langle v; E_X[P(X)XX^\top]^{-1} Q(x_s)x_s y_s\rangle] \right| \\
&\quad \sqrt{\frac{p}{2}\,(\mathrm{Var}(\langle v; E_X[P(X)XX^\top]^{-1} Q(x_s)x_s y_s\rangle))\,\frac{\log(\frac{2}{})}{n=2}} \qquad \text{(with probability at least } 1 \quad \text{if } n \quad 4\log(2=\ )) \\
&\quad \sqrt{\frac{4(B^2 + \ ^2)\log(\frac{2}{})}{n}}\, k\ vk_{(E_X[P(X)XX^\top]^{-1}} \\
&= kvk_{E_X[nP(X)XX^\top]^{-1}} \sqrt{\frac{4(B^2 + \ ^2)\log(2=\ )}{}}.
\end{aligned}
$$

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

Lemma 3. Holds

$$
\max_{z;z^0 \in Z_\ } \frac{kz \quad z^0 k^2_{E_X[P(X)XX^\top]^{-1}}}{\ ^2} \quad 64 \max_{z \in Z \setminus z} \frac{kz \quad z\ k^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z \quad z\ ; \quad \rangle^2}
$$

Proof. Let $S_\ = \{z \in Z\ : \langle z \quad z; \quad \rangle \quad 4\ \}$. We have

$$
\begin{aligned}
\max_{z;z^0 \in Z_\ } \frac{kz \quad z^0 k^2_{E_X[P(X)XX^\top]^{-1}}}{\ ^2} &\quad \max_{z;z^0 \in S_\ } \frac{kz \quad z^0 k^2_{E_X[P(X)XX^\top]^{-1}}}{\ ^2} \\
&= 16 \max_{z;z^0 \in S_\ } \frac{kz \quad z^0 k^2_{E_X[P(X)XX^\top]^{-1}}}{(4\ )^2} \\
&\quad 64 \max_{z \in S_\ } \frac{kz \quad z\ k^2_{E_X[P(X)XX^\top]^{-1}}}{(4\ )^2} \\
&= 64 \max_{z \in S_\ \setminus z} \frac{kz \quad z\ k^2_{E_X[P(X)XX^\top]^{-1}}}{\max\{(4\ )^2; \langle z \quad z\ ; \quad \rangle^2\}} \\
&\quad 64 \max_{z \in Z \setminus z} \frac{kz \quad z\ k^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z \quad z\ ; \quad \rangle^2}.
\end{aligned}
$$

□

B.2.1 Reparameterization

Proposition 2. Fix $\ \in \Delta_X$ and any $\ \in \Delta_X$. De ne $k = \ _{k_1} = \sup_{x \in X} \ (x)/\ (x)$ and $(\ ) = \max_{z \in Z \setminus z} \frac{kz \quad z\ k^2_{E_X[XX^\top]^{-1}}}{\langle z \quad z; \quad \rangle^2}$. For any $t; \quad \in \mathbb{R}_+$ the following optimization problems achieve the same value

- $\min_{P: X \to [0;1]} t\, E_X[P(X)]$ subject to $\max_{z \in Z \setminus z} \frac{kz \quad z\ k^2_{E_X[P(X)XX^\top]^{-1}}}{\langle z \quad z; \quad \rangle^2} \quad t$
- $\min_{\ \in \Delta_X} (\ )$ subject to $k = \ _{k_1} (\ ) \quad t$

Let us rst prove a simple lemma.

Lemma 4. Let $\mathcal{P}$ denote the set of all functions $P: X \to [0;1]$. And for any $\ \in \Delta_X$ with support X let $\mathcal{P}^0 = \{ \ _x = \ _x : \ \in \Delta_X; \quad 0 : \ _x = \ _x \in [0;1]\}$. Then $\mathcal{P} = \mathcal{P}^0$.

Proof. Fix any $P \in \mathcal{P}$. If $\ _x = P_x\ _x = kP \quad k_1$ and $\ = kP \quad k_1$ then $\ = \ \in \mathcal{P}^0$ and is equal to P. This implies $\mathcal{P} \quad \mathcal{P}^0$.

For the other direction, x any $\ \in \Delta_X$ and $\quad 0$ such that $\ _x = \ _x \in [0;1]$ for all x. If $P = \ = \ $ then $P \in \mathcal{P}$ which implies $\mathcal{P}^0 \quad \mathcal{P}$ and concludes the proof. □

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

$$\min_{P:X\to[0;1]} t\,E_X\quad[P(X)]\quad\text{subject to}\quad \max_{z\in\mathcal{Z}}\frac{\|z\quad z\|_{E_X\quad[P(X)XX^{>}]^{1}}^{2}}{\langle z\quad z;\quad\rangle^{2}}\quad t$$

is equivalent to

$$\min_{0;\ 24_X} t\,E_X\quad[\quad(X)=\ (X)]\quad\text{subject to}\quad \max_{z\in\mathcal{Z}}\frac{\|z\quad z\|_{E_X\quad[\quad(X)=\ (X)XX^{>}]^{1}}^{2}}{\langle z\quad z;\quad\rangle^{2}}\quad t$$
$$(x)=\ (x)\quad 1\quad 8x\ 2\ X$$

which is equal to, after simplifying,

$$\min_{0;\ 24_X} t\qquad\text{subject to}\quad \max_{z\in\mathcal{Z}}\frac{\|z\quad z\|_{E_X\quad[XX^{>}]^{1}}^{2}}{\langle z\quad z;\quad\rangle^{2}}\quad t$$
$$(x)=\ (x)\quad 1\quad 8x\ 2\ X$$

which is equal to

$$\min_{u\quad 0;\ 24_X} u\quad\text{subject to}\quad (\ )\quad u$$
$$k=\ k_1\quad\frac{t}{u}:$$

