Semiparametric Contextual Bandits

Akshay Krishnamurthy [1] **Zhiwei Steven Wu** [1] **Vasilis Syrgkanis** [2]

Abstract

This paper studies *semiparametric contextual bandits*, a generalization of the linear stochastic bandit problem where the reward for an action is modeled as a linear function of known action features confounded by a non-linear action-independent term. We design new algorithms that achieve $\tilde{O}(d\sqrt{T})$ regret over T rounds, when the linear function is d-dimensional, which matches the best known bounds for the simpler unconfounded case and improves on a recent result of Greenewald et al. (2017). Via an empirical evaluation, we show that our algorithms outperform prior approaches when there are non-linear confounding effects on the rewards. Technically, our algorithms use a new reward estimator inspired by doubly-robust approaches and our proofs require new concentration inequalities for self-normalized martingales.

1. Introduction

A number of applications including online personalization, mobile health, and adaptive clinical trials require that an agent repeatedly makes decisions based on user or patient information with the goal of optimizing some metric, typically referred to as a reward. For example, in online personalization problems, we might serve content based on user history and demographic information with the goal of maximizing user engagement with our service. Since counterfactual information is typically not available, these problems require algorithms to carefully balance *exploration*—making potentially suboptimal decisions to acquire new information—with *exploitation*—using collected information to make better decisions. Such problems are often best modeled with the framework of *contextual bandits*, which captures the exploration-exploitation tradeoff and enables rich decision making policies but ignores the long-term temporal effects that make general reinforcement learning challenging. Contextual bandit algorithms have seen recent success in applications, including news recommendation (Li et al., 2010) and mobile health (Tewari & Murphy, 2017).

Contextual bandit algorithms can be categorized as either *parametric* or *agnostic*, depending on whether they model the relationship between the reward and the decision or not. Parametric approaches typically assume that the reward is a (generalized) linear function of a known decision-specific feature vector (Filippi et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013). Once this function is known to high accuracy, it can be used to make near-optimal decisions. Exploiting this fact, algorithms for this setting focus on learning the parametric model. Unfortunately, fully parametric assumptions are often unrealistic and challenging to verify in practice, and these algorithms may perform poorly when the assumptions do not hold.

In contrast, agnostic approaches make no modeling assumptions about the reward and instead compete with a large class of decision-making policies (Langford & Zhang, 2008; Agarwal et al., 2014). While these policies are typically parametrized in some way, these algorithms provably succeed under weaker conditions and are generally more robust than parametric ones. On the other hand, they typically have worse statistical guarantees, are conceptually much more complex, and have high computational overhead, technically requiring solving optimization problems that are NP-hard in the worst case. This leads us to a natural question:

> *Is there an algorithm that inherits the simplicity and statistical guarantees of the parametric methods* and *the robustness of the agnostic ones?*

Working towards an affirmative answer to this question, we consider a semiparametric contextual bandit setup where the reward is modeled as a linear function of the decision confounded by an additive non-linear perturbation that is independent of the decision. This setup significantly generalizes the standard parametric one, allowing for complex, non-stationary, and non-linear rewards (See Section 2 for a precise formulation). On the other hand, since this perturbation is just a baseline reward for all decisions, it has no influence on the optimal one, which depends only on the unknown linear function. In the language of econometrics

[1] Microsoft Research, New York, New York [2] Microsoft Research, Cambridge, Massachusetts. Correspondence to: Akshay Krishnamurthy <akshay@cs.umass.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

and causal modeling, the *treatment effect* is linear.

In this paper, we design new algorithms for the semiparametric contextual bandits problem. When the linear part of the reward is d-dimensional, our algorithms achieve $\tilde{O}(d\sqrt{T})$ regret over T rounds, even when the features and the confounder are chosen by an adaptive adversary. This guarantee matches the best results for the simpler linear stochastic bandit problem up to logarithmic terms, showing that there is essentially no statistical price to pay for robustness to confounding effects. On the other hand, our algorithm and analysis is quite different, and it is not hard to see that existing algorithms for stochastic bandits fail in our more general setting. Our regret bound also improves on a recent result of Greenewald et al. (2017), who consider the same setup but study a weaker notion of regret. Our algorithm, main theorem, and comparisons are presented in Section 3.

We also compare our algorithm to approaches from both parametric and agnostic families in an empirical study (we use a linear policy class for agnostic approaches). In Section 5, we evaluate several algorithms on synthetic problems where the reward is (a) linear, and (b) linear with confounding. In the linear case, our approach learns, but is slightly worse than the baselines. On the other hand, when there is confounding, our algorithm significantly outperforms both parametric and agnostic approaches. As such, these experiments demonstrate that our algorithm represents a favorable trade off between statistical efficiency and robustness.

On a technical level, our algorithm and analysis require several new ideas. First, we derive a new estimator for linear models in the presence of confounders, based on recent and classical work in semiparametric statistics and econometrics (Robinson, 1988; Chernozhukov et al., 2016). Second, since standard algorithms using optimism principles fail to guarantee consistency of this new estimator, we design a new randomized algorithm, which can be viewed as an adaptation of the action-elimination method of Even-Dar et al. (2006) to the contextual bandits setting. Finally, analyzing the semiparametric estimator requires an intricate deviation argument, for which we derive a new self-normalized inequality for vector-valued martingales using tools from de la Peña et al. (2008; 2009).

