Proceedings of Machine Learning Research vol 167:1–37, 2022 33rd International Conference on Algorithmic Learning Theory

Leveraging Initial Hints for Free in Stochastic Linear Bandits

Ashok Cutkosky CUTKOSKY@GOOGLE.COM
Chris Dann CHRISDANN@GOOGLE.COM
Abhimanyu Das ABHIDAS@GOOGLE.COM
Qiuyi (Richard) Zhang QIUYIZ@GOOGLE.COM
Google AI

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

We study the setting of optimizing with bandit feedback with additional prior knowledge provided to the learner in the form of an initial hint of the optimal action. We present a novel algorithm for stochastic linear bandits that uses this hint to improve its regret to $\widetilde{O}(\sqrt{T})$ when the hint is accurate, while maintaining a minimax-optimal $\widetilde{O}(d\sqrt{T})$ regret independent of the quality of the hint. Furthermore, we provide a Pareto frontier of tight tradeoffs between best-case and worst-case regret, with matching lower bounds. Perhaps surprisingly, our work shows that leveraging a hint shows provable gains without sacrificing worst-case performance, implying that our algorithm adapts to the quality of the hint for free. We also provide an extension of our algorithm to the case of m initial hints, showing that we can achieve a $\widetilde{O}(m^{2/3}\sqrt{T})$ regret.

1. Introduction

A variety of problems across different disciplines involve making sequential decisions based on noisy observations, often modeled as bandit problems. Learners interacting with a sequential decision making task naturally have access to some "hint", side information, or prior knowledge about what might be a good action or policy for that task. This could be, for example, gleaned from previous or related tasks completed by that learner, or simply some prior domain knowledge about the task. In other cases, the hint could be an existing baseline against which an optimizer should be competitive.

For example, consider the problem of hyperparameter tuning of a neural network on a dataset. If hyperparameter tuning has previously been performed on similar datasets earlier, it would be reasonable to use the optimal hyperparameters from such tasks as a hint for which hyperparameters might be good for the current task. Such offline knowledge transfer is at the heart of emerging fields of learning research, such as transfer learning (Torrey and Shavlik, 2010), metalearning (Vanschoren, 2018), and offline RL (Levine et al., 2020).

In practice, such hints might be inaccurate or imprecise for the current task at hand. A natural question to ask then is whether we can design bandit algorithms that utilize such hints to improve their regret, but, at the same time, are "robust" to imprecisions in the hints. That is, if the hint is "accurate", it obtains better regret than an algorithm that does not have access to this hint. However, if the hint is "inaccurate", then the algorithm should not perform much worse than an algorithm that simply ignores the hint.

More concretely, in this paper we focus on the *stochastic linear bandit* problem on the unit ball (Lattimore and Szepesvári, 2020). Standard regret bounds for this problem is $O(d\sqrt{T})$, and this is known to be unimprovable in the worst case (Dani et al., 2008). To model the notion of a "hint",

© 2022 A. Cutkosky, C. Dann, A. Das & Q.(. Zhang.

we further suppose that the algorithm is provided with a vector $h \in \mathbb{R}^d$ that is some (possibly imprecise) estimate of the optimal action $a_\star$, but does not know in advance how close h is to $a_\star$. In this paper, we address the following question for the stochastic linear bandit setting: Can we design a bandit algorithm that uses this hint to obtain $o(d\sqrt{T})$ regret if the hint is sufficiently accurate (i.e "close" to $a_\star$), but maintains the same worst-case $O(d\sqrt{T})$ guarantee, even if the hint is inaccurate?

The analogous question has been previously studied in the multi-armed bandit setting (MAB) (Lattimore, 2015) and answered in the negative. Essentially, Lattimore (2015) shows that any MAB algorithm over k arms that obtains $o(\sqrt{kT})$ regret if the hint is accurate, must suffer $\omega(\sqrt{kT})$ worst case regret if the hint is inaccurate. Specifically, for some hint arm i, let R_i^T denote the regret with respect to arm i, and R^T denote the regret with respect to the best arm, so $R^T = \max_i R_i^T$. Then, there are matching upper and lower bounds that show that $R^T R_i^T = \Theta(kT)$, which implies an inherent tradeoff on a Pareto frontier.

1.1. Our Contributions

Perhaps surprisingly, in case of stochastic linear bandits where the action space forms a unit ball ($\{a \in \mathbb{R}^d \colon \|a\| \leq 1\}$), we can answer our main question in the affirmative. We introduce an algorithm that obtains a dimension-independent regret of $O(\sqrt{T})$ if the hint is sufficiently accurate, while maintaining an $O(d\sqrt{T})$ worst-case regret bound, even if the hint is arbitrarily inaccurate. Specifically, we can characterize our algorithm's performance in terms of obtaining an $R_h^T = \widetilde{O}(\sqrt{T})$ *hint-based regret*, i.e. regret with respect to the hint action, while maintaining a worst case $R^T = \widetilde{O}(d\sqrt{T})$ regret with respect to the optimal action (see Theorem 11). Equivalently, let r_h be the instantaneous regret of playing the hint h, then our algorithm achieves total regret $\min(r_h T + \widetilde{O}(\sqrt{T}), \widetilde{O}(d\sqrt{T}))$.

Our algorithm exploits the geometry of the action set to quickly approximate r_h by playing small perturbations of h. These perturbations are on the order of $O(1/\sqrt{T})$ so as to ensure $O(\sqrt{T})$ hint-based regret, but are tightly controlled as to ensure fast statistical inference of r_h, allowing us to bound our worst-case regret by switching to any $O(d\sqrt{T})$ linear bandit algorithm. We emphasize that all upper and lower bounds in our paper hold in the regime when $\|\theta^\star\| = \Theta(1)$, which enforces a fixed signal-to-noise ratio and our main variable then becomes purely the quality of the hint.

Furthermore, similar to Lattimore (2015) we characterize the inherent explore-exploit tradeoff when balancing superior performance if the hint is sufficiently accurate with inferior performance when the hint is misspecified, in terms of a tight *Pareto Regret Frontier* for our algorithm. Specifically, our algorithm can achieve any hint-based regret $R_h^T < \sqrt{T}$, while maintaining a worst case regret of $O(dT/R_h^T)$ (see Theorem 13), and we obtain matching lower bounds to show that this Pareto Regret Frontier cannot be improved in general (see Theorem 1). This is particularly surprising since by restricting the action set to axis-aligned actions, we can derive MAB over d arms, which has the same Pareto frontier as the more general linear bandit problem when $R_h < \sqrt{T}$. This somewhat confounds the conventional wisdom that linear bandits should suffer an extra $O(\sqrt{d})$ complexity due to the larger action set.

Lastly, we also extend our analysis to the case of multiple hints, where the algorithm is now provided m hints, $h_1, h_2, \ldots, h_m$. Although this is quite practical, as optimization tasks often have multiple prior tasks to learn hints from, this setup in the stochastic setting has not been studied before, according to our knowledge. We show that we can achieve hint-based regret of $\widetilde{O}(m^{2/3}\sqrt{T})$ with respect to the best hint $h^\star$ (see Lemma 15) while maintaining the usual worst case $O(d\sqrt{T})$

regret (see Lemma 14). To do this, we use a clever combination of MAB on the m hints while simultaneously performing perturbations of each hint to estimate its instantaneous regret. To achieve the sublinear rate, we perform a careful balancing between the sample complexity of two elimination forces: the first coming from a suboptimality compared to $h^\star$ and the second coming from suboptimality compared to $a^\star$.

We summarize our contributions as follows:

- Introduce novel algorithm for stochastic linear bandits on the unit ball that achieves $R_h^T = \widetilde{O}(\sqrt{T})$ while maintaining the usual $\widetilde{O}(d\sqrt{T})$ regret. Equivalently, our total regret is the best of both worlds: $\min(r_h T + \widetilde{O}(\sqrt{T}), \widetilde{O}(d\sqrt{T}))$.
- Characterize a Pareto frontier of (R_h^T, R^T) tradeoffs with matching upper and lower bounds, up to log factors, that $R_h^T R^T = \Theta(dT)$. Surprisingly, this matches the same Pareto frontier for MAB with d arms, implying that linear bandits may not suffer inherently higher regret than MAB in some settings.
- Generalize our algorithm to the practical yet novel m-hint setting and show sublinear hint-based regret $R_{h^*}^T = \widetilde{O}(m^{2/3}\sqrt{T})$ while maintaining usual $\widetilde{O}(d\sqrt{T})$ regret. Equivalently, our total regret in this case is: $\min(r_{h^*} T + \widetilde{O}(m^{2/3}\sqrt{T}), \widetilde{O}(d\sqrt{T}))$

1.2. Related work

1.2.1. CONSERVATIVE EXPLORATION

Conservative exploration, introduced in (Kazerouni et al., 2016; Wu et al., 2016), attempts to maximize reward while keeping its performance relative to a baseline action, which is analogous to an initial hint, above a certain multiplicative threshold. While similar to our setting, the problem is typically viewed as constrained optimization and the regret bounds given are usually of the form $O(d\sqrt{T} + S)$, where S is the additional cost of following the baseline constraint. In our setting, we do not enforce any constraint; instead, we want to simultaneously minimize hint-based additive regret while diverting away from playing the hint when its instantaneous regret is too large.

1.2.2. TUNING THE LEARNING RATE IN MIRROR DESCENT

A standard worst-case optimal algorithm for our linear bandit problem is to employ mirror descent with an appropriate regularizer combined with a one-point gradient estimator Abernethy et al. (2008). Such algorithms typically have a *learning rate* parameter that could in principle be tuned to the problem at hand: if $h = a_\star$, we should center the regularizer at h and set the learning rate to 0. If h is far from $a_\star$, we should instead optimize the learning rate for the worst-case scenario. Thus, the problem of effectively using the hint is essentially a problem of tuning the learning rate. Unfortunately, it is extremely unclear how and if it is possible to do this without prior knowledge of some unavailable quality value such as $\|h - a_\star\|$. Although the corresponding tuning problem for the *full-information* online linear optimization problem can be solved (Streeter and McMahan, 2012; Cutkosky and Orabona, 2018; Chen et al., 2021), these techniques do not extend in any obvious way to the bandit setting.

1.2.3. PRIORS WITH UPPER CONFIDENCE BOUND

A second standard algorithm for solving the stochastic linear bandit problem is to employ upper-confidence-bound (UCB) and the optimism principle Abbasi-Yadkori et al. (2011). The standard technique involves solving a regularized least-squares subproblem in order to generate a confidence ellipsoid for the true parameter $\theta^\star$. Intuitively, the regularizer in the least-squares subproblem plays a similar role to the *prior* in methods based on Thompson sampling Agrawal and Goyal (2013); Abeille and Lazaric (2017). Thus, a natural strategy is to have the hint h inform the regularizer or the prior. For example, instead of the standard regularized least-squares problem in which the regularizer is $\frac{\lambda}{2}\|x\|^2$, we could use $\frac{\lambda}{2}\|x-h\|^2$, which would encourage the confidence ellipsoid to be centered at h. Unfortunately, the influence of the regularizer on the final regret bounds for UCB seems rather small: λ only appears inside a logarithmic term. There does not appear to be any clear way to set the regularizer in such a way that there is significant improvement when $h = a^\star$ while maintaining reasonable regret for $h \neq a^\star$ (the former naively requires exponentially large λ, which disallows the latter).

1.2.4. MODEL SELECTION

Yet another natural approach to incorporating a hint is through *model selection*. Model selection in contextual bandits is a more general problem than the one we explore here and has been intensely studied in recent years (Foster et al., 2019; Ghosh et al., 2020; Pacchiano et al., 2020; Arora et al., 2021; Agarwal et al., 2017; Chatterji et al., 2020; Bibaut et al., 2020; Krishnamurthy et al., 2021; Cutkosky et al., 2021). In its most expansive formulation, we consider a set of K "base" bandit algorithms, each of which may or may not achieve a good regret bound. The goal is to combine these algorithms in a black-box manner so as to produce a single bandit algorithm whose regret is guaranteed to be not much more than the best possible regret we would have been able to obtain by exclusively employing any one of the base algorithms.

To use such a result for our problem, we could consider $K = 2$ algorithms: one algorithm ignores all feedback and simply plays the hint action h at all time steps. The other is any standard linear bandit algorithm that ignores the hint but uses the feedback. If the h is indeed the optimal action, then the first algorithm achieves 0 regret, but otherwise has linear regret. Alternatively, the second algorithm may always guarantee $d\sqrt{T}$ regret, which is the optimal bound for d-dimensional linear bandits. Unfortunately, if we were to combine these algorithms using these black-box approaches there would be significant overhead. Typical bounds (e.g. Agarwal et al. (2017)) would yield an algorithm whose regret when the hint is correct is $\sqrt{T}$, but decays to $d^2\sqrt{T}$ when the hint is incorrect. Thus, we must pay a significant price in the worst-case regret for improved performance in the best-case. Our goal is to instead leverage the additional structure of our problem (stochasticity, linearity, unit-ball domain) to achieve improved results when the hint is correct without compromising on worst-case performance.

1.2.5. HINTS IN ADVERSARIAL CONTEXT

In the adversarial context, there has been many areas of work that tries to incorporate once-per-round hints to bound worst case regret. We note that since we work in the stochastic context, most of these results do not apply since our hints *do not predict* per-round stochastic noise and the presented bounds are too weak. Nevertheless, recent works (Purohit et al., 2018; Lykouris and Vassilvtiskii, 2018) introduced algorithms that tradeoff a notion of consistency (competitive ratio

when hint is good or perfect) and a notion of robustness (worst-case competitive ratio). These results are supplemented with lower bounds (Wei and Zhang, 2020) and Pareto frontier characterizations (Angelopoulos et al., 2019). Furthermore, a recent result (Wei et al., 2020) shows how to leverage hint estimators in contextual adversarial MAB, including the setting with multiple estimators.

2. Problem Setting and Notation

We consider the classic *stochastic linear bandit learning setting*. In each round $t \in \mathbb{N}$, the learner chooses an action a_t from the action set $\mathcal{A} = \{a \in \mathbb{R}^d \colon \|a\| \leq 1\}$ and receives a reward $y_t = \langle a_t, \theta^\star \rangle + \xi_t$ where ξ_t is independent 1-sub-Gaussian noise and $\theta^\star \in \mathbb{R}^d$ is the unknown true parameter vector. The performance of a learner is measured by its (pseudo-) regret

$$R^T = \mathrm{Reg}(T) = \mathrm{Reg}_{a^\star}(T) = \sum_{t=1}^{T} r(a^\star, a_t) = \sum_{t=1}^{T} \langle \theta^\star, a^\star - a_t \rangle$$

where $a^\star = \frac{\theta^\star}{\|\theta^\star\|}$ is the optimal action and $r(a, a') = \langle \theta^\star, a \rangle - \langle \theta^\star, a' \rangle$ is the instantaneous regret of action a' with respect to action a.

Initial action hint: We study the problem where the learning is provided with a hint $h \in \mathcal{A}$ in the form of an action before interacting with the bandit instance. This hint is supposed to be a guess of the optimal action available through prior knowledge. The goal of the learner is to use this hint to achieve better regret when the quality of the hint is good. The quality can be measured by the regret $r_h = r(a^\star, h)$ of h w.r.t. the optimal action. To assess the degree to which a learner can leverage a good hint, we look at its regret w.r.t. the hint (*hint-based regret*)

$$R_h^T = \mathrm{Reg}_h(T) = R(\mathbf{a}, \mathbf{h}) = \sum_{t=1}^{T} r(h, a_t) = \sum_{t=1}^{T} \langle \theta^\star, h - a_t \rangle \ .$$

Note that the regret w.r.t. the best action is the sum of the total regret of the hint and the hint-based regret, $\mathrm{Reg}(T) = T \cdot r(a^\star, h) + \mathrm{Reg}_h(T)$. This implies that when the hint is sufficiently good, i.e. $r(a^\star, h) \leq \frac{\widetilde{R}}{T}$, the learner with hint-based regret $\widetilde{R} \ll d\sqrt{T}$ also achieves total regret $\mathrm{Reg}(T) \leq \widetilde{R}$. Note that a naive exploitative algorithm that achieves no hint-based regret simply plays h in all rounds. Thought it performs well if $h \approx a^\star$, it suffers linear regret in problems where the hint is not good, i.e., $r(a^\star, h) \gg \frac{d}{\sqrt{T}}$, which is undesirable.

