Active Learning for Stochastic Contextual Linear Bandits

Emma Brunskill
ebrun@stanford.edu

Ishani Karmarkar
ishanik@stanford.edu

Zhaoqi Li
zli9@stanford.edu

Abstract

A key goal in stochastic contextual linear bandits is to efficiently learn a near-optimal policy. Prior algorithms for this problem learn a policy by strategically sampling actions and naively (passively) sampling contexts from the underlying context distribution. However, in many practical scenarios—including online content recommendation, survey research, and clinical trials—practitioners can actively sample or recruit contexts based on prior knowledge of the context distribution. Despite this potential for *active learning*, the role of strategic context sampling in stochastic contextual linear bandits is underexplored. We propose an algorithm that learns a near-optimal policy by strategically sampling rewards of context-action pairs. We prove *instance-dependent* theoretical guarantees demonstrating that our active context sampling strategy can improve over the minimax rate by up to a factor of $\sqrt{d}$, where d is the linear dimension. We also show empirically that our algorithm reduces the number of samples needed to learn a near-optimal policy, in tasks such as warfarin dose prediction and joke recommendation.

1 Introduction

In many applications, algorithm designers seek to leverage reward feedback to develop contextualized decision policies. For example, in healthcare, clinical trial outcomes may be used to design medication dosages adapted to patients' demographics and health conditions [39, 43]. Similarly, in recommendation systems, user interaction data from a small sub-population can help inform content recommendations tailored to specific users' preferences [2, 36]. Personalized decision-making may also aid in AI alignment, as preference data could steer models to become more widely useful [6, 30].

Contextual bandits offer a natural framework to formalize such decision-making problems. In a contextual bandit, we have a finite context set $\mathcal{X}$ and a finite action set $\mathcal{A}$. Contexts are drawn from a distribution $p \in \Delta^{\mathcal{X}}$ (for any $k \in \mathbb{Z}_{>0}$, Δ^k is the k-dimensional simplex.) Each context-action pair yields a (stochastic) reward $r(x, a)$. In applications, contexts correspond, for example, to patients or users, while rewards might reflect medical outcomes, user engagement, or preference alignment.

In the exploration setting, or experiment design setting [13, 24, 28, 46], the goal is to design an *exploration algorithm* which observes sampled rewards $\{r(x_t, a_t)\}_{t=1}^T$ of T context-action pairs $\{(x_t, a_t)\}_{t=1}^T$ and uses these to learn a policy $\hat{\pi}$. We measure the quality of $\hat{\pi}$ by its *(simple) regret*,

$$R(\hat{\pi}) = \max_{\pi \in \Pi} \mathbb{E}[r(x, \pi(x))] - \mathbb{E}[r(x, \hat{\pi}(x))], \tag{1}$$

where $\Pi := \{\pi : \mathcal{X} \to \mathcal{A}\}$ is the space of (deterministic) policies. To design efficient exploration algorithms with theoretical regret bounds, prior works often consider the setting where the rewards are generated by a noisy d-dimensional linear model. This is known as the *stochastic contextual linear bandits* (SCLBs) setting [1, 13, 24, 26, 33, 46]. Prior works on SCLBs learn a near-optimal policy by sampling rewards of context-action pairs, where the contexts are sampled randomly from the distribution p (what we call *passive context sampling*) and actions are sampled strategically.

Preprint. Under review.

However, in many real-world applications, practitioners may already know the context distribution p and can leverage the fact that p is known a-priori to *strategically* select *both* observed contexts *and* observed actions using this knowledge. This approach (which we call *active context sampling*) is naturally examples in clinical trial design [12, 14] and consumer marketing [25]. In such settings, one might hope that by *jointly* optimizing the observed context and action pairs $\{(x_t, a_t)\}_{t\in[T]}$, an exploration algorithm would be able to better leverage existence of context-action pairs that disproportionately reveal useful information about the underlying reward model. Thus, in this work we explore the following question: *can active context sampling reduce the sample complexity needed to learn a high performing policy for stochastic contextual linear bandits?*

A natural question is: does allowing the exploration algorithm to choose both contexts and actions reduce the problem to a linear bandit or classic active learning? This is not the case, because, **while the algorithm can select contexts and actions during exploration, at deployment, the environment produces contexts according to the distribution** p. So, we require a *context-dependent* policy to achieve low regret, and the problem does not directly reduce to a linear bandit.[1] Our techniques build on prior literature on active learning for regression, but our objective is different. In active learning for regression, the loss is measured by mean squared error, so continuous convex optimization results [15] apply. In contrast, in SCLBs, the loss is the *suboptimality* of the policy, which is a discontinuous loss and requires new insights (see Appendix F). We discuss related work in Appendix B.

Motivated by practical applications, we design an exploration algorithm which actively samples contexts *and* actions to learn a near-optimal policy for an SCLB. (See Figure 3.) We provide a *polynomial-time* active context sampling exploration algorithm for SCLBs and prove an *instance-dependent* regret bound for our algorithm. We prove that our instance-dependent regret bound matches the minimax-optimal rate in the worst-case. We demonstrate the power of active context sampling by constructing a class of SCLBs where our instance-dependent regret bounds provably improve—by a $\sqrt{d}$-factor—over the state-of-the-art rates obtained by other polynomial-time algorithms for this problem. We support our theoretical analysis with numerical experiments on warfarin dosage prescription and joke recommendations. These show our active context sampling exploration algorithm significantly reduces the samples needed to learn a good policy, compared to baselines.

2 Our approach and main theoretical results

An SCLB is a tuple $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$. Here, $\mathcal{X}$ is the context set, $\mathcal{A}$ is the action set, $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d_{\neq 0}$ is a known d-dimensional feature mapping, and $p \in \Delta^{\mathcal{X}}$ is a known context distribution. We also have a 1-subgaussian noise distribution ν over $\mathbb{R}$ and *unknown* linear parameter $\theta^\star \in \mathbb{R}^d$. The rewards are modeled as a noisy linear function in the feature mapping, i.e. $r(x, a) \sim \phi(x, a)^\top \theta^\star + \eta$ where $\eta \sim \nu$. We assume that the problem is normalized such that $\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x, a)\| \leq L$ and $\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \mathbb{E}[r(x, a)] \in [0, 1]$, where $\|\cdot\|$ denotes the Euclidean norm. To ensure the problem is well-posed, we always assume the feature mapping is full-rank, i.e., $\text{span}(\{\phi(x, a)\}) = \mathbb{R}^d$. For SCLBs, the regret (1) simply reduces to

$$R(\hat{\pi}) = \mathop{\mathbb{E}}_{x\sim p}[\max_{a\in\mathcal{A}} \phi(x, a)^\top \theta^\star - \phi(x, \hat{\pi}(x))^\top \theta^\star]. \tag{2}$$

Our goal is to design a way to sample *rewards* of a set of (x, a) tuples so that the resulting dataset of context-action-rewards (x, a, r) tuples allows us to learn a near-optimal contextual policy π. This policy π would be used for future deployment, when contexts are sampled from a known distribution p (Recall Figure 3 and see Fig 4.) In this section, we fix $\lambda > 0$ to be a regularization parameter, $T > 0$ to be a sample budget, and $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ to be any SCLB. Recall we assume p is known, in order to inform active context sampling (in Appendix E we consider relaxing this requirement).

Notation. For any $A \in \mathbb{R}^{d\times d}$ with $A \succ 0$ and $x \in \mathbb{R}^d$, we denote $\|x\|_{A^{-1}} := (x^\top A^{-1} x)^{1/2}$. For any set $\mathcal{S} \subset \mathcal{X} \times \mathcal{A}$, we define the covariance matrix $\Sigma_{\mathcal{S}} := \lambda I + \sum_{(x,a)\in\mathcal{S}} \phi(x, a)\phi(x, a)^\top$. Suppose we use ridge regression to learn a policy $\hat{\pi}$. A key quantity in the regret is the following:

$$\Gamma(\mathcal{S}) := \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}}. \tag{3}$$

[1] One could treat each context as a separate linear bandit, but this naive approach would introduce a sample complexity dependence on the number of contexts, which in general can be large relative to the feature dimension.

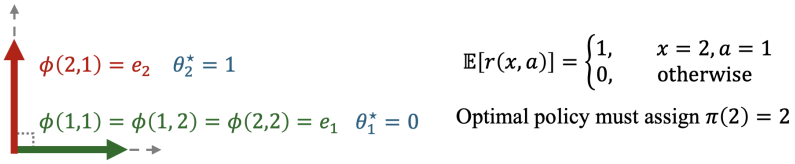


Figure 1: Visualization of a 2D instance with two contexts.

It quantifies how the design of the dataset $\mathcal{S}$ influences the regret bound. Prior works [1, 46] design algorithms (RFLinUCB and Planner-Sampler) which construct a set $\mathcal{S}$ by *passively* sampling contexts $x_t \sim p$ and strategically sampling actions $a_t \sim \pi_t(x_t)$ (where each $\pi_t \in \Delta^{\mathcal{A}}$) such that

$$\Gamma(\mathcal{S}) = \mathop{\mathbb{E}}_{s\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} \leq O(d/T) \text{ with high probability.} \tag{4}$$

Using a standard ridge regression regret bound (restated in Lemma I.1), [1, 46] obtain an overall minimax-optimal regret bound of $\tilde{O}(\sqrt{\beta d/T})$, where β is the usual quantity appeared in ridge regression defined in (17). Our following theorem formalizes a sense in which the uncertainty bound in (4) is tight if one restricts to constructing the dataset $\mathcal{S}$ by passively sampling $x_t \sim p$ (passive context sampling) *regardless* of how the actions a_t are chosen.

Theorem 2.1. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be* any *SCLB. For each $t \in [T]$, let $\pi_t : \mathcal{X} \to \Delta^{\mathcal{A}}$ be arbitrary. Let $\mathcal{S} = \{(x_1, a_1), ..., (x_T, a_T)\} \subset \mathcal{X} \times \mathcal{A}$ such that for each $t \in [T]$, $x_t \sim p$ and $a_t \sim \pi_t(x_t)$. Then, as $\lambda \to 0$, $\mathbb{E}_{\mathcal{S}}[\Gamma(\mathcal{S})] \geq d/T$.*

In light of this obstacle, one hope is that active context sampling—strategically sampling the contexts x_t when constructing $\mathcal{S}$—might allow us to leverage the structure of a *given* SCLB $\mathcal{B}$ to obtain a more fine-grained, *instance-dependent* analysis than the standard minimax-optimal $\tilde{O}(\sqrt{\beta d/T})$ bound. In particular, we aim to design an algorithm that leverages the existence of any disproportionately informative context-action pairs $(x, a) \in \mathcal{X} \times \mathcal{A}$ to obtain instance-dependent rates that obtain regret as low as $\tilde{O}(\sqrt{\beta/T})$ in the best-case, yet no higher than $\tilde{O}(\sqrt{\beta d/T})$ in the worst-case.

A natural approach is to construct $\mathcal{S}$ so that (3) is as small as possible. Although, we might hope to find the optimal $\mathcal{S}^\star = \operatorname{argmin}_{\mathcal{S}\subset\mathcal{X}, |\mathcal{S}|=T} \Gamma(\mathcal{S})$, this is NP-hard [8, 44], so we instead consider a *fractional relaxation*. We seek an optimal sampling *distribution* $w^\star$ with objective value $\mathcal{C}_{\mathcal{B}}$:

$$w^\star := \operatorname*{argmin}_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \text{ and } \mathcal{C}_{\mathcal{B}} := \min_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}. \tag{5}$$

Theorem I.4 shows (5) is a semi-definite program (SDP), enabling polynomial-time solution. Intuitively, $T \cdot w^\star(x, a)$ is the optimal "fraction" of context (x, a) that should be included in $\mathcal{S}$. Such a fractional sampling procedure is not implementable, but we can *simulate* by sampling set $\mathcal{S}$ as follows: for each $t \in [T]$, let $(x_t, a_t) \sim w^\star$ i.i.d. . For T sufficiently large we can expect:

$$\Gamma(\mathcal{S}) = \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} \approx 1/T \cdot \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{w^\star}^{-1}} = \mathcal{C}_{\mathcal{B}}/T. \tag{6}$$

If (6) holds, Lemma I.1 implies a regret of $\tilde{O}(\sqrt{\mathcal{C}_{\mathcal{B}}\beta/T})$! Indeed, Appendix C formally converts the intuition laid out above into a polynomial-time algorithm (Algorithm 1) achieving a regret of $\tilde{O}(\sqrt{\mathcal{C}_{\mathcal{B}}\beta/T})$. As we discuss further in Appendix C, the instance-dependent quantity $\mathcal{C}_{\mathcal{B}}$ can be as low as $O(1)$ (indicating that our result can improve the regret bounds of Planner-Sampler and RFLinUCB by up to a $\sqrt{d}$ factor) is *never* worse than d (ensuring our algorithm is minimax-optimal.)

Theorem 2.2. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB, $\lambda > 0$, $\delta \in (0, 1)$, and $T > 0$ be a sample budget. Invoke Algorithm 1 with $\alpha \leftarrow 1/2$. Let β be as defined in* (17)*. There exists $T_0 = \tilde{O}(d^2)$ so that whenever $T \geq T_0$, with probability $1 - \delta$, Algorithm 1 outputs $\hat{\pi}$ with $R(\hat{\pi}) \leq \tilde{O}(\sqrt{\beta\mathcal{C}_{\mathcal{B}}/T})$. This regret bound is always at most $\tilde{O}(\sqrt{\beta d/T})$ Moroever, the algorithm runs in polynomial time.*

We demonstrate the instance-dependent bound in Theorem 2.2 leads to quantifiably improved performance by the adapting the family of hard instances from [5]. Details of the family are in Appendix D. Figure 1 provides a visualization for a $d = 2$ example with two contexts, one with high-probability $x = 1$ and one with low-probability $x = 2$. The high probability context reveals $\theta_1^\star = 0$, so either action is optimal, but only $a = 1$ is optimal in $x = 2$. As $p(1) \gg p(2)$, passive context sampling repeatedly encounters $x = 1$, and requires many samples to learn the best policy. Meanwhile, active context sampling upsamples $x = 2$ and learns the best policy more efficiently. Appendix D formalizes this to show our approach is stronger than passive sampling by a $\sqrt{d}$ factor on instances like Figure 1.

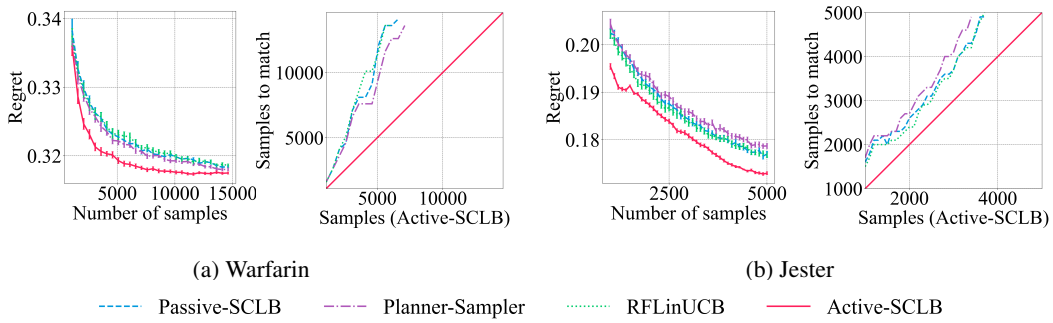


Figure 2: Real-world datasets. The left plots show regret vs. number of samples (mean ± 2 standard errors over 100 trials); the right plots show the minimum number of samples required for baselines' mean regret (over 100 trials) to match Active-SCLB's mean regret for a given sample budget. The *naive baseline* of providing a single best arm to all has a regret of 0.382 (Warfarin) and 0.219 (Jester).

3 Numerical experiments

We implement our exploration policy (Algorithm 1) with $\alpha \leftarrow 0$ (denoted Active-SCLB). We compare Active-SCLB to RF-LinUCB [1, 46] and the Planner-Sampler method of [46]. We also compare to a third baseline (Passive-SCLB), which is a natural passive learning analog of our Active-SCLB, where in Line 2 we enforce additional constraints to force the contexts to be sampled passively from p. Appendix H includes more detail on each baseline and synthetic experimental results.

Here we focus on evaluation of our method on two real-world datasets. We report the regret of a naive baseline, which reports the regret of the naive policy that selects the *same action for each context* ($\pi_{\text{naive}}(x) = \text{argmax}_{a \in \mathcal{A}} \mathbb{E}_{x' \sim p}\, r(x', a)$ for all $x \in \mathcal{X}$). This baseline performs poorly, which certifies that on these real-world datasets, the contextual information is useful. To more realistically model real-world settings where context-action pairs would generally be drawn *without* replacement, for these real-world datasets we (slightly) modify each of the methods to sample context-action pairs without replacement (using rejection sampling) when reporting our results.

For Warfarin, we use the Warfarin Pharmacogenetics Consortium dataset in the warfit-learn package, which is a cleaned clinical dataset of 5650 patients taking blood thinner warfarin [9, 38] (after removing duplicates). Each patient is associated to 31 features corresponding to demographics and health history. The task is to select the best dosage (action) of {"low", "medium", and "high"} for a given patient (context), corresponding to 3 actions and 17,223 context-action pairs. The context distribution p is uniform over contexts. The reward of a context-action (patient-dosage) pair (x, a) is $+1$ if the dosage is correct (0 otherwise). The regret of a policy is the fraction of patients *mis-dosed*.

For Jester, we use the cleaned version [23] of the Jester dataset [19], which contains ratings of 48,447 users on 100 jokes (we subsample down to 2,000 users to keep experiments tractable.) Similar to Kong et al. [23], we hold out the top 5 "gold" jokes with the highest average ratings. For each user, we create a 30-dimensional feature vector by multiplying their 95-dimensional feature vector of joke ratings (for the remaining jokes) with a 95×30 matrix whose entries are iid $\mathcal{N}(0, 1)$ and applying a sigmoid to each of the resulting 30 values. The task is to predict the best "gold" joke (action) for each user (context), resulting in 10,000 context-action pairs. We model the distribution p as uniform across contexts. The reward of (x, a) is the user x's rating of gold joke a.

Figure 2 summarizes findings for $\lambda =$ 1e-6 (other λ are in Appendix H.) Active-SCLB consistently outperforms other baselines (all of which perform similarly.) Active-SCLB often requires 5,000 fewer samples to achieve similar regret to baselines; on Jester, it sometimes requires 1,500 fewer samples.

Summary. We presented an active context sampling algorithm for SCLBs. Our approach enjoys an *instance-dependent* regret guarantee which in the worst-case, matches the minimax-optimal rate and in the best-case is up to a $\sqrt{d}$-factor tighter than comparable works [1, 46]. Two limitations of our work are that scaling SDPs to massive problems could be computationally expensive and that our SCLB setting makes a linear realizability assumption. We discuss these further in Appendix E.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NeurIPS)*, 2011.

[2] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part III 19*, 2012.

[3] Clément L Canonne. A short note on learning discrete distributions. In *arXiv preprint arXiv:2002.11457*, 2020.