2. Preliminaries

We study a generalization of the linear stochastic bandit problem with *action-dependent* features and *action-independent* confounder. The learning process proceeds for T rounds, and in round t, the learner receives a context $x_t \triangleq \{z_{t,a}\}_{a\in\mathcal{A}}$ where $z_{t,a} \in \mathbb{R}^d$ and $\mathcal{A}$ is the action set, which we assume to be large but finite. The learner then chooses an action $a_t \in \mathcal{A}$ and receives reward

$$r_t(a_t) \triangleq \langle \theta, z_{t,a_t}\rangle + f_t(x_t) + \xi_t, \qquad (1)$$

where $\theta \in \mathbb{R}^d$ is an unknown parameter vector, $f_t(x_t)$ is a confounding term that depends on the context x_t but, crucially, does not depend on the chosen action a_t, and ξ_t is a noise term that is centered and independent of a_t.

For each round t, let $a_t^\star \triangleq \mathrm{argmax}_{a\in\mathcal{A}}\langle\theta, z_{t,a}\rangle$ denote the optimal action for that round. The goal of our algorithm is to minimize the regret, defined as

$$\mathrm{Reg}(T) \triangleq \sum_{t=1}^{T} r_t(a_t^\star) - r_t(a_t) = \sum_{t=1}^{T}\langle\theta, z_{t,a_t^\star} - z_{t,a_t}\rangle.$$

Observe that the noise term ξ_t, and, more importantly, the confounding term $f_t(x_t)$ are absent in the final expression, since they are independent of the action choice.

We consider the challenging setting where the context x_t and the confounding term $f_t(\cdot)$ are chosen by an adaptive adversary, so they may depend on all information from previous rounds. This is formalized in the following assumption.

Assumption 1 (Environment). *We assume that $x_t = \{z_{t,a}\}_{a\in\mathcal{A}}, f_t, \xi_t$ are generated at the beginning of round t, before a_t is chosen. We assume that x_t and f_t are chosen by an adaptive adversary, and that ξ_t satisfies $\mathbb{E}[\xi_t|x_t, f_t] = 0$ and $|\xi_t| \le 1$.*

We also impose mild regularity assumptions on the parameter, the feature vectors, and the confounding functions.

Assumption 2 (Boundedness). *Assume that $\|\theta\|_2 \le 1$ and that $\|z_{t,a}\|_2 \le 1$ for all $a \in \mathcal{A}, t \in [T]$. Further assume that $f_t(\cdot) \in [-1, 1]$ for all $t \in [T]$.*

For simplicity, we assume an upper bound of 1 in these conditions, but our algorithm and analysis can be adapted to more generic regularity conditions.

Related work. Our setting is related to linear stochastic bandits and several variations that have been studied in recent years. Among these, the closest is the work of Greenewald et al. (2017) who consider the same setup and provide a Thompson Sampling algorithm using a new reward estimator that eliminates the confounding term. Motivated by applications in medical intervention, they consider a different notion of regret from our more-standard notion and, as such, the results are somewhat incomparable. For our notion of regret, their analysis can produce a $T^{2/3}$-style regret bound, which is worse than our optimal $\sqrt{T}$ bound. See Section 3.3 for a more detailed comparison.

Other results for linear stochastic bandits include upper-confidence bound algorithms (Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), Thompson sampling algorithms (Agrawal & Goyal, 2013; Russo & Van Roy, 2014), and extensions to generalized linear models (Filippi et al., 2010; Li et al., 2017). However, none of these models accommodate arbitrary and nonlinear confounding effects. Moreover, apart from Thompson sampling, all of these algorithms use deterministic action-selection policies (conditioning on the history), which provably incurs $\Omega(T)$ regret in our setting, as we will see.

One can accommodate confounded rewards via an agnostic-learning approach to contextual bandits (Auer et al., 2002; Langford & Zhang, 2008; Agarwal et al., 2014). In this framework, we make no assumptions about the reward, but rather compete with a class of parameterized policies (or experts). Since a d-dimensional linear policy is optimal in our setting, an agnostic algorithm with a linear policy class addresses precisely our notion of regret. However there are two disadvantages. First, agnostic algorithms are all computationally intractable, either because they enumerate the (infinitely large) policy class, or because they assume access to optimization oracles that can solve NP-hard problems in the worst case. Second, most agnostic approaches have regret bounds that grow with $\sqrt{K}$, the number of actions, while our bound is completely independent of K.

We are aware of one approach that is independent of K, but it requires enumeration of an infinitely large policy class. This method is based on ideas from the adversarial linear and combinatorial bandits literature (Dani et al., 2008; Abernethy et al., 2008; Bubeck et al., 2012; Cesa-Bianchi & Lugosi, 2012). Writing $\theta_t \triangleq (\theta, f_t(x_t)) \in \mathbb{R}^{d+1}$ and $z'_{t,a} \triangleq (z_{t,a}, 1) \in \mathbb{R}^{d+1}$, our setting can be re-formulated in the adversarial linear bandits framework. However, standard linear bandit algorithms compete with the best fixed action vector in hindsight, rather than the best policy with time-varying action sets. To resolve this, one can use the linear bandits reward estimator (Swaminathan et al., 2017) in a contextual bandit algorithm like EXP4 (Auer et al., 2002), but this approach is not computationally tractable with the linear policy class. For our setting, we are not aware of any computationally efficient approaches, even oracle-based approaches, that achieve $\text{poly}(d)\sqrt{T}$ regret with no dependence on the number of actions.