Note, there exists a feasible $(;u\ )$ precisely when there exists a $2\ 4_X$ such that $k=\ k_1\quad(\ )\quad t$, in which case the optimization problem is equal to

$$\min_{24_X}\ (\ )\qquad\text{subject to}\quad k=\ k_1\quad(\ )\quad t$$

□

C Analysis of the Optimization Problem

C.1 Proof of Theorem 4

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

The proof relies on analyzing another function $\bar{D}: R^d\ _0^d\ 7!\ R$. For simplicity, rst, we de ne

$$h\ (x)=P\ (x)\qquad(\log(1\quad P\ (x))+\log(P\ (x)))\quad P\ (x)x^{>}\quad x: \tag{9}$$

Recall that our dual objective $D(\ )=E_X\quad[h\ (X)]+\frac{1}{c^2}\sum_{y2Y\cdot}y^{>}\ _y y$. Since the rst term in $E_X\quad[h\ (X)]$ only depends on $=\sum_{y2Y\cdot}\ _y$, we can consider the following optimization problem.

$$\begin{aligned} f(\ )=\max_{\ _y}\ & \sum_{y2Y\cdot}y^{>}\ _y y\\ \text{subject to}\ & \sum_{y2Y\cdot}\ _y=\\ & _y\quad 0;\quad 8y\ 2\ Y\cdot: \end{aligned} \tag{10}$$

Then, the alternative dual objective $\bar{D}(\ )$ is de ned as $\bar{D}(\ )=\ E_X\quad[h\ (X)]+\frac{1}{c^2}f(\ )$. We can immediately see that maximizing $\bar{D}(\ )$ is equivalent to maximizing $D(\ )$. In particular, let $\quad 2\ \operatorname{argmax}_{\ 0}\bar{D}(\ )$ and $_y\ _{y2Y\cdot}$ be the set of PSD matrices that solve problem (10) and evaluate $f(\quad)$. We can see that $_y\ _{y2Y\cdot}$ also maximizes $D(\ )$. Conversely, for $=\ _y\ _{y2Y\cdot}\ 2\ \operatorname{argmax}_{\ _y\ 0;8y}D(\quad)$, we also have $\sum_{y2Y\cdot}\ _y\ 2\ \operatorname{argmax}_{\ 0}\bar{D}(\ )$.

Further, we also de ne their empirical version D_E and $\bar{D}_E$ with extra i.i.d. samples $x_1;\cdots;x_u$ as

$$D_E(\quad)=\frac{1}{u}\sum_{i=1}^{u}h\ (x_i)+\frac{1}{c^2}\sum_{y2Y\cdot}y^{>}\ _y y\quad\text{and}\quad\bar{D}_E(\ )=\frac{1}{u}\sum_{i=1}^{u}h\ (x_i)+\frac{1}{c^2}f(\ ): \tag{11}$$

Recall that the problem Algorithm 2 tries to solve is

$$\begin{aligned} \min_P \quad & \mathbb{E}_X \quad [P(X) \quad (\log(1 \quad P(X)) + \log(\ P(X)))] \\ \text{subject to} \quad & \mathbb{E}_X \quad P(X)XX^{>} \quad \frac{1}{c^2} yy^{>}; \quad \forall y \in Y^{\backslash}: \end{aligned} \tag{12}$$

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

Theorem 5. Suppose $\|x\|_2 \quad M$ for any $x \in \text{supp}(\ )$ and $= \mathbb{E}_X \quad XX^{>}$ is invertible. Let $\in \text{argmax}_{y \quad 0; \forall y} D(\ )$ and $(\ ) = \frac{\max(\ )}{\min(\ )}$ be condition number. Assume $\|\|_F > 0$ and define $! = \min_{\in S^d : \|\|_F = 1} \mathbb{E}_X \quad X^{>} \quad X^{\ 2}$, where S^d is the set of $d \quad d$ symmetric matrices. Let $\|Y^{\backslash}\|_{C^2} = \frac{1}{c^2} \sum_{y \in Y^{\backslash}} \|y\|_2^4$.

Then, $= \sum_{y \in Y^{\backslash}} \ _y$ is unique. Further, for any > 0 and > 0, suppose it holds that

$$\min\left(\frac{\sqrt{\frac{3(\ )\|\|\|\|_F M^2}{8} \quad \frac{1+}{}}}{}; \frac{4}{9}\|\|\|\|_F^2 M^4; \frac{1}{2\sqrt{3}} \right)$$

$$K \quad \frac{288(\ )^2 \|Y^{\backslash}\|^3 \|\|\|\|_F^4 M^4 (M^4 + C^2) \quad 2\|\|\|\|_F M^2 + 1\ ^4 \log(6/\ )}{!^2 \ 6} \quad \frac{1+}{}^2$$

$$u \quad \frac{576(\ )^2 \|\|\|\|_F^2 M^8 \quad 2\|\|\|\|_F M^2 + 1\ ^4 \log(6/\ )}{!^2 \ 6} \quad \frac{1+}{}^2 :$$

Then, with probability at least $1 \quad$, Algorithm 2 will output ^{e} that satisfies

- $y^{>} \mathbb{E}_X \quad P_e(X)XX^{>} \quad {}^1 y \quad (1+ \ )c^2; \quad \forall y \in Y^{\backslash}$.
- $\mathbb{E}_X \quad P_e(X) \quad \mathbb{E}_X \quad P^{\ e}(X) \quad + 4\sqrt{\ }$, where P^e is the optimal solution to problem (20).