[4] Ian Char, Youngseog Chung, Willie Neiswanger, Kirthevasan Kandasamy, Andrew O Nelson, Mark Boyer, Egemen Kolemen, and Jeff Schneider. Offline contextual bayesian optimization. In *Advances in Neural Information Processing Systems*, 2019.

[5] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

[6] Zekai Chen, Po-Yu Chen, and Francois Buet-Golfouse. Online personalizing white-box llms generation with neural bandits. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024.

[7] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[8] Ali Civril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. In *Theoretical Computer Science*, 2009.

[9] International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. In *New England Journal of Medicine*, 2009.

[10] Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. In *arXiv preprint arXiv:2402.10500*, 2024.

[11] Nicollas de Campos Silva et al. Active learning in contextual bandits: handling the uncertainty about the user's preferences in interactive recommendation systems. In *Universidade Federal de Minas Gerais*, 2023.

[12] Kun Deng, Joelle Pineau, and Susan Murphy. Active learning for personalizing treatment. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2011.

[13] Aniket Anand Deshmukh, Srinagesh Sharma, James W Cutler, Mark Moldwin, and Clayton Scott. Simple regret minimization for contextual bandits. In *arXiv preprint arXiv:1810.07371*, 2018.

[14] Zoe Fowler, Kiran Premdat Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib. Clinical trial active learning. In *Proceedings of the 14th ACM international conference on bioinformatics, computational biology, and health informatics*, 2023.

[15] Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on learning theory*, 2015.

[16] Katsuki Fujisawa, Hitoshi Sato, Satoshi Matsuoka, Toshio Endo, Makoto Yamashita, and Maho Nakata. High-performance general solver for extremely large-scale semidefinite programming problems. In *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2012.

[17] Claudio Gentile, Zhilei Wang, and Tong Zhang. Fast rates in pool-based batch active learning. In *Journal of Machine Learning Research*, 2024.

[18] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. In *Machine learning*, 2013.

[19] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. In *Information Retrieval*, 2001.

[20] Paul J Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. In *arXiv preprint arXiv:2405.12762*, 2024.

[21] Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*, 2020.

[22] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[23] Weihao Kong, Emma Brunskill, and Gregory Valiant. Sublinear optimal policy value estimation in contextual bandits. In *International conference on artificial intelligence and statistics*, 2020.

[24] Sanath Kumar Krishnamurthy, Ruohan Zhan, Susan Athey, and Emma Brunskill. Proportional response: Contextual bandits for simple and cumulative regret minimization. In *Advances in Neural Information Processing Systems*, 2023.

[25] Lauren I Labrecque, Ereni Markos, and Aron Darmody. Addressing online behavioral advertising and privacy implications: A comparison of passive versus active learning approaches. In *Journal of Marketing Education*, 2021.

[26] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. In *Cambridge University Press*, 2020.

[27] Xiang Li, Viraj Mehta, Johannes Kirschner, Ian Char, Willie Neiswanger, Jeff Schneider, Andreas Krause, and Ilija Bogunovic. Near-optimal policy identification in active reinforcement learning. In *arXiv preprint arXiv:2212.09510*, 2022.

[28] Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. In *Advances in Neural Information Processing Systems*, 2022.

[29] Zhenwei Lin, Zikai Xiong, Dongdong Ge, and Yinyu Ye. Pdcs: A primal-dual large-scale conic programming solver with gpu enhancements. In *arXiv preprint arXiv:2505.00311*, 2025.

[30] Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, and Min Lin. Sample-efficient alignment for llms. In *arXiv preprint arXiv:2411.01493*, 2024.

[31] Friedrich Pukelsheim. Optimal design of experiments. In *SIAM*, 2006.

[32] Julian Rodemann, Christoph Jansen, and Georg Schollmeyer. Reciprocal learning. In *Advances in Neural Information Processing Systems*, 2024.

[33] Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

[34] Burr Settles. Active learning literature survey. In *University of Wisconsin-Madison Department of Computer Sciences*, 2009.

[35] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in neural information processing systems*, volume 27, 2014.

[36] Liang Tang, Yexi Jiang, Lei Li, and Tao Li. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, 2014.

[37] Joel A Tropp. User-friendly tail bounds for sums of random matrices. In *Foundations of computational mathematics*, 2012.

[38] Gianluca Truda and Patrick Marais. Evaluating warfarin dosing models on multiple datasets with a novel software framework and evolutionary optimisation. In *Journal of Biomedical Informatics*, 2021.

[39] Yogatheesan Varatharajah and Brent Berry. A contextual-bandit-based approach for informed decision-making in clinical trials. In *Life*, 2022.

[40] Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. In *Advances in Neural Information Processing Systems*, 2022.

[41] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, 2022.

[42] Bingyan Wang, Yuling Yan, and Jianqing Fan. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. In *Advances in neural information processing systems*, 2021.

[43] Jixian Wang and Ram Tiwari. Optimal dose selection in phase i/ii dose finding trial with contextual bandits: a case study and practical recommendations. In *Journal of Biopharmaceutical Statistics*, 2025.

[44] William J Welch. Algorithmic complexity: three np-hard problems in computational statistics. In *Journal of Statistical Computation and Simulation*, 1982.

[45] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088, 2006.

[46] Andrea Zanette, Kefan Dong, Jonathan N Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits. In *Advances in Neural Information Processing Systems*, 2021.

A Organization of Appendices

We briefly provide an outline of the contents in this appendix. In Appendix B we discuss related work. In Appendix C we provide additional description and discussion of Algorithm 1. In Appendix D we provide formal guarantees on active versus passive context sampling on a hard instance. In Appendix E we discuss how some of the SCLB modeling assumptions in the main body can be relaxed to still yield interesting results. In Appendix F we include a helpful visualization to compare our setting of active context sampling for contextual bandits with prior work on linear bandits, passive context sampling for contextual bandits, and traditional active learning for regression. In Appendix G we discuss additional theoretical findings, including an active context sampling version of the ContextualRAGE algorithm proposed in [28]. In Appendix H we discuss additional implementational details, including additional experiments on Warfarin dosage and Joke recommendation with different values of the regularization parameter λ. In Appendix I we present all proofs omitted in the main body.

Link to code for experiments. We have made code available at the following anonymous repository: `https://anonymous.4open.science/r/ACLB_release-1B6E/README.md`. Our code will also be released publicly in the final version of the paper.

B Related work

Related work on SCLBs. Exploration algorithms for SCLBs are well-studied [13, 24, 28, 33, 46]. All of these prior works all design algorithms which use *passive context sampling*. In particular, [46] introduced two polynomial-time algorithms, reward-free LinUCB (RFLinUCB) (adapted from the LinUCB algorithm of [1]) and the Planner-Sampler algorithm. Both algorithms use T reward observations $\{r(x_t, a_t)\}_{t\in[T]}$ to learn a policy $\hat{\pi}$ such that $R(\hat{\pi}) \leq \tilde{O}(\sqrt{d\beta/T})$ with high probability.[2] Here, β is a parameter, which under mild assumptions satisfies $\sqrt{\beta} = O(\sqrt{d})$ (defined formally in Lemma I.1, (17)). This rate is known to be minimax-optimal (up to polylogarithmic factors) [7, 46].

Inspired by *practical applications* where active context sampling may be beneficial, our work focuses on leveraging the power of active context sampling to develop a *polynomial-time* algorithm with tighter instance-dependent-regret bounds. Li et al. [28] also studied instance-dependent regret bounds for SCLBs; however, their algorithm again uses passive context sampling and requires exponential time. Thus, it does not lend itself well to practical applications. We discuss Li et al. [28] further in Section D and Appendix G. An orthogonal line of work studies algorithms minimizing the *cumulative regret*, for SCLBs; however this is not directly comparable to our exploration setting (see [1, 26] as well as Appendix E for further discussion of the cumulative regret versus simple regret setting.)

Other related work. SCLBs are a type of linear Markov Decision Process (MDP) with effective horizon of 1, where contexts are the MDP states. Some works [40, 41] obtain instance-dependent rates for MDPs, but neither considers actively sampling states/contexts, and their algorithms are generally computationally hard to implement. Wang et al. [42] and Gheshlaghi Azar et al. [18] consider MDPs in the generative model setting, where states are sampled actively *but* rewards are known a-priori; this is not useful for SCLBs, as the SCLB problem is trivial if rewards are known.

Our work relates to the broad literature on experiment design and active learning (see, [11, 31, 32, 34]), which seeks to design experiments that best reduce statistical uncertainty in unknown variables. Our algorithmic techniques are related to active learning for ridge regression [5, 45] and G-optimal experiment design for linear bandits [26, 31]. Although we are not aware of prior work specifically on active context sampling for SCLBs, there has been some work on active learning for other contextual decision making models. Char et al. [4], Kirschner et al. [22], Li et al. [27] studied contextual Bayesian optimization, where the reward is modeled as a Gaussian process (as opposed to a linear reward model). Their guarantees are weaker, requiring either super-linear dependence on the context-action space [4], or explicit dependence on the dimension d [27]. Lastly, Das et al. [10] studied active context sampling in a preference framework with a Bradley-Terry-Luce reward model.

[2] We use $\tilde{O}(\cdot)$ to hide polylogarithmic factors in the variables $d, 1/\delta$ and L.

C Additional description and discussion of Algorithmic Details

Algorithm 1: Active-SCLB

Input: SCLB $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$, regularization parameter $\lambda > 0$, sample budget $T \in Z_{>0}$, smoothing parameter $\alpha \in [0, 1]$.

// Compute smoothing distribution $\hat{q}$ (see discussion of implementability below).

1 Compute $q \in \Delta^{\mathcal{X}\times\mathcal{A}}$ such that $\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_q^{-1}} \leq 2d$

// Compute a 2-multiplicative approximation of the solution to (7) using an SDP solver.

2 Compute $w \in \Delta^{\mathcal{X}\times\mathcal{A}}$ such that $w(x,a) \geq \alpha q(x,a)$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, and $\mathbb{E}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \leq 2\,\mathbb{E}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{w^*}^{-1}}$

3 $\mathcal{S} \leftarrow \{(x_1,a_1),...,(x_T,a_T)\}$ where each $(x_t,a_t) \sim w$ i.i.d. // Build dataset $\mathcal{S}$ of size T.

4 For $t \in [T]$ sample $r_t \leftarrow r(x_t,a_t)$ // Sample reward observations.

5 $\hat{\theta} \leftarrow \Sigma_{\mathcal{S}}^{-1} \sum_{t\in[T]} \phi(x_t,a_t) r_t$ // Perform ridge regression.

return: $\hat{\pi}(x) \leftarrow x \mapsto \operatorname{argmax}_{a\in\mathcal{A}} \phi(x,a)^\top \hat{\theta}$

First, Line 1 computes a distribution q that *approximates* the *G-optimal distribution* $q^\star$, defined as follows:

Theorem C.1 (Kiefer-Wolfowitz Theorem, Theorem 21.1 of [26], restated)**.** *Suppose that* $q^\star := \operatorname{argmin}_{q\in\Delta^{\mathcal{X}\times\mathcal{A}}} \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|_{\Sigma_q^{-1}}$. *Then* $\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{q^\star}^{-1}} = d$.

There are various polynomial-time algorithms to find q which is a 2-multiplicative approximation to $q^\star$ in the sense of Line 1 (e.g., Franke-Wolfe or an SDP solver; see Chapter 21 of [26]).

Second, in Line 2, for some pre-specified constant $\alpha \in (0,1)$, the algorithm computes a distribution w, which approximates a "α-smoothed" version of $w^\star$ (which we denote by w^*):

$$w^* = \operatorname*{argmin}_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \mathbb{E}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}, \quad \text{subject to } w(x,a) \geq \alpha \cdot q(x,a),\ \forall (x,a) \in \mathcal{X}\times\mathcal{A}. \tag{7}$$

This optimization problem for w^* is *identical* to the one for $w^\star$ in (5), *except* for the constraint that w^* must dominate αq. This constraint is for a technical reason: it ensures that w^* is *well-conditioned* so that matrix-concentration arguments carry through without additional dependencies on L and λ. Fortunately, as the next lemma shows, the objective values attained by w, w^*, and $w^\star$ are close; so we can replace $w^\star$ with w in the approach outlined in the main body, at the cost only constants in sample complexity.

Lemma C.2. *Let* $w^\star$, w^* *be as in* (5), (7)*, respectively. Moreover, let* w *be as in Line 2. Then,* $\mathbb{E}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \leq 2\,\mathbb{E}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{w^*}^{-1}} \leq 2/(1-\alpha) \cdot \mathcal{C}_{\mathcal{B}}$.

Theorem I.4 proves that (7) can be expressed as a semi-definite program (SDP). The proof leverages properties of Schur complements (Lemma I.3.) It is well-known SDPs can be solved to high accuracy in polynomial time (e.g., [21]). Thus, one can compute a w satisfying Line 2 in polynomial time. There are also many practical solvers (e.g., [20]). In experiments, we used MOSEK (see Appendix H.)

Finally, the algorithm draws T samples from w to construct $\mathcal{S}$, performs ridge regression, and outputs a policy using the procedure of Lemma I.1.

We now highlight some advantageous features of Algorithm 1.

- **Choice of parameter** α: The choice of $\alpha = 1/2$ in Theorem 2.2 is purely to enable a finite-sample matrix concentration analysis in the proof. In theory, any constant $\alpha \in (0,1)$ would suffice for the analysis to go through (at the cost of constants in the sample complexity). In practice, we find even $\alpha = 0$ works well and avoids the need to ever compute q. Hence, we set $\alpha \leftarrow 0$ in our experiments.
- **SDP approximation**: The theoretical analysis only requires a 2-multiplicative approximation to Line 2, but in practice, SDPs can be solved to high accuracy. Tighter approximation would lead to tighter constants in our regret bound. In our experiments, we solve the SDP to convergence.
- **Batching**: For $t \in [T]$, Algorithm 1 queries rewards for $\{(x_t,a_t)\}$ drawn i.i.d. from w in Line 3. Importantly, the t-th pair (x_t,a_t) is *not* dependent on the history $\{(x_j,a_j)\}_{j\in[t-1]}$. This ensures the reward observations can be parallelized. Hence, our algorithm falls under the batch-learning paradigm in active learning [17] and non-adaptivity paradigm in SCLBs [46]. This is important

in settings where each reward requires long observation horizons. For example, in drug trials, it may be unreasonable to run trials iteratively on one subject at a time to inform selection of the next subject—hence parallelizing reward observations is important.

- **Robustness to approximation of p**: In some settings, the context distribution p may only be known approximately, i.e., one replaces p in Algorithm 1 with an approximation $\hat{p} \approx p$. For example, in applications, $\hat{p}$ might be constructed as an *empirical distribution* from some historical dataset of sampled contexts. Note that building such a $\hat{p}$ *only* requires *context* data (no reward data), which we believe is available in many applicatoons. In Appendix E, we show that the theoretical guarantees of Algorithm 1 decay *smoothly* as a function of the total-variation distance between p and $\hat{p}$.

D Formal guarantees on active versus passive context sampling

In this section, we formally demonstrate that the instance-dependent bound in Theorem 2.2 leads to quantifiably improved performance on an instance-dependent basis. We describe the following family of SCLB instances (parameterized by A and d) where active context sampling (Algorithm 1) achieves quantitatively improved regret bounds compared to the minimax rate. This instance is adapted from the hard instance for passive learning in maximum likelihood estimation [5].

Definition D.1. Let $d \in \mathbb{Z}_{>0}, A \in \mathbb{Z}_{>1}$. Let $\mathcal{B}^{\star}_{d,A} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^{\star})$ be an SCLB with $\mathcal{X} = [d]$, $\mathcal{A} = [A]$, $\nu = \mathcal{N}(0, 1)$ and $\phi, p, \theta^{\star}$ defined as follows, where e_i is the i-th standard basis vector:

$$\phi(x, a) = \begin{cases} e_x & \text{if } a = 1, \\ e_1 & \text{otherwise} \end{cases}, \quad p(x) = \begin{cases} 1 - \frac{d-1}{d^2} & \text{if } x = 1, \\ 1/d^2 & \text{otherwise} \end{cases}, \quad [\theta]^{\star}_i = \begin{cases} 0, & \text{if } i = 1 \\ 1, & \text{otherwise} \end{cases}.$$

This SCLB has one high-probability context $x = 1$ and $(d - 1)$ low-probability contexts. The high probabilty context *always* reveals $\theta^{\star}_1$, which is 0. Any algorithm can only learn about high-reward actions when it queries a context $x \neq 1$. If an algorithm passively samples contexts, this happens rarely (only with probability roughly $1/d$.) On the other hand, an active context sampling algorithm can actively *upsample* these rarer contexts in order to gain information about $\theta^{\star}$'s remaining $(d - 1)$ coordinates—-allowing it to more efficiently determine the best action to sample for these contexts. This intuition suggests that active context sampling should perform well on $\mathcal{B}_{d,A}$. We formalize this by showing $\mathcal{C}_{B^{\star}_{d,\mathcal{A}}}$ is *independent of* d and λ, as follows.

Lemma D.2. *For any $d \in \mathbb{Z}_{>0}, A \in \mathbb{Z}_{>1}$, $\mathcal{C}_{\mathcal{B}^{\star}_{d,A}} \leq 4$.*

Thus, on $\mathcal{B}^{\star}_{d,A}$ Theorem 2.2 gives a regret bound of $\tilde{O}(\sqrt{\beta/T})$. In contrast, the prior polynomial-time algorithms (RFLinUCB and Planner-Sampler) passively sample contexts and only guarantee a bound of $\tilde{O}(\sqrt{\beta d/T})$, which is worse by a factor of $\sqrt{d}$ [46]. This illustrates a concrete case where our result can be stronger than prior work by up to a $\sqrt{d}$ factor! We validate this improvement empirically in the next section. In addition, in Appendix E, we show that this $\sqrt{d}$ improvement remains *even* if p is unknown but *approximated* from roughly $\Theta(d^2)$ i.i.d. *context samples* from p.

Comparison to adaptive sampling SCLB bounds. Li et al. [28] use passive context sampling with a *data-adaptive* exploration policy to obtain instance-dependent rates, which, in some cases might be tighter than Theorem 2.2. Their algorithm uses passive context sampling and strategically sample actions with a reward gap data-adaptively. Li et al. [28] extends prior work on data-adaptive best-arm identification for stochastic non-contextual bandits to stochastic contextual linear bandits (SCLBs). However, to our knowledge there is no polynomial-time implementation of the algorithm from Li et al. [28]. Motivated by applications, our focus is on designing polynomial-time algorithms.

Nonetheless, introducing active context sampling would *still* improve over the passive-learning data-adaptive sampling algorithm in [28]! Appendix G presents a data-adaptive active context sampling analog of Li et al. [28]'s exponential-time algorithm and proves this would improve over the regret bounds attained in [28]—in particular, on the class of SCLBs proposed in Definition D.1, our active-context sampling variant of Li et al. [28]'s original passive-context sampling algorithm tightens the resulting bound by a factor of $\sqrt{d}$, as we find for our polynomial-time algorithm. While we omit this analysis in the main body (as our focus is on tractable algorithms) our results highlight that active context sampling has potential to strengthen contextual bandit algorithms more broadly.

