Adapting to Misspecification in Contextual Bandits

Dylan J. Foster[⋆] dylanf@mit.edu
Claudio Gentile[†] cgentile@google.com
Mehryar Mohri[†‡] mohri@google.com
Julian Zimmert[†] zimmert@google.com

Abstract

A major research direction in contextual bandits is to develop algorithms that are computationally efficient, yet support flexible, general-purpose function approximation. Algorithms based on modeling rewards have shown strong empirical performance, yet typically require a well-specified model, and can fail when this assumption does not hold. Can we design algorithms that are efficient and flexible, yet degrade gracefully in the face of model misspecification? We introduce a new family of oracle-efficient algorithms for ε-misspecified contextual bandits that adapt to unknown model misspecification—both for finite and infinite action settings. Given access to an *online oracle* for square loss regression, our algorithm attains optimal regret and—in particular—optimal dependence on the misspecification level, with *no prior knowledge*. Specializing to linear contextual bandits with infinite actions in d dimensions, we obtain the first algorithm that achieves the optimal $\tilde{\mathcal{O}}(d\sqrt{T} + \varepsilon\sqrt{d}T)$ regret bound for unknown ε.

On a conceptual level, our results are enabled by a new optimization-based perspective on the regression oracle reduction framework of Foster and Rakhlin [21], which we believe will be useful more broadly.

1 Introduction

The contextual bandit (CB) problem is an extension of the standard multi-armed bandit problem that is relevant to a variety of applications in practice, including health services [43], online advertisement [35, 4] and recommendation systems [8]. In the contextual bandit setting, at each round, the learner observes a feature vector (or *context*) and an action set. The learner must select an action out of that set and only observes the reward of that action. To make its selection, the learner has access to a family of hypotheses (or *policies*), which map contexts to actions. The objective of the learner is to achieve a cumulative reward that is close to that of the best hypothesis in hindsight for that specific sequence of contexts and action sets.

A common approach to the contextual bandit problem consists of reducing it to a supervised learning task such as classification or regression [33, 20, 6, 7, 42, 8, 36]. Recently, Foster and Rakhlin [21] proposed SquareCB, an efficient reduction from K-armed contextual bandits to square loss regression under *realizability* assumptions. One open question that comes up after this work is whether their approach can be generalized to action spaces with many (or infinite) actions in d-dimensions. Another open question is whether one can seamlessly shift from realizability to misspecified models without requiring prior knowledge of the amount of misspecification. This is precisely the setup we study here, where the action set is large or infinite, but where the learner has a 'good' feature representation available up to some *unknown* amount of misspecification.

Adequately handling misspecification has been a subject of intense recent interest even for the simple special case of linear contextual bandits. Du et al. [19] questioned whether "good" is indeed enough,

[⋆]Massachusetts Institute of Technology.
[†]Google Research.
[‡]Courant Institute of Mathematical Sciences.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

that is, whether we can learn efficiently even without realizability. Lattimore et al. [34] gave a positive answer to that question, provided the misspecification level ε is known in advance, and showed that the price of misspecification (for regret) is roughly $\varepsilon\sqrt{d}T$, where d is the dimension and T is the time horizon. However, they left the adapting to unknown ε as an open question.

Our results. We provide an affirmative answer to all of these questions. We generalize SquareCB to infinite action sets, and use this strategy to adapt to unknown misspecification ε by combining it with a *bandit model selection* procedure akin to the one proposed by Agarwal et al. [9]. Our algorithm is oracle-efficient, and adapts to misspecification efficiently and optimally whenever it has access to an online oracle for square loss regression. When specialized to linear contextual bandits, it answers the question of Lattimore et al. [34].

An important conceptual contribution of our work is to show that one can view the action selection scheme used by SquareCB as an approximation to a log-barrier regularized optimization problem, which paves the way for a generalization to infinite action spaces. Another by-product of our results is a generalization of the original CORRAL algorithm [9] for combining bandit algorithms, which is simpler, flexible, and enjoys improved logarithmic factors.

1.1 Related Work

The contextual bandit is a well-studied problem, and misspecification in bandits and reinforcement learning has been the subject of intense recent interest. We mention a few works which are closely related to our results.

For linear bandits in d dimensions, Lattimore et al. [34] gave an algorithm with regret $\mathcal{O}(d\sqrt{T} + \varepsilon\sqrt{d}T)$, and left adapting to unknown misspecification for changing action sets as an open problem. Concurrent work of Pacchiano et al. [38] solves this problem for the special case where contexts/action sets are stochastic, and also leverages CORRAL-type aggregation of contextual bandit algorithms. Our results resolve this question in the general adversarial setting.

Within the literature general-purpose contextual bandit algorithms, our approach builds on a recent line of research that provides reductions to offline/online square loss regression [22, 21, 39, 46, 24].

Besides the standard references on oracle-based agnostic contextual bandits (e.g., [33, 20, 6, 7]), ε-misspecification is somewhat related to the recent stream of literature on bandits with adversarially-corrupted feedback [37, 27, 14]. See the discussion in Appendix A.

2 Problem Setting

We consider the following contextual bandit protocol: At every round $t = 1, \ldots, T$, the learner first observes a context $x_t \in \mathcal{X}$ and an action set $\mathcal{A}_t \subseteq \mathcal{A}$, where $\mathcal{A} \subseteq \mathbb{R}^d$ is a compact action space; for simplicity, we assume throughout that $\mathcal{A} = \{a \in \mathbb{R}^d \colon \|a\| \leq 1\}$, but place no restriction on $(\mathcal{A}_t)_{t=1}^T$. Given the context and action set, the learner chooses action $a_t \in \mathcal{A}_t$, then observes a stochastic loss $\ell_t \in [-1, +1]$ depending on the action selected. We assume that the sequence of context vectors x_t and the associated sequence of action sets $\mathcal{A}_t$ are generated by an oblivious adversary.

We let $\mu(a, x) := \mathbb{E}[\ell_t \,|\, x_t = x, a_t = a]$ denote the mean loss function, which is unknown to the learner. We adopt a semi-parametric approach to modeling the losses, in which $\mu(a, x)$ is modelled a (nearly) linear in the action a, but can depend on the context x arbitrarily [21, 46, 15]. In particular, we assume the learner has access to a class of functions $\mathcal{F} \subseteq \{f \colon \mathcal{X} \to \mathbb{R}^d\}$, where for each $f \in \mathcal{F}$, $\langle a, f(x)\rangle$ attempts to predict the value of $\mu(a, x)$. In a well-specified/realizable setting, one would assume that there exists some $f^\star \in \mathcal{F}$ such that $\mu(a, x) = \langle a, f^\star(x)\rangle$. In this paper, we make no such assumption, but the regret incurred by our algorithms depends on how far this is from being true. For each $f \in \mathcal{F}$, we let $\pi_f(\cdot, \cdot)$ denote the *induced policy*, whose action at time t is given by $\pi_f(x_t, \mathcal{A}_t) := \operatorname{argmin}_{a \in \mathcal{A}_t} \langle a, f(x_t)\rangle$.

The goal of the learner is to minimize its pseudoregret $\mathsf{Reg}(T)$ against the best unconstrained policy:

$$\mathsf{Reg}(T) := \mathbb{E}\Big[\textstyle\sum_{t=1}^T \mu(a_t, x_t) - \inf_{a \in \mathcal{A}_t} \mu(a, x_t)\Big].$$

Here, and for the remainder of the paper, we use $\mathbb{E}[\cdot]$ to denote the expectation with respect to both the randomized choices of the learner and the stochastic realization of the losses ℓ_t.

This setup recovers the usual finite-arm contextual bandit with K arms setting by taking $\mathcal{A}_t = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Another important special case is the well-studied linear contextual bandit setting, which corresponds to the case where $\mathcal{F}$ consists of constant vector-valued functions that do not depend on $\mathcal{X}$. Specifically, for any $\Theta \subseteq \mathbb{R}^d$, we can take $\mathcal{F} = \{x \mapsto \theta \,|\, \theta \in \Theta\}$. In this case, the prediction $\langle a, f(x)\rangle$ simplifies to $\langle a, \theta\rangle$, a constant linear function of the action space $\mathcal{A}$. This special case recovers the most widely studied version of linear contextual bandits [3, 12, 1, 16, 2, 10, 17], as well as Gaussian process extensions [40, 31, 18, 41].

2.1 Misspecification

Contextual bandit algorithms based on modeling rewards typically rely on the assumption of a *well-specified model* (or, "realizability"): That is, existence of a function $f^\star \in \mathcal{F}$ such that $\mu(a, x) = \langle a, f^\star(x)\rangle$ for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$ [16, 1, 6, 22]. Since this assumption may not be realistic in practice, a more recent line of work has begun to develop algorithms for misspecified models. In particular, Crammer and Gentile [17], Ghosh et al. [26], Lattimore et al. [34] and Foster and Rakhlin [21] consider a uniform ε-misspecified setting in which

$$\inf_{f\in\mathcal{F}} \sup_{a\in\mathcal{A}, x\in\mathcal{X}} |\mu(a, x) - \langle a, f(x)\rangle| \leq \varepsilon, \tag{1}$$

for some misspecification level $\varepsilon > 0$. Notably, Lattimore et al. [34] show that for the linear setting, regret must grow as $\Omega(d\sqrt{T} + \varepsilon\sqrt{d}T)$. Since $d\sqrt{T}$ is the optimal regret for a well-specified model, $\varepsilon\sqrt{d}T$ may be thought of as the price of misspecification.

In this paper, we consider a weaker average-case notion of misspecification. Given a sequence $S = (x_1, \mathcal{A}_1), \ldots, (x_T, \mathcal{A}_T)$ of context-action set pairs, we define the average misspecification level $\varepsilon_T(S)$ as

$$\varepsilon_T(S) := \inf_{f\in\mathcal{F}} \left(\frac{1}{T}\sum_{t=1}^T \sup_{a\in\mathcal{A}_t} (\langle a, f(x_t)\rangle - \mu(a, x_t))^2\right)^{1/2}. \tag{2}$$

This quantity measures the misspecification level for the specific sequence S at hand. Of course, the uniform bound in Eq. (1) directly implies $\varepsilon_T(S) \leq \varepsilon$ for all S in Eq. (2), and $\varepsilon_T(S) = 0$ whenever the model is well-specified.

We provide regret bounds that optimally adapt to $\varepsilon_T(S)$ for any given realization of the sequence S, with no prior knowledge of the misspecification level. The issue of adapting to unknown misspecification has not been addressed even for the stronger uniform notion (1). Indeed, previous efforts typically use prior knowledge of ε to tune the exploration-exploitation scheme to encourage conservative exploration when misspecification is large; see Lattimore et al. [34, Appendix E], Foster and Rakhlin [21, Section 5.1], Crammer and Gentile [17, Section 4.2], and Zanette et al. [47] for examples. Naively adapting such schemes using, e.g., doubling tricks, presents difficulties because the quantity in Eq. (2) does not appear to be estimable without knowledge of μ.

2.2 Regression Oracles

Following Foster and Rakhlin [21], we assume access to an *online regression oracle* SqAlg, which is simply an algorithm for sequential prediction with the square loss, using $\mathcal{F}$ as a benchmark class. More precisely, the oracle operates under the following protocol. At each round $t \in [T]$, the algorithm receives a context $x_t \in \mathcal{X}$, outputs a predictor $\hat{y}_t \in \mathbb{R}^d$ (in particular, we interpret $\langle a, \hat{y}_t\rangle$ as the predicted loss for action a), then observes an action $a_t \in \mathcal{A}$ and loss $\ell_t \in [-1, +1]$ and incurs loss $(\langle a_t, \hat{y}_t\rangle - \ell_t)^2$.[4] Formally, we make the following assumption.

Assumption 1. The regression oracle SqAlg guarantees that for any (potentially adaptively chosen) sequence $\{(x_t, a_t, \ell_t)\}_{t=1}^T$,

$$\sum_{t=1}^T (\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - \inf_{f\in\mathcal{F}} \sum_{t=1}^T (\langle a_t, f(x_t)\rangle - \ell_t)^2 \leq \mathsf{Reg}_{\mathsf{Sq}}(T),$$

for some (non-data-dependent) upper bound $\mathsf{Reg}_{\mathsf{Sq}}(T)$.

[4] As in Foster and Rakhlin [21], the *square loss* itself does not play a crucial role, and can be replaced by other loss that is strongly convex with respect to the learner's predictions.

For the finite-action setting, this definition coincides with that of Foster and Rakhlin [21]. To simplify the presentation of our results, we assume throughout the paper that $\mathrm{Reg}_{\mathrm{Sq}}(T)$ is a non-decreasing function of T.

While this type of oracle suffices for all of our results, our algorithms are stated more naturally in terms of a stronger oracle which supports *weighted* online regression. In this model, we follow the same protocol as in Assumption 1, except that at each time t, the regression oracle observes a weight $w_t \geq 0$ at the same time as the context x_t, and the loss incurred is now $w_t \cdot (\langle a_t, \hat{y}_t \rangle - \ell_t)^2$. For technical reasons, we also allow the oracle for this model to be randomized. We make the following assumption.

Assumption 2. The weighted regression oracle SqAlg guarantees that for any (potentially adaptively chosen) sequence $\{(w_t, x_t, a_t, \ell_t)\}_{t=1}^T$,

$$\mathbb{E}\left[\sum_{t=1}^T w_t(\langle a_t, \hat{y}_t \rangle - \ell_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^T w_t(\langle a_t, f(x_t) \rangle - \ell_t)^2\right] \leq \mathbb{E}\Big[\max_{t \in [T]} w_t\Big] \cdot \mathrm{Reg}_{\mathrm{Sq}}(T),$$

for some upper bound $\mathrm{Reg}_{\mathrm{Sq}}(T)$, where the expectation is taken with respect to the oracle's randomization.

We show in Appendix B (Algorithm 5)that any unweighted regression oracle satisfying Assumption 1 can be transformed into a randomized oracle for weighted regression that satisfies Assumption 2, with no overhead in runtime. Hence, to simplify exposition, for the remainder of the paper we state our results in terms of weighted regression oracles satisfying Assumption 2.

Online regression has been well-studied, and many efficient algorithms are known for standard classes $\mathcal{F}$. One example, which is important for our applications, is when $\mathcal{F}$ is linear.

Example 1 (Linear Models)**.** Suppose $\mathcal{F} = \{x \mapsto \theta \,|\, \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is a convex set with $\|\theta\| \leq 1$. Then the online Newton step algorithm [28] satisfies Assumption 1 with $\mathrm{Reg}_{\mathrm{Sq}}(T) = \mathcal{O}(d \log(T))$ and—via our reduction (Algorithm 5)—can be augmented to satisfy Assumption 2.

Further examples include kernels [45], generalized linear models [29], and standard nonparametric classes [25]. We refer to Foster and Rakhlin [21] for a more extensive discussion.

Additional notation. We make use of the following additional notation. Given a set X, we let $\Delta(X)$ denote the set of all probability distributions over X. If X is continuous, we restrict $\Delta(X)$ to distributions with *countable* support. We let $\|x\|$ denote the euclidean norm for $x \in \mathbb{R}^d$. For any positive definite matrix $H \in \mathbb{R}^{d \times d}$, we denote the induced norm on $x \in \mathbb{R}^d$ by $\|x\|_H^2 = \langle x, Hx \rangle$. For functions $f, g : X \to \mathbb{R}_+$, we write $f = \mathcal{O}(g)$ if there exists some constant $C > 0$ such that $f(x) \leq Cg(x)$ for all $x \in X$. We write $f = \tilde{\mathcal{O}}(g)$ if $f = \mathcal{O}(g \max\{1, \mathrm{polylog}(g)\})$, and define $\tilde{\Omega}(\cdot)$ analogously.

3 Adapting to Misspecification: An Oracle-Efficient Algorithm

We now present our main result: an efficient reduction from contextual bandits to online regression that adapts to unknown misspecification $\varepsilon_T(S)$ and supports infinite action sets. Our main theorem is as follows.

Theorem 1. *Suppose we have access to a weighted regression oracle* SqAlg *that satisfies Assumption 2 for class $\mathcal{F}$. Then there exists an efficient reduction which guarantees that for any sequence $S = (x_1, \mathcal{A}_1), \ldots, (x_T, \mathcal{A}_T)$ with misspecification level $\varepsilon_T(S)$,*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)\log(T)} + \varepsilon_T(S)\sqrt{d}T\right).$$

The algorithm has building blocks: First, we extend the reduction of [21] to infinite action sets via a new optimization-based perspective, and we show that the resulting algorithm has favorable dependence on misspecification level when it is known in advance. Then, we combine this reduction with a scheme which aggregates multiple instances to adapt to unknown misspecification. If the time required for a single query to SqAlg is $\mathcal{T}_{\mathrm{SqAlg}}$, then the per-step runtime of our algorithm is $\tilde{\mathcal{O}}(\mathcal{T}_{\mathrm{SqAlg}} + |\mathcal{A}_t| \cdot \mathrm{poly}(d))$.

As an important application, we solve an open question recently posed by Lattimore et al. [34]: we exhibit an efficient algorithm for infinite-action linear contextual bandits which optimally adapts to unknown misspecification.

Corollary 1. *Let* $\mathcal{F} = \{x \mapsto \theta \,|\, \theta \in \mathbb{R}^d, \|\theta\| \le 1\}$. *Then there exists an efficient algorithm that, for any sequence* $S = (x_1, \mathcal{A}_1), \ldots, (x_T, \mathcal{A}_T)$*, satisfies*

$$\mathsf{Reg}(T) = \mathcal{O}\left(d\sqrt{T}\log(T) + \varepsilon_T(S)\sqrt{d}T\right).$$

This result immediately follows from Theorem 1 by applying online Newton step algorithm as the regression oracle, as in Example 1. Modulo logarithmic factors, this bound coincides with the one achieved by Lattimore et al. [34] for the simpler non-contextual linear bandit problem, for which the authors also present a matching lower bound.

