Complex Bandit Problems and Thompson Sampling

Aditya Gopalan
Department of Electrical Engineering
Technion, Israel
aditya@ee.technion.ac.il

Shie Mannor
Department of Electrical Engineering
Technion, Israel
shie@ee.technion.ac.il

Yishay Mansour
School of Computer Science
Tel Aviv University, Israel
mansour@tau.ac.il

Abstract

We study stochastic multi-armed bandit settings with complex actions derived from the basic bandit arms, e.g., subsets or partitions of basic arms. The decision maker is faced with selecting at each round a complex action instead of a basic arm. We allow the reward of the complex action to be some function of the basic arms' rewards, and so the feedback observed may not necessarily be the reward per-arm. For instance, when the complex actions are subsets of bandit arms, we may only observe the maximum reward over the chosen subset. Feedback from playing (complex) actions can thus be indicative of rewards from other actions, and leveraging this coupled feedback becomes important to the decision maker in order to learn efficiently. We propose applying Thompson Sampling – a Bayesian-inspired algorithm for the standard multi-armed bandit – for minimizing regret in complex bandit problems. We derive the first general, frequentist regret bound for Thompson sampling in complex bandit settings, that holds without specific structural assumptions on the prior used by the algorithm. The regret bound exhibits the standard logarithmic scaling with time but with a non-trivial multiplicative constant that encodes the coupled information structure of the complex bandit. As applications, we show improved regret bounds (compared to treating the complex actions as independent) for a class of complex, subset-selection bandit problems. Using particle filters for computing posterior distributions that often lack an explicit closed-form, we apply Thompson-sampling algorithms for subset selection and job-scheduling problems and present numerical results.

Keywords: Multi-armed bandit, Complex bandit, Thompson sampling, Posterior Sampling

1 Introduction

The Multi-Armed Bandit (MAB) is a classical framework in machine learning and optimization. In the basic MAB setting, there is a finite set of actions, each of which has a reward derived from some stochastic process, and a learner selects actions to optimize long-term performance. The MAB framework gives a crystallized abstraction of a fundamental decision problem – whether to explore or exploit in the face of uncertainty. Bandit problems have been extensively studied, and several well-performing methods now exist for regret minimization [1, 2, 3]. However, the requirement that the actions' rewards be independent is often a severe limitation, as in the following examples.

(a) Web Advertising: Assume a publisher controlling advertisements on a web-site, selecting each time a (small) subset of ads to be displayed to the user. As the publisher is paid per click, it would like to maximize its revenue or equivalently the click-through probability for the set of displayed ads. However, the dependency between different ads causes the problem not to decompose nicely. For example, showing two car ads might not significantly increase the click probability over a single car ad. *(b) Job Scheduling:* Assume we have a small number of resources (say, machines) and in each time step we receive a set of jobs (the basic "arms"), where the duration of each job follows some fixed but unknown distribution. The latency of a machine is the sum of the latencies of the jobs (basic arms) assigned to it, and the makespan of the system is the maximum latency over the machines. Here, the decision maker's complex action is to partition the jobs (basic arms) between the machines to minimize the makespan.

Such examples motivate settings where a more complex model than the simple MAB is required. A feature of these problems is that instead of knowing the rewards of basic actions, we can only hope to receive an aggregate reward for the complex action taken, which is typically "coupled" across basic arms (e.g., the net latency in the job scheduling problem contains information about individual job durations). Using this aggregate feedback to reason and make decisions about complex actions is thus a non-trivial problem.

A natural algorithmic prescription for such situations is Thompson sampling [4, 5, 6]: Maintain a fictitious prior distribution over the basic parameters, draw a random sample of parameters, play the best (complex) action for the sample and update the prior to the posterior. The main advantage of a "pseudo-Bayesian" approach such as Thompson sampling, compared to other MAB methodologies such as UCB, is that it can handle a wide range of information models that go beyond observing the individual rewards alone. For example, suppose we observe only the final makespan for scheduling jobs on machines as above. In Thompson sampling, we merely need to compute a posterior given this observation and use it. In contrast, it seems difficult to adapt an algorithm such as UCB to handle this case without having a potentially exponential dependence on the number of basic arms [1]. The Bayesian view taken by Thompson sampling also allows us to use efficient numerical algorithms such as particle filtering [8] to estimate and track posterior distributions.