E Discussion of SCLB model, limitations, and relaxations

In this section, we discuss in more detail the role of some assumptions of our SCLB model, and how some of them can be relaxed under appropriate conditions.

E.1 Relaxing the assumption that the context distribution p is known.

One assumption in our work is that the context distribution p is known a-priori. Although we believe this to be reasonable for some applications (see for example, the motivating examples presented in Section 1); in other settings p may only be known *approximately* from *prior historical data from the context distribution*.

Consequently, a natural question is the following. Suppose that the exact context distribution p is unknown a priori and that the learner only has access to an approximate context distribution $\hat{p} \in \Delta^{\mathcal{X}}$. Suppose that one runs Algorithm 1 with $\hat{p}$ in place of p. How do the theoretical guarantees decay as a function of the error between p and $\hat{p}$?

We address this question below. In Section E.1.1 we show that the theoretical guarantees of our algorithm decay smoothly as a function of the total variation distance between p and $\hat{p}$, and we quantify the amount of historical context data needed to control this error. In Section E.1.2 we show that on the bandit instance from Definition D.1, a coarse approximation of $\hat{p}$ constructed from a dataset of $\Omega(d^2)$ sampled contexts is sufficient to maintain the $\sqrt{d}$ improvement over passive context sampling.

Lastly, a perhaps another related question is: when constructing w in Line 2) is necessary or would taking $w \leftarrow q^\star$ (recall that $q^\star$ does not depend on p) obtain just as strong a rate. This is *not* the case, as we discuss in Section E.2.

E.1.1 Theoretical guarantees decay with the total variation distance between p and $\hat{p}$

In the following, we use $\mathrm{tv}(p, p') := 1/2 \cdot \sum_{x \in \mathcal{X}} |p(x) - p'(x)|$ to denote the total variation distance between two discrete distributions over a context set $\mathcal{X}$.

Theorem E.1. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB, $\lambda > 0$, $\delta \in (0, 1)$, and $T > 0$ be a sample budget. Suppose that p is unknown but that one has access to an arbitrary approximation $\hat{p} \in \Delta^{\mathcal{X}}$. Let $\hat{\mathcal{B}} = (\mathcal{X}, \mathcal{A}, \phi, \hat{p}, \nu, \theta^\star)$ be the corresponding approximate bandit instance and β be as defined in (17). There exists $T_0 = \tilde{O}(d^2)$ so that whenever $T \geq T_0$, with probability $1 - \delta$, Algorithm 1 outputs $\hat{\pi}$ such that the regret evaluated on the true bandit instance $\mathcal{B}$ satisfies*

$$R(\hat{\pi}) = \mathop{\mathbb{E}}_{x \sim p}[\max_{a \in \mathcal{A}} r(x, a) - r(x, \hat{\pi}(x))] \leq \tilde{O}\left(\sqrt{\beta \cdot \frac{\mathcal{C}_{\mathcal{B}} + d\,\mathrm{tv}(p, p')}{T}}\right).$$

Moroever, the algorithm runs in polynomial time.

To prove the theorem, we first prove the following helper lemma.

Lemma E.2. *Consider the setting of Theorem E.1. Let q be as in Line 1. Let $\hat{w}$ be the choice that Line 2 of Algorithm 1 would select when invoked on $\hat{\mathcal{B}}$ with $\alpha \leftarrow 1/2$. Then,*

$$\mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma^{-1}_{\hat{w}}} \leq 4\mathcal{C}_{\mathcal{B}} + 32d\,\mathrm{tv}(p, \hat{p}).$$

Proof. For notational convenience, for any bandit $\mathcal{B}' = (\mathcal{X}, \mathcal{A}, \phi, p', \nu, \theta^\star)$ and for any distribution $w \in \Delta^{\mathcal{X}, \mathcal{A}}$, define

$$\kappa(\mathcal{B}', w) := \mathop{\mathbb{E}}_{x \sim p'} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma^{-1}_w}.$$

Consider any w which is feasible for (7) for $\alpha = 1/2$. Then, by the constraint that w dominates $\alpha \cdot q = 1/2q$ we have that $\Sigma_w \succeq 1/2 \cdot \Sigma_q$, and consequently,

$$\left|\kappa(\mathcal{B}, w) - \kappa(\hat{\mathcal{B}}, w)\right| = \left|\mathbb{E}_{x \sim p}[\max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma^{-1}_w}] - \mathbb{E}_{x \sim \hat{p}}[\max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma^{-1}_w}]\right|$$

$$
\begin{aligned}
&= \left| \sum_{x \in \mathcal{X}} [p(x) - \hat{p}(x)] \cdot \max_{a \in \mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \right| \\
&\leq \sum_{x \in \mathcal{X}} |p(x) - \hat{p}(x)| \cdot \max_{a \in \mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \\
&\leq 4d \sum_{x \in \mathcal{X}} |p(x) - \hat{p}(x)| \\
&\leq 8d \cdot \mathrm{tv}(p, \hat{p})
\end{aligned}
$$

where the second-to-last inequality holds because $\Sigma_w \succeq 1/2 \cdot \Sigma_q$ ensures $\|\phi(x,a)\|^2_{\Sigma_w^{-1}} \leq 4d$ and the last inequality holds by definition of the total-variation distance $\mathrm{tv}(p, \hat{p})$.

Now, $\hat{w}^*$ and w^* be the choices of distributions that (7) would select when invoked on $\hat{\mathcal{B}}$ and $\mathcal{B}$ respectively. That is,

$$
\begin{aligned}
w^* &= \operatorname*{argmin}_{w \in \Delta^{\mathcal{X} \times \mathcal{A}}} \mathbb{E}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}, \quad \text{subject to } w(x,a) \geq \alpha \cdot q(x,a),\ \forall (x,a) \in \mathcal{X} \times \mathcal{A} \\
\hat{w}^* &= \operatorname*{argmin}_{w \in \Delta^{\mathcal{X} \times \mathcal{A}}} \mathbb{E}_{x \sim \hat{p}} \max_{a \in \mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}, \quad \text{subject to } w(x,a) \geq \alpha \cdot q(x,a),\ \forall (x,a) \in \mathcal{X} \times \mathcal{A}.
\end{aligned}
$$

Then, by the argument above, we have that

$$
\kappa(\mathcal{B}, \hat{w}^*) \leq \kappa(\hat{\mathcal{B}}, \hat{w}^*) + 8d\mathrm{tv}(p, \hat{p}) \leq \kappa(\hat{\mathcal{B}}, w^*) + 8d\mathrm{tv}(p, \hat{p}) \leq \kappa(\mathcal{B}, w^*) + 16d\mathrm{tv}(p, \hat{p}),
$$

where the first step follows because $|\kappa(\mathcal{B}, \hat{w}^*) - \kappa(\hat{\mathcal{B}}, \hat{w}^*)| \leq 8d\mathrm{tv}(p, p')$, the second step follows because of the optimality conditions for $\hat{w}^*$, and the third step follows because $|\kappa(\mathcal{B}, w^*) - \kappa(\hat{\mathcal{B}}, w^*)| \leq 8d\mathrm{tv}(p, p')$.

Noting the definition of $\hat{w}$, we conclude

$$
\kappa(\mathcal{B}, \hat{w}) \leq 2\kappa(\mathcal{B}, \hat{w}^*) \leq 2\kappa(\mathcal{B}, w^*) + 32d\mathrm{tv}(p, \hat{p}).
$$

The result now follows by Lemma C.2, which ensures that $\kappa(\mathcal{B}, w^*) \leq 2\mathcal{C}_{\mathcal{B}}$. □

Proof of Theorem E.1. The proof follows identically as that of Theorem 2.2, except that the second-to-last display instead holds (by Lemma E.2) with

$$
\mathbb{E}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x,a)\|_{\Sigma_{\mathcal{S}}^{-1}} \leq \frac{2}{T} \mathbb{E}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \leq \frac{8}{T} \mathcal{C}_{\mathcal{B}} + 64d \cdot \mathrm{tv}(p, p').
$$

□

We also optain the following useful corollary.

Corollary E.3. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB, $\lambda > 0$, $\delta \in (0,1)$, and $T > 0$ be a sample budget. Suppose that p is unknown but that one has access to $M = \tilde{\Theta}(|\mathcal{X}|\, d^2 \epsilon^{-2})$ historical iid samples from p. Let $\hat{p}$ be the empirical context distribution constructed from these M samples.*

Let $\hat{\mathcal{B}} = (\mathcal{X}, \mathcal{A}, \phi, \hat{p}, \nu, \theta^\star)$ be the corresponding approximate bandit instance and β be as defined in (17)*. There exists $T_0 = \tilde{O}(d^2)$ so that whenever $T \geq T_0$, with probability $1 - \delta$, Algorithm 1 outputs $\hat{\pi}$ such that the regret evaluated on the true bandit instance $\hat{\mathcal{B}}$ satisfies*

$$
R(\hat{\pi}) = \mathbb{E}_{x \sim p}[\max_{a \in \mathcal{A}} r(x,a) - r(x, \hat{\pi}(x))] \leq \tilde{O}\left(\sqrt{\beta \cdot \frac{\mathcal{C}_{\mathcal{B}} + \epsilon}{T}} \right).
$$

Moroever, the algorithm runs in polynomial time.

Proof. The proof follows by applying Theorem E.1 and noting that $M = \tilde{O}(|\mathcal{X}| d^2 \epsilon^{-2})$ is sufficient samples such that with high probability, $\mathrm{tv}(p, p') \leq \epsilon / d$ [3]. □

E.1.2 Sustained improvement on the bandit instance from Definition D.1.

A natural and important question is whether the gains to active context sampling over passive sampling strategies still persist when we use an empirical estimate of the context distribution. We expect this to be true, and we now show on the example from Definition 5.1, that optimizing with respect to an empirical context distribution (instead of the true context distribution) still enables our approach to recover our $\sqrt{d}$-factor improvement over passive context sampling. For the example instance in Definition 5.1, suppose that we build an empirical estimate $\hat{p}$ using M samples of *only contexts*. We believe it is very common for there to be prior large datasets of contexts – consider recommendation systems, etc. Recall the multiplicative Chernoff bounds: for $X_{1\cdots M}$ independent Bernoulli random variables in (0,1) with mean p_k, and $\epsilon < 1$ then

$$\mathbb{P}\left(\frac{1}{M}\sum_i X_i \geq (1+\epsilon)p_k\right) \leq e^{-\epsilon^2 M p_k/(2+\epsilon)}$$

and

$$\mathbb{P}\left(\frac{1}{M}\sum_i X_i \leq (1-\epsilon)p_k\right) \leq e^{-\epsilon^2 M p_k/2}$$

Set $\epsilon = .25$. Thus, $M_k = \frac{32}{p_k}\log(2/\delta)$ is sufficient to ensure the resulting estimate $\mathbb{P}(\hat{p}_k \geq (1+\epsilon)p_k) + \mathbb{P}(\hat{p}_k \leq (1-\epsilon)p_k) \leq \delta$, where $\hat{p}_k$ is the empirical estimate $\frac{1}{M}\sum_i X_i$. We use a union bound to ensure this holds for all d contexts, and chose the minimum probability $p_k = \frac{1}{d^2}$, $M = 32d^2\log(2d/\delta)$. This ensures that with probability at least $1-\delta$,

$$\mathbb{P}\left[\max_{x\in\mathcal{X}} |p(x) - \hat{p}(x)| \leq p(x)/4\right].$$

Condition on the above event in the remainder of this argument. Note that under the above event, we have that for all $x > 1$,

$$\hat{p}(1) \geq \frac{3}{4}\left(1 - \frac{d-1}{d^2}\right) > \frac{5}{4}\cdot\frac{1}{d^2} \geq \hat{p}(x).$$

Now, consider the SDP problem we solve in Algorithm 1 if we use $\hat{p}$ in place of p. Note that the feature e_1 can be accessed through the context-action pair $(x=1, a=1)$ so we may assume without loss of generality that $w(x,a)$ is only supported at $a=1$. We have

$$\Sigma_w = \sum_{x,a} w(x,a)\phi(x,a)\phi(x,a)^\top = \sum_{i=1}^d z_i e_i e_i^\top$$

where $z_i := w(i,1)$ for each $i \in [d]$. Then

$$\Sigma_w^{-1} = \sum_{i=1}^d z_i^{-1} e_i e_i^\top,$$

and consequently,

$$\max_a \|\phi(x,a)\|^2_{\Sigma_w^{-1}} = \max_a \phi(x,a)^\top \Sigma_w^{-1}\phi(x,a) = \sum_{i=1}^d \max_a z_i^{-1}\|\phi(x,a)\|_2^2.$$

Recall the definition that $\phi(x,a) = e_x$ if $a = 1$ and e_1 otherwise. Hence, the max becomes $\min(z_1, z_x)^{-1}$. Plugging this in the formulation of SDP gives

$$\mathop{\mathbb{E}}_{x\sim\hat{p}} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} = \sum_x \hat{p}(x)\cdot\frac{1}{\min(z_1, z_x)}.$$

Next, we prove that $z_1 \geq z_i$ for all $i \neq 1$. Indeed, suppose for the sake of contradiction that $z_1 < z_i$ for some $i \neq 1$. Then consider $z' \in \Delta^{\mathcal{X}}$ such that $z'_x = \frac{1}{2}(z_1 + z_i)$ if $x = 1$ or $x = i$, and $z'_x = z_x$

for $x \neq 1$ and $x \neq i$. Note that $z'_1 = z'_i > z_1$ by this construction. Let w and w' be the weights correspond to z and z'. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\hat{p}} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} &= \frac{\hat{p}(1)}{z_1} + \frac{\hat{p}(i)}{z_1} + \sum_{x\neq 1, x\neq i} \hat{p}(x)\cdot\frac{1}{\min(z_1, z_x)} \\
&> \frac{\hat{p}(1)}{\frac{1}{2}(z_1+z_i)} + \frac{\hat{p}(i)}{\frac{1}{2}(z_1+z_i)} + \sum_{x\neq 1, x\neq i} \hat{p}(x)\cdot\frac{1}{\min(z_1, z_x)} \\
&= \frac{\hat{p}(1)}{z'_1} + \frac{\hat{p}(i)}{z'_i} + \sum_{x\neq 1, x\neq i} \hat{p}(x)\cdot\frac{1}{\min(z'_1, z'_x)} \\
&= \mathop{\mathbb{E}}_{x\sim\hat{p}} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{w'}^{-1}}.
\end{aligned}$$

Therefore, such w (and thus such z) cannot be a minimizer of the expectation. Hence, the SDP reduces to

$$\min_{z\in\Delta^n} \sum_{x\in\mathcal{X}} \hat{p}(x)\cdot\frac{1}{z_x}.$$

Using Lagrange multipliers we get the closed form solution, $z_x = \frac{\sqrt{\hat{p}(x)}}{\sum_x \sqrt{\hat{p}(x)}}$, and optimal value

$$\left(\sum_{x\in\mathcal{X}} \sqrt{\hat{p}(x)}\right)^2.$$

Moreover, using the guarantee above on the accuracy of the learned $\hat{p}(x)$ parameters given M prior contexts, we prove

$$\left(\sum_{x\in\mathcal{X}} \sqrt{\hat{p}(x)}\right)^2 \leq \left(\sum_{x\in\mathcal{X}} \sqrt{1.25\cdot p(x)}\right)^2 = \left(\sqrt{1.25\cdot\frac{d^2-d+1}{d^2}} + 1.25\cdot\sum_{x=2}^{d}\frac{1}{d}\right)^2 = O(1).$$

Therefore, given $M = 32d^2\log(2d/\delta)$ samples, with high probability the active context sample algorithm using the empirically-derived context probabilities will also improve over passive context sampling by a $\sqrt{d}$ factor. This shows that at least in some settings, using the empirical distribution to estimate the context distribution will yield the same $\sqrt{d}$ improvement in the resulting regret bounds, compared to passive context sampling methods.

E.2 Using the G-optimal distribution directly is insufficient

A perhaps natural question is the following. When constructing w in Line 2) is the SDP solution w truly necessary, or would taking $w \leftarrow \hat{q}$ (which does not depend on p) obtain just as strong a rate?

This is *not* the case. Indeed, Theorem C.1 shows that the G-optimal design $\hat{q}$ satisfies $\max_{(x,a)} \|\phi(x,a)\|_{\Sigma_{\hat{q}}^{-1}}^2 = d$. By Lemma I.1, this in general yields a bound of $\sqrt{d\beta/T}$. This gives no improvement over passive context sampling.

As an illustrative example, consider again the bandit instance from Definition D.1. The G-optimal design can select $(x, 1)$ with probability $1/d$ for each $x \in [d]$, but that upsamples the rare contexts too much, causing a d-dependence in the bound. Indeed, we find that

$$\mathbb{E}_{x\sim p}\|e_x\|^2_{\Sigma_{\hat{q}}^{-1}} = \left(\frac{d^2-d+1}{d^2}\right)\cdot d + \sum_{i=2}^{d}\frac{1}{d^2}\cdot d = \Omega(d),$$

and hence, Lemma I.1 suggests that even on this simple bandit instance, if one were to modify Algorithm 1 to use $w \leftarrow \hat{q}$ in Line 2, then one would again obtain a worse d-dependence than using the active context sampling distribution $w = \alpha\hat{q} + (1-\alpha)\hat{w}$ as in our original Algorithm 1.

E.3 Assumption that the reward model is linear.

Our method is immediately amenable to some non-linear reward models in the sense that one could always lift the feature vector into a higher-dimensional space which captures non-linear relationships between the original feature entries. To briefly explain why this is the case, recall that one can always use linear regression algorithms to fit a quadratic regression model just by modifying the feature vector to include the transformed variable x^2 in addition to x as a feature. What is important is just that the reward is linear (or approximately so) in the lifted space. See, for example, the empirical section of Kong et al. whom our paper cites.

In principle, the idea of active context sampling could also apply to kernelized contextual bandits. However, the main challenge is that in these more complex reward models, it is not clear whether the optimization problem to solve for the optimal sampling distribution is computationally tractable (e.g., solvable in polynomial time.) Some prior works (see Line 118) have discussed greedily sampling contexts for the kernelized settings, but this does not result in tight guarantees that avoid a dependence. Thus, while developing theoretically-grounded techniques for very general reward models might be challenging, we hope that our work provides potentially useful insights towards tackling more sophisticated reward models as future work. Moreover, as discussed in Section 3, our empirical results on real-world data seem robust to model misspecification, which indicates that our methods could outperform their theoretical guarantees even if the true reward model is not linear.

E.4 Pure exploration setting and choice of simple regret as the objective function.

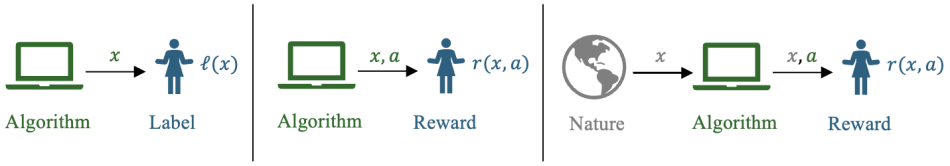
In this section, we discuss the motivation for our pure exploration setting as well as our choice of regret function.

