Online Learning and Bandits with Queried Hints

Aditya Bhaskara ✉
School of Computing, University of Utah, Salt Lake City, UT, USA

Sreenivas Gollapudi ✉
Google Research, Mountain View, CA, USA

Sungjin Im ✉
University of California, Merced, CA, USA

Kostas Kollias ✉
Google Research, Mountain View, CA, USA

Kamesh Munagala ✉
Computer Science Department, Duke University, Durham, NC, USA

─── **Abstract** ───

We consider the classic online learning and stochastic multi-armed bandit (MAB) problems, when at each step, the online policy can *probe* and find out which of a small number (k) of choices has better reward (or loss) before making its choice. In this model, we derive algorithms whose regret bounds have exponentially better dependence on the time horizon compared to the classic regret bounds. In particular, we show that probing with $k = 2$ suffices to achieve time-independent regret bounds for online linear and convex optimization. The same number of probes improve the regret bound of stochastic MAB with independent arms from $O(\sqrt{nT})$ to $O(n^2 \log T)$, where n is the number of arms and T is the horizon length. For stochastic MAB, we also consider a stronger model where a probe reveals the reward values of the probed arms, and show that in this case, $k = 3$ probes suffice to achieve parameter-independent constant regret, $O(n^2)$. Such regret bounds cannot be achieved even with full feedback *after* the play, showcasing the power of limited "advice" via probing *before* making the play. We also present extensions to the setting where the hints can be imperfect, and to the case of stochastic MAB where the rewards of the arms can be correlated.

1 Introduction

In this paper, we consider two problems that form the cornerstone of sequential analysis and decision theory, a field first developed by Wald [53] in the 1940's. The first is the online linear/convex optimization problem that was initially studied in the context of repeated games by Hannan [25] and Blackwell [13] in the 1950's. In this problem, there is a possibly infinite space of potential actions in a high-dimensional space. At each step, a decision maker needs to choose one action, that is sometimes called an "arm". Subsequently, nature presents an adversarially chosen linear (or convex) loss function, and the decision maker incurs the evaluation of this loss function at the chosen action. Subsequently, the decision maker is told the loss function at that time step. The goal is to compete with an omniscient policy that

knows all the loss functions in advance, but is restricted to choosing one fixed action for all time steps. The difference between the loss of the policy and that of the omniscient policy is termed *regret*. Over the decades, several policies have been developed [28, 36, 57] that achieve regret $O(\sqrt{T})$, where T is the horizon length, and the $O(\cdot)$ hides problem-dependent parameters. Such a dependence on T is also optimal [26]. A well-known specialization is the *experts problem* where the actions or arms form a discrete set, and each arm incurs an arbitrary loss at each time step that is unrelated to the losses of the other arms [36, 16].

The second problem is the stochastic multi-armed bandit (MAB) problem that was first formulated by Robbins [45] around 1950, though the widely used Thompson Sampling policy for this problem dates back to the 1930's [52]. In this problem, a decision maker is faced with n independent arms that yield *i.i.d.* rewards from unknown underlying distributions. As before, the goal is to design a policy or allocation rule to sequentially play these arms to maximize reward. In a sense, this can be viewed as a stochastic version of the experts problem; however, the key difference is that the decision maker only learns the reward of the chosen arm at the end of the time step, and not the rewards of all the arms. The regret is measured against an omniscient policy that knew the reward distributions of the arms (but not the reward values), and chooses the single arm with maximum expected reward at all time steps. In seminal work, Lai and Robbins [32] showed an optimal allocation policy along with tight lower bounds on the regret incurred by any policy, assuming a parametric form on the distributions. This result was subsequently generalized by Auer, Cesa-Bianchi, and Fisher [5], who derived similar upper bounds without assuming a parametric form on the distributions.

Both the online convex optimization and the stochastic MAB problem have found numerous applications in areas ranging from clinical trial design to ad-word allocations to recommendation systems, and continue to be extensively studied in the fields of statistics and machine learning. For more on the history and variants of this problem, we refer the reader to several excellent books [15, 48, 34, 26].

1.1   Probe Model and Motivation

In this paper, we study the following twist on these problems. Suppose before playing, we are allowed to query or *probe* an oracle with k options, and the oracle responds with the best of these at that step – either telling the algorithm the identity of this option and nothing more, or telling the algorithm the rewards/losses of all the options. We subsequently play the option suggested by the oracle. How should these probes be chosen, and can we obtain much better regret?

Our motivation for such a probe model comes from the recent literature on designing algorithms that can leverage machine learning (ML) based *predictions*. This paradigm has been used to obtain improved guarantees for many classic online algorithmic problems [44, 38, 33, 43, 22] (see also the survey [39]). In these settings, the online algorithm is assumed to have access to an auxiliary ML model that predicts properties of the arriving inputs. The goal is to derive improved bounds assuming the predictions are correct, whilst doing nearly as well as worst-case algorithms when the predictions are incorrect. Of particular relevance is recent work on "parsimonious hints", where the ML model provides as little information to the online algorithm as possible. For instance, recent work has considered online linear optimization, where the hint is a direction with a strictly positive dot product with the cost vector [17, 10]. In such a model, for optimization over a sphere, the regret bound improves from $\Omega(\sqrt{T})$ to $O(\ln T)$, even with hints at $O(\sqrt{T})$ time steps. Similarly, recent work [27] has considered randomized caching with parsimonious hints about next request time of a few

cached pages.

The probe model can now be viewed as a parsimonious ML hint, where the algorithm queries the predictor with a few options, asking it either for the identity of the best option among these, or for the losses/rewards of all these options. As an example, consider modeling the problem of shortest path routing using stochastic bandits (where each of m paths has its length drawn from a distribution at each step). In this case, a routing engine can query for the length of a few paths, by possibly querying users who have chosen this path, before routing other users.

1.2   Results and Conceptual Contribution

In this paper, we study the following question:

> Can we obtain improved regret bounds for online learning and MAB problems against the classic best arm in hindsight benchmark, if the algorithm is allowed the power of using ML advice via a few probes before making its decision to play?

We answer this question in the affirmative: We show that just $k = 2$ probes (or $k = 3, 4$ for stronger results) suffice to drastically improve the known regret guarantees. Indeed, our main results (in Sections 3 and 5) show a *constant regret bound independent of the time horizon*, assuming the hints are correct!

Before proceeding further, it is instructive to compare our results with [17]. They consider online linear optimization where the domain is the sphere (or more generally, strictly convex) and show that if the hint is a direction with positive "correlation" with the cost vector, the regret improves to $O(\ln T)$. They also show that such an improvement cannot be obtained if the decision space is more general, say the ℓ_∞ ball, where the regret remains $\Omega(\sqrt{T})$. In contrast, we show that if the algorithm is allowed to choose a direction and ask the predictor if the cost along that direction is increasing or not, then the regret improves to constant; further this result holds for linear optimization over *any domain* (not just strictly convex), and convex optimization over *any* convex space.

In the literature on algorithms with ML-based predictions, one key requirement is for the algorithms to not be "thrown off" by incorrect predictions. This aspect is referred to as robustness [39], and the goal is to recover worst-case guarantees (ones possible without any hints) even if the hints are adversarial. In the context of linear optimization, this was studied in [9]. They extend the result of [17] again assuming the optimization is over a sphere; however, their result also has a dependence of $O(\ln T)$ on the horizon length. As our second contribution, with B imperfect hints, our probe model improves this to obtain regret $O(\sqrt{B+1})$ (with no dependence on the horizon T), again holding for any underlying space over which the linear optimization is performed, as well as for convex optimization over any convex space. We finally note that there was no extension known for [17] to the MAB problem, and one of our contributions is to develop stochastic bandit algorithms with improved regret bounds under parsimonious hints.

At a conceptual level, our work shows that querying or probing options via comparisons is far more powerful than more 'passive' models for hints. We believe that such query based hint models may find other applications beyond online learning. At an even higher level, our work is reminiscent of the *power of two choices* in online load balancing [7], where choosing the lesser loaded of two random bins leads to an exponential improvement in the expected maximum load. This paradigm has found numerous applications, such as hashing, congestion control, and distributed memory management. Our paper is in a similar vein – we show that allowing a few queries or hints suffices to give an exponential improvement in regret.

1.3   Overview and Technical Highlight

We present our model and summary of results in Section 2, where we also place the regret bounds we obtain in context. Our main probing model is the BESTPROBE model, where the online policy probes a set of arms or options ahead of the play, and is told the best of these options (without revealing the actual losses/rewards of these options). For the MAB problem, we also consider the ALLPROBE model where the policy also observes the rewards of all arms it probes. We present algorithms and regret bounds for online linear/convex optimization in the BESTPROBE model in Section 3. Our results extend to the case where the hints can be imperfect, and we present this in Section 3.3. We present analogous results for the MAB problem in Section 4. We consider the MAB problem in the ALLPROBE model in Sections 5 and 6.

At a technical level, our results require development of new probabilistic tools to lower bound (resp. upper bound) the maximum (resp. minimum) of independent random variables(See Lemmas 5, 7, 9 and 10.) We term these as "reverse prophet inequalities", since they are in some sense the reverse of well-known prophet inequality results [30, 46] that upper bound the expected maximum of a set of independent random variables by a sum of quantities related to individual distributions. These technical lemmas are crucial to both our algorithms as well as our improved regret analyses. Much like prophet inequalities, these lemmas are of independent interest as stand-alone probability tools.

1.4   Other Related Work

Stochastic Probing. The question of adaptively or non-adaptively probing independent distributions has been widely studied, with applications to database query optimization [21, 42, 18, 37], wireless communication [24], and traffic routing [12]. Much like our model, a probe reveals the true underlying value drawn from the distribution; however, this line of work largely focuses on algorithm design as opposed to learning. It is shown in [21] that the problem of computing the best set of k distributions to probe in order to maximize the expected value of the maximum of the probed set is NP-HARD when k is not a constant. A related problem is the *Pandora's problem* [55, 24, 8], where there is no bound on the number of probes, but we seek to maximize the largest value found minus the total probing cost spent in discovering the value. A general adaptive greedy algorithm for such problems, which probes the next distribution conditioned on the values seen so far, was presented in [23]. Our work is different in that we assume k is a small constant (so that NP-HARDNESS is not an issue); instead, we seek to understand the power of such probes in repeated bandit interactions.