We resolve the challenge of confounded rewards with an estimator from the semiparametric statistics literature (Tsiatis, 2007), which focuses on estimating functionals of a nonparametric model. Most estimators are based on *Neyman Orthogonalization* (Neyman, 1979), which uses moment equations that are insensitive to nuisance parameters in a method-of-moments approach (Chernozhukov et al., 2016). These orthogonal moments typically involve a linear correction to an initial nonparametric estimate using so-called *influence functions* (Bickel et al., 1998; Robins et al., 2008). Robinson (1988) used this approach for the offline version of our setting (known as the partially linear regression (PLR) model) where he demonstrated a form of *double-robustness* (Robins & Rotnitzky, 1992) to poor estimation of the nuisance term (in our case $f_t(x_t)$). We generalize Robinson's work to the online setting, showing how orthogonalized estimators can be used for adaptive exploration. This requires several new techniques, including a novel action selection policy and a self-normalized inequality for vector-valued martingales.

3. Algorithm and Results

In this section, we describe our algorithm and present our main theoretical result, an $\tilde{O}(d\sqrt{T})$ regret bound for the semiparametric contextual bandits problem.

3.1. A Lower Bound

Before turning to the algorithm, we first present a lower bound against deterministic algorithms. Since the functions f_t may be chosen by an adaptive adversary, it is not hard to show that this setup immediately precludes the use of deterministic algorithms.

Proposition 3. *Consider an algorithm that, at round t, chooses an action a_t as a deterministic function of the observable history $H_t \triangleq \{x_{1:t}, a_{1:t-1}, r_{1:t-1}\}$. There exists a semiparametric contextual bandit instance with $d = 2$ and $K = 2$ where the regret of the algorithm is at least $T/2$.*

See Appendix B for the proof, which resembles the standard argument against deterministic online learning algorithms (Cover, 1965). The main difference is that the adversary uses the confounding term to corrupt the information that the learner receives, whereas, in the standard proof, the adversary chooses the optimal action in response to the learner. In fact, deterministic algorithms can succeed in the full information version of our setting, since taking differences between rewards eliminates the confounder. Thus, bandit feedback plays a crucial role in our construction and the bandit setting is considerably more challenging than the full information analogue.

We emphasize that, except for the Thompson Sampling approach (Agrawal & Goyal, 2013), essentially all algorithms for the linear stochastic bandit problem use deterministic strategies, so they provably fail in the semiparametric setting. As we mentioned, Thompson Sampling was analyzed in our setting by Greenewald et al. (2017), but they do not obtain the optimal $\sqrt{T}$-type regret bound (See Section 3.3 for a more quantitative and detailed comparison). In contrast, our algorithm is quite different from all of these approaches; it ensures enough randomization to circumvent the lower bound and also achieves the optimal $\sqrt{T}$ regret.

To conclude this discussion, we remark that the $\Omega(d\sqrt{T})$ lower bound for linear stochastic bandits (Dani et al., 2008), which also applies to randomized algorithms, holds in our more general setting as well.

3.2. The Algorithm

Pseudocode for the algorithm, which we call BOSE, for "Bandit Orthogonalized Semiparametric Estimation," is displayed in Algorithm 1. The algorithm maintains an estimate $\hat{\theta}$ for the true parameter θ, which it uses in each round to select an action via two steps: (1) an action elimination step that removes suboptimal actions, and (2) an optimization step that finds a good distribution over the surviving actions. The algorithm then samples and plays an action from this distribution and uses the observed reward to update the parameter estimate $\hat{\theta}$. This parameter estimation step is the third main element of the algorithm. We now describe each of these three components in detail.

Parameter estimation. For simplicity, we use $z_t \triangleq z_{t,a_t}$ to denote the feature vector for the action that was chosen at round t, and similarly we use $r_t \triangleq r_t(a_t)$. Using all previously collected data, specifically $\{z_\tau, r_\tau\}_{\tau=1}^t$ at the end of round t, we would like to estimate the parameter θ. First, if $f_\tau(x_\tau)$ were identically zero, by exploiting the linear parametrization we could use ridge regression, which with some $\lambda > 0$ gives

$$\hat{\theta}_{\text{Ridge}} \triangleq \left(\lambda I + \sum_{\tau=1}^t z_\tau z_\tau^\top\right)^{-1} \sum_{\tau=1}^t z_\tau r_\tau.$$

This estimator appears in most prior approaches for linear stochastic bandits (Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Unfortunately, since $f_\tau(x_\tau)$ is non-zero, $\hat{\theta}_{\text{Ridge}}$ has non-trivial and non-vanishing bias, so even in benign settings it is not a consistent estimator for θ.[1]

Our approach to eliminating the bias from the confounding term $f_\tau(x_\tau)$ is to center the feature vectors z_τ. Intuitively, in the ridge estimator, if z_τ is centered, then $z_\tau(r_\tau - \langle \theta^\star, z_\tau\rangle)$ is mean zero, even when there is non-negligible bias in the second term. As such, the error of the corresponding estimator can be expected to concentrate around zero. In the semiparametric statistics literature, this is known as *Neyman Orthogonalization* (Neyman, 1979), which was analyzed in the context of linear regression by Robinson (1988) and in a more general setting by Chernozhukov et al. (2016).