Proof. First Bullet Point. Fix some > 0. Let $\hat{\ }$ and corresponding $\hat{\ } = \sum_{y \in Y^{\backslash}} \hat{\ }_y$ 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 $\hat{\ }_y$ to each $y \in Y^{\backslash}$ has been optimized by solving problem (10). That is, we have $D(\hat{\ }) = \overline{D}(\hat{\ })$ and $D_E(\hat{\ }) = \overline{D}_E(\hat{\ })$. Let ^{e} and ^{e} be the ones after the re-scaling step. Then, by Theorem 3.13 of [21], with probability at least $1 \quad \frac{}{3}$, it holds that

$$\overline{D}(\ ) \quad \overline{D}(\hat{\ }) = D(\ ) \quad D(\hat{\ }) \quad \frac{\text{Reg}(K) + 2\sqrt{2K \log(6/\ )}}{K};$$

where $\text{Reg}(K)$ is the regret of running projected stochastic gradient ascent for K steps with $\sqrt{k}$ specified in Algorithm 2. Meanwhile, by Theorem 4.14 of [21] also, we have $\text{Reg}(K) = \sqrt{2B^2 \sum_{k=1}^{K} \sum_{y \in Y^{\backslash}} \|g_{k;y}\|_2^2}$, where $B = \sqrt{|Y^{\backslash}|} \|\|_F$ bound the norm of $= \ _y \ _{y \in Y^{\backslash}}$. Since $g_{k;y} = \frac{yy^{>}}{c^2} \quad P_{\hat{\ }(k)}(x_k)x_k x_k^{>}$, we can easily get $\sum_{y \in Y^{\backslash}} \|g_{k;y}\|_2^2 \quad 2|Y^{\backslash}| M^4 + \frac{2}{c^2} \sum_{y \in Y^{\backslash}} \|y\|_2^4 = 2|Y^{\backslash}| M^4 + 2|Y^{\backslash}| C^2$. Thus, we have

$$\text{Reg}(K) \quad 2|Y^{\backslash}| \|\|_F^2 \sqrt{|Y^{\backslash}| M^4 + |Y^{\backslash}| C^2} \quad \sqrt{K} := C_{\text{Reg}} \sqrt{K} \tag{13}$$

$$\Rightarrow \quad \overline{D}(\ ) \quad \overline{D}(\hat{\ }) \quad \frac{C_{\text{Reg}} + 2\sqrt{2\log(6/\ )}}{\sqrt{K}}; \tag{14}$$

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

By Lemma 5, we know that is unique and as long as $\frac{1}{2}+\frac{1}{3}$, $\overline{D}(\ )$ is G-strongly concave with respect to ℓ_2 norm over $S = \{\ \ 0 : k\ \ k_F\ \ 2k\ \ k_F\}$, where G is de ned in Eq.(21). Thus, by Lemma 11, if K is large enough such that

$$\overline{D}(\ \ ) \quad \overline{D}(\hat{\ }) \quad \frac{C_{Reg} + 2\sqrt{2\log(6=\ )}}{\sqrt{K}} \quad \frac{G\,k\ \ k_F}{2};$$

then $\hat{\ } \quad k\ \ k_F$, which implies $\hat{\ } \quad 2k\ \ k_F$. That is, $\hat{\ } \ 2\ S$. Then, under this condition, by using Lemma 8, when $\frac{4}{9}k\ \ k_F\ M^4$ and

$$u \quad \left(\frac{6\ (\ )\ k\ \ k_F\ M^4\ \ 2 + \sqrt{2\log(6=\ )}}{G^2}\ \ \frac{1+}{}\right)^{1/2}; \tag{15}$$

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

$$\overline{D}_E(\hat{\ }) \quad \overline{D}(\hat{\ }) \quad \frac{G^2}{3M^2\ (\ )}\ \ \frac{}{1+} \quad \text{and} \quad \overline{D}_E(e) \quad \overline{D}(e) \quad \frac{G^2}{3M^2\ (\ )}\ \ \frac{}{1+} \tag{16}$$

$$\begin{aligned}
\Rightarrow \quad \overline{D}(\ \ ) \quad \overline{D}(e) \quad & \overline{D}(\ \ ) \quad \overline{D}(\hat{\ }) + \overline{D}(\hat{\ }) \quad \overline{D}(e) \\
& \overline{D}(\ \ ) \quad \overline{D}(\hat{\ }) + \overline{D}(\hat{\ }) \quad \overline{D}_E(\hat{\ }) + \overline{D}_E(e) \quad \overline{D}(e) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Since } \overline{D}_E(\hat{\ }) \quad \overline{D}_E(e)\ ) \\
& \frac{C_{Reg} + 2\sqrt{2\log(6=\ )}}{\sqrt{K}} + \frac{2G^2}{3M^2\ (\ )}\ \ \frac{}{1+}: \quad (\text{By Eq. (14) and (16)})
\end{aligned}$$

Since e is a smaller re-scaling of $\hat{\ }$, we have $e\ \ 2\ S$, which implies $\frac{G}{2}\ \ e\ \ _F \quad \overline{D}(\ \ ) \quad \overline{D}(e)$ by property of strongly concave function [3]. Therefore, by Lemma 12, to guarantee an at most multiplicative constraint violation, it is suf cient to choose K such that

$$\begin{aligned}
\frac{G}{2}\ \ e\ \ _F \quad & \overline{D}(\ \ ) \quad \overline{D}(e) \\
& \frac{C_{Reg} + 2\sqrt{2\log(6=\ )}}{\sqrt{K}} + \frac{2G^2}{3M^2\ (\ )}\ \ \frac{}{1+} \\
& \min\ \ \frac{4G^2}{3M^2\ (\ )}\ \ \frac{}{1+};\ \frac{G\,k\ \ k_F}{2} \\
= \ & \frac{4G^2}{3M^2\ (\ )}\ \ \frac{}{1+}: \qquad \left(\text{If } \sqrt{\frac{3\ (\ )\ k\ \ k_F\ M^2}{8}\ \ \frac{1+}{}}\right)
\end{aligned}$$

An algebraic rearrangement gives us

$$K \quad \left(\frac{3\ (\ )\ M^2\ \ C_{Reg} + 2\sqrt{2\log(6=\ )}}{2G^2}\ \ \frac{1+}{}\right)^{1/2}: \tag{17}$$

Second Bullet Point. We then prove the upper bound for primal objective value $\mathbb{E}_X\ \ P_e(X)$, which explains the reason why an extra re-scaling step is needed. Denote $g(s) = D_E(s\ \ e)$. By construction, we know that $g(s)$ is maximized at $s = 1$ because $e = s\ \ \hat{\ }$, where $s = \operatorname{argmax}_{s\,2\,[0;1]} D_E(s\ \ \hat{\ })$. Therefore, we have $g^0(1) \quad 0$, which in turn gives us

$$g^0(1) = \frac{1}{c^2}\sum_{y\,2\,Y\ \cdot} y^{>}\ e_y y \quad \frac{1}{u}\sum_{i=1}^{u} P_e(x_i)x_i^{>}\ e x_i \quad 0:$$