Bandits with Probes. In the *cascading bandits* model [31, 49], a recommendation system (such as search engine or streaming service) needs to choose k items (or arms) to show to a user and obtains feedback on what item the user clicks on. Similarly, in the bandits with pre-observation model [58], the arms are wireless channels of unknown quality, and a user needs to sequentially probe the channels until a good channel is found. This is a bandit version of the Pandora's problem. Finally, motivated by job scheduling applications, the online budgeted submodular coverage problem [51] considers the more general problem where the reward obtained by the player is an unknown submodular function of the probed arms. Such problems have also been recently considered in the experts setting [41] where the policy gets feedback about all arms after the play, but is allowed to make a bounded number of probes before the play. One commonality in all the above works is that they consider *policy*

regret: The benchmark is an omniscient solution that not only knows the rewards/losses of the arms, but is also allowed as many probes per step as the online policy. In contrast, motivated by machine learning hints, we consider probing as providing the online policy *more power* compared to an omniscient benchmark and our goal is to study the resulting improvement in regret bounds.

Dueling Bandits. Our problem is also related to *dueling bandits* [56], where the only information available on playing a pair of arms is the result of a noisy comparison. Assuming the noisy comparison model is Condorcet consistent (e.g., the Bradley-Terry model), the goal is to minimize the error incurred in not playing the best arm. Though our problem is superficially similar in that we allow plays of multiple arms, in our case, we observe the reward of at least one of the played arms. Further, it is easily possible to incur zero (or even negative) regret in our problem by playing two sub-optimal arms. This makes our problem technically very different. Nevertheless, it is an interesting question whether techniques from dueling bandits can be used to improve some of our results.

Predictable Sequences. In the field of online algorithms in general, recent research has focused on incorporating machine learning predictions to obtain more optimistic bounds if the predictions are correct, but preserve the robustness of the classic online model in case the predictions turn out to be inaccurate. Of particular relevance is recent work [44, 54, 50] on bandits and experts with "predictable sequences", where improved regret bounds are shown when the algorithm is given a prediction that is close on average to the true reward at each step. Our work is similar in spirit if we view the probe as a perfect prediction; however, in our case, the probes are both *interactive* and *parsimonious*, meaning that the policy itself has to decide on the predictions to obtain each step, and further, these are few in number.

Bandits with Limited Advice. A related model to predictable sequences is the *experts* model, where in each step, each of m experts makes a prediction about the best arm at that step. In the limited advice setting [47, 29], the policy can choose only $k \ll m$ of the experts in each round and obtain their predicted best arm; subsequently, the policy plays one of the arms and obtains its reward. The goal of the policy is to compete with reward of following the best expert in hindsight. However, the predictions of these experts can be arbitrary, and unrelated to which arm was actually the best arm. Therefore, even if a policy obtains the predictions of all experts, it cannot avoid $\Omega(\sqrt{nT})$ regret, making this problem fundamentally different from our setting.

2 Model and Results

In the online linear optimization problem, we are given a a finite set $\mathcal{W} \subseteq \mathbb{R}^d$ of options (or arms). Assume that for all $w \in \mathcal{W}$, we have $w \in [-1, 1]^d$. At each step t, the algorithm chooses action w^t and is subsequently presented a cost vector $\ell^t \in [-1, 1]^d$, for which it incurs loss $\langle \ell^t, w^t \rangle \in [-d, d]$. Note that the choice of action w^t only depends on the cost vectors ℓ^q for $q < t$. We assume cost vectors ℓ^t are generated by an *oblivious* adversary that does not know the internal randomness used by the algorithm in choosing w^t. We further assume that any linear function can be efficiently minimized over the set $\mathcal{W}$. We describe the related experts and online convex optimization settings in the full paper [11].

There is a horizon of T steps; we assume T is known, but this assumption can be removed using standard techniques. We measure regret with respect to the hindsight optimum as:

$$\text{Regret} = \sum_{t=1}^{T} \langle \ell^t, w^t \rangle - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} \langle \ell^t, w \rangle.$$

We denote $\mathtt{OPT} = \min_{w \in \mathcal{W}} \sum_{t=1}^{T} \langle \ell^t, w \rangle$.

In the stochastic MAB problem, there are n bandit arms. Arm i yields *i.i.d.* rewards X_i drawn from an independent distribution $D_i \in [0,1]$. Let $\mu_i = \mathbf{E}[X_i]$. The policy can play one arm i_t at each step t, this choice can depend on the observed rewards till time $t-1$. Unlike online linear optimization, in the MAB problem, at the end of time step t, the policy *only* learns the reward $r_{i_t,t} \sim D_{i_t}$ of the arm i_t that it plays at time t. The hindsight optimum plays the arm with highest expected reward each step so that $\mathtt{OPT} = T \max_i \mathbf{E}[X_i]$. We consider pseudo-regret, which is simply $\mathtt{OPT} - \mathbf{E}[\sum_{t=1}^{T} r_{i_t t}]$. In Section 6, we extend this to the case where there is a joint (correlated) reward distribution over arms and the rewards are drawn *i.i.d.* across time from this joint distribution.

Probe Model. There are two models of probing and feedback that we consider. These models are parameterized by a number $k \leq n$, which captures the number of probes allowed. The first model applies to both online learning and stochastic MAB, while the second applies only to MAB.

Probes with Best Arm Feedback (BESTPROBE). Any policy probes a set S_t of at most k arms or options at any time step t, and learns which arm in S_t will incur the lowest loss or maximum reward at step t (but not the reward/loss values). The policy plays the arm $i \in S_t$ with largest reward (resp. minimum loss). For instance, in the MAB problem, if the reward of arm i at time t at step t is $r_{it} \sim D_i$, the policy's reward is $\hat{R}_t = \max_{i \in S_t} r_{it}$.

Probes with All Feedback (ALLPROBE). For the stochastic MAB problem, we also consider a model that gives the online policy more information. In this model, the policy is actually told the rewards $r_{it} \sim D_i$ for each arm $i \in S_t$, and it subsequently plays the arm with largest reward, incurring reward $\hat{R}_t = \max_{i \in S_t} r_{it}$ at that step.

When $k = 1$, these models reduce to the classic versions of online linear optimization and stochastic MAB. Note that our regret measure is a departure from work on policy regret: In our case, $\mathtt{OPT}$ remains the same whether $k = 1$ or $k > 1$. In that sense, we seek to understand the power of increasing k on the regret of the online policy. As mentioned before, the motivation comes from viewing the $k > 1$ setting as ML hints received by the online policy.

2.1   Our Results

Our main results are as follows:

- In Section 3, for the online linear optimization problem in the BESTPROBE model, we show a regret bound of $O\left(d^2 \ln d\right)$, independent of T with $k = 2$ probes. Using similar techniques, in the full paper [11], we improve to $O(\ln n)$ for the experts model with n experts, and we also present a horizon-independent regret bound for online convex optimization.
- In Section 3.3, we extend the model to the setting where B of the hints could be imperfect. We show an algorithm with regret bound $O\left(d^2 \ln d \sqrt{B+1}\right)$. The dependence on B is trivially optimal, as can be seen by considering $B = T$. Similar results can be obtained for online convex optimization.
- In Section 4, for the stochastic MAB problem in the BESTPROBE model, using $k = 2$ probes, we obtain a *parameter independent* regret bound of $O(n^2 \log T)$.
- In Section 5, for the stochastic MAB problem in the ALLPROBE model, we show a regret bound of $O(n^2)$, independent of T, with only $k = 3$ probes.

- In Section 6, we consider the stochastic MAB problem in AllProbe model when the rewards of arms can be correlated at each time step (whilst still being independent across time). Here, we show a $\tilde{O}(n^{8/3}T^{1/3})$ regret bound using $k = 4$ probes.

Our regret bounds in Sections 4 and 5 for the MAB problem are *parameter independent*, meaning that there is no dependence of the regret bound on the means and variances of the individual arms. All of our bounds except those in Section 6 are an exponential improvement in T over bounds without probes, since there is a lower bound of $\Omega(\sqrt{nT})$ [6] for the stochastic MAB problem, and a lower bound of $\Omega(\sqrt{T \ln n})$ [16] for the experts problem with $k = 1$ probes.

We also remark that the difference between the last two results is a subtle one. Typically in the stochastic MAB problem, the rewards of the arms are assumed to be independent, but most known results carry over to the case where the rewards can be dependent (as long as the samples are independent across time). Informally, this is because the algorithm only receives a feedback about one arm at each step. This is not the case when $k > 1$, and leads to the main open problem of understanding if we can improve upon the $\Theta(\sqrt{T})$ regret for adversarial bandits in the AllProbe or BestProbe models.

3 Online Linear Optimization in the BestProbe Model

We now consider online linear optimization in the BestProbe model. Recall that in standard online linear optimization, the algorithm needs to play an option (or arm) $w \in \mathcal{W}$ each step, and learns the linear loss function ℓ^t at the end of each step t. We show that with an oracle that can return the better of $k = 2$ options at each step t, we can achieve regret that is independent of T.

For simplicity, we will assume $\mathcal{W}$ is a convex polytope in $\mathbb{R}^d$, though our result easily extends to the case where $\mathcal{W}$ is any finite set of options over which linear functions can be efficiently optimized. In more generality, our results hold as long as there are a finite set of options $S \subseteq \mathcal{W}$, such that for $w \in S$, the set $C_w = \{\vec{v} \in \mathbb{R}^d | w = \mathrm{argmin}_{s \in \mathcal{W}} \langle s, \vec{v} \rangle \}$ has positive volume, and further, $\cup_{w \in S} C_w = \mathbb{R}^d$.

Our results also extend to the more general case of online convex optimization with arbitrary convex domains. For conceptual simplicity, we present only the linear case here and defer the general case to the full paper [11]. In the *special case* of online learning with experts (where the domain $\mathcal{W}$ is the n-dimensional unit simplex), we present an improved regret bound that only depends logarithmically on the number of experts, whilst still being independent of T. The logarithmic dependence on the number of experts is analogous to the standard regret bounds for the experts problem, and the details are presented in the full paper [11].