To center the feature vector, we will, at round t, choose action a_t by sampling from some distribution $\pi_t \in \Delta(\mathcal{A})$. Let $\mu_t \triangleq \mathbb{E}_{a_t\sim\pi_t}[z_{t,a_t}|x_t]$ denote the mean feature vector, taking expectation only over our random action choice. With this notation, the orthogonalized estimator is

$$\Gamma = \lambda I + \sum_{\tau=1}^t (z_\tau - \mu_\tau)(z_\tau - \mu_\tau)^\top,$$
$$\hat{\theta} = \Gamma^{-1}\sum_{\tau=1}^t (z_\tau - \mu_\tau) r_\tau.$$

$\hat{\theta}$ is a Ridge regression version of Robinson's classical semiparametric regression estimator (Robinson, 1988). The estimator was originally derived for observational studies where one might not know the *propensities* μ_τ exactly, and the standard description involves estimates $\hat{f}_\tau$ and $\hat{\mu}_\tau$ for the confounding term f_τ and the propensities μ_τ respectively. Informally, the estimator achieves a form of double-robustness, in the sense that it is accurate if either of these auxilliary estimators are. In our case, since we know the propensities μ_τ exactly, we can use an inconsistent estimator for the confounding term, so we simply set $\hat{f}_\tau(x_\tau) \equiv 0$. In Lemma 5, we prove a precise finite sample concentration inequality for this orthogonalized estimator, showing that the confounding term $f_t(x_t)$ does not introduce any bias. While the estimator has been studied in prior works (Robinson, 1988), to our knowledge, our error guarantee is novel.

The convergence rate of the orthogonalized estimator depends on the eigenvalues of the matrix Γ, and we must carefully select actions to ensure these eigenvalues are sufficiently large. To see why, notice that any deterministic action-selection approach with the orthogonalized estimator (including confidence based approaches), will fail, since $z_t = \mu_t$, so the eigenvalues of Γ do not grow rapidly and in fact the estimator is identically 0. This argument motivates our new action selection scheme which ensure substantial conditional covariance.

Action selection. Our action selection procedure has two main elements. First using our estimate $\hat{\theta}$, we eliminate any action that is provably suboptimal. Based on our analysis for the estimator $\hat{\theta}$, at round t, we can certify action a is suboptimal, if we can find another action b such that

$$\langle \hat{\theta}, z_{t,b} - z_{t,a}\rangle > \gamma(T)\|z_{t,b} - z_{t,a}\|_{\Gamma^{-1}}.$$

Here $\gamma(T)$ is the constant specified in the algorithm, and $\|x\|_M \triangleq \sqrt{x^\top M x}$ denotes the Mahalanobis norm. Using our confidence bound for $\hat{\theta}$ in Lemma 5 below, this inequality certifies that action b has higher expected reward than action a, so we can safely eliminate a from consideration.

The next component is to find a distribution over the surviving actions, denoted $\mathcal{A}'_t$ at round t, with sufficient covariance. The distribution $\pi_t \in \Delta(\mathcal{A}'_t)$ that we use is the

[1] A related estimator *can* be used to evaluate the reward of a policy, as in linear and combinatorial bandits (Cesa-Bianchi & Lugosi, 2012), but to achieve adequate exploration, one must operate over the policy class, which leads to computational intractability. We would like to use $\hat{\theta}$ to drive exploration, and this seems to require a consistent estimator. See Appendix A for a simple example demonstrating how using a biased estimator in a confidence-based approach results in linear regret.

Algorithm 1: BOSE (Bandit orthogonalized semiparametric estimation)

```
Input: $T, \delta \in (0,1)$.
Set $\lambda \leftarrow 4d\log(9T) + 8\log(4T/\delta)$ and $\gamma(T) \leftarrow \sqrt{\lambda} + \sqrt{27d\log(1+2T/d) + 54\log(4T/\delta)}$.
Initialize $\hat{\theta} \leftarrow 0 \in \mathbb{R}^d, \Gamma \leftarrow \lambda I_{d\times d}$.
for $t = 1, \ldots, T$ do
    Observe $x_t = \{z_{t,a}\}_{a\in\mathcal{A}}$
    Filter
        $$\mathcal{A}_t \leftarrow \left\{ a \in \mathcal{A} \mid \forall b \in \mathcal{A}, \langle\hat{\theta}, z_{t,b} - z_{t,a}\rangle \le \gamma(T)\|z_{t,a} - z_{t,b}\|_{\Gamma^{-1}} \right\}. \quad (2)$$
    Find distribution $\pi_t \in \Delta(\mathcal{A}_t)$ such that $\forall a \in \mathcal{A}_t$ (We use $\mathrm{Cov}_{b\sim\pi_t}(z_{t,b}) \triangleq \mathbb{E}[z_{t,b}z_{t,b}^\top] - (\mathbb{E}z_{t,b})(\mathbb{E}z_{t,b})^\top$.)
        $$\|z_{t,a} - \mathbb{E}_{b\sim\pi_t} z_{t,b}\|^2_{\Gamma^{-1}} \le \mathrm{tr}(\Gamma^{-1} \underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b})). \quad (3)$$
    Sample $a_t \sim \pi_t$ and play $a_t$. Observe $r_t(a_t)$. ($r_t(a_t) = \langle\theta, z_{t,a_t}\rangle + f_t(x_t) + \xi_t$.)
    Let $\mu_t = \mathbb{E}_{a\sim\pi_t}[z_{t,a} \mid x_t]$ and update parameters
        $$\Gamma \leftarrow \Gamma + (z_{t,a_t} - \mu_t)(z_{t,a_t} - \mu_t)^\top, \qquad \hat{\theta} \leftarrow \Gamma^{-1}\sum_{\tau=1}^{t}(z_{\tau,a_\tau} - \mu_\tau)r_\tau(a_\tau). \quad (4)$$
```

solution to the following feasibility problem

$$\forall a \in \mathcal{A}'_t, \quad \|z_{t,a} - \mathbb{E}_{b\sim\pi_t} z_{t,b}\|^2_{\Gamma^{-1}} \le \mathrm{tr}(\Gamma^{-1} \underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b})).$$

For intuition, the left hand side of the constraint for action a is an upper bound on the expected regret if a is the optimal action on this round. Thus, the constraints ensure that the regret is related to the covariance of the distribution, which means that if we incur high regret, the covariance term $\mathrm{Cov}_{b\sim\pi_t}(z_{t,b})$ will be large. Since we use a sample from π_t to update our parameter estimate, this means that whenever the instantaneous regret is large, we must learn substantially about the parameter. In this way, the distribution π_t balances exploration and exploitation. We will see in Lemma 8 that this program is convex and always has a feasible solution.

