Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles

Dylan J. Foster [1] **Alexander Rakhlin** [1]

Abstract

A fundamental challenge in contextual bandits is to develop flexible, general-purpose algorithms with computational requirements no worse than classical supervised learning tasks such as classification and regression. Algorithms based on regression have shown promising empirical success, but theoretical guarantees have remained elusive except in special cases. We provide the first universal and optimal reduction from contextual bandits to *online* regression. We show how to transform any oracle for online regression with a given value function class into an algorithm for contextual bandits with the induced policy class, with no overhead in runtime or memory requirements. We characterize the minimax rates for contextual bandits with general, potentially nonparametric function classes, and show that our algorithm is minimax optimal whenever the oracle obtains the optimal rate for regression. Compared to previous results, our algorithm requires no distributional assumptions beyond realizability, and works even when contexts are chosen adversarially.

1. Introduction

We consider the design of practical, provably efficient algorithms for contextual bandits, where a learner repeatedly receives contexts and makes decisions on the fly so as to learn a policy that maximizes their total reward. Contextual bandits have been successfully applied in user recommendation systems (Agarwal et al., 2016) and mobile health applications (Tewari & Murphy, 2017), and in theory they are perhaps simplest reinforcement learning problem that embeds the full complexity of statistical learning with function approximation.

A key challenge in contextual bandits is to develop flexible, general purpose algorithms that work for arbitrary, user-specified classes of policies and come with strong theoretical guarantees on performance. Depending on the task, a user might wish to try decision trees, kernels, neural nets, and beyond to get the best performance. General-purpose contextual bandit algorithms ensure that the user doesn't have to design a new algorithm from scratch every time they encounter a new task.

Oracle-based algorithms constitute the dominant approach to general-purpose contextual bandits. Broadly, these algorithms seek to reduce the contextual bandit problem to basic supervised learning tasks such as classification and regression so that off-the-shelf algorithms can be applied. However, essentially all oracle-based contextual bandit algorithms suffer from one or more of the following issues:

1. Difficult-to-implement oracle.
2. Strong assumptions on hypothesis class or distribution.
3. High memory and runtime requirements.

Agnostic oracle-efficient algorithms (Langford & Zhang, 2008; Dudik et al., 2011; Agarwal et al., 2014) require few assumptions on the distribution, but reduce contextual bandits to *cost-sensitive classification*. Cost-sensitive classification is intractable even for simple hypothesis classes (Klivans & Sherstov, 2009), and in practice implementations are forced to resort to heuristics to implement the oracle (Agarwal et al., 2014; Krishnamurthy et al., 2016).

Foster et al. (2018) recently showed that a variant of the UCB algorithm for general function classes (Russo & Van Roy, 2014) can be made efficient in terms of calls to an oracle for *supervised regression*. Regression alleviates some of the practical issues with classification because it can be solved in closed form for simple classes and is amenable to gradient-based methods. Indeed, Foster et al. (2018) and Bietti et al. (2018) found that this algorithm typically outperformed algorithms based on classification oracles across a range of datasets. However, the theoretical analysis of the algorithm relies on strong distributional assumptions that are difficult to verify in practice, and it can indeed fail pathologically when these assumptions fail to hold.

All of the provably optimal general-purpose algorithms described above—both classification- and regression-based—are memory hungry: they keep the entire dataset in memory

[1] Massachusetts Institute of Technology. Correspondence to: Dylan Foster <dylanf@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

and repeatedly augment it before feeding it into the oracle. Even if the oracle itself is online in the sense that it admits streaming or incremental updates, the resulting algorithms do not have this property. At this point it suffices to say that—to our knowledge—no general-purpose algorithm with provably optimal regret has made it into a large-scale contextual bandit deployment in the real world (e.g., Agarwal et al. (2016)).

In this paper, we address issues (1), (2), and (3) simultaneously: We give a new contextual bandit algorithm which is efficient in terms of queries to an *online* oracle for *regression*, and which requires *no assumptions* on the data-generating process beyond a well-specified model.