Our main analytical result is a general regret bound for Thompson sampling in complex bandit settings. The bound for this general setting scales logarithmically with time, as is standard in stochastic bandit results. But more interestingly, the constant for this logarithmic scaling can be explicitly characterized in terms of the bandit's KL divergence geometry, and captures the "information complexity" of the bandit problem. Recent work has shown the regret-optimality of Thompson sampling for the basic MAB [6, 9], and has even provided regret bounds for a very special complex bandit setting – the linear bandit case where the reward is a linear function of the actions [10]. However, the analysis of general complex bandits poses challenges that cannot be overcome using the techniques in existing work. Indeed, these existing proof techniques rely heavily on the structure of the prior and posterior distributions, and break down when analyzing the evolution of the (often not closed-form) posterior distributions from complex observations. In contrast, we develop a new proof technique for analyzing Thompson sampling that, with almost no structural assumptions, allows us to derive regret bounds for complex feedback [2]. Our result thus generalizes performance results for posterior sampling algorithms.

We apply the general regret bound to derive corollaries for a class of subset selection complex bandit problems, which show significant gains over treating all actions independently. Our theoretical findings are complemented by numerical studies of Thompson sampling for two complex bandit scenarios – subset selection from a bandit and job scheduling.

Setup and Notation: Consider a general stochastic model $X_1, X_2, ...$ of independent and identically distributed random variables drawn from $\mathbb{R}^{\mathsf{N}}$ (N is taken to be the dimension of the underlying MAB). The distribution of each X_t is parametrized by $\theta^* \in \Theta$, where Θ denotes the set of candidate parameters. At each time t, an action A_t is played from a set of candidate actions $\mathcal{A}$, following which the decision maker obtains a stochastic observation $Y_t = f(X_t, A_t) \in \mathcal{Y}$ and a reward $g(f(X_t, A_t)) \in \mathbb{R}$. Here, f and g are general fixed functions, and we will often denote $g \circ f$ by the function h. We denote by $l(y; a, \theta)$ the likelihood of observing y upon playing action a, when the distribution parameter is θ. For $\theta \in \Theta$, let $a^*(\theta)$ be an action that yields the highest expected reward for a model with parameter θ,

[1] Dani et al. [7] extended the UCB framework to the case of complex actions with *linear* rewards. However, for bandits with nonlinear reward functions, it is unclear how UCB-like algorithms can be applied apart from treating all the complex actions independently.

[2] The complete proofs of all results can be found in [11].

Algorithm 1 Thompson Sampling

Input: Parameter space Θ, action space $\mathcal{A}$, output space $\mathcal{Y}$, likelihood $l(y; a, \theta)$, **Parameter:** (Prior) Distribution π over Θ. **Initialize** $\pi_0 = \pi$. **At each time** $t \geq 1$, **Draw** $\theta_t \in \Theta$ according to the distribution π_{t-1}, **Play** $A_t = a^*(\theta_t)$, **Observe** $Y_t = f(X_t, A_t)$, and **Update** π_{t-1} to π_t: $\pi_t(d\theta) \propto l(Y_t; A_t, \theta)\pi_{t-1}(d\theta), \quad \pi_t(\Theta) = 1$.

i.e., $a^*(\theta) := \arg\max_{a \in \mathcal{A}} \mathbb{E}_\theta[h(X_1, a)$, with arbitrary tie-breaking [3]. We use $e^{(j)}$ to denote the j-th unit vector in finite-dimensional Euclidean space. The goal is to play an action at each time t to minimize the (expected) *regret* over T rounds: $R_T := \sum_{t=1}^{T} h(X_t, a^*(\theta^*)) - h(X_t, A_t)$, or alternatively, the number of plays of suboptimal actions[4]: $\sum_{t=1}^{T} \mathbb{1}\{A_t \neq a^*\}$.