Our action selection scheme bears some resemblance to action-elimination approaches that have been studied in various bandit settings (Even-Dar et al., 2006). The main differences are that we adapt these ideas to the contextual setting and carefully choose a distribution over the surviving actions to balance exploration and exploitation.

3.3. The Main Result

We now turn to the main result, a regret guarantee for BOSE.

Theorem 4. *Consider the semiparametric contextual bandit problem under Assumption 1 and Assumption 2. For any parameter $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 has regret at most $O(d\sqrt{T}\log(T/\delta))$.*

The constants, and indeed a bound depending on λ and $\gamma(T)$ can be extracted from the proof, provided in the appendix. To interpret the regret bound, it is worth comparing with several related results:

Comparison with linear stochastic bandits. While most algorithms for linear stochastic bandits provably fail in our setting (via Proposition 3), the best regret bounds here are $O(\sqrt{dT\log(TK/\delta)})$ (Chu et al., 2011) and $O(d\sqrt{T}\log(T) + \sqrt{dT\log(T)\log(1/\delta)})$ (Abbasi-Yadkori et al., 2011) depending on whether we assume that the number of actions K is small or not. This latter result is optimal when the number of actions is large (Dani et al., 2008), which is the setting we are considering here. Since our bound matches this optimal regret up to logarithmic factors, and since linear stochastic bandits are a special case of our semiparametric setting, our result is therefore also optimal up to logarithmic factors. An interesting open question is whether an $\tilde{O}(\sqrt{dT\log(K/\delta)})$ regret bound is achievable in the semiparametric setting.

Comparison with agnostic contextual bandits. The best oracle-based agnostic approaches achieve $\tilde{O}(\sqrt{dKT})$ regret (Agarwal et al., 2014), incurring a polynomial dependence on the number of actions K, although there is one inefficient method that can achieve $\tilde{O}(d\sqrt{T})$,[2] as we discussed previously. To date, all efficient methods in the agnostic setting require some form of i.i.d. (Agarwal et al., 2014) or transductive assumption (Syrgkanis et al., 2016; Rakhlin & Sridharan, 2016) on the contexts, which we do not assume here.

Comparison with Greenewald et al. (2017). Greenewald et al. (2017) consider a very similar setting to ours, where rewards are linear with confounding, but where one default action a_0 always has $z_{t,a_0} \equiv 0 \in \mathbb{R}^d$. Applications in mobile health motivate a restriction that the algorithm

[2] This follows easily by combining ideas from Auer et al. (2002) and Cesa-Bianchi & Lugosi (2012).

choose the a_0 action with probability $\in [p, 1-p]$ for some small $p \in (0,1)$. Their work also introduces a new notion of regret where they compete with the policy that also satisfies this constraint but otherwise chooses the optimal action $a_t^\star$. In this setup, they obtain an $\tilde{O}(d^2\sqrt{T})$ regret bound, which has a worse dimension dependence than Theorem 4.

While the setup is somewhat different, we can still translate our result into a regret bound in their setting, since BOSE can support the probability constraint, and by coupling the randomness between BOSE and the optimal policy, the regret is unaffected.[3] On the other hand, since the constant in their regret bound scales with $1/p$, their results as stated are vacuous when $p = 0$ which is precisely our setting. For our more challenging regret definition, their analysis can produce a suboptimal $T^{2/3}$-style regret bound, and in this sense, Theorem 4 provides a quantitative improvement.

Summary. BOSE achieves essentially the same regret bound as the best linear stochastic bandit methods, but in a much more general setting. On the other hand, the agnostic methods succeed under even weaker assumptions, but have worse regret guarantees and/or are computationally intractable. Thus, BOSE broadens the scope for computationally efficient contextual bandit learning.

4. Proof Sketch

We sketch the proof of Theorem 4 in the two-action case ($|\mathcal{A}| = 2$), which has a much simpler proof that preserves the main ideas. The technical machinery needed for the general case is much more sophisticated, and we briefly describe some of these steps at the end of this section, with a complete proof in the Appendix.

In the two arm case, one should set $\gamma(T) \triangleq \sqrt{\lambda} + \sqrt{9d\log(1+T/(d\lambda)) + 18\log(T/\delta)}$ and $\lambda = O(1)$, which differs slightly from the algorithm pseudocode for the more general case. Additionally, note that with two actions, the uniform distribution over $\mathcal{A}_t$ is always feasible for Problem (3). Specifically, if the filtered set has cardinality 1, we simply play that action deterministically, otherwise we play one of the two actions uniformly at random.

The proof has three main steps. First we analyze the orthogonalized regression estimator defined in (4). Second, we study the action selection mechanism and relate the regret incurred to the error bound for the orthogonalized estimator. Finally, using a somewhat standard potential argument, we show how this leads to a $\sqrt{T}$-type regret bound. For the proof, let $\hat{\theta}_t, \Gamma_t$ be the estimator and covariance matrix used on round t, both based on $t-1$ samples.