1.1. Setup

We consider the following contextual bandit protocol, which occurs over T rounds. At each round $t \in [T]$, Nature selects a context $x_t \in \mathcal{X}$ and loss function $\ell_t : \mathcal{A} \to [0,1]$, where $\mathcal{A} = [K]$ is the learner's action space. The learner then selects an action $a_t \in \mathcal{A}$ and observes $\ell_t(a_t)$. We allow the contexts x_t to be chosen arbitrarily by an adaptive adversary, but we assume that each loss ℓ_t is drawn independently from a fixed distribution $\mathbb{P}_{\ell_t}(\cdot \mid x_t)$, where $\mathbb{P}_{\ell_1}, \ldots, \mathbb{P}_{\ell_T}$ are selected a-priori by an oblivious adversary.

We assume that the learner has access to a class of value functions $\mathcal{F} \subset (\mathcal{X} \times \mathcal{A} \to [0,1])$ (such as linear models or neural networks) that models the mean of the reward distribution. Specifically, we make the following standard *realizability* assumption (Chu et al., 2011; Agarwal et al., 2012; Foster et al., 2018).

Assumption 1 (Realizability)**.** There exists a regressor $f^\star \in \mathcal{F}$ such that for all t, $f^\star(x,a) = \mathbb{E}[\ell_t(a) \mid x_t = x]$.

The learner's goal is to compete with the class of policies induced by the model class $\mathcal{F}$. For each regression function $f \in \mathcal{F}$, we let $\pi_f(x) = \arg\min_{a \in \mathcal{A}} f(x,a)$ be the induced policy. Then aim of the learner is to minimize their *regret* to the optimal policy:

$$\mathrm{Reg}_{\mathsf{CB}}(T) = \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi^\star(x_t)), \qquad (1)$$

where $\pi^\star := \pi_{f^\star}$. Going forward, we let $\Pi = \{\pi_f \mid f \in \mathcal{F}\}$ denote the induced policy class.

1.2. Contributions

We introduce the notion of an *online regression oracle*. At each time t, an online regression oracle, which we denote SqAlg (for "square loss regression algorithm"), takes as input a tuple (x_t, a_t), produces a real-valued prediction $\widehat{y}_t \in \mathbb{R}$, and then receives the true outcome y_t. The goal of the oracle is to predict the outcomes as well as the best function in a class $\mathcal{F}$, in the sense that for every sequence of outcomes the *square loss regret* is bounded:

$$\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(x_t, a_t) - y_t)^2 \leq \mathrm{Reg}_{\mathsf{Sq}}(T). \quad (2)$$

Our main algorithm, SquareCB (Algorithm 1), is a reduction that efficiently and optimally turns any online regression oracle into an algorithm for contextual bandits in the realizable setting.

Theorem 1 (informal). *Suppose Assumption 1 holds. Then* SquareCB*, when invoked with an online regression oracle with square loss regret* $\mathrm{Reg}_{\mathsf{Sq}}(T)$*, ensures that with high probability*

$$\mathrm{Reg}_{\mathsf{CB}}(T) \leq C \cdot \sqrt{KT \cdot \mathrm{Reg}_{\mathsf{Sq}}(T)},$$

where $C > 0$ *is a small numerical constant. Moreover,* SquareCB *inherits the memory and runtime requirements of the oracle.*

We show (Section 3) that SquareCB is *optimal*, in the sense that for every class $\mathcal{F}$, there exists a choice for the oracle SqAlg such that SquareCB attains the minimax optimal rate for $\mathcal{F}$. For example, when $|\mathcal{F}| < \infty$, one can choose SqAlg such that $\mathrm{Reg}_{\mathsf{Sq}}(T) \leq 2\log|\mathcal{F}|$, and so SquareCB enjoys the optimal rate $\mathrm{Reg}_{\mathsf{CB}}(T) \leq C\sqrt{KT \log|\mathcal{F}|}$ for finite classes (Agarwal et al., 2012). On the other hand, the reduction is black-box in nature, so on the practical side one can simply choose SqAlg to be whatever works best.

An advantage of working with 1) regression and 2) online oracles is that we can instantiate SquareCB reduction to give new provable end-to-end regret guarantees for concrete function classes of interest. In Section 2 we flesh this direction out and provide new guarantees for high-dimensional linear classes, generalized linear models, and kernels. SquareCB is also robust to model misspecification: we show (Section 5.1) that the performance gracefully degrades when the realizability assumption is satisfied only approximately.