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

$$u \quad \left(\frac{2k\ \ k_F\ M^2\ \ k\ \ k_F\ M^2 + \sqrt{2\log(6=\ )}}{3=2}\right)^{1/2}; \tag{18}$$

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

$$E_X \quad P\ (X)X^{>} \quad X \quad \frac{1}{u}\sum_{i=1}^{u} P\ (x_i)x_i^{>} \quad x_i \quad \frac{p}{}$$

$$=) \quad \frac{1}{c^2}\sum_{y2Y^{`}} y^{>} e_y y \quad E_X \quad \left[P_e(X)X^{>} e_X\right] + \frac{p}{} \quad 0: \tag{19}$$

Now, let P^e be the optimal solution of problem (20) and P^{Δ} be the optimal solution of the same problem with bound constraint $P(x) \quad 1$.

$$\begin{aligned} &\min_P \quad E_X \quad [P(X)] \\ &\text{subject to} \quad y^{>} E_X \quad P(X)XX^{>} \quad {}^{1} y \quad c^2; \quad 8y 2 Y^{`}; \\ &\qquad 0 \quad P(x) \quad 1 \quad ; \quad 8x 2 X: \end{aligned} \tag{20}$$

Then, we can notice that

$$\begin{aligned}
&E_X \quad P_e(X) \\
&E_X \quad P_e(X) \quad (\log(1 \quad P_e(X)) + \log(P_e(X))) \\
&E_X \quad P_e(X) \quad (\log(1 \quad P_e(X)) + \log(P_e(X))) \\
&\quad + \frac{1}{c^2}\sum_{y2Y^{`}} y^{>} e_y y \quad E_X \quad \left[P_e(X)X^{>} e_X\right] + \frac{p}{} \qquad \text{(By Eq. (19))} \\
&= \inf_P \ L(P; {}^{e}) + \frac{p}{} \qquad \text{(By de nition of Lagrangian function and how we solve for } {}^{e}) \\
&\quad \max_{y \quad 0;8y2Y^{`}} \ \inf_P L(P; \quad) + \frac{p}{} \\
&= E_X \quad [P \quad (X) \quad (\log(1 \quad P \quad (X)) + \log(P \quad (X)))] + \frac{p}{} \\
&E_X \quad \left[P^{\Delta}(X) \quad \log(1 \quad P^{\Delta}(X)) \quad \log P^{\Delta}(X)\right] + \frac{p}{} \\
&\qquad \text{(Since } P^{\Delta} \text{ is feasible to problem (12))} \\
&E_X \quad \left[P^{\Delta}(X)\right] + 3\frac{p}{}; \qquad \text{(Since } a\log(a) \quad \frac{p}{} a \text{ for } a\ 2\ (0;1)) \\
&E_X \quad \left[P^e(X)\right] + 4\frac{p}{}: \qquad \text{(Since } P^{\Delta}(x) \text{ can have at most more contribution than } P^e)
\end{aligned}$$

Therefore, in summary, Suppose K and u satisfy conditions speci ed in Eq. (17), (15) and (18) and $\quad \min \left\{ \frac{\sqrt{3\ (\ )\ k \quad k_F M^2 \quad 1+}}{8}; \frac{4}{9} k \quad k_F^2 M^4; \frac{1}{2^{p}\frac{}{3}} \right\}$, where C_{Reg} and G are de ned in Eq. (13) and (21), respectively. Then. by applying a simple union bound, with probability at least $1 \quad$, the output of Algorithm 2 ^{e} satis esy$^{>} E_X \quad P(X)XX^{>} \quad {}^{1} y \quad (1+ \quad)c^2; 8y 2 Y^{`}$ and $E_X \quad P_e(X) \quad E_X \quad \left[P^e(X)\right] + 4\frac{p}{}$. □

C.2 Relevant Lemmas

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

Lemma 5. As long as $\frac{1}{2^{p}\frac{}{3}}$, $\overline{D}(\ )$ is G-strongly concave with respect to ℓ_2-norm on the bounded region $S = \{ \quad 0 : k \quad k_F \quad 2k \quad k_F \}$ with coef cient

$$G = \frac{}{2(2k \quad k_F M^2 + 1)^2} \min_{2 S^d : k \quad k_F = 1} E_X \quad \left[X^{>} \quad X \quad {}^{2}\right]: \tag{21}$$

Because of this, as a corollary, will be unique.

Proof. By Lemma 6, since $f(\ )$ is concave in , it is suf cient to prove that $E_X \quad [h \quad (X)]$ is G-strongly concave on S, where $h \quad (x)$ is de ned in Eq. (9). Then, we have

$$r \quad {}^{2} E_X \quad [h \quad (X)] = E_X \quad \frac{dP}{dq}(X)\text{vec}\ XX^{>} \quad \text{vec}\ XX^{>}\ {}^{>}:$$

Since $\|x\|_2 \quad M$, for any $\quad 2 S$, we have $q(x) = x^{>} \quad x \quad 1 \quad 2k \quad k_F M^2 + 1$. By Lemma, 14, we know that if $12^2 \quad 2k \quad k_F M^2 + 1^2$, which can be done by choosing $\frac{1}{2^{\sqrt{3}}}$, we have $\frac{dP}{dq}(x) \quad \frac{}{2(2k \quad k_F M^2+1)^2}$ for any $x 2 X$ and $\quad 2 S$. Therefore, we have

$$r^2 E_X [h(X)] \quad E_X \left[\mathrm{vec}\, XX^{>} \quad \mathrm{vec}\, XX^{>} \right]^{>}$$

Now, let S be the set of all $d \quad d$ symmetric matrices. It is obvious that S is a subspace of the vector space of all $d \quad d$ matrices and $S \quad S$. Thus, by applying Lemma 7, we can conclude that $E_X [h(X)]$ is G-strongly concave on S with respect to ℓ_2 norm and

$$G = \frac{}{2(2k \quad k_F M^2 + 1)^2} \min_{2 S^d : \|\| k_F = 1} \mathrm{vec}(\,)^{>} E_X \left[\mathrm{vec}\, XX^{>} \quad \mathrm{vec}\, XX^{>} \right]^{>} \mathrm{vec}(\,)$$

$$= \frac{}{2(2k \quad k_F M^2 + 1)^2} \min_{2 S^d : \|\| k_F = 1} E_X \left[X^{>} \quad X \right]^2 :$$

Thus the proof is complete. □

Lemma 6. $f(\,)$ de ned in Eq.(10) is concave in $\,$.