For the estimator, we prove the following lemma for the two action case. The main technical ingredient is a self-normalized inequality for vector-valued martingales, which can be obtained using ideas from de la Peña et al. (2009).

Lemma 5. *Under Assumption 1 and Assumption 2, let $K = 2$ and $\gamma(T) \triangleq \sqrt{\lambda} + \sqrt{9d\log(1+T/(d\lambda)) + 18\log(T/\delta)}$. Then, with probability at least $1-\delta$, the following holds simultaneously for all $t \in [T]$:*

$$\|\hat{\theta}_t - \theta\|_{\Gamma_t} \le \gamma(T).$$

Proof. Using the definitions and Assumption 1, it is not hard to re-write

$$\hat{\theta}_t = \Gamma_t^{-1}(\Gamma_t - \lambda I)\theta + \Gamma_t^{-1}\sum_{\tau=1}^{t-1} Z_\tau \zeta_\tau,$$

where $Z_\tau \triangleq z_{\tau,a_\tau} - \mu_\tau$ and $\zeta_\tau \triangleq \langle \theta, \mu_\tau\rangle + f_\tau(x_\tau) + \xi_\tau$. Further define $S_t \triangleq \sum_{\tau=1}^{t-1} Z_\tau\zeta_\tau$. Then, applying the triangle inequality the error is at most

$$\|\hat{\theta}_t - \theta\|_{\Gamma_t} \le \|\lambda\theta\|_{\Gamma_t^{-1}} + \|S_t\|_{\Gamma_t^{-1}}.$$

The first term here is at most $\sqrt{\lambda}$ since $\Gamma_t \succeq \lambda I$. To control the second term, we need to use a self-normalized concentration inequality, since Z_τ is a random variable, and the normalizing term $\Gamma_t = \lambda I + \sum_{\tau=1}^{t-1} Z_\tau Z_\tau^\top$ depends on the random realizations. In Lemma 10 in the appendix, we prove that with probability at least $1-\delta$, for all $t \in [T]$

$$\|S_t\|_{\Gamma_t^{-1}}^2 \le 9d\log(1+T/(d\lambda)) + 18\log(T/\delta). \qquad (5)$$

The lemma follows from straightforward calculations. □

Before proceeding, it is worth commenting on the difference between our self-normalized inequality (5) and a slightly different one used by Abbasi-Yadkori et al. (2011) for the linear case. In their setup, they have that ζ_τ is conditionally centered and sub-Gaussian, which simplifies the argument since after fixing the Z_τs (and hence Γ_t), the randomness in the ζ_τs suffices to provide concentration. In our case, we must use the randomness in Z_τ itself, which is more delicate, since Z_τ affects the numerator S_t, but also the normalizer Γ_t. In spite of this additional technical challenge, the two self-normalized processes admit similar bounds.

Next, we turn to the action selection step, where recall that either a single action is played deterministically, or the actions are played uniformly at random.

Lemma 6. *Let $\mu_t \triangleq \mathbb{E}_{a\sim\pi_t} z_{t,a}$ where π_t is the solution to (3), and assume that the conclusion of Lemma 5 holds. Then with probability at least $1-\delta$*

$$\textit{Reg}(T) \le \sqrt{2T\log(1/\delta)} + 2\gamma(T)\sum_{t=1}^{T}\sqrt{\operatorname{tr}(\Gamma_t^{-1}\operatorname*{Cov}_{b\sim\pi_t}(z_{t,b}))}.$$

[3] Technically it is actually smaller by a factor of $(1-p)$.

Proof. We first study the instantaneous regret, taking expectation over the random action. For this, we must consider two cases. First, with Lemma 5, if $|\mathcal{A}_t| = 1$, we argue that the regret is actually zero. This follows from the Cauchy-Schwarz inequality since assuming $\mathcal{A}_t = \{a\}$ we get

$$\langle \theta, z_{t,a} - z_{t,b}\rangle \ge \langle \hat{\theta}_t, z_{t,a} - z_{t,b}\rangle - \gamma(T)\|z_{t,a} - z_{t,b}\|_{\Gamma_t^{-1}}$$

which is non-negative using the fact that b was eliminated. Therefore a is the optimal action and we incur no regret. Since π_t has no covariance, the upper bound holds.

On the other rounds, we set $\pi_t = \mathrm{Unif}(\{a, b\})$ and hence $\mu_t = (z_{t,a} + z_{t,b})/2$. Assuming again that a is the optimal action, the expected regret is

$$\begin{aligned}
&\langle \theta, z_{t,a} - \mu_t\rangle = \frac{1}{2}\langle \theta, z_{t,a} - z_{t,b}\rangle \\
&\le \frac{1}{2}\left(\langle \hat{\theta}_t, z_{t,a} - z_{t,b}\rangle + \gamma(T)\|z_{t,a} - z_{t,b}\|_{\Gamma_t^{-1}}\right) \\
&\le \gamma(T)\|z_{t,a} - z_{t,b}\|_{\Gamma_t^{-1}} \le 2\gamma(T)\sqrt{\mathrm{tr}(\Gamma_t^{-1} \underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b}))}.
\end{aligned}$$

Here the first inequality uses Cauchy-Schwarz, the second uses (2), since neither action was eliminated, and the third uses (3). This bounds the expected regret, and the lemma follows by Azuma's inequality. □

The last step of the proof is to control the sequence

$$\sum_{t=1}^{T}\sqrt{\mathrm{tr}(\Gamma_t^{-1} \underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b}))}.$$