Pure exploration setting. First, we note that if one's goal is to minimize cumulative regret, alternate algorithms may be preferable. We do not have bounds on the cumulative regret. The reason we focus on pure exploration is that it is better aligned with some important settings where a fixed experimental period is more common. For example:

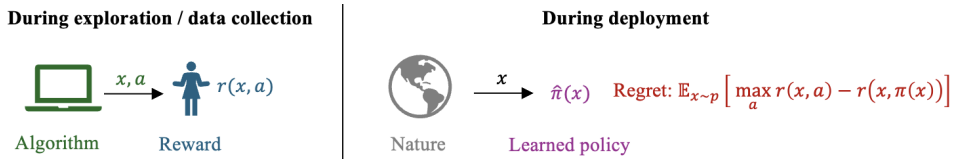
- Social media companies often run experimental studies on users. A company might be willing to test out several different strategies (knowing that it may temporarily lose on engagement) in order to quickly learn a high-quality strategy to deploy in production long-term.
- In medical settings, one might design a monitored clinical trial in which patients are actively monitored in case of any adverse consequences to the administered treatment. Such monitoring is not possible long-term or at-scale, however, the willingness to take a measured risk to quickly explore new treatments during a short-term trial can be valuable for eventually identifying good treatments to prescribe to the general public.
- Another natural setting where the pure exploration setting is well-suited is applications where we have the ability to simulate the reward of a particular context-action pair by running an (expensive) simulation. This may arise in scientific decision making where physical or medical simulations are available. In such cases, one is unconcerned with the regret during exploration.

To summarize, during the exploration period, our regret may be high, however, this freedom to explore actions freely enables us to quickly (i.e., with few samples) converge upon a near-optimal policy which can be deployed at scale after the exploration stage. In contrast, if one tries to balance exploration and exploitation as in the online bandit setting, convergence to an optimal policy could be slower. This is precisely what motivates the extensive prior work on the pure exploration model [13, 24, 46].

As an important related note, there is a sense in which active context sampling *only* makes sense in the pure exploration setting and does not make sense in the cumulative regret minimization setting. As a thought experiment, suppose we modify the cumulative regret minimization setting to allow the learner to actively select the context at each round t. Then, there is a trivial way to achieve low cumulative regret: the algorithm could just select the *same* context x' in every single round and learn an optimal action for context x'. This would have low cumulative regret, since the learner only needs to learn a good action in context x'. However, clearly, this is not a very interesting setting, since the learner would never learn to perform well on the real-world distribution p.



(a) Active context sampling for contextual bandits (center) lies between classic active learning (left), e.g., for regression, and passive context sampling for contextual bandits (right).



(b) During exploration, active context sampling allows selecting *both* contexts and actions. However, the goal is to learn a policy that performs well in *deployment*, when contexts will be sampled from p.

Figure 3: Active context sampling for contextual bandits.

In contrast, our work allows the learner to actively sample contexts freely during the exploration phase; however, it must learn a policy that will perform well in the eventual real-world distribution p. This is diagrammed in Figure 4a.

Simple regret vs best arm identification. In PAC-learning for best-arm identification, the goal is to find the best arm in every context [35]. That is, the goal would be to identify an approximately optimal action in every context.

In contrast, in simple regret, we are okay with using a suboptimal arm in some rare contexts as long as the learned policy performs well on average over the distribution of contexts. In this sense, best-arm-identification is a very strict "solution concept" which requires strong performance uniformly over every context. Meanwhile, simple regret is more relaxed and often more realistic in that it only asks for learning context-to-action policy which performs well (on average) over the context distribution. In practical scenarios, we are often satisfied with a policy which works very well on average (even if it is suboptimal in some rare contexts). However, to truly find the best arm in every context as in best-arm identification, we might need to spend many more samples.

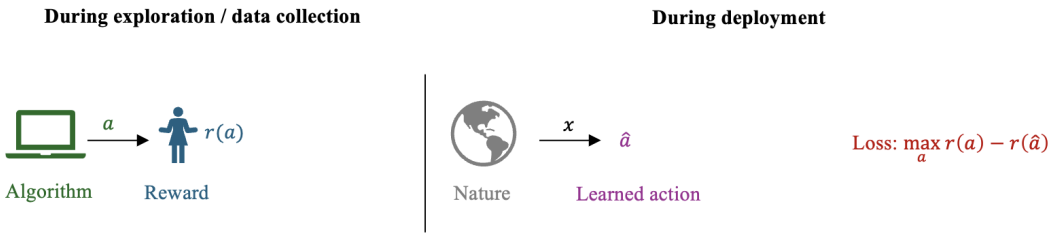
This is why prior works have also looked at simple regret under the SCLB model to better capture real-world scenarios where average-performance of a policy is the key performance indicator (Zanette et al., Li et al., Deshmukh et al., Krishnamurthy et al. as cited in our paper).

E.5 Limitations of SDP solving and linear realizability.

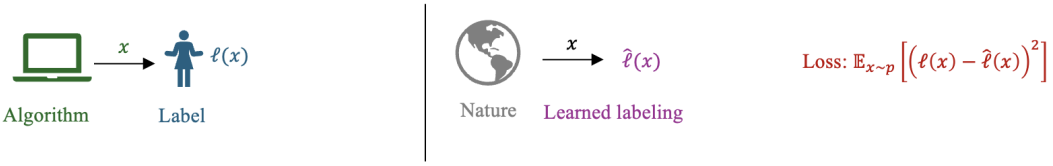
One limitation is that SCLBs make a linear realizability assumption [1, 46]. However, on the real-world Warfarin and Jester datasets, Algorithm 1 performs well even though we expect the reward models are nonlinear, suggesting our procedure can perform well empirically *even* if the problem is *misspecified* as an SCLB. Misspecification may also explain why the regret is not strictly decreasing in Figure 2. Another limitation is that, while SDP-solving is fast for moderately-sized SDPs, SDP-solving could be expensive in massive context-action spaces. Our experiments on Warfarin and Jester demonstrate our method can easily scale to at least 15,000 context-action pairs. On both datasets, SDP solving runs in under an hour. To scale to massive problems, one might consider parallel/GPU-accelerated SDP solvers [16, 29]. Practitioners might also experiment with heuristic approximations of Line 2 (e.g., sub-sampling the context-action pairs prior to solving the SDP); we conjecture it is nontrivial to provide theoretical guarantees for such heuristics.

F Comparison of active learning settings

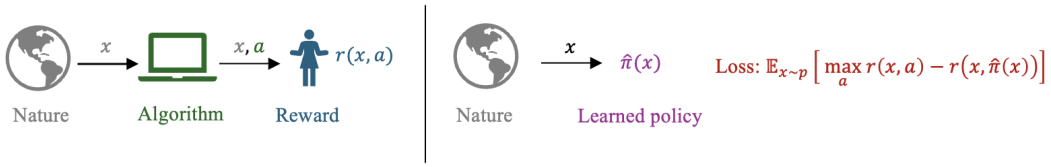
In this section, we include a helpful visualizations (Figures 3 and 4) to compare linear bandits, active learning (e.g., for regression), passive learning for SCLBs, and active learning for SCLBs.



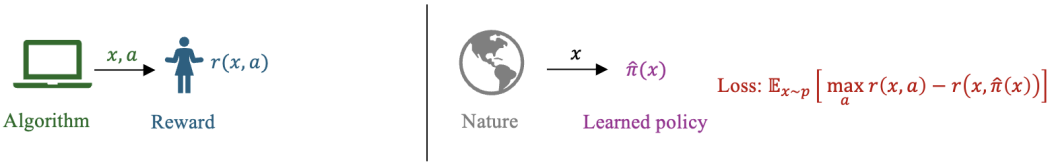
(a) Linear bandits



(b) Active learning (e.g., for regression)



(c) Passive context sampling for SCLBs



(d) Active context sampling for SCLBs

Figure 4: Four learning paradigms. Figure 4a describes the typical linear bandit setting, where there are no *contexts* and the learner's goal is to learn a good universal action. Figure 4b shows active learning for regression, with the mean-squared loss. During exploration, the learner actively samples data x to learn a label function $\hat{\ell}(x)$; but during deployment, inputs arrive according to the distribution p. In other settings, such as active learning for classification, this mean-squared loss might be replaced with another *smooth, convex* loss function. In contrast, in contextual bandits (Figures 4c and 4d), the learner needs to learn a policy-to-action mapping. In this case, the loss function function is the sub-optimality of the policy—which is discontinuous—and consequently, traditional techniques from continuous optimization do not immediately apply.

G Additional theoretical results

In this appendix, we expand on additional results which were omitted from the main body.

G.1 Probably-approximately-correct guarantee

Here, we state a probably-approximately-correct (PAC)-style version of our main result (Theorem 2.2). The following theorem is an immediate corollary of Theorem 2.2, and makes an assumption on $\theta^\star, \lambda$ in order to more directly compare to the PAC-learning versions reported in [1, 46].

Corollary G.1 (PAC-learning version of Theorem 2.2)**.** *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB and $\epsilon, \delta \in (0, 1)$. Assume $\|\theta^\star\| \leq \tilde{O}(1)$ and $\lambda \leq \tilde{O}(1)$. When initialized with a sample budget of $T \geq T_1 = \tilde{O}(\mathcal{C}_\mathcal{B} \cdot d\epsilon^{-2} + d^2)$ and $\alpha = 1/2$, Algorithm 1 returns a policy $\hat{\pi}$ such that with probability $1 - \delta$, $R(\hat{\pi}) \leq \epsilon$.*

Proof. From the definition of β (17), it is clear that so long as $\lambda, \|\theta^\star\| \leq \tilde{O}(1)$, $\beta \leq \tilde{O}(d)$. The corollary now follows immediately Theorem 2.2. □

In comparison, the methods from Zanette et al. [46] and Abbasi-Yadkori et al. [1] (RFLinUCB and Planner-Sampler) require $\tilde{O}(d^2\epsilon^{-2})$ samples to achieve ϵ regret under the same assumptions.

Recalling that $\mathcal{C}_\mathcal{B} \leq d$, we see that our sample-complexity guarantee of $\tilde{O}(\mathcal{C}_\mathcal{B} d\epsilon^{-2})$ is always *at least* as strong as that of Planner-Sampler and RFLinUCB Abbasi-Yadkori et al. [1], Zanette et al. [46] and may be as low as $\tilde{O}(d^2 + d/\epsilon^2)$ when $\mathcal{C}_\mathcal{B} = O(1)$ (recall Section D which gives an example where this occurs.) Thus, in the PAC-learning setting, we improve over Planner-Sampler and RFLinUCB by up to a dimension factor.

Remark G.2. Note that our restriction to $\epsilon \in (0, 1)$ is without loss of generality. Because we assume that rewards are bounded between $[0, 1]$ in expectation, if $\epsilon > 1$, then *any* policy has regret at most ϵ, and consequently, the problem is trivial.

G.2 Instance-dependent guarantees and comparison to [28]

In this section, we compare in more detail against the instance-dependent rates of [28] for SCLBs. We did not discuss this in detail in the main body, because the instance-dependent sample complexity rates of [28] for SCLBs are obtained using a computationally intractable algorithm (see, e.g., discussion around Theorem 2.14 of and conclusion of [28].) In contrast, our goal in this work is to focus on practical applications where active context sampling may be helpful; hence, our goal was to design polynomial-time implementable algorithms which obtain instance-dependent rates for SCLBs. Because it is not known how to implement or even approximately implement the guarantees of [28] in polynomial time, note that the result of Theorem 2.14 in [28] may be an unfair comparison to our result Theorem 2.2.

Correspondingly, our goal of this section, is to discuss the result of [28] in greater detail and show that their rates can also be improved with active context learning—if we disregard the concerns over computational tractability. In Section G.2.1, we first state the main result of [28], which uses passive context sampling to design an instance-dependent algorithm (ContextualRAGE) for SCLBs. In Section G.2.2, we will show that using active context sampling, we can generalize (ContextualRAGE) to design a new active context sampling algorithm, which we call Active-ContextualRAGE. Finally, in Section G.2.3 we show that our family of SCLBs from Definition D.1 remains an instance where Active-ContextualRAGE outperforms ContextualRage by a $\sqrt{d}$ factor in its regret bound (which corresponds to a d factor in the PAC-learning sample complexity).

The main takeaway for the results in this section is (1) theoretically, we can prove that active context sampling improves over the rates of Li et al. [28], however, the algorithm achieving this improved rate is not computationally intractable; and (2) even if we consider computationally intractable algorithms as in [28], active context sampling improves over the rates achieved by passive context sampling on the family of SCLBs proposed in Section D.

G.2.1 Restating the main result of [28] for SCLBs

To aid in the statement of the main result of [28], we first introduce some additional notation. For any policy $\pi \in \Pi$, we denote $\phi_\pi := \mathbb{E}_{x\sim p}\, \phi(x, \pi(x))$.

Theorem G.3 (Theorem 2.14 of [28], restated)**.** *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB, $\|\theta^\star\| \leq 1$, and $\epsilon, \delta \in (0, 1)$. Define*

$$\mathcal{F} := \left\{ w \in \Delta^{\mathcal{X}\times\mathcal{A}} : \forall x \in \mathcal{X}, \sum_{a\in\mathcal{A}} w(x, a) = p(x) \right\}.$$

Let $\pi^\star : \mathcal{X} \to \mathcal{A}$ be the optimal policy given by

$$\pi^\star(x) := \operatorname*{argmax}_{a\in\mathcal{A}} \phi(x, a)^\top \theta^\star. \tag{8}$$

Define

$$\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon} := \min_{w\in\mathcal{F}} \max_{\pi:\pi\neq\pi^\star} \frac{\|\phi_\pi - \phi^\star_\pi\|^2_{\Sigma_w^{-1}}}{\max(\epsilon, \langle \phi_\pi - \phi_{\pi^\star}, \theta^\star\rangle)^2}.$$

Moroever, let

$$\Delta_\epsilon = \max\left(\epsilon, \min_{\pi\in\Pi:\pi\neq\pi^\star} \langle \phi_{\pi^\star} - \phi_\pi, \theta^\star\rangle\right). \tag{9}$$

Then, with probability at least $1-\delta$*, ContextualRAGE (Algorithm 1 of [28]) returns a policy* $\hat{\pi}$ *such that* $R(\hat{\pi}) \leq \epsilon$ *after making at most*

$$O\left(\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon} \log(\min\{d\log(1/\epsilon), \log|\Pi|\} + \log(1/\delta))\log(\Delta_\epsilon^{-1})\right) \tag{10}$$

reward observations. Moreover, it is always *the case that* (10) *is upper bounded by* $\tilde{O}(d^2/\epsilon^2)$.

This result gives a fine-grained PAC-learning guarantee, which may be much stronger than the minimax-rate of $\tilde{O}(d^2/\epsilon^2)$ when $\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon} \ll d^2/\epsilon^2$.

However, as discussed in Section 3.3 and Section 4 of Li et al. [28], the ContextualRAGE algorithm presented in [28] is computationally ineffcient because it requires maintaining a set of policies Π_ℓ from round to round. Because $|\Pi| = \mathcal{X}^{\mathcal{A}}$, even maintaining Π_1 at the start of the algorithm requires exponential-time. Thus, while Theorem G.3 is very interesting information theoretically, it does not directly lend itself well to practical applications of contextual bandits.

Nonetheless, in the following section, we show that if we disregard computational implementability, we can obtain a similar result to our Theorem G.3: in particular, we present a new active-learning variant of [28]'s ContextualRAGE that achieves an *even tighter* instance-dependent rate than Theorem G.3.

G.2.2 Developing tighter instance-dependent rates using active context sampling

Here, we propose an algorithm that takes advantage of active context sampling to achieve a better instance-dependent PAC learning guarantee than the passive instance-dependent sample complexity bound stated in [28]. The pseudocode is shown in Active-ContextualRAGE (Algorithm 2). As with the ContextualRAGE algorithm (Theorem G.3), Active-ContextualRAGE is computationally infeasible for the same reason that ContextualRAGE is infeasible: it requires explicitly maintaining a set of policies which could be as large as Π in every iteration.

Algorithm 2: Active-ContextualRAGE

Require: $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d$, $\delta \in (0,1)$

Initialize $\Pi_1 = \Pi$

for $\ell = 1, 2, \cdots, \lceil \log_2(1/\epsilon)\rceil$ **do**

  $\epsilon_\ell := 2^{-\ell}$, $\delta_\ell := \delta/(2\ell^2|\Pi|)$

  Let n_ℓ be the minimum value s.t.:

$$\min_{w\in\triangle_{\mathcal{X}\times\mathcal{A}}} \max_{\pi,\pi'\in\Pi_\ell} \frac{\|\mathbb{E}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))]\|^2_{\Sigma_w^{-1}} \log(1/\delta_\ell)}{n_\ell} \leq \epsilon_\ell^2$$

  with argmin given by $w^{(\ell)}$.

  For each $t \in [n_\ell]$, pull $(c_t, a_t) \sim w^{(\ell)}$, observe reward r_t

  Compute $O_t = \Sigma^{-1}_{w^{(\ell)}} \phi(c_t, a_t) r_t$

  For $\pi, \pi' \in \Pi_\ell$

$$\hat{\Delta}_\ell(\pi,\pi') = \mathsf{Cat}(\{\langle \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))], O_i\rangle\}_{i=1}^{n_\ell})$$

  Update

$$\Pi_{\ell+1} = \Pi_\ell \setminus \{\pi' \in \Pi_l \mid \max_{\pi\in\Pi_\ell} \hat{\Delta}_\ell(\pi,\pi') > \epsilon_\ell\}$$

end for

return $\Pi_{\ell+1}$

We first state a lemma that guarantees the optimal policy is inside the candidate policy set Π_ℓ after elimination, and all policies in Π_ℓ have small gaps. The following lemma is the same as Lemma 3.1 in [28] and the proof follows the identical argument as in [28].

Lemma G.4 (Lemma 3.1 of [28], restated)**.** *In the execution of Algorithm 2, with probability at least $1-\delta$, for all $\ell > 1$, $\pi_* \in \Pi_\ell$ and $\max_{\pi\in\Pi_\ell} V(\pi^*) - V(\pi) \le 4\epsilon_\ell$.*

Next, we state a theorem that gives a PAC upper bound for Algorithm 2 using active context sampling. Note that the upper bound $\tilde{O}(d^2/\epsilon^2)$ in this theorem is quite loose, and we show in Section G.2.3 that the active complexity bound can be much smaller than the passive one.