Proof. To show its concavity, consider $^{(1)} \quad 0$, $^{(2)} \quad 0$ and some $\ 2 (0; 1)$. Let $(\ ^{(i)}_y)_{y 2 Y^{\cdot}}$ be the optimal solution obtained by evaluating $f(\ ^{(i)})$ for $i 2 \{1; 2\}$. Then, we can notice that

$$f(\ ^{(1)}) + (1 \quad)f(\ ^{(2)}) = \sum_{y 2 Y^{\cdot}} y^{>} \ ^{(1)}_y y + (1 \quad) \sum_{y 2 Y^{\cdot}} y^{>} \ ^{(2)}_y y$$

$$= \sum_{y 2 Y^{\cdot}} y^{>} (\ ^{(1)}_y + (1 \quad) \ ^{(2)}_y) y$$

$$f(\ ^{(1)} + (1 \quad) \ ^{(2)}):$$

The last inequality above holds because $\sum_{y 2 Y^{\cdot}} \ ^{(i)}_y = \ ^{(i)}$ for $i \quad 2 \{1; 2\}$ and thus $\sum_{y 2 Y^{\cdot}} \ ^{(1)}_y + (1 \quad) \ ^{(2)}_y = \ ^{(1)} + (1 \quad) \ ^{(2)}$, which means that $(\ ^{(1)}_y + (1 \quad) \ ^{(2)}_y)_{y 2 Y^{\cdot}}$ is a feasible solution for problem (10) with parameter $\ ^{(1)} + (1 \quad) \ ^{(2)}$. Therefore, we can conclude that $f(\,)$ is concave in $\,$. □

Lemma 7. Let $f : R^d \ 7! \ R$ be a convex and twice differentiable function. If for some subspace $S \quad R^d$, we have $\min_{w 2 S : \|w\|_2 = 1} w^{>} r^2 f(x) w \quad > \ 0, 8x 2 S$, then f is $\,$-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 $x 2 S$, there exists unique $z 2 R^m$ such that $x = V z$, where $V = [v_1 \quad \cdots \quad v_m]$. That is, there is one-to-one correspondence between S and R^m.

Now, we de ne $g : R^m \ 7! \ R$ as $g(z) = f(V z)$. It is easy to compute $r^2 g(z) = V^{>} r^2 f(V z) V$. Then, notice that for any $w^0 2 R^m$ such that $\|w^0\|_2 = 1$, we have $V w^0 2 S$ and $\|V w^0\|_2 = \sqrt{w^{0>} V^{>} V w^0} = \sqrt{w^{0>} w^0} = 1$. Thus, we have

$$\min_{w^0 2 R^m : \|w^0\|_2 = 1} w^{0>} r^2 g(z) w^0 = \min_{w^0 2 R^m : \|w^0\|_2 = 1} w^{0>} V^{>} r^2 f(V z) V w^0$$

$$= \min_{w 2 S : \|w\|_2 = 1} w^{>} r^2 f(V z) w \quad :$$

Therefore, g is $\,$-strongly convex with respect to ℓ_2 norm. Then, for any $x_1; x_2 2 S$, there exists unique $z_1; z_2 2 R^m$ such that $x_1 = V z_1$ and $x_2 = V z_2$. Notice that $\|z_1 \quad z_2\|_2 = \|x_1 \quad x_2\|_2$ since V preserves the norm. Further, by de nition of strong convexity, for any $2 [0; 1]$, we have

$$g(\ z_1 + (1 \quad)z_2) + \frac{}{2} (1 \quad) \|z_1 \quad z_2\|_2^2 \quad g(z_1) + (1 \quad)g(z_2)$$

$$\Rightarrow \quad f(\ V z_1 + (1 \quad)V z_2) + \frac{}{2} (1 \quad) \|x_1 \quad x_2\|_2^2 \quad f(V z_1) + (1 \quad)f(V z_2)$$

$$\Rightarrow \quad f(\ x_1 + (1 \quad)x_2) + \frac{}{2} (1 \quad) \|x_1 \quad x_2\|_2^2 \quad f(x_1) + (1 \quad)f(x_2):$$

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

C.2.2 Concentration Inequalities

Lemma 8. Let $x_1; \dots; x_u$ be i.i.d. samples. If $\hat{\ }_F \quad 2k \quad k_F, kxk_2 \quad M$ for any $x \ 2 \ X$ and $\frac{4}{9}k \quad k_F^2 \ M^4$, then with probability at least $1 \quad \frac{2}{3}$, it holds for any $\ 2 \ = \ s \ \hat{\ }: \ s \ 2 \ [0; 1]$ simultaneously that

$$E_X \quad [h \ (X)] \quad \frac{1}{u} \sum_{i=1}^{u} h \ (x_i) \quad \frac{2k \quad k_F \ M^2 \ 2 + \sqrt{2 \log(6= \ )}}{\sqrt{u}}$$

$$E_X \quad P \ (X)X^> \ X \quad \frac{1}{u} \sum_{i=1}^{u} P \ (x_i)x_i^> \ x_i \quad \frac{2k \quad k_F \ M^2 \ k \quad k_F \ M^2 + \sqrt{2 \log(6= \ )}}{\sqrt{u}}:$$