2 Regret Performance: Formal Results

Our main result is a high-probability regret bound for Thompson sampling for large enough time horizons. We prove the bound under the following assumptions about the parameter space Θ, action space $|\mathcal{A}|$, observation space $|\mathcal{Y}|$, and the fictitious prior π.

Assumption 1 (Finitely many actions, observations). *The action and observation spaces $\mathcal{A}$ and $\mathcal{Y}$ are finite: $|\mathcal{A}|, |\mathcal{Y}| < \infty$.*

Assumption 2 (Bounded rewards). [5] $\forall x \in \mathcal{X}, a \in \mathcal{A} : h(x, a) \in [0, 1]$.

Assumption 3 (Discrete prior, and "Grain of truth"). *The prior distribution π is supported over a discrete set of particles: $\Theta = \{\theta_1, \ldots, \theta_L\}$, with $\theta^* \in \Theta$ and $\pi(\theta^*) > 0$. Furthermore, there exists $\Gamma \in (0, 1/2)$ such that $\Gamma \leq l(y; a, \theta) \leq 1 - \Gamma$ $\forall \theta \in \Theta, a \in \mathcal{A}, y \in \mathcal{Y}$.*

Assumption 4 (Unique best action). [6] $\mathbb{E}[h(X_1, a^*)] > \max_{a \in \mathcal{A}, a \neq a^*} \mathbb{E}[h(X_1, a)]$.

We denote by $D(\theta_a^* || \theta_a)$ the Kullback-Leibler divergence between the output distributions $\{l(y; a, \theta^*) : y \in \mathcal{Y}\}$ and $\{l(y; a, \theta) : y \in \mathcal{Y}\}$. For each action $a \in \mathcal{A}$, let $S_a := \{\theta \in \Theta : a^*(\theta) = a\}$ be the set of parameters for which playing a is optimal. For $a \neq a^*$, let $S'_a := \{\theta \in S_a : D(\theta_{a^*}^* || \theta_{a^*}) = 0\}$. We can now state our main result – a regret bound for Thompson sampling for general complex bandits.

Theorem 1 (Thompson Sampling, General Regret Bound). *Under Assumptions 1-4, the following holds for the Thompson Sampling algorithm. For $\delta, \epsilon \in (0, 1)$, there exists $T^\star \geq 0$ such that for all $T \geq T^\star$, with probability at least $1 - \delta$, $\sum_{a \neq a^*} N_T(a) \leq \mathsf{B} + \mathsf{C}(\log T)$, where $\mathsf{B} \equiv \mathsf{B}(\delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ is a problem-dependent constant that does not depend on T, and* [7]*:*

$$\mathsf{C}(\log T) := \max\left\{ \sum_{k=1}^{|\mathcal{A}|-1} z_k(a_k) : \forall j, k \in [|\mathcal{A}| - 1] \min_{\theta \in S'_{a_k}} \langle z_k, D_\theta \rangle \geq \frac{1+\epsilon}{1-\epsilon} \log T, \min_{\theta \in S'_{a_k}} \langle z_k - e^{(j)}, D_\theta \rangle < \frac{1+\epsilon}{1-\epsilon} \log T \right\}. \quad (1)$$

The maximum above is taken over sequences of suboptimal actions $\{a_k\}_{k=1}^{|\mathcal{A}|-1} = \mathcal{A} \setminus \{a^*\}$, and integer vectors $z_k \in \mathbb{Z}_+^{|\mathcal{A}|-1} \times \{0\}$ satisfying $z_i \succeq z_k, z_i(a_k) = z_k(a_k)$ $\forall i \geq k$.

The usefulness of Theorem 1 lies in the fact that it can couple information across complex actions and give better leading constants for regret scaling than the standard decoupled MAB.

2.1 Application: Playing Subsets of Bandit Arms, "Full Information"

Let us take a standard N-armed Bernoulli bandit with arm parameters $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_N$. Suppose the (complex) actions are all size M subsets of the N arms. Following the choice of a subset, we get to observe the rewards of all the M chosen arms (thus the output space is $\{0, 1\}^M$), and receive some bounded reward of the chosen arms.