The remainder of this section is dedicated to proving Theorem 1. The roadmap is as follows. First, we revisit the reduction from K-armed contextual bandits to online regression by Foster and Rakhlin [21] and provide a new optimization-based perspective. This new viewpoint leads to a natural generalization from the K-armed case to the infinite action case. We then provide an aggregation-type procedure which combines multiple instances of this algorithm to adapt to unknown misspecification, and finally put all the pieces together to prove the main result. As an extension, we also give a variant of the algorithm which enjoys improved bounds when the action sets $\mathcal{A}_t$ lie in low-dimensional subspaces of $\mathbb{R}^d$. Going forward, we abbreviate $\varepsilon_T(S)$ to ε_T whenever the sequence S is clear from context.

3.1 Oracle Reductions with Finite Actions: An Optimization-Based Perspective

An important special case of our setting, is the finite-arm contextual bandit problem, where $\mathcal{A}_t = \mathcal{K} := \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. For this setting, Foster and Rakhlin [21] proposed an efficient and optimal reduction called SquareCB, which is displayed in Algorithm 1. At each step, queries the oracle SqAlg with the current context x_t and receives a loss predictor $\hat{\theta}_t \in \mathbb{R}^K$ (so that $(\hat{\theta}_t)_i$ predicts the loss of action i). The algorithm then samples an action from a probability distribution introduced by Abe and Long [3]. Specifically for any $\theta \in \mathbb{R}^K$ and learning rate $\gamma > 0$, we define $\mathsf{abe\text{-}long}(\theta, \gamma)$ as the distribution $p \in \Delta([K])$ obtained by first selecting any $i^\star \in \operatorname{argmin}_{i \in [K]} \theta_i$, then defining

$$p_i = \begin{cases} \frac{1}{K + \gamma(\theta_i - \theta_{i^\star})}, & \text{if } i \neq i^\star, \\ 1 - \sum_{i' \neq i^\star} p_i, & \text{otherwise.} \end{cases} \tag{3}$$

By choosing $\gamma \propto \sqrt{KT/(\mathsf{Reg}_{\mathrm{Sq}}(T) + \varepsilon_T)}$, this algorithm guarantees that

$$\mathsf{Reg}(T) \le \mathcal{O}\Big(\sqrt{KT\mathsf{Reg}_{\mathrm{Sq}}(T)} + \varepsilon_T\sqrt{K}T\Big).$$

Algorithm 1: SquareCB [21]

Input: Learning rate γ, time horizon T.
Initialize Regression oracle SqAlg.
for $t = 1, \ldots, T$ **do**
- Receive context x_t.
- Let $\hat{\theta}_t$ be the oracle's prediction for x_t.
- Sample $I_t \sim \mathsf{abe\text{-}long}(\hat{\theta}_t, \gamma)$.
- Play $a_t = \mathbf{e}_{I_t}$ and observe loss ℓ_t.
- Update SqAlg with (x_t, a_t, ℓ_t).

Since this approach is the starting point for our results, it will be useful to sketch the proof. For $p \in \Delta(\mathcal{A})$, let $H_p := \mathbb{E}_{a \sim p}[aa^\top]$ be the correlation matrix, and $\bar{a}_p := \mathbb{E}_{a \sim p}[a]$ be the expected action. Let the sequence S be fixed, and let $f^\star \in \mathcal{F}$ be any regression function which attains the value of $\varepsilon_T(S)$ in Eq. (2).[5] With $a_t^\star := \pi_{f^\star}(x_t, \mathcal{A}_t)$ and $\theta_t^\star := f^\star(x_t)$, we have

$$\begin{aligned} \mathbb{E}\left[\textstyle\sum_{t=1}^T \mu(a_t, x_t) - \inf_{a \in \mathcal{A}_t} \mu(a, x_t)\right] &\le \mathbb{E}\left[\textstyle\sum_{t=1}^T \langle a_t - a_t^\star, \theta_t^\star \rangle\right] + 2\varepsilon_T T \\ &= \mathbb{E}\left[\textstyle\sum_{t=1}^T \langle \bar{a}_{p_t} - a_t^\star, \theta^\star \rangle - \frac{\gamma}{4}\|\theta^\star - \hat{\theta}_t\|_{H_{p_t}}^2\right] + \mathbb{E}\left[\textstyle\sum_{t=1}^T \frac{\gamma}{4}\|\theta^\star - \hat{\theta}_t\|_{H_{p_t}}^2\right] + 2\varepsilon_T T. \end{aligned}$$

The first expectation term above is bounded by $\mathcal{O}(KT/\gamma)$, which is established by showing that $\mathsf{abe\text{-}long}(\hat{\theta}, \gamma)$ is an approximate solution to the per-round minimax problem

$$\min_{p \in \Delta(\mathcal{K})} \max_{\theta \in \mathbb{R}^K} \max_{a^\star \in \mathcal{K}} \langle \bar{a}_p - a^\star, \theta \rangle - \frac{\gamma}{4}\|\hat{\theta} - \theta\|_{H_p}^2, \tag{4}$$

[5] If the infimum is not obtained, we can simply apply the argument that follows with a limit sequence.

with value $\mathcal{O}(K/\gamma)$. The second expectation term is bounded by $\mathcal{O}(\gamma \cdot (\mathsf{Reg}_{\mathsf{Sq}}(T) + \varepsilon_T T))$, which follows almost immediately from the definition of the square loss regret in Assumption 1 (see the proof of Theorem 3 for details). Choosing γ to balance the terms leads to the result.

As a first step toward generalizing this result to infinite actions, we propose a new distribution which *exactly* solves the minimax problem (4). This distribution is the solution to a dual optimization problem based on log-barrier regularization, and provides a new principled approach to deriving reductions.

Lemma 1. For any $\theta \in \mathbb{R}^K$ and $\gamma > 0$, the unique minimizer of Eq. (4) coincides with the unique minimizer of the $\mathsf{log\text{-}barrier}(\theta, \gamma)$ optimization problem defined by

$$\mathsf{log\text{-}barrier}(\theta, \gamma) = \operatorname{argmin}_{p \in \Delta([K])} \left\{ \langle p, \theta \rangle - \tfrac{1}{\gamma} \textstyle\sum_{a \in [K]} \log(p_a) \right\} = \left(\tfrac{1}{\lambda + \gamma \theta_i} \right)_{i=1}^{K}, \qquad (5)$$

where λ is the unique value that ensures that the weights on the right-hand side above sum to one.

The abe-long distribution is closely related to the log-barrier distribution: Rather than finding the optimal Lagrange multiplier λ that solves the log-barrier problem, the abe-long strategy simply plugs in $\lambda = K - \gamma \min_{i'} \theta_{i'}$, then shifts weight to $p_{i^\star}$ to ensure the distribution is normalized. Since the log-barrier strategy solves the minimax problem Eq. (4) exactly, plugging it into the results of Foster and Rakhlin [21] and Simchi-Levi and Xu [39] in place of abe-long leads to slightly improved constants. More importantly, this new perspective leads to a principled way to extend these reductions to infinite actions.

3.2 Moving to Infinite Action Sets: The Log-Determinant Barrier

We generalize the log-barrier distribution to infinite action sets using the log-determinant function. For any $p \in \Delta(\mathcal{A})$, denote $\bar{a}_p = \mathbb{E}_{a \sim p}[a]$ and $H_p = \mathbb{E}_{a \sim p}[aa^T]$. Furthermore we use $\dim(\mathcal{A})$ to denote the dimension of the smallest affine linear subspace that contains $\mathcal{A}$. When $\dim(\mathcal{A}) < d$, we adopt the convention that $\det(\cdot)$ takes the product of only the first $\dim(\mathcal{A})$ eigenvalues of the matrix in its argument, so that the solution bwlow is well-defined. Our *logdet-barrier* distributions are defined as follows.

Algorithm 2: SquareCB.Inf

Input: Learning rate γ, time horizon T.
Initialize Regression oracle SqAlg.
for $t = 1, \ldots, T$ **do**
- Receive context x_t.
- Let $\hat{\theta}_t$ be the oracle's prediction for x_t.
- Play $a_t \sim \mathsf{logdet\text{-}barrier}(\hat{\theta}_t, \gamma; \mathcal{A}_t)$.
- Observe loss ℓ_t.
- Update SqAlg with (x_t, a_t, ℓ_t).

Definition 1. For any $\theta \in \mathbb{R}^d$, action set $\mathcal{A} \subset \mathbb{R}^d$, and $\gamma > 0$, the set of $\mathsf{logdet\text{-}barrier}(\theta, \gamma; \mathcal{A})$ distributions are defined as the solutions to

$$\operatorname{argmin}_{p \in \Delta(\mathcal{A})} \left\{ \langle \bar{a}_p, \theta \rangle - \gamma^{-1} \log \det(H_p - \bar{a}_p \bar{a}_p^T) \right\}. \qquad (6)$$

In general, Eq. (6) does not admit a unique solution; all of our results apply to *any* minimizer. Our key result is that these logdet-barrier distributions solve a minimax problem analogous to that of Eq. (4).

Lemma 2. Any solution to $\mathsf{logdet\text{-}barrier}(\hat{\theta}, \gamma; \mathcal{A})$ satisfies

$$\max_{\theta \in \mathbb{R}^d} \max_{a^\star \in \mathcal{A}} \langle \bar{a}_p - a^\star, \theta \rangle - \tfrac{\gamma}{4} \|\hat{\theta} - \theta\|_{H_p}^2 \leq \gamma^{-1} \dim(\mathcal{A}). \qquad (7)$$

By replacing the abe-long distribution with the logdet-barrier distribution in Algorithm 1, we obtain an optimal reduction for infinite action sets. This algorithm, which we call SquareCB.Inf, is displayed in Algorithm 2.

Theorem 2. *Given a regression oracle* SqAlg *that satisfies Assumption 1 for class $\mathcal{F}$,* SquareCB.Inf *with learning rate $\gamma \propto \sqrt{dT/(\mathsf{Reg}_{\mathsf{Sq}}(T) + \varepsilon)}$ guarantees for all sequences S with $\varepsilon_T(S) \leq \varepsilon$ that*

$$\mathsf{Reg}(T) = \mathcal{O}\left(\sqrt{dT \mathsf{Reg}_{\mathsf{Sq}}(T)} + \varepsilon \sqrt{d} T \right).$$

The logdet-barrier optimization problem is closely related to the D-optimal experimental design problem and to finding the John ellipsoid [30, 44], which correspond to the case where $\theta = 0$ in Eq. (6) [32]. By adapting specialized optimization algorithms for these problems (in particular, a Frank-Wolfe-type scheme), we can efficiently solve the logdet-barrier problem. In particular, we have the following proposition.

Proposition 1. *An approximation to* (6) *that achieves the same regret bound up to a constant factor can be computed in time* $\tilde{\mathcal{O}}(|\mathcal{A}_t| \cdot \mathrm{poly}(d))$ *and memory* $\tilde{\mathcal{O}}(\log|\mathcal{A}_t| \cdot \mathrm{poly}(d))$ *per round.*

The algorithm and a full analysis for runtime and memory complexity, as well as the impact on the regret, is provided in Appendix E.

3.3 Adapting to Misspecification: Algorithmic Framework

The regret bound of SquareCB.Inf in Theorem 2 achieves optimal dependence on dimension and on the misspecification level, but requires an a-priori upper bound on $\varepsilon_T(S)$ to set the learning rate. We now turn our attention to adapting to this parameter.

At a high level, our approach is to run multiple instances of SquareCB.Inf, each tuned to a different level of misspecification, then run an aggregation procedure on top to learn the best instance. Specifically, if we initialize a collection of $M := \lfloor \log(T) \rfloor$ instances of Algorithm 2 in which the learning rate for instance m is tuned for misspecification level $\varepsilon'_m := \exp(-m)$ (that is, we follow a geometric grid), then it is straightforward to show that there exists $m^\star \in [M]$ such that the $m^\star$th instance would enjoy optimal regret if we ran it on the sequence S. Since $m^\star$ is not known a-priori, we run an aggregation (or, "Corralling") procedure [9] to select the best instance. This approach is, in general, not suitable for model selection, since it typically requires prior knowledge of the optimal regret bound to tune certain parameters appropriately [23]. We show that adaptation to misspecification is an exception to this rule, and provides a simple setting where model selection for contextual bandits is possible.

We consider the aggregation scheme in Algorithm 3, which is a generalization of the CORRAL algorithm of Agarwal et al. [9]. The algorithm is initialized with M *base* algorithms, and uses a multi-armed bandit algorithm with M arms as a *master* algorithm responsible for choosing which base algorithm to follow at each round.

Algorithm 3: Corralling [9]

Input: Master algorithm Master, T
Initialize $(\mathsf{Base}_m)_{m=1}^M$
for $t = 1, \ldots, T$ **do**
  Receive context x_t.
  Receive A_t, q_{t,A_t} from Master.
  Pass $(x_t, \mathcal{A}_t, q_{t,A_t}, \rho_{t,A_t})$ to Base_{A_t}.
  Base_{A_t} plays a_t and observes ℓ_t.
  Update Master with $\tilde{\ell}_{t,A_t} = (\ell_t + 1)$.

The master maintains a distribution $q_t \in \Delta([M])$ over the base algorithms. At each round t, it samples an algorithm $A_t \sim q_t$ and passes the current context x_t into this algorithm, as well as the sampling probability q_{t,A_t} and a weight ρ_{t,A_t}, where we define $\rho_{t,m} := 1/\min_{s \le t} q_{s,m}$ for each m. The base algorithm A_t now plays a regular contextual bandit round: Given the context x_t, it proposes an arm a_t, which is pulled, receives the loss ℓ_t, and updates its internal state. Finally, the master updates its state with the action-loss pair $(A_t, \tilde{\ell}_{t,A_t})$, where $\tilde{\ell}_{t,A_t} := \ell_t + 1$ (for technical reasons, it is useful to shift the loss by 1 to ensure non-negativity).

Let $\mathsf{Reg}^m_{\mathrm{Imp}}(T) := \mathbb{E}\Big[\sum_{t=1}^T \frac{\mathbb{I}\{A_t = m\}}{q_{t,m}} (\mu(a_t, x_t) - \inf_{a \in \mathcal{A}_t} \mu(a, x_t))\Big]$ denote the *importance-weighted regret* for base algorithm m, which is simply the pseudoregret incurred in the rounds where Algorithm 3 follows this base algorithm, weighted inversely proportional to the probability that this occurs. It is straightforward to show that for any choice of master and base algorithms, this scheme guarantees that

$$\mathsf{Reg}(T) = \mathbb{E}\Big[\sum_{t=1}^T \tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,m^\star}\Big] + \mathsf{Reg}^{m^\star}_{\mathrm{Imp}}(T)\ , \tag{8}$$

where $\tilde{\ell}_{t,m}$ henceforth denotes the loss the algorithm would have suffered at round t if we had $A_t = m$. That is, the regret of Algorithm 3 is equal to the regret $\mathsf{Reg}_M(T) := \mathbb{E}[\sum_{t=1}^T \tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,m^\star}]$ of the master algorithm, plus the importance-weighted regret of the optimal base algorithm $m^\star$.

The difficulty in instantiating this general scheme lies in the fact that the important-weighted regret of the best base typically scales with $\mathbb{E}[\rho^\alpha_{T,m^\star}] \cdot \mathsf{Reg}^{m^\star}_{\mathrm{Unw}}(T)$, where $\alpha \in [\frac{1}{2}, 1]$ is an algorithm-dependent parameter and $\mathsf{Reg}^m_{\mathrm{Unw}}(T) := \mathbb{E}[\sum_{t=1}^T \mathbb{I}\{A_t = m\} (\mu(a_t, x_t) - \inf_{a \in \mathcal{A}_t} \mu(a, x_t))]$ denotes the unweighted regret of algorithm m. A-priori, the $\mathbb{E}[\rho^\alpha_{T,m^\star}]$ can be unbounded, leading to large regret. The key to the analysis of Agarwal et al. [9], and the approach we follow here, is to use a master algorithm with *negative regret* proportional to $\mathbb{E}[\rho^\alpha_{T,m^\star}]$, allowing to cancel this factor.

Base algorithm. As the first step towards instantiating the aggregation scheme above, we specify the base algorithm. We use a modification to SquareCB.Inf based importance weighting, which is designed to ensure that the importance-weighted regret in Eq. (8) is bounded. Pseudocode for the mth base algorithm is given in Algorithm 4.

Let the instance m be fixed, and let $Z_{t,m} = \mathbb{I}\{A_t = m\}$ indicate the event that this instance gets to select an arm; note that we have $Z_{t,m} \sim \mathrm{Ber}(q_{t,m})$ marginally. When $Z_{t,m} = 1$, instance m receives $q_{t,m}$ and $\rho_{t,m} = \max_{s\le t} q_{s,m}^{-1}$ from the master algorithm. The instance then follows the same update scheme as in the vanilla version of SquareCB.Inf, except that i) it uses an adaptive learning rate $\gamma_{t,m}$, which is tuned based on $\rho_{t,m}$, and ii) it uses a weighted square loss regression oracle (Assumption 2), with the weight w_t set as a function of $\gamma_{t,m}$ and $q_{t,m}$.

Algorithm 4: SquareCB.Imp (for base alg. m)

Input: T, $\mathrm{Reg}_{\mathrm{Sq}}(T)$
Initialize Weighted regression oracle SqAlg.
for $t = (\tau_1, \tau_2, \ldots) \subset [T]$ **do**
- Receive context x_t and $(q_{t,m}, \rho_{t,m})$.
- Set $\gamma_{t,m} = \min\left\{\frac{\sqrt{d}}{\varepsilon'_m}, \sqrt{dT/(\rho_{t,m}\mathrm{Reg}_{\mathrm{Sq}}(T))}\right\}$.
- Set $w_t = \gamma_{t,m}/q_{t,m}$.
- Compute oracle's prediction $\hat{\theta}_t$ for x_t, w_t.
- Sample $a_t \sim$ logdet-barrier$(\theta_t, \gamma_{t,m}; \mathcal{A}_t)$.
- Play a_t and observe loss ℓ_t.
- Update SqAlg with (w_t, x_t, a_t, ℓ_t).