Compared to previous methods, which either maintain global confidence intervals, version spaces, or distributions over feasible hypotheses, our method applies a simple mapping proposed by Abe & Long (1999) from scores to action probabilities at each step. This leads to the method's efficient runtime guarantee. In Section 5.2 we show that this type of reduction extend beyond the finite actions by designing a variant of SquareCB that has $\mathrm{Reg}_{\mathsf{CB}}(T) \leq C\sqrt{d_{\mathcal{A}} T \cdot \mathrm{Reg}_{\mathsf{Sq}}(T)}$ for the setting where actions live in the $d_{\mathcal{A}}$-dimensional unit ball in ℓ_2.

1.3. Towards Learning-Theoretic Guarantees for Contextual Bandits

The broader goal of this work is to develop a deeper understanding of the algorithmic principles and statistical complexity of contextual bandit learning in the "large-$\mathcal{F}$, small-$\mathcal{A}$" regime, where the goal is to learn from a rich, potentially nonparametric function class with a small number of actions. We call this setting "**C**ontextual **B**andits with **Ri**ch **C**lasses of **H**ypotheses", or RichCBs.

Beyond providing a general algorithmic principle for RichCBs (SquareCB), we resolve two central questions regarding the statistical complexity of RichCBs.

1. What are the minimax rates for RichCBs when $|\mathcal{F}| = \infty$?
2. Can we achieve logarithmic regret for RichCBs when the underlying instance has a gap?

Recall that for general finite classes $\mathcal{F}$, the gold standard here is $\mathrm{Reg}_{\mathsf{CB}}(T) \le \sqrt{KT\log|\mathcal{F}|}$, with an emphasis on the logarithmic scaling in $|\mathcal{F}|$. For the first point, we characterize (Section 3) the minimax rates for infinite classes $\mathcal{F}$ as a function of *metric entropy*, a fundamental complexity measure in learning theory. We also show that SquareCB is universal, in the sense that it can always be instantiated with a choice of SqAlg to achieve the minimax rate. Interestingly, we show that for general classes with metric entropy $\mathcal{H}(\mathcal{F},\varepsilon)$, the minimax rate is $\widetilde{\Theta}(T\cdot\varepsilon_T)$, where ε_T satisfies the classical balance

$$\varepsilon_T^2 \asymp \mathcal{H}(\mathcal{F},\varepsilon_T)/T,$$

found throughout the literature on nonparametric estimation (Yang & Barron, 1999; Tsybakov, 2008).

For the second point, we show (Section 4), that for general function classes $\mathcal{F}$ with $|\mathcal{F}| < \infty$, obtaining logarithmic regret when there is a gap between the best and second-best action is impossible if we insist that regret scales with $\mathrm{polylog}|\mathcal{F}|$: There exist instances with constant gap and polynomially large hypothesis class for which any algorithm must experience $\sqrt{T}$-regret.

This last point suggests that designing optimal algorithms for RichCBs seems to require new algorithmic ideas. Indeed, two of the dominant strategies for the realizable setting, generalized UCB and Thompson sampling (Russo & Van Roy, 2013), always adapt to the gap to get logarithmic regret, but without strong structural assumptions on $\mathcal{F}$ they can have regret $\Omega(|\mathcal{F}|)$.

1.4. Related Work

Our algorithm builds off of the work of Abe & Long (1999) (see also Abe et al. (2003)). Our key insight is that a particular action selection scheme used in these works for linear contextual bandits actually yields an algorithm for general function classes when combined with the idea of an online regression oracle. Interestingly, while Abe & Long (1999) contains essentially the first formulation of the contextual bandit problem, the techniques used within seem to have been forgotten by time in favor of more recent approaches to linear contextual bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011); see further discussion in Section 2.

As discussed in the introduction, our results build on a long line of work on oracle-efficient contextual bandit algorithms. We discuss some important points of comparison below.

Agnostic algorithms. The longest line of research on oracle-efficient CBs focuses on the agnostic i.i.d. setting (Langford & Zhang, 2008; Dudik et al., 2011; Agarwal et al., 2014). All of these algorithms assume access to an *offline* cost-sensitive classification oracle for the policy class which, given a dataset $(x_1,\ell_1),\ldots,(x_n,\ell_n)$, solves

$$\arg\min_{\pi\in\Pi}\sum_{t=1}^{n}\ell_t(\pi(x_t)). \tag{3}$$

In particular, the ILOVETOCONBANDITS (ILTCB) algorithm (Agarwal et al., 2014) enjoys optimal $\sqrt{KT\log|\Pi|}$ regret given such an oracle. This type of oracle has two drawbacks. First, classification for arbitrary datasets is intractable for most policy classes, so implementations typically resort to heuristics to implement (3). Second, because the oracle is *offline*, the memory required by ILTCB scales linearly with T (the algorithm repeatedly generates augmented versions of the dataset and feeds them into the oracle). To deal with this issue, the implementation of ILTCB in Agarwal et al. (2014) resorts to heuristics in order to make use of an online oracle classification, but the resulting algorithm has no guarantees, and analyzing it was left as an open problem.