Finally, all of these algorithms can be extended to the case when hints/probes can be incorrect at a small number of steps. Once again, we describe the algorithm only for online linear optimization, in Section 3.3.

3.1 Algorithm: Differentially Private Regularization

Let us first recap the algorithmic framework of randomized regularization [28] for the setting without probes. Let $L^{t-1} = \sum_{q=1}^{t-1} \ell^q$ denote the sum of the cost vectors till time t. The algorithm chooses a d-dimensional random cost vector x of sufficiently large variance upfront and at step t, chooses the *regularized optimum* action,

$$w^t = \mathrm{argmin}_{w \in \mathcal{W}} \langle L^{t-1} + x, w \rangle.$$

The analysis proceeds in two parts. First it is shown that if L^t were hypothetically used instead of L^{t-1} in the above step, the only regret would be due to adding noise x, and this is independent of the time horizon T. Next, it is shown that since x has large variance, using L^{t-1} instead of L^t produces almost the same distribution of the regularized optimum w^t. These steps trade-off, since the larger the variance of noise, the worse the first step and better the second. The optimal trade-off yields a $O(\sqrt{T})$ bound.

In the probing model, our algorithm LwC will simply sample *two* random vectors x, y and compute the regularized optimal solutions as above. The algorithm will find out which of these solutions has smaller loss at time t, and then choose this solution as its action w^t. Our key lemma (Lemma 5) shows that if the noise vector is chosen so as to satisfy a *differential privacy* property, then the error in the first step above (comparing with L^t) goes away! In other words, the better of two samples produced using the regularized distribution obtained using L^{t-1} will be as good as a sample obtained using L^t.

We note that the use of differentially private noise was first considered in [2], who observed that viewing randomized regularization as differential privacy of the loss across time leads to simpler analysis and somewhat stronger regret bounds. We show an *algorithmic* application of this approach for probing. Indeed, though classical regularization does not require differentially private noise (indeed, it does not even require randomization [57]), this seems critical to achieving our bounds.

Algorithm. We now formally describe the algorithm. Recall that the distribution $\texttt{Laplace}(\beta)$ has density function $f(x) = \frac{1}{2\beta}\exp(-|x|/\beta)$ for $x \in \mathbb{R}$. For $\eta \le 0.4$ being a constant, the algorithm performs these steps at time t.

- Choose $x_j \sim \texttt{Laplace}(d/\eta)$ for each $j \in \{1, \ldots, d\}$; set $a^t = \operatorname{argmin}_{w \in \mathcal{W}}\langle L^{t-1} + x, w\rangle$.
- Choose $y_j \sim \texttt{Laplace}(d/\eta)$ for each $j \in \{1, 2, \ldots, d\}$; set $b^t = \operatorname{argmin}_{w \in \mathcal{W}}\langle L^{t-1} + y, w\rangle$.
- Let $A^t = \langle a^t, \ell^t\rangle$ and $B^t = \langle b^t, \ell^t\rangle$. Probe to learn $w^t = \operatorname{argmin}_{a^t, b^t}\{A^t, B^t\}$.
- Play w^t as the action at time t, incurring actual loss $\min(A^t, B^t)$.

We will show the following theorem:

▶ **Theorem 1.** *The constant $\eta \in (0, 0.4]$, the regret of the* LwC *algorithm is $O\left(d^2 \ln d\right)$.*

3.2 Analysis

As in the classic analysis of regularization [28], define a hypothetical "Be the Regularized Leader" (BTRL) algorithm: Choose $x_j \sim \texttt{Laplace}(d/\eta)$ independently for each $j \in \{1, 2, \ldots, d\}$. At step t, use $c^t = \operatorname{argmin}_{w \in \mathcal{W}}\langle L^{t-1} + \ell^t + x, w\rangle$ as the action taken at step t. Note that BTRL is *not realizable*. Let $D = \max_{w, w' \in \mathcal{W}} |w - w'|_1$. The next lemma restates the classic "be the leader" result from [28].

▶ **Lemma 2** ([28]). *For any $\eta \ge 0$, the regret of the* BTRL *algorithm is at most $D \cdot \mathbf{E}[\max_{j=1}^d |x_j|]$.*

For the specific setting of $x_j \sim \texttt{Laplace}(d/\eta)$, we have $\mathbf{E}[\max_{j=1}^d |x_j|] = O\left(\frac{d}{\eta}\ln d\right)$ assuming η is a constant. Further, we have $D = 2d$, so that we obtain:

▶ **Corollary 3.** *For constant $\eta > 0$, the regret of the* BTRL *algorithm is $O\left(\frac{d^2 \ln d}{\eta}\right)$*

In the rest of the analysis, we focus on a particular step t, and omit the superscript t. Let $\mathcal{D}_1$ denote the distribution of the regularized optimum a^t (resp. b^t) using L^{t-1} in the LwC algorithm, and $\mathcal{D}_2$ denote the distribution of the regularized optimum c^t using L^t in

the BTRL algorithm. The following lemma is a consequence of the well-known Laplace mechanism in differential privacy [19], and we present a proof for completeness.

▶ **Lemma 4** (η-Differential Privacy). *For all $w \in \mathcal{W}$, we have:* $\exp(-\eta) \leq \frac{\Pr[\mathcal{D}_1 = w]}{\Pr[\mathcal{D}_2 = w]} \leq \exp(\eta)$.

Proof. Consider $x, y \sim \texttt{Laplace}(d/\eta)$. Let X_i and Y_i denote the random variables $L_i^{t-1} + x_i$ and $L_i^{t-1} + \ell_i^t + y_i$ respectively. For any fixed value v, for any dimension i, their density functions are related as:

$$\frac{f_{X_i}(v)}{f_{Y_i}(v)} \leq \exp(\ell_i^t \eta/d) \leq \exp(\eta/d),$$

since $\ell_i^t \leq 1$. Let X denote the d-dimensional random variable whose i^{th} dimension is X_i, and similarly define Y. Using the above, if $\vec{x}, \vec{y}$ have components drawn independently from $\texttt{Laplace}(d/\eta)$, then for any $\vec{v} \in \mathbb{R}^d$, the density functions of X and Y are related as

$$\frac{f_X(\vec{v})}{f_Y(\vec{v})} \leq \prod_{i=1}^d \exp(\ell_i^t \eta/d) \leq \exp(\eta).$$

A similar argument shows $\frac{f_Y(\vec{v})}{f_X(\vec{v})} \leq \exp(\eta)$. Since $\mathcal{W}$ is a convex polytope in $\mathbb{R}^d$, the optimum solution for any $\vec{v}$ is achieved at a vertex $w \in \mathcal{W}$. Further, w.l.o.g., the set of $\vec{v}$ whose optimum corresponds to $w \in \mathcal{W}$ define a convex cone in $\mathbb{R}^d$ with positive volume. The lemma follows by integrating the density functions f_X, f_Y over this cone.[1]   ◀

We will now overload notation and use $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) to refer to the distribution of the losses $\langle \ell^t, w \rangle$ for w chosen according to $\mathcal{D}_1$ (resp. $\mathcal{D}_2$). Note that these new distributions are discrete with support size equal to the number of $w \in \mathcal{W}$ that are optimum for some $\vec{v} \in \mathbb{R}^d$, and also satisfy the previous lemma.

The crux of the analysis is the following lemma, which shows that the expected min of the two losses of the regularized optima using L^{t-1} is at most that of the regularized optimum using L^t, that is, the per-step loss of LwC is at most that of the BTRL algorithm. We can view this as a "reverse prophet inequality" that upper bounds the expected minimum instead of lower bounding it.

▶ **Lemma 5** (Reverse Prophet Inequality for Private Noise). *Let A, B be losses drawn independently from $\mathcal{D}_1$ and let C be a loss drawn from $\mathcal{D}_2$, where $\mathcal{D}_1, \mathcal{D}_2$ are of bounded support and satisfy η-differential privacy (Lemma 4) for $\eta \in (0, 0.4]$. Then $\mathbf{E}[\min(A, B)] \leq \mathbf{E}[C]$.*

Proof. By shifting the distributions if necessary, we may assume that the support of $\mathcal{D}_1$ and $\mathcal{D}_2$ is the set of non-negative real numbers. Let $G(x) = \Pr[\mathcal{D}_1 \geq x]$ and $\hat{G}(x) = \Pr[\mathcal{D}_2 \geq x]$. Similarly, let $F(x) = \Pr[\mathcal{D}_1 \leq x]$ and $\hat{F}(x) = \Pr[\mathcal{D}_2 \leq x]$. Let $q = \text{argmin}_x \{x | G(x) \leq \exp(-\eta)\}$. Therefore,

$$\mathbf{E}[C] = \int_{x=0}^q (1 - \hat{F}(x))dx + \int_{x=q}^\infty \hat{G}(x)dx = q - \int_{x=0}^q \hat{F}(x)dx + \int_{x=q}^\infty \hat{G}(x)dx,$$

$$\mathbf{E}[\min(A, B)] = q - \int_{x=0}^q F(x)(2 - F(x))dx + \int_{x=q}^\infty (G(x))^2 dx.$$

---

[1] Note that the proof also extends to the case where $\mathcal{W}$ is any finite set of outcomes, since w.l.o.g., there is a subset $S \subseteq \mathcal{W}$ of outcomes such that the set of $\vec{v}$ for which $w \in S$ is the optimal outcome is continuous and convex (and hence has positive volume), and the union over S of these sets of $\vec{v}$ spans $\mathbb{R}^d$.