Objective: Our goal is to devise an algorithm that is able to leverage a hint but is also robust to its quality. Specifically, this algorithm should simultaneously guarantee a worst-case regret rate $\mathrm{Reg}(T) \leq R$ and a hint-based regret rate $\mathrm{Reg}_h(T) \leq R_h$ for all hints h and bandit instances (with high probability).

Multiple hints: We also consider the setting where multiple initial hints $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$ are provided. In this case, the algorithm should maintain small hint-based regret with respect to the best hint $h^\star = \mathrm{argmin}_{h \in \mathcal{H}} r(a^\star, h)$ while ensuring a worst-case regret rate $\mathrm{Reg}(T) \leq R$.

Additional notation: Our analysis makes heavy use on the geometry of the action and parameter space. We assume the action set $\mathcal{A}$ to be the unit ball and all norms are ℓ_2, but our results can be generalized to general ellipsoids. When the action set consists only of axis-aligned unit vectors, this

setting reduces to the *multi-armed bandit setting*. To that end, it is useful to define the following projection operators and recall some of their properties. For a vector $v \in \mathbb{R}^d$, let P_v, $P_v^\perp$ be projection operators onto the span of v and its orthogonal complement, respectively,

$$P_v u = \frac{\langle u, v\rangle v}{\|v\|^2} \qquad\qquad P_v^\perp u = u - P_v u,$$

where $u \in \mathbb{R}^d$. Also note that $\|P_v u\|\|v\| = |\langle v, u\rangle| = \|P_u v\|\|u\|$ and $\|P_v^\perp u\|\|v\| = \|P_u^\perp v\|\|u\|$.

3. Regret Lower Bounds

Algorithms that can effectively exploit prior information, while maintaining good worst-case performance have an inherent explore-exploit tradeoff throughout optimization. In this section, we look at lower bounds that show an inherent limitation on the ability of algorithms to explore and exploit well effectively at the same time. Specifically, we will show that if an algorithm has superior hint-based regret, it must lack the explorative capabilities to bound the worst-case total regret. As discussed in the introduction, for the MAB setting with k arms, Lattimore (2015) shows that the Pareto frontier is lower bounded by $R_h^T R^T = \Omega(kT)$. For linear bandit setting, we can show analogous lower bounds of $R_h^T R^T = \Omega(dT)$ via axis-aligned perturbations of a true parameter θ_0. In the following, we make the bandit instance θ for regret and hint-based regret explicit by superscripts θ.

Theorem 1 *For any action hint $h \in \mathcal{A} \subseteq \mathbb{R}^{d+1}$ with $\|h\| = 1$, horizon T and learning algorithm, there is a family Θ of stochastic linear bandit instances so that the following holds. All parameters $\theta \in \Theta$ have $\Theta(1)$ norm. Let $R_h = \max_{\theta\in\Theta} \mathbb{E}[\mathrm{Reg}_h^\theta(T)]$ be largest hint-based regret in family Θ. Then there is a bandit instance $\theta \in \Theta$ such that the expected regret w.r.t. the best arm is*

$$\mathbb{E}[\mathrm{Reg}^\theta(T)] \geq \frac{1}{4} + \frac{T}{4}\min\left(1, \frac{d-1}{R_h}\right) \geq \Omega\left(\frac{dT}{R_h}\right)$$

See Appendix A for the complete proof.

Theorem 2 *Let $T \geq \min(d^2/2^8, 2^6)$ and $\Delta^2 = \frac{d}{2^8\sqrt{T}}$. For any hint $h \in \mathbb{R}^{d+1}$ with $\|h\| = 1$ and any algorithm, there is a linear bandit instance with parameter $\theta^\star \in \mathbb{R}^{d+1}$ satisfying $\|\theta^\star - \frac{h}{2}\| \leq \Delta$ with optimal action $a^\star = \frac{\theta^\star}{\|\theta^\star\|}$ satisfying $\|a^\star - h\| \leq 4\Delta$ and $r(a^\star, h) = \langle \theta^\star, a^\star - h\rangle \leq 972\Delta^2 \leq O(d/\sqrt{T})$ such that*

$$\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] \geq \frac{d\sqrt{T}}{512}$$

This theorem provides a fundamental limit on how "good" a hint needs to be in order for it to provide a significant advantage over a worst-case algorithm. In particular, even if the distance between the hint h and the optimal action $a^\star$ is $O(\sqrt{d}/T^{1/4})$, and even simply playing the hint on its own already achieves $T \cdot r(a^\star, h) \leq O(d\sqrt{T})$ regret, we still cannot leverage the hint to actually improve on the worst-case bounds. That is, the hint must be very high quality in order to be useful. This lower bound also suggests how one should go about attempting to use the hint: since a hint can only be useful if $r(a^\star, h)$ is small enough that playing just the hint for all T rounds would already improve upon the worst-case regret, the fundamental question to answer is *how good is the hint?* If we can answer this question using a small number of observations, then we would be able to either simply play the hint if it is sufficiently good, or fall back to a standard worst-case algorithm that ignores the hint otherwise.

4. New Algorithm for Stochastic Linear Bandits with Action Hint

In linear bandits with a given action hint, the goal of a learning algorithm is to use the hint to achieve lower regret if possible but still achieve close minimax-optimal regret even when the hint is misleading. To what degree an algorithm should rely a hint depends on its quality as measured by its instantaneous regret $r_h = r(a^\star, h)$. In fact, if we know r_h, then the following simple switching algorithm achieves the desired learning properties:

Algorithm 1: SWITCH$(h, \widehat{r}_h, T, R_{LB})$: Simple Switching Algorithm

Input: hint $h \in \mathbb{R}^d$, approximation of instantaneous regret $\widehat{r}_h \leq r_h \in \mathbb{R}^+$, number of rounds $T \in \mathbb{Z}$, maximum regret of hint-agnostic linear bandit algorithm $R_{LB} \in \mathbb{R}^+$

if $\widehat{r}_h T \leq R_{LB}$ **then**
  play h
else
  play hint-agnostic linear bandit algorithm $\mathcal{L}$

Lemma 3 *Let $\widehat{r}_h$ be a α-approximation to the instantaneous regret r_h of h, i.e., $\widehat{r}_h \leq r_h \leq \alpha \widehat{r}_h$ and $R_{LB} \in \mathbb{R}^+$ an upper-bound on the regret of the linear bandit algorithm $\mathcal{L}$ played for T episodes, i.e., $R_{\mathcal{L}}^T \leq R_{LB}$. Then the worst case regret running* SWITCH$(h, \widehat{r}_h, T, R_{LB})$ *for up to T rounds is bounded by αR_{LB}.*

If α is a constant and we employ a minimax-optimal hint-agnostic algorithm such as OFUL (Abbasi-Yadkori et al., 2011) that satisfies $R_{\mathcal{L}}^T \leq R_{LB} = \widetilde{O}(d\sqrt{T})$ with high probability, then SWITCH always retains the minimax-rate $\widetilde{O}(d\sqrt{T})$ with high probability as well. Moreover, since the instantaneous reward estimator is a lower bound $\widehat{r}_h \leq r_h$, SWITCH also guarantees zero regret with respect to the hint. Unfortunately, SWITCH cannot be used directly since no approximation $\widehat{r}$ to the instantaneous regret r_h is known to the learner. The main idea of our algorithm is to first estimate r_h up to a constant factor without incurring too much regret and then call the switch procedure.

4.1. Estimating Regret of Hint Action

Our estimation procedure for $r_h = \langle a^\star - h, \theta^\star \rangle$ is based on the following decomposition:

Lemma 4 *Let $a \in \mathbb{R}^d$ be any action with $\|a\| = 1$ and $\langle a, \theta^\star \rangle \geq -\|\theta^\star\|/2$. Then the instantaneous regret of this action is bounded as*

$$\frac{1}{2}\frac{\|P_a^\perp \theta^\star\|^2}{\|\theta^\star\|} \quad \leq \quad \langle a^\star - a, \theta^\star \rangle \quad \leq \quad 3\frac{\|P_a^\perp \theta^\star\|^2}{\|\theta^\star\|} \ .$$

Instantiating this lemma with $a = h$ shows that estimating $\|\theta^\star\|$ and $\|P_h^\perp \theta^\star\|$ is sufficient to compute a constant-factor approximation of r_h, as long as h is not an extremely bad hint. In the following, we will first present our procedure for estimating $\|P_h^\perp \theta^\star\|$ while incurring small regret with respect to both h and the optimal action $a^\star$. We will also be able to estimate $\|\theta^\star\|$ by applying essentially the same procedure with $h = 0$, because $P_0^\perp \theta^\star = \theta^\star$. By combining these estimates, we obtain an estimate of $\|P_h^\perp \theta^\star\|^2 / \|\theta^\star\|$, which is in turn an estimate of r_h.

Since $P_h^\perp \theta^\star$ is orthogonal to the hint h, we gain no information about it in rounds where h itself was played. We would gain the most information if we played actions that are orthogonal to h but those incur very large regret when the hint h is aligned with $\theta^\star$. To balance regret and information gain, we therefore play perturbations of the hint h

$$\textsc{Perturb}(h, p) = \frac{h + p}{\sqrt{1 + \|p\|^2}}$$

where $p \in \mathbb{R}^d$ is an orthogonal perturbation to h ($\langle p, h\rangle = 0$).

We will show that by choosing the size of the perturbation correctly and applying careful statistical analysis, we can incur small hint-based regret while quickly estimating $\|P_h^\perp \theta^\star\|$. Our perturbations will be sampled from projection Gaussian distributions, defined below:

Definition 5 *Let the projected Gaussian distribution $\mathcal{N}_h(0, \mathbf{I})$ be the distribution of $P_h^\perp g$, where $g \sim \mathcal{N}(0, \mathbf{I})$ and $P_h^\perp$ is the orthogonal projection onto the complement of h. When h is a unit vector, the distribution is equivalent to $\mathcal{N}(0, \mathbf{I} - hh^\top)$. For ease of notation, we will often drop $\mathbf{I}$ and $\mathcal{N}_h(0, \Delta) := \mathcal{N}(0, \Delta(\mathbf{I} - \frac{hh^\top}{\|h\|^2}))$*

Next, we analyze the regret of playing a perturbation for a single round, both with respect to the hint h (Lemma 12), and also with respect to the optimal action (Lemma 7):

Lemma 6 (Hint-based Regret of Perturbed Action on the Unit Ball) *Let $h \in \mathbb{R}^d$ a hint action with $\|h\| = 1$ and let $p \in \mathbb{R}^d$ be any orthogonal perturbation of h with $\|p\| < \frac{1}{8}$. Then, the regret w.r.t. hint h of the perturbed action $a = \textsc{Perturb}(h, p)$ is bounded as*

$$\langle h - a, \theta^\star\rangle \leq \|\theta^\star\|\|p\|^2 + |\langle p, P_h^\perp \theta^\star\rangle|.$$

Furthermore, if $p \sim \mathcal{N}_h(0, \frac{\Delta^2}{d-1})$, then with probability at least $1 - \delta$,

$$\langle h - a, \theta^\star\rangle = O\left(\|\theta^\star\| \log(1/\delta) \left(\Delta^2 + \Delta \frac{\|P_h^\perp \theta^\star\|}{\|\theta^\star\|\sqrt{d-1}}\right)\right)$$

Lemma 7 (Regret of Perturbed Hint on the Unit Ball) *Let $\theta^\star \in \mathbb{R}^d$ be the true parameter and $h \in \mathbb{R}^d$ an action hint with $\|h\| = 1$ and $\langle h, \theta^\star\rangle \geq -\|\theta^\star\|/4$. Further, let $p \in \mathbb{R}^d$ be any orthogonal perturbation of h with $\|p\| < \frac{1}{8}$. Then, the instantaneous regret of the perturbed prior action $a = \textsc{Perturb}(h, p)$ is bounded as*

$$\langle a^\star - a, \theta^\star\rangle \leq 12 \frac{\|P_h^\perp \theta^\star\|^2}{\|\theta^\star\|} + 3\|\theta^\star\|\|p\|^2 \,.$$

Now, we present our algorithm to estimate the norm $\|P_h^\perp \theta^\star\|$ based on perturbing the hint (Algorithm 2. First, we present an algorithm that yields an accurate approximation with some moderate constant probability of failure while maintaining small regret. Our procedure requires playing a fixed underlying action for the perturbation to obtain optimal sample complexity. Later, we will use a median of means approach to amplify this constant probability provide a high-probability guarantee (Algorithm 4). The analysis for the constant-probability algorithm is provided in Lemma 8 and Lemma 9, while the analysis for the high-probability algorithm is provided in Lemma 10.

Algorithm 2: ESTIMATENORM(h, Δ): Low Regret ℓ_2-norm Estimation

Input: reference action $h \in \mathbb{R}^d$, perturbation magnitude $\Delta \in \mathbb{R}^+$

Initialize $n = 0$

Sample perturbation $p \in \mathbb{R}^d$ from $\mathcal{N}_h(0, \frac{\Delta^2}{d'}) := \mathcal{N}(0, \frac{\Delta^2}{d'}(\mathbf{I} - \frac{hh^\top}{\|h\|^2}))$ where $d' = d - \mathbf{1}\,(h \neq 0)$

Procedure PLAYANDUPDATE()

  Increment $n \leftarrow n + 1$

  Play hint h, observe reward y_n and update average $\overline{y}_n = \frac{1}{n}\sum_j y_j$

  Play $\frac{h+p}{\sqrt{\|p\|^2+\|h\|^2}}$, observe reward z_n and update average $\overline{z}_n = \frac{1}{n}\sum_j z_j$

  Compute average difference $\overline{x}_n = \overline{z}_n\sqrt{\|p\|^2 + \|h\|^2} - \overline{y}_n$

  Compute confidence width $b_n = \sqrt{\frac{3(1+\|p\|^2+\Delta^2)\ln(40\ln(2n))}{n}}$

  if $|x_n| \geq 2b_n$ **then**

    return $\frac{\sqrt{d'}}{\Delta}|\overline{x}_n|$

Lemma 8 *With probability at least* 0.7*, Algorithm 2 (*ESTIMATENORM(h, Δ)*) returns* r*, a constant-factor approximation to* $\|P_h^\perp\theta^\star\|$*, i.e.,* $0.06\|P_h^\perp\theta^\star\| \leq r \leq 5\|P_h^\perp\theta^\star\|$*, after* $\widetilde{O}\left(\frac{d(1+\Delta^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$ *calls to its* PLAYANDUPDATE *procedure and also satisfies* $\|p\| \leq 3\Delta$*.*

Lemma 9 *In the event where Lemma 8 holds, if* $\Delta \leq 1/24$ *and* $\langle h, \theta^\star\rangle \geq -\frac{\|\theta^\star\|}{4}$*, then the total regret incurred by Algorithm 2 is*

$$\widetilde{O}\left(\frac{d}{\Delta^2\|\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right)$$

and the regret w.r.t. hint vector h *is*

$$\widetilde{O}\left(\frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right).$$

To gain some intuition for these lemmas, let us look ahead a bit: we will eventually wish to set $\Delta^2 = \Theta(\|\theta^\star\|/\sqrt{T})$. Further, consider the case that $\|P_h^\perp\theta^\star\| = \Theta(d\|\theta^\star\|/\sqrt{T})$ - as suggested by the lower bound Theorem 2, this is a "transition case" beyond which we may be able to leverage the hint. With these settings, Lemma 9 implies that the regret is $\widetilde{O}(\sqrt{T})$, so that we are able to estimate $\|P_h^\perp\theta^\star\|$ well (as implied by Lemma 8) without incurring significant regret.

The next lemma amplifies these results to a high-probability guarantee by running $O(\log(1/\delta))$ parallel instances of the constant-probability estimation and taking medians of the returned estimates. It is the main work-horse of our final algorithm analysis (see Algorithm 4 in Appendix).