The importance weighted regret $\mathrm{Reg}^m_{\mathrm{Imp}}(T)$ for this scheme is bounded as follows.

Theorem 3. *When invoked within Algorithm 3 with a regression oracle satisfying Assumption 2, the importance-weighted regret for each instance $m \in [M]$ of Algorithm 4 satisfies*

$$\mathrm{Reg}^m_{\mathrm{Imp}}(T) \le \tfrac{3}{2}\mathbb{E}[\sqrt{\rho_{T,m}}]\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)} + \left(\left(\tfrac{\varepsilon'_m}{\varepsilon_T} + \tfrac{\varepsilon_T}{\varepsilon'_m}\right)\sqrt{d} + 2\right)\varepsilon_T T. \quad (9)$$

The key feature of this regret bound is that only the leading term involving $\mathrm{Reg}_{\mathrm{Sq}}(T)$ depends on the importance weights, not the second term involving the misspecification. This is allows us to get away with tuning the master algorithm using only d, T, and $\mathrm{Reg}_{\mathrm{Sq}}(T)$, but not ε_T, which is critical to adapt without prior knowledge. Another important detail is that if ε'_m is within a constant factor of ε_T, the second term simplifies to $\mathcal{O}(\varepsilon_T\sqrt{d}T)$ as desired.

3.4 Improved Master Algorithms for Combining Bandit Algorithms

It remains to provide a master algorithm for use within Algorithm 3. While it turns out the master algorithm proposed in Agarwal et al. [9] suffices for this task, we go a step further and propose a new master algorithm called (α, R)*–hedged FTRL* which is simpler and enjoys slightly improved regret, removing logarithmic factors. While this is not the focus of the paper, we believe that it to be a useful contribution on its own, because it provides a new approach to designing master algorithms for bandit aggregation. We hope that it will find use more broadly.

The (α, R)–hedged FTRL algorithm is parameterized by a regularizer and two scale parameters $\alpha \in (0, 1)$ and $R > 0$. We defer a precise definition and analysis to Appendix D, and state only the relevant result for our aggregation setup here. This result concerns a specific instance of the (α, R)–hedged FTRL algorithm that we call (α, R)–hedged Tsallis-INF, which instantiates the framework using the Tsallis entropy as a regularizer [11, 5, 48]. The key property of the algorithm is that the regret with respect to a policy playing a fixed arm m contains a negative contribution of magnitude $\rho^{\alpha}_{T,m}R$. The following result is a corollary of a more general theorem, Theorem 6.

Corollary 2. *Consider the adversarial multi-armed bandit problem with M arms and losses $\tilde{\ell}_{t,m} \in [0, 2]$. For any $\alpha \in (0, 1)$ and $R > 0$, the (α, R)–hedged Tsallis-INF algorithm with learning rate $\eta = \sqrt{1/(2T)}$ guarantees that for all $m^\star \in [M]$,*

$$\mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,m^\star}\right] \le 4\sqrt{2MT} + \mathbb{E}\left[\min\left\{\tfrac{1}{1-\alpha}, 2\log(\rho_{T,m^\star})\right\} M^\alpha - \rho^\alpha_{T,m^\star}\right]\cdot R. \quad (10)$$

3.5 Putting Everything Together

Crucially, the regret bound in Corollary 2 has a negative $R \cdot \rho^{\alpha}_{T,m^\star}$ term which, for sufficiently large R and appropriate α, can be used to offset the regret incurred from importance-weighting the base algorithms. In particular, $\left(\frac{1}{2}, \frac{3}{2}\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)}\right)$–hedged Tsallis-INF has exactly the negative regret contribution needed to cancel the importance weighting term in Eq. (9) if we use SquareCB.Imp as the base algorithm. In more detail, we combine the regret for the master and base algorithms as follows to prove Theorem 1.

Proof sketch for Theorem 1. Using Eq. (8), it suffices to bound the regret of the bandit master $\mathrm{Reg}_M(T)$ and the important-weighted regret $\mathrm{Reg}_{\mathrm{Imp}}^{m^\star}(T)$ for the optimal instance $m^\star$. By Corollary 2, using $\left(\frac{1}{2}, \frac{3}{2}\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)}\right)$–hedged Tsallis-INF as the master algorithm gives

$$\mathrm{Reg}_M(T) \leq \mathcal{O}\left(\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)\log(T)}\right) - \tfrac{3}{2}\mathbb{E}[\sqrt{\rho_{T,m^\star}}]\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)}.$$

Whenever the misspecification level is not trivially small, the geometric grid ensures that there exists $m^\star \in [M]$ such that $e^{-1}\varepsilon_T \leq \varepsilon'_{m^\star} \leq \varepsilon_T$. For this instance, Theorem 3 yields

$$\mathrm{Reg}_{\mathrm{Imp}}^{m^\star}(T) \leq \tfrac{3}{2}\mathbb{E}[\sqrt{\rho_{T,m^\star}}]\sqrt{dT\mathrm{Reg}_{\mathrm{Sq}}(T)} + \mathcal{O}(\varepsilon_T\sqrt{d}T).$$

Summing the two bounds using Eq. (8) completes the proof. □

3.6 Extension: Adapting to the Average Dimension

A canonical application of linear contextual bandit is the problem of online news article recommendation, where the context x_t is taken to be a feature vector containing information about the user, and each action $a \in \mathcal{A}_t$ is the concatenation of x_t with a feature representation for a candidate article (e.g., Li et al. [35]). In this application and others like it, it is often the case that while examples lie in high-dimensional space, the true dimensionality $\dim(\mathcal{A}_t)$ of the action set is small, so that $d_{\mathrm{avg}} := \frac{1}{T}\sum_{t=1}^{T}\dim(\mathcal{A}_t) \ll d$. If we have prior knowledge of d_{avg} (or an upper bound thereof), we can exploit this low dimensionality for tighter regret. In fact, following the proof of Theorem 3 and Theorem 1, and bounding $\sum_{t=1}^{T}\dim(A_t)$ by $d_{\mathrm{avg}}T$ instead of dT, it is fairly immediate to show that Algorithm 3 enjoys improved regret $\mathrm{Reg}(T) = \mathcal{O}(\sqrt{d_{\mathrm{avg}}T\mathrm{Reg}_{\mathrm{Sq}}(T)\log(T)} + \varepsilon_T\sqrt{d_{\mathrm{avg}}}T)$, so long as d_{avg} is replaced by d in the algorithm's various parameter settings. Our final result shows that it is possible to adapt to unknown d_{avg} and unknown misspecification simultaneously. The key idea to apply a doubling trick on top of Algorithm 3

Theorem 4. *There exists an algorithm that, under the same conditions as Theorem 1, satisfies* $\mathrm{Reg}(T) = \mathcal{O}\left(\sqrt{d_{\mathrm{avg}}T\mathrm{Reg}_{\mathrm{Sq}}(T)\log(T)} + \varepsilon_T\sqrt{d_{\mathrm{avg}}}T\right)$ *without prior knowledge of* d_{avg} *or* ε_T.

We remark that while the bound in Theorem 4 replaces the d factor in the reduction with the data-dependent quantity d_{avg}, the oracle's regret $\mathrm{Reg}_{\mathrm{Sq}}(T)$ may itself still depend on d unless a sufficiently sophisticated algorithm is used.

4 Discussion

We have presented the first general-purpose, oracle-efficient algorithms for contextual bandits that adapt to unknown model misspecification. For infinite-action linear contextual bandits, our results yield the first optimal algorithms that adapt to unknown misspecification with changing action sets. Our results suggest a number of interesting conceptual questions:

- Can our optimization-based perspective lead to new oracle-based algorithms for more rich types of infinite action sets? Examples include nonparametric actions and structured (e.g., sparse) linear actions.
- Can our reduction-based techniques be lifted to more sophisticated interactive learning settings such as reinforcement learning?

On the technical side, we anticipate that our new approach to reductions will find broader use; natural extensions include reductions for offline oracles [39] and adapting to low-noise conditions [24].

Lastly, we recall that in passing, we have derived a novel class of master algorithms for combining bandit algorithms which enjoys more flexibility, an improvement in logarithmic factors, and a greatly simplified analysis. We hope this result will be useful for future work on model selection in contextual bandits.

Acknowledgements

DF acknowledges the support of NSF TRIPODS grant #1740751. We thank Teodor Marinov and Alexander Rakhlin for discussions on related topics.

Broader Impact

This paper concerns contextual bandit algorithms that adapt to unknown model misspecification. Because of their efficiency and ability to adapt to the amount of misspecification contained with no prior knowledge, our algorithms are robust, and may be suitable for large-scale practical deployment. On the other hand, our work is at the level of foundational research, and hence its impact on society is shaped by the applications that stem from it. We will focus our brief discussion on the applications mentioned in the introduction.

Health services [43] offer an opportunity for potential positive impact. Contextual bandits can be used to propose medical interventions that lead to a better health outcomes. However, care must be taken to ethically implement the explore-exploit tradeoff in this sensitive setting, and more research is required. Online advertisements [4, 35] and recommendation systems [8] are another well-known application. While improved, robust algorithms can lead to increased profits here, it is important to recognize that this may positively impact society as a whole.

Lastly, we mention that predictive algorithms like contextual bandits become more and more powerful as more information is gathered about users. This provides a clear incentive toward collecting as much information as possible. We believe that the net benefit of research on contextual bandit outweighs the harm, but we welcome regulatory efforts to produce a legal framework that steers the usage of machine learning algorithms, including in contextual bandits, in a direction which is respects of the privacy rights of users.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, NIPS, pages 2312–2320. Curran Associates, Inc., 2011.

[2] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Proc. of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9, 2012.

[3] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the 16th International Conference on Machine Learning*, ICML, pages 3–11, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[4] N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

[5] J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, NIPS, pages 2197–2205. Curran Associates, Inc., 2015.

[6] A. Agarwal, M. Dudik, S. Kale, J. Langford, and R. Schapire. Contextual bandit learning with predictable rewards. In *Proc. of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[7] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1638–1646, 22–24 Jun 2014.

[8] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, and O. Ribas. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[9] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

[10] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proc. of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135, 2013.

[11] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009.

[12] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[14] I. Bogunovic, A. Krause, and J. Scarlett. Corruption-tolerant Gaussian Process bandit optimization. In *Proc. of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[15] V. Chernozhukov, M. Demirer, G. Lewis, and V. Syrgkanis. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, pages 15065–15075, 2019.

[16] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proc. of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 208–214. PMLR, 2011.

[17] K. Crammer and C. Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.

[18] J. Djolonga, A. Krause, and V. Cevher. High-dimensional Gaussian Process bandits. In *Proc. 27th NIPS*, pages 1025–1033, 2013.

[19] S. S. Du, S. M. Kakade, R. Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

[20] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 169–178, 2011.

[21] D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

[22] D. J. Foster, A. Agarwal, M. Dudik, H. Luo, and R. Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548, 2018.

[23] D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752, 2019.

[24] D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[25] P. Gaillard and S. Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory*, pages 764–796, 2015.

[26] A. Ghosh, S.R. Chowdhury, and A. Gopalan. Misspecified linear bandits. In *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[27] A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proc. of Conference on Learning Theory*, pages 1562–1578, 2019.

[28] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[29] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *NIPS*, pages 927–935, 2011.

[30] L. G. Khachiyan and M. J. Todd. On the complexity of approximating the maximal inscribed ellipsoid for a polytope. Technical report, Cornell University Operations Research and Industrial Engineering, 1990.

[31] A. Krause and C.S. Ong. Contextual Gaussian process bandit optimization. In *Proc. 25th NIPS*, 2011.

[32] P. Kumar and E. A. Yildirim. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126(1):1–21, 2005.

[33] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20*, NIPS, pages 817–824. 2008.

[34] T. Lattimore, C. Szepesvari, and W. Gellert. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

[35] K. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web*, pages 661–670, 2010.

[36] H. Luo, C-Y. Wei, A. Agarwal, and J. Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1739–1776, 2018.

[37] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proc. of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122. ACM, 2018.

[38] A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. Neural Information Processing Systems (NeurIPS), 2020.

[39] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN*, 2020.

[40] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022, June 2010.

[41] Y. Sui, A. Gotovos, J. Burdick, and A. Krause. Safe exploration for optimization with gaussian processes. In *Proc. of the 32nd International Conference on Machine Learning*, volume 37, pages 997–1005, 2015.

[42] V. Syrgkanis, A. Krishnamurthy, and R. Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2159–2168, 2016.

[43] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[44] Michael J Todd and E Alper Yıldırım. On khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Applied Mathematics*, 155(13):1731–1744, 2007.

[45] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proc. of the 29th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 654–663, 2013.

[46] Y. Xu and A. Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

[47] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent Bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

[48] Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 467–475. PMLR, 2019.

Algorithm 5: Randomized reduction from weighted to unweighted online regression

Input: Online regression oracle SqAlg satisfying Assumption 1.
Initialize $w_{\max} \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  Receive weight w_t and x_t.
  if $w_t > w_{\max}$ **then**
    Reset SqAlg.
    $w_{\max} \leftarrow 2w_t$.
  Predict $\hat{y}_t$, where $\hat{y}_t$ is the prediction from SqAlg given x_t.
  Observe a_t and ℓ_t.
  if $u_t \sim \mathrm{Ber}(w_t/w_{\max}) = 1$ **then**
    Update SqAlg with (x_t, a_t, ℓ_t).

A Additional Related Work

In particular, our work builds on and provides a new perspective on the online square loss oracle reduction of Foster and Rakhlin [21]. The infinite-action setting we consider was introduced in Foster and Rakhlin [21], but algorithms were only given for the special case where the action set is the sphere; our work extends this to arbitrary action sets. Concurrent work of Xu and Zeevi [46] gives a reduction to offline oracles for infinite action sets. This result is not strictly comparable: On one hand, an online oracle can always be converted to an offline oracle through online-to-batch conversion, but when an online oracle *is* available our algorithm is significantly more efficient.

Misspecification in contextual bandits can be formalized in different ways that go beyond the setting we consider. First, we mention a long line of work which reduces stochastic contextual bandits to oracles for cost-sensitive classification [33, 20, 6, 7]. These results are agnostic, meaning they make no assumption on the model. However, in the presence of misspecification, the type of guarantee is somewhat different than what we provide here: rather than giving a bound on regret to the true optimal policy, these results give bounds on the regret to the best-in-class policy.

Another line of works consider a model in which the feedback received by the learning algorithm at each round may be arbitrarily corrupted by an adaptive adversary [37, 27, 14]. Typical results for this setting pick up additive error $\mathcal{O}(C)$, where C is the total number of corrupted rounds. While this model was original introduced for non-contextual stochastic bandits, it has recently been extended to Gaussian process bandit optimization, which is closely related to the contextual bandit setting (though these results only tolerate $C \leq \sqrt{T}$). While this is not the focus of our paper, we mention in passing that our notion of misspecification satisfies $\varepsilon_T(S) \leq \sqrt{C/T}$, and hence our main theorem (Theorem 1) picks up additive error $\sqrt{CT}$ for this corrupted setting (albeit, only with an oblivious adversary).

B Reducing Weighted to Unweighted Regression

In this section we show how to transform any unweighted online regression oracle SqAlg satisfying Assumption 1 into a weighted oracle satisfying Assumption 2. The reduction is stated in Algorithm 5.

Theorem 5. *If the oracle* SqAlg *satisfies Assumption 1 with regret bound* $\mathrm{Reg}_{\mathrm{Sq}}(T)$*, then Algorithm 5 satisfies Assumption 2 with regret bound* $\mathrm{Reg}_{\mathrm{Sq}}(T)$.

Proof. Let $D_t = (w_t, x_t, a_t, \ell_t)$ and define the filtration $\mathfrak{F}_t = \sigma(D_{1:t})$, with the convention $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathfrak{F}_t]$. Let $\tau_1, \tau_2 \ldots, \tau_I$ denote the timesteps at which the algorithm doubles $w_{\max}$ and resets SqAlg, with the convention $\forall n > I : \tau_n = T + 1$. Note that these random variables are stopping times with respect to the filtration $\mathfrak{F}_{1:T}$, and hence $\mathfrak{F}_{\tau_i}$ is well-defined for each $i \in \mathbb{N}$. It will also be helpful to note that we always have $\tau_{i+1} > \tau_i$ for all $i \leq I$ by construction and otherwise $\tau_{i+1} = \tau_i$. We also observe that $\tau_1 = 1$ unless $w_1 = 0$.