Theorem G.5 (Active-ContextualRAGE)**.** *Let $\epsilon, \delta \in (0,1)$, $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d$, $\|\theta^\star\| \le 1$, and let $\Delta_\epsilon, \pi^\star$ be as in* (8) *and* (9)*, respectively. Define*

$$\rho^{\mathrm{act}}_{\mathcal{B},\epsilon} := \min_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \max_{\pi:\pi\neq\pi^\star} \frac{\|\phi_\pi - \phi^\star_\pi\|^2_{\Sigma_w^{-1}}}{\max(\epsilon, \langle \phi_\pi - \phi_{\pi^\star}, \theta^\star\rangle)^2}.$$

Then, with probability at least $1-\delta$, Active-ContextualRAGE (Algorithm 2) returns a policy $\hat{\pi}$ such that $R(\hat{\pi}) \le \epsilon$ after making at most

$$O\left(\rho^{\mathrm{act}}_{\mathcal{B},\epsilon}(\min\{d\log(1/\epsilon), \log|\Pi|\} + \log(1/\delta))\log(\Delta_\epsilon^{-1})\right) \tag{11}$$

reward observations, which is itself upper bounded by $\tilde{O}(d^2/\epsilon^2)$.

Proof. For notational convenience, for any $\pi \in \Pi$, let

$$V(\pi) = \mathop{\mathbb{E}}_{x\sim p} \phi(x, \pi(x))^\top \theta^\star.$$

Define $S_\ell = \{\pi \in \Pi : V(\pi^*) - V(\pi) \le 4\epsilon_\ell\}$. Lemma G.4 implies that with probability at least $1-\delta$ we have $\bigcap_{\ell=1}^\infty \{\Pi_\ell \subseteq S_\ell\}$. Observe that if for any $\mathcal{V} \subset \Pi$ we define

$$\rho(w^{(\ell)}, \mathcal{V}) := \min_{w\in\Delta_{\mathcal{X}\times\mathcal{A}}} \max_{\pi,\pi'\in\Pi_\ell} \left\| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))] \right\|^2_{\Sigma_w^{-1}},$$

then

$$\begin{aligned}
\rho(w^{(\ell)}, \Pi_\ell) &= \min_{w\in\Delta_{\mathcal{X}\times\mathcal{A}}} \max_{\pi,\pi'\in\Pi_\ell} \| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))]\|^2_{\Sigma_w^{-1}} \\
&\le \min_{w\in\Delta_{\mathcal{X}\times\mathcal{A}}} \max_{\pi,\pi'\in S_\ell} \| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))]\|^2_{\Sigma_w^{-1}} =: \rho(S_\ell).
\end{aligned}$$

By line 4 of Algorithm 2, we know that for each ℓ, $n_\ell = \lceil \rho(w^{(\ell)}, \Pi_\ell)\log(1/\delta_\ell)\epsilon_\ell^{-2}\rceil$. Also, for $\ell \ge \lceil \log_2(4\Delta_\epsilon^{-1})\rceil$ we have for all $\pi \in S_\ell$, $V(\pi^*) - V(\pi) \le \epsilon$ (by Lemma G.4 and Definition of S_ℓ and ϵ_ℓ), thus the sample complexity to identify any π in S_ℓ is

$$\begin{aligned}
\sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} n_\ell &= \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \lceil 4\epsilon_\ell^{-2}\rho(w^{(\ell)}, \Pi_\ell)\log(2\ell^2|\Pi|/\delta)\rceil \\
&\le \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} 4\epsilon_\ell^{-2}\rho(S_\ell)\log(2\ell^2|\Pi|/\delta) + 1 \\
&\le c\log(\log(\Delta_\epsilon^{-1})|\Pi|/\delta) \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \epsilon_\ell^{-2}\rho(S_\ell)
\end{aligned}$$

for some absolute constant $c > 0$. We now note that

$$\begin{aligned}
&\min_{w\in\Delta_{\mathcal{X}\times\mathcal{A}}} \max_{\pi\in\Pi\setminus\pi^*} \frac{\|\mathbb{E}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi^*(x))]\|^2_{\Sigma_w^{-1}}}{\langle \mathbb{E}_{x\sim p}[\phi(x,\pi^*(x)) - \phi(x,\pi(x))], \theta^*\rangle^2 \vee \epsilon^2} \\
&= \min_{w\in\Delta_{\mathcal{X}\times\mathcal{A}}} \max_{\ell\le\lceil\log_2(4\Delta_\epsilon^{-1})\rceil} \max_{\pi\in S_\ell} \frac{\|\mathbb{E}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi^*(x))]\|^2_{\Sigma_w^{-1}}}{\langle \mathbb{E}_{x\sim p}[\phi(x,\pi^*(x)) - \phi(x,\pi(x))], \theta^*\rangle^2 \vee \epsilon^2}
\end{aligned}$$

$$\geq \frac{1}{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \min_{w \in \triangle_{\mathcal{X}\times\mathcal{A}}} \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \max_{\pi \in S_\ell} \frac{\|\mathbb{E}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi^*(x))]\|^2_{\Sigma_w^{-1}}}{\langle \mathbb{E}_{x\sim p}[\phi(x,\pi^*(x)) - \phi(x,\pi(x))], \theta^*\rangle^2 \vee \epsilon^2}$$

(max lower bounded by average)

$$\geq \frac{1}{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} (4\epsilon_\ell)^{-2} \min_{w \in \triangle_{\mathcal{X}\times\mathcal{A}}} \max_{\pi \in S_\ell} \| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi^*(x))]\|^2_{\Sigma_w^{-1}}$$

($\pi \in S_\ell$ implies that gap less than ϵ_ℓ and $4\epsilon_\ell \geq \epsilon$)

$$\geq \frac{1}{64\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \epsilon_\ell^{-2} \min_{w \in \triangle_{\mathcal{X}\times\mathcal{A}}} \max_{\pi,\pi' \in S_\ell} \| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))]\|^2_{\Sigma_w^{-1}}$$

$$= \frac{1}{64\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} \epsilon_\ell^{-2} \rho(S_\ell)$$

where for the last inequality, we have used the fact that for any $\pi, \pi' \in S_\ell$,

$$\begin{aligned}
&\| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi'(x))]\|^2_{\Sigma_w^{-1}} = \|\phi_\pi - \phi_{\pi'}\|^2_{\Sigma_w^{-1}} \\
&= (\phi_\pi - \phi_{\pi'})^\top \Sigma_w^{-1} (\phi_\pi - \phi_{\pi'}) \\
&= (\phi_\pi - \phi_{\pi^*} + \phi_{\pi^*} - \phi_{\pi'})^\top \Sigma_w^{-1} (\phi_\pi - \phi_{\pi^*} + \phi_{\pi^*} - \phi_{\pi'}) \\
&= (\phi_\pi - \phi_{\pi^*})^\top \Sigma_w^{-1} (\phi_\pi - \phi_{\pi^*}) + (\phi_{\pi^*} - \phi_{\pi'})^\top \Sigma_w^{-1} (\phi_{\pi^*} - \phi_{\pi'}) + 2(\phi_\pi - \phi_{\pi^*})^\top \Sigma_w^{-1} (\phi_{\pi^*} - \phi_{\pi'}) \\
&\leq 2 \max_{\pi'' \in S_\ell} \|\phi_{\pi''} - \phi_{\pi^*}\|^2_{\Sigma_w^{-1}} + 2(\phi_\pi - \phi_{\pi^*})^\top \Sigma_w^{-1/2} \Sigma_w^{-1/2} (\phi_{\pi^*} - \phi_{\pi'}) \\
&\leq 2 \max_{\pi'' \in S_\ell} \|\phi_{\pi''} - \phi_{\pi^*}\|^2_{\Sigma_w^{-1}} + 2\|\phi_\pi - \phi_{\pi^*}\|_{\Sigma_w^{-1}} \|\phi_{\pi^*} - \phi_{\pi'}\|_{\Sigma_w^{-1}} \qquad \text{(Cauchy-Schwarz)} \\
&\leq 4 \max_{\pi \in S_\ell} \| \mathop{\mathbb{E}}_{x\sim p}[\phi(x,\pi(x)) - \phi(x,\pi^*(x))]\|^2_{\Sigma_w^{-1}}.
\end{aligned}$$

Therefore,

$$\sum_{\ell=1}^{\lceil \log_2(4\Delta_\epsilon^{-1})\rceil} n_\ell \leq 64c \log(\log(\Delta_\epsilon^{-1})|\Pi|/\delta) \lceil \log_2(4\Delta_\epsilon^{-1})\rceil \rho^{\mathrm{act}}_{\mathcal{B},\epsilon}. \tag{12}$$

We now use a discretization argument to show the bound in (10). Define ϵ-ball $\mathcal{T}_\epsilon := \{\pi : \forall \pi, \pi', \langle \phi_\pi - \phi_{\pi'}, \theta^*\rangle \leq \epsilon\}$ and let $\mathcal{T}$ be a cover for Π using those balls, i.e. $\mathcal{T} := \{\mathcal{T}_{\epsilon,i}\}_{i=1}^{|\mathcal{T}|}$ Since $\theta^* \in \mathbb{R}^d$, the covering number $|\mathcal{T}| \leq O((1/\epsilon)^d)$. Let $\Pi_\mathcal{T} := \{\pi_i : \pi_i \in \mathcal{T}_{\epsilon,i}\}_{i=1}^{|\mathcal{T}|}$ be a collection of policies where we take one policy from each ϵ-ball in the cover $\mathcal{T}$. Then $|\Pi_\mathcal{T}| = |\mathcal{T}| \leq O((1/\epsilon)^d)$. The exact argument holds for identifying the optimal policy in $\Pi_\mathcal{T}$, i.e. it takes at most

$$64c \log(\log(\Delta_\epsilon^{-1})|\Pi_\mathcal{T}|/\delta) \lceil \log_2(4\Delta_\epsilon^{-1})\rceil \rho^{\mathrm{act}}_{\mathcal{B},\epsilon} = O((d\log(1/\epsilon) + \log(1/\delta)) \log_2(\Delta_\epsilon^{-1}) \rho^{\mathrm{act}}_{\mathcal{B},\epsilon}) \tag{13}$$

samples to identify $\pi^*_\mathcal{T} \in \Pi_\mathcal{T}$. However, note that the global optimal policy π^* must lie in some ball $\mathcal{T}_{\epsilon,i_0}$, so $V(\pi^*) - V(\pi^*_\mathcal{T}) \leq V(\pi^*) - V(\pi_{i_0}) \leq \epsilon$, so $\pi^*_\mathcal{T}$ is an ϵ-optimal policy. The first part of the statement follows by taking the minimum of (12) and (13). As for the inequality bounding the sample complexity by $\tilde{O}(d^2/\epsilon^2)$, note that $\rho^{\mathrm{act}}_{\mathcal{B},\epsilon}$ is upper bounded by $\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon}$, so this upper bound follows directly from the second part of Theorem 2.14 of [28]. □

Theorem G.5 improves over Theorem G.3 in the following sense. Note that because the minimization in $\varrho_{\mathcal{B},\epsilon}$ is over a *strictly larger* distribution class ($\Delta^{\mathcal{X}\times\mathcal{A}}$)than the minimization in $\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon}$ (which is only minimized over $\mathcal{F}$), we have that

$$\rho^{\mathrm{act}}_{\mathcal{B},\epsilon} \leq \rho^{\mathrm{pas}}_{\mathcal{B},\epsilon},$$

In the following section, we demonstrate that that on the hard instance for passive context sampling described in Section D (Definition D.1), we have that $\rho^{\mathrm{act}}_{\mathcal{B},\epsilon}$ is indeed a $\Theta(d)$-factor smaller than $\rho^{\mathrm{pas}}_{\mathcal{B},\epsilon}$.

G.2.3 Demonstrating the power of active context sampling

The main result of this section is the following.

Lemma G.6. *Let $A, d \in \mathbb{Z}_{>1}$. Let $\mathcal{B}^\star_{d,A}$ be the SCLB instance from Definition D.1. Then, for any $\epsilon \in (0,1)$ we have that*

$$\rho^{\mathrm{act}}_{\mathcal{B}^\star_{d,A},\epsilon} \leq \frac{8}{d} \cdot \rho^{\mathrm{pas}}_{\mathcal{B}^\star_{d,A},\epsilon}.$$

Proof. For notational convenience, we let $\mathcal{B} = \mathcal{B}^\star_{d,A}$ inside this proof. The optimal policy $\pi^\star$ is given by $\pi^\star(x) = 1, \forall x \in \mathcal{X}$. From the definitions of $\rho_{\mathcal{B},\epsilon}$, $\varrho_{\mathcal{B},\epsilon}$, and $\mathcal{B}$, it is enough to show that

$$\min_{w \in \Delta^{\mathcal{X}\times\mathcal{A}}} \max_{\pi:\pi\neq\pi^\star} \|\phi_\pi - \phi_{\pi^\star}\|_{\Sigma_w^{-1}}{}^2 \leq \frac{8}{d} \cdot \min_{w\in\mathcal{F}} \max_{\pi:\pi\neq\pi^\star} \|\phi_\pi - \phi^\star_\pi\|_{\Sigma_w^{-1}}{}^2. \tag{14}$$

Note that due to the structure of the SCLB—wherein all actions besides $a = 1$ reveal the same feature e_1—one maximizing policy in the inner maximization will be given by $\pi(x) = 2, \forall x \in \mathcal{A}$. Thus, without loss of generality, we can fix

$$\phi_\pi - \phi_{\pi^\star} = e_1 - \left(\left(1 - \frac{d-1}{d^2}\right) e_1 + \frac{1}{d^2}\sum_{i=2}^d e_i\right) = \frac{d-1}{d^2} e_1 - \sum_{i=2}^d \frac{1}{d^2} e_i. \tag{15}$$

and show that

$$\min_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \|\phi_\pi - \phi^\star_\pi\|_{\Sigma_w^{-1}}{}^2 \leq \frac{8}{d} \cdot \min_{w\in\mathcal{F}} \|\phi_\pi - \phi^\star_\pi\|_{\Sigma_w^{-1}}{}^2. \tag{16}$$

The remainder of the proof is devoted to proving (16).

Now, consider the right-hand-side of (16). We need to reason about the optimal choice of w. Note that all actions for $a \neq 1$ reveal the same feature vector—which corresponds to 0 reward; so, without loss of generality we can restrict the minimizing sampling distribution's support to $a = 1$. Moreover, because $w \in \mathcal{F}$ constrains the marginal of w with respect to x to be equal to p, this indicates that one minimizing sampling distribution for the right-hand-side of (16) is given by

$$w(x,a) = \begin{cases} 1 - (d-1)/d^2, & x = 1, a = 1 \\ 1/d^2 & x \neq 1, a = 1 \\ 0 & \text{otherwise} \end{cases}.$$

In this case, Σ_w^{-1} is a diagonal matrix with $\frac{d^2-d+1}{d^2}$ in the first entry, and d^2 on the remainder of the diagonal. Hence, we have

$$\begin{aligned} \min_{w\in\mathcal{F}} \|\phi_\pi - \phi^\star_\pi\|^2_{\Sigma_w^{-1}} &= \frac{d^2-d+1}{d^2} \cdot \frac{(d-1)^2}{d^4} + d^2 \sum_{i=2}^d \frac{1}{d^4} \\ &= \frac{d^2-d+1}{d^2} \cdot \frac{(d-1)^2}{d^4} + \frac{d-1}{d^2} \\ &\geq \frac{1}{2d}, \end{aligned}$$

where the last inequality uses that $d > 1$.

Next, we turn to the left-hand-side of (16). Consider the (active) sampling distribution

$$w'(x,a) = \begin{cases} \frac{1}{2}, & x = 1, a = 1 \\ \frac{1}{2(d-1)} & x \neq 1, a = 1 \\ 0 & \text{otherwise} \end{cases}.$$

In this case, $\Sigma_{w'}^{-1}$ is a diagonal matrix with 2 in the first entry, and $2(d-1)$ on the remainder of the diagonal. This distribution is actively sampling contexts in the sense that its marginal with respect to x is *not* p. A similar calculation as above shows that

$$\min_{w\in\Delta^{\mathcal{X}\times\mathcal{A}}} \|\phi_\pi - \phi^\star_\pi\|^2_{\Sigma_w^{-1}} \leq \|\phi_\pi - \phi^\star_\pi\|^2_{\Sigma_{w'}^{-1}} = 2\frac{(d-1)^2}{d^4} + 2(d-1)\sum_{i=2}^d \frac{1}{d^4}$$

$$= 2\frac{(d-1)^2}{d^4} + \frac{2(d-1)^2}{d^4} \leq \frac{4}{d^2}.$$

Thus, we conclude that

$$\min_{w \in \Delta^{\mathcal{X}\times\mathcal{A}}} \|\phi_\pi - \phi_\pi^\star\|_{\Sigma_w^{-1}}{}^2 \leq \frac{4}{d} = 8 \cdot \frac{1}{2d} \leq 8 \min_{w \in \mathcal{F}} \|\phi_\pi - \phi_\pi^\star\|_{\Sigma_w^{-1}}{}^2,$$

as desired. □

This result shows that from the PAC-learning perspective, Active-ContextualRAGE (Algorithm 2) once again improves the dimension dependence from [28] on an instance-dependent basis. However, as neither is known to be implementable in polynomial time, it remains a future work to explore possible implementations or heuristics to approximately simulate these algorithms. We hope our work provides insight on how to incorporate active context sampling, if a polynomial-time implementation or approximation of Contextual-RAGE is eventually developed in future research.

H Additional experimental details

This appendix contains additional experimental details to aid in reproducing our empirical results. As an overview, we use each method (Active-SCLB and baselines) and collect a dataset of T samples and train a policy using ridge regression (recall Lemma D.2) and examine how regret (1) decays with T. All experiments were performed on a CPU machine with 12 cores and 36 GB RAM. Additional details explaining these choice of benchmarks are as follows.

H.1 Overview of passive context sampling baselines

Here, we briefly describe the two baselines against which we compare our empirical results.

Reward-free LinUCB (RFLinUCB). RFLinUCB is a slight modification (proposed in [46]) of the LinUCB algorithms (proposed by [1]) in order to adapt it for the pure exploration setting. The RFLinUCB algorithm implements the standard LinUCB algorithm of Abbasi-Yadkori et al. [1] with the reward function set to 0. The original LinUCB algorithm contains the reward function in its exploration policy because it is designed to minimize the online regret (rather than the simple regret, which is of interest in the pure exploration setting), and this is an unncessary in the pure exploration setting [46].

Planner-Sampler Zanette et al. [46] point out that one limitation of the RFLinUCB algorithm is that the algorithm is *adaptive* to the observed context-action pairs, in the sense that the t-th observed context depends on the history $\{x_1, a_1\}, ..., \{x_t, a_t\}$. This can be challenging to implement in certain real-world scenarios, since it requires *sequentially* querying context-action pairs. To address this, the planner-sampler algorithm of Zanette et al. [46] is designed to match the regret bound of RFLinUCB with a *fixed, non-adaptive* policy. The Planner-Sampler algorithm proceeds in two stages.

First, a "Planner" algorithm observes the features (but not the rewards) of T_0 contexts $\{\phi(x_t, a_t)\}_{t\in[T_0],a\in\mathcal{A}}$ for an offline set of contexts drawn independently $x_t \sim p$. The Planner uses these observed features to design a stochastic exploration policy $\pi' : \mathcal{X} \to \mathcal{A}$.