A parallel line of work focuses on algorithms for the *adversarial* setting where losses are also arbitrary (Rakhlin & Sridharan, 2016; Syrgkanis et al., 2016a;b). Notably, the BISTRO algorithm (Rakhlin & Sridharan, 2016) essentially gives a reduction from adversarial CBs to a particular class of "relaxation-based" online learning algorithms for cost-sensitive classification, but the algorithm has sub-optimal $T^{3/4}$ regret for finite classes.

Realizability-based algorithms. Under the realizability assumption, Foster et al. (2018) provide a version of the UCB strategy for general function classes (Russo & Van Roy, 2014) that makes use of a *offline regression oracle* that solves

$$\arg\min_{f\in\mathcal{F}}\sum_{t=1}^{n}(f(x_t,a_t)-\ell_t(a_t))^2. \tag{4}$$

While this is typically an easier optimization problem than (3)—it can be solved in closed form for linear classes and is amenable to gradient-based methods—the algorithm only attains optimal regret under strong distributional assumptions (beyond just realizability) or when the class $\mathcal{F}$ has bounded eluder dimension (Russo & Van Roy, 2013), and it can have linear regret when these assumptions fail to hold (Foster et al., 2018, Proposition 1).

Thompson sampling and posterior sampling are closely related to UCB and have similar regret guarantees (Russo & Van Roy, 2014). These algorithms are only efficient for certain simple classes $\mathcal{F}$, and implementations for general classes resort to heuristics such as bootstrapping, which do not have strong theoretical guarantees except in special cases (Vaswani et al., 2018; Kveton et al., 2019).

We mention in passing that under our assumptions (realizability, online regression oracle), one can design an online oracle-efficient variant of ε-Greedy with $T^{2/3}$-type regret; $\mathsf{SquareCB}$ appears to be strictly superior.

Other square loss-related reductions. Abernethy et al. (2013) consider the related problem reducing realizable contextual bandits with general function classes $\mathcal{F}$ *and* large action spaces to knows-what-it-knows (KWIK) learning oracles (Li et al., 2011). KWIK learning is much stronger property than regret minimization, and KWIK learners only exist for certain structured hypotheses classes. Interestingly though, this work also provides a computational lower bound which suggests that efficient reductions of the type we provide here ($\mathsf{SquareCB}$) are *not* possible if one insists on $\log K$ dependence rather than $\mathrm{poly}(K)$ dependence.

Abbasi-Yadkori et al. (2012) develops contextual bandit algorithms that use online regression algorithms to form confidence sets for use within UCB-style algorithms. Ultimately these algorithms inherit the usual drawbacks of UCB, namely that they require either strong assumptions on the structure of $\mathcal{F}$ or strong distributional assumptions.

1.5. Additional Notation

We adopt non-asymptotic big-oh notation: For functions $f, g : \mathcal{X} \to \mathbb{R}_+$, we write $f = \mathcal{O}(g)$ if there exists some constant $C > 0$ such that $f(x) \leq Cg(x)$ for all $x \in \mathcal{X}$. We write $f = \widetilde{\mathcal{O}}(g)$ if $f = \mathcal{O}(g \max\{1, \mathrm{polylog}(g)\})$.

For a vector $x \in \mathbb{R}^d$, we let $\|x\|_2$ denote the euclidean norm and $\|x\|_\infty$ denote the element-wise ℓ_∞ norm. For a matrix A, we let $\|A\|_{\mathrm{op}}$ denote the operator norm. If A is symmetric, we let $\lambda_{\min}(A)$ denote the minimum eigenvalue. When $P \succ 0$ is a positive definite matrix, we let $\|x\|_P = \sqrt{\langle x, Px \rangle}$ denote the induced weighted euclidean norm.