For the first step, we show that the conditional regret of Algorithm 5 between any pair of doubling steps is bounded. Let $i \leq I$ and $f \in \mathcal{F}$ be fixed, and observe that $i \leq I$ holds iff $\tau_i \leq T$, which is

$\mathfrak{F}_{\tau_i}$-measurable. Hence,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=\tau_i}^{\tau_{i+1}-1} w_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big) \mid \mathfrak{F}_{\tau_i}\right] \\
&= \mathbb{E}\left[2w_{\tau_i} \sum_{t=\tau_i}^{\tau_{i+1}-1} \frac{w_t}{2w_{\tau_i}}\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big) \mid \mathfrak{F}_{\tau_i}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[2w_{\tau_i} \sum_{t=\tau_i}^{\tau_{i+1}-1} \mathbb{E}_t\big[u_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big)\big] \mid \mathfrak{F}_{\tau_i}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[2w_{\tau_i} \sum_{t=\tau_i}^{\tau_{i+1}-1} u_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big) \mid \mathfrak{F}_{\tau_i}\right] \\
&\overset{(c)}{\leq} \mathbb{E}[2w_{\tau_i} \mid \mathfrak{F}_{\tau_i}] \cdot \mathsf{Reg}_{\mathrm{Sq}}(T)\,,
\end{aligned}$$

where (a) follows from the conditional independence of u_t, (b) is by the tower rule of expectation, and (c) uses Assumption 1 on the set $\{t \in \{\tau_i, \dots \tau_{i+1} - 1\} \mid u_t = 1\}$ (in particular, that regret is bounded by $\mathsf{Reg}_{\mathrm{Sq}}(T)$ on every sequence with probability 1 and $\mathsf{Reg}_{\mathrm{Sq}}(T)$ is non-decreasing in T). For $i > I$, the term is 0 since the sum is empty. To complete the proof that Algorithm 5 satisfies Assumption 2, we sum the bound above across all epochs as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} w_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big)\right] \\
&\overset{(d)}{=} \mathbb{E}\left[\sum_{i=1}^{\infty}\sum_{t=\tau_i}^{\tau_{i+1}-1} w_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big)\right] \\
&\overset{(e)}{=} \mathbb{E}\left[\sum_{i=1}^{\infty}\mathbb{E}\left[\sum_{t=\tau_i}^{\tau_{i+1}-1} w_t\big((\langle a_t, \hat{y}_t\rangle - \ell_t)^2 - (\langle a_t, f(x_t)\rangle - \ell_t)^2\big) \mid \mathfrak{F}_{\tau_i}\right]\right] \\
&\overset{(f)}{\leq} \mathbb{E}\left[\sum_{i=1}^{I}\mathbb{E}[2w_{\tau_i} \mid \mathfrak{F}_{\tau_i}]\right]\mathsf{Reg}_{\mathrm{Sq}}(T) \\
&\overset{(g)}{=} 2\mathbb{E}\left[\sum_{i=0}^{I} w_{\tau_i}\right]\mathsf{Reg}_{\mathrm{Sq}}(T) \\
&\overset{(h)}{\leq} 2\mathbb{E}[2w_{\tau_I}]\mathsf{Reg}_{\mathrm{Sq}}(T) \overset{(i)}{\leq} 4\mathbb{E}\left[\max_{t\in[T]} w_t\right]\mathsf{Reg}_{\mathrm{Sq}}(T)\,,
\end{aligned}$$

where (d) uses that all $t < \tau_1$ have $w_t = 0$, (e) uses the tower rule of expectation, (f) applies the conditional bound between stopping times above, (g) uses the tower rule of expectation again, (h) holds because the weights at least double between doubling steps, and (i) follows because τ_I is a random variable with support over $[T]$. □

C Proofs from Section 3

In this section we provide complete proofs for all of the algorithmic results from Section 3.

C.1 Proofs from Section 3.1

Proof of Lemma 1. We begin by showing that the log-barrier distribution takes the form claimed in Eq. (5). The minimization problem of Lemma 1 is strictly convex and the value is ∞ on the boundary. Hence the unique solution lies in the interior of the domain. By the K.K.T. conditions, the partial derivatives in each coordinate must coincide for the minimizer $p^\star$. There exists a $\tilde{\lambda} \in \mathbb{R}$ such that

$$\forall a \in [K]: \frac{\partial}{\partial p_a}\left(\langle p^\star, \theta\rangle - \frac{1}{\gamma}\sum_{a\in[K]}\log(p^\star_a)\right) = \theta_a - \frac{1}{\gamma p^\star_a} = \tilde{\lambda}\,.$$

Substituting $\tilde{\lambda} = \min_{a\in[K]}\theta_a - 1/\gamma$ and rearranging finishes the proof.

Next we show that the log-barrier distribution indeed solves the minimax problem Eq. (4), which we rewrite as

$$\begin{aligned}
&\min_{p\in\Delta([K])}\sup_{\theta\in\mathbb{R}^K}\max_{i^\star\in[K]}\langle \bar{a}_p - \mathbf{e}_{i^\star}, \theta\rangle - \frac{\gamma}{4}\|\hat{\theta}-\theta\|^2_{H_p} \\
&= \min_{p\in\Delta([K])}\max_{i^\star\in[K]}\sup_{\delta\in\mathbb{R}^K}\langle \bar{a}_p - \mathbf{e}_{i^\star}, \hat{\theta}+\delta\rangle - \frac{\gamma}{4}\|\delta\|^2_{H_p}\,. \qquad (11)
\end{aligned}$$

For any choice of p and $i^\star$, taking the derivative of the expression in Eq. (11) with respect to δ, we have

$$\frac{\partial}{\partial\delta}\left[\langle \bar{a}_p - \mathbf{e}_{i^\star}, \delta\rangle - \frac{\gamma}{4}\|\delta\|^2_{H_p}\right] = \bar{a}_p - \mathbf{e}_{i^\star} - \frac{\gamma}{2}H_p\delta\,. \qquad (12)$$

For p on the boundary of $\Delta([K])$ (i.e. there exists $i \in [K]$ such that $p_i = 0$), the gradient is constant and the supremum has value $+\infty$. Hence, we only need to consider the case where p lies in the interior of $\Delta([K])$, which implies $H_p \succ 0$. In this case Eq. (12) is strongly convex in δ and the unique maximizer is given by $\delta^\star = \frac{2}{\gamma}H_p^{-1}(\bar{a}_p - \mathbf{e}_{i^\star})$. Hence, we can rewrite (11) as

$$\begin{aligned}
&\min_{p\in\Delta([K])}\max_{i^\star\in[K]}\max_{\delta\in\mathbb{R}^K}\langle \bar{a}_p - \mathbf{e}_{i^\star}, \hat{\theta}+\delta\rangle - \frac{\gamma}{4}\|\delta\|^2_{H_p} \\
&= \min_{\substack{p\in\Delta([K])\\ H_p\succ 0}}\max_{i^\star\in[K]}\langle \bar{a}_p - \mathbf{e}_{i^\star}, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - \mathbf{e}_{i^\star}\|^2_{H_p^{-1}} \\
&\geq \min_{\substack{p\in\Delta([K])\\ H_p\succ 0}}\mathbb{E}_{i^\star\sim p}\left[\langle \bar{a}_p - \mathbf{e}_{i^\star}, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - \mathbf{e}_{i^\star}\|^2_{H_p^{-1}}\right] \qquad (13)\\
&= \min_{\substack{p\in\Delta([K])\\ H_p\succ 0}}\mathbb{E}_{i^\star\sim p}\left[\frac{1}{\gamma}\left(\mathrm{tr}(H_pH_p^{-1}) - \|\bar{a}_p\|^2_{H_p^{-1}}\right)\right] = \frac{K-1}{\gamma}\,.
\end{aligned}$$

Now consider the inequality (13). If we can show that there exists a unique solution p such that this step in fact holds with equality, then we have identified the minimizer over $p \in \Delta([K])$. Consider an arbitrary candidate solution p on the interior of $\Delta([K])$. Then, letting $W_i := \langle \bar{a}_p - \mathbf{e}_{i^\star}, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - \mathbf{e}_{i^\star}\|^2_{H_p^{-1}}$, the step (13) lower bounds $\max_{i\in[K]} W_i$ by $\mathbb{E}_{i\sim p}[W_i]$. This step holds with equality if and only if $\mathbb{E}_{i\sim p}[W_i - \max_{i'\in[K]} W_{i'}] = 0$. Since all probabilities p_i are strictly positive, this can happen if and only if

$$\exists\tilde{\lambda}\in\mathbb{R} \quad \text{such that} \quad \forall i \in [K]: W_i = \langle \bar{a}_p - \mathbf{e}_i, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - \mathbf{e}_i\|^2_{H_p^{-1}} = \tilde{\lambda}\,.$$

Basic algebra shows that

$$\langle \bar{a}_p - \mathbf{e}_i, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - \mathbf{e}_{i^\star}\|^2_{H_p^{-1}} = \sum_{i'\in[K]} p_{i'}\hat{\theta}_i - \hat{\theta}_i - \frac{1}{\gamma} + \frac{1}{\gamma p_i} = \tilde{\lambda}\,.$$

Substituting $\tilde{\lambda} = \sum_{i'\in[K]} p_{i'}\hat{\theta}_i - \min_j\hat{\theta}_j - \frac{1}{\gamma} + \lambda\gamma$, rearranging and picking the unique value such that this is a probability distribution leads to the log-barrier distribution. □

C.2 Proofs from Section 3.2

Recall that $\dim(\mathcal{A})$ is the dimension of the smallest affine linear subspace containing $\mathcal{A}$. In other words $\forall a \in \mathcal{A}: \dim(\mathcal{A}) = \dim(\mathrm{span}(\mathcal{A} - a))$. Our main result in this section is the following slightly stronger version of Lemma 2.

Lemma 3. Any solution $p \in \Delta(\mathcal{A})$ to the problem $\mathsf{logdet\text{-}barrier}(\hat{\theta}, \gamma; \mathcal{A})$ in Eq. (6) satisfies

$$\max_{a^\star \in \mathcal{A}} \sup_{\theta \in \mathbb{R}^d} \langle \bar{a}_p - a^\star, \theta \rangle - \frac{\gamma}{4} \|\hat{\theta} - \theta\|^2_{H_p - \bar{a}_p \bar{a}_p^\top} \leq \gamma^{-1} \dim(\mathcal{A}) \,.$$

Since $-\|\hat{\theta}-\theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top} = -\|\hat{\theta}-\theta\|^2_{H_p} + \langle \hat{\theta}-\theta, \bar{a}_p\rangle^2 \geq -\|\hat{\theta}-\theta\|^2_{H_p}$, Lemma 2 is a direct corollary of Lemma 3.

Before proving Lemma 3, we discuss in detail the case where $\mathcal{A}$ does not span $\mathbb{R}^d$.

C.2.1 Handling the case where $\dim(\mathcal{A}) < d$.

We first show that if $\dim(\mathcal{A}) < d$, there exists a bijection of $\mathcal{A}$ to a set $\tilde{\mathcal{A}} \subset \mathbb{R}^{\dim(\mathcal{A})}$ and a projection P of the loss estimator into $\mathbb{R}^{\dim(\mathcal{A})}$, such that $\mathsf{logdet\text{-}barrier}(\theta, \gamma; \mathcal{A})$ and $\mathsf{logdet\text{-}barrier}(P(\theta), \gamma; \tilde{\mathcal{A}})$ are (up to the bijection) identical, and such that the objective in Lemma 3 coincides. This implies for all subsequent sections, we can assume w.l.o.g. that $\dim(\mathcal{A}) = d$, since if this does not hold we can work in the subspace outlined in this section.

We pick an arbitrary anchor $\mathfrak{a} \in \mathcal{A}$, let P be the projection onto $\mathrm{span}(\mathcal{A} - \mathfrak{a})$ represented in a fixed arbitrary orthonormal basis of $\mathrm{span}(\mathcal{A} - \mathfrak{a})$. Denote $\tilde{\mathcal{A}} = P(\mathcal{A} - \mathfrak{a})$ and for $p \in \Delta(\mathcal{A})$ let $\tilde{p} \in \Delta(\tilde{\mathcal{A}})$ be such that $\tilde{p}_{P(a-\mathfrak{a})} = p_a$ (recall that we define $\Delta(\mathcal{A})$ to have countable support). Let $\hat{\theta} \in \mathbb{R}^d$ be arbitrary, then

$$\langle \bar{a}_p, \hat{\theta} \rangle = \mathbb{E}_{a\sim p}\Big[\langle P(a - \mathfrak{a}), P(\hat{\theta})\rangle\Big] + \langle \mathfrak{a}, \hat{\theta}\rangle = \langle \bar{a}_{\tilde{p}}, P(\hat{\theta})\rangle + \langle \mathfrak{a}, \hat{\theta}\rangle \,.$$

Recall that we define the $\det$ in $\mathsf{logdet\text{-}barrier}$ as the product over the first $\dim(\mathcal{A})$ eigenvalues of $H_p - \bar{a}_p\bar{a}_p^\top$. Let $(\nu_i)_{i=1}^{\dim(\mathcal{A})}$ denote the corresponding eigenvectors (note that this requires $\nu_i \in \mathrm{span}(\mathcal{A} - \mathfrak{a})$). We have

$$\begin{aligned}
&\log\det(H_p - \bar{a}_p\bar{a}_p^\top) = \sum_{i=1}^{\dim(\mathcal{A})} \log(\|\nu_i\|^2_{H_p - \bar{a}_p\bar{a}_p^\top}) = \sum_{i=1}^{\dim(\mathcal{A})} \log(\mathbb{E}_{a\sim p}[\langle \nu_i, a - \bar{a}_p\rangle^2]) \\
&= \sum_{i=1}^{\dim(\mathcal{A})} \log(\mathbb{E}_{a\sim p}[\langle \nu_i, a - \mathfrak{a} - \mathbb{E}_{a'\sim p}(a' - \mathfrak{a})\rangle^2]) = \sum_{i=1}^{\dim(\mathcal{A})} \log(\mathbb{E}_{a\sim p}[\langle P(\nu_i), P(a - \mathfrak{a}) - \bar{a}_{\tilde{p}}\rangle^2]) \\
&= \sum_{i=1}^{\dim(\mathcal{A})} \log(\|P(\nu_i)\|^2_{H_{\tilde{p}} - \bar{a}_{\tilde{p}}\bar{a}_{\tilde{p}}^\top}) = \log\det(H_{\tilde{p}} - \bar{a}_{\tilde{p}}\bar{a}_{\tilde{p}}^\top) \,,
\end{aligned}$$

where we use the fact that P only changes the representation on $\mathrm{span}(\mathcal{A} - \mathfrak{a})$ and does not change the identity of the eigenvalues. Combining these two results immediately shows that for any $p \in \mathsf{logdet\text{-}barrier}(\hat{\theta}, \gamma; \mathcal{A})$ it follows that $\tilde{p} \in \mathsf{logdet\text{-}barrier}(P(\hat{\theta}), \gamma; \tilde{\mathcal{A}})$ and vice versa.

For the objective of Lemma 3, we have

$$\langle \bar{a}_p - a^\star, \theta \rangle = \langle \mathbb{E}_{a\sim p}[P(a - \mathfrak{a})] - P(a^\star - \mathfrak{a}), P(\theta)\rangle = \langle \bar{a}_{\tilde{p}} - (P(a^\star - \mathfrak{a})), P(\theta)\rangle.$$

For the quadratic term, following the same steps as above for ν_i, we have

$$\|\hat{\theta} - \theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top} = \|P(\hat{\theta}) - P(\theta)\|^2_{H_{\tilde{p}} - \bar{a}_{\tilde{p}}\bar{a}_{\tilde{p}}^\top} \,.$$

and

$$\langle \bar{a}_p - a^\star, \theta \rangle - \frac{\gamma}{4}\|\hat{\theta} - \theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top} = \langle \bar{a}_{\tilde{p}} - P(a^\star - \mathfrak{a}), P(\theta)\rangle - \frac{\gamma}{4}\|P(\hat{\theta}) - P(\theta)\|^2_{H_{\tilde{p}} - \bar{a}_{\tilde{p}}\bar{a}_{\tilde{p}}^\top} \,.$$

Hence, we have

$$\max_{a^\star \in \mathcal{A}} \sup_{\theta \in \mathbb{R}^d} \langle \bar{a}_p - a^\star, \theta \rangle - \frac{\gamma}{4}\|\hat{\theta} - \theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top} = \max_{\tilde{a}^\star \in \tilde{\mathcal{A}}} \sup_{\tilde{\theta} \in \mathbb{R}^{\dim(\mathcal{A})}} \langle \bar{a}_{\tilde{p}} - \tilde{a}^\star, \tilde{\theta}\rangle - \frac{\gamma}{4}\|P(\hat{\theta}) - \tilde{\theta}\|^2_{H_{\tilde{p}} - \bar{a}_{\tilde{p}}\bar{a}_{\tilde{p}}^\top} \,.$$

C.2.2 Handling the case where $\dim(\mathcal{A}) = d$.

Lemma 4. When $\dim(\mathcal{A}) = d$, any solution $p \in \Delta(\mathcal{A})$ to the problem $\textsf{logdet-barrier}(\theta, \gamma; \mathcal{A})$ in Eq. (6) satisfies

$$\forall a \in \mathcal{A}: \langle \bar{a}_p - a, \theta \rangle + \frac{1}{\gamma} \|\bar{a}_p - a\|^2_{H_p^{-1} - \bar{a}_p \bar{a}_p^\top} \leq \frac{\dim(\mathcal{A})}{\gamma}.$$

Proof. We first observe that any solution $p \in \Delta(\mathcal{A})$ to the problem $\textsf{logdet-barrier}(\hat{\theta}, \gamma; \mathcal{A})$ must be positive definite in the sense that $H_p - \bar{a}_p \bar{a}_p^\top \succ 0$, since otherwise the objective has value ∞; note that $\dim(\mathcal{A}) = d$ implies that there exists p with $H_p - \bar{a}_p \bar{a}_p^\top \succ 0$. Going forward we work only with p for which $H_p - \bar{a}_p \bar{a}_p^\top \succ 0$.