A natural (discrete) prior for this problem can be obtained by discretizing each of the N basic dimensions and putting uniform mass over all points: $\Theta = \left\{\beta, 2\beta, \ldots \left(\lfloor \frac{1}{\beta} \rfloor - 1\right)\beta\right\}^N$, $\beta \in (0, 1)$, $\pi(\theta) = \frac{1}{|\Theta|}$ $\forall \theta \in \Theta$. We can then show:

Corollary 1. *Suppose $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_N) \in \Theta$ and $\mu_{N-M} < \mu_{N-M+1}$. Then, the following holds for the Thompson sampling algorithm. For $\delta, \epsilon \in (0, 1)$, there exists $T^\star \geq 0$ such that for all $T \geq T^\star$, with probability at least $1 - \delta$, $\sum_{a \neq a^*} N_T(a) \leq \mathsf{B}_2 + \left(\frac{1+\epsilon}{1-\epsilon}\right) \sum_{i=1}^{N-M} \frac{1}{D(\mu_i || \mu_{N-M+1})} \log T$, where $\mathsf{B}_2 \equiv \mathsf{B}_2(\delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ does not depend on T.*

[3] The subscript θ denotes the probability measure parametrized by θ, and by default, the absence of a subscript is to be understood as working with the parameter θ^*.

[4] We refer to this latter objective as regret since, under bounded rewards, both the objectives scale similarly with the problem size.

[5] In general, any upper bound on the absolute value of the reward function suffices.

[6] This assumption is made only for the sake of notational convenience and does not affect the essence of this paper's results.

[7] $\mathsf{C}(\log T) \equiv \mathsf{C}(T, \delta, \epsilon, \mathcal{A}, \mathcal{Y}, \Theta, \pi)$ as well, but we suppress the dependence on the problem parameters since we are mainly concerned with the constant for the time scaling.

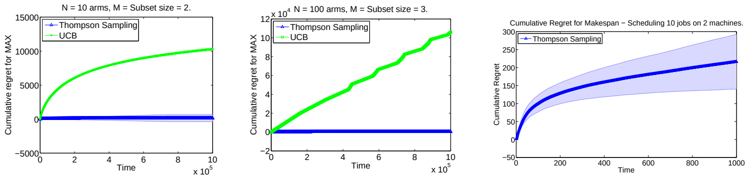
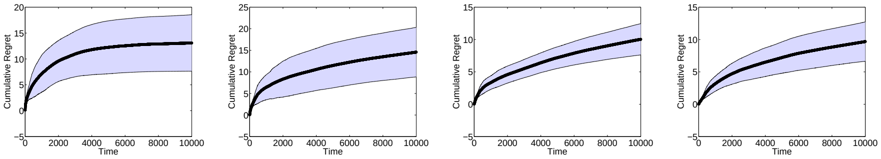


Figure 1: **Left and center:** Cumulative regret with observing the maximum of a pair out of 10 arms (left), and that of a triple out of 100 arms (center), for (a) Thompson sampling using a particle filter, and (b) UCB treating each subset as a separate actions. The arm means are chosen to be equally spaced in $[0, 1]$. The regret is averaged across 150 runs, and the confidence intervals shown are ± 1 standard deviation. **Right:** Cumulative regret with respect to the best makespan with particle-filter-based Thompson sampling, for scheduling 10 jobs on 2 machines. The job means are chosen to be equally spaced in $[0, 10]$. The best job assignment gives an expected makespan of 31. The regret is averaged across 150 runs, and the confidence intervals shown are ± 1 standard deviation.



(a) $N = 10, M = 3$ (b) $N = 10, M = 4$ (c) $N = 10, M = 5$ (d) $N = 10, M = 6$

Figure 2: Cumulative regret with observing the maximum value of M out of $N = 10$ arms for Thompson sampling. The prior is uniform over the discrete domain $\{0.1, 0.3, 0.5, 0.7, 0.9\}^N$, with the arms' means lying in the same domain (setting of Theorem 1). The regret is averaged across 10 runs, and the confidence intervals shown are ± 1 standard deviation.