2. The Reduction: SquareCB

We now describe our main algorithm, $\mathsf{SquareCB}$, and state our main regret guarantee and some consequences for concrete function classes. To give the guarantees, we first formalize the concept of an online regression oracle, as sketched in the introduction.

2.1. Online Regression Oracles

We assume access to an oracle SqAlg for the standard online learning setting with the square loss (Cesa-Bianchi & Lugosi, 2006, Ch. 3). The oracle performs real-valued online regression with features in $\mathcal{Z} := \mathcal{X} \times \mathcal{A}$, and is assumed to have a prediction error guarantee relative to the regression function class $\mathcal{F}$. We consider the following model:

For $t = 1, \ldots, T$:

- Nature chooses input instance $z_t = (x_t, a_t)$.
- Algorithm chooses prediction $\widehat{y}_t$.
- Nature chooses outcome y_t.

Formally, we model the algorithm as a sequence of mappings $\mathsf{SqAlg}_t : \mathcal{Z} \times (\mathcal{Z} \times \mathbb{R})^{t-1} \to [0,1]$, so that $\widehat{y}_t = \mathsf{SqAlg}_t(z_t\,; (z_1, y_1), \ldots, (z_{t-1}, y_{t-1}))$ in the protocol above. Each such algorithm induces a mapping

$$\widehat{y}_t(x,a) := \mathsf{SqAlg}_t(x, a\,; (z_1, y_1), \ldots, (z_{t-1}, y_{t-1})), \quad (5)$$

which corresponds to the prediction the algorithm would make at time t if we froze its internal state and fed in the feature vector (x, a).

The simplest condition under which our reduction works posits that SqAlg enjoys a regret bound for individual sequence prediction.

Assumption 2a. *The algorithm* SqAlg *guarantees that for every (possibly adaptively chosen) sequence* $z_{1:T}, y_{1:T}$, *regret is bounded as*

$$\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(z_t) - y_t)^2 \leq \mathrm{Reg}_{\mathsf{Sq}}(T). \quad (6)$$

While there is a relatively complete theory characterizing what regret bounds $\mathrm{Reg}_{\mathsf{Sq}}(T)$ can be achieved for this setting for general classes $\mathcal{F}$ (Rakhlin & Sridharan, 2014), the requirement that the regret bound holds for arbitrary sequences $y_{1:T}$ may be restrictive for some classes, at least as far as efficient algorithms are concerned. The following relaxed assumption also suffices.

Assumption 2b. *Under Assumption 1, the algorithm* SqAlg *guarantees that for every (possibly adaptively chosen) sequence* $\{(x_t, a_t)\}_{t=1}^{T}$, *we have*

$$\sum_{t=1}^{T} (\widehat{y}_t - f^\star(x_t, a_t))^2 \leq \mathrm{Reg}_{\mathsf{Sq}}(T). \quad (7)$$

Assumption 2b holds with high probability whenever Assumption 2a holds and the problem is realizable, but it is a weaker condition that allows for algorithms tailored toward realizability; we shall see examples of this in the sequel. This formulation shows that the choice of square loss in (6) does not actually play a critical role: Any algorithm that attains a regret bound of the form (6) with the square loss replaced by a *strongly convex* loss such as the log loss implies a bound of the type (7) under realizability.

2.2. The Algorithm

Our main algorithm, SquareCB, is presented in Algorithm 1. At time t, the algorithm receives the context x_t and computes the oracle's predicted scores $\widehat{y}_t(x_t, a)$ for each action. Then, following the probability selection scheme of Abe & Long (1999), it computes the action with the lowest score (b_t) and assigns a probability to every other action inversely proportional to the gap between the action's score and that of b_t. Finally, the algorithm samples its action a_t from this distribution, observes the loss $\ell_t(a_t)$, and feeds the tuple $((x_t, a_t), \ell_t(a_t))$ into the oracle. The main guarantee for the algorithm is as follows.

Theorem 1. *Suppose Assumption 1 and Assumption 2a/b hold. Then for any* $\delta > 0$*, by setting* $\mu = K$ *and* $\gamma = \sqrt{KT/(\mathrm{Reg}_{\mathsf{Sq}}(T) + \log(2\delta^{-1}))}$, SquareCB *guarantees that with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}_{\mathsf{CB}}(T) \le 4\sqrt{KT \cdot \mathrm{Reg}_{\mathsf{Sq}}(T)} + 8\sqrt{KT \log(2\delta^{-1})}. \tag{8}$$

Let us discuss some key features of the algorithm and regret bound.