Recall $p = \textsf{logdet-barrier}(\hat{\theta}, \gamma; \mathcal{A})$ is any solution to

$$\operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \Big\{ \langle \bar{a}_p, \hat{\theta} \rangle - \gamma^{-1} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \Big\},$$

where $\Delta(\mathcal{A})$ is the set of distributions over countable subsets of $\mathcal{A}$. Hence we can write

$$\Delta(\mathcal{A}) = \left\{ \sum_{i=1}^{\infty} w_i \mathbf{e}_{A_i} \,|\, w \in \mathbb{R}_+^{\mathbb{N}}, A \in \mathcal{A}^{\mathbb{N}}, \sum_{i=1}^{\infty} w_i = 1 \right\},$$

where $\mathbf{e}_a$ denotes the distribution that selects a with probability 1. By first-order optimality, p is a solution to Eq. (6) if and only if

$$\forall p' \in \Delta(\mathcal{A}): \sum_{a \in \mathrm{supp}(p) \cup \mathrm{supp}(p')} (p'_a - p_a) \frac{\partial}{\partial p_a} \left[\langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \right] \geq 0\,.$$

By the K.K.T. conditions, this is the case if and only if there exists some $\tilde{\lambda} \in \mathbb{R}$ such that

$$\forall a \in \mathrm{supp}(p): \frac{\partial}{\partial p_a} \left[\langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \right] = \tilde{\lambda} \tag{14}$$

$$\forall a \in \mathcal{A}: \frac{\partial}{\partial p_a} \left[\langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \right] \geq \tilde{\lambda}\,. \tag{15}$$

To find $\tilde{\lambda}$, we calculate the partial derivative with the chain rule:

$$\begin{aligned}
&\frac{\partial}{\partial p_a} \left[\langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \right] \\
&= \langle a, \hat{\theta} \rangle - \frac{\det(H_p - \bar{a}_p \bar{a}_p^\top) \operatorname{tr}((H_p - \bar{a}_p \bar{a}_p^\top)^{-1}(aa^\top - \bar{a}_p a^\top - a \bar{a}_p^\top))}{\gamma \det(H_p - \bar{a}_p \bar{a}_p^\top)} \\
&= \langle a - \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \|a - \bar{a}_p\|^2_{(H_p - \bar{a}_p \bar{a}_p^\top)^{-1}} + \frac{1}{\gamma} \|\bar{a}_p\|^2_{(H_p - \bar{a}_p \bar{a}_p^\top)^{-1}} + \langle \bar{a}_p, \hat{\theta} \rangle\,.
\end{aligned}$$

Using Eq. (14) and taking the expectation over p yields

$$\tilde{\lambda} = \mathbb{E}_{a \sim p}\left[\frac{\partial}{\partial p_a} \left[\langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det(H_p - \bar{a}_p \bar{a}_p^\top) \right] \right] = -\frac{d}{\gamma} + \frac{1}{\gamma} \|\bar{a}_p\|^2_{(H_p - \bar{a}_p \bar{a}_p^\top)^{-1}} + \langle \bar{a}_p, \hat{\theta} \rangle\,.$$

Finally, plugging this into Eq. (15), we get

$$\forall a \in \mathcal{A}: \langle a - \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \|a - \bar{a}_p\|^2_{(H_p - \bar{a}_p \bar{a}_p^\top)^{-1}} \geq -\frac{d}{\gamma}\,.$$

Rearranging finishes the proof. □

Proof of Lemma 3. As mentioned in the previous proof, for any solution $p \in \Delta(\mathcal{A})$ to the problem $\textsf{logdet-barrier}(\hat{\theta}, \gamma; \mathcal{A})$ the matrix $H_p - \bar{a}_p \bar{a}_p^\top$ is positive definite. In this case, for any fixed $a^\star \in \mathcal{A}$, the function

$$\theta \mapsto \langle \bar{a}_p - a^\star, \theta \rangle - \frac{\gamma}{4} \|\hat{\theta} - \theta\|^2_{H_p - \bar{a}_p \bar{a}_p^\top}$$

is strictly concave in θ and the maximizer $\theta^\star$ is found by setting the derivative with respect to θ to 0:

$$\frac{\partial}{\partial\theta}\Big[\langle \bar{a}_p - a^\star, \theta\rangle - \frac{\gamma}{4}\|\hat{\theta}-\theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top}\Big] = \bar{a}_p - a^\star + \frac{\gamma}{2}(H_p - \bar{a}_p\bar{a}_p^T)(\hat{\theta}-\theta)$$

$$\theta^\star = \hat{\theta} + \frac{2}{\gamma}(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}(a_p - a^\star)\,.$$

Substituting in this choice, we have that

$$\max_{a^\star\in\mathcal{A}} \sup_{\theta\in\mathbb{R}^d} \langle \bar{a}_p - a^\star, \theta\rangle - \frac{\gamma}{4}\|\hat{\theta}-\theta\|^2_{H_p - \bar{a}_p\bar{a}_p^\top} = \max_{a^\star\in\mathcal{A}}\langle \bar{a}_p - a^\star, \hat{\theta}\rangle + \frac{1}{\gamma}\|\bar{a}_p - a\|^2_{(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}}\,.$$

To complete the proof, we apply Lemma 4 to the right-hand side above. □

C.3 Proofs from Section 3.3

Proof of Theorem 3. Let m be fixed. To keep notation compact, we abbreviate $q_t \equiv q_{t,h}$, $\rho_t \equiv \rho_{t,m}$, $\gamma_t \equiv \gamma_{t,m}$, $Z_t \equiv Z_{t,m}$, and so forth.

Let the sequence S be fixed, and let $f^\star$ be any predictor achieving the value of $\varepsilon_T(S)$. If the infimum is not achieved, we can consider a limit sequence; we omit the details. Recall that since we assume an oblivious adversary, $f^\star$ is fully determined before the interaction protocol begins. Let us abbreviate $\theta^\star_t = f^\star(x_t)$, $a^\star_t = \pi_{f^\star}(x_t)$, and $\pi^\star_t(x_t) = \operatorname{argmin}_{a\in\mathcal{A}_t}\mu(a, x_t)$, where ties are broken arbitrarily. Then we can bound

$$\begin{aligned}
\mathsf{Reg}_{\mathrm{Imp}}(T) &= \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\left(\mu(a_t, x_t) - \mu(\pi^\star_t(x_t), x_t)\right)\right]\\
&\le \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\left(\langle a_t - \pi^\star_t(x_t), \theta^\star_t\rangle + 2\max_{a\in\mathcal{A}_t}|\mu(a, x_t) - \langle a, \theta^\star_t\rangle|\right)\right]\\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\left(\langle a_t - \pi^\star_t(x_t), \theta^\star_t\rangle\right)\right] + 2\varepsilon_T T\\
&\overset{(b)}{\le} \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\langle a_t - a^\star_t, \theta^\star_t\rangle\right] + 2\varepsilon_T T\\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\left(\langle \bar{a}_{p_t} - a^\star_t, \theta^\star_t\rangle - \frac{\gamma_t}{4}\|\hat{\theta}_t - \theta^\star\|^2_{H_{p_t}} + \frac{\gamma_t}{4}\|\hat{\theta}_t - \theta^\star\|^2_{H_{p_t}}\right)\right] + 2\varepsilon_T T\\
&\overset{(d)}{\le} \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\left(\frac{\dim(\mathcal{A}_t)}{\gamma_t} + \frac{\gamma_t}{4}\|\hat{\theta}_t - \theta^\star\|^2_{H_{p_t}}\right)\right] + 2\varepsilon_T T\\
&\overset{(e)}{\le} \mathbb{E}\left[\max_{t\in[T]}\gamma_t^{-1}\right]\sum_{t=1}^T \dim(\mathcal{A}_t) + \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\frac{\gamma_t}{4}(\langle a_t, \hat{\theta}_t\rangle - \langle a_t, \theta^\star_t\rangle)^2\right] + 2\varepsilon_T T\,.
\end{aligned}$$

Here (a) follows from the fact that $\mathbb{E}[Z_t] = q_t$ and the Cauchy-Schwarz inequality, together with the definition of ε_T; (b) follows from the definition of the policy $\pi_{f^\star}$; (c) is due to the fact that, conditioned on $Z_t = 1$, we sample $a_t \sim p_t$ with $\mathbb{E}_{a_t\sim p_t}[a_t] = \bar{a}_{p_t}$; (d) uses Lemma 2; (e) uses $\mathbb{E}_{a_t\sim p_t}[a_t a_t^\top] = H_{p_t}$. Continuing with squared error term above, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\gamma_t(\langle a_t, \hat{\theta}_t\rangle - \langle a_t, \theta^\star_t\rangle)^2\right]\\
&= \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\gamma_t\Big((\langle a_t, \hat{\theta}_t\rangle - \ell_t)^2 - (\langle a_t, \theta^\star_t\rangle - \ell_t)^2 + 2(\ell_t - \langle a_t, \theta^\star_t\rangle)\langle a_t, \hat{\theta}_t - \theta^\star_t\rangle\Big)\right]\\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^T \frac{Z_t}{q_t}\gamma_t\Big((\langle a_t, \hat{\theta}_t\rangle - \ell_t)^2 - (\langle a_t, \theta^\star_t\rangle - \ell_t)^2 + 2(\mu(a_t, x_t) - \langle a_t, \theta^\star_t\rangle)\langle a_t, \hat{\theta}_t - \theta^\star_t\rangle\Big)\right],
\end{aligned}$$

where (a) uses that ℓ_t is conditionally independent of Z_t and a_t. We bound the term involving the difference of squares as

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t((\langle a_t,\hat{\theta}_t\rangle-\ell_t)^2-(\langle a_t,\theta_t^\star\rangle-\ell_t)^2)\right]\le\mathbb{E}\left[\max_{t\in[T]}\frac{\gamma_t}{q_t}\right]\mathsf{Reg}_{\mathrm{Sq}}(T),$$

by Assumption 2, which also holds if SqAlg runs for less than T timesteps, since we could extend the sequence with 0 weight until time T. For the linear term, we apply the sequence of inequalities

$$\begin{aligned}
&2\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t(\mu(a_t,x_t)-\langle a_t,\theta_t^\star\rangle)\langle a_t,\hat{\theta}_t-\theta_t^\star\rangle\right]\\
&\overset{(a)}{\le} 2\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t((\mu(a_t,x_t)-\langle a_t,\theta_t^\star\rangle)^2+\frac{1}{4}\langle a_t,\hat{y}_t-\theta_t^\star\rangle^2\right]\\
&\le 2\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t\max_{a\in\mathcal{A}_t}((\mu(a,x_t)-\langle a,\theta_t^\star\rangle)^2+\frac{1}{4}\langle a_t,\hat{y}_t-\theta_t^\star\rangle^2\right]\\
&\overset{(b)}{\le} 2\mathbb{E}\left[\max_{t\in[T]}\gamma_t\right]\varepsilon_T^2T+\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t(\langle a_t,\hat{\theta}_t\rangle-\langle a_t,\theta_t^\star\rangle)^2\right],
\end{aligned}$$

where (a) is by the AM-GM inequality: $2ab\le 2a^2+\frac{1}{2}b^2$; (b) follows from the fact that Z_t is conditionally independent of γ_t, and the definition of ε_T.

Altogether, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t(\langle a_t,\hat{\theta}_t\rangle-\langle a_t,\theta_t^\star\rangle)^2\right]\\
&\le\mathbb{E}\left[\max_{t\in[T]}\frac{\gamma_t}{q_t}\right]\mathsf{Reg}_{\mathrm{Sq}}(T)+2\mathbb{E}\left[\max_{t\in[T]}\gamma_t\right]\varepsilon_T^2T+\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t(\langle a_t,\hat{\theta}_t\rangle-\langle a_t,\theta_t^\star\rangle)^2\right].
\end{aligned}$$

Rearranging yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{Z_t}{q_t}\gamma_t(\langle a_t,\hat{\theta}_t\rangle-\langle a_t,\theta_t^\star\rangle)^2\right]\le 2\mathbb{E}\left[\max_{t\in[T]}\frac{\gamma_t}{q_t}\right]\mathsf{Reg}_{\mathrm{Sq}}(T)+4\mathbb{E}\left[\max_{t\in[T]}\gamma_t\right]\varepsilon_T^2T\,.$$

Combining all of the developments so far, we have

$$\mathsf{Reg}_{\mathrm{Imp}}(T)\le\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t^{-1}\right]\dim(\mathcal{A}_t)+\frac{1}{2}\mathbb{E}\left[\max_{t\in[T]}\frac{\gamma_t}{q_t}\right]\mathsf{Reg}_{\mathrm{Sq}}(T)+\mathbb{E}\left[\max_{t\in[T]}\gamma_t\right]\varepsilon_T^2T+2\varepsilon_TT\,. \tag{16}$$

The proof is completed by noting that the learning rate $\gamma_t=\min\left\{\frac{\sqrt{d}}{\varepsilon'},\sqrt{dT/(\rho_t\mathsf{Reg}_{\mathrm{Sq}}(T))}\right\}$ is non-increasing, but $\gamma_t\rho_t\ge\frac{\gamma_t}{q_t}$ is non-decreasing. Hence, we can upper bound the expression above by

$$\begin{aligned}
\mathsf{Reg}_{\mathrm{Imp}}(T)&\le\mathbb{E}\left[\gamma_T^{-1}\right]dT+\frac{1}{2}\mathbb{E}\left[\gamma_T\rho_T\right]\mathsf{Reg}_{\mathrm{Sq}}(T)+\mathbb{E}[\gamma_1]\varepsilon_T^2T+2\varepsilon_TT\\
&\le\left(\frac{\varepsilon'}{\sqrt{d}}+\mathbb{E}[\sqrt{\rho_T}]\sqrt{\frac{\mathsf{Reg}_{\mathrm{Sq}}(T)}{dT}}\right)dT+\frac{1}{2}\mathbb{E}[\sqrt{\rho_T}]\sqrt{dT\mathsf{Reg}_{\mathrm{Sq}}(T)}+\frac{\sqrt{d}}{\varepsilon'}\varepsilon_T^2T+2\varepsilon_TT\,.
\end{aligned}$$

□

Proof of Theorem 1. Let $m^\star:=\operatorname{argmin}_{m\in[M]}\frac{\varepsilon_T}{\varepsilon'_m}+\frac{\varepsilon'_m}{\varepsilon_T}$ if $\varepsilon_T\ge T^{-1}$ and $m^\star=M$ otherwise. We begin by formally verifying the claim

$$\mathsf{Reg}(T)=\mathbb{E}\left[\sum_{t=1}^{T}\tilde{\ell}_{t,A_t}-\tilde{\ell}_{t,m^\star}\right]+\mathsf{Reg}_{\mathrm{Imp}}^{m^\star}(T)\,. \tag{17}$$

By the definition $\tilde{\ell}_{t,A_t} := \ell_t + 1$, we have

$$\mathbb{E}\left[\tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,m^\star}\right] = \mathbb{E}\left[\ell_t + 1 - \frac{Z_{t,m^\star}}{p_{t,m^\star}}(\ell_t + 1)\right] = \mathbb{E}\left[\mu(a_t, x_t) - \frac{Z_{t,m^\star}}{p_{t,m^\star}}\mu(a_t, x_t)\right].$$

The second term is

$$\mathsf{Reg}^{m^\star}_{\mathrm{Imp}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \frac{Z_{t,m^\star}}{p_{t,m^\star}}(\mu(a_t, x_t) - \mu(\pi^\star_t(x_t), x_t))\right] = \mathbb{E}\left[\sum_{t=1}^{T} \frac{Z_{t,m^\star}}{p_{t,m^\star}}\mu(a_t, x_t) - \mu(\pi^\star_t(x_t), x_t)\right].$$

Combining both lines leads to Eq. (17). The losses $\tilde{\ell}$ satisfy $\forall m \in [M] : \tilde{\ell}_{t,m} \in [0, 2]$, since $\ell_t \in [-1, 1]$ and we shift the loss by 1. Hence we can apply Corollary 2 with $\alpha = \frac{1}{2}$ and $R = \frac{3}{2}\sqrt{dT\mathsf{Reg}_{\mathrm{Sq}}(T)}$ to obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,m^\star}\right] \leq 4\sqrt{2MT} + 3\sqrt{dT\mathsf{Reg}_{\mathrm{Sq}}(T)M} - \frac{3}{2}\mathbb{E}[\sqrt{\rho_{T,a^\star}}]\sqrt{dT\mathsf{Reg}_{\mathrm{Sq}}(T)},$$

and by Theorem 3,

$$\mathsf{Reg}^{m^\star}_{\mathrm{Imp}}(T) \leq \left(\left(\frac{\varepsilon'_{m^\star}}{\varepsilon_T} + \frac{\varepsilon_T}{\varepsilon'_{m^\star}}\right)\sqrt{d} + 2\right)\varepsilon_T T + \frac{3}{2}\mathbb{E}[\sqrt{\rho_{T,a^\star}}]\sqrt{dT\mathsf{Reg}_{\mathrm{Sq}}(T)}.$$

Either $\varepsilon_T > T^{-1}$, in which case we can pick $m^\star$ such that $\varepsilon'_{m^\star} \in [\varepsilon_T, e\varepsilon_T]$ and $\left(\frac{\varepsilon'_{m^\star}}{\varepsilon_T} + \frac{\varepsilon_T}{\varepsilon'_{m^\star}}\right) \leq e + e^{-1}$, or we pick $\varepsilon'_{m^\star} = T^{-1}$ and the misspecification term is bounded by

$$\left(\left(\frac{\varepsilon'_{m^\star}}{\varepsilon_T} + \frac{\varepsilon_T}{\varepsilon'_{m^\star}}\right)\sqrt{d} + 2\right)\varepsilon_T T = \left(\left(\varepsilon'_{m^\star} + \frac{\varepsilon_T^2}{\varepsilon'_{m^\star}}\right)\sqrt{d} + 2\varepsilon_T\right)T \leq 2\sqrt{d} + 2.$$

Summing the regret bounds for the base and master algorithms completes the proof. □

C.4 Proofs from Section 3.6

The procedure runs in episodes. At the begin of episode 1, the algorithm assumes $D_1 = \sum_{t=1}^{T} \dim(\mathcal{A}_t) \leq 2T$ and initializes its learning rate accordingly. Within each episode $i \geq 1$, if the agent observes at time t that $\sum_{s=\tau_i}^{t} \dim(\mathcal{A}_s) > D_i$, it restarts the algorithm with $D_{i+1} = 2D_i$; we denote this time by $\tau_{i+1} = t$. We can assume $\dim(\mathcal{A}) \leq d < T$ or the result is trivial, hence we never need to double more than once at each time step. For technical reasons, we require that the bound of Theorem 1 also holds when the agent plays only on a subset of time steps.