Note that for $x \geq q$, $G(x) \leq \exp(-\eta)$. Since $\hat{G}(x) \geq \exp(-\eta)G(x)$, this implies $(G(x))^2 \leq \hat{G}(x)$. Further, for $x \in [0, q]$, we have $F(x) \leq 1 - \exp(-\eta)$, and further, $\hat{F}(x) \leq \exp(\eta)F(x)$. Therefore

$$F(x)(2 - F(x)) \geq \exp(-\eta)(1 + \exp(-\eta))\hat{F}(x) \geq \hat{F}(x)$$

for $\eta \in (0, 0.4]$. Putting this together, we infer $\mathbf{E}[C] \geq \mathbf{E}[\min(A, B)]$, completing the proof. ◀

We note that the proof of the above lemma crucially needs the two-sided bound in Lemma 4. In contrast, the classical regret results for random regularization, for instance, in [28], only require the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ be at most η. However, this weaker condition is insufficient for proving the lemma, as can be seen by the following example: $\mathcal{D}_1$ is a deterministic value d, while $\mathcal{D}_2$ is d with probability $1 - \eta$ and 0 otherwise. Then, $\mathbf{E}[\min(A, B)] = \mathbf{E}[A] = d$, while $\mathbf{E}[C] = d(1 - \eta)$. This leads to a regret of ηdT over T steps against the BTRL algorithm.

Proof of Theorem 1. The proof is now immediate. Lemma 5 implies that at step t, the expected loss of LwC is at most that of the BTRL algorithm. Using linearity of expectation over time steps and combining with the regret bound for BTRL from Corollary 3, we have proved Theorem 1.

3.3   Handling Imperfect Hints

We now consider the setting where at most B of the T hints (comparisons) yield incorrect answers. We show an algorithm that yields regret $O(d^2 \ln d \sqrt{B + 1})$. At one extreme, when $B = T$, this recreates the $O(\sqrt{T})$ regret guarantee for classical online linear optimization, while at the other extreme, when $B = 0$, this recovers Theorem 1. It is also easy to show that such a dependence on B is optimal for any T. To see this, simply construct an instance where the loss function at steps where the hints are correct is identically zero, so that the hints are vacuous. The only relevant steps are the ones where the hints are incorrect, so that any algorithm's regret is lower-bounded by the regret of classical online linear optimization over B steps.

Algorithm. In the sequel, we assume B is known; the case for unknown B follows by the standard doubling trick where we maintain a guess for B and restart the algorithm once B doubles. This can be done since we are in the full information regime and we get to know if a query answer was incorrect; we omit the details. Our algorithm is nearly identical to LwC for $B = 0$, and proceeds as follows. We will set $\eta = \frac{1}{5\sqrt{B+1}}$ in this algorithm. We also have a parameter $p = 5\eta$. The main difference in the algorithm is the following: After LwC probes to learn $w^t = \operatorname{argmin}_{a^t, b^t}\{A^t, B^t\}$, the new algorithm plays w^t as the action with probability p and plays a^t with probability $1 - p$. In other words, the algorithm now uses the hint (action w^t) with probability $1 - p$, else it ignores the hint at that step and mimics classical follow the regularized leader (action a^t). Note that at the end of every step, the algorithm learns if the hint was correct. This allows the algorithm to keep track of the number of mis-predictions and is critical to using the doubling trick for unknown B.

Analysis. We will show the following theorem. We note that the generalization to the experts and online convex optimization settings (as described in the full paper [11]), to yield regret with a $O(\sqrt{B + 1})$ dependence on B, follows the same outline and is hence omitted.

▶ **Theorem 6.** *The regret of the modified* LwC *algorithm with B incorrect hints is $O(d^2 \ln d \sqrt{B + 1})$.*

Our analysis hinges on the following generalization of the reverse prophet inequality (Lemma 5) to mimic the behavior of the algorithm.

▶ **Lemma 7** ((General Reverse Prophet Inequality.)). *Let $p > 0$ be a given parameter, and let $\mathcal{D}_1, \mathcal{D}_2$ be bounded support probability distributions (over losses) that satisfy η-differential privacy (Lemma 4) for some $\eta < p/4$. Let C be a random sample from $\mathcal{D}_2$, and let Z be a random variable obtained as follows: two samples A, B are drawn from $\mathcal{D}_1$. With probability $(1 - p)$, Z is set to be A. With probability p, $Z = \min(A, B)$. Then we have $\mathbf{E}[Z] \leq \mathbf{E}[C]$.*

Proof. Define $G(x)$ and $\hat{G}(x)$ as in Lemma 5. Once again, suppose that the support of the distributions is the set of non-negative reals. By definition, we have

$$\mathbf{E}[C] = \int_0^\infty \hat{G}(x)dx \quad \text{and} \quad \mathbf{E}[Z] = \int_0^\infty \left((1 - p)G(x) + pG(x)^2\right) dx.$$

By the differential privacy property and the choice of η, we have that $G(x), \hat{G}(x)$ are within a factor of $(1 \pm \frac{p}{2})$ of one another for every x. Now consider two cases.

First, suppose $G(x) \leq 1/2$. In this case, $(1 - p)G(x) + pG(x)^2 \leq (1 - \frac{p}{2})G(x) \leq \hat{G}(x)$.

Next, suppose $G(x) > 1/2$. Now, writing $F(x) = (1 - G(x))$ for convenience, we have

$$\begin{aligned}
(1 - p)G(x) + pG(x)^2 &= (1 - F(x))(1 - pF(x)) \\
&= 1 - F(x) - pF(x) + pF(x)^2 \\
&< 1 - F(x) - \frac{p}{2}F(x) = 1 - (1 + \frac{p}{2})F(x).
\end{aligned}$$

In the last inequality, we used $F(x) < 1/2$. By the privacy property, the final expression is $\leq 1 - \hat{F}(x) = \hat{G}(x)$. Plugging this into the above integral completes the proof of the lemma. ◀

We now prove Theorem 6 by bounding the regret against the BTRL algorithm.

Proof of Theorem 6. Consider the BTRL algorithm with the same value of $\eta = \frac{1}{5\sqrt{B+1}}$. By Corollary 3, this has regret $O(d^2 \ln d\sqrt{B + 1})$. We now bound the regret of our algorithm against the BTRL algorithm. Towards this end, set S_1 denote the set of time steps where the hints are correct. By Lemma 7, the loss of our algorithm at these steps is at most that of the BTRL algorithm.

Let S_2 denote the set of steps where the hints are incorrect. At each of these steps, with probability p the algorithm plays w^t, and incurs loss $O(d^2)$. There are $B \cdot p = O(\sqrt{B + 1})$ such steps in expectation, yielding total loss (and regret) at most $O(d^2\sqrt{B + 1})$ against the BTRL algorithm. For the remaining steps, the algorithm plays action a^t. Since Laplace noise is η-differentially private (Lemma 4), the probability that $a^t \neq c^t$ is at most η, where c^t is the action taken by the BTRL algorithm. In the case where $a^t \neq c^t$, the algorithm incurs regret $O(d^2)$. Therefore the total regret against BTRL due to these steps is $O(B\eta d^2) = O(d^2\sqrt{B + 1})$. Combining with the regret of $O(d^2 \ln d\sqrt{B + 1})$ of the BTRL algorithm, this shows our algorithm also has regret $O(d^2 \ln d\sqrt{B + 1})$ against the best fixed action in hindsight. This completes the proof. ◀

4 Stochastic MAB in the BESTPROBE Model with $k = 2$ Probes

Recall that in the stochastic MAB problem in the BESTPROBE model, the policy can probe k arms and learn the identity of the arm with the maximum reward. At the end of the play,

it only learns the reward of the arm that was played that step, and not the rewards of the other arms that were probed that step. We will show a parameter independent regret bound of $O(n^2 \log T)$ for a horizon of T steps using only $k = 2$ probes. As mentioned before, this is an exponential improvement over the $\Omega(\sqrt{nT})$ regret necessary with $k = 1$ probes.

4.1 The `Meta UCB-V` Algorithm

The key algorithmic idea is the following: If we use an optimistic estimate of the sample mean (similar to UCB [45]), probe the top two arms based on this estimate and choose the arm with the higher reward, we obtain a guaranteed advantage (in expectation) over simply choosing the top arm based on the estimate. Our key technical lemmas, Lemma 9 and 10 show that if the gap between the means of the arms played and the mean of the optimal arm is small, the expected regret is actually ≤ 0. This is key to achieving our improved regret bounds.

Formally, define a new classical bandit instance (with no probes) as follows. For every pair of arms (i, j), we have a meta-arm (i, j). Therefore, the new instance as $\binom{n}{2}$ meta-arms. If X_i and X_j denote the random variables corresponding to the rewards of arms i and j respectively, the reward of meta-arm (i, j) is $\max(X_i, X_j)$. Playing the meta-arm (i, j) corresponds to probing the pair of arms i and j and obtaining/observing the value $X_{ij} = \max(X_i, X_j)$.

Our algorithm runs the UCB-V policy [4] on the new bandit instance. We call this algorithm `Meta UCB-V`. For completeness, this algorithm works as follows: At step t, suppose (i, j) has been played s_{ij}^t times. Let m_{ij}^t and V_{ij}^t denote the sample mean and sample variance over the s_{ij}^t plays:

$$m_{ijt} = \frac{\sum_{q=1}^{s_{ijt}} X_{ijq}}{s_{ijt}} \qquad V_{ijt} = \frac{\sum_{q=1}^{s_{ijt}} (X_{ijq} - m_{ijt})^2}{s_{ijt}}. \tag{1}$$

Define the quantity $\texttt{UCB}_{ij}^t$ as: $\texttt{UCB}_{ij}^t = m_{ij}^t + \sqrt{\frac{2.4 V_{ij}^t \log t}{s_{ij}^t}} + \frac{3.6 \log t}{s_{ij}^t}$. At time step t, the meta-arm (i, j) with the highest value of $\texttt{UCB}_{ij}^t$ is played.

We will show the following parameter-independent regret bound as our main result; recall that we are considering pseudo-regret throughout this paper.

▶ **Theorem 8.** *The* `Meta UCB-V` *algorithm has regret* $O(n^2 \log T)$ *with* $k = 2$ *probes.*

Recall that $\mu_i = \mathbf{E}[X_i]$, $\mu^* = \max_i \mathbf{E}[X_i]$, and the benchmark is $\texttt{OPT} = T\mu^*$. We now define analogous quantities for the meta-arms. Recall that $X_{ij} = \max(X_i, X_j)$ is the random variable corresponding to the reward of meta-arm (i, j). Let $\mu_{ij} = \mathbf{E}[X_{ij}]$ and $\sigma_{ij}^2 = \text{Var}[X_{ij}]$. Let $M^* = \max_{(i,j)} \mu_{ij}$ and $\Delta_{ij} = M^* - \mu_{ij}$.