Discussion: The result illustrates the power of additional information from observing several arms of a bandit. Even though the total number of actions $\binom{N}{M}$ can be exponential in M, the regret bound still scales as $O((N-M)\log T)$. Note also that for $M = 1$ (the standard MAB), the regret scaling is essentially $\sum_{i=1}^{N-M} \frac{1}{D(\mu_i || \mu_{N-M+1})} \log T$, which is interestingly the best known regret bound for standard Bernoulli bandits obtained by specialized, regret-optimal algorithms such as KL-UCB [3], and more recently, Thompson Sampling with the Beta prior [9].

2.2 Application: Playing Subsets of Bandit Arms, MAX Reward

Using the same setting and size-M subset actions as before but *not* being able to observe all the individual arms' rewards results in much more interesting bandit settings. Here, we assume that we get to observe as reward the *maximum* reward from the M chosen arms of a standard N-armed Bernoulli bandit. The feedback is still aggregated across basic arms but at the same time very different from the full information case – observing a reward of 0 is very uninformative whereas a value of 1 is highly informative about the constituent arms. We can apply the general machinery of Theorem 1 to obtain a non-trivial regret bound in this complex feedback setting. Let $\beta \in (0, 1)$, and suppose that $\Theta = \{1 - \beta^R, 1 - \beta^{R-1}, \ldots, 1 - \beta^2, 1 - \beta\}^N$, for positive integers R and N. As before, let $\mu \in \Theta$ denote the basic arms' parameters, and let $\mu_{min} := \min_{a \in \mathcal{A}} \prod_{i \in a} (1 - \mu_i)$.

Corollary 2. *For $0 \le M \le N, M \neq \frac{N}{2}$, $\delta, \epsilon \in (0, 1)$, there exists $T^\star \ge 0$ such that Thompson sampling satisfies: for all $T \ge T^\star$, with probability at least $1 - \delta$, $\sum_{a \neq a^*} N_T(a) \le \mathsf{B}_3 + (\log 2) \left(\frac{1+\epsilon}{1-\epsilon}\right) \left[1 + \binom{N-1}{M}\right] \frac{\log T}{\mu_{\min}^2 (1-\beta)}$.*

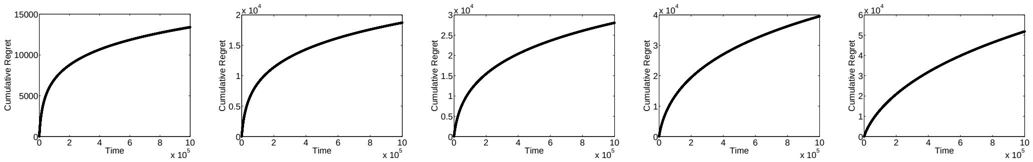
Discussion: This regret bound is of the order of $\binom{N-1}{M} \frac{\log T}{\mu_{\min}^2}$, which is significantly smaller than the standard MAB bound of $|\mathcal{A}| \frac{\log T}{\mu_{\min}^2} = \binom{N}{M} \frac{\log T}{\mu_{\min}^2}$ by a multiplicative factor of $\frac{\binom{N-1}{M}}{\binom{N}{M}} = \frac{N-M}{N}$, or by an additive factor of $\binom{N-1}{M-1} \frac{\log T}{\mu_{\min}^2}$. In fact, though this is a provable reduction in the regret scaling, the actual reduction is likely to be much better in practice – the experimental results in Section 3 attest to this. The proof of the corollary uses sharp combinatorial estimates relating to vertices on the N-dimensional hypercube.

3 Regret Performance: Numerical Results

We evaluate the performance of Thompson sampling (Algorithm 1) on two complex bandit settings – (a) Playing subsets of arms with the MAX reward function, and (b) Job scheduling over machines to minimize makespan. Where the posterior distribution is not closed-form, we use a particle filter approximation [8] for efficient posterior updates.