Lemma 10 *With probability at least* $1 - \delta$*, Algorithm 4 (*ESTIMATENORMHP(h, Δ, δ)*) returns* r*, a constant-factor approximation to* $\|P_h^\perp\theta^\star\|$*, i.e.,* $0.06\|P_h^\perp\theta^\star\| \leq r \leq 5\|P_h^\perp\theta^\star\|$*, after at most*

$n \leq \widetilde{O}\left(\frac{d(1+\Delta^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$ *calls to its* PLAYANDUPDATE *procedure. Further, if* $\Delta \leq 1/224$*, then the regret incurred with respect to the hint* h *is:*

$$O\left(n\|\theta^\star\|\Delta^2 + n\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d}}\right) \leq \widetilde{O}\left(\frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta} + \frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|}\ln\frac{1}{\delta}\right)$$

If in addition, $\langle h, \theta^\star\rangle \geq -\frac{\|\theta^\star\|}{4}$*, then the total regret incurred is*

$$O\left(n\frac{\|P_h^\perp\theta^\star\|^2}{\|\theta^\star\|} + n\|\theta^\star\|\Delta^2\right) \leq \widetilde{O}\left(\frac{d(1+\|h\|)}{\Delta^2\|\theta^\star\|}\ln\frac{1}{\delta} + \frac{d(1+\|h\|)\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta}\right).$$

Remark: Notice that during the norm estimation process, we play the *same* perturbed action over and over again, as opposed to re-sampling the perturbation p afresh for each action. Perhaps counterintuitively, fixing the perturbation p is actually crucial for a low sample complexity in de-noising the norm estimate, since resampling p adds significant noise into the observations.

To see how this might be, consider observing samples of $y \sim N(\mu, 1)$, where $\mu \sim N(0, \sigma^2)$ in one instance and $\mu = 0$ in another instance. Our goal is to distinguish which instance we are in from our sampled observations and $\sigma < 1$ is small, so it's a relatively difficult distinguishing task. In this case, we also allow the observer to resample μ from the mean distribution, if they wish, before observing the final sample.

If we fix the randomness of μ, we are distinguishing between samples from $N(0, 1)$ and $N(\mu, 1)$, where $|\mu|$ is likely to be at least σ, so the total variational distance between those distribution is on the order of σ.

However, if we vary μ and resample at each observation, then we are distinguishing between $N(0, 1)$ vs $N(0, 1 + \sigma^2)$, which is significantly harder since the total variational distance between these distribution is on the order of σ^2, not σ.

Now, we are in a position to put together our main algorithm and analysis. The algorithm has three distinct phases. The first two phases apply our low-regret norm estimation procedure (Algorithm 4) to estimate $\|\theta^\star\|$ (by setting $h = 0$) and $\|P_h^\perp\theta^\star\|$ respectively. The last phase simply combines the values to estimate $r(a^\star, h)$, and then calls Algorithm 1. However, there is a significant subtlety that must be overcome when estimating $\|P_h^\perp\theta^\star\|$: if h is actually a very poor hint (i.e. $\langle\theta^\star, h\rangle \leq -\|\theta^\star\|/4$), then Algorithm 4 will actually incur a large regret. In order to avoid this issue, we observe that at least one of h and $-h$ must be positively correlated with $\theta^\star$, so that at least one of these must be usable with Algorithm 4. Further, if either one (say h), satisfies $\langle\theta^\star, h\rangle \leq -\|\theta^\star\|/4$, then there will be a large gap between the rewards for actions h and $-h$. This means that we can quickly select which of the two is positively correlated.

Finally, performing this two-arm selection procedure while simultaneously norm estimating allows norm estimation to succeed extremely quickly before the two-arm selection ends. To get around this, note that we only fully eliminate either $h, -h$ if it is clearly a bad choice according to the norm estimate. Otherwise, we will continue calling `PlayAndUpdate()`, although it is important to note that we will no longer play the perturbed hint after the instance has returned a norm estimate, so as to minimize hint-based regret. The full description is provided in the pseudocode, with analysis in Theorem 11 below:

Theorem 11 (Main Regret Bound) *Suppose we instantiate Algorithm 1 with any standard worst-case optimal linear bandit algorithm. Then there exists an absolute constant* W *such that with*

Algorithm 3: PARETOBANDIT(h, T, δ, W): Pareto-Optimal Bandit Algorithm on Unit Ball

Input: hint $h \in \mathbb{R}^d$, number of rounds T, failure probability δ, bound on constant factor worst-case regret scalings for three phases W.

// Phase 1: Estimate norm $\|\theta^\star\|$

Initialize $C_0 \leftarrow$ ESTIMATENORMHP$(0, 1, \frac{\delta}{4})$ (from Algorithm 4)

Call C_0.PLAYANDUPDATE() until it returns a value r

// Phase 2: Estimate norm of orthogonal complement $\|P_h^\perp \theta^\star\|$

Set exploration radius $\Delta = \frac{1}{\sqrt{r}T^{1/4}}$

Initialize $C_+ \leftarrow$ ESTIMATENORMHP$(+h, \Delta, \frac{\delta}{4})$ and $C_- \leftarrow$ ESTIMATENORMHP$(-h, \Delta, \frac{\delta}{4})$

Initialize active set $\mathcal{S} = \{C_+, C_-\}$

repeat

  foreach *active instance* $C_i \in \mathcal{S}$ **do**

    // If C_i already returned a norm estimate, PlayAndUpdate no longer plays an perturbation.

    Call C_i.PLAYANDUPDATE()

    // Maintain CI of hint's expected reward

    $\mathcal{R}_i \leftarrow$ all reward samples obtained by C_i so far playing unperturbed hint

    Compute confidence interval $Y_i = (\bar{y}_i - b_i, \bar{y}_i + b_i)$ with $\bar{y}_i = \frac{1}{|\mathcal{R}_i|}\sum_{y \in \mathcal{R}_i} y$ and $b_i = \sqrt{\frac{3\ln(40\ln(2|\mathcal{R}_i|)/\delta)}{|\mathcal{R}_i|}}$

  // Eliminate worse hint if possible

  if $Y_+ \cap Y_- = \varnothing$ **then**

    Remove C_i with smaller $\bar{y}_i$ from active set $\mathcal{S}$

until *any instance in* S *returns a value* $r_\perp$ *satisfying* $\frac{0.06}{2\cdot 5^2} \cdot \frac{r_\perp^2}{r} \geq 10Wd\log(T)/\sqrt{T}$ *or* $|S| = 1$ *and the lone instance returns* $r_\perp$

// Phase 3: Commit to hint or ignore it.

For all remaining rounds, call SWITCH$(\widehat{h}, \frac{0.06}{2\cdot 5^2} \cdot \frac{r_\perp^2}{r}, T, Wd\log(T)\sqrt{T})$ from Algorithm 1 with hint $\widehat{h}$ randomly chosen from a surviving active instance in $\mathcal{S}$

probability at least $1-\delta$*, Algorithm 3 has worst case regret at most* $O(d\sqrt{T}\log(T/\delta))$*. Further, if* $\|\theta^\star\| \geq \max(d, 3734)/\sqrt{T}$ *, then the hint based regret simultaneously satisfies* $R_h^T = O(\sqrt{T}\log(T/\delta))$*.*

5. Other Trade-offs on the Pareto Frontier

Unsurprisingly, we can in fact achieve anything on the Pareto frontier of the hint-based vs total regret. Specifically, we show that we can maintain a hint-based regret of $R_h^T = O(G\log(T))$ while maintaining a worse-case regret of $R^T = O(dT\log(T)/G)$. As shown by our lower bounds, this tradeoff is in fact tight, up to log factors. First, we prove a more generic version of Lemma 9 by simply adjusting the values of Δ^2 and removing the dependence on $\|P_h^\perp a\|$ via a worst-case analysis.

Lemma 12 *Suppose* R *is a constant factor approximation to* $\|\theta^\star\|$ *and* $G = \|\theta^\star\|\Delta^2$*. Running* ESTIMATENORM$(p, \Delta g)$*, with at least constant probability over random* g*, will incur a hint-based regret of*

$$R_h^T = O\left(GT\log(T) + \frac{\log(T)}{G}\right)$$

Furthermore, if $\langle h, \theta^\star\rangle \geq -\|\theta^\star\|/4$*, then with probability at least* 0.7*, we maintain a worst case regret of*

$$R^T = O\left(dGT\log(T) + \frac{d\log(T)}{G}\right)$$

We note that there is a general segment in our algorithm that plays both $h, -h$ and requires running a 2-arm bandit process that chooses between $h, -h$. Specifically, we can view the $h, -h$ as the arms and each pull of the arm corresponds to playing the hint and its perturbation, encapsulated by a call to PLAYANDUPDATE. While the pull plays the perturbed action, the returned reward ignores the perturbed action and is simply the reward of playing the hint. By modifying the bandit algorithm to favor choosing h by following the hint-based MAB algorithm presented in Lattimore (2015), we can bound the hint-based regret and worst case regret tradeoff and surprisingly derive the same Pareto frontier upper bounds for linear bandits as for multi-armed bandits when $k = d$ (see Algorithm 5).

Theorem 13 (Pareto Frontier) *Let* $G \leq \sqrt{T}$ *and* $\|\theta^\star\| \geq d/G$*. Then, there is an alteration of Algorithm 3 with constants* $c_0 > c_1$ *(Algorithm 5 with MAB algorithm described in Lattimore (2015)) that has hint based regret*

$$R_h^T = O(G\log(T))$$

and worst case regret at most

$$R^T = O(dT\log(T)/G)$$

6. Multiple Hints

Suppose that we are given m separate hints $h_1, ..., h_m$ and our goal is to perform as well as the best hint, while maintaining standard worst-case regret bounds even when none of the hints are trustworthy. Specifically, let $h^\star$ be the best hint, then we want to bound $R^T_{h^*}$ and R^T. For now, m should be a small polynomial power of d and if m gets too large, it will likely become futile to perform hint selection.

For now, consider simply selecting the best hint in a multi-armed bandit setting, where each arm is the hint $h_1, ...h_m$, and assume our loss is still linear. Note that the best known algorithms give an upper bound and matching lower bound of $R^T_{h^\star} = O(\sqrt{mT})$, up to $\log$ factors, at least in the regime when $m = o(d)$. Since we want to perform well with respect to the best hint, our algorithm will rely on a multi-armed bandit analysis, where each arm is a separate hint. However, we cannot spend all of our sample budget on selecting among $\{h_1, ..., h_m\}$ as we may incur large worst case regret since our hints may have similar performance but all of them are incur large instantaneous regret.

Therefore, we will necessarily have to add a suboptimality estimation phase for each hint, at least until the hint is clearly not $h^\star$. A naive analysis would suggest that this may imply that $R^T_{h^\star} = O(m\sqrt{T})$ but we show surprisingly that we can achieve a sublinear rate $R^T_{h^\star} = \widetilde{O}(m^{2/3}\sqrt{T})$ while maintaining the usual worst case guarantees of $\widetilde{O}(d\sqrt{T})$. This implies that we can perform some tradeoff between the sample complexity of two elimination forces: the first coming from a suboptimality compared to $h^\star$ and the second coming from suboptimality compared to $a^\star$.

Remark: As noted throughout the paper, we can compare to the multi-armed bandit setting, in which we have m hint arms $h_1, ...h_m$ that we want to perform well with respect to, as well as some additional $k - m$ arms $a_{m+1}, ..., a_k$. Note that the optimal action $a^\star$ may not be a hint arm. This specific problem was studied in Lattimore (2015) and a quick calculation shows that there are matching upper and lower bounds that gives $R^T_{h^*} = \Theta(\sqrt{mT})$, while maintaining a worst case regret of $R^T = \Theta(\frac{k}{\sqrt{m}}\sqrt{T})$. Note that the product of the hint-based and worst case regret is still $\Omega(kT)$.

Throughout this section, we will simplify and focus on the main ideas by assuming that $\|\theta^\star\| = \Omega(1)$. Our main algorithm is present in the appendix (see Algorithm 6). By following the algorithmic alterations in the previous section, we note that adding a $\|\theta^\star\|$ estimation step is straightforward and one might be able to apply suitable tradeoffs by using the hint-based MAB algorithm in the remark above to obtain a full Pareto frontier.

Lemma 14 *Assume there are absolute constants $c_1, c_2 \in \mathbb{R}^+$ such that $c_1 \leq \|\theta\| \leq c_2$ and assume Algorithm 6 is called with a hint set $\mathcal{H}$ that contains $-h$ for each $h \in \mathcal{H}$. With high constant probability, the regret of Algorithm 6 with $|\mathcal{H}| = m \leq d$ after T rounds is bounded as*

$$R^T = \widetilde{O}(d\sqrt{T}).$$

Lemma 15 *By setting $B = m^{1/3}$ and c_0 sufficiently large in Algorithm 6, the hint-based regret of our algorithm is bounded with high probability by*

$$R^T_{h^*} = \widetilde{O}(m^{2/3}\sqrt{T})$$

7. Conclusion

The problem of improving best-case performance without sacrificing worst-case performance has proven to be surprisingly difficult in the bandit setting, in comparison to the full-information setting. In this paper, we provided an intriguing positive construction for the case of stochastic linear bandits over the unit ball. We show that, when provided with a sufficiently high-quality *hint* for the optimal action, we are able to significantly improve the regret to only $\widetilde{O}(\sqrt{T})$ larger than what would be incurred by simply playing the hint for all time steps, while gracefully decaying to the worst-case optimal bound of $\widetilde{O}(d\sqrt{T})$ when the hint is only mediocre *without prior knowledge of the hint quality*. We provide lower bound demonstrating optimality of our construction, but some open questions remain. Notably, our improved rates only arise when the parameter $\theta^\star$ is sufficiently large, although we show in Theorem 2 that even with large $\theta^\star$, a hint is still required to go beyond worst-case bounds. Furthermore, the Pareto frontier is poorly understood when the actions sets have different geometry or in the multi-hint case when m is significantly greater than d. It is our hope that the techniques presented here may shed an optimistic light on when it is possible to go beyond worst-case in further bandit problems.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Jacob Abernethy, Elad E Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, pages 263–273, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

Spyros Angelopoulos, Christoph Dürr, Shendan Jin, Shahin Kamali, and Marc Renault. Online computation with untrusted advice. *arXiv preprint arXiv:1905.05655*, 2019.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.

Aurélien F Bibaut, Antoine Chambaz, and Mark J van der Laan. Rate-adaptive model selection over a collection of black-box contextual bandit algorithms. *arXiv preprint arXiv:2006.03632*, 2020.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854. PMLR, 2020.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, 2021.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32:14741–14752, 2019.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Problem-complexity adaptive model selection for stochastic linear bandits. *arXiv preprint arXiv:2006.02612*, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*. PMLR, 2021.

Tor Lattimore. The pareto regret frontier for bandits. *Advances in Neural Information Processing Systems*, 28:208–216, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Thodoris Lykouris and Sergei Vassilvtiskii. Competitive caching with machine learned advice. In *International Conference on Machine Learning*, pages 3296–3305. PMLR, 2018.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ml predictions. *Advances in Neural Information Processing Systems*, 31:9661–9670, 2018.