First, recall that

$$\underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b}) \triangleq \mathbb{E}_{b\sim\pi_t}\left[(z_{t,b} - \mu_t)(z_{t,b} - \mu_t)^\top\right]$$

with $\mu_t \triangleq \mathbb{E}_{b\sim\pi_t}[z_{t,b}]$. Since in the two-arm case π_t either chooses an arm deterministically or uniformly randomizes between the two arms, the following always holds:

$$\underset{b\sim\pi_t}{\mathrm{Cov}}(z_{t,b}) = (z_{t,a_t} - \mu_t)(z_{t,a_t} - \mu_t)^\top.$$

It follows that $\Gamma_{t+1} \triangleq \Gamma_t + \mathrm{Cov}_{b\sim\pi_t}(z_{t,b})$, and with $\Gamma_1 \triangleq \lambda I$, the standard potential argument for online ridge regression applies. We state the conclusion here, and provide a complete proof in the appendix.

Lemma 7. *Let Γ_t, π_t be defined as above and define $M_t \triangleq (z_{t,a_t} - \mu_t)(z_{t,a_t} - \mu_t)^\top$. Then*

$$\sum_{t=1}^{T}\sqrt{\mathrm{tr}(\Gamma_t^{-1}M_t)} \le \sqrt{dT(1 + 1/\lambda)\log(1 + T/(d\lambda))}.$$

Combining the three lemmas establishes a regret bound of

$$\mathrm{Reg}(T) \le O\left(\sqrt{Td\log(T/\delta)\log(T/d)} + d\sqrt{T}\log(T/d)\right)$$

with probability at least $1 - \delta$ in the two-action case.

Extending to many actions. Several more technical steps are required for the general setting. First, the martingale inequality used in Lemma 5 requires that the random vectors are symmetric about the origin. This is only true for the two-action case, and in fact a similar inequality does not hold in general for the non-symmetric situation that arises with more actions. In the non-symmetric case, both the empirical and the population covariance must be used in the normalization, so the analogue of (5) is instead

$$\|S_t\|^2_{(\Gamma_t + \mathbb{E}\Gamma_t)^{-1}} \le 27d\log(1 + 2T/d) + 54\log(4T/\delta).$$

On the other hand, the error term for our estimator depends only on the empirical covariance Γ_t. To correct for the discrepancy, we use a covering argument[4] to establish

$$\lambda I + \Gamma_t \succeq (\lambda - 6d\log(T/\delta))I + (\Gamma_t + \mathbb{E}\Gamma_t)/3.$$

With this semidefinite inequality, we can translate from the Mahalanobis norm in the weaker self-normalized bound to one with just Γ_t, which controls the error for the estimator.

We also argue that problem (3) is always feasible, which is the contents of the following lemma.

Lemma 8. *Problem (3) is convex and always has a feasible solution. Specifically, for any vectors $z_1, \ldots, z_n \in \mathbb{R}^d$ and any positive definite matrix M, there exists a distribution $w \in \Delta([n])$ with mean $\mu_w \triangleq \mathbb{E}_{b\sim w}z_b$ such that*

$$\forall i \in [n], \|z_i - \mu_w\|_M^2 \le \mathrm{tr}(M \underset{b\sim w}{\mathrm{Cov}}(z_b)).$$

The proof uses convex duality. Integrating these new arguments into the proof for the two-action case leads to Theorem 4.

5. Experiments

We conduct a simple experiment to compare BOSE with several other approaches[5]. We simulate three different environments that follow the semiparametric contextual bandits model with $d = 10$, $K = 2$. In the first setting the reward is linear and the action features are drawn uniformly from the unit sphere. In the latter two settings, we set $f_t(x_t) = -\max_a\langle\theta, z_{t,a}\rangle$, which is related to the construction in the proof of Proposition 3. One of these semiparametric settings has action features sampled from the unit sphere, while for the other, we sample from the intersection of the unit sphere and the positive orthant.

In Figure 1, we plot the performance of Algorithm 1 against four baseline algorithms: (1) OFUL: the optimistic algorithm for linear stochastic bandits (Abbasi-Yadkori et al.,

[4] For technical reasons, the Matrix Bernstein inequality does not suffice here since it introduces a dependence on the maximal variance. See Appendix for details.

[5] Our code is publicly available at `http://github.com/akshaykr/oracle_cb/`.

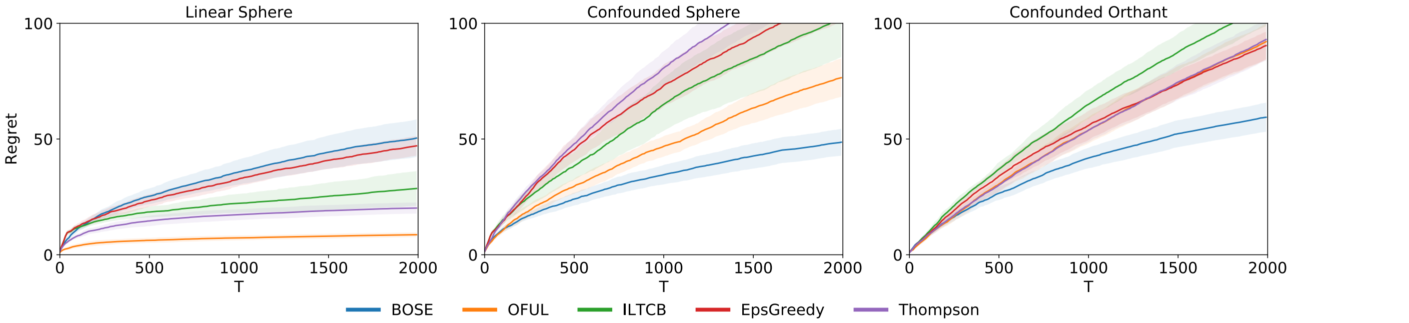

Figure 1. Synthetic experiments with $d = 10, K = 2$. Left: A linear environment where action-features are uniformly from the unit sphere. Center: A confounded environment with features from the sphere. Right: A confounded environment with features from the sphere intersected with the positive orthant. Algorithms are BOSE, OFUL (Abbasi-Yadkori et al., 2011), ILTCB (Agarwal et al., 2014), EPSGREEDY (Langford & Zhang, 2008), and THOMPSON (Agrawal & Goyal, 2013). Agnostic approaches use a linear policy class.