Corollary 3. *Let $\mathcal{T} \subset [T]$ be a subset of the time horizon chosen oblivious to the actions of the agent. Then the upper bound of Theorem 1 for* SquareCB.Imp *on the sequence S is also an upper bound for running the algorithm on the sub-sequence $S_{\mathcal{T}}$. This also holds for any refinement of the bound based on the average dimension d_{avg} instead of d.*

Proof. We extend the sequence S by adding an "end" sequence $E = (\{0\}, \mathfrak{x})_{t=1}^{T}$, where $\mathfrak{x} \in \mathcal{X}$ is picked such that $\mu(0, \mathfrak{x}) = 0$ (If there is no such context, we add a context with that property to $\mathcal{X}$). Let the enhanced sequence be $S' = S + E$ and consider the sequence $\tilde{S} = S'_{\mathcal{T}\cup\{T+1,\ldots,2T-|\mathcal{T}|\}}$, which is of length T. The regret contribution from playing on the E section on the sequence is always 0, since there is only one action. Furthermore $\varepsilon_T(\tilde{S}) \leq \varepsilon_T(S)$ Hence we have by Theorem 1

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}} \mu(a_t, x_t) - \min_{a\in\mathcal{A}_t} \mu(a, x_t)\right] = \mathbb{E}\left[\sum_{t\in\mathcal{T}\{T+1,\ldots,2T-|\mathcal{T}|\}} \mu(a_t, x_t) - \min_{a\in\mathcal{A}_t} \mu(a, x_t)\right]$$
$$\leq \mathcal{O}\left(\sqrt{d}\varepsilon_T(S)T + \sqrt{d\mathsf{Reg}_{\mathrm{Sq}}(T)\log(T)}\right).$$

Since $d_{\mathrm{avg}}(\tilde{S}) \leq d_{\mathrm{avg}}(S)$, this argument also extends to the refined version where d is replaced by d_{avg} if the algorithm parameters are tuned accordingly. □

Proof of Theorem 4. Let $\tau_1, \ldots, \tau_L$ denote the times where the algorithm is restarted, with $\tau_1 = 1$ and $\tau_{L+1} = T + 1$ by convention.

Since the adversary fixes the action sets in advance, these doubling times are deterministic. The regret is given by

$$\begin{aligned}\mathsf{Reg}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \mu(a_t, x_t) - \min_{a \in \mathcal{A}_t} \mu(a, x_t)\right] \\ &\leq \sum_{i=1}^{L} \mathbb{E}\left[\sum_{t=\tau_i}^{\tau_{i+1}-1} \mu(a_t, x_t) - \min_{a \in \mathcal{A}_t} \mu(a, x_t)\right].\end{aligned}$$

By Corollary 3 applied to each episode

$$\mathbb{E}\left[\sum_{t=\tau_{i-1}+1}^{\tau_i} \mu(a_t, x_t) - \min_{a \in \mathcal{A}_t} \mu(a, x_t)\right] = \sqrt{2^i} \cdot \mathcal{O}\Big(\varepsilon_T(S)T + \sqrt{\mathsf{Reg}_{\mathrm{Sq}}(T)\log(T)}\Big).$$

Summing over these terms and observing that

$$\sum_{i=1}^{L} 2^{i/2} = \mathcal{O}(2^{L/2}) = \mathcal{O}(1) \cdot \sqrt{\frac{1}{T}\sum_{t=1}^{T} \dim(\mathcal{A}_t)} = \mathcal{O}\Big(d_{\mathrm{avg}}^{1/2}\Big),$$

completes the proof. □

D Improved Master Algorithms for Bandit Aggregation

In this section, we present a general class of algorithms that can be used for the master algorithm within the framework of Algorithm 3. For the remainder of this section, we are working in a generic adversarial multi-armed bandit setting, where the agent selects an action $A_t \in [M]$ at any time step and observes the associated loss $\ell_{t,A_t} \in [0, L]$. Compared to the original CORRAL algorithm of Agarwal et al. [9], our new algorithms are simpler to analyze, more flexible, and improve in terms of logarithmic factors.

The CORRAL algorithm is a special case of Algorithm 3 that uses a bandit variant of Online Mirror Descent (OMD) algorithm with log-barrier regularization as the master.[6] The bandit variant of the OMD algorithm used within CORRAL is parameterized by a Legendre potential $F(x) = \sum_{i=1}^{d} \eta_i^{-1} f(x_i)$ where $\eta_1, \ldots, \eta_d$ are per-coordinate learning rates. It initializes the distribution $p_1 = \operatorname{argmin}_{p \in \Delta([M])} F(p)$. At each time t, the bandit OMD algorithm samples arm $A_t \sim p_t$, observes ℓ_t, and constructs an unbiased importance-weighted loss estimator $\hat{\ell}_t = \frac{\ell_{t,A_t}}{p_{t,A_t}} \mathbf{e}_{A_t}$. It then updates the action distribution as

$$p_{t+1} = \operatorname{argmin}_{p \in \Delta([M])} \langle p, \hat{\ell} \rangle + D_F(p, p_t), \tag{18}$$

where $D_F(x, y) := F(x) - F(y) - \langle x - y, \nabla F(y) \rangle$ is the Bregman divergence associated with F.

A key to the performance of the CORRAL master is an time-dependent learning rate[7] schedule for each of the per-arm learning rates, which increases the learning rate for each arm whenever the probability for that arm falls below a certain threshold.

An algorithm closely related to OMD is the Follow-the-Regularized-Leader (FTRL) algorithm. In particular, for any sequence of loss vector estimates $(\hat{\ell}_t)_{t=1}^T$, there exists a sequence of (vector) biases b_t, such that FTRL running on the loss sequence $(\hat{\ell}_t - b_t)_{t=1}^T$ using the same learning rate as its OMD counterpart has an identical trajectory of plays p_t.

We can view the CORRAL master through the lens of FTRL: the algorithm performs two steps whenever it increases the learning rate of arm i. First it subtracts a bias $b_{t,i} > 0$ from the loss

[6]Note that the use of the log-barrier in CORRAL is not related to our use of the log-barrier within the contextual bandit framework.

[7]For time-dependent learning rates, replace η by η_t in the update rule of Eq. (18).

estimates for arm i. Then it increases the learning rate for that arm. We argue that only the former step is actually relevant to the performance of CORRAL, while the latter is unnecessary, and ends up complicating the analysis. This motivates the (α, R)-hedged FTRL algorithm, which achieves a slightly improved guarantee by removing the per-coordinate learning rates.

D.1 The Hedged FTRL Algorithm

Following the intuition in the prequel we propose (α, R)-hedged FTRL, a modified variant of the FTRL algorithm with strong guarantees for aggregating bandit algorithms. We begin by defining a basic bandit variant of FTRL algorithm.

The FTRL family algorithms of algorithms is parameterized by a potential F and learning rate $\eta > 0$. At each round t, the algorithm selects

$$p_t = \operatorname{argmin}_{p\in\Delta([M])} \langle p, \hat{L}_{t-1}\rangle + \eta^{-1}F(p)\,, \quad \text{where} \quad \hat{L}_t = \textstyle\sum_{s=1}^t \hat{\ell}_s\,.$$

Two relevant properties of F that arise in our analysis are *stability* and *diameter*. Define

$$\bar{F}^\star_\eta(-L) = \max_{p\in\Delta([M])} \langle p, -L\rangle - \eta^{-1}F(p)\,.$$

The stability $\operatorname{stab}(F)$ and diameter $\operatorname{diam}(F)$ of F for a loss range $[0, L]$ are define as follows:

$$\operatorname{stab}(F) = \sup_{\eta>0} \sup_{x\in\Delta([M])} \sup_{\ell\in[0,L]^M} \eta^{-1}\mathbb{E}_{A\sim x}\left[D_{\bar{F}^\star_\eta}\left(\eta^{-1}\nabla F(x) - \frac{\ell_A}{x_A}\mathbf{e}_A, \eta^{-1}\nabla F(x)\right)\right],$$
$$\operatorname{diam}(F) = \max_{p\in\Delta([M])} F(p) - \min_{p\in\Delta([M])} F(p)\,.$$

Given a potential with bounded $\operatorname{stab}(F)$ and $\operatorname{diam}(F)$, setting the learning rate as $\eta = \sqrt{\operatorname{diam}(F)/(\operatorname{stab}(F)T)}$ leads to a regret bound $2\sqrt{\operatorname{stab}(F)\operatorname{diam}(F)T}$ for FTRL [5].[8] Well-known algorithms that arise as special cases of this result include:

- EXP3 [13] is an instance of FTRL with $F(x) = \sum_{i=1}^M x_i \log(x_i)$, $\operatorname{diam}(F) = \log(M)$ and $\operatorname{stab}(F) \le \frac{L^2M}{2}$.
- Tsallis-INF [11, 5, 48] is the instance of FTRL with the best known regret bound. It is given by $F(x) = -2\sum_{i=1}^M \sqrt{x_i}$ with $\operatorname{diam}(F) \le 2\sqrt{M}$ and $\operatorname{stab}(F) \le L^2\sqrt{M}$.

We can now present the (α, R)-hedged FTRL algorithm. The algorithm augments the basic FTRL strategy using an additional pair of parameters $(\alpha, R) \in (0, 1) \times \mathbb{R}$. The algorithm initializes $(B_{0,i})_{i=1}^M$ with $B_{0,i} = \rho^\alpha_{1,i}R$. At each step t, it plays $A_t \sim p_t$ and computes

$$\tilde{p}_{t+1} = \operatorname{argmin}_{p\in\Delta([M])} \langle p, \hat{L}_t - (B_{t-1} - B_0)\rangle + \eta^{-1}F(p)\,, \quad \text{where} \quad \hat{L}_t = \textstyle\sum_{s=1}^t \hat{\ell}_s\,.$$

If $\tilde{p}^{-\alpha}_{t+1,A_t}R \le B_{t-1,A_t}$, the algorithm sets $B_t = B_{t-1}$ and $p_{t+1} = \tilde{p}_{t+1}$. Otherwise it chooses the unique $b_t > 0$, such that for $B_t = B_{t-1} + b_t\mathbf{e}_{A_t}$ it holds simultaneously

$$p_{t+1} = \operatorname*{argmin}_{p\in\Delta([M])} \langle p, \hat{L}_t - (B_t - B_0)\rangle + \eta^{-1}F(p) \qquad \text{and} \qquad p^{-\alpha}_{t+1,A_t}R = B_{t,A_t}\,.$$

This algorithm is always well defined (see Appendix D.2 for details). The main regret guarantee is as follows. Let $\rho_{t,i} = \max_{s\le t} p^{-1}_{s,i}$.

Theorem 6. *Then for any potential F with $\operatorname{stab}(F), \operatorname{diam}(F) < \infty$, the pseudo-regret $\mathsf{Reg}_M(T) = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t} - \ell_{t,a^\star}\right]$ of (α, R)-hedged FTRL run with learning rate $\eta = \sqrt{\operatorname{diam}(F)/(\operatorname{stab}(F)T)}$ against any arm $a^\star \in [M]$ is bounded as follows:*

$$\mathsf{Reg}_M(T) \le 2\sqrt{\operatorname{stab}(F)\operatorname{diam}(F)T} + \left[\frac{\alpha}{1-\alpha}\sum_{i=1}^M \left(\rho^{\alpha-1}_{1,i} - \mathbb{E}[\rho^{\alpha-1}_{T,i}]\right) + \rho^\alpha_{1,a^\star} - \mathbb{E}[\rho^\alpha_{T,a^\star}]\right]\cdot R\,.$$

The algorithm may be viewed "hedging" against the event that the arm $a^\star$ experiences a very small probability, as this guarantees a negative regret contribution of $\rho^{-\alpha}_{T,a^\star}R$.

[8] Abernethy et al. [5] present this result slightly differently. See our proof of Theorem 6 with $R = 0$ for an alternative.

D.2 Proofs

Before proving the main result, we first established that the (α, R)-hedged FTRL is well-defined. The algorithm initializes with B_0 such that $\nabla \bar{F}^\star_\eta(B_0)_i^{-\alpha} R = B_{0,i}$. For symmetric potentials $F(x) = \sum_{i=1}^M f(x_i)$, $\nabla \bar{F}^\star_\eta(c\mathbf{1}_M) = \frac{1}{M}\mathbf{1}_M$ for any $c \in \mathbb{R}$. Hence $B_0 = M^{-\alpha}R\mathbf{1}_M$ satisfies the initialization condition. Otherwise a solution exists by the observation that $\nabla \bar{F}^\star_\eta(B_0)_i^{-\alpha} R$ is a continuous, decreasing function in $B_{0,i}$ that has positive values at $B_0 = 0$. Hence a solution to the equation must exist.

The same argument holds during the update at subsequent rounds t. Only the arm that was played can decrease in probability, which means we only need to ensure that $\rho^\alpha_{t+1,A_t} R = B_{t,A_t}$. The LHS is continuously decreasing with increasing b_t, while the RHS is increasing. The optimal value must exist, it is unique and lays in $[0, \hat{\ell}_{t,A_t}]$.

Proof of Theorem 6. We follow the standard FTRL analysis. Let $\tilde{B}_t = B_t - B_0$ and note that $p_t = \nabla \bar{F}^\star_\eta(-\hat{L}_{t-1} + \tilde{B}_{t-1})$, so $\langle p_t, \hat{\ell}_t\rangle = \langle \nabla \bar{F}^\star_\eta(-\hat{L}_{t-1} + \tilde{B}_{t-1}), \hat{L}_t - \hat{L}_{t-1}\rangle$. Hence, we can write

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t} - \ell_{t,a^\star}\right] &= \mathbb{E}\left[\sum_{t=1}^T \langle p_t, \hat{\ell}_t\rangle - \hat{\ell}_{t,a^\star}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T D_{\bar{F}^\star_\eta}(-\hat{L}_t + \tilde{B}_{t-1}, -\hat{L}_{t-1} + \tilde{B}_{t-1})\right] \\
&\qquad + \mathbb{E}\left[\sum_{t=1}^T \Big(-\bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_{t-1}) + \bar{F}^\star_\eta(-\hat{L}_{t-1} + \tilde{B}_{t-1}) - \hat{\ell}_{t,a^\star}\Big)\right].
\end{aligned}
$$

Note that there exists λ such that $-\hat{L}_{t-1} + \tilde{B}_{t-1} = \lambda\mathbf{1}_M + \eta^{-1}\nabla F(p_t)$. Furthermore, adding or subtracting the same $\lambda\mathbf{1}_M$ term to both arguments does not change the value of the Bregman divergence, because $\bar{F}_\eta(-L + \lambda\mathbf{1}_M) = F_\eta(-L) + \lambda$. Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T D_{\bar{F}^\star_\eta}(-\hat{L}_t + \tilde{B}_{t-1}, -\hat{L}_{t-1} + \tilde{B}_{t-1})\right] \\
&= \eta\mathbb{E}\left[\sum_{t=1}^T \eta^{-1} D_{\bar{F}^\star_\eta}(\eta^{-1}\nabla F(p_t) - \hat{\ell}_t, \eta^{-1}\nabla F(p_t))\right] \le \eta\,\mathrm{stab}(F)T\,.
\end{aligned}
$$

Rearranging the second term gives

$$
\begin{aligned}
&\sum_{t=1}^T \Big(-\bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_{t-1}) + \bar{F}^\star_\eta(-\hat{L}_{t-1} + \tilde{B}_{t-1}) - \hat{\ell}_{t,a^\star}\Big) \\
&= \bar{F}^\star_\eta(0) - \bar{F}^\star_\eta(-\hat{L}_T + \tilde{B}_{T-1}) - \hat{L}_{T,a^\star} + \sum_{t=1}^{T-1} \bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_t) - \bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_{t-1})\,.
\end{aligned}
$$

Note that $\bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_t) = \langle p_{t+1}, -\hat{L}_t + \tilde{B}_t\rangle + \eta^{-1}F(p_{t+1})$. Furthermore we have the bounds

$$
-\bar{F}^\star_\eta(-\hat{L}_T + \tilde{B}_{T-1}) \le -\Big(\langle \mathbf{e}_{a^\star}, -\hat{L}_T + \tilde{B}_{T-1}\rangle - \eta^{-1}F(\mathbf{e}_{a^\star})\Big)\,,
$$

and

$$
-\bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_{t-1}) \le -\Big(\langle p_{t+1}, -\hat{L}_t + \tilde{B}_{t-1}\rangle - \eta^{-1}F(p_{t+1})\Big)\,.
$$

Plugging these in leads to

$$\begin{aligned}
&\bar{F}^\star_\eta(0) - \bar{F}^\star_\eta(-\hat{L}_T + \tilde{B}_{T-1}) - \hat{L}_{T,a^\star} + \sum_{t=1}^{T-1} \bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_t) - \bar{F}^\star_\eta(-\hat{L}_t + \tilde{B}_{t-1}) \\
&\leq \frac{F(\mathbf{e}_{a^\star}) - F(p_1)}{\eta} - \tilde{B}_{T-1,a^\star} + \sum_{t=1}^{T-1} \langle p_{t+1}, \tilde{B}_t - \tilde{B}_{t-1}\rangle \\
&\leq (\rho^\alpha_{1,a^\star} - \rho^\alpha_{T,a^\star})R + \frac{\operatorname{diam}(F)}{\eta} + \sum_{t=1}^{T-1} \langle p_{t+1}, B_t - B_{t-1}\rangle \,.
\end{aligned}$$