Matthew Streeter and H Brendan McMahan. No-regret algorithms for unconstrained online convex optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pages 2402–2410, 2012.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

Alexander Wei and Fred Zhang. Optimal robustness-consistency trade-offs for learning-augmented online algorithms. *arXiv preprint arXiv:2010.11443*, 2020.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pages 3583–3634. PMLR, 2020.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

Contents of main article and appendix

1 Introduction **1**
1.1 Our Contributions . . . 2
1.2 Related work . . . 3
1.2.1 Conservative Exploration . . . 3
1.2.2 Tuning the Learning Rate in Mirror Descent . . . 3
1.2.3 Priors with Upper Confidence Bound . . . 4
1.2.4 Model Selection . . . 4
1.2.5 Hints in Adversarial Context . . . 4

2 Problem Setting and Notation **5**

3 Regret Lower Bounds **6**

4 New Algorithm for Stochastic Linear Bandits with Action Hint **7**
4.1 Estimating Regret of Hint Action . . . 7

5 Other Trade-offs on the Pareto Frontier **12**

6 Multiple Hints **13**

7 Conclusion **14**

A Proofs of Regret Lower Bounds **18**

B Proof of Regret Upper Bounds **25**
B.1 Algorithms and proofs for Pareto and Multi-Hint settings . . . 32

C Auxiliary Technical Lemmas **36**

Appendix A. Proofs of Regret Lower Bounds

Theorem 1 *For any action hint $h \in \mathcal{A} \subseteq \mathbb{R}^{d+1}$ with $\|h\| = 1$, horizon T and learning algorithm, there is a family Θ of stochastic linear bandit instances so that the following holds. All parameters $\theta \in \Theta$ have $\Theta(1)$ norm. Let $R_h = \max_{\theta\in\Theta} \mathbb{E}[\mathrm{Reg}_h^\theta(T)]$ be largest hint-based regret in family Θ. Then there is a bandit instance $\theta \in \Theta$ such that the expected regret w.r.t. the best arm is*

$$\mathbb{E}[\mathrm{Reg}^\theta(T)] \geq \frac{1}{4} + \frac{T}{4} \min\left(1, \frac{d-1}{R_h}\right) \geq \Omega\left(\frac{dT}{R_h}\right)$$

Proof Without loss of generality, let $h = e_1$ be axis-aligned for ease of notation. Set $\theta_0 = \rho h$ for a $\rho \in \mathbb{R}^+$ defined later and consider the following family of linear bandit instances, identified by their reward parameters,

$$\Theta = \{\theta_0\} \cup \{\theta_0 + \Delta e_i, \theta_0 - \Delta e_i \colon i = 2, 3, \ldots, d\} ,$$

where $\Delta \in \mathbb{R}^+$ will be defined later. Further, the reward distribution for any action a is a standard normal random variable centered at $\langle a, \theta\rangle$. Essentially, this family contains θ_0 and parameters that deviate from it in any one dimension (except the first). First note that since $\theta_0 \in \Theta$, we have

$$\mathbb{E}[\mathrm{Reg}_h^{\theta_0}(T)] = \sum_{t=1}^{T} \rho(1 - \mathbb{E}_{\theta_0}[a_{t1}]) \leq R_h \tag{1}$$

where a_{t1} is the first component of the action a_t played in round t by the algorithm.

Next, we consider any other instance $\theta^\star = \theta_0 \pm \Delta e_i \in \Theta$ for some i. Since the action space is the unit ball, the optimal action in this instance is

$$a^\star = \frac{\theta^\star}{\|\theta^\star\|} = \frac{\theta^\star}{\sqrt{\rho^2 + \Delta^2}} = \frac{\theta^\star}{r},$$

where $r = \sqrt{\rho^2 + \Delta^2}$. The expected regret after T rounds in bandit instance $\theta^\star$ is

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] &= \sum_{t=1}^{T} \mathbb{E}\left[\langle a^\star - a_t, \theta^\star\rangle\right] \\
&= \sum_{t=1}^{T} \rho\left(\frac{\rho}{r} - \mathbb{E}[a_{t1}]\right) + \sum_{t=1}^{T} \Delta\mathbb{E}\left(\frac{\Delta}{r} - a_{ti} \cdot \mathrm{sign}(\theta_i^\star)\right) \\
&= \sum_{t=1}^{T} \rho\left(\frac{\rho}{r} - \mathbb{E}[b_{t1}]\right) + \sum_{t=1}^{T} \Delta\mathbb{E}\left(\frac{\Delta}{r} - b_{ti}\right) \qquad (2)
\end{aligned}$$

where $b_{t1} = a_{t1}$ and $b_{ti} = a_{ti} \cdot \mathrm{sign}(\theta_i^\star)$ and expectations are taken w.r.t. the bandit instance $\theta^\star$. Notice that

$$\frac{r}{2\Delta}\left(\frac{\Delta}{r} - b_{ti}\right)^2 = \left[\frac{\Delta}{2r} - b_{ti} + \frac{r}{2\Delta} b_{ti}^2\right] \overset{(i)}{\leq} -\frac{\Delta}{2r} + \frac{r}{2\Delta}(1 - b_{t1}^2) + \frac{\Delta}{r} - b_{ti}$$

where step (i) follows from $1 \geq \|b_t\|^2 = b_{t1}^2 + \sum_{i=2}^{d} b_{ti}^2$. Now, denote $X_t := \frac{\rho}{r} - b_{t1}$ and $Y_t := -\frac{\Delta}{2r} + \frac{r}{2\Delta}(1 - b_{t1}^2)$. Then

$$
\begin{aligned}
2Y_t &= -\frac{\Delta}{r} + \frac{r}{\Delta}\left(1 - \left(\frac{\rho}{r} - X_i\right)^2\right) = -\frac{\Delta}{r} + \frac{r}{\Delta}\left(1 - \frac{\rho^2}{r^2} + 2\frac{\rho}{r}X_t - X_t^2\right) \\
&= \underbrace{-\frac{\Delta}{r} + \frac{r}{\Delta}\left(1 - \frac{\rho^2}{r^2}\right)}_{=0} + 2\frac{\rho}{\Delta}X_t - \frac{r}{\Delta}X_t^2 = 2\frac{\rho}{\Delta}X_t - \frac{r}{\Delta}X_t^2
\end{aligned}
$$

and thus $Y_t = \frac{\rho}{\Delta}X_t - \frac{r}{2\Delta}X_t^2$. Using these identities, we rewrite the regret in Equation 6 as

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] &\geq \sum_{t=1}^{T} \rho\mathbb{E}[X_t] + \sum_{t=1}^{T} \Delta\mathbb{E}\left(\frac{r}{2\Delta}\left[\frac{\Delta}{r} - b_{ti}\right]^2 - Y_t\right) \\
&= \sum_{t=1}^{T} \rho\mathbb{E}[X_t] + \sum_{t=1}^{T} \mathbb{E}\left(\frac{r}{2}\left[\frac{\Delta}{r} - b_{ti}\right]^2 - \rho X_t + \frac{r}{2}X_t^2\right) \\
&= \frac{r}{2}\sum_{t=1}^{T} \mathbb{E}\left[X_t^2\right] + \frac{r}{2}\sum_{t=1}^{T} \mathbb{E}\left[\left(\frac{\Delta}{r} - b_{ti}\right)^2\right] \\
&\geq \frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\Delta}{r} - b_{ti}\right)^2\right].
\end{aligned}
$$

We now follow the analysis of Theorem 24.2 by Lattimore and Szepesvári (2020). Define $\tau_i = T \wedge \min\{t\colon \sum_{s=1}^{t} a_{si}^2 \geq T\Delta^2/r^2\}$, which is a stopping time and $U_i(x) = \sum_{t=1}^{\tau_i}\left(\frac{\Delta}{r} - a_{ti}x\right)^2$ for $x \in \{-1, +1\}$. Then we can lower-bound the expression above as

$$
\frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\Delta}{r} - b_{ti}\right)^2\right] \geq \frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{\tau_i}\left(\frac{\Delta}{r} - b_{ti}\right)^2\right] = \frac{r}{2}\mathbb{E}\left[U_i(\operatorname{sign}\theta_i^\star)\right]. \tag{3}
$$

Let $\mathbb{P}_0$ be the action distribution of the algorithm in bandit instance θ_0 and $\mathbb{P}_i$ be the action distribution in bandit instance $\theta^\star = \theta_0 + \Delta e_i$ up to round τ_i. By Pinsker's inequality, we have

$$
|\mathbb{E}_{\theta^\star}[U_i(1)] - \mathbb{E}_{\theta_0}[U_i(1)]| \leq \sqrt{\frac{1}{2}D(\mathbb{P}_0, \mathbb{P}_i)}\sup U_i(1) \leq \left(4T\frac{\Delta^2}{r^2} + 2\right)\sqrt{\frac{1}{2}D(\mathbb{P}_0, \mathbb{P}_i)}
$$

where the second inequality follows from the following bound on $U_i(1)$

$$
U_i(1) = \sum_{t=1}^{\tau_i}\left(\frac{\Delta}{r} - a_{ti}\right)^2 \leq 2\tau_i\frac{\Delta^2}{r^2} + 2\sum_{t=1}^{\tau_i} a_{ti}^2 \leq 4T\frac{\Delta^2}{r^2} + 2.
$$

Since $\theta^\star$ and θ_0 only differ on the i-th coordinate, and the noise in the observed rewards for any action has a standard normal distribution, we can bound the KL divergence of the output distribution of the t-th action as $\frac{\Delta^2}{2}a_{ti}^2$. Now, by the chain rule, we bound

$$
D(\mathbb{P}_0, \mathbb{P}_i) \leq \frac{\Delta^2}{2}\mathbb{E}_{\theta_0}\left[\sum_{t=1}^{\tau_i} a_{ti}^2\right]
$$

Consider now the index $j \in \{2, 3, \ldots, d\}$ which minimizes $\mathbb{E}_{\theta_0}\left[\sum_{t=1}^{\tau_i} a_{tj}^2\right]$. For this index, we can further bound

$$D(\mathbb{P}_0, \mathbb{P}_j) \leq \frac{\Delta^2}{2}\frac{1}{d-1}\sum_{i=2}^{d}\mathbb{E}_{\theta_0}\left[\sum_{t=1}^{\tau_i} a_{ti}^2\right] \leq \frac{\Delta^2}{2}\frac{1}{d-1}\sum_{t=1}^{\tau_i}\mathbb{E}_{\theta_0}[1-a_{t1}^2] \leq \frac{\Delta^2}{d-1}\sum_{t=1}^{\tau_i}\mathbb{E}_{\theta_0}[1-a_{t1}]$$

where the last inequality follows from $1 - x^2 \leq 2(1-x)$ for $x \in [-1, 1]$. Finally, the RHS above can be upper-bounded by $\frac{\Delta^2}{d-1}\frac{R_h}{\rho}$ using Equation 1. Thus, we have shown that

$$|\mathbb{E}_{\theta^\star}[U_j(1)] - \mathbb{E}_{\theta_0}[U_j(1)]| \leq \left(4T\frac{\Delta^2}{r^2} + 2\right)\sqrt{\frac{\Delta^2 R_h}{2\rho(d-1)}}$$

and by symmetry, we can also show that the same upper-bound holds for $|\mathbb{E}_{\theta^\star_-}[U_j(-1)] - \mathbb{E}_{\theta_0}[U_j(-1)]|$ where $\theta^\star_- = \theta_0 - \Delta e_j$. Therefore, we have

$$\begin{aligned}
\mathbb{E}_{\theta^\star}[U_j(1)] + \mathbb{E}_{\theta^\star_-}[U_j(-1)] &\geq \mathbb{E}_{\theta_0}[U_j(1) + U_j(-1)] - 2\left(4T\frac{\Delta^2}{r^2} + 2\right)\sqrt{\frac{\Delta^2 R_h}{2\rho(d-1)}} \\
&= 4\mathbb{E}_{\theta_0}\left[\frac{\tau_j\Delta^2}{r^2} + \sum_{t=1}^{\tau_j} a_{ti}^2\right] - 2\left(4T\frac{\Delta^2}{r^2} + 2\right)\sqrt{\frac{\Delta^2 R_h}{2\rho(d-1)}} \\
&\geq 2\left(4T\frac{\Delta^2}{r^2} + 2\right) - 2\left(4T\frac{\Delta^2}{r^2} + 2\right)\sqrt{\frac{\Delta^2 R_h}{2\rho(d-1)}}
\end{aligned}$$

Now, set $\Delta^2 = \frac{(d-1)\rho}{2R_h}$, yielding:

$$\begin{aligned}
\mathbb{E}_{\theta^\star}[U_j(1)] + \mathbb{E}_{\theta^\star_-}[U_j(-1)] &\geq 4T\frac{\Delta^2}{r^2} + 2 \\
&= 4T\frac{(d-1)\rho}{2R_h r^2} + 2
\end{aligned}$$

Use $r^2 = \Delta^2 + \rho^2 = \frac{(d-1)\rho + 2R_h\rho^2}{2R_h}$:

$$= \frac{4T(d-1)\rho}{(d-1)\rho + 2R_h\rho^2} + 2$$

Now, we need to set ρ in order to both enforce the constraint $r^2 \leq 1$, and also achieve our desired regret lower bounds. To this end, we consider two cases, either $d - 1 \leq R_h$ or not. If $d - 1 \leq R_h$, then set $\rho = \frac{1}{2}$ to obtain $r^2 = \frac{\frac{d-1}{2} + \frac{R_h}{2}}{2R_h} \leq \frac{d-1}{4R_h} + \frac{1}{4} \leq 1/2 \leq 1$. In this case, we also have

$$\begin{aligned}
\mathbb{E}_{\theta^\star}[U_j(1)] + \mathbb{E}_{\theta^\star_-}[U_j(-1)] &\geq \frac{4T(d-1)}{(d-1) + R_h} + 2 \\
&\geq \frac{2T(d-1)}{R_h} + 2
\end{aligned}$$

This implies that at least one of $\mathbb{E}_{\theta^\star}[U_j(1)]$ or $\mathbb{E}_{\theta^\star_-}[U_j(-1)]$ is greater than $\frac{T(d-1)}{R_h}+1$. Thus by Equation 3, there is some $\theta^\star$ such that

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] &\geq \frac{r}{2}\left(\frac{T(d-1)}{R_h}+1\right)\\ &\geq \frac{T(d-1)}{4R_h}+\frac{1}{4}\end{aligned}$$

where we have used $r \geq \rho = 1/2$.

Otherwise (if $d-1 \geq R_h$), set $\rho = \frac{R_h}{2(d-1)}$ to obtain $r^2 = \frac{R_h}{2(d-1)}$ to obtain $r^2 = \frac{R_h/2+R_h^3/2(d-1)^2}{2R_h} \leq \frac{1}{2}+\frac{R_h^2}{2(d-1)} \leq 1$ again. Further, this also implies the upper bound $r^2 \geq 1/4$. Thus, we now have:

$$\begin{aligned}\mathbb{E}_{\theta^\star}[U_j(1)]+\mathbb{E}_{\theta^\star_-}[U_j(-1)] &\geq \frac{2TR_h}{R_h/2+R_h^3/2(d-1)^2}+2\\ &\geq 2T+2\end{aligned}$$

This implies that there is some $\theta^\star$ for which:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] &\geq \frac{r}{2}(T+1)\\ &\geq \frac{T}{4}+\frac{1}{4}\end{aligned}$$

Putting both cases together provides the desired lower bound.

■

Lemma 16 *Let $\theta_0 \in \mathbb{R}^{d+1}$, $T \in \mathbb{N}$ and $\Delta \in (0, \sqrt{3/4})$. Then for any algorithm, there is a linear bandit instance with unit-ball action set ($\mathcal{A} = \{a \in \mathbb{R}^{d+1} \colon \|a\|_2 \leq 1\}$) and parameter $\theta^\star \in \mathbb{R}^{d+1}$ with $\|\theta^\star - \theta_0\|_2 = \Delta$ such that the expected regret after T rounds is at least*

$$\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] \geq \frac{T\Delta}{2\sqrt{1+\frac{\|\theta_0\|^2}{\Delta^2}}} - \sqrt{d}\Delta\frac{\sqrt{\Delta^2+\|\theta_0\|^2}}{8}\left(\frac{4T}{\frac{d\|\theta_0\|^2}{\Delta^2}+d}+2\right)\sqrt{\frac{T}{\frac{d\|\theta_0\|^2}{\Delta^2}+d}+1} \quad (4)$$

$$\geq \frac{T\Delta^2}{2} - \frac{\sqrt{d}\Delta}{4}\left(\frac{32}{d}T\Delta^2+1\right)^{3/2} \quad (5)$$

where the second form holds when $\Delta \leq \sqrt{\frac{3}{4}}$ and $\|\theta_0\| = 1/2$.