- The algorithm enjoys $\widetilde{\mathcal{O}}(\sqrt{T})$-regret whenever the oracle SqAlg gets a fast $\log T$-type rate for online regression. This holds for finite classes ($\mathrm{Reg}_{\mathsf{Sq}}(T) = \log|\mathcal{F}|$) as well as parametric classes such as linear functions in $\mathbb{R}^d$ ($\mathrm{Reg}_{\mathsf{Sq}}(T) = d \log(T/d)$). We sketch some more examples below, and we show in Section 3 that the regret is optimal whenever SqAlg is optimal.
- The algorithm inherits the runtime and memory requirements of the oracle SqAlg up to lower order terms. If $\mathcal{T}_{\mathsf{SqAlg}}$ denotes per-round runtime for SqAlg and $\mathcal{M}_{\mathsf{SqAlg}}$ denotes the maximum memory, then the per-round runtime of SquareCB is $\mathcal{O}(\mathcal{T}_{\mathsf{SqAlg}} \cdot K)$, and the maximum memory is $\mathcal{O}(\mathcal{M}_{\mathsf{SqAlg}} \cdot K)$.
- The regret scales as $\sqrt{K}$ in the number of actions. This is near-optimal in the sense that any algorithm that works *uniformly* for all oracles must pay a $\widetilde{\Omega}(\sqrt{K})$ factor: For multi-armed bandits, one can achieve $\mathrm{Reg}_{\mathsf{Sq}}(T) = \log K$,[1] yet the optimal bandit regret is $\Omega(\sqrt{KT})$. However, for specific function classes, the dependence on K may be suboptimal.[2]

Algorithm 1 SquareCB

```
parameters:
    Learning rate $\gamma > 0$, exploration parameter $\mu > 0$.
    Online regression oracle SqAlg.
for $t = 1, \dots, T$ do
    Receive context $x_t$.
    // Compute oracle's predictions (Eq. (5)).
    For each action $a \in \mathcal{A}$, compute $\widehat{y}_{t,a} := \widehat{y}_t(x_t, a)$.
    Let $b_t = \arg\min_{a \in \mathcal{A}} \widehat{y}_{t,a}$.
    For each $a \neq b_t$, define $p_{t,a} = \frac{1}{\mu + \gamma(\widehat{y}_{t,a} - \widehat{y}_{t,b_t})}$,
      and let $p_{t,b_t} = 1 - \sum_{a \neq b_t} p_{t,a}$.
    Sample $a_t \sim p_t$ and observe loss $\ell_t(a_t)$.
    Update SqAlg with example $((x_t, a_t), \ell_t(a_t))$.
```

At a conceptual level, the proof (which, beyond the idea of using a generic regression oracle and taking advantage of modern martingale tail bounds, closely follows Abe & Long (1999)) is interesting because it is agnostic to the structure of the class $\mathcal{F}$. We show that at each timestep, the instantaneous bandit regret is upper bounded by the instantaneous square loss regret of SqAlg. No structure is shared across timesteps, and all of the heavy lifting regarding generalization is taken care of by Assumption 2a/Assumption 2b.

One important point to discuss is the assumption that the bound (7) holds for every sequence $\{(x_t, a_t)\}_{t=1}^{T}$. While the assumption that the bound holds for adaptively chosen contexts x can be removed if contexts are i.i.d., the analysis critically uses that the regret bound holds when the actions $a_1, \dots, a_T$ are chosen adaptively (since actions selected in early rounds are used by SquareCB to determine the action distribution at later rounds). On a related note, even when contexts are i.i.d., it is not clear that one can implement an online regression oracle that satisfies the requirements of Theorem 1 via calls to an offline regression oracle, and offline versus online regression oracles appear to be incomparable assumptions. Whether optimal regret can be attained via reduction to an offline oracle is an open question.

2.3. Examples and Applications

Online square loss regression is a well-studied problem, and efficient algorithms with provable regret guarantees are known for many classes (Vovk, 1998; Azoury & Warmuth, 2001; Vovk, 2006; Gerchinovitz, 2013; Rakhlin & Sridharan, 2014; Gaillard & Gerchinovitz, 2015). Here we