For the `Meta UCB-V` policy, let T_{ij} denote the expected number of times meta-arm (i, j) is played. Let $R_{ij} = T_{ij}\Delta_{ij}$ denote the expected regret against M^* due to playing meta-arm (i, j). The main result of Audibert *et al.* [4] is the following.[2]

$$R_{ij} \leq 10 \left(\frac{\sigma_{ij}^2}{\Delta_{ij}} + 1 \right) \ln T. \tag{2}$$

---

[2] We note that the result in [4] assumes the arms are independent, while our meta-arms are correlated. However, the expected regret bounds in [4] hold as is when the arms are correlated, provided the samples from the arms are *i.i.d.* across time. This suffices for our purposes.

4.2 Reverse Prophet Inequalities

Our main technical lemmas *lower bound* the expected maximum of a pair of random variables in terms of the mean and variance of the individual variables. These lemmas effectively show that if the gap between the means of the arms played and the mean of the optimal arm is small, the expected regret is actually ≤ 0. As mentioned before, these can be viewed as the reverse of standard prophet inequalities [30, 46] that upper bound the expected maximum. These lemmas forms the crux of our analysis both in this section and in Section 5, and may be of independent interest in related settings.

▶ **Lemma 9.** *Let X and Y be independent random variables supported on $[0,1]$ whose means satisfy $\mu_Y \geq \mu_X$, and let $Z = \max(X, Y)$. Then $\mathbf{E}[Z] \geq \mu_X + \sigma_X^2/2$.*

Proof. We can write $\mathbf{E}[Z] = \int_{t=0}^1 \mathbf{E}[Z|X = t]f_X(t)dt$, where f_X is the pdf of X. We split the integral into two sums, $t \in [0, \mu_X]$ and $t \in [\mu_X, 1]$. For the first integral, we use $Z \geq Y$, and thus

$$\int_{t=0}^{\mu_X} \mathbf{E}[Z|X = t]f_X(t)dt \geq \int_{t=0}^{\mu_X} \mathbf{E}[Y]f_X(t)dt \geq \int_{t=0}^{\mu_X} \mu_X f_X(t)dt.$$

Here, the first inequality uses the independence of X and Y and the second inequality uses $\mu_Y \geq \mu_X$. For the second integral, we use $Z \geq X$, and obtain

$$\int_{t=\mu_X}^1 \mathbf{E}[Z|X = t]f_X(t)dt \geq \int_{t=\mu_X}^1 t f_X(t)dt.$$

Thus, writing the t as $\mu_X + (t - \mu_X)$, we get that the sum of the two integrals is bounded as

$$\mathbf{E}[Z] \geq \mu_X + \mathbf{E}[(X - \mu_X)_+], \tag{3}$$

where $(X - \mu_X)_+$ is the random variable $\max(0, X - \mu_X)$.

Likewise, define $(X - \mu_X)_- = \max(0, \mu_X - X)$. By the definition of the expectation, we have $\mathbf{E}[(X - \mu_X)_+] = \mathbf{E}[(X - \mu_X)_-]$. On the other hand, because our random variables are bounded on $[0, 1]$, we have $\mathbf{E}[(X - \mu_X)_+] \geq \mathbf{E}[(X - \mu_X)_+^2]$ and likewise for $\mathbf{E}[(X - \mu_X)_-]$. Therefore,

$$\mathbf{E}[(X - \mu_X)_+] + \mathbf{E}[(X - \mu_X)_-] \geq \mathbf{E}[(X - \mu_X)^2] = \sigma_X^2. \tag{4}$$

Thus both the terms are at least $\sigma_X^2/2$, implying the lemma. ◀

We next extend Lemma 9 to pairs of arms. Note that this does not follow from Lemma 9 by simply replacing "arms" with "pairs of arms", since different pairs of arms are no longer independent.

▶ **Lemma 10.** *Let pair (i, j) be such that $\mu_{ij} < \mu^*$. Then $M^* - \mu_{ij} \geq \frac{\sigma_{ij}^2}{4}$.*

Proof. For notational convenience, let $q = \text{argmax}_s \mu_s$. By assumption, we have $\mathbf{E}[X_q] = \mu^* > \mu_{ij}$.

Since $M^* = \max_{(q,r)} \mu_{qr}$, we have $M^* \geq \max(\mu_{iq}, \mu_{jq})$. Therefore,

$$M^* - \mu_{ij} \geq \max(\mu_{iq}, \mu_{jq}) - \mu_{ij}. \tag{5}$$

We now use the same argument as the proof of Eq (3). Set $Z = \max(X_i, X_q)$ and $X = X_i$, but split the integral at $t = \mu^*$ instead of $t = \mathbf{E}[X_i]$. This yields

$$\mu_{iq} = \mathbf{E}[Z] \geq \int_{t=0}^{\mu^*} \mu^* f_{X_i}(t) dt + \int_{t=\mu^*}^{1} ((t - \mu^*) + \mu^*) f_{X_i}(t) dt = \mu^* + \mathbf{E}[(X_i - \mu^*)_+],$$

A similar inequality holds for μ_{jq}. Combining with Eq (5), we have

$$M^* - \mu_{ij} \geq \max\left\{\mathbf{E}[(X_i - \mu^*)_+], \mathbf{E}[(X_j - \mu^*)_+]\right\} + (\mu^* - \mu_{ij}).$$

Since $\mu^* \geq \mu_{ij}$ by assumption, the above implies

$$M^* - \mu_{ij} \geq \max\left\{\mathbf{E}[(X_i - \mu_{ij})_+], \mathbf{E}[(X_j - \mu_{ij})_+]\right\}$$
$$\geq \frac{1}{2}\left(\mathbf{E}[(X_i - \mu_{ij})_+] + \mathbf{E}[(X_j - \mu_{ij})_+]\right) \geq \frac{1}{2}\mathbf{E}[(X_{ij} - \mu_{ij})_+]$$

To see the last inequality, simply observe that for any value $a \geq 0$, we have $\Pr[X_{ij} \geq \mu_{ij} + a] \leq \Pr[X_i \geq \mu_{ij} + a] + \Pr[X_j \geq \mu_{ij} + a]$. Finally, using Eq (4) with $X = X_{ij}$, we obtain

$$M^* - \mu_{ij} \geq \frac{1}{2}\mathbf{E}[(X_{ij} - \mu_{ij})_+] \geq \frac{1}{4}\sigma_{ij}^2.$$

This completes the proof.   ◀

4.3   Proof of Theorem 8

We split the set of meta-arms into two types. The first type is arms (i, j) for which $\mu_{ij} \geq \mu^*$. Since $\mu^* - \mu_{ij} \leq 0$, the expected regret (against the benchmark μ^*) of playing (i, j) is non-positive.

We therefore focus on arms (i, j) with $\mu_{ij} < \mu^*$. By Lemma 10, we have $\Delta_{ij} = M^* - \mu_{ij} \geq \frac{\sigma_{ij}^2}{4}$. Combining with Eq (2), we have $R_{ij} \leq 10(4 + 1)\ln T = 50\ln T$. However, we have $\Delta_{ij} \geq \mu^* - \mu_{ij}$, so that the regret of (i, j) against μ^* is at most that against M^*. Putting all this together, we have:

$$\text{Regret} = \sum_{(i,j):\mu_{ij} < \mu^*} T_{ij}(\mu^* - \mu_{ij}) \leq \sum_{(i,j):\mu_{ij} < \mu^*} T_{ij}(M^* - \mu_{ij})$$
$$= \sum_{(i,j):\mu_{ij} < \mu^*} T_{ij}\Delta_{ij} = \sum_{(i,j):\mu_{ij} < \mu^*} R_{ij} \leq \sum_{(i,j):\mu_{ij} < \mu^*} 50\ln T = O(n^2 \ln T)$$

This completes the proof of Theorem 8.

5   Stochastic MAB in the ALLPROBE Model

We will now switch to the ALLPROBE model for the stochastic MAB problem, where a policy can probe k arms, and receive as feedback the rewards of all these arms at the current time step. In this model, we will show a regret bound of $O(n^2)$ when we are allowed $k = 3$ probes per step.

In addition to using Lemma 9 from the previous section, the main observation is that with $k = 3$ probes allowed, we can have one probe dedicated to *exploring* the arms in a round-robin fashion. Thus ALLPROBE allows us to overcome the explore/exploit trade-off and obtain better concentration bounds, albeit with the loss of an additional n factor in the regret. The overall algorithm is reminiscent of the UCB-V algorithm of Audibert *et al.* [4] from Section 4, but differs in how we construct the optimistic estimate, as well as the analysis.

Simultaneous Explore-Exploit Algorithm. Our algorithm uses one probe per step to play the n arms in a round robin fashion. We call this the *exploration probe* and it enables the algorithm to maintain the sample mean and sample variance for each arm. Therefore, at time $t \geq 1$, any arm i is observed $s_{it} \geq \lfloor \frac{t}{n} \rfloor$ steps by the exploration probe. Let $X_{i1}, X_{i2}, \ldots, X_{is_{it}}$ denote these observations. Analogous to Eq. (1), let m_{it}, V_{it} denote the sample mean and variance of arm i after s_{it} observations. Choose $\epsilon = 0.1$; any other small constant will work equally well. Let

$$\text{UCB}_{it} = m_{it} + \epsilon V_{it}. \tag{6}$$

The remaining two probes at time step t are used to probe the two arms with the largest values of UCB_{it}. We call these the *exploitation* probes. This completes the description of the algorithm.

Note that in our analysis, we will assume the policy finally plays one of the two exploitation-probed arms and not the exploration-probed arm. We use the exploration probe only to update the estimates of m_{it} and V_{it}, and the results of these probes could be obtained at the end of that time step. This is sufficient to get constant regret. We show the following theorem below.