Proof Let $\bar{\Delta} = \frac{\Delta}{\sqrt{d}}$. Without loss of generality, let $\theta_0 = \theta e_1$ be axis-aligned where $\theta = \|\theta_0\|_2$. We consider parameters $\theta^\star \in \{(\theta, \pm\bar{\Delta}, ..., \pm\bar{\Delta}) \in \mathbb{R}^{d+1}\}$. Since the action space is the unit ball, the our optimal action is

$$a^\star = \frac{\theta^\star}{\|\theta^\star\|} = \frac{\theta^\star}{\sqrt{\theta^2+\Delta^2}} = \frac{\theta^\star}{r},$$

where $r = \sqrt{\theta^2 + \Delta^2}$ is the norm of all considered optimal parameters. The expected regret after T rounds in bandit instance $\theta^\star$ is

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] &= \sum_{t=1}^{T} \mathbb{E}\left[\langle a^\star - a_t, \theta^\star\rangle\right] \\
&= \sum_{t=1}^{T} \theta \left(\frac{\theta}{r} - \mathbb{E}[a_{t1}]\right) + \sum_{t=1}^{T}\sum_{i=2}^{d+1} \bar{\Delta}\mathbb{E}\left(\frac{\bar{\Delta}}{r} - a_{ti}\cdot \mathrm{sign}(\theta_i^\star)\right) \\
&= \sum_{t=1}^{n} \theta \left(\frac{\theta}{r} - \mathbb{E}[a_{t1}]\right) + \sum_{t=1}^{T}\sum_{i=2}^{d+1} \bar{\Delta}\mathbb{E}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right) \qquad (6)
\end{aligned}$$

where $b_{t1} = a_{t1}$ and $b_{ti} = a_{ti}\cdot \mathrm{sign}(\theta_i^\star)$ for $2 \le i \le d+1$. Notice that

$$\frac{r}{2\bar{\Delta}}\sum_{i=2}^{d+1}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right)^2 = \sum_{i=2}^{d+1}\left[\frac{\bar{\Delta}}{2r} - b_{ti} + \frac{r}{2\bar{\Delta}}b_{ti}^2\right] \overset{(i)}{\le} -\frac{\bar{\Delta}d}{2r} + \frac{r}{2\bar{\Delta}}(1 - b_{t1}^2) + \sum_{i=2}^{d+1}\left[\frac{\bar{\Delta}}{r} - b_{ti}\right],$$

where step (i) follows from $1 \ge \|b_t\|^2 = b_{t1}^2 + \sum_{i=2}^{d+1} b_{ti}^2$. Now, denote $X_t := \frac{\theta}{r} - b_{t1}$ and $Y_t := -\frac{\bar{\Delta}d}{2r} + \frac{r}{2\bar{\Delta}}(1 - b_{t1}^2)$. Then

$$\begin{aligned}
2Y_t &= \frac{\bar{\Delta}d}{r} + \frac{r}{\bar{\Delta}}\left(1 - \left(\frac{\theta}{r} - X_i\right)^2\right) = \frac{\bar{\Delta}d}{r} + \frac{r}{\bar{\Delta}}\left(1 - \frac{\theta^2}{r^2} + 2\frac{\theta}{r}X_t - X_t^2\right) \\
&= \underbrace{\frac{\bar{\Delta}d}{r} + \frac{r}{\bar{\Delta}}\left(1 - \frac{\theta^2}{r^2}\right)}_{=0} + 2\frac{\theta}{\bar{\Delta}}X_t - \frac{r}{\bar{\Delta}}X_t^2 = 2\frac{\theta}{\bar{\Delta}}X_t - \frac{r}{\bar{\Delta}}X_t^2
\end{aligned}$$

and thus $Y_t = \frac{\theta}{\bar{\Delta}}X_t - \frac{r}{2\bar{\Delta}}X_t^2$. Using these identities, we rewrite the regret in Eq. (6) as

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] &\ge \sum_{t=1}^{T}\theta\mathbb{E}[X_t] + \sum_{t=1}^{T}\bar{\Delta}\mathbb{E}\left(\sum_{i=2}^{d+1}\frac{r}{2\bar{\Delta}}\left[\frac{\bar{\Delta}}{r} - b_{ti}\right]^2 - Y_t\right) \\
&= \sum_{t=1}^{T}\theta\mathbb{E}[X_i] + \sum_{t=1}^{T}\mathbb{E}\left(\frac{r}{2}\sum_{i=2}^{d+1}\left[\frac{\bar{\Delta}}{r} - b_{ti}\right]^2 - \theta X_t + \frac{r}{2}X_t^2\right) \\
&= \frac{r}{2}\sum_{t=1}^{T}\mathbb{E}\left[X_t^2\right] + \sum_{i=2}^{d+1}\frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right)^2\right] \\
&\ge \sum_{i=2}^{d+1}\frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right)^2\right].
\end{aligned}$$

We now follow the analysis of Theorem 24.2 by Lattimore and Szepesvári (2020). Define $\tau_i = T \wedge \min\{t\colon \sum_{s=1}^{t} a_{si}^2 \ge T\bar{\Delta}^2/r^2\}$, which is a stopping time and $U_i(x) = \sum_{t=1}^{\tau_i}\left(\frac{\bar{\Delta}}{r} - a_{ti}x\right)^2$ for $x \in \{-1, +1\}$. Then we can lower-bound the expression above as

$$\sum_{i=2}^{d+1}\frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right)^2\right] \ge \sum_{i=2}^{d+1}\frac{r}{2}\mathbb{E}\left[\sum_{t=1}^{\tau_i}\left(\frac{\bar{\Delta}}{r} - b_{ti}\right)^2\right] = \sum_{i=2}^{d+1}\frac{r}{2}\mathbb{E}\left[U_i(\mathrm{sign}\,\theta_i^\star)\right].$$

Let θ' an alternative to $\theta^\star$. By Pinsker's inequality, we have

$$\mathbb{E}_{\theta^\star}[U_i(1)] \geq \mathbb{E}_{\theta'}[U_i(1)] - \sqrt{\frac{1}{2}D(\mathbb{P}^\star, \mathbb{P}')}\sup U_i(1) \geq \mathbb{E}_{\theta'}[U_i(1)] - \left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{1}{2}D(\mathbb{P}^\star, \mathbb{P}')}$$

where the second inequality follows from the following bound on $U_i(1)$

$$U_i(1) = \sum_{t=1}^{\tau_i}\left(\frac{\bar{\Delta}}{r} - a_{ti}\right)^2 \leq 2\tau_i\frac{\bar{\Delta}^2}{r^2} + 2\sum_{t=1}^{\tau_i} a_{ti}^2 \leq 4T\frac{\bar{\Delta}^2}{r^2} + 2.$$

We bound $D(\mathbb{P}^\star, \mathbb{P}') \leq \mathbb{E}_{\theta^\star}[T_i(\tau_i)]D(\mathbb{P}_i^\star, \mathbb{P}_i') \leq \frac{\bar{\Delta}^2}{2}\mathbb{E}_{\theta^\star}\left[\sum_{t=1}^{\tau_i} a_{ti}^2\right] \leq \frac{\bar{\Delta}^2}{2}\left(\frac{T\bar{\Delta}^2}{r^2} + 1\right)$. Then we have

$$\begin{aligned}
\mathbb{E}_{\theta^\star}[U_i(1)] + \mathbb{E}_{\theta'}[U_i(-1)] &\geq \mathbb{E}_{\theta'}[U_i(1) + U_i(-1)] - \frac{\bar{\Delta}}{2}\left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{T\bar{\Delta}^2}{r^2} + 1} \\
&= 2\mathbb{E}_{\theta'}\left[\frac{\tau_i\bar{\Delta}^2}{r^2} + \sum_{t=1}^{\tau_i} a_{ti}^2\right] - \frac{\bar{\Delta}}{2}\left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{T\bar{\Delta}^2}{r^2} + 1} \\
&\geq 2T\frac{\bar{\Delta}^2}{r^2} - \frac{\bar{\Delta}}{2}\left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{T\bar{\Delta}^2}{r^2} + 1} =: G.
\end{aligned}$$

We denote by $G = 2T\frac{\bar{\Delta}^2}{r^2} - \frac{\bar{\Delta}}{2}\left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{T\bar{\Delta}^2}{r^2} + 1}$. Then by the randomization hammer:

$$\begin{aligned}
\sum_{\theta' \in \{\pm\bar{\Delta}\}^d} \mathbb{E}[\mathrm{Reg}_{\theta'}(T)] &\geq \frac{r}{2}\sum_{\theta' \in \{\pm\bar{\Delta}\}^d}\sum_{i=2}^{d+1}\mathbb{E}_{\theta'}\left[U_i(\operatorname{sign}\theta_i')\right] \\
&= \frac{r}{2}\sum_{i=2}^{d+1}\sum_{\theta'_{-i} \in \{\pm\bar{\Delta}\}^{d-1}}\sum_{\theta'_i \in \{\pm\bar{\Delta}\}}\mathbb{E}_{\theta'}\left[U_i(\operatorname{sign}\theta_i')\right] \\
&\geq \frac{r}{2}\sum_{i=2}^{d+1}\sum_{\theta'_{-i} \in \{\pm\bar{\Delta}\}^{d-1}} G = 2^d\frac{Grd}{4}.
\end{aligned}$$

Since the average regret over all θ' is at least $Grd/4$, there is at least one $\theta^\star$ such that

$$\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] \geq \frac{Grd}{4} = \frac{Td\bar{\Delta}^2}{2r} - \frac{rd\bar{\Delta}}{8}\left(4T\frac{\bar{\Delta}^2}{r^2} + 2\right)\sqrt{\frac{T\bar{\Delta}^2}{r^2} + 1} \tag{7}$$

$$= \frac{T\Delta^2}{2r} - \frac{r\sqrt{d}\Delta}{8}\left(4T\frac{\Delta^2}{dr^2} + 2\right)\sqrt{\frac{T\Delta^2}{dr^2} + 1} \tag{8}$$

$$\geq \frac{T\Delta^2}{2r} - \frac{r\sqrt{d}\Delta}{4}\left(2T\frac{\Delta^2}{dr^2} + 1\right)^{3/2}. \tag{9}$$

Assume now that $0 \leq \Delta \leq \sqrt{\frac{3}{4}}$. Then $r \in [0.25, 1]$ and the lower-bound above simplifies to

$$\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] \geq \frac{T\Delta^2}{2} - \frac{\sqrt{d}\Delta}{4}\left(\frac{32}{d}T\Delta^2 + 1\right)^{3/2}. \tag{10}$$

■

Theorem 2 *Let $T \geq \min(d^2/2^8, 2^6)$ and $\Delta^2 = \frac{d}{2^8\sqrt{T}}$. For any hint $h \in \mathbb{R}^{d+1}$ with $\|h\| = 1$ and any algorithm, there is a linear bandit instance with parameter $\theta^\star \in \mathbb{R}^{d+1}$ satisfying $\|\theta^\star - \frac{h}{2}\| \leq \Delta$ with optimal action $a^\star = \frac{\theta^\star}{\|\theta^\star\|}$ satisfying $\|a^\star - h\| \leq 4\Delta$ and $r(a^\star, h) = \langle \theta^\star, a^\star - h\rangle \leq 972\Delta^2 \leq O(d/\sqrt{T})$ such that*

$$\mathbb{E}[\mathrm{Reg}^{\theta^\star}(T)] \geq \frac{d\sqrt{T}}{512}$$

Proof Notice that our assumptions on Δ, T and h imply $\|h/2\| = 1/2$ and $\Delta \leq \frac{1}{4} \leq \sqrt{\frac{3}{4}}$. Thus, by Lemma 16, there is some $\theta^\star$ satisfying $\|\theta^\star - \frac{h}{2}\| \leq \Delta$ and $\langle \theta^\star, a^\star - h\rangle \leq \Delta^2$ such that

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_{\theta^\star}(T)] &\geq \frac{T\Delta^2}{2} - \frac{\sqrt{d}\Delta}{4}\left(\frac{32}{d}T\Delta^2 + 1\right)^{3/2} \\
&\geq \frac{T\Delta^2}{2} - \frac{\sqrt{d}\Delta}{4}\left(\frac{\sqrt{T}}{8} + 1\right)^{3/2}
\end{aligned}$$

use $T \geq 2^6$:

$$\begin{aligned}
&\geq \frac{d\sqrt{T}}{256} - \frac{d}{64T^{1/4}}\left(\frac{\sqrt{T}}{4}\right)^{3/2} \\
&= \frac{d\sqrt{T}}{512}
\end{aligned}$$

Furthermore, notice that $1/2 - \Delta = \|h/2\| - \Delta \leq \|\theta^\star\| \leq \|h/2\| + \Delta = 1/2 + \Delta$. Now, since $f(x) = 1/x$ is convex for positive x, we have:

$$\frac{1}{1/2 + \Delta} \geq 2 - \Delta$$

Further, by our conditions on T, $\Delta \leq 1/4$. Thus $|f'(x)| \leq 2$ for $x \in [1/2 - \Delta, 1/2 + \Delta]$ so that:

$$\frac{1}{1/2 - \Delta} \leq 2 - 2\Delta$$

Putting all this together:

$$\begin{aligned}
\left\|\frac{\theta^\star}{\|\theta^\star\|} - h\right\| &\leq \|h - 2\theta^\star\| + \|\theta^\star\|\left|2 - \frac{1}{\|\theta^\star\|}\right| \\
&\leq 4\Delta
\end{aligned}$$

Finally, we need to bound $\langle \theta^\star, a^\star - h\rangle$. To this end, observe that $\|h - 2\theta^\star\| \leq 2\Delta$. Then, since $\Delta \leq 1/4$ and $\|h\| = 1$, we must have $\langle \theta^\star, h\rangle \geq 2\|\theta^\star\|^2 - 2\Delta\|\theta^\star\| \geq -\|\theta^\star\|/2$. Therefore, by Lemma 4, we have

$$\langle \theta^\star, a^\star - h\rangle \leq 3\frac{\|P_h^\perp \theta^\star\|^2}{\|\theta^\star\|}$$

Now, let us define $\epsilon = h - 2\theta^\star$. Notice that $\|\epsilon\| \le 2\Delta \le 1/2$. Observe that

$$\begin{aligned} P_h^\perp \theta^\star &= \theta^\star - \frac{\langle h, \theta^\star\rangle h}{\|h\|^2} \\ &= \theta^\star - 2\langle h, \theta^\star\rangle \theta^\star - \langle h, \theta^\star\rangle \epsilon \\ &= \theta^\star - 4\|\theta^\star\|^2\theta^\star - (\langle h, \theta^\star\rangle + 2\|\theta^\star\|^2)\epsilon \end{aligned}$$

Now, we have $(1/2 - \Delta)^2 \le \|\theta^\star\|^2 \le (1/2 + \Delta)^2$. Since $\Delta \le 1/4$, this yields $1/4 - \Delta \le \|\theta^\star\| \le 1/4 + \frac{3}{2}\Delta$ so that

$$\|P_h^\perp \theta^\star\| \le 9\Delta$$

Thus overall we obtain $\langle \theta^\star, a^\star - h\rangle \le 972\Delta^2$ ■

Appendix B. Proof of Regret Upper Bounds

Lemma 17 *Let $a \in \mathbb{R}^d$ be any action with $\|a\| = 1$ and $\langle a, \theta^\star\rangle \ge -\|\theta^\star\|/2$. Then the instantaneous regret of this action is bounded as*

$$\frac{1}{2}\frac{\|P_a^\perp\, \theta^\star\|^2}{\|\theta^\star\|} \quad \le \quad \langle a^\star - a, \theta^\star\rangle \quad \le \quad 3\frac{\|P_a^\perp\, \theta^\star\|^2}{\|\theta^\star\|}\,.$$

Proof Note that $P_{\theta^*} a = \frac{\langle a, \theta^*\rangle \theta^*}{\|\theta^*\|^2}$.

$$\begin{aligned} \langle a^* - a, \theta^*\rangle &= \left\langle \frac{\theta^*}{\|\theta^*\|} - a, \theta^* \right\rangle = \left\langle \frac{\theta^*}{\|\theta^*\|} - \frac{\langle a, \theta^*\rangle \theta^*}{\|\theta^*\|^2}, \theta^* \right\rangle = \|\theta^*\| \left\langle \frac{\theta^*}{\|\theta^*\|^2} - \frac{\langle a, \theta^*\rangle \theta^*}{\|\theta^*\|^3}, \theta^* \right\rangle \\ &= \|\theta^*\| \left(1 - \frac{\langle a, \theta^*\rangle}{\|\theta^*\|}\right) \end{aligned}$$

Now, use $\frac{\langle a, \theta^*\rangle}{\|\theta^*\|} \in [-1/2, 1]$ and the observation $1 - x \le 3(1 - |x|)$ for $x \in [-1/2, 1]$:

$$\le 3\|\theta^*\| \left(1 - \frac{|\langle a, \theta^*\rangle|}{\|\theta^*\|}\right) = 3\|\theta^*\| \left(1 - \|P_{\theta^*} a\|\right) = 3\|\theta^*\| \left(1 - \sqrt{1 - \|P_{\theta^*}^\perp a\|^2}\right)$$

now, since $1 - \sqrt{1 - x} \le x$ for $x \in [0, 1]$

$$\le 3\|\theta^*\| \|P_{\theta^*}^\perp a\|^2 = 3\frac{\|P_a^\perp \theta^*\|}{\|\theta^*\|}\,.$$

Conversely, since $1 - x \ge 1 - |x|$ and $\frac{x}{2} \le 1 - \sqrt{1 - x}$, we also have

$$\langle a^* - a, \theta^*\rangle = \|\theta^*\| \left(1 - \frac{\langle a, \theta^*\rangle}{\|\theta^*\|}\right) \ge \|\theta^*\| \left(1 - \frac{|\langle a, \theta^*\rangle|}{\|\theta^*\|}\right) \ge \frac{1}{2}\|\theta^*\| \left(\|P_{\theta^*}^\perp a\|^2\right)\,.$$

Finally use the identity $\|P_{\theta^\star}^\perp a\| \|\theta_\star\| = \|P_a^\perp \theta^\star\| \|a\|$. ■

Proof [Proof of Lemma 6] We can write the regret of a with respect to h as

$$\langle h-a,\theta^\star\rangle = \langle h - \frac{h+p}{\sqrt{1+\|p\|^2}},\theta^\star\rangle = (1-\frac{1}{\sqrt{1+\|p\|^2}})\langle h,\theta^\star\rangle + \frac{1}{\sqrt{1+\|p\|^2}}\langle p,\theta^\star\rangle$$

Since $\langle h,\theta^\star\rangle \le \|h\|\|\theta^\star\| = \|\theta^\star\|$ and for any $y\in(0,1)$ we have $1-\frac{1}{\sqrt{1+y}} \le y \Leftrightarrow (1-y)\sqrt{1+y} \le 1 \Leftrightarrow (1-y^2)(1-y)\le 1$, we can upper-bound this by

$$\le \|p\|^2\|\theta^\star\| + |\langle p,\theta^\star\rangle| \le \|p\|^2\|\theta^\star\| + |\langle p, P_h^\perp\theta^\star\rangle|$$

where the final inequality holds because $\langle p,h\rangle = 0$.