In the second stage, a "Sampler" algorithm samples the rewards of T context-action pairs drawn according to the Planner's policy π'. That is, the Sampler observes $r(x'_1, a'_1), ..., r(x'_T, a'_T)$ where each $x'_t \sim p$ and $a'_t \sim \pi'(x'_t)$ independently. Zanette et al. [46] study various tradeoffs of T_0 and T.

In our experiments, we generously set $T_0 = |\mathcal{X}|$ in order to ensure fair comparison with RF-LinUCB and our active context sampling approach, which use knowledge of the full feature mapping. We then measure the regret as a function of the number of reward observations, T required by the "Sampler".

Active-SCLB and Passive-SCLB. Recall that in our experiments, we run Algorithm 1 with $\alpha \leftarrow 0$ and solve the SDP in Line 2 of Algorithm 1 to convergence. In this setting, there is a natural passive context sampling analogue of our Algorithm: instead of (approximately) solving

$$\hat{w} \approx \operatorname*{argmin}_{w\in\Delta^{\mathcal{S}\times\mathcal{A}}} \|\phi(x,a)\|^2_{\Sigma^{-1}_{\hat{w}}},$$

we can solve the same problem with the additional constraint that the marginal distribution of w with respect to x must match p. That is, we can define

$$\mathcal{F} = \left\{ w \in \Delta^{\mathcal{X}\times\mathcal{A}} : \forall x \in \mathcal{X}, \sum_{a\in\mathcal{A}} w(x,a) = p(x) \right\}$$

and solve

$$\hat{w} \approx \operatorname*{argmin}_{w\in\mathcal{F}} \|\phi(x,a)\|^2_{\Sigma^{-1}_{\hat{w}}}.$$

Note that restricting the feasible space to $\mathcal{F}$ only requires some additional linear constraints, and consequently it is easy to see that the problem can still be formulated as an SDP. For completeness, we include a full pseudocode of the versions used in our numerical experiments in Algorithm 3 and Algorithm 4.

Algorithm 3: Active-SCLB

Input: SCLB $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$, regularization parameter $\lambda > 0$, sample budget $T \in Z_{>0}$.
Output: A policy $\hat{\pi} : \mathcal{X} \to \mathcal{A}$

// Approximately solve the following using an SDP solver.

Compute $\hat{w} = \operatorname{argmin}_{w \in \Delta^{\mathcal{X} \times \mathcal{A}}} \mathbb{E}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_w^{-1}}$

// Draw samples and apply ridge regression to compute a policy $\hat{\pi}$. Recall in synthetic experiments, we sampled with replacement, while in experiments on real-world datasets, we sampled without replacement using rejection sampling.

$\mathcal{S} \leftarrow \{(x_1, a_1), ..., (x_T, a_T)\}$ where each $(x_t, a_t) \sim \hat{w}$ i.i.d.

$\hat{\theta} \leftarrow \Sigma_{\mathcal{S}}^{-1} \sum_{(x_t, a_t) \in \mathcal{S}} \phi(x_t, a_t) r_t$

return: $\hat{\pi}(x) \leftarrow x \mapsto \operatorname{argmax}_{a \in \mathcal{A}} \phi(x, a)^\top \hat{\theta}$

Algorithm 4: Passive-SCLB

Input: SCLB $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$, regularization parameter $\lambda > 0$, sample budget $T \in Z_{>0}$.
Output: A policy $\hat{\pi} : \mathcal{X} \to \mathcal{A}$

// Approximately solve the following using an SDP solver.

Compute $\hat{w} = \operatorname{argmin}_{w \in \mathcal{F}} \mathbb{E}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_w^{-1}}$

// Draw samples and apply ridge regression to compute a policy $\hat{\pi}$. Recall in synthetic experiments, we sampled with replacement, while in experiments on real-world datasets, we sampled without replacement using rejection sampling.

$\mathcal{S} \leftarrow \{(x_1, a_1), ..., (x_T, a_T)\}$ where each $(x_t, a_t) \sim \hat{w}$ i.i.d.

$\hat{\theta} \leftarrow \Sigma_{\mathcal{S}}^{-1} \sum_{(x_t, a_t) \in \mathcal{S}} \phi(x_t, a_t) r_t$

return: $\hat{\pi}(x) \leftarrow x \mapsto \operatorname{argmax}_{a \in \mathcal{A}} \phi(x, a)^\top \hat{\theta}$

H.2 Implementation details for SDP solving

In all of our experiments, we used CVXPY (an open source Python-embedded modeling language for convex optimization problems) to model the SDP variables, objectives, and constraints. Within CVXPY, we used the Mosek (Version 10) SDP solver for solving the SDPs (described further in the following paragraph.) All solver hyperparameters were fixed to their default values in CVXPY.

MOSEK is a software package for solving structured optimizations such as SDPs. Although it is not open-sourced, MOSEK is available to academic users for free Additionally, for non-academic users, MOSEK offers a free trial period.

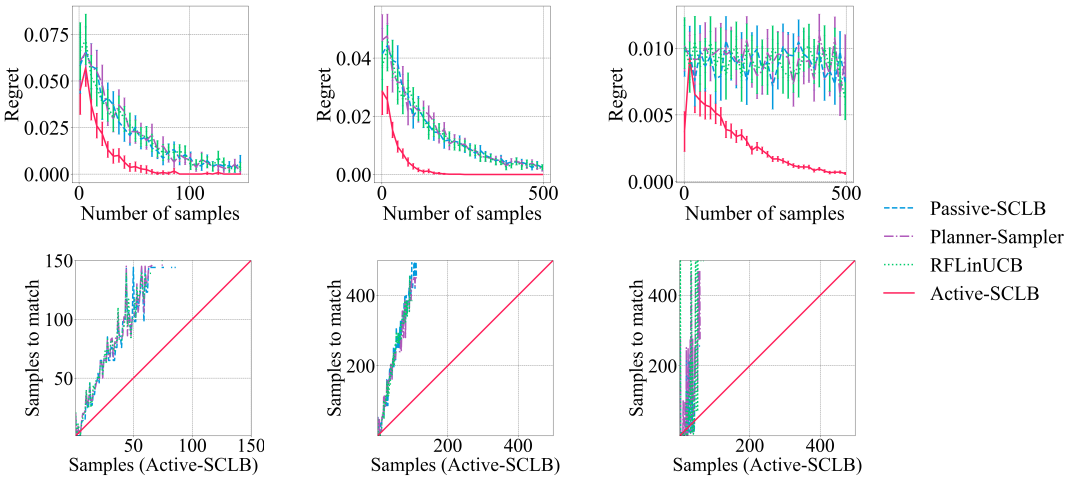
To ensure numerical stability, when applying the SDP solver, we (1) included a small amount of numerical regularization for the SDP constraints, on the order of 1e-6, to avoid numerical stability errors and (2) normalized all features such that $\|\phi(x, a)\|_2 \leq 1$. For consistency, the feature normalization was applied to *all* baselines as well as Active-SCLB in our experiments.

H.3 Synthetic experimental results

We experiment on the SCLB instance in Definition D.1. We fix $\lambda = 1\text{e}{-6}$ and vary $d \in [5, 10, 50]$. Figure 5 shows our results. The top row shows the difference in the rate of regret decay between our method (Active-SCLB) and others is more pronounced as d grows. This is consistent with the $O(\sqrt{d})$ theoretical gap between the regret bound of Active-SCLB and other methods on $\mathcal{C}_{\mathcal{B}^\star_{A,d}}$, as dicussed in Section D. The bottom row of the figure shows that other baselines usually require far more data to match the performance of Active-SCLB at a given sample budget.

H.4 Experiments with different regularization amounts.

In this section, we include results on our real-world datasets for differing values of the regularizer λ. The choice of λ is typically a design choice, which practitioners use to balance bias-variance



(a) $d = 5$ (b) $d = 10$ (c) $d = 50$

Figure 5: Experiments on $\mathcal{B}^{\star}_{d,10}$ (Definition D.1). Top: Regret vs. number of samples (mean $\pm$ 2 standard errors over 100 trials). Bottom: Minimum number of samples required for baselines' mean regret (over 100 trials) to match Active-SCLB's mean regret for a given sample budget.

trade-offs (see also, the discussion in [46]). Smaller values of λ reduce bias at the risk of increased variance; larger values of λ increase bias (often at the benefit of reduced variance).

Figures 6 and 7 show that our improvements remain largely consistent as we vary $\lambda \in \{\text{1e-6}, \text{1e-4}, \text{1e-2}\}$. These results were omitted in the main body for brevity, because the results are largely consistent across different values of λ.

I Omitted proofs

This section contains proofs which are omitted in the main body. Recall that when applied to matrices, $\|\cdot\|$ always denotes the spectral norm. When applied to vectors, $\|\cdot\|$ denotes the Euclidean norm.

Organization of proofs. In Section I.3 we collect some established results in SCLBs and PSD matrix theory, which we use in our analysis. In Section I.4 we prove Theorem 2.2. In Section I.5 we include omitted proofs from Section D.

Notation. We use $[n]$ to denote the set $\{1, ..., n\}$ and $\|\cdot\|$ to denote the Euclidean norm (when applied to vectors) and the spectral norm (when applied to matrices). We use $[v]_i$ for the i-the entry of $v \in \mathbb{R}^d$. We use $\succeq, \succ$ for the Loewner ordering: matrix A satisifes $A \succeq 0$ ($A \succ 0$) if and only if A is positive semi-definite (positive definite), respectively (see also, Lemma I.2.) For any $A \in \mathbb{R}^{d\times d}$ with $A \succ 0$ and $x \in \mathbb{R}^d$, we denote $\|x\|_{A^{-1}} := (x^\top A^{-1} x)^{1/2}$. For any set $\mathcal{S} \subset \mathcal{X} \times \mathcal{A}$, we define the covariance matrix $\Sigma_{\mathcal{S}} := \lambda I + \sum_{(x,a)\in\mathcal{S}} \phi(x,a)\phi(x,a)^\top$. For any $w \in \Delta^{\mathcal{X}\times\mathcal{A}}$, let $\Sigma_w := \lambda/T \cdot I + \mathbb{E}_{(x,a)\sim w} \phi(x,a)^\top \phi(x,a)$ denote the w-weighted covariance matrix of the feature mapping ϕ. Correspondingly, we also denote the T-sample *empirical* covariance matrix as follows

$$\hat{\Sigma}_{w,T} := \lambda/T \cdot I + 1/T \sum_{t\in[T]} \phi(x_t, a_t)\phi(x_t, a_t)^\top \text{ where each } (x_t, a_t) \sim w \text{ independently.}$$

I.1 Ridge regression regret bound.

We first restate a standard result from the SCLB literature. The next lemma bounds the regret of a learned policy obtained by fitting a ridge regression model to dataset $\mathcal{S}$ [26] and is the basis for prior works which obtain polynomial-time algorithms with minimax-optimal regret for SCLBs [1, 46].

Lemma I.1 (Ridge regression regret bound)**.** *Let $\lambda > 0$ be any regularization parameter, $\mathcal{S} = \{(x_1, a_1), ..., (x_T, a_T)\} \subset \mathcal{X} \times \mathcal{A}$, and $\delta \in (0, 1)$ be a failure probability. For each $t \in [T]$, let*

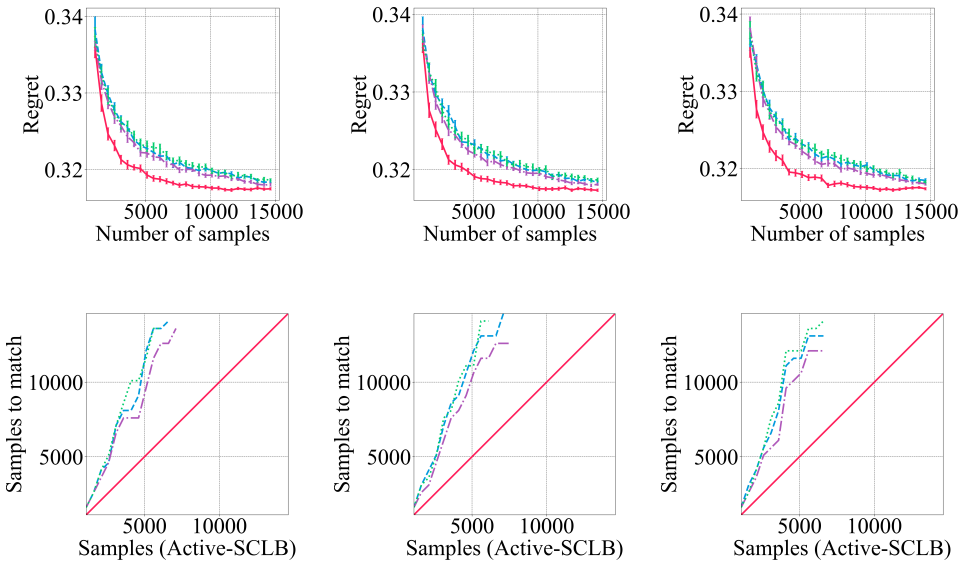


Figure 6: Warfarin dataset. *Top:* Regret vs. number of samples (mean $\pm$ 2 standard errors over 100 trials). *Bottom:* Minimum number of samples required for baselines' mean regret (over 100 trials) to match Active-SCLB's mean regret for a given sample budget.

$r_t \sim \phi(x_t, a_t)^\top \theta^\star + \eta_t$ *where each* $\eta_t \sim \nu$*, independently. Let* $\hat{\theta}$ *solve the* ridge regression *problem, i.e.,* $\hat{\theta} := \Sigma_{\mathcal{S}}^{-1} \sum_{t \in [T]} \phi(x_t, a_t) r_t$ *and* $\hat{\pi} := x \mapsto \operatorname{argmax}_{a \in \mathcal{A}} \phi(x, a)^\top \hat{\theta}$*. Moreover, define*

$$\sqrt{\beta} := 2 \cdot \min\Big(\underbrace{\sqrt{d \log\left(2(1 + TL^2/\lambda)/\delta\right)} + \sqrt{\lambda}\|\theta^\star\|_2}_{\text{tighter when } |\mathcal{X} \times \mathcal{A}| \text{ large}}, \underbrace{2\sqrt{2}\sqrt{\log\left(12T^2|\mathcal{X}||\mathcal{A}|/(\pi^2\delta)\right)}}_{\text{tighter when } |\mathcal{X} \times \mathcal{A}| \text{ small}} \Big). \tag{17}$$

Define the uncertainty measure

$$\Gamma(\mathcal{S}) := \mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}}. \tag{18}$$

Then, with probability $1 - \delta/2$*,* $R(\hat{\pi}) \le \sqrt{\beta \Gamma(\mathcal{S})}$.

I.2 Omitted proof sketch for Theorem 4.3.

Proof sketch. In light of the regret bound (3) in Lemma I.1, to prove Theorem 2.2, it suffices to show that for $\mathcal{S}$ as constructed in Line 3, the following holds with high probability:

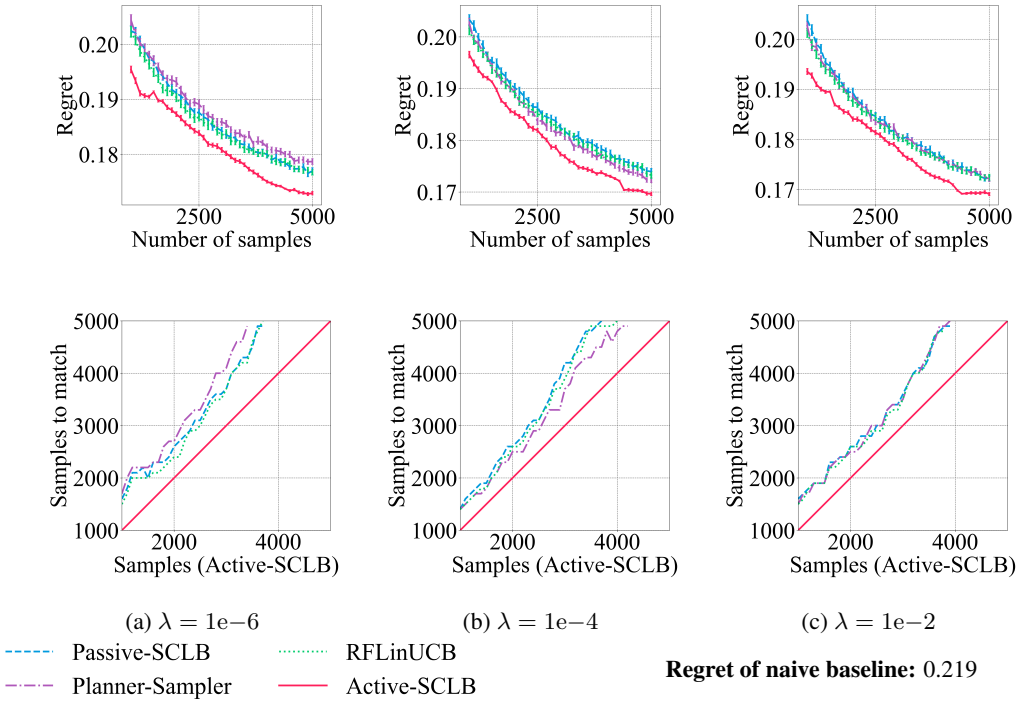
$$\Gamma(\mathcal{S}) = \mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} = 8/T\mathcal{C}_{\mathcal{B}}.$$

We prove this bound in three stages. First, by construction of $\mathcal{S}$ in Line 3, $\Sigma_{\mathcal{S}} = T\hat{\Sigma}_{w,T}$. Inverting both sides, we have $T\Sigma_{\mathcal{S}}^{-1} = \hat{\Sigma}_{w,T}^{-1}$. Second, we use matrix concentration guarantees [37] to show that for T sufficiently large, with high probability,

$$T\Sigma_{\mathcal{S}}^{-1} = \hat{\Sigma}_{w,T}^{-1} \preceq 1/\alpha \cdot \Sigma_w^{-1}. \tag{19}$$

This concentration argument curcially uses that q is constructed to be well-conditioned to the features, in that $\max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_q^{-1}} \le 2d$ and that $\Sigma_w \succeq \alpha \Sigma_q$ (since w dominates αq.) Third, setting $\alpha \leftarrow 1/2$ in the above display, we have that for any $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$T\|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} = T \cdot \phi(x, a)^\top (\Sigma_{\mathcal{S}})^{-1} \phi(x, a) \le 2 \cdot \phi(x, a)^\top \Sigma_w^{-1} \phi(x, a) = 2 \cdot \|\phi(x, a)\|^2_{\Sigma_w^{-1}}.$$



(a) $\lambda = 1\text{e}{-6}$ (b) $\lambda = 1\text{e}{-4}$ (c) $\lambda = 1\text{e}{-2}$

Figure 7: Jester dataset. *Top:* Regret vs. number of samples (mean $\pm$ 2 standard errors over 100 trials). *Bottom:* Minimum number of samples required for baselines' mean regret (over 100 trials) to match Active-SCLB's mean regret for a given sample budget.