▶ **Theorem 11.** *The regret of the simultaneous explore-exploit algorithm is $O(n^2)$ for the* ALLPROBE *model with $k = 3$ probes.*

We remark that the bound has a slightly worse dependence on n, the number of arms, than the standard UCB bounds [15]. Improving the bound to $O(n)$ is an interesting open direction. Further, it is easy to show examples where playing the top two arms with highest UCB1 scores [5], or running Thompson Sampling twice [3], has regret that is polynomial in T. However, we conjecture that the policy that simply plays the top two arms according to m_{it} also has constant regret, and we leave showing this as an interesting open question.

5.1 Some Tail Bounds

Before presenting the proof of Theorem 11, we present some well-known tail bounds for the sample mean and variance.

We will use tail bounds that follow from Audibert *et al.* [4, 20]. Fix a time t and some arm i whose distribution D_i has mean μ_i and variance σ_i^2. Note that we assume $D_i \in [0, 1]$. Consider the s_{it} exploration probes and the resulting estimates m_{it} and V_{it} of the sample mean and variance respectively:

$$m_{it} = \frac{\sum_{q=1}^{s_{it}} X_{iq}}{s_{it}} \qquad V_{it} = \frac{\sum_{q=1}^{s_{it}} (X_{iq} - m_{it})^2}{s_{it}}. \tag{7}$$

▶ **Theorem 12** (Implicit in Audibert *et al.*[4]). *For any time $t \geq 1$, we have the following tail bounds on m_{it} and V_{it} after s_{it} probes of arm i.*

1. *For $q \geq \frac{1}{18}$, we have:* $\Pr\left[|m_{it} - \mu_i| > q\sigma_i^2\right] \leq 3e^{-\frac{q\sigma_i^2 s_{it}}{23}}$.

2. *For $q \geq \frac{1}{18}$, we have:* $\Pr\left[V_{it} > (1+q)\sigma_i^2\right] \leq 3e^{-\frac{q\sigma_i^2 s_{it}}{23}}$.

3. $\Pr[V_{it} < 0.65\sigma_i^2] \leq 3e^{-0.01\sigma_i^2 s_{it}}$.

We present a proof sketch of this theorem in the full paper [11].

5.2 Proof of Theorem 11

Let $t_0 = 4n$; we will ignore the first t_0 steps in the analysis, and they contribute $O(n)$ to the regret since the rewards are bounded in $[0, 1]$. This implies $s_{it} \geq \frac{t}{2n}$ for $t \geq t_0$.

Let $q = \text{argmax}_i \mu_i$. Let $\mu^* = \mu_q$ and $\sigma_*^2 = \sigma_q^2$. At any time step t, suppose a pair of arms (i, j) with $\mu_i \geq \mu_j$ is played by the exploitation probes. Let $\Delta_j = \mu^* - \mu_j$. Suppose $\mathbf{E}[\max(X_i, X_j)] = \mu_{ij} \geq \mu^*$, then the expected regret is zero, otherwise, the expected regret is at most Δ_j. We charge this regret to arm j. We will now compute the probability with which a regret of Δ_j is charged to arm j in time step t. Below, we will omit t from the subscript when the connotation is obvious.

Let $\Delta_j = \ell\sigma_j^2 = \rho\sigma_*^2$. We split the analysis into three cases. Note that the key hurdle with obtaining constant regret is that the tail bounds in Theorem 12 that decay exponentially with time at rate depending on variance only hold when the deviation from the mean is at least a constant times the variance of the arm. We therefore split our analysis based on whether Δ_j is at least a constant times σ_j^2 or not. In the latter case, we use Lemma 9 to argue that the regret is already non-positive. In the former case, we obtain exponentially decaying regret in time, which is sufficient to obtain overall constant regret.

In the following, for notational brevity we may drop t from m_{jt}, V_{jt} and s_{jt}.

Case 1. $\ell \leq \frac{1}{2}$. In this case, setting $X = X_j$ and $Y = X_i$ in Lemma 9, we have

$$\mu_{ij} \geq \mu_j + \sigma_j^2/2 \geq \mu_j + \Delta_j \geq \mu^*.$$

Therefore, playing (i, j) incurs non-positive regret.

Case 2. $\ell \geq \frac{1}{2}$ **and** $\rho \geq \frac{1}{2}$. Let $w = \mu_j + \frac{\Delta_j}{2}$. Recall the definition of UCB_{it} from Eq (6), and define the *good event* as $\text{UCB}_j \leq w$ **and** $\text{UCB}_q \geq w$. In this case, if j is probed, then $q = i$, so that there is no regret. We upper bound the probability of the good event not happening by a union of *bad* events.

The first bad event is that $V_j > 0.65\sigma_j^2$, where V_j is as defined in Eq (7). By Theorem 12, there exists constant $c_1 > 0$ such that

$$\Pr[V_j > 0.65\sigma_j^2] \leq 3e^{-c_1 \ell \sigma_j^2 s_j} = 3e^{-c_1 \Delta_j s_j}.$$

Assuming this event do not happen, the second bad event is $\text{UCB}_j > w$, which is equivalent to $m_j > \mu_j + \frac{\ell}{2}\sigma_j^2 - \epsilon V_j$, where m_j is as defined in Eq (7). Since $V_j \leq 0.65\sigma_j^2$, it suffices to bound the probability of the event

$$m_j - \mu_j > (\ell/2 - 0.65\epsilon)\sigma_j^2 > 0.3\ell\sigma_j^2,$$

where we have used $\epsilon = 0.1$. By Theorem 12, there is a constant $c_2 > 0$ so that

$$\Pr[\text{UCB}_j > w | V_j \leq \ell\sigma_j^2] \leq 3e^{-c_2 \ell \sigma_j^2 s_j} = 3e^{-c_2 \Delta_j s_j}.$$

Analogously, the third bad event is that $\text{UCB}_q < w = \mu_j + \frac{\Delta_j}{2} = \mu^* - \frac{\Delta_j}{2}$. This implies $m_q < \mu^* - \frac{\rho}{2}\sigma_*^2$. Repeating the same argument as the second bad event, we obtain

$$\Pr[\text{UCB}_q < w] \leq 3e^{-c_2 \rho \sigma_*^2 s_q} = 3e^{-c_2 \Delta_j s_q}.$$

Denote the regret charged to arm j at step t as R_{jt}. Let $c_3 = \frac{\min(c_1, c_2)}{2}$. Note next that $s_{jt}, s_{qt} \geq \frac{t}{2n}$. We now take the union bound over the three bad events above, and note that the regret charged to arm j conditioned on the bad event is at most Δ_j. This implies:

$$\mathbf{E}[R_{jt}] \leq 9\Delta_j e^{-c_3 \Delta_j \frac{t}{n}}.$$

Case 3. $\ell \geq \frac{1}{2}$ **and** $\rho < \frac{1}{2}$**.** We define the *good* event has having $\mathtt{UCB}_j \leq \mu^*$, and $\mathtt{UCB}_q \geq \mu^*$. In this case, if arm j is played, then arm q is also played so that the regret is zero. As before, we upper bound the probability of the good event not happening by a union of *bad* events.

As in Case (2), $\mathtt{UCB}_j > \mu^*$ is captured by two bad events $V_j > \ell\sigma_j^2$ and $\mathtt{UCB}_j > \mu^*$ given $V_j \leq \ell\sigma_j^2$. It is easy to check that the probability of these events are upper bounded by those derived for Case (2), so that:

$$\Pr[\mathtt{UCB}_j > \mu^*] \leq 6e^{-c_3\ell\sigma_j^2 s_j} = 6e^{-c_3\Delta_j s_j}.$$

To capture $\mathtt{UCB}_q < \mu^*$, we consider two other bad events $V_q < 0.65\sigma_*^2$ and $\mathtt{UCB}_q < \mu^*$ given $V_q \geq 0.65\sigma_*^2$. Using Theorem 12 and the fact that $\Delta_j < \frac{\sigma_*^2}{2}$, we have:

$$\Pr[V_q < 0.65\sigma_*^2] \leq 3e^{-0.01\sigma_*^2 s_q} \leq 3e^{-c_4\Delta_j s_q}$$

for some constant $c_4 > 0$. Assume therefore that $V_q \geq 0.65\sigma_*^2$. The last bad event is $\mathtt{UCB}_q < \mu^*$, which is equivalent to $\mu^* - m_q > \epsilon V_q$, which implies $\mu^* - m_q > 0.065\sigma_*^2$, since $\epsilon = 0.1$. Using Theorem 12 and the fact that $\Delta_j < \frac{\sigma_*^2}{2}$, there is a constant $c_5 > 0$ such that

$$\Pr[\mathtt{UCB}_q < \mu^* | V_q \geq 0.65\sigma_*^2] \leq 3e^{-\frac{0.065}{23}\sigma_*^2 s_q} \leq 3e^{-c_5\Delta_j s_q}.$$

As before, we take the union of all these bad events and set $c_6 = \frac{1}{2}\min(c_3, c_4, c_5)$ to obtain:

$$\mathbf{E}[R_{jt}] \leq 12\Delta_j e^{-c_6\Delta_j \frac{t}{n}}.$$

Given the tail bounds derived in each of the three cases, by linearity of expectation over all time steps t and sub-optimal arms j to which the regret can be charged, we have:

$$\mathbf{E}[\text{Regret}] \leq O(n) + \sum_{j \neq q}\sum_{t=t_0}^{T} 12\Delta_j e^{-c_6\Delta_j \frac{t}{n}} = O(n^2),$$

where use a regret of $t_0 = O(n)$ for the first t_0 steps, and use linearity of expectation beyond that. This completes the proof of Theorem 11.

6   Handling Correlation Between Arms in the ALLPROBE Model

We now consider the ALLPROBE model when the rewards on the arms can be correlated. In other words, the reward vector $\vec{r}_t$ at any time t is drawn from a joint distribution over $[0,1]^n$. These draws are *i.i.d.* across time steps. Note that the analysis for the independent case presented above crucially needs Lemma 9, which does not hold when arms can be correlated. We now show a different algorithm and analysis (in the ALLPROBE model) that uses $k = 4$ probes and achieves a regret bound of $\widetilde{O}(T^{1/3}) \cdot \text{poly}(n)$. Again note that such a dependence on T cannot be achieved in the standard bandit model with $k = 1$ probes.