Furthermore, if $p\sim\mathcal{N}(0,\frac{1}{d}\mathbf{I})$, then we see that with probability at least $1-\delta$, $\|p\|^2 \le O(\log(1/\delta))$ and since $\langle p, P_h^\perp a^\star\rangle \sim N(0,\frac{\|P_h^\perp a^\star\|^2}{d})$, we have that with probability at least $1-\delta$,

$$|\langle p, P_h^\perp a^\star\rangle| \le \frac{\|P_h^\perp a^\star\|\sqrt{\log(1/\delta)}}{\sqrt{d}}$$

Therefore, we conclude that with probability at least $1-\delta$, our instantaneous regret bound follows by scaling down p by Δ and applying the same argument in the projected $d-1$ dimension subspace given by the projection $P_h^\perp = \mathbf{I} - hh^\top$. Note that we may rewrite $\|P_h^\perp a^\star\| = \|P_h^\perp\theta^\star\|/\|\theta^\star\|$ to get our final theorem.

■

Proof [Proof of Lemma 7] Since $\langle h, \frac{\theta^\star}{\|\theta^\star\|}\rangle \ge -1/4$ and $\|p\| < 1/8$, we have $\langle a, \frac{\theta^\star}{\|\theta^\star\|}\rangle \ge -1/2$. Therefore, we can apply Lemma 4 and bound the instantaneous regret of a as

$$\langle a^\star - a,\theta^\star\rangle \le 3\|\theta^\star\|\|P_{\theta^\star}^\perp a\|^2 = 3\|\theta^\star\|(1-\|P_{\theta^\star}a\|^2).$$

To complete the proof, it suffices to show that $\|P_{\theta^\star}a\|^2 \ge 1 - 4\frac{\|P_h^\perp\theta^\star\|^2}{\|\theta^\star\|} - \|p\|^2$ which we do in the following. To simplify notation let $\theta^\star = \alpha h + v$ where $\alpha\in\mathbb{R}$ and $v = P_h^\perp\theta^\star \in \mathbb{R}^d$ with $\langle v,h\rangle = 0$. Then bound

$$\begin{aligned}
\|P_{\theta^\star}a\|^2 &= \left\langle \frac{\theta^\star}{\|\theta^\star\|}, a\right\rangle^2 = \frac{1}{\alpha^2+\|v\|^2}\langle \alpha h + v, a\rangle^2 \\
&\ge \frac{\alpha^2}{\alpha^2+\|v\|^2}\langle h,a\rangle^2 - \frac{2}{\alpha^2+\|v\|^2}|\alpha\langle v,a\rangle| + \frac{1}{\alpha^2+\|v\|^2}\langle v,a\rangle^2 \\
&\ge (1-\frac{\|v\|^2}{\alpha^2+\|v\|^2})\langle h,a\rangle^2 - \frac{2|\alpha|}{\alpha^2+\|v\|^2}|\langle v,p\rangle| \\
&\ge \left(1-\frac{\|v\|^2}{\|\theta^\star\|^2}\right)\frac{1}{1+\|p\|^2} - \frac{2|\alpha|\|p\|\|v\|}{\|\theta^\star\|^2}
\end{aligned}$$

using $\frac{1}{1+x}\le 1-x/2$ for $x\in[0,1]$:

$$\begin{aligned}
&\ge 1 - \frac{\|v\|^2}{\|\theta^\star\|^2} - \frac{\|p\|^2}{2} - \frac{2|\alpha|\|p\|\|v\|}{\|\theta^\star\|^2} \\
&\ge 1 - \frac{\|v\|^2}{\|\theta^\star\|^2} - \frac{\|p\|^2}{2} - \frac{2\|p\|\|v\|}{\|\theta^\star\|}
\end{aligned}$$

applying young inequality $2xy \leq \frac{x^2}{\lambda} + \lambda y^2$ with $\lambda = 2$

$$\geq 1 - 4\frac{\|v\|^2}{\|\theta^\star\|^2} - \|p\|^2$$

■

Lemma 18 *With probability at least* 0.7, *Algorithm 2 (*ESTIMATENORM(h, Δ)*) returns r, a constant-factor approximation to $\|P_h^\perp \theta^\star\|$, i.e., $0.06\|P_h^\perp \theta^\star\| \leq r \leq 5\|P_h^\perp \theta^\star\|$, after $\widetilde{O}\left(\frac{d(1+\Delta^2+\|h\|^2)}{\Delta^2\|P_h^\perp \theta^\star\|^2}\right)$ calls to its* PLAYANDUPDATE *procedure and also satisfies $\|p\| \leq 3\Delta$.*

Proof We first show correctness and then bound the number of calls before the a value is returned.

Correctness We first compute the expectation of the sample averages $\bar{y}_n$, $\bar{z}_n$ and $\bar{x}_n$ for all n, assuming p to be fixed,

$$\begin{aligned}
\mathbf{E}[\bar{y}_n] &= \mathbf{E}[y_i] = \langle \theta^\star, h \rangle, \\
\mathbf{E}[\bar{z}_n] &= \mathbf{E}[z_i] \left\langle \theta^\star, \frac{h+p}{\sqrt{\|h\|^2 + \|p\|^2}} \right\rangle = \frac{\langle \theta^\star, h \rangle + \langle \theta^\star, p \rangle}{\sqrt{\|h\|^2 + \|p\|^2}}, \\
\mathbf{E}[\bar{x}_n] &= \mathbf{E}[\bar{z}_n]\sqrt{\|h\|^2 + \|p\|^2} - \mathbf{E}[\bar{y}_n] = \langle \theta^\star, p \rangle \ .
\end{aligned}$$

We can write $\bar{x}_n$ as an average of $x_i = \sqrt{\|h\|^2 + \|p\|^2} z_i - y_i$. Since z_i and y_i are each 1-sub-Gaussian random variables, x_i is $\sqrt{1 + \|h\|^2 + \|p\|^2}$-sub-Gaussian (each after being centered). We can now apply an anytime-version of the standard Hoeffding concentration argument (see Lemma 22) to get that with probability at least 0.9 for all $n \in \mathbb{N}$

$$|\langle \theta^\star, p \rangle - \bar{x}_n| \leq \sqrt{\frac{3(1 + \|h\|^2 + \|p\|^2)\ln(40\ln(2n))}{n}} =: b_n \ . \tag{11}$$

The algorithm returns a value if and only if the magnitude of the empirical average is at least twice the confidence width, i.e., $|\bar{x}_n| \geq 2b_n$. Using this condition and Equation 11, we have

$$\frac{|\bar{x}_n|}{2} = |\bar{x}_n| - \frac{1}{2}|\bar{x}_n| \leq |\bar{x}_n| - b_n \leq |\langle \theta^\star, p \rangle| \leq |\bar{x}_n| + b_n \leq |\bar{x}_n| + \frac{1}{2}|\bar{x}_n| = \frac{3|\bar{x}_n|}{2} \ , \tag{12}$$

that is, $\|\bar{x}_n\|$ is a constant factor approximation of $|\langle \theta^\star, p \rangle|$. We now argue using the distribution of p that $|\langle \theta^\star, p \rangle|$ is a constant factor approximation of $\|P_h^\perp \theta^\star\| \frac{\Delta}{\sqrt{d'}}$. This holds because $\frac{1}{\Delta}p$ is a d'-dimensional isotropic Gaussian random variable (in $\mathbb{R}^d$ when $h = 0$ and the $d-1$-dimensional orthogonal complement of h when $h \neq 0$). We can therefore show that $\frac{\langle \theta^\star, p \rangle^2}{\Delta^2 \|P_h^\perp \theta^\star\|^2}$ follows a χ_1^2 for which we can bound its tail probabilities (see Lemma 21) to get that

$$0.1\frac{\Delta\|P_h^\perp \theta^\star\|}{\sqrt{d'}} \leq |\langle \theta^\star, p \rangle| \leq 2.33\frac{\Delta\|P_h^\perp \theta^\star\|}{\sqrt{d'}} \tag{13}$$

holds with probability at least 0.9. Combining Equation 12 and Equation 13, we get the desired constant-factor approximation of the return value

$$|x_n|\frac{\sqrt{d-1}}{\Delta} \in \left[0.06\|P_h^\perp \theta^\star\|, 5\|P_h^\perp \theta^\star\|\right] \ . \tag{14}$$

Note that this result holds with probability at least 0.8, by taking a union bound over the events of Equation 11 and Equation 13.

Number of rounds. We now bound the number of rounds n until the return condition $|\bar{x}_n| \geq 2b_n$ is satisfied in the events considered above. If this condition is violated in round n, i.e., $|\bar{x}_n| < 2b_n$, then by rearranging this inequality, we have $|\langle \theta^\star, p\rangle| \leq |\bar{x}_n| + b_n < 2b_n + b_n = 3b_n$. Combining this with Equation 13 gives that if the algorithm does not return a value in round n, then

$$\frac{0.1\Delta}{\sqrt{d'}}\|P_h^\perp \theta^\star\| < 3b_n = 3\sqrt{\frac{3(1+\|h\|^2+\|p\|^2)\ln(40\ln(2n))}{n}}$$

and thus $n = O\left(\frac{d(1+\|p\|^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\ln\ln\frac{d(1+\|p\|^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right) = \widetilde{O}\left(\frac{d(1+\|p\|^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$. Finally, by Lemma 21, we have $\|p\| \leq 3\Delta$ with probability at least 0.9 and thus $n = \widetilde{O}\left(\frac{d(1+\Delta^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$. ■

Lemma 19 *In the event where Lemma 8 holds, if $\Delta \leq 1/24$ and $\langle h, \theta^\star\rangle \geq -\frac{\|\theta^\star\|}{4}$, then the total regret incurred by Algorithm 2 is*

$$\widetilde{O}\left(\frac{d}{\Delta^2\|\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right)$$

and the regret w.r.t. hint vector h is

$$\widetilde{O}\left(\frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right).$$

Proof Since $\|p\| \leq 3\Delta \leq 1/8$, we can apply Lemma 7 and bound this regret as

$$n\left(24\frac{\|P_h^\perp\theta^\star\|^2}{\|\theta^\star\|} + 3\|\theta^\star\|\Delta^2\right) = \widetilde{O}\left(\frac{d}{\Delta^2\|\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right).$$

The regret with respect to the reference action h is bounded as by Lemma 6 as

$$\begin{aligned} n(\langle h-a, \theta^\star\rangle) &\leq n\|\theta^\star\|\|p\|^2 + n\|\theta^\star\||\langle p, a^\star\rangle| \\ &\leq 9n\|\theta^\star\|\Delta^2 + 2.33n\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d'}} \\ &= \widetilde{O}\left(\frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|} + \frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\right) \end{aligned}$$

because $\|\theta^\star\||\langle p, a^\star\rangle| = |\langle p, \theta^\star\rangle| \leq 2.33\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d'}}$ ■

Lemma 20 *With probability at least $1-\delta$, Algorithm 4 (*ESTIMATENORMHP(h, Δ, δ)*) returns r, a constant-factor approximation to $\|P_h^\perp\theta^\star\|$, i.e., $0.06\|P_h^\perp\theta^\star\| \leq r \leq 5\|P_h^\perp\theta^\star\|$, after at most $n \leq \widetilde{O}\left(\frac{d(1+\Delta^2+\|h\|^2)}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$ calls to its* PLAYANDUPDATE *procedure. Further, if $\Delta \leq 1/224$, then the regret incurred with respect to the hint h is:*

$$O\left(n\|\theta^\star\|\Delta^2 + n\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d}}\right) \leq \widetilde{O}\left(\frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta} + \frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|}\ln\frac{1}{\delta}\right)$$

Algorithm 4: ESTIMATENORMHP(h, Δ, δ): High Probability Low Regret ℓ_2-norm Estimation

Input: reference action $h \in \mathbb{R}^d$, perturbation magnitude $\Delta \in \mathbb{R}^+$, failure probability δ

Set $k = 560 \ln(1/\delta)$, Initialize active set $\mathcal{S} = [k]$ and return set $\mathcal{R} = \varnothing$
Initialize k instances of Algorithm 2 as $C_i = \text{ESTIMATENORM}(h, \Delta)$ for $i \in [k]$
Procedure PLAYANDUPDATE()
  if $|\mathcal{R}| \geq 0.67k$ **then**
    play hint h, observe reward y_n.
  else
    for $i \in \mathcal{S}$ **do**
      Call $r_i = C_i$.PLAYANDUPDATE()
      if r_i *is not none* **then**
        $\mathcal{S} \leftarrow \mathcal{S} \setminus \{C_i\}$ and $\mathcal{R} \leftarrow \mathcal{R} \cup \{r_i\}$
  if $|\mathcal{R}| \geq 0.67k$ **then**
    return median$(\mathcal{R})$

If in addition, $\langle h, \theta^\star \rangle \geq -\frac{\|\theta^\star\|}{4}$, then the total regret incurred is

$$O\left(n\frac{\|P_h^\perp \theta^\star\|^2}{\|\theta^\star\|} + n\|\theta^\star\|\Delta^2\right) \leq \widetilde{O}\left(\frac{d(1+\|h\|)}{\Delta^2\|\theta^\star\|}\ln\frac{1}{\delta} + \frac{d(1+\|h\|)\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta}\right).$$

Proof Denote by F_i the event where the ith instance of Algorithm 2 fails, i.e., where the statement in Lemma 8 does not hold. All failure events are independent from each other and have probability at most $\mathbb{P}(F_i) \leq 0.3$ by Lemma 8. We here consider the event E where at least $0.67k$ instances succeed. The probability of this event is at least

$$1 - \mathbb{P}\left(\sum_{i=1}^{k} \mathbf{1}\left(F_i^c\right) < 0.67k\right) \geq 1 - \exp\left(-2k\left(0.7 - \frac{0.67k}{k}\right)^2\right) \geq 1 - \delta$$

by Hoeffding's inequality. We know that in E, Algorithm 4 returns a value r after at most $\widetilde{O}\left(\frac{d}{\Delta^2\|P_h^\perp\theta^\star\|^2}\right)$ calls to its PLAYANDUPDATE procedure. Further, since at most $0.33k$ instances fail in E, the majority of entries in the return set $\mathcal{R}$ was generated by a succeeding instance. Hence, the constant approximation guarantee of Algorithm 2 also holds for Algorithm 4.