1. Subset Plays, MAX Reward: We assume the setup of Section 2.2 where one plays a size-M subset in each round and observes the maximum value. The individual arms' reward parameters are taken to be equi-spaced in $(0, 1)$. It is observed that Thompson sampling outperforms standard "decoupled" UCB by a wide margin in the cases we consider (Figure 1, left and center). The differences are especially pronounced for the larger problem size $N = 1000, M = 3$, where UCB, that sees $\binom{N}{M}$ separate actions, appears be in the exploratory phase throughout. Figure 2 affords a closer look at the regret for the above problem, and presents the results of using a flat prior over a uniformly discretized grid of models in $[0, 1]^{10}$ – the setting of Theorem 1.

2. Subset Plays, Average Reward: We apply Thompson sampling again to the problem of choosing the best M out of N basic arms of a Bernoulli bandit, but this time receiving a reward that is the *average value* of the chosen subset. This specific form of the feedback makes it possible to use a *continuous, Gaussian prior* density over the space of basic parameters that is updated to a Gaussian posterior assuming a fictitious Gaussian likelihood model [10]. This is a fast, practical alternative to UCB-style deterministic methods [7, 12] which require performing a convex optimization every instant. Figure 3 shows the regret of Thompson sampling with a Gaussian prior/posterior for choosing various size M subsets $(3, 5, 10, 20, 50)$ out of $N = 100$ arms. It is practically impossible to naively apply a decoupled bandit algorithm over such a problem due to the very large number of complex actions (e.g., there are $\approx 10^{13}$ actions even for $M = 10$) [8]. However, Thompson sampling merely samples from a $N = 100$ dimensional Gaussian and picks the best M coordinates of the sample, which yields a dramatic reduction in running time. (Note that the constant factors in the regret curves are seen to be modest when compared to the total number of complex actions.



(a) $(100, 3)$ (b) $(100, 5)$ (c) $(100, 10)$ (d) $(100, 20)$ (e) $(100, 50)$

Figure 3: Cumulative regret for (N, M): Observing the average value of M out of N arms for Thompson sampling. The prior is a standard normal independent density over N dimensions, and the posterior is also normal under a Gaussian likelihood model. The regret is averaged across 10 runs. Confidence intervals are ± 1 standard deviation.

3. Job Scheduling: We consider a stochastic job-scheduling problem in order to illustrate the versatility of Thompson sampling for bandit settings more complicated than subset actions. There are $N = 10$ types of jobs and 2 machines. Every job type has a different, unknown mean duration, with the job means taken to be equally spaced in $[0, N]$, i.e., $\frac{iN}{N+1}$, $i = 1, \ldots, N$. At each round, one job of each type arrives to the scheduler, with a random duration that follows the exponential distribution with the corresponding mean. All jobs must be scheduled on one of two possible machines. The loss suffered upon scheduling is the *makespan*, i.e., the maximum of the two job durations on the machines. Once the jobs in a round are assigned to the machines, only the *total* durations on the machines machines can be observed. Figure 1 (right) shows the results of applying Thompson sampling with an exponential prior for the jobs' means along with a particle filter.

References

[1] J. C. Gittins, K. D. Glazebrook, and R. R. Weber, *Multi-Armed Bandit Allocation Indices*. Wiley, 2011.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," *Journal of Machine Learning Research - Proceedings Track*, vol. 19, pp. 359–376, 2011.

[4] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 24, no. 3–4, pp. 285–294, 1933.

[5] S. Scott, "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol. 26, pp. 639–658, 2010.

[6] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem.," *Journal of Machine Learning Research - Proceedings Track*, vol. 23, pp. 39.1–39.26, 2012.

[7] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *COLT*, pp. 355–366, 2008.

[8] A. Doucet, N. D. Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[9] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Proceedings of the Twenty-third International Conference on Algorithmic Learning Theory*, 2012.

[10] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

[11] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," tech. rep., Technion, May 2013. http://webee.technion.ac.il/people/aditya/ts-complexbandits-techreport.pdf.

[12] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

[8] Though existing algorithms [7, 12] account for linear feedback across coupled actions via tight confidence sets, they require (at least as stated) a computationally expensive search over all complex actions.