6.1   Correlation-Exploitation Algorithm

As before the idea is to explore and exploit simultaneously, but the algorithm now plays pairs of arms, and thus can keep track of the *gain* offered by playing two arms simultaneously. More formally, the algorithm uses $k = 4$ probes in every step, and consists of two exploit and two explore probes.

Explore Probes. The explore probes pull every pair (i, j) of arms in a round-robin fashion. Using these probes, the algorithm maintains estimates of the mean reward μ_i, and additionally, estimates of quantities

$$G_{ji} := \mathbb{E}[(X_j - X_i)_+]. \tag{8}$$

Let us call the estimates $\hat{\mu}_i$ and $\hat{G}_{ji}$ respectively. The latter estimates the gain that arm j offers over arm i when played together. These estimates turn out to be crucial in handling correlations in the rewards. Note that since the $\hat{\mu}_i$ uses the same samples used to estimate $\hat{G}_{ji}$, these estimates can be dependent.

Exploit Probes. For every arm i, define its "partner" as $\operatorname{argmax}_j \hat{G}_{ji}$. The partner can change with time, and is a random variable that depends on the rewards obtained so far. At every time t, for the exploit probes, the algorithm pulls the arm i that has the highest value of $\hat{\mu}_i$ – we call this the *primary* arm at time t – and its partner.

6.2 Analysis

We next turn to the analysis of the algorithm described above. Recall that $\Delta_i = \mu_q - \mu_i$, where q is the arm with highest expected reward. We will show the following theorem. Though the improvement is not as impressive as for the independent reward case, we note that such a dependence on T cannot be obtained with $k = 1$ probes.

▶ **Theorem 13.** *The regret of the correlation-exploitation algorithm is bounded by*

$$Regret \leq O\left(n^3 \cdot \sum_{i \neq q} \sqrt{\frac{\log(1/\Delta_i)}{\Delta_i}}\right),$$

which implies a parameter independent regret bound of $\tilde{O}(n^{8/3}T^{1/3})$.

The rest of this section is devoted to proving the above theorem. Towards the end, we will show that the analysis is tight and the dependence on $1/\sqrt{\Delta}$ cannot be improved for this algorithm.

Our proof uses standard Chernoff bounds (see, e.g., [14, 40]). For completeness, we state the version we use.

▶ **Theorem 14.** *Let* $X_1, X_2, \ldots, X_n$ *be independent random variables with support* $[0, 1]$. *Let* $p_i = \mathbf{E}[X_i]$, *and let* $\mu = \sum_i p_i$. *Then we have the following:*
1. *(Small deviation) For any* $\delta \in [0, 1]$,

$$\Pr[X \geq (1 + \delta)\mu] \leq e^{-\mu\delta^2/3},$$
$$\Pr[X \leq (1 - \delta)\mu] \leq e^{-\mu\delta^2/3}.$$

2. *(Large deviation) For* $\delta \geq 1$, *we have*

$$\Pr[X \geq (1 + \delta)\mu] \leq e^{-\mu\delta/3}.$$

Define $Z_{i,t}$ to be the random variable indicating if arm i is the primary arm at time t. The first observation is that for all arms with $\Delta_i > 0$,

$$\Pr[Z_{i,t} = 1] \leq 2 \exp(-t\Delta_i^2/4n). \tag{9}$$

To see this, note that in order to choose arm i over arm q, we must have either $\hat{\mu}_i \geq \mu_i + \frac{\Delta_i}{2}$ or $\hat{\mu}_q \leq \mu_q - \frac{\Delta_i}{2}$. Since we have at least t/n *i.i.d.* samples for each arm and since the rewards are in $[0,1]$, the probability of each of these events can be bounded using Bernstein's inequality, and taking the union over the two events implies (9).

Next, we turn to the analysis of the quantities G_{ij} and their estimates $\hat{G}_{ij}$. The first observation is the following:

▶ **Observation 15.** *For any two arms* (i,j), *we have* $\mathbb{E}[\max(X_i, X_j)] = \mu_i + \mathbb{E}[(X_j - X_i)_+] = \mu_i + G_{ji}$. *Furthermore, for every arm* i, *there exists* j *such that* $G_{ji} \geq \Delta_i$.

This implies that when we play i, if we are able to identify its "optimal partner" j, then we will incur zero regret in expectation. However, since we only estimate G_{ji}, the actual regret can be higher. To analyze this difference, let us define $D_i(t)$ to be $\mathbb{E}[(X_j - X_i)_+]$, where j is the partner of i at time t and the expectation is over the reward distribution. As j is a random variable (depending on the rewards observed at times $t' < t$), so is $D_i(t)$. The key observation is that the expected regret at time t conditioned on i being the explore arm is bounded by $\Delta_i - D_i(t)$.

Our overall approach is to bound the expected regret as

$$\sum_i \sum_t \mathbb{E}[Z_{i,t}(\Delta_i - D_i(t))].$$

To bound this, we fix an index i and analyze the sum over t. We can further bound the sum as:

$$\sum_{t \leq T_i} \mathbb{E}[\Delta_i - D_i(t)] + \sum_{t > T_i} \mathbb{E}[Z_{i,t}]\Delta_i, \tag{10}$$

where $T_i = \frac{4n^2 \log(1/\Delta_i)}{\Delta_i^2}$. The second summation is bounded easily using (9):

$$\sum_{t > T_i} \mathbb{E}[Z_{i,t}]\Delta_i \leq 2\Delta_i \int_{t=T_i}^{\infty} e^{-t\Delta_i^2/4n} \, dt = 2\Delta_i \cdot \frac{4n}{\Delta_i^2} \cdot e^{-T_i\Delta_i^2/4n} \leq O(1).$$

Let us thus focus on the first sum. For the first $\frac{12n^2}{\Delta_i}$ time steps, we simply bound the expectation by Δ_i, which makes the sum add up to $O(n^2)$. As t increases, the following lemma shows that $(\Delta_i - D_i(t))$ becomes much smaller than Δ_i with high probability.

▶ **Lemma 16.** *Assume w.l.o.g. that* $\Delta_i < 1$. *Suppose that* $t \geq 12n^2/\Delta_i$. *Then for any* $c \geq 0$,

$$\Pr[\Delta_i - D_i(t) \geq c] \leq \Pr[\Delta_i - D_i(t) > c\Delta_i] \leq ne^{-c^2\Delta_i t/36n^2}.$$

Proof. First note that we can assume $0 < c < 1$; the bound for $c \geq 1$ is trivial. The proof proceeds in two parts, both of which use the fact that after t steps, the explore arms (which perform round robin) use at least (t/n^2) samples for computing each of the $\hat{G}_{ji}$. First, we show that for the optimal arm q, since $G_{qi} \geq \Delta_i$ from Observation 15,

$$\Pr[\Delta_i - \hat{G}_{qi} > \frac{c}{2}\Delta_i] \leq \Pr[\hat{G}_{qi} < (1 - \frac{c}{2})G_{qi}] \leq e^{-c^2\Delta_i t/12n^2}.$$

This is a direct application of Theorem 14. Next, we consider any arm j for which G_{ji} is $\leq \Delta_i(1-c)$. In this case, we wish to argue that $\hat{G}_{ij} < \Delta_i - \frac{c}{2}\Delta_i$ with high probability. For this, we consider two cases.

Case 1. $G_{ji} < \Delta_i/4$. In this case, $\Delta_i - \frac{c}{2}\Delta_i > \Delta_i/2 > 2G_{ji}$, and thus to bound $\Pr[\hat{G}_{ji} \geq 2G_{ji}]$, we can use the "high deviation" regime of Chernoff bounds (see Theorem 14) to conclude that

$$\Pr[\hat{G}_{ji} \geq \Delta_i - \frac{c}{2}\Delta_i] \leq \Pr[\hat{G}_{ji} \geq \Delta_i/2] \leq e^{-t\Delta_i/6n^2}.$$

Case 2. $G_{ji} \geq \Delta_i/4$. In this case, since $\Delta_i - \frac{c}{2}\Delta_i \geq (1 + \frac{c}{2})G_{ji}$, we can use a Chernoff bound again, to obtain

$$\Pr[\hat{G}_{ji} \geq \Delta_i - \frac{c}{2}\Delta_i] \leq e^{-c^2 G_{ji}t/12n^2} \leq e^{-c^2\Delta_i t/48n^2}.$$

Combining the two parts and taking a union bound, we have that with probability at least $1 - ne^{-c^2\Delta_i t/48n^2}$, we have that (a) $\max_j \hat{G}_{ji} \geq \Delta_i - \frac{c}{2}\Delta_i$, and (b) the max is not attained by any j with $G_{ij} \leq G_{ji}(1-c)$. If both (a) and (b) hold, then $\Delta_i - D_i(t) \leq c\Delta_i$, and this completes the proof of the lemma.   ◄

Lemma 16 can be used to bound the first term of (10), using the following technical lemma.

▶ **Lemma 17.** *Let $n, \beta \geq 1$ be parameters, and Y be a random variable that satisfies the condition:*

$$\forall c > 0, \Pr[Y \geq c] \leq ne^{-c^2\beta}.$$

Then $\mathbb{E}[Y] \leq \frac{2n}{\sqrt{\beta}}$.

Proof. Since we are only interested in an upper bound on $\mathbb{E}[Y]$, we can ignore potential negative values of Y, and write

$$\mathbb{E}[Y] \leq \int_{c=0}^{\infty} \Pr[Y \geq c]\, dc \leq n\int_{c=0}^{\infty} e^{-c^2\beta} dc.$$

We then split the integrals into a sum over the intervals $c \in [0, \frac{1}{\sqrt{\beta}}], [\frac{1}{\sqrt{\beta}}, \frac{2}{\sqrt{\beta}}], [\frac{2}{\sqrt{\beta}}, \frac{3}{\sqrt{\beta}}], \ldots$. As the integrand in the $(i+1)$th interval is bounded by $\exp(-i^2)$, the sum can be bounded as desired.   ◄

Using the lemmas, we can bound the first summation in (10). The main observation is that by using Lemma 16, if $t = \beta\frac{48n^2}{\Delta_i}$, the hypothesis of Lemma 17 is satisfied for $Y = \Delta_i - D_i(t)$. This implies that we can bound

$$\mathbb{E}[\Delta_i - D_i(t)] \leq \frac{2n}{\sqrt{\beta}} = 12n^2\sqrt{\frac{\Delta_i}{t}}.$$

Summing this between $t = \frac{48n^2}{\Delta_i}$ (or even $t = 1$) and $t = T_i = \frac{4n^2\log(1/\Delta_i)}{\Delta_i^2}$, we obtain a bound of $O\left(n^2\sqrt{\Delta_i T_i}\right)$. Plugging in the value of T_i then completes the proof of Theorem 13.