Next, we can apply Lemma 6 to bound the regret with respect to h as:

$$\begin{aligned}
n(\langle h - a, \theta^\star\rangle) &\leq n\|\theta^\star\|\|p\|^2 + n\|\theta^\star\||\langle p, P_h^\perp a^\star\rangle| \\
&\leq 9n\|\theta^\star\|\Delta^2 + n|\langle p, \theta^\star\rangle| \\
&\leq 9n\|\theta^\star\|\Delta^2 + 2.33n\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d'}} \\
&= \widetilde{O}\left(\frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta} + \frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|}\ln\frac{1}{\delta}\right)
\end{aligned}$$

because $\|\theta^\star\||\langle p, a^\star\rangle| = |\langle p, \theta^\star\rangle| \leq 2.33\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d'}}$ by Equation 13.

Finally, the total number of calls to PLAYANDUPDATE of all C_i instances (failing and succeeding) is $n = \widetilde{O}\left(\frac{d}{\Delta^2\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta}\right)$, and since $\|p\| \leq 3\Delta \leq 1/8$, each of the $2n$ samples collected satisfies the conditions of Lemma 7. We can therefore bound the total regret as

$$2n\left(24\frac{\|P_p^\perp\theta^\star\|^2}{\|\theta^\star\|} + 3\|\theta^\star\|\Delta^2\right) = \widetilde{O}\left(\frac{d(1+\|h\|)}{\Delta^2\|\theta^\star\|}\ln\frac{1}{\delta} + \frac{d(1+\|h\|)\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\frac{1}{\delta}\right).$$

■

Proof [Proof of Theorem 11] First, observe that if $\|\theta^\star\| \leq \max(d, 3374)/\sqrt{T}$, then any sequence of actions would obtain regret $O(d\sqrt{T})$. Thus, the interesting regime is $\|\theta^\star\| \geq \max(d, 3374)/\sqrt{T}$, which we consider for the remainder of the proof.

We split our regret analysis into 3 phases, as labeled in the psuedocode: in the first phase we estimate $\|\theta^\star\|$, in the second phase we estimate $\|P_h^\perp\theta^\star\|$, and finally in the last phase we call SWITCH. Note that our algorithm may not execute all three segments before we reach our iteration budget T.

Notice that if $\langle\theta^\star, h\rangle \leq -\|\theta^\star\|/4$, then to show that $R_h^T \leq \widetilde{O}(\sqrt{T})$, it suffices to instead show that $R_{-h}^T \leq \widetilde{O}(\sqrt{T})$. Thus by possibly swapping $-h$ and h, we may assume $\langle\theta^\star, h\rangle \geq -\|\theta^\star\|/4$.

Let E_0, E_+ and E_- be the events such that the conclusion of Lemma 10 holds for C_0, C_+ and C_- respectively. Notice that by Lemma 10, each of these events has probability at least $1-\delta/4$. Further, let $E_{Y,+}$ and $E_{Y,-}$ be the respective events that $\langle\theta^\star, h\rangle \in Y_+$ and $\langle\theta^\star, -h\rangle \in Y_-$ for all n. By Lemma 22, we have that $E_{Y,+}$ and $E_{Y,-}$ each occur with probability at least $1-\delta/10$. Let E_{LB} be the event that the linear bandit algorithm used by SWITCH has regret at most $W_3 d\sqrt{t}$ for all $t \leq T$. There exists an absolute constant W_3 such that E_{LB} occurs with probability at least $1-\delta/20$. Let E be the union of all these events. Clearly E has probability at least $1-\delta$. We condition the rest of our argument on this event.

Phase 1: For the first phase, since $\|\theta^\star\| \geq d/\sqrt{T}$, by Lemma 10, we obtain r satisfying $0.06\|\theta^\star\| \leq r \leq 5\|\theta^\star\|$ after $\widetilde{O}(d\ln(1/\delta)/\|\theta^\star\|^2)$ rounds, incurring $\widetilde{O}\left(\frac{d}{\|\theta^\star\|}\ln\frac{1}{\delta}\right) = \widetilde{O}(\sqrt{T}\ln(1/\delta))$ regret. Since the total regret is bounded by $\widetilde{O}(\sqrt{T}\ln(1/\delta))$, the hint-based regret is also similarly bounded. Further, since $\|\theta^\star\| \geq 3374/\sqrt{T}$, this implies $\Delta \leq \frac{1}{224}$.

Phase 2: For the second phase, we call Algorithm 4 on h and $-h$ as instances C_+ and C_-. We need to verify three facts: first, the value $r_\perp$ produced by this phase needs to be a constant-factor approximation to $\|P_h^\perp\theta^\star\|$. Second, the hint-based regret incurred during this phase must be at most $\widetilde{O}(\sqrt{T})$. Finally, the worst-case regret incurred during this phase must be at most $\widetilde{O}(d\sqrt{T})$.

We will consider two broad cases: either we exit phase 2 with at most one of C_+, C_- having returned, or not. Let us first consider the case that at most one instance returns. Notice that since $\langle\theta^\star, h\rangle \geq \langle\theta^\star - h\rangle$, under our assumed events $E_{Y,\pm}$, we will never eliminate C_+ in line 12 of the algorithm. Thus, the only way this case can occur is if the first algorithm to return immediately triggers one of the conditions on line 13 and exits Phase 2 (or no instances return).

Notice that $P_h^\perp = P_{-h}^\perp$. Thus by Lemma 10, the return values of both C_+ and C_- will satisfy $0.06\|P_h^\perp\theta^\star\| \leq r_\perp \leq 5\|P_h^\perp\theta^\star\|$, so regardless of which subroutine provides $r_\perp$, it will be a constant-factor approximation to $\|P_h^\perp\theta^\star\|$.

Now, it remains to bound the regret. Let $N+1$ be the number of times C_-.PLAYANDUPDATE() is called, so that $Y_+ \cap Y_i \neq \varnothing$ after N calls. Notice that $\langle \theta^\star, h\rangle \in Y_+$ and $\langle \theta^\star, -h\rangle \in Y_-$, and after N calls, $|Y_+| = |Y_-| = 2\sqrt{\frac{3\ln\frac{40\ln 2N}{\delta}}{N}}$. Thus, $|\langle \theta^\star, 2h\rangle| \leq 4\sqrt{\frac{3\ln\frac{40\ln 2N}{\delta}}{N}}$. If the total number of samples taken during this phase is $n = O(N)$, this means that $n|r(h,-h)| \leq \widetilde{O}\left(\sqrt{T}\ln\frac{1}{\delta}\right)$. Now, if we could show that the regret of C_+ with respect to h and C_- with respect to $-h$ were both also $\widetilde{O}\left(\sqrt{T}\ln\frac{1}{\delta}\right)$, this would establish our regret bound with respect to the hint during this phase (because $r(h,a) = r(-h,a) + r(h,-h)$ for all a). To this end, consider two cases: $\|P_h^\perp\theta^\star\|^2 \geq d\|\theta^\star\|/\sqrt{T}$ or not. If $\|P_h^\perp\theta^\star\|^2 \geq d\|\theta^\star\|/\sqrt{T}$, then by Lemma 10, we have that C_+ and C_- have regret with respect to h and $-h$ of:

$$\widetilde{O}\left(\frac{d\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta} + \frac{\sqrt{d}}{\Delta\|P_h^\perp\theta^\star\|}\ln\frac{1}{\delta}\right) = \widetilde{O}\left(\sqrt{T}\ln\frac{1}{\delta}\right)$$

Alternatively, if $\|P_h^\perp\theta^\star\|^2 \leq d\|\theta^\star\|/\sqrt{T}$, then again by Lemma 10, the same regret values are bounded by:

$$O\left(T\|\theta^\star\|\Delta^2 + T\frac{\Delta\|P_h^\perp\theta^\star\|}{\sqrt{d}}\right) \leq O(\sqrt{T})$$

This establishes the desired regret bounds with respect to the hint.

To establish the bounds with respect to the optimal action, notice that Lemma 10, implies that C_+ achieves regret:

$$\widetilde{O}\left[\min\left(T\frac{\|P_h^\perp\theta^\star\|^2}{\|\theta^\star\|} + T\|\theta^\star\|\Delta^2,\ \frac{d(1+\|h\|)}{\Delta^2\|\theta^\star\|}\ln\frac{1}{\delta} + \frac{d(1+\|h\|)\|\theta^\star\|}{\|P_h^\perp\theta^\star\|^2}\ln\frac{1}{\delta}\right)\right]$$

Now by once considering two cases depending on whether $\|P_h^\perp\theta^\star\| \geq d\|\theta^\star\|/\sqrt{T}$, we see that this result implies a total regret (for C_+) of $\widetilde{O}(d\sqrt{T}\ln\frac{1}{\delta})$.

Now, let us tackle the regret of C_-. We again consider two cases, either $\langle\theta^\star, h\rangle \leq \|\theta^\star\|/4$ or not. If $\langle\theta^\star, h\rangle \leq \|\theta^\star\|/4$, then $\langle\theta^\star, -h\rangle \geq -\|\theta^\star\|/4$ and so the last conclusion of Lemma 10 applies to C_- as well so that the same argument as in the previous paragraph shows that the regret of C_- with respect to the optimal action is $\widetilde{O}\left(d\sqrt{T}\ln\frac{1}{\delta}\right)$. Alternatively, if $\langle\theta^\star, h\rangle \geq \|\theta^\star\|/4$, notice that since $|\langle\theta^\star, 2h\rangle| \leq 4\sqrt{\frac{3\ln\frac{40\ln 2N}{\delta}}{N}}$, we must have

$$N \leq \widetilde{O}\left(\frac{\ln\frac{1}{\delta}}{|\langle\theta^\star, h\rangle|}\right) \leq \widetilde{O}\left(\frac{\ln\frac{1}{\delta}}{\|\theta^\star\|}\right) \leq \widetilde{O}\left(\frac{\ln\frac{1}{\delta}\sqrt{T}}{d}\right)$$

Thus the regret obtained by C_- cannot be more than $\widetilde{O}\left(\frac{\sqrt{T}}{d}\ln\frac{1}{\delta}\right)$.

This completes the analysis of Phase 2 when at most one instance returns. Let's now consider the case: both instances return, but the first one to return provides a value of $r_\perp$ that such that $\frac{0.06r_\perp^2}{2\cdot 5^2}\cdot\frac{r_\perp^2}{r} \leq Wd\log(T)/\sqrt{T}$. For this case, notice that since $\langle\theta^\star, h\rangle \geq \langle\theta^\star - h\rangle$, under our assumed events $E_{Y,\pm}$, we will never eliminate C_+ in line 12 of the algorithm. Thus since both instances return, it must be that the first-returning instance was C_-. Now, after returning, C_- clearly

incurs zero additional regret with respect to $-h$. Thus, by the same argument as in the previous case, the $Y_+ \cup Y_- = \varnothing$ test will trigger before the regret with respect to the hint exceeds $\widetilde{O}\left(\sqrt{T}\ln\frac{1}{\delta}\right)$. For the worst-case regret, notice that $r_h \leq \widetilde{O}(\ln(1/\delta)/\sqrt{T})$ (because the return value for C_- did not trigger the test in line 11), we have that the worst-case regret is the regret with respect to h plus $Tr_h \leq \widetilde{O}\left(\sqrt{T}\ln\frac{1}{\delta}\right)$.

This completes the analysis of Phase 2.

Phase 3: Now, we focus on the last phase. At this point, we have established that $r_\perp$ is a constant-factor approximation of $\|P_h^\perp\theta^\star\|$ and r is a constant-factor approximation of $\|\theta^\star\|$. Therefore $r_\perp^2/r = \Theta(\|P_h^\perp\theta^\star\|^2/\|\theta^\star\|) = \Theta(r_h)$, where the final equality follows from Lemma 4. Moreover, since $r \geq 0.06\|\theta^\star\|$ and $r_\perp < 5\|P_h^\perp\theta^\star\|$, $\frac{0.06r_\perp^2}{2\cdot 5^2}\cdot\frac{r_\perp^2}{r} \leq \frac{1}{2}\cdot\|P_h^\perp\theta^\star\|^2/\|\theta^\star\| \leq r_h$. Further, since we cannot eliminate h, we must have that either $|S| = 1$ and h is the remaining hint, or $|S| = 2$, but $r_\perp$ is such that SWITCH would not choose to play the hint in any event. Thus SWITCH incurs no regret with respect to h, while always maintaining a regret of $O(d\sqrt{T})$ with respect to the optimal action. ■

B.1. Algorithms and proofs for Pareto and Multi-Hint settings

Algorithm 5: Pareto Frontier: Bandit Algorithm on Unit Ball

Input: hint $h \in \mathbb{R}^d$, number of rounds T, failure probability δ, target total hint regret: G

// Phase 1: Estimate norm $\|\theta^\star\|$

1 Initialize $C_0 \leftarrow$ ESTIMATENORMHP$(0, 1, \frac{\delta}{4})$

2 Call C_0.PLAYANDUPDATE() until it returns a value r

// Phase 2: Estimate norm of orthogonal complement $\|P_h^\perp\theta^\star\|$

3 Set exploration radius $\Delta = \frac{\sqrt{G}}{\sqrt{rT}}$

4 Initialize $C_+ \leftarrow$ ESTIMATENORMHP$(+h, \Delta, \frac{\delta}{4})$ and $C_- \leftarrow$ ESTIMATENORMHP$(-h, \Delta, \frac{\delta}{4})$

5 Initialize active set $\mathcal{S} = \{C_+, C_-\}$

// Elements of S are arms, a pull corresponds calling `PlayAndUpdate`, reward is the unperturbed hint loss

6 Run `MultiArmBandit(S, G)` until any instance in S returns a value $r_\perp$ satisfying $r_\perp^2/r \geq c_0 * d\log(T)/G$ or $|S| = 1$ and the lone instance returns $r_\perp$

// Phase 3: Commit to hint or ignore it. Note $c_0 > c_1$.

7 For remaining rounds, call SWITCH$(h, r_\perp^2/r, T, c_1 * d\log(T)T/G)$ for some $h \in S$ randomly chosen

Proof [Proof of Lemma 12] This lemma follows similarly to Lemma 9.

By Lemma 6, we see that our hint-based regret is bounded by high constant probability by:

$$T_e * \|\theta^\star\|\left(\Delta^2 + \Delta\frac{\|P_h^\perp a^\star\|}{\sqrt{d}}\right)$$

where T_e is the number of iterations of `EstimateNorm`. Let us write $\theta^\star = \alpha h + v$, then note $\|P_h^\perp a^\star\| = \|v\|/\|\theta^\star\|$, so if $\|v\|^2 \leq \|\theta^\star\|^2 d\log(T)\Delta^2$, then we have that

$$R_h^T \leq O(\|\theta^\star\|\Delta^2 T\sqrt{\log(T)})$$

Furthermore, in this case, note that the instantaneous regret of playing h is, by Lemma 4, given by $\Theta(\|v\|^2/\|\theta^\star\|) = \|\theta^\star\| d\log(T)\Delta^2$. Therefore, the full regret is bounded by $O(d\|\theta^\star\|\Delta^2 T\log(T))$.

Otherwise, we have $\|v\|^2 \geq \|\theta^\star\|^2 d\log(T)\Delta^2$. Then by Lemma 10, the maximum rounds of iterations is $T_e = O(\frac{d\log(T)}{\|v\|^2\Delta^2})$, with probability at least $2/3$, so we can bound our hint-based regret by

$$R_h^T = O\left(\frac{d\log(T)}{\|v\|^2\Delta^2}\right) * \|\theta^\star\| \left(\Delta^2 + \Delta\frac{\|v\|}{\|\theta^\star\|\sqrt{d}}\right) = O\left(\frac{\log(T)}{\|\theta^\star\|\Delta^2}\right)$$

For the worst case regret, if $\langle h, \theta^\star\rangle \geq -\|\theta^\star\|/4$, then we incur regret at most

$$O\left(\frac{d\|\theta^\star\|\log(T)}{\|v\|^2} + \frac{d\log(T)}{\Delta^2\|\theta^\star\|}\right) = O\left(\frac{d\log(T)}{\Delta^2\|\theta^\star\|}\right)$$

by our Lemma 10.