To bound the final sum above, note that for each coordinate i, the difference $B_{t,i} - B_{t-1,i}$ can be non-zero only if $p_{t+1,i}$ achieves $p_{t+1,i} = \rho^{-1}_{t+1,i}$. Therefore, we have

$$\begin{aligned}
p_{t+1,i}(B_{t,i} - B_{t-1,i}) &= R\rho^{-1}_{t+1,i}\left(\rho^\alpha_{t+1,i} - \rho^\alpha_{t,i}\right) \\
&= \alpha R \int_{\rho_{t,i}}^{\rho_{t+1,i}} x^{\alpha-1}\rho^{-1}_{t+1}\, dx \\
&\leq \alpha R \int_{\rho_{t,i}}^{\rho_{t+1,i}} x^{\alpha-2}\, dx \\
&= \frac{\alpha R}{1-\alpha}(\rho^{\alpha-1}_{t,i} - \rho^{\alpha-1}_{t+1,i}) \,.
\end{aligned}$$

Applying this bound to each coordinate, we have

$$\sum_{t=1}^{T-1} \langle p_{t+1}, B_t - B_{t-1}\rangle = \sum_{i=1}^{M} \frac{\alpha R}{1-\alpha}\left(\rho^{\alpha-1}_{1,i} - \rho^{\alpha-1}_{T,i}\right) = \sum_{i=1}^{M} \frac{\alpha R}{1-\alpha}\left(\rho^{\alpha-1}_{1,i} - \rho^{\alpha-1}_{T,i}\right) \,.$$

Combining all of the bounds so far concludes the proof. □

Proof of Corollary 2. The Tsallis regularizer is

$$F(x) = -\sum_{i=1}^{M} 2\sqrt{x_i}\,,$$

with a stability for the loss range $[0, L]$ of $L^2\sqrt{M}$ and a diameter of $2\sqrt{M}$[48][9]. Due to the symmetry of the potential, we have $\forall i:\ p_{1,i} = 1/M$. Using Theorem 6 with the loss range $[0, 2]$ leads to

$$\begin{aligned}
\mathsf{Reg}_M(T) &\leq 4\sqrt{2MT} + \left[\frac{\alpha}{1-\alpha}\sum_{i=1}^{M}(M^{\alpha-1} - \mathbb{E}[\rho^{\alpha-1}_{T,i}]) + M^\alpha - \mathbb{E}[\rho^\alpha_{T,m^\star}]\right] R \\
&\leq 4\sqrt{2MT} + \left[\frac{\alpha}{1-\alpha}M^\alpha\left(1 - M^{1-\alpha}\min_{j\in[M]}\mathbb{E}[\rho^{\alpha-1}_{T,j}]\right) + M^\alpha - \mathbb{E}[\rho^\alpha_{T,m^\star}]\right] R \,.
\end{aligned}$$

Dropping the negative $-M^{1-\alpha}\min_{j\in[M]}\mathbb{E}[\rho^{\alpha-1}_{T,j}]$ term leads to the first part of the $\min\{\cdot\}$ expression in Eq. (10). For the other term in the $\min\{\cdot\}$, note that the function

$$\alpha \mapsto \frac{\alpha}{1-\alpha}\left(1 - z^{\alpha-1}\right)$$

is monotonically increasing in α with

$$\lim_{\alpha\to 1}\frac{\alpha}{1-\alpha}\left(1 - z^{\alpha-1}\right) = \log(z)\,.$$

Absorbing $\log(\max_{j\in[M]}\mathbb{E}[\rho_{T,j}]/M) + 1$ by $2\log(\max_{j\in[M]}\mathbb{E}[\rho_{T,j}])$ (using that $\rho_{1,i} = M$) completes the proof. □

[9]This has been shown for $L = 1$ but the extension to general L is trivial.

E Approximation Algorithms for the Log-Determinant Barrier Problem

Recall that at every step, SquareCB.Inf (Algorithm 2) needs to sample from any distribution in logdet-barrier$(\hat{\theta}, \gamma; \mathcal{A})$, which is defined as

$$p^\star \in \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \gamma \langle \bar{a}_p, \hat{\theta} \rangle - \frac{1}{\gamma} \log\det\left(H_p - \bar{a}_p \bar{a}_p^\top\right) , \tag{19}$$

where $\bar{a}_p = \mathbb{E}_{a\sim p}[a]$ and $H_p = \mathbb{E}_{a\sim p}[aa^\top]$. In this section, we develop optimization algorithms to efficiently find approximate solutions to the problem Eq. (19). In particular, our main result will be to prove Proposition 1.

While this is a convex optimization problem, developing efficient algorithms presents a number of technical difficulties. First, the optimization problem is non-smooth due to the presence of the log-determinant function, which prevents us from applying standard first-order methods such as gradient descent out of the box. Second, representing distributions in $\Delta(\mathcal{A})$ naively requires $\Omega(|\mathcal{A}|)$ memory. To get the result in Proposition 1, we employ a specialized Frank-Wolfe-type method, which maintains a sparse distribution and requires only $\mathcal{O}(\log|\mathcal{A}|)$ memory.

As a first step toward solving the problem numerically, we move to an equivalent but slightly more convenient formulation which lifts the actions to $d+1$ dimensions. Define the *lifting* operator, which adds a new coordinate with 1 to each vector, by

$$\tilde{a} := \begin{pmatrix} a \\ 1 \end{pmatrix},$$

and define

$$\tilde{a}_p := \mathbb{E}_{a\sim p}[\tilde{a}], \quad \tilde{H}_p := \mathbb{E}_{a\sim p}\left[\tilde{a}\tilde{a}^\top\right], \quad \tilde{\theta} := \begin{pmatrix} \hat{\theta} \\ 0 \end{pmatrix}, \quad \text{and} \quad \tilde{d} := d+1 .$$

Furthermore, we define

$$G(p) = \langle \tilde{a}_p, \tilde{\theta} \rangle - \frac{1}{\gamma} \log\det(\tilde{H}_p). \tag{20}$$

Proposition 2. *The set of solutions for the lifted problem*

$$\operatorname*{argmin}_{p\in\Delta(\mathcal{A})} G(p) = \operatorname*{argmin}_{p\in\Delta(\mathcal{A})} \langle \tilde{a}_p, \tilde{\theta} \rangle - \frac{1}{\gamma} \log\det(\tilde{H}_p) , \tag{21}$$

is identical to the set of solutions for Eq. (19), and vice-versa.

Proof. By Lemma 4, any solution $p^\star$ to Eq. (19) must satisfy the optimality condition

$$\forall a \in \mathcal{A} \colon \langle \bar{a}_{p^\star} - a, \hat{\theta} \rangle + \frac{1}{\gamma} \|\bar{a}_{p^\star} - a\|^2_{(H_{p^\star} - \bar{a}_{p^\star}\bar{a}_{p^\star}^\top)^{-1}} \le \frac{d}{\gamma} .$$

Now, let $p^\star$ be a minimizer for the optimization problem in (21). By first order optimality, we have

$$\forall p' \in \Delta(\mathcal{A}) \colon \sum_{a \in \operatorname{supp}(p^\star)\cup\operatorname{supp}(p')} (p'_a - p^\star_a)\left(\langle \tilde{a}, \tilde{\theta} \rangle - \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}}\right) \ge 0 .$$

By the K.K.T. conditions, this condition holds if and only if there exists $\lambda \in \mathbb{R}$ such that

$$\forall a \in \operatorname{supp}(p^\star) \colon \langle \tilde{a}, \tilde{\theta} \rangle - \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}} = \lambda \tag{22}$$

and

$$\forall a \in \mathcal{A} \colon \langle \tilde{a}, \tilde{\theta} \rangle - \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}} \ge \lambda . \tag{23}$$

Note that Eq. (22) implies that

$$\mathbb{E}_{a\sim p^\star}\left[\langle \tilde{a}, \tilde{\theta} \rangle - \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}}\right] = \langle \tilde{a}_{p^\star}, \hat{\theta} \rangle - \frac{\tilde{d}}{\gamma} = \lambda.$$

Combining this identity with Eq. (23) and rearranging, we conclude that

$$\forall a \in \mathcal{A} : \langle \tilde{a}_{p^\star} - a, \hat{\theta} \rangle + \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}_{p^\star}^{-1}} \leq \frac{\tilde{d}}{\gamma}. \tag{24}$$

Finally, observe that for any $p \in \Delta(\mathcal{A})$

$$\tilde{H}_p = \begin{pmatrix} H_p & \bar{a}_p \\ \bar{a}_p^\top & 1 \end{pmatrix}, \quad \text{and} \quad \tilde{H}_p^{-1} = \begin{pmatrix} \left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1} & -\left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1} \bar{a}_p \\ -\bar{a}_p^\top \left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1} & 1 + \|\bar{a}_p\|^2_{\left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1}} \end{pmatrix},$$

where the second expression uses the identity for the Schur complement. Using the latter expression, we have that

$$\begin{aligned} \|\tilde{a}\|^2_{\tilde{H}_p^{-1}} &= \|a\|^2_{(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}} - 2a^\top \left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1} \bar{a}_p + \|\bar{a}_p\|^2_{\left(H_p - \bar{a}_p\bar{a}_p^\top\right)^{-1}} + 1 \\ &= \|a - \bar{a}_p\|^2_{(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}} + 1\,. \end{aligned} \tag{25}$$

By plugging this expression into Eq. (24), it follows that the optimality conditions for the problems (21) and (19) are identical. Any solution $p^\star$ to the problem (21) yields a solution to the problem (19), and vice-versa. □

In light of Proposition 2, we will work exclusively with the lifted problem going forward. Before stating our algorithm, we introduce the following approximate version of the optimality condition in Eq. (4), which quantifies the quality of a candidate solution $p \in \Delta(\mathcal{A})$.

Definition 2. For any action set $\mathcal{A}$, parameter $\hat{\theta} \in \mathbb{R}^d$ and learning rate $\gamma > 0$, a distribution $p \in \Delta(\mathcal{A})$ is called an η*-rounding* if it satisfies

$$\forall a \in \mathcal{A}: \quad \frac{1}{\gamma} \|\tilde{a}\|^2_{\tilde{H}_p^{-1}} \leq (1+\eta) \left(\frac{\tilde{d}}{\gamma} + \langle \tilde{a} - \tilde{a}_p, \tilde{\theta} \rangle \right). \tag{26}$$

The following lemma quantifies the loss in regret incurred by sampling from an η-rounding for the logdet-barrier objective rather than an exact solution.

Lemma 5. Suppose that for all steps t, we sample from an η-rounding for logdet-barrier$(\mathcal{A}_t, \hat{\theta}_t, \gamma/(1+\eta))$ within Algorithm 2. Then the bound from Lemma 4 will increase by at most a factor of $1 + 2\eta$.

Lemma 5 implies that to achieve the regret bound from Theorem 2 up to a factor of 2, it suffices to find a $1/2$-rounding.

Proof. We first prove an analogue of the inequality in Lemma 4. Let t be fixed and abbreviate $\hat{\theta} \equiv \hat{\theta}_t$. Assume without loss of generality that $d = \dim(\mathcal{A}_t)$. For an η-rounding p that satisfies Eq. (26) with learning rate $\gamma' := \gamma/(1+\eta)$, by the identity (25), the following inequalities are equivalent:

$$\begin{aligned} &\frac{1}{\gamma'} \|\tilde{a}\|^2_{\tilde{H}_p^{-1}} \leq (1+\eta) \left(\frac{\tilde{d}}{\gamma'} + \langle a - \bar{a}_p, \hat{\theta} \rangle \right) \\ \iff &\frac{1+\eta}{\gamma} \|\tilde{a}\|^2_{\tilde{H}_p^{-1}} \leq (1+\eta) \left(\frac{\tilde{d}(1+\eta)}{\gamma} + \langle a - \bar{a}_p, \hat{\theta} \rangle \right) \\ \iff &\frac{1}{\gamma} \left(\|a - \bar{a}_p\|^2_{(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}} + 1 \right) \leq \frac{(d+1)(1+\eta)}{\gamma} + \langle a - \bar{a}_p, \hat{\theta} \rangle \\ \iff &\langle \bar{a}_p - a, \hat{\theta} \rangle + \frac{1}{\gamma} \|a - \bar{a}_p\|^2_{(H_p - \bar{a}_p\bar{a}_p^\top)^{-1}} \leq \frac{d}{\gamma} \left(1 + \eta + \frac{\eta}{d} \right). \end{aligned}$$

It follows that the bound from Lemma 4 increases by at most a factor of $(1 + \eta + \frac{\eta}{d}) < 1 + 2\eta$ if we use an η-rounding rather than an exact solution. □

E.1 Algorithm

Preliminaries. To keep notation compact, throughout this section we drop the learning rate parameter and work with

$$G(p) := \langle \tilde{a}_p, \tilde{\theta} \rangle - \log\det(\tilde{H}_p), \quad \text{and} \quad p^\star \in \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} G(p). \tag{27}$$

Note that this suffices to capture the case where $\gamma \neq 1$ (Eq. (20)), since we can multiply both sides by γ and absorb a gamma factor into θ. Consequently, for the remainder of the section we work under the assumption that $\|\theta\| \leq \gamma$ rather than $\|\theta\| \leq 1$. The definition of an η-rounding remains unaffected, since we can multiply both sides in Eq. (26) by γ.

Additional notation. For each $a \in \mathcal{A}$, let $\mathbf{e}_a \in \Delta(\mathcal{A})$ be the distribution that selects a with probability 1. For any distributions $p_1, p_2 \in \Delta(\mathcal{A})$, let $\mathrm{conv}[p_1, p_2] = \{\lambda p_1 + (1-\lambda)p_2 \mid \lambda \in [0,1]\}$ be their convex hull. To improve readability, we abbreviate $\|\cdot\|_{\tilde{H}_p^{-1}}$ to $\|\cdot\|_p$ in this section.

Algorithm. Our main algorithm is stated in Algorithm 6. The algorithm is a generalization of Khachiyan's algorithm for optimal design [30]. It maintains a finitely supported distribution over arms in $\mathcal{A}$ and adds a single arm to the support at each step.

In more detail, the algorithm proceeds as follows. At step k, the algorithm checks whether the current iterate p_{k-1} is an η-rounding. If this is the case, the algorithm simply terminates, as we are done. Otherwise, with $a^\star := \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \theta \rangle$, the algorithm first checks whether the current distribution satisfies $\tilde{d} + \langle a^\star - \bar{a}_{p_{k-1}}, \theta \rangle \geq 1$. If that condition is violated, we define a new distribution p'_{k-1} by choosing the distribution in $\mathrm{conv}[p_{k-1}, e_{a^\star}]$ that minimizes $G(p)$. This ensures that $\frac{\partial}{\partial \lambda}[G(p'_{k-1} + x(\mathbf{e}_{a^\star} - p_{k-1})](0) = 0$ the same as the one along p'_{k-1}, i.e.

$$\langle a^\star, \theta \rangle - \|a^\star\|^2_{p'_{k-1}} = \mathbb{E}_{a \sim p'_{k-1}}\left[\langle a, \theta \rangle - \|a\|^2_{p'_{k-1}}\right] = \langle \bar{a}_{p'_{k-1}}, \theta \rangle - \tilde{d}\,,$$

and hence $\min_{a \in \mathcal{A}} \tilde{d} + \langle a - \bar{a}_{p'_{k-1}}, \theta \rangle = \|a^\star\|^2_{p'_{k-1}} \geq 1$. This ensures in particular that

$$\eta_k := \max_{a \in \mathcal{A}} \|\tilde{a}\|^2_{p'_{k-1}} / (d + \langle a - \bar{a}_{p'_{k-1}}, \theta \rangle) \tag{28}$$

is well defined. To conclude the iteration, the algorithm selects an action a_k that attains the maximum in Eq. (28) and adds it to the support of p'_{k-1}, yielding p_k.

Algorithm 6: Frank-Wolfe for minimizing the logdet-barrier objective

---

Input: $p_0 \in \Delta(\mathcal{A}), \mathcal{A}, \theta, \eta$
Let $a^\star = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \theta \rangle$, $k = 1$.
while p_{k-1} is not an η-rounding (Eq. (26)) **do**
  if $\tilde{d} + \langle a^\star - \bar{a}_{p_{k-1}}, \theta \rangle < 1$ **then**
    Solve $p'_{k-1} = \operatorname{argmin}_{p \in \mathrm{conv}[p_{k-1}, \mathbf{e}_{a^\star}]} G(p)$.
  else
    $p'_{k-1} = p_{k-1}$.
  Pick any $a_k \in \operatorname{argmax} \|\tilde{a}\|^2_{p'_{k-1}} / (d + \langle a - \bar{a}_{p'_{k-1}}, \theta \rangle)$ (ties broken arbitrarily).
  Solve $p_k = \operatorname{argmin}_{p \in \mathrm{conv}[p'_{k-1}, \mathbf{e}_{a_k}]} G(p)$.
  Increment k.

E.2 Analysis

In this section we prove a number of intermediate results used to bound the iteration complexity of Algorithm 6, culminating in our main convergence guarantee, Theorem 7. The total computational complexity is summarized at the end of the section in Appendix E.2.1.

We begin by relating the η-rounding property to the suboptimality gap for the objective $G(p)$.