2011), (2) THOMPSON sampling for linear contextual bandits (Agrawal & Goyal, 2013), (3) EPSGREEDY: the ϵ-greedy approach (Langford & Zhang, 2008) with a linear policy class, (4) ILTCB: a more sophisticated agnostic algorithm (Agarwal et al., 2014) with linear policy class. The first algorithm is deterministic, so can have linear regret in our setting, but is the natural baseline and one we hope to improve. Thompson Sampling is another natural baseline, and a variant was used by Greenewald et al. (2017) in essentially the same setting as ours. The latter two have $(Kd)^{1/3}T^{2/3}$ and $\sqrt{KdT}$ regret bounds respectively under our assumptions, but require solving cost-sensitive classification problems, which are NP-hard in general. Following prior empirical evaluations (Krishnamurthy et al., 2016), we use a surrogate loss formulation based on square loss minimization in the implementation.

The results of the experiment are displayed in Figure 1, where we plot the cumulative regret against the number of rounds T. All algorithms have a single parameter that governs the degree of exploration. In BOSE and OFUL, this is the constant $\gamma(T)$ in the confidence bound, in THOMPSON it is the variance of the prior, and in ILTCB and EPSGREEDY it is the amount of uniform exploration performed by the algorithm. For each algorithm we perform 10 replicates for each of 20 values of the corresponding parameter, and we plot the best average performance, with error bars corresponding to ±2 standard deviations.

In the linear experiment (Figure 1, left panel), BOSE performs the worst, but is competitive with the agnostic approaches, demonstrating a price to pay for robustness. The experimental setup in the center panel is identical except with confounding, and BOSE is robust to this confounding, with essentially the same performance, while the three baselines degrade dramatically. Finally, when the features lie in the positive orthant (right panel), OFUL degrades further, while BOSE remains highly effective.

Regarding the baselines, we make two remarks:

1. Intuitively, the positive orthant setting is more challenging for OFUL since there is less inherent randomness in the environment to overcome the confounding effect.
2. The agnostic approaches, despite strong regret guarantees, perform somewhat poorly in our experiments, and we believe this for three reasons. First, our surrogate-loss implementation is based on an implicit realizability assumption, which is not satisfied here. Second, we expect that the constant factors in their regret bounds are significantly larger than those of BOSE or OFUL. For computational reasons, we only solve the optimization problem in ILTCB every 50 rounds, which causes a further constant factor increase in the regret.

Overall, while BOSE is worse than other approaches in the linear environment, the experiment demonstrates that when the environment is not perfectly linear, approaches based on realizability assumptions (either explicitly like in OFUL, or implicitly like in implementations of ILTCB and EPSGREEDY), can fail. We emphasize that linear environments are rare in practice, and such assumptions are typically impossible to verify. We therefore believe that trading off a small loss in performance in the specialized linear case for significantly more robustness, as BOSE demonstrates, is desirable.

6. Discussion

This paper studies a generalization of the linear stochastic bandits setting, where rewards are confounded by an adaptive adversary. Our new algorithm, BOSE, achieves the optimal regret, and also matches (up to logarithmic factors) the best algorithms for the linear case. Our empirical evaluation shows that BOSE offers significantly more robustness than prior approaches, and performs well in several environments.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Abernethy, J. D., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandit algorithms with supervised learning guarantees. In *International Conference Artificial Intelligence and Statistics*, 2011.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. *Efficient and adaptive estimation for semiparametric models*. Springer New York, 1998.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 2012.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. M. Double machine learning for treatment and causal parameters. *arXiv:1608.00060*, 2016.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Cover, T. M. Behavior of sequential predictors of binary sequences. In *Conference on Information Theory, Statistical Decision Functions and Random Processes*, 1965.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

de la Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and statistical applications*. Springer Science & Business Media, 2008.

de la Peña, V. H., Klass, M. J., and Lai, T. L. Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 2009.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, 2010.

Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability*, 1975.

Greenewald, K., Tewari, A., Murphy, S., and Klasnja, P. Action centered contextual bandits. In *Advances in Neural Information Processing Systems*, 2017.

Krishnamurthy, A., Agarwal, A., and Dudík, M. Contextual semibandits via supervised learning oracles. In *Advances in Neural Information Processing Systems*, 2016.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, 2008.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, 2010.

Li, L., Lu, Y., and Zhou, D. Provable optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, 2017.

Neyman, J. C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 1979.

Rakhlin, A. and Sridharan, K. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, 2016.

Robins, J. M. and Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*. Springer, 1992.

Robins, J. M., Li, L., Tchetgen Tchetgen, E., and van der Vaart, A. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, 2008.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 1988.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 2010.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 2014.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 1958.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, 2017.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. E. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, 2016.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*. Springer, 2017.

Tsiatis, A. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.