■

Proof [Proof of Theorem 13] This proof follows directly from combining Lemma 12 with the same reasoning as Theorem 11. We borrow the same notation and split our regret analysis into 3 phases, as labeled in the psuedocode: in the first phase we estimate $\|\theta^\star\|$, in the second phase we estimate $\|P_h^\perp\theta^\star\|$, and finally in the last phase we call SWITCH. Note that our algorithm may not execute all three segments before we reach our iteration budget T.

Phase 1: For the first phase, since $\|\theta^\star\| \geq d/G$, by Lemma 10, we obtain r satisfying $0.06\|\theta^\star\| \leq r \leq 5\|\theta^\star\|$ after $\widetilde{O}(d\ln(1/\delta)/\|\theta^\star\|^2)$ rounds, incurring $\widetilde{O}\left(\frac{d}{\|\theta^\star\|}\ln\frac{1}{\delta}\right) = \widetilde{O}(G\ln(1/\delta))$ regret. Since the total regret is bounded by $\widetilde{O}(G\ln(1/\delta))$, the hint-based regret is also similarly bounded.

Phase 2: For the second phase, we call Algorithm 4 on h and $-h$ as instances C_+ and C_-. We need to verify the hint-based and worst case regrets.

For the hint-based regret, we first consider the non-perturbed actions. In this case, we need to bound $R(\mathbf{h}, -\mathbf{h})$ in the 2-arm bandit game, we play the hint-based MAB algorithm with total hint regret G against the arm corresponding to h (i.e. C_+) and worst case regret $O(dT/G)$ (as described in the upper bounds in Lattimore (2015)). If the MAB terminates early, we bound the hint-based regret by using the worst case regret. Note that if some C returns $r_\perp$ with $r_\perp^2/r \geq c_0 * d\log(T)/(G\sqrt{T})$, then since the worst case bound is $R = O(dT\log(T)/G)$ for some constant and $r_\perp^2/r$ is a constant approximation to r_h, we can find c_0 such that $r_h T \geq R$, which shows that our hint-based regret must be negative.

Now, it suffices to add the perturbations and note that by the additivity property of the regret, we can simply bound the hint-based regret of C_-, C_+ to each hint respectively. To do this, we use Lemma 12 with $\Delta^2 = G/rT$ to bound the hint-based regret by $O(\|\theta^\star\|\Delta^2 T\log(T)) = O(G\log(T))$ for $G \leq \sqrt{T}$. Similarly, we bound the worst case regret by the same lemma by $O(dT\log(T)/G)$.

Phase 3: Now, we focus on the last phase. At this point, we have established that $r_\perp$ is a constant-factor approximation of $\|P_h^\perp\theta^\star\|$ and r is a constant-factor approximation of $\|\theta^\star\|$. Therefore $r_\perp^2/r = \Theta(\|P_h^\perp\theta^\star\|^2/\|\theta^\star\|) = \Theta(r_h)$, where the final equality follows from Lemma 4. Moreover, since $r \geq 0.06\|\theta^\star\|$ and $r_\perp < 5\|P_h^\perp\theta^\star\|$, $\frac{0.06r_\perp^2}{2\cdot 5^2}\cdot\frac{r_\perp^2}{r} \leq \frac{1}{2}\cdot\|P_h^\perp\theta^\star\|^2/\|\theta^\star\| \leq r_h$. Further, since we cannot eliminate h, we must have that either $|S| = 1$ and h is the remaining hint, or $|S| = 2$, but $r_\perp$ is such that SWITCH would not choose to play the hint in any event. Thus SWITCH incurs no

regret with respect to h, while always maintaining a regret of $O(dT\log(T)/G)$ with respect to the optimal action.

■

Proof [Proof of Lemma 14] Let $\mathcal{E}$ be the event where the statement of Lemma 10 holds all instances C_i, where $\langle h_i, \theta^\star\rangle \in Y_i$ at all times for all $h_i \in \mathcal{H}$. By Lemma 22, Lemma 10 and a union bound, the probability of $\mathcal{E}$ is at least $1-\delta/2$. In this event, the number of rounds of all instances C_i of ESTIMATENORMHP is bounded as

$$n_i \le \frac{T}{mB} \wedge \widetilde{O}\left(\frac{d(1+\Delta^2+\|h_i\|^2)}{\Delta^2\|P_{h_i}^{\perp}\theta^\star\|^2}\right)$$

and when $\langle h_i, \theta^\star\rangle \ge -1/4\|\theta^\star\|$, the total regret of C_i is bounded as

$$O\left(n_i\frac{\|P_h^{\perp}\theta^\star\|^2}{\|\theta^\star\|} + n_i\|\theta^\star\|\Delta^2\right) \le \widetilde{O}\left(\frac{d}{\Delta^2} + \frac{T\Delta^2}{mB}\right) .$$

Note that for any h_i with $\langle h_i, \theta^\star\rangle < -1/4\|\theta^\star\|$, there is $-h_i \in \mathcal{H}$ with $r(-h_i, h_i) \ge 1/2\|\theta^\star\| \ge c_1/2$. Also, we know that $\langle h_i, \theta^\star\rangle \in Y_i$ for both $h_i, -h_i$ and the interval length of Y_i is $O(\log(N/\delta)/\sqrt{N})$ after N calls to C_i^+, C_i^-. Therefore, by the interval width of Y_i, we conclude that $r(-h_i, h_i) \le c\frac{\log(N/\delta)}{\sqrt{N}}$. Therefore, we pull C_i^+ at most $N = O(\log(T))$ times before it is eliminated (since $-h_i$ is assumed in this case to have a higher reward).

Summing over all m (or $2m$) priors, we get a final regret bound bound of

$$\widetilde{O}\left(\frac{dm}{\Delta^2} + \Delta^2\frac{T}{B}\right)$$

as long as $d/\Delta^2 \ge \log(T)$.

Since $\Delta^2 = m/\sqrt{T}$, the first term in the expression to be bounded by $d\sqrt{T}$. For $m \le d$, this setting of Δ^2 also implies that $\Delta^2 T/B \le dm/\Delta^2$ for any positive integral value of B.

For the last $T - T/B$ rounds, since we reduced to the 1-prior case, our worst case regret is $\widetilde{O}(d\sqrt{T})$ by Theorem 11. ■

Proof [Proof of Lemma 15] The proof follows extremely closely to Theorem 11 and Lemma 14. We again condition on $\mathcal{E}$ be the event where the statement of Lemma 10 holds all instances C_i, where $\langle h_i, \theta^\star\rangle \in Y_i$ at all times for all $h_i \in \mathcal{H}$, which holds with probability at least $1-\delta$. For the first part of the algorithm, note that we are essentially want to bound the hint-based regret of playing multi-arm bandit with the m hints for T/B rounds. First consider the case when the best hint $h^\star$ was not eliminated and when we consider the regret when playing a non-perturbed action.

Then, we claim that for the first T/B rounds,

$$\sum_j r(h_{a_j}, h^\star) \le \sqrt{mT/B}\log(T)$$

Let $s_i = r(h_i, h^\star)$, then since the intervals of Y_i are shrinking like $O(1/\sqrt{N})$, we conclude that each suboptimal arm is pulled at most $O(\log(T)/s_i^2)$ times as long as $s_i \ge \sqrt{mB\log(T)/T}$ and would have been eliminated with high probability by the reward confidence interval comparison.

Algorithm 6: Multi-Hint Bandit Algorithm on Unit Ball

Input: hints $\{h_i \in \mathbb{R}^d\}_{i=1}^m$, number of rounds T, failure probability δ, exploration ratio B, worst case regret scaling W

Set exploration radius $\Delta = \frac{\sqrt{m}}{T^{1/4}}$ and active set $\mathcal{S} = \varnothing$

foreach *hint* h_i **do**

  Initialize $C_i^+ \leftarrow$ ESTIMATENORMHP$(h_i, \Delta, \frac{\delta}{4m})$

  Initialize $C_i^- \leftarrow$ ESTIMATENORMHP$(-h_i, \Delta, \frac{\delta}{4m})$

  Add C_i^+, C_i^- to active set $\mathcal{S}$

repeat

  foreach *active instance* $C_i \in \mathcal{S}$ **do**

    Call C_i.PLAYANDUPDATE()

    // Maintain CI of hint's expected reward

    $\mathcal{R}_i \leftarrow$ all reward samples obtained by C_i so far playing unperturbed hint

    Compute confidence interval $Y_i = (\bar{y}_i - b_i, \bar{y}_i + b_i)$ with $\bar{y}_i = \frac{1}{|\mathcal{R}_i|}\sum_{y \in \mathcal{R}_i} y$ and $b_i = \sqrt{\frac{3\ln(40m\ln(2|\mathcal{R}_i|)/\delta)}{|\mathcal{R}_i|}}$

  // Eliminate worse hint if possible

  if $Y_i \cap Y_j = \varnothing$ *for any* i, j **then**

    Remove C_i with smaller $\bar{y}_i$ from active set $\mathcal{S}$

  if $|\mathcal{S}| > 1$ *and* C_i *has returned a value* $r_\perp$ *satisfying* $\frac{r_\perp^2}{r} \geq \frac{c_0 W d \log(T)}{\sqrt{T}}$ **then**

    Remove C_i from $\mathcal{S}$

until $\frac{T}{mB}$ *iterations or* $|\mathcal{S}| = 1$

// Commit to single hint after $\frac{T}{mB}$ iterations

Choose h randomly from $\mathcal{S}$ and call PARETOBANDIT(h, T, δ, W)

Let R be a regret threshold, then the total regret with respect to the best hint is $R(T/B) + \frac{m\log(T)}{R}$, where the first term of the regret captures the regret for all arms with $s_i \leq R$ and the second term captures that for $s_i \geq R$. By setting $R = \sqrt{mB\log(T)/T}$, we get our final regret bound of $\sqrt{mT/B}\log(T)$.

Therefore, we see that hint-based regret is bounded by $\widetilde{O}(\sqrt{mT/B})$, if we do not play a perturbed action. However, since we are playing a perturbed action at each step with an orthogonal perturbation of h_i, the hint-based regret due to the perturbation is $\widetilde{O}(\Delta^2 * (T/B) + \sum_i \Delta\|P^{\perp}_{h_{a_i}}\theta^\star\|/\sqrt{d})$ by direct application of Lemma 6 and using the fact that $r(a_i, h^\star) \leq r(h_i, h^\star) + r(a_i, h_i)$, where a_i is the perturbed action for hint h_i.

Now as in Theorem 11, consider two cases: $\|P^{\perp}_h\theta^\star\|^2 \geq d\|\theta^\star\|/\sqrt{T}$ or not. If $\|P^{\perp}_h\theta^\star\|^2 \leq md\|\theta^\star\|/\sqrt{T}$, then since $\Delta^2 = m/\sqrt{T}$, we directly deduce that the hint-based regret is $\widetilde{O}(\Delta^2(T/B) + (m/\sqrt{T})(T/B)) = \widetilde{O}((m/B)\sqrt{T})$. Otherwise, if $\|P^{\perp}_h\theta^\star\|^2 \geq md\|\theta^\star\|/\sqrt{T}$, then by Lemma 10, we have that C_+ and C_- have regret with respect to h and $-h$ of:

$$\widetilde{O}\left(\frac{d\|\theta^\star\|}{\|P^{\perp}_h\theta^\star\|^2}\ln\frac{1}{\delta} + \frac{\sqrt{d}}{\Delta\|P^{\perp}_h\theta^\star\|}\ln\frac{1}{\delta}\right) = \widetilde{O}\left(\frac{1}{m}\sqrt{T}\ln\frac{1}{\delta}\right)$$

Since there are at most m hints, the total hint-based regret is bounded by $\widetilde{O}(\sqrt{T}\log(T))$. Therefore, we conclude that the hint-based regret is dominated by $\widetilde{O}((m/B)\sqrt{T})$.

For the second part of the algorithm when h is chosen randomly from $\mathcal{S}$, note that since all hints with $s_i \geq \Omega(\sqrt{mB\log(T)/T})$ would have been eliminated, deferring to Theorem 11, the hint-based regret for the remaining rounds is at most $\widetilde{O}(s_iT + \sqrt{T}) = \widetilde{O}(\sqrt{mBT})$, with respect to h^*.

Therefore, our total hint-based regret is

$$\widetilde{O}(\sqrt{mBT} + (m/B)\sqrt{T})$$

Setting $B = m^{1/3}$ gives our result.

Finally, if the best hint has been eliminated, then it must be the case that we can find c_0 such that $r_\perp^2/r \geq c_0\frac{d\log(m)\log(T)}{\sqrt{T}}$ implies that $r_h \geq R/T$ since $r_\perp^2/r = \Theta(r_h)$, where R is our worst case bound from above, which shows that our hint-based regret must be negative. ■

Appendix C. Auxiliary Technical Lemmas

Lemma 21 *Let $G \sim \mathcal{N}(0, \mathbf{I}_d/d)$ be an isotropic Gaussian random variable and $v \in \mathbb{R}^d$ be arbitrary. Then the events*

$$\left\{\|G\|^2 \leq 1 + 2\sqrt{\frac{\ln(1/\delta)}{d}} + 2\frac{\ln(1/\delta)}{d}\right\} \qquad \left\{\|G\|^2 \geq 1 - 2\sqrt{\frac{\ln(1/\delta)}{d}}\right\}$$

$$\left\{\langle v, G\rangle^2 \leq \frac{2\|v\|^2}{d}\min\left\{\Phi^{-1}(1-\delta/2)^2, \frac{1}{2} + \ln\frac{1}{\delta} + \sqrt{\ln\frac{1}{\delta}}\right\}\right\}$$

$$\left\{\langle v, G\rangle^2 \geq \frac{\|v\|^2}{d}\max\left\{\Phi^{-1}(1/2+\delta/2)^2, 1 - 2\sqrt{\ln\frac{1}{\delta}}\right\}\right\}$$

each have probability at least $1-\delta$.

Proof We know that $X = d\|G\|^2$ is χ^2_d-distributed and thus, by the tail bounds of Laurent and Massart (2000), we have

$$\mathbb{P}\left(X - d \geq 2\sqrt{d \ln \frac{1}{\delta}} + 2\ln\frac{1}{\delta}\right) \leq \delta \qquad \text{and} \qquad \mathbb{P}\left(X - d \leq -2\sqrt{d \ln \frac{1}{\delta}}\right) \leq \delta\,.$$

Rearranging the first event gives that for $\delta \leq \exp(-1/2)$

$$\|G\|^2 \geq \frac{1}{d} + 4\ln(1/\delta) \geq 2\sqrt{\ln(1/\delta)} + 2\ln(1/\delta) \geq 2\sqrt{\frac{\ln(1/\delta)}{d}} + 2\frac{\ln(1/\delta)}{d}$$

happens with probability at most δ. Rearranging the second event gives

$$\|G\|^2 \leq 1 - 2\sqrt{\ln(1/\delta)} \leq 1 - 2\sqrt{\frac{\ln(1/\delta)}{d}}$$

happens with probability at most δ. Further, for any $v \in \mathbb{R}^d$, the distribution of $Y = d\langle v, G\rangle^2$ is χ^2_1 since, w.l.o.g. $v = e_1$ and $\langle v, G\rangle = G_1 \sim \mathcal{N}(0, 1/d)$ and thus we can apply the tail bound above again to obtain the desired statement for the remaining events. ■

Lemma 22 *Let* $(X_i)_{i\in\mathbb{N}}$ *be a sequence of independent* σ*-sub-Gaussian random variables. Then with probability at least* $1-\delta$ *for all* $n \in \mathbb{N}$ *jointly*

$$\left|\sum_{i=1}^{n} X_i\right| \leq \sigma\sqrt{\frac{3\ln\frac{4\ln(2n)}{\delta}}{n}}\,.$$

Proof Follows directly from Theorem 1 by Howard et al. (2021) (see their Equation (2)). ■