Proof. To prove the first inequality, first, notice that we have $h \ (x) \ = \ P \ (x)q \ (x) \ (\log(1 \quad P \ (x)) + \log(P \ (x)))$, where $q \ (x) = x^> \ x \quad 1$. Since $P \ (x)$, defined in Eq. (7), explicitly only depends on $q \ (x)$ instead of x directly, we can treat h as a function of q and define a function class $F \ = \ \{x \ 7! \ x^> (s \ \hat{\ }) \ x : s \ 2 \ [0; 1]\}$. It is well-known that if h is L_1-Lipschitz in q and $|h \ (x)| \quad R_1$ for any $\ 2$ and x, then, with probability at least $1 \quad \frac{}{3}$, it holds simultaneously for all $\ 2$ that [2, 19]

$$E_X \quad [h \ (X)] \quad \frac{1}{u} \sum_{i=1}^{u} h \ (x_i) \quad 2L_1 \ R_u(F) + R_1 \sqrt{\frac{2 \log(6= \ )}{u}}; \tag{22}$$

where $R_u(F)$ is the Rademacher complexity of F.

To find L_1, we can compute

$$\begin{aligned} \frac{dh}{dq} &= \frac{dP}{dq} q \quad P + \frac{dP}{dq} \frac{}{1 \quad P} \quad \frac{}{P} \\ &= \frac{dP}{d \ q} q \quad P + \frac{dP}{dq} \ q \qquad \text{(Since } P \text{ satisfies Eq. (6))} \\ &= P \end{aligned}$$

Therefore, we have $\frac{dh}{dq} \ 2 \quad 1; \ \frac{}{3}$ by Lemma 14. Therefore, we can set $L_1 = 1$.

Let h_0 be the value of h when $q \ = \ 1$, which means $x^> \ x = 0$. To find R_1, notice that since $\frac{dh}{dq} \ 2 \quad 1; \ \frac{}{3}$, we must have $q \ + h_0 \quad h \quad \frac{}{3}q \ + h_0$. By Lemma 14, we know that $h_0 \ 2 \ 0; 2\sqrt{\ }$. Therefore, we have $x^> \ x \quad h \ (x) \quad \frac{}{3}x^> \ x + 3\sqrt{\ }$ for any $x \ 2 \ X$ and $\ 2$. Since $k \ k_F \ \hat{\ }_F \quad 2k \quad k_F$, we have $|h \ (x)| \quad 2k \quad k_F \ M^2 := R_1$, which holds when $\frac{4}{9}k \quad k_F^2 \ M^4$. Then, by Lemma 9, we know that $R_u(F) \quad \frac{2k \quad k_F \ M^2}{\sqrt{u}}$. Thus, plugging in values of L_1, R_1 and $R_u(F)$ 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 $\tilde{h} \ (x) = P \ (x)x^> \ x = P \ (x)q \ (x) + P \ (x)$. Then, with probability at least $1 \quad \frac{}{3}$, it holds simultaneously for any $\ 2$ that

$$E_X \left[\tilde{h} \ (X) \right] \quad \frac{1}{u} \sum_{i=1}^{u} \tilde{h} \ (x_i) \quad 2L_2 \ R_u(F) + R_2 \sqrt{\frac{2 \log(6= \ )}{u}}; \tag{23}$$

where $\tilde{h} \ (x) \quad R_2$ for any $x \ 2 \ X$, $\ 2$ and $\tilde{h}$ is L_2-Lipschitz in q.

To find L_2, we can compute

$$\frac{d\tilde{h}}{dq} = P \ + \frac{dP}{dq} \ x^> \ x:$$

By Lemma 14, we know that $\frac{dP}{dq} \in \left[0, \frac{1}{8}\right]$. Thus, we have $\frac{d\bar{n}}{dq} \leq 1 + \frac{k \, k_F M^2}{4} := L_2$. It is obvious that $\bar{n}(x) \leq 2k \, k_F M^2 := R_2$. Thus, by plugging the values of L_2, R_2 and $R_u(F)$ into Eq. (23), we can obtain the second concentration inequality.

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

Lemma 9. If $\hat{\ } \leq 2k \, k_F$, then, we have $R_u(F) \leq \sqrt{\frac{E_X[(X^\top \hat{\ } X)^2]}{u}} \leq \frac{2k \, k_F M^2}{\sqrt{u}}$, where $F = \left\{ x \mapsto \langle \hat{\ }, x \rangle^2 : s \in [0,1] \right\}$.

Proof. Let $_1, \ldots, _u$ be i.i.d. Rademacher random variables, which are uniform over $\{-1, +1\}$. Let $x_1, \ldots, x_u$ be i.i.d. samples. Then, by de nition of Rademacher complexity, we have

$$
\begin{aligned}
R_u(F) &= E\left[\sup_{q \in F} \frac{1}{u} \sum_{i=1}^{u} {}_i q(x_i)\right] \\
&= E\left[\sup_{s \in [0,1]} \frac{1}{u} \sum_{i=1}^{u} {}_i \langle x_i^\top, (s\hat{\ }) \rangle x_i\right] && \text{(By de nition of } F\text{)} \\
&\overset{(i)}{=} \frac{1}{u} E\left[\left(1 \sum_{i=1}^{n} {}_i \langle x_i^\top, \hat{\ } \rangle x_i \quad 0 \sum_{i=1}^{n} {}_i \langle x_i^\top, \hat{\ } \rangle x_i\right)_+\right] \\
&\leq \frac{1}{u} E\left[\sum_{i=1}^{u} {}_i \langle x_i^\top, \hat{\ } \rangle x_i\right] \\
&\leq \frac{1}{u} \sqrt{E\left[\left(\sum_{i=1}^{u} {}_i \langle x_i^\top, \hat{\ } \rangle x_i\right)^2\right]} && \text{(By Jensen's inequality)} \\
&= \frac{1}{u} \sqrt{E\left[\sum_{i=1}^{u} \langle x_i^\top, \hat{\ } \rangle x_i\right]^2} && \text{(Since } {}_i\text{'s are i.i.d. and } E[{}_i] = 0\text{)} \\
&= \sqrt{\frac{E_X[\langle X^\top, \hat{\ } \rangle^2]}{u}} \leq \frac{2k \, k_F M^2}{\sqrt{u}}.
\end{aligned}
$$

Here, the equality (i) holds because when $\sum_{i=1}^{n} {}_i \langle x_i^\top, \hat{\ } \rangle x_i < 0$, 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 $f(\cdot)$ is linear in scalar multiplication.