Applying expectation and max to the above display and then invoking Lemma C.2, we conclude

$$T \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} \le 2 \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \le 8\mathcal{C}_{\mathcal{B}}.$$

Thus, $\Gamma(\mathcal{S}) \le 8\mathcal{C}_{\mathcal{B}}/T$ as desired, and Lemma I.1 yields the regret bound $R(\hat{\pi}) \le \tilde{O}(\sqrt{\beta\mathcal{C}_{\mathcal{B}}/T})$. Finally, to show that $\mathcal{C}_{\mathcal{B}} \le d$, we let $q^\star$ be the G-optimal design as in Theorem C.1. Then, note that

$$\mathcal{C}_{\mathcal{B}} \le \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{q^\star}^{-1}} \le \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{q^\star}^{-1}} \le d. \quad \square$$

I.3 Omitted proofs of standard results in SCLBs and PSD matrix theory.

We first prove some standard facts about the Lowener order. We believe the facts in the following Lemma I.2 are well-known; however, we collect them here for the sake of completeness.

Lemma I.2 (Lowener order facts)**.** *The following facts hold whenever $A, B, C \in \mathbb{R}^{d\times d}$ are positive semidefinite.*

(i) $A \succeq B$ *if and only if for all* $x \in \mathbb{R}^d$, $x^\top Ax \ge x^\top Bx$.

(ii) If $A \succ 0$, $A^{-1} \succ 0$.

(iii) If $A \succ (\succeq) B$, *then for any symmetric* $F \in \mathbb{R}^{d\times d}$, *we have* $FAF \succ (\succeq) FBF$.

(iv) If $A \succeq B \succ 0$, *then* $A^{-1} \preceq B^{-1}$.

(v) If $A = B + C$ *then* $A \succeq B$.

(vi) If $\|B^{-1/2}(A-B)B^{-1/2}\| \le \epsilon$ *for some* $\epsilon \in (0,1)$, *then* $(1-\epsilon)B \preceq A \preceq (1+\epsilon)B$.

Proof. We prove the facts one-by-one.

(i) If $A \succeq B$, then $A - B \succeq 0$. For any $x \in \mathbb{R}^d$, we have

$$x^\top (A - B)x \geq 0 \quad \text{implies} \quad x^\top Ax \geq x^\top Bx.$$

Conversely, if $x^\top Ax \geq x^\top Bx$ for all $x \in \mathbb{R}^d$, then for any $x \in \mathbb{R}^d$

$$x^\top (A - B)x = x^\top Ax - x^\top Bx \geq 0,$$

and hence $A - B \succeq 0$—thus, we have $A \succeq B$.

(ii) Since $A \succ 0$ (i.e., A is positive definite), it is invertible and its inverse is also positive definite. This follows because for any $x \neq 0$,

$$x^\top A^{-1}x = (A^{-1}x)^\top A(A^{-1}x) > 0.$$

Thus, $A^{-1} \succ 0$.

(iii) Suppose $A \succeq B$. Then for any $x \in \mathbb{R}^d$,

$$x^\top Ax \geq x^\top Bx.$$

Now, for any $y \in \mathbb{R}^d$ let $x = Fy$. We have that

$$x^\top Ax \geq x^\top Bx,$$

and consequently using that F is symmetric, we have

$$y^\top FAFy \geq y^\top FBFy,$$

The proof if $A \succ B$ is identical, replacing the $\geq$ with $>$.

(iv) Suppose $A \succeq B \succ 0$. Since $B \succ 0$, B is invertible. For any nonzero vector $x \in \mathbb{R}^d$, define $y = B^{-1/2}x$. Then, note that

$$x^\top A^{-1}x \leq x^\top B^{-1}x \quad \text{if and only if} \quad y^\top B^{1/2}A^{-1}B^{1/2}y \leq \|y\|^2.$$

Define the matrix $M = B^{1/2}A^{-1}B^{1/2}$. We aim to show that $M \preceq I$, which, by the above argument, implies $A^{-1} \preceq B^{-1}$.

Since $A \succeq B \succ 0$, then $B^{-1/2}AB^{-1/2} \succeq I$, by (iii). This ensures that the smallest eigenvalue of $B^{-1/2}AB^{-1/2} \succeq I$ is at least 1. Consequently, $(B^{-1/2}AB^{-1/2})^{-1} \preceq I$. Now,

$$(B^{-1/2}AB^{-1/2})^{-1} \preceq I \quad \text{implies} \quad B^{1/2}A^{-1}B^{1/2} \preceq I.$$

Therefore, by (i), we have that for all $x \in \mathbb{R}^d$,

$$x^\top A^{-1}x = y^\top My \leq y^\top y = x^\top B^{-1}x,$$

and hence $A^{-1} \preceq B^{-1}$, as desired.

(v) If $A = B + C$ and $C \succeq 0$, then $A - B = C \succeq 0$, so $A \succeq B$.

(vi) If $\|B^{-1/2}(A - B)B^{-1/2}\| \leq \epsilon$, then for any $x \in \mathbb{R}^d$,

$$-\epsilon \cdot x^\top x \leq x^\top B^{-1/2}(A - B)B^{-1/2}x \leq \epsilon \cdot x^\top x$$

Consequently, by (i), we have that

$$-\epsilon I \preceq B^{-1/2}(A - B)B^{-1/2} \preceq \epsilon I$$

Now, applying (iii) we have

$$-\epsilon B \preceq (A - B) \preceq \epsilon B$$

Rearranging, we obtain

$$(1 - \epsilon)B \preceq A \preceq (1 + \epsilon)B.$$

□

Next, we provide a proof of Lemma I.1. The proof follows the exposition in Section 3 of [46] closely.

Lemma I.1 (Ridge regression regret bound)**.** *Let $\lambda > 0$ be any regularization parameter, $\mathcal{S} = \{(x_1, a_1), ..., (x_T, a_T)\} \subset \mathcal{X} \times \mathcal{A}$, and $\delta \in (0, 1)$ be a failure probability. For each $t \in [T]$, let $r_t \sim \phi(x_t, a_t)^\top \theta^\star + \eta_t$ where each $\eta_t \sim \nu$, independently. Let $\hat{\theta}$ solve the* ridge regression *problem, i.e., $\hat{\theta} := \Sigma_{\mathcal{S}}^{-1} \sum_{t \in [T]} \phi(x_t, a_t) r_t$ and $\hat{\pi} := x \mapsto \operatorname{argmax}_{a \in \mathcal{A}} \phi(x, a)^\top \hat{\theta}$. Moreover, define*

$$\sqrt{\beta} := 2 \cdot \min\Big(\underbrace{\sqrt{d \log\left(2(1 + TL^2/\lambda)/\delta\right)} + \sqrt{\lambda}\|\theta^\star\|_2}_{\text{tighter when } |\mathcal{X} \times \mathcal{A}| \text{ large}}, \underbrace{2\sqrt{2}\sqrt{\log\left(12T^2|\mathcal{X}||\mathcal{A}|/(\pi^2\delta)\right)}}_{\text{tighter when } |\mathcal{X} \times \mathcal{A}| \text{ small}} \Big). \tag{17}$$

Define the uncertainty measure

$$\Gamma(\mathcal{S}) := \mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}}. \tag{18}$$

Then, with probability $1 - \delta/2$, $R(\hat{\pi}) \leq \sqrt{\beta \Gamma(\mathcal{S})}$.

Proof. Recall that by (2),

$$R(\hat{\pi}) = \mathop{\mathbb{E}}_{x \sim p}[\max_{a \in \mathcal{A}} \phi(x, a)^\top \theta^\star - \phi(x, \hat{\pi}(x))^\top \theta^\star].$$

Let $\pi^\star : x \to \operatorname{argmax}_{a \in \mathcal{A}} \phi(x, a)^\top \theta^\star$. Now, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\phi(x, \pi^\star(x))^\top \theta^\star - \phi(x, \hat{\pi}(x))^\top \theta^\star &= \phi(x, \pi^\star(x))^\top \hat{\theta} - \phi(x, \hat{\pi}(x))^\top \theta^\star + \phi(x, \pi^\star(x))^\top \theta^\star - \phi(x, \pi^\star(x))^\top \hat{\theta} \\
&\leq \phi(x, \hat{\pi}(x))^\top \hat{\theta} - \phi(x, \hat{\pi}(x))^\top \theta^\star + \phi(x, \pi^\star(x))^\top \theta^\star - \phi(x, \pi^\star(x))^\top \hat{\theta} \\
&= \phi(x, \hat{\pi}(x))^\top (\hat{\theta} - \theta^\star) + \phi(x, \pi^\star(x))^\top (\theta^\star - \hat{\theta}),
\end{aligned}$$

where the inequality follows because by construction of $\hat{\pi}$, we know that

$$\phi(x, \hat{\pi}(x))^\top \hat{\theta} \geq \phi(x, \pi^\star(x))^\top \hat{\theta}.$$

By Proposition 1 and 2 of [35] (or Theorem 19.2 of [26], derived originally from [1]), we have that with probability $1 - \delta/2$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$|\phi(x, a)^\top (\hat{\theta} - \theta^\star)| \leq \sqrt{\beta}/2 \cdot \max_{a \in \mathcal{A}} \|\phi(x, a)\|_{\Sigma_{\mathcal{S}}{}^{-1}}.$$

Condition on this event in the remainder of the proof.

Now, combining the preceding two displays and applying a triangle inequality, we can conclude

$$\begin{aligned}
R(\hat{\pi}) &\leq \mathop{\mathbb{E}}_{x \sim p} \phi(x, \hat{\pi}(x))^\top (\hat{\theta} - \theta^\star) + \phi(x, \pi^\star(x))^\top (\theta^\star - \hat{\theta}) \\
&\leq 2 \mathop{\mathbb{E}}_{x \sim p}[\max_{a \in \mathcal{A}} |\phi(x, a)^\top (\hat{\theta} - \theta^\star)|] \leq \sqrt{\beta} \cdot \mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|_{\Sigma_{\mathcal{S}}^{-1}}.
\end{aligned}$$

Now, note that Jensen's inequality ensures

$$\left(\mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|_{\Sigma_{\mathcal{S}}^{-1}} \right)^2 \leq \mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}},$$

or equivalently,

$$\mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|_{\Sigma_{\mathcal{S}}^{-1}} \leq \sqrt{\mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}}}.$$

Thus,

$$R(\hat{\pi}) \leq \sqrt{\beta} \cdot \sqrt{\mathop{\mathbb{E}}_{x \sim p} \max_{a \in \mathcal{A}} \|\phi(x, a)\|^2_{\Sigma_{\mathcal{S}}^{-1}}}. \tag{20}$$

□

The following lemma about *Schur complements* will later be helpful for justifying the SDP formulation of Line 2 in Algorithm 1. We believe the following Lemma I.3 is well-known, but we include the proof for completeness.

Lemma I.3 (Schur complements)**.** *Let $M \in \mathbb{R}^{(d+1)\times(d+1)}$ be a symmetric matrix of the form*

$$M = \begin{pmatrix} a & b^\top \\ b & C \end{pmatrix}$$

where $a \in \mathbb{R}, b \in \mathbb{R}^d, C \in \mathbb{R}^{d\times d}$ and $C \succ 0$. Then, $M \succeq 0$ if and only if $a - b^\top C^{-1} b \geq 0$. The quantity $a - b^\top C^{-1} b$ is called the Schur complement *of block C.*

Proof. First, note that $C \succ 0$ is invertible, and so let

$$G = \begin{pmatrix} 1 & -b^\top C^{-1} \\ 0 & I \end{pmatrix}.$$

Note that G is invertible, and in particular, it is easy to verify that

$$G^{-1} = \begin{pmatrix} 1 & b^\top C^{-1} \\ 0 & I \end{pmatrix}.$$

Next, observe that

$$GMG^\top = \begin{pmatrix} a - b^\top C^{-1} b & 0 \\ 0 & C \end{pmatrix}.$$

Now, to prove the claim, first, suppose that $M \succeq 0$. Then, consider any $z \in \mathbb{R}^d$ and note that

$$z^\top GMG^\top z = (G^\top z)^\top M (G^\top z) \geq 0,$$

and hence $GMG^\top \succeq 0$. However, $GMG^\top$ is also block-diagonal, so we can conclude that each block must be positive semi-definite. Thus, $a - b^\top C^{-1} b \geq 0$.

On the other hand, suppose that $a - b^\top C^{-1} b \geq 0$. Then, because $GMG^\top$ is block-diagonal and $C \succ 0$, we have that $GMG^\top \succeq 0$. Then, consider any $z \in \mathbb{R}^{d+1}$ and note that

$$0 \leq ((G^{-1})^\top z)^\top GMG^\top ((G^{-1})^\top z) = z^\top M z.$$

Thus, we conclude that $M \succeq 0$. □

I.4 Proof of Theorem 2.2

In this section, we prove Theorem 2.2. Throughout this section, we fix $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ to be any SCLB instance.

We divide our proof of Theorem 2.2 into two sub-sections. First, in Section I.4.1 we explain how to formulate (5) as an SDP in order to implement Line 2 of Algorithm 1. This, along with the fact that Theorem C.1 can be formulated as an SDP [26] ensures that the algorithm is implementable in polynomial time.

Second, in Section I.4.2 we carry out the regret analysis, which was sketched in the main body.

I.4.1 Expressing (5) as an SDP to implement Line 2

In this section, we show how to express the optimization problem (5) as an SDP. First, we require the following technical lemma.

Theorem I.4 (SDP formulation)**.** *Let $h \in \mathbb{R}_{\geq 0}^{\mathcal{X}\times\mathcal{A}}$ be such that $\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} h(x,a) \leq 1$. Then, the optimization problem*

$$\begin{aligned} \text{minimize:} \quad & \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}, \\ \text{subject to:} \quad & w \in \Delta^{\mathcal{X}\times\mathcal{A}}, \\ & w(x,a) \geq h(x,a) \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned} \tag{21}$$

is equivalent to the following semi-definite program (SDP):

$$\begin{aligned} \text{minimize:} \quad & \sum_{x\in\mathcal{X}} t(x)\, p(x), \\ \text{subject to:} \quad & w \in \Delta^{\mathcal{X}\times\mathcal{A}}, \\ & \begin{pmatrix} t(x) & \phi(x,a)^\top \\ \phi(x,a) & \Sigma_w \end{pmatrix} \succeq 0 \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}, \\ & w(x,a) \geq h(x,a) \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned} \tag{22}$$

Proof. By Lemma I.3 the optimization problem (22) is equivalent to

$$\begin{aligned} \text{minimize:} \quad & \sum_{x\in\mathcal{X}} t(x)\, p(x), \\ \text{subject to:} \quad & w \in \Delta^{\mathcal{X}\times\mathcal{A}}, \\ & t(x) - \phi(x,a)^\top \Sigma_w^{-1} \phi(x,a) \geq 0 \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}, \\ & w(x,a) \geq h(x,a) \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned}$$

Next, note that $t(x) - \phi(x,a)^\top \Sigma_w^{-1} \phi(x,a) \geq 0$ if and only if $t(x) \geq \|\phi(x,a)\|^2_{\Sigma_w^{-1}}$. Thus, (22) is further equivalent to

$$\begin{aligned} \text{minimize:} \quad & \sum_{x\in\mathcal{X}} t(x)\, p(x), \\ \text{subject to:} \quad & w \in \Delta^{\mathcal{X}\times\mathcal{A}}, \\ & t(x) \geq \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \quad \forall\, x \in \mathcal{X}, \\ & w(x,a) \geq h(x,a) \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned}$$

From the above display we conclude that (22) is equivalent to

$$\begin{aligned} \text{minimize:} \quad & \sum_{x\in\mathcal{X}} p(x) \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}, \\ \text{subject to:} \quad & w \in \Delta^{\mathcal{X}\times\mathcal{A}}, \\ & w(x,a) \geq h(x,a) \quad \forall\, (x,a) \in \mathcal{X}\times\mathcal{A}. \end{aligned}$$

which is equivalent to (5). □

Remark I.5. In the special case where $q \equiv 0$, (21) is equivalent to (5). Meanwhile, it is also easy to see that (21) generalizes (7).

I.4.2 Regret analysis

In this section I.4.2 we carry out the regret analysis outlined in the main body.

Matrix concentration. Our regret analysis relies on the following standard matrix version of Hoeffding's inequality.

Theorem I.6 (Matrix Hoeffding (Theorem 1.3 of [37]))**.** *Consider a finite sequence of independent, symmetric, random matrices $X_1, ..., X_T \in \mathbb{R}^{d\times d}$ and $\gamma > 0$ such that*

$$\mathbb{E}[X_t] = 0, \text{ and } X_t^2 \preceq \gamma^2 I \text{ almost surely.}$$

Then, for all $\eta \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum_{t\in[T]} X_t \right) \geq \eta \right\} \leq d \exp\left(-\frac{\eta^2}{8\gamma^2 T} \right).$$

In fact, we will only need to use the following (simple) corollary of Theorem I.6.

Corollary I.7 (Matrix Hoeffding Corollary)**.** *Consider a finite sequence of symmetric matrices $X_1, ..., X_T \in \mathbb{R}^{d\times d}$ and $\gamma > 0$ such that*

$$\mathbb{E}[X_t] = 0, \text{ and } \|X_t\| \leq \gamma \text{ almost surely.}$$

Then, for all $\eta \geq 0$,

$$\mathbb{P}\left\{\left\|\frac{1}{T}\sum_{t\in[T]} X_t\right\| \geq \eta\right\} \leq 2d\exp\left(-\frac{\eta^2 T}{8\gamma^2}\right).$$

Proof. Note that $\|X_t\| \leq \gamma$ implies that $X_t^2 \preceq \gamma^2 I$. Thus, for all $\eta \geq 0$

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\frac{1}{T}\sum_{t\in[T]} X_t\right) \geq \eta\right\} &= \mathbb{P}\left\{\lambda_{\max}\left(\sum_{t\in[T]} X_t\right) \geq T\eta\right\} \\ &\leq d\exp\left(-\frac{\eta^2 T^2}{8\gamma^2 T}\right) = d\exp\left(-\frac{\eta^2 T}{8\gamma^2}\right),\end{aligned}$$

where the inequality holds by Theorem I.6. Applying the same analysis to $-X_t$, we have that

$$\mathbb{P}\left\{\lambda_{\max}\left(\frac{1}{T}\sum_{t\in[T]} -X_t\right) \geq \eta\right\} = \mathbb{P}\left\{\lambda_{\min}\left(\frac{1}{T}\sum_{t\in[T]} X_t\right) \leq -\eta\right\} \leq d\exp\left(-\frac{\eta^2 T}{8\gamma^2}\right).$$

The corollary now follows by a union bound over the events in the two preceding displays. □

With these concentration inequalities, we are prepared to analyze the concentration of $\hat{\Sigma}_{w,T}$ to Σ_w.

Concentration of $\hat{\Sigma}_{w,T}$ to Σ_w. To aid the analysis, we define some additional notation. Let $w^\star, q^\star$ be as in (5), Theorem C.1 respectively; w^* be as in (7), w, q be as in Algorithm 1 and

$$M := \Sigma_q^{-1/2}\Sigma_w\Sigma_q^{-1/2}.$$

We collect some useful properties in the following lemma.