Lemma 6. If $p \in \Delta(\mathcal{A})$ is an η-rounding, then

$$G(p) - G(p^\star) \leq \log(1+\eta)\tilde{d}\,.$$

Proof of Lemma 6. By the optimality conditions in Eqs. (22) to (24), we are guaranteed that

$$\forall a \in \mathrm{supp}(p^\star):\ \tilde{d} + \langle a, \theta \rangle = \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}} + \langle \bar{a}_{p^\star}, \theta \rangle\,.$$

Hence, combining this statement with the η-rounding condition for p, we have that

$$\forall a \in \mathrm{supp}(p^\star):\ \|\tilde{a}\|^2_{\tilde{H}^{-1}_{p}} \le (1+\eta)\left(\|\tilde{a}\|^2_{\tilde{H}^{-1}_{p^\star}} + \langle \bar{a}_{p^\star} - \bar{a}_p, \theta \rangle\right).$$

Taking the expectation over $a \sim p^\star$ on both sides above and rearranging leads to

$$\langle \bar{a}_p - \bar{a}_{p^\star}, \theta \rangle \le \tilde{d} - \frac{\mathrm{tr}(\tilde{H}_{p^\star}\tilde{H}^{-1}_p)}{1+\eta} = \tilde{d} - \frac{\mathrm{tr}(\tilde{H}^{\frac{1}{2}}_{p^\star}\tilde{H}^{-1}_p\tilde{H}^{\frac{1}{2}}_{p^\star})}{1+\eta}\,.$$

From the definition of $G(p)$, this implies that

$$G(p) - G(p^\star) \le \tilde{d} - \frac{\mathrm{tr}(\tilde{H}^{\frac{1}{2}}_{p^\star}\tilde{H}^{-1}_p\tilde{H}^{\frac{1}{2}}_{p^\star})}{1+\eta} + \log\det(\tilde{H}^{\frac{1}{2}}_{p^\star}\tilde{H}^{-1}_p\tilde{H}^{\frac{1}{2}}_{p^\star}),$$

where we recall that $\det(\tilde{H}^{\frac{1}{2}}_{p^\star}\tilde{H}^{-1}_p\tilde{H}^{\frac{1}{2}}_{p^\star}) = \det(\tilde{H}_{p^\star}\tilde{H}^{-1}_p) > 0$, since $\tilde{H}_{p^\star}$, $\tilde{H}_p \succ 0$. Now, let $(\lambda_i)_{i=1,\dots,\tilde{d}}$ be the eigenvalues of $\tilde{H}^{\frac{1}{2}}_{p^\star}\tilde{H}^{-1}_p\tilde{H}^{\frac{1}{2}}_{p^\star}$. Then we have

$$G(p) - G(p^\star) = \sum_{i=1}^{\tilde{d}} 1 - \frac{\lambda_i}{1+\eta} + \log(\lambda_i) \le \tilde{d}\max_{\lambda>0}\left\{1 - \frac{\lambda}{1+\eta} + \log(\lambda)\right\} = \tilde{d}\log(1+\eta)\,.$$

□

Our next lemma lower bounds the rate at which the suboptimality gap improves at each iteration.

Lemma 7. In each iteration of Algorithm 6, the suboptimality gap improves by at least

$$G(p_{k-1}) - G(p_k) \ge \Omega\left(\min\{\eta_k, 1\}^2/d\right), \tag{29}$$

where we recall that $\eta_k := \|a_k\|^2_{p'_{k-1}}/(\tilde{d} + \langle a_k - \bar{a}_{p'_{k-1}}\rangle)$. Furthermore, if $\eta_k \ge 2\tilde{d}$, then it also holds that

$$G(p_k) - G(p^\star) \le \left(1 - \frac{1}{2\tilde{d}}\right)(G(p_{k-1}) - G(p^\star))\,. \tag{30}$$

Proof. We first prove that Eq. (29) holds. Let k be fixed, and let $\alpha \in [0,1]$ such that $p_k = (1-\alpha)p'_{k-1} + \alpha\mathbf{e}_{a_k}$. Then we have

$$\begin{aligned}
G(p_k) &= \langle \bar{a}_{p_k}, \theta \rangle - \log\det\left(\tilde{H}_{p_k}\right) \\
&= (1-\alpha)\langle \bar{a}_{p'_{k-1}}, \theta \rangle + \alpha\langle \tilde{a}_k, \theta \rangle - \log\det\left((1-\alpha)\tilde{H}_{p'_{k-1}} + \alpha\tilde{a}_k\tilde{a}_k^\top\right) \\
&= \langle \bar{a}_{p'_{k-1}}, \theta \rangle + \alpha\langle \tilde{a}_k - \bar{a}_{p'_{k-1}}, \theta \rangle - \log\left(\det\left((1-\alpha)\tilde{H}_{p'_{k-1}}\right)\cdot\left(1 + \frac{\alpha}{1-\alpha}\|\tilde{a}_k\|^2_{p'_{k-1}}\right)\right) \\
&= G(p'_{k-1}) + \alpha\langle \tilde{a}_k - \bar{a}_{p'_{k-1}}, \theta \rangle - (\tilde{d}-1)\log(1-\alpha) - \log\left(1 - \alpha + \alpha\|\tilde{a}_k\|^2_{p'_{k-1}}\right),
\end{aligned}$$

where the third equality uses the matrix determinant lemma. Now, recall that by the definition of a_k, we have $\|\tilde{a}_k\|^2_{p'_{k-1}} = (1+\eta_k)(\tilde{d} + \langle \tilde{a}_k - \bar{a}_{p'_{k-1}}, \theta \rangle)$. Let us abbreviate $Z_k := \|\tilde{a}_k\|^2_{p'_{k-1}} \ge 1 + \eta_k$. We proceed as

$$\begin{aligned}
G(p_{k-1}) - G(p_k) &\ge G(p'_{k-1}) - G(p_k) \\
&= \alpha\langle \bar{a}_{p'_{k-1}} - \tilde{a}_k, \theta \rangle + (\tilde{d}-1)\log(1-\alpha) + \log\left(1 - \alpha + \alpha\|\tilde{a}_k\|^2_{p'_{k-1}}\right) \\
&= \alpha\left(\tilde{d} - \frac{Z_k}{1+\eta_k}\right) + (\tilde{d}-1)\log(1-\alpha) + \log\left(1 + \alpha(Z_k - 1)\right) \\
&= \max_{\alpha' \in [0,1]}\left\{\alpha'\left(\tilde{d} - \frac{Z_k}{1+\eta_k}\right) + (\tilde{d}-1)\log(1-\alpha') + \log\left(1 + \alpha'(Z_k - 1)\right)\right\},
\end{aligned} \tag{31}$$

where the last equality uses that α is chosen such that $G(p_k)$ is minimized. Next, recalling the elementary fact that for all $x \geq -\frac{1}{2}$, $\log(1+x) \geq x - x^2$, we have in particular that

$$\begin{aligned}
&G(p_{k-1}) - G(p_k) \\
&\geq \max_{\alpha' \geq \frac{1}{2}} \left\{ \alpha' \left(\tilde{d} - \frac{Z_k}{1+\eta_k} \right) + (\tilde{d}-1)(-\alpha' - \alpha'^2) + \alpha'(Z_k - 1) - \alpha'^2 (Z_k - 1)^2 \right\} \\
&= \max_{\alpha' \geq \frac{1}{2}} \left\{ \alpha' \frac{\eta_k Z_k}{1+\eta_k} - \alpha'^2 \left(\tilde{d} - 1 + (Z_k - 1)^2 \right) \right\}.
\end{aligned}$$

Note that $\tilde{d} \geq 3$ and $\max_{x>0} \frac{x}{2+(x-1)^2} \leq 1$, so if we choose

$$\alpha' = \frac{\eta_k Z_k}{2(1+\eta_k)\left(\tilde{d} - 1 + (Z_k-1)^2\right)} \leq \frac{1}{2},$$

we get the lower bound

$$G(p_{k-1}) - G(p_k) \geq \frac{\eta_k^2 Z_k^4}{4(1+\eta_k)^2 \left(\tilde{d} - 1 + (Z_k - 1)^2\right)}.$$

The proof of Eq. (29) now follows by noting that $\frac{x^2}{d+(x-1)^2} \geq \frac{1}{d}$ for all $x \geq 1$.

We now prove that the second part of the lemma, Eq. (30), holds. Suppose $\eta_k > 2\tilde{d}$. We return to Eq. (31) and this time select

$$\alpha' = \frac{\sqrt{\eta_k}}{Z_k - 1} \leq \frac{1}{\sqrt{\eta_k}} \leq \frac{1}{2}.$$

Using the approximation $\log(1+x) \geq x - x^2$ only for the first term in (31), we get

$$\begin{aligned}
G(p_{k-1}) - G(p_k) &\geq \alpha' \left(\tilde{d} - \frac{Z_k}{1+\eta_k} \right) - (\tilde{d}-1)(\alpha' + \alpha'^2) + \log\left(1 + \alpha'(Z_k - 1)\right) \\
&\geq -\frac{\sqrt{\eta_k}}{1+\eta_k} - \frac{\tilde{d}-1}{\eta_k} + \log(1+\sqrt{\eta_k}) \\
&= -\frac{\sqrt{\eta_k}}{1+\eta_k} - \frac{\tilde{d}-1}{\eta_k} + \log(1+\sqrt{\eta_k}) - \frac{1}{4}\log(1+\eta_k) + \frac{1}{4}\log(1+\eta_k) \\
&\geq -\frac{\sqrt{\eta_k}}{1+\eta_k} - \frac{1}{2} + \frac{1}{\eta_k} + \log(1+\sqrt{\eta_k}) - \frac{1}{4}\log(1+\eta_k) + \frac{1}{4}\log(1+\eta_k),
\end{aligned}$$

where the last line uses that $\eta_k \geq 2\tilde{d}$. Now observe that for $x \geq 6$

$$\begin{aligned}
&\frac{\partial}{\partial x}\left(-\frac{\sqrt{x}}{1+x} + \frac{1}{x} + \log(1+\sqrt{x}) - \frac{1}{4}\log(1+x) \right) \\
&= \frac{x-1}{2\sqrt{x}(1+x)^2} - \frac{1}{x^2} + \frac{1}{2(\sqrt{x}+x)} - \frac{1}{4(1+x)} \\
&= \frac{x^{\frac{7}{2}} + x^3 + 5x^{\frac{5}{2}} - 7x^2 - 12x^{\frac{3}{2}} - 8x - 4x^{\frac{1}{2}} - 4}{4x^2(1+\sqrt{x})(1+x)^2} \\
&\geq \frac{7x^2 + 60x^{\frac{3}{2}} - 7x^2 - 12x^{\frac{3}{2}} - 8x - 4x^{\frac{1}{2}} - 4}{4x^2(1+\sqrt{x})(1+x)^2} \geq 0.
\end{aligned}$$

Hence

$$\begin{aligned}
&-\frac{\sqrt{\eta_k}}{1+\eta_k} - \frac{1}{2} + \frac{1}{\eta_k} + \log(1+\sqrt{\eta_k}) - \frac{1}{4}\log(1+\eta_k) \\
&\geq -\frac{\sqrt{6}}{1+6} - \frac{1}{2} + \frac{1}{6} + \log(1+\sqrt{6}) - \frac{1}{4}\log(1+6) > 0.
\end{aligned}$$

It follows that

$$G(p_{k-1}) - G(p_k) \geq \frac{1}{4}\log(1+\eta_k).$$

□

The next lemma ensures we can efficiently find a good initial distribution p_0.

Lemma 8 (Kumar and Yildirim [32], Lemma 3.1)**.** There exists an algorithm that terminates in $\mathcal{O}(|\mathcal{A}|d^2)$ time and finds a distribution $p_0 \in \Delta(\mathcal{A})$ with support $|\operatorname{supp}(p_0)| \le 2\tilde{d}$ such that

$$-\log\det(\tilde{H}_{p_0}) + \min_{p\in\Delta(\mathcal{A})} \log\det(\tilde{H}_p) = \mathcal{O}(d\log(d))\,.$$

The memory requirement of this routine is $\mathcal{O}\left(d^2 + \log(|\mathcal{A}|d)\right)$.

Corollary 4. *The distribution of Lemma 8 has an initial suboptimality gap of*

$$G(p_0) - G(p^\star) = \mathcal{O}(d\log(d) + \gamma)\,.$$

Proof. Recall that

$$G(p_0) - G(p^\star) = \langle \bar{a}_{p_0} - \bar{a}_{p^\star}, \theta\rangle - \log\det(\tilde{H}_{p_0}) + \log\det(\tilde{H}_{p^\star}).$$

The difference between the log-det terms is bounded by $\mathcal{O}(d\log(d))$ using Lemma 8, while the difference between the linear terms is bounded by

$$\langle \bar{a}_{p_0} - \bar{a}_{p^\star}, \theta\rangle \le \|\bar{a}_{p_0} - \bar{a}_{p^\star}\| \cdot \|\theta\| \le 2\gamma\,.$$

□

Theorem 7. *If Algorithm 6 is initialized using the distribution of Lemma 8, then it requires $\mathcal{O}(d(\log(d) + \log(\gamma))$ iterations to reach a $2d$-rounding. Moreover,*

- *After reaching the $2d$-rounding above, the algorithm requires $\mathcal{O}\big(\log(d)d^2\big)$ additional iterations to reach a 1-rounding.*
- *After reaching such a 1-rounding, the algorithm requires $\mathcal{O}\big(d^2/\eta\big)$ additional iterations to reach an η-rounding for any $\eta < 1$.*

Altogether, for any $\eta > 0$, Algorithm 6—when initialized using Lemma 8—requires

$$\mathcal{O}(d\log(\gamma) + d^2(\log(d) + 1/\eta))$$

total steps to reach an η-rounding.

Proof. By Corollary 4 we know that the initial distribution p_0 sastisfies

$$G_0 := G(p_0) - G(p^\star) = \mathcal{O}(d\log(d) + \gamma)\,.$$

We first consider bound the number of steps required to reach a $2d$-rounding. Let k_0 denote the first step k in which p_k is a $2d$-rounding. Then every $k < k_0$ has $\eta_k > 2d$, so in light of Lemma 7, all such k have

$$G(p_k) - G(p_0) \le \left(1 - \frac{1}{2\tilde{d}}\right)(G(p_{k-1}) - G(p_0))$$

and

$$G(p_k) \le G(p_{k-1}) - \Omega(1/d).$$

It follows that as long as $\eta_k > 2d$, the suboptimality gap will reach 1 in most $\mathcal{O}(d\log(G_0)) = \mathcal{O}(d(\log(d) + \log(\gamma)))$ iterations. Moreover, since the absolute decrease in function value is at least $\Omega(1/d)$, the gap will reach zero after another $\mathcal{O}(d)$ iterations. We conclude that after $\mathcal{O}(d(\log(d) + \log(\gamma)))$ iterations, the algorithm must find a $2d$-rounding.

We now bound the number of steps to reach a 1-rounding from the first step where we have a $2d$-rounding. By Lemma 6, the suboptimality gap of any $2d$-rounding is at most $\mathcal{O}(d\log(d))$. Moreover, as long as we haven't reached a 1-rounding, Lemma 7 guarantees that the suboptimality gap will decrease by $\Omega(1/d)$ per step. Hence, we must reach a 1-rounding within $\mathcal{O}(d^2\log(d))$ iterations.

Finally we bound the number of steps required to reach an η-rounding for any $\eta < 1$, starting from the first iteration where we reach a 1-rounding. We adapt an argument of Kumar and Yildirim [32]. Given an η_k-rounding for $\eta_k \le 1$, we need $\mathcal{O}\big(d^2/\eta_k\big)$ iterations to reach an $(\eta_k/2)$-rounding. This follows from the same argument as above: the suboptimality gap is at most $\mathcal{O}(d\eta_k)$ by Lemma 6 (using that $\log(1+\eta_k) \le \eta_k$) and we reduce it by $\Omega(\eta_k^2/d)$ as long as we have not found an $(\eta_k/2)$-rounding (by Lemma 7). Summing up the required number of iterations to get from precision 1 to $1/2$ to $1/4$ to ... to $1/2^{\lceil\log_2(1/\eta)\rceil}$ shows that $\mathcal{O}(d^2/\eta)$ total iterations suffice. □

E.2.1 Total Computational Complexity

The computational complexity per iteration for our method is comparable to similar algorithms for the D-optimal design problem (the case $\theta = 0$) [30, 32, 44]. We walk through the computation complexity step-by-step for completeness, and to handle differences arising from our generalization to the $\theta \neq 0$ case.

The first difference is that we have an intermediate optimization along the line $\mathrm{conv}(p_{k-1}, \mathbf{e}_{a^\star})$. This step increases the computational complexity by a factor of 2. At each iteration, Algorithm 6 computes

$$\operatorname*{argmax}_{a \in \mathcal{A}} \frac{\|\tilde{a}\|^2_{p'_{k-1}}}{d + \langle a - \bar{a}_{p'_{k-1}}, \theta \rangle} \,.$$

For generic action sets, this can be computed in time $\mathcal{O}(|\mathcal{A}|d^2)$, given that $\tilde{H}^{-1}_{p'_{k-1}}$ has already been computed.

In the next step, the algorithm solves the one dimensional optimization problem

$$\max_{\alpha' \in [0,1]} \left(\alpha' \left(\tilde{d} - \frac{Z_k}{1+\eta_k} \right) + (\tilde{d} - 1)\log(1 - \alpha') + \log\left(1 + \alpha'(Z_k - 1)\right) \right) \,,$$

where $Z_k = \|\tilde{a}_k\|^2_{p'_{k-1}}$. This can be done in time $\mathcal{O}(1)$, since it is equivalent to solving the quadratic problem

$$\left(\tilde{d} - \frac{Z_k}{1+\eta_k} \right) - \frac{\tilde{d} - 1}{1 - x} + \frac{Z_k - 1}{1 + x(Z_k - 1)} = 0 \,.$$

Finally we need to update $\bar{a}_p$, which costs $\mathcal{O}(d)$, and update $\tilde{H}_p^{-1}$, which can be done in time $\mathcal{O}(d^2)$ using a rank-one update.

Across all iterations, we require a total of $\tilde{\mathcal{O}}(d^4|\mathcal{A}|)$ arithmetic operations, with p_k never exceeding a support of $\mathcal{O}\big(d^2\log(d) + d\log(\gamma)\big)$, since we maximally add one arm to the support in any iteration. We can store p_k as a sparse vector of key and value pairs, where each entry has a memory complexity of $\mathcal{O}(\log(|\mathcal{A}|))$ to represent the keys.