Lemma 10. If $D_E(\hat{\ }) = \bar{D}_E(\hat{\ })$, with $\hat{\ } = \sum_{y \in Y^\ast} \hat{\ }_y$, then, for any $s \geq 0$, it holds that $D_E(s\hat{\ }) = \bar{D}_E(s\hat{\ })$, where D_E and $\bar{D}_E$ are de ned in Eq. (11).

Proof. It suf ces to show that if $\sum_{y \in Y^\ast} y^\top \hat{\ }_y y = f(\hat{\ })$, then $\sum_{y \in Y^\ast} y^\top (s\hat{\ }_y) y = f(s\hat{\ })$ for any $s > 0$. By de nition, we have

$$
\begin{aligned}
f(s\hat{\ }) = \max_{\ } \quad & \sum_{y \in Y^\ast} y^\top \ {}_y y \\
\text{subject to} \quad & \sum_{y \in Y^\ast} {}_y = s\hat{\ } \\
& {}_y \geq 0, \quad \forall y \in Y^\ast.
\end{aligned}
$$

For the above optimization problem, we can do a change of variable by setting $\frac{1}{s}\,{}^{0}_{y} = {}_{y}$ $\Rightarrow$ ${}_{y} = s\,{}^{0}_{y}$. Then, we have

$$f(s\,\hat{\ }) = \max_{y} \sum_{y \in Y} y^{>}(s\,{}^{0}_{y})y \quad \text{subject to} \quad \sum_{y \in Y} s\,{}^{0}_{y} = s\,\hat{\ }, \quad s\,{}^{0}_{y} \geq 0; \quad \forall y \in Y:$$

$$\Rightarrow \quad f(s\,\hat{\ }) = \max_{y} s \sum_{y \in Y} y^{>}\,{}^{0}_{y}y \quad \text{subject to} \quad \sum_{y \in Y} {}^{0}_{y} = \hat{\ }, \quad {}^{0}_{y} \geq 0; \quad \forall y \in Y:$$

$$\Rightarrow \quad f(s\,\hat{\ }) = s\,f(\hat{\ }) = s \sum_{y \in Y} y^{>}\,{}_{y}y = \sum_{y \in Y} y^{>}(s\,\hat{\ }_{y})y:$$

Thus, the proof is complete. □

Lemma 11. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a concave function with maximizer x^0 over the convex set C. Further, assume that f is G-strongly concave with respect to ℓ_2 norm in region $S \setminus C$, where $S = \{x : \|x - x^0\|_2 \leq A\}$. If $f(x^0) - f(x) \geq \frac{AG}{2}$ and $c \in C$, then $x \in S$.

Proof. By property of strong concavity, we know that $f(x^0) - f(x) \geq \frac{G}{2}\|x - x^0\|_2$ for any $x \in S \setminus C$. Now, suppose x^0 satisfies $f(x^0) - f(x) \geq \frac{AG}{2}$, $x^0 \in C$ and $x^0 \notin S$. Then, we must have $\|x^0 - x^0\|_2 > A$.

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

$$\frac{AG}{2} = \frac{G}{2}\|z - x^0\|_2 \leq f(x^0) - f(z) < f(x^0) - f(x) \leq \frac{AG}{2}:$$

This is a contradiction and thus we must have $x^0 \in S$. □

The following lemma quantitatively describes how close θ and θ^* needs to be to ensure an at most multiplicative constraint violation.

Lemma 12. Assume $\|x\|_2 \leq M$ for any $x \in X$. Let $= E_X[XX^{>}] \succ 0$ and $= \operatorname{argmax}_{\theta \in \Theta} \bar{D}(\theta)$. Then, for any > 0, if we have

$$e \quad {}_{F} \quad \frac{8\,{}^{2}\,{}_{\min}(\,)}{3M^2\,{}_{\max}(\,)} \quad \frac{}{1 + };$$

then it holds that $y^{>} E_X[P_e(X)XX^{>}]^{-1} y \leq (1 +)c^2$ for any $y \in Y$.

Proof. Fix some > 0. First, notice that if we regard P as a function of $q(x) = x^{>} x - 1$, it then holds that

$$\|r_{\theta} P(x)\|_2 = \left\|\frac{dP}{dq} r_{\theta} q(x)\right\|_2 \leq \left\|\frac{dP}{dq}\right\| \|xx^{>}\|_2 \leq \left\|\frac{dP}{dq}\right\| M^2 \leq \frac{M^2}{8};$$

where we obtain the last inequality by using Lemma 14. Therefore, for any $x \in X$ and $e \geq 0$, we have $P_e(x) \leq P(x) + \frac{M^2}{8} e \quad {}_{F}$ by mean value theorem and Cauchy-Schwartz. inequality.