6.2.1 Tight Instance

We now show an instance where the above algorithm has regret $\Omega(n/\sqrt{\Delta})$. There are three arms and $n - 3$ dummy arms for large n. Arm 1 has reward X_t that is drawn *i.i.d.* from a Bernoulli distribution that is $1/3$ with probability $1/2$ and $2/3$ with probability $1/2$. Arm 2 has reward $Y_t = X_t + A_t$, where A_t is *i.i.d.* drawn from Bernoulli$(1/3, 3\Delta)$. Arm 3 has

reward $Z_t = X_t + B_t$, where B_t is *i.i.d.* drawn from Bernoulli$(1/3, 3\Delta(1 - \sqrt{\Delta}))$, with $\Pr[B_t = 0 | A_t = 0] = 1$.

The dummy arms have reward zero at all time steps. We assume that at every step, one pair of arms (i, j) is sampled, and these samples are used to estimate G_{ij}, G_{ji}, μ_i, and μ_j. The dummy arms ensure the estimates $\hat{\mu}_i$ are approximately independent for all $i, j \in \{1, 2, 3\}$. Further, $\hat{G}_{ji}$ and $\hat{G}_{ki}$ are independent for all $i, j, k \in \{1, 2, 3\}$, since these estimates are constructed at different time steps.

On this instance, the regret is with respect to arm 2, with $\mathbf{E}[Y] = \frac{1}{2} + \Delta$. However, the construction of reward distributions ensures that $\mathbf{E}[\max(X, Y)] = \mathbf{E}[\max(Z, Y)] = \mathbf{E}[Y]$. Therefore, at any step, the expected regret of any strategy that plays pairs of arms is non-negative.

Using the tightness of Chernoff bounds on Bernoulli distributions, we can check that on this instance, with constant probability, the following two events happen for all $t \le \Omega(n/\Delta^2)$:
- $\hat{\mu}_1 > \max(\hat{\mu}_2, \hat{\mu}_3)$, so that the algorithm plays arm 1;
- $\hat{G}_{31} > \hat{G}_{21}$, so that arm 3 is the partner of arm 1 and gets played.

In this event, the algorithm incurs regret $\Delta^{3/2}$ against the optimal arm 2 each step, for a total regret of $\Omega(n/\sqrt{\Delta})$. This shows the analysis above is tight.

7 Conclusion

We conclude with some open questions. The main open question is whether the stochastic assumptions are needed for the MAB results. In other words, can we obtain improved regret guarantees for the adversarial MAB problem [6]. We make progress in this direction with our results for correlated MAB in Section 6; however, we believe our results even for this case can be improved.

For the stochastic MAB problem, one intriguing open question is whether constant regret is possible for $k = 2$ probes in the AllProbe model. Note that Theorem 8 implies a regret of $O(n^2 \log T)$. However, unlike the celebrated Lai-Robbins result [32] that shows the $\log T$ factor is necessary when $k = 1$, we have not been able to show a lower bound requiring such dependence on T for $k = 2$, either for the AllProbe model or for the BestProbe model. We leave this as an interesting open question. Another interesting question is to extend our bandit results to the case with imperfect hints.

At a higher level, it would be interesting to explore the power of a few probes in more complex bandit settings. One example is the linear contextual bandit problem [1, 35] where the stochastic arms correspond to latent variables. At any step, a decision space is given and the policy needs to choose a linear combination of these variables from the decision space, obtaining that linear combination of the reward of the arms as its reward. Now suppose the latent space of variables has small dimension, then does having multiple probes help with the regret bounds?

References

1 Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, page 2312–2320, Red Hook, NY, USA, 2011. Curran Associates Inc.

2 Jacob D Abernethy, Young Hun Jung, Chansoo Lee, Audra McMillan, and Ambuj Tewari. Online learning via the differential privacy lens. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

3 Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

4 Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

5 Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002.

6 Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, jan 2003.

7 Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, sep 1999.

8 Hedyeh Beyhaghi and Robert Kleinberg. Pandora's problem with nonobligatory inspection. In Anna Karlin, Nicole Immorlica, and Ramesh Johari, editors, *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 131–132. ACM, 2019.

9 Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online learning with imperfect hints. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 822–831. PMLR, 13–18 Jul 2020.

10 Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Logarithmic regret from sublinear hints. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28222–28232. Curran Associates, Inc., 2021. URL: `https://proceedings.neurips.cc/paper/2021/file/edb947f2bbceb132245fdde9c59d3f59-Paper.pdf`.

11 Aditya Bhaskara, Sreenivas Gollapudi, Sungjin Im, Kostas Kollias, and Kamesh Munagala. Online learning and bandits with queried hints, 2022. URL: `https://arxiv.org/abs/2211.02703`, `doi:10.48550/ARXIV.2211.02703`.

12 Aditya Bhaskara, Sreenivas Gollapudi, Kostas Kollias, and Kamesh Munagala. Adaptive probing policies for shortest path routing. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

13 David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1 – 8, 1956.

14 Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. `doi:10.1093/acprof:oso/9780199535255.001.0001`.

15 Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

16 Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, may 1997.

17 Ofer Dekel, Arthur Flajolet, Nika Haghtalab, and Patrick Jaillet. Online learning with a hint. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper/2017/file/22b1f2e0983160db6f7bb9f62f4dbb39-Paper.pdf`.

18 Amol Deshpande, Lisa Hellerstein, and Devorah Kletenik. Approximation algorithms for stochastic submodular set cover with applications to boolean function evaluation and min-knapsack. *ACM Trans. Algorithms*, 12(3):42:1–42:28, 2016.

19 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 2014.

20 David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

21 A. Goel, S. Guha, and K. Munagala. Asking the right questions: Model-driven optimization using probes. In *Proc. of the 2006 ACM Symp. on Principles of Database Systems*, 2006.

22 Sreenivas Gollapudi and Debmalya Panigrahi. Online algorithms for rent-or-buy with expert advice. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2319–2327, 2019.

23 Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artif. Int. Res.*, 42(1):427–486, September 2011.

24 S. Guha, K. Munagala, and S. Sarkar. Optimizing transmission rate in wireless channels using adaptive probes. In *SIGMETRICS/Performance*, pages 381–382, 2006.

25 James Hannan. *Approximation to BAYES risk in repeated play*, pages 97–140. Princeton University Press, 2016.

26 Elad Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019. URL: `http://arxiv.org/abs/1909.05207`.

27 Sungjin Im, Ravi Kumar, Aditya Petety, and Manish Purohit. Parsimonious learning-augmented caching. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9588–9601. PMLR, 17–23 Jul 2022.

28 Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005. Learning Theory 2003.

29 Satyen Kale. Multiarmed bandits with limited expert advice. *CoRR*, abs/1306.4653, 2013. URL: `http://arxiv.org/abs/1306.4653`, arXiv:1306.4653.

30 Ulrich Krengel and Louis Sucheston. Semiamarts and finite values. *Bulletin of the American Mathematical Society*, 83(4):745 – 747, 1977.

31 Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 767–776, Lille, France, 07–09 Jul 2015. PMLR.

32 T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

33 Silvio Lattanzi, Thomas Lavastida, Benjamin Moseley, and Sergei Vassilvitskii. Online scheduling via learned weights. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1859–1877. SIAM, 2020.

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

35 Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery.

36   N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. URL: `https://www.sciencedirect.com/science/article/pii/S0890540184710091`, `doi:https://doi.org/10.1006/inco.1994.1009`.

37   Zhen Liu, Srinivasan Parthasarathy, Anand Ranganathan, and Hao Yang. Near-optimal algorithms for shared filter evaluation in data stream systems. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, page 133–146, New York, NY, USA, 2008.

38   Thodoris Lykouris and Sergei Vassilvtiskii. Competitive caching with machine learned advice. In *International Conference on Machine Learning*, pages 3302–3311, 2018.

39   Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 646–662. Cambridge University Press, 2020. `doi:10.1017/9781108637435.037`.

40   Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995. `doi:10.1017/cbo9780511814075`.

41   Samrat Mukhopadhyay, Sourav Sahoo, and Abhishek Sinha. k-experts - online policies and fundamental limits. *CoRR*, abs/2110.07881, 2021. URL: `https://arxiv.org/abs/2110.07881`, `arXiv:2110.07881`.

42   K. Munagala, S. Babu, R. Motwani, and J. Widom. The pipelined set cover problem. *Proc. Intl. Conf. Database Theory*, 2005.

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

44   Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 993–1019. JMLR.org, 2013.

45   Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. URL: `https://doi.org/`, `doi:bams/1183517370`.

46   Ester Samuel-Cahn. Comparison of threshold stop rules and maximum for independent nonnegative random variables. *The Annals of Probability*, 12(4):1213 – 1216, 1984. `doi:10.1214/aop/1176993150`.

47   Yevgeny Seldin, Koby Crammer, and Peter Bartlett. Open problem: Adversarial multiarmed bandits with limited advice. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 1067–1072, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

48   Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

49   Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Ranked bandits in metric spaces: Learning diverse rankings over large document collections. *J. Mach. Learn. Res.*, 14(1):399–436, feb 2013.

50   Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proc. the 31th International Conference on Machine Learning, ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1593–1601, 2014.

51   Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL: `https://proceedings.neurips.cc/paper/2008/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf`.

52   William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

53   A. Wald. *Sequential analysis*. John Wiley, 1947.

54   Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory,*

COLT 2018, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

55 Martin L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979.

56 Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

57 Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003.

58 Jinhang Zuo, Xiaoxi Zhang, and Carlee Joe-Wong. Observe before play: Multi-armed bandit with pre-observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):7023–7030, Apr. 2020.