Lemma I.8 (Properties to aid concentration analysis)**.** *Let w, q be as in Algorithm 1. Then, the following hold true.*

(i) $\Sigma_w \succeq \alpha\Sigma_q$.

(ii) $M \succeq \alpha I$.

(iii) For each $t \in [T]$, let Z_t be a random matrix defined as follows. Draw $(x_t, a_t) \sim w$ (i.i.d. for each t) and let

$$Z_t = M^{-1/2}\left[\Sigma_q^{-1/2}\left(\frac{\lambda}{T}I + \phi(x_t,a_t)\phi(x_t,a_t)^\top\right)\Sigma_q^{-1/2} - M\right]M^{-1/2}.$$

Then, $\mathbb{E}[Z_t] = 0$.

(iv) $\|Z_t\| \leq \frac{3d}{\alpha}$.

Proof. We prove the claims one-by-one. Recall that $w, q \in \Delta^{\mathcal{X}\times\mathcal{A}}$ and consequently, we can write $w = (1-\alpha)\bar{w} + \alpha q$ for some $\bar{w} \in \Delta^{\mathcal{X}\times\mathcal{A}}$.

(i) By Lemma I.2 (v), we have that

$$\Sigma_w = (1-\alpha)\Sigma_{\bar{w}} + \alpha\Sigma_q \succeq \alpha\Sigma_q.$$

(ii) Expanding out M, by Lemma I.2 (v), we have

$$\begin{aligned} M &= \Sigma_q^{-1/2}((1-\alpha)\Sigma_{\bar{w}} + \alpha\Sigma_q)\Sigma_q^{-1/2} \\ &= (1-\alpha)\Sigma_q^{-1/2}\Sigma_{\bar{w}}\Sigma_q^{-1/2} + \alpha I \succeq \alpha I, \end{aligned}$$

where the last inequality used Lemma I.2 (v).

(iii) To see that $\mathbb{E}[Z_t] = 0$ note that by linearity,

$$\mathbb{E}\left[\frac{\lambda}{T}I + \phi(x_t, a_t)\phi(x_t, a_t)^\top\right] = \Sigma_w.$$

So, by linearity of expectation and the definition of M, we have

$$\mathbb{E}\left[\Sigma_q^{-1/2}\left(\frac{\lambda}{T}I + \phi(x_t, a_t)\phi(x_t, a_t)^\top\right)\Sigma_q^{-1/2} - M\right] = 0.$$

Applying linearity of expectation once more,

$$\mathbb{E}\left[M^{-1/2}\left[\Sigma_q^{-1/2}\left(\frac{\lambda}{T}I + \phi(x_t, a_t)\phi(x_t, a_t)^\top\right)\Sigma_q^{-1/2} - M\right]M^{-1/2}\right] = 0.$$

(iv) Notice that we can expand Z_t into three terms as follows.

$$Z_t = \frac{\lambda}{T}M^{-1/2}\Sigma_q^{-1}M^{-1/2} + [M^{-1/2}\Sigma_q^{-1/2}\phi(x_t, a_t)][M^{-1/2}\Sigma_q^{-1/2}\phi(x_t, a_t)]^\top - I.$$

We analyze this term by term and apply triangle inequality.

- The spectral norm of the first term is bounded using submultiplicativity:

$$\left\|\frac{\lambda}{T}M^{-1/2}\Sigma_q^{-1}M^{-1/2}\right\| \leq \frac{\lambda}{T}\|M^{-1}\|\|\Sigma_q^{-1}\| \leq \frac{\lambda}{T}\cdot\frac{1}{\alpha}\cdot\frac{T}{\lambda} = \frac{1}{\alpha}.$$

  The last inequality in the display above used the fact that $\Sigma_q \succeq \frac{\lambda}{T}I$ and the fact from part (ii) that $M \succeq \alpha I$ to deduce (using Lemma I.2 (iv)) that $\Sigma_q^{-1} \preceq \frac{T}{\lambda}I$ and $M^{-1} \preceq \frac{1}{\alpha}I$.
- Meanwhile, the second term is a rank-one matrix, and hence,

$$\begin{aligned} \left\|[M^{-1/2}\Sigma_q^{-1/2}\phi(x_t, a_t)][M^{-1/2}\Sigma_q^{-1/2}\phi(x_t, a_t)]^\top]\right\| &= \|M^{-1/2}\Sigma_q^{-1/2}\phi(x_t, a_t)\|^2 \\ &= \phi(x_t, a_t)^\top\Sigma_q^{-1/2}M^{-1}\Sigma_q^{-1/2}\phi(x_t, a_t) \\ &= \phi(x_t, a_t)^\top\Sigma_w^{-1}\phi(x_t, a_t). \end{aligned}$$

  However, note that by (i) and Lemma I.2 (iv), we know $\Sigma_w^{-1} \preceq \frac{1}{\alpha}\Sigma_q^{-1}$. Thus, by Lemma I.2 (i) we have that

$$\begin{aligned} \phi(x_t, a_t)^\top\Sigma_w^{-1}\phi(x_t, a_t) &\leq \frac{1}{\alpha}\phi(x_t, a_t)^\top\Sigma_q^{-1}\phi(x_t, a_t) \\ &= \frac{1}{\alpha}\|\phi(x_t, a_t)\|_{\Sigma_q^{-1}} \leq \frac{2d}{\alpha}, \end{aligned}$$

  where the last inequality holds by construction of q in Line 1 of Algorithm 1.
- The last term is $-I$, whose spectral norm is 1.

Consequently, by triangle inequality,

$$\|Z_t\| \leq \frac{1}{\alpha} + 1 + \frac{2d}{\alpha} \leq \frac{4d}{\alpha},$$

where the second inequality used that $d \geq 1, \alpha < 1$ implies $1 + 1/\alpha < 2d/\alpha$.

□

Lemma I.9 (Application of matrix Hoeffding)**.** *Let w, q, α be as in Algorithm 1. Define $T_0 = 512 \cdot d^2/\alpha^2 \log(4d/\delta)$. Then, for any $T \geq T_0$, with probability $1 - \delta/2$,*

$$\|M^{-1/2}[\Sigma_q^{-1/2}\hat{\Sigma}_{w,T}\Sigma_q^{-1/2} - M]M^{-1/2}\| \leq \frac{1}{2}.$$

Proof. Let Z_t be as in Lemma I.8. We will apply Corollary I.7 with $X_t \leftarrow Z_t, \gamma \leftarrow 4d/\alpha, \eta \leftarrow 1/2$. Note that

$$M^{-1/2}[\Sigma_q^{-1/2}\hat{\Sigma}_{w,T}\Sigma_q^{-1/2} - M]M^{-1/2} = \frac{1}{T}\sum_{t\in[T]} Z_t,$$

and hence Lemma I.8 along with Corollary I.7 ensures that

$$\mathbb{P}\left\{\|M^{-1/2}[\Sigma_q^{-1/2}\hat{\Sigma}_{w,T}\Sigma_q^{-1/2} - M]M^{-1/2}\| \geq 1/2\right\} \leq 2d\exp\left(-\frac{T}{32\gamma^2}\right).$$

Whenever $T \geq T_0 = 32\gamma^2\log(4d/\delta)$, the right-hand side is at most $\delta/2$. □

Next, we fill out the proof of Lemma C.2.

Lemma C.2. *Let $w^\star$, w^* be as in* (5)*,* (7)*, respectively. Moreover, let w be as in Line 2. Then, $\mathbb{E}_{x\sim p}\max_{a\in\mathcal{A}}\|\phi(x,a)\|^2_{\Sigma_w^{-1}} \leq 2\,\mathbb{E}_{x\sim p}\max_{a\in\mathcal{A}}\|\phi(x,a)\|^2_{\Sigma_{w^*}^{-1}} \leq 2/(1-\alpha)\cdot\mathcal{C}_\mathcal{B}$.*

Proof. The first inequality holds immediately by Line 2. For the second inequality, let $w' = (1-\alpha)w^\star + \alpha q$. Note that w' is feasible for (7). Since $\Sigma_{w'} \succeq (1-\alpha)\cdot\Sigma_{w^\star}$ implies $\Sigma_{w'}^{-1} \preceq 1/(1-\alpha)\Sigma_{w^\star}^{-1}$,

$$\mathop{\mathbb{E}}_{x\sim p}\max_{a\in\mathcal{A}}\|\phi(x,a)\|^2_{\Sigma_{w^*}^{-1}} \leq \mathop{\mathbb{E}}_{x\sim p}\max_{a\in\mathcal{A}}\|\phi(x,a)\|^2_{\Sigma_{w'}^{-1}} \leq 1/(1-\alpha)\mathop{\mathbb{E}}_{x\sim p}\max_{a\in\mathcal{A}}\|\phi(x,a)\|^2_{\Sigma_{w^\star}^{-1}}$$

where the first inequality is by the optimality of w^* and the second inequality is by the Loewner ordering. The lemma now follows by definition of $\mathcal{C}_\mathcal{B}$. □

Finally, we are prepared to prove Theorem 2.2.

Theorem 2.2. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be an SCLB, $\lambda > 0$, $\delta \in (0,1)$, and $T > 0$ be a sample budget. Invoke Algorithm 1 with $\alpha \leftarrow 1/2$. Let β be as defined in* (17)*. There exists $T_0 = \tilde{O}(d^2)$ so that whenever $T \geq T_0$, with probability $1-\delta$, Algorithm 1 outputs $\hat{\pi}$ with $R(\hat{\pi}) \leq \tilde{O}(\sqrt{\beta\mathcal{C}_\mathcal{B}/T})$. This regret bound is always at most $\tilde{O}(\sqrt{\beta d/T})$ Moroever, the algorithm runs in polynomial time.*

Proof. Take $\alpha = 1/2$ and T_0 as in Lemma I.9. Then, by Lemma I.9, with probability $1-\delta/2$,

$$\|M^{-1/2}[\Sigma_q^{-1/2}\hat{\Sigma}_{w,T}\Sigma_q^{-1/2} - M]M^{-1/2}\| \leq 1/2.$$

Condition on this event in the remainder of the proof. Now, by Lemma I.2 (vi),

$$1/2\cdot M \preceq \Sigma_q^{-1/2}\hat{\Sigma}_{w,T}\Sigma_q^{-1/2} \preceq 3/2\cdot M.$$

Multiplying through by $\Sigma_q^{1/2}$ on the left and right and applying Lemma I.2 (iii), we have

$$1/2\cdot\Sigma_q^{1/2}M\Sigma_q^{1/2} \preceq \hat{\Sigma}_{w,T} \preceq 3/2\cdot\Sigma_q^{1/2}M\Sigma_{\hat{q}}^{1/2}.$$

Substituting in the definition $M = \Sigma_q^{-1/2}\Sigma_w\Sigma_q^{-1/2}$, we can simplify the above display to obtain

$$1/2\cdot\Sigma_w \preceq \hat{\Sigma}_{w,T} \preceq 3/2\cdot\Sigma_w.$$

Consequently, by Lemma I.2 (iv), it follows that

$$\hat{\Sigma}_{w,T}^{-1} \preceq 2\cdot\Sigma_w^{-1}.$$

Now, note that for $\mathcal{S}$ as defined in Line 3, $\Sigma_\mathcal{S} = T\hat{\Sigma}_{w,T}$ and hence, $T\Sigma_\mathcal{S}^{-1} = \hat{\Sigma}_{w,T}^{-1}$. Thus,

$$\Sigma_\mathcal{S}^{-1} \preceq \frac{2}{T}\Sigma_w^{-1}. \tag{23}$$

Consequently, Lemma I.2 (i), (23) ensures that for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, we have

$$\|\phi(x,a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} = \phi(x,a)^\top \Sigma_{\mathcal{S}}^{-1} \phi(x,a) \le \frac{2}{T} \phi(x,a)^\top \Sigma_w^{-1} \phi(x,a) = \frac{2}{T} \|\phi(x,a)\|^2_{\Sigma_w^{-1}}.$$

Taking expectation and max and applying the above display point-wise, we have

$$\mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{\mathcal{S}}^{-1}} \le \frac{2}{T} \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \le \frac{8}{T} \mathcal{C}_{\mathcal{B}},$$

where the last inequality holds due to Lemma C.2.

Now, by applying a union bound with the guarantee of Lemma I.1, we see that with probability $1-\delta$, $R(\hat{\pi}) \le \tilde{O}(\sqrt{\mathcal{C}_{\mathcal{B}} \beta / T})$.

Finally, to show that $\mathcal{C}_{\mathcal{B}} \le d$, we let $q^\star$ be the G-optimal design as in the Kiefer-Wolfowitz Theorem (Theorem C.1). Then, by the definition of $\mathcal{C}_{\mathcal{B}}$, (recall (6)) we have

$$\mathcal{C}_{\mathcal{B}} \le \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{q^\star}} \le \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_{q^\star}} \le d.$$

□

I.5 Miscellaneous omitted proofs

Lemma D.2. *For any $d \in \mathbb{Z}_{>0}, A \in \mathbb{Z}_{>1}$, $\mathcal{C}_{\mathcal{B}^\star_{d,A}} \le 4$.*

Proof. Consider $w(x,a)$ defined as follows

$$w(x,a) = \begin{cases} 1-(d-1)/(2d), & x=1, a=1 \\ 1/(2d), & x \ne 1, a = 1, . \\ 0, & \text{otherwise} \end{cases}$$

Note that because $\mathcal{X} = [d]$, we have

$$\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a) = \sum_{x\in[d]} w(x,1) = 1 - \frac{(d-1)}{2d} + (d-1)\cdot\frac{1}{2d} = 1.$$

Then, we can see that

$$C := \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a)\phi(x,a)\phi(x,a)^\top,$$

is a diagonal matrix with

$$C_{ii} = \begin{cases} 1-(d-1)/(2d), & i=1 \\ 1/(2d), & i \ne 1 \end{cases}.$$

Thus,

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \phi(x,a)^\top C^{-1} \phi(x,a) &\le \frac{2d}{d+1} + \sum_{i=2}^{d} \frac{1}{d^2} \max\left(2d, \frac{2d}{d+1}\right) \\ &\le \frac{2d}{d+1} + 2(d-1)\max\left(\frac{1}{d}, \frac{1}{d(d+1)}\right) \\ &\le 2 + \frac{2(d-1)}{d} \le 4, \end{aligned}$$

where the second-to-last inequality holds because $d(d+1) \ge d$. □

Theorem 2.1. *Let $\mathcal{B} = (\mathcal{X}, \mathcal{A}, \phi, p, \nu, \theta^\star)$ be* any *SCLB. For each $t \in [T]$, let $\pi_t : \mathcal{X} \to \Delta^{\mathcal{A}}$ be arbitrary. Let $\mathcal{S} = \{(x_1, a_1), ..., (x_T, a_T)\} \subset \mathcal{X} \times \mathcal{A}$ such that for each $t \in [T]$, $x_t \sim p$ and $a_t \sim \pi_t(x_t)$. Then, as $\lambda \to 0$, $\mathbb{E}_{\mathcal{S}}[\Gamma(\mathcal{S})] \ge d/T$.*

Proof. Note that

$$\mathbb{E}[\Sigma_{\mathcal{S}}] = \lambda I + \sum_{t\in[T]} \sum_{x\in\mathcal{X}} p(x) \sum_{a\in\mathcal{A}} [\pi_t(x)]_a \phi(x,a)\phi(x,a)^\top,$$

and consequently,

$$\begin{aligned}\frac{1}{T}\mathbb{E}[\Sigma_{\mathcal{S}}] &= \frac{\lambda}{T} I + \sum_{x\in\mathcal{X}} p(x) \frac{1}{T} \sum_{t\in[T]} \sum_{a\in\mathcal{A}} [\pi_t(x)]_a \phi(x,a)\phi(x,a)^\top \\ &= \frac{\lambda}{T} I + \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} p(x) \left(\frac{1}{T} \sum_{t\in[T]} [\pi_t(x)]_a \right) \phi(x,a)\phi(x,a)^\top.\end{aligned}$$

So, for each $(x,a) \in \mathcal{X}\times\mathcal{A}$, let

$$\begin{aligned}[\bar{\pi}(x)]_a &:= \frac{1}{T} \sum_{t\in[T]} [\pi_t(x)]_a, \\ w(x,a) &:= p(x)[\bar{\pi}(x)]_a.\end{aligned}$$

We then observe that

$$\frac{1}{T}\mathbb{E}[\Sigma_{\mathcal{S}}] = \Sigma_w, \quad T\left(\mathbb{E}[\Sigma_{\mathcal{S}}]\right)^{-1} = \Sigma_w^{-1}. \tag{24}$$

Thus, we can observe that

$$\begin{aligned}T \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{(\mathbb{E}[\Sigma_{\mathcal{S}}])^{-1}} &= \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \\ &\geq \mathop{\mathbb{E}}_{x\sim p} \mathop{\mathbb{E}}_{a\sim\bar{\pi}(x)} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \\ &= \mathop{\mathbb{E}}_{(x,a)\sim w} \|\phi(x,a)\|^2_{\Sigma_w^{-1}} \\ &= \mathop{\mathbb{E}}_{(x,a)\sim w} \phi(x,a)^\top \Sigma_w^{-1} \phi(x,a) \\ &= \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} \mathrm{tr}\left(w(x,a)\phi(x,a)\phi(x,a)^\top \Sigma_w^{-1}\right) \\ &= \mathrm{tr}\left(\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a)\phi(x,a)\phi(x,a)^\top \Sigma_w^{-1} \right).\end{aligned}$$

where the inequality follows from the simple fact that the expectation is always at most the maximum; the second-to-last equality uses the cyclic property of the trace; and the last equality uses linearity of the trace. Substituting in the formula for Σ_w and dividing both sides of the above display by T,

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{(\mathbb{E}[\Sigma_{\mathcal{S}}])^{-1}} \geq \frac{1}{T}\mathrm{tr}\left(\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a)\phi(x,a)\phi(x,a)^\top \Sigma_w^{-1} \right) \\ &= \frac{1}{T}\mathrm{tr}\left(\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a)\phi(x,a)\phi(x,a)^\top \left[\frac{\lambda}{T} I + \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} w(x,a)\phi(x,a)\phi(x,a)^\top \right]^{-1} \right) \\ &\underset{\lambda\to 0}{\to} \frac{d}{T}.\end{aligned}$$

Finally, using that $\mathbb{E}[\Sigma_{\mathcal{S}}^{-1}] \succeq (\mathbb{E}[\Sigma_{\mathcal{S}}])^{-1}$ and Lemma I.2 (i), we have

$$\mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{(\mathbb{E}[\Sigma_{\mathcal{S}}])^{-1}} \leq \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\mathbb{E}[\Sigma_{\mathcal{S}}^{-1}]}.$$

Combining the two above displays, we conclude that

$$\lim_{\lambda\to 0} \mathop{\mathbb{E}}_{x\sim p} \max_{a\in\mathcal{A}} \|\phi(x,a)\|^2_{\mathbb{E}[\Sigma_{\mathcal{S}}^{-1}]} \geq \frac{d}{T},$$

as desired. □