Therefore, if we have $e \quad {}_{F}$, then

$$P_e(x) \leq P(x) + \frac{M^2}{8} \Rightarrow P_e(x) \leq P(x) + \frac{M^2}{8}$$

$$\Rightarrow \quad E_X \quad P_e(X)XX^{>} \quad E_X \quad P \quad (X)XX^{>} \quad \frac{M^2}{8} E_X \quad XX^{>} \; :$$

By Lemma 13, we know that

$$y^{>} E_X \quad P_e(X)XX^{>} \quad {}^{1} y \quad c^2(1+\ ) \; \Leftrightarrow \; E_X \quad P_e(X)XX^{>} \quad \frac{yy^{>}}{(1+\ )c^2}: \qquad (24)$$

Let $= E_X \quad P \quad (X)XX^{>}$. Therefore, to guarantee the condition in Eq. (24), it is suf cient to guarantee that $\frac{M^2}{8} \quad \frac{yy^{>}}{(1+\ )c^2}$, which is equivalent to

$$w^{>} \quad w \quad \frac{M^2}{8} w^{>} \quad w \quad \frac{(w^{>} y)^2}{c^2(1+\ )}; \quad 8 \text{unit vector} w \ 2 \ R^d$$

$$\Leftrightarrow \quad \frac{1}{w^{>} \quad w} \; w^{>} \quad \frac{yy^{>}}{(1+\ )c^2} \quad w \quad \frac{M^2}{8}; \quad 8 \text{unit vector} w \ 2 \ R^d:$$

Therefore, it is suf cient to choose such that

$$\frac{M^2}{8} \quad \frac{1}{\max(\ )} \quad \min \quad \frac{yy^{>}}{c^2(1+\ )} \quad \min_{w:kwk_2=1} \frac{1}{w^{>} \quad w} \; w^{>} \quad \frac{yy^{>}}{(1+\ )c^2} \quad w:$$

Since P satis es the constraint de ned in problem (12), we have $\frac{yy^{>}}{c^2}$. Meanwhile, by Lemma 14, we know that $P \quad (x) \quad \frac{\ }{3}$ for any $x \ 2 \ X$, which means that $\frac{\ }{3}$. That is, for any unit vector $w \ 2 \ R^d$, we have

$$w^{>} \quad w \quad \frac{w^{>} y \ ^2}{c^2} \quad \text{and} \quad w^{>} \quad w \quad \frac{\ }{3} \min(\ ) \; ;$$

which together implies $w^{>} \quad w \quad \max \frac{\ }{3} \min(\ ) ; \frac{(w^{>} y)^2}{c^2}$. Therefore, it holds that

$$w^{>} \quad w \quad \frac{w^{>} y \ ^2}{(1+\ )c^2} \quad \max \left(\frac{\ }{3} \min(\ ) ; \frac{w^{>} y \ ^2}{c^2} \right) \quad \frac{w^{>} y \ ^2}{(1+\ )c^2}$$

$$= \max \left(\frac{\ }{3} \min(\ ) \quad \frac{w^{>} y \ ^2}{(1+\ )c^2}; \frac{w^{>} y \ ^2}{(1+\ )c^2} \right)$$

$$\frac{\ }{3(1+\ )} \min(\ )$$

$$\Rightarrow \quad \min \quad \frac{yy^{>}}{c^2(1+\ )} \quad \frac{\ }{3(1+\ )} \min(\ ) \; :$$

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

$$\frac{M^2}{8} = \frac{\min(\ )}{3(1+\ ) \max(\ )} \Rightarrow \quad = \frac{8 \ ^2 \min(\ )}{3M^2 \max(\ )} \quad \frac{\ }{1+\ };$$

Thus, the proof is complete. □

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

Lemma 13. If $E_X \quad P(X)XX^{>}$ is invertible and $c > 0$, then

$$y^{>} E_X \quad P(X)XX^{>} \quad {}^{1} y \quad c^2 \; \Leftrightarrow \; E_X \quad P(X)XX^{>} \quad \frac{yy^{>}}{c^2}:$$

Proof. For simplicity, let $A = E_X \quad P(X)XX^{>} \quad 0$. Then, we consider the block matrix $\begin{matrix} A & y \\ y^{>} & c^2 \end{matrix} \ 2 \ R^{(d+1) \quad (d+1)}$. Let $[u \quad a]^{>} \ 2 \ R^{d+1}$ with $u \ 2 \ R^d$ be some vector.

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

$$[u \quad a] \begin{matrix} A & y \\ y^\top & c^2 \end{matrix} \begin{matrix} u \\ a \end{matrix} = u^\top A u + 2au^\top y + 2c^2 a^2 := r(u; a):$$

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)

$$u = -aA^{-1}y \Rightarrow r(u^*; a) = a^2(c^2 - y^\top A^{-1} y):$$

Since $y^\top A^{-1} y \le c^2$, we know that $r(u^*; a) \ge 0$, which means $r(u; a) \ge 0$ for any $[u \quad a]^\top \in \mathbb{R}^{d+1}$.

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

$$a = \frac{u^\top y}{c^2} \Rightarrow r(u; a^*) = u^\top A u - \frac{(u^\top y)^2}{c^2}:$$

Since $r(u; a) \ge 0$ for any $[u \quad a]^\top \in \mathbb{R}^{d+1}$, we know that $u^\top A u - \frac{(u^\top y)^2}{c^2} \ge 0$ for any $u \in \mathbb{R}^d$.

That is, we have $A \succeq \frac{yy^\top}{c^2}$.

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

C.2.4 Properties of P

A visualization of P is given in Figure 2.

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

Lemma 14. The function $P(x)$ defined in (7), if regarding as a function of $q(x) = x^\top x - 1 \in [-1, \infty)$, satisfies

- $\lim_{q \to 0} P = \frac{1}{2}$ for any $\in (0; 1)$
- When $q = -1$, $P = \frac{1}{2} + \frac{\sqrt{1+4\ ^2}}{2} \ge \frac{1}{3}$ and $P \ge (\log(1 - P) + \log(P)) \ge 2^{\sqrt{-}}$ for any $\in (0; 1)$.
- $\frac{dP}{dq} = \frac{\sqrt{q^2+4\ ^2} - 2\ ^2}{q^2 + \sqrt{q^2+4\ ^2}}$ decreases as q^2 increases. Further, $\frac{dP}{dq} \in [0; \frac{1}{8}]$. Thus, P increases monotonically as q increases and $P(x) \ge \frac{1}{3}$ for any $x \in X$ and ≥ 0.
- $\frac{dP}{dq}\Big|_{q=-1} \ge \frac{1}{10}$ and $\frac{dP}{dq} \ge \frac{1}{2q^2}$ when $q^2 \ge 12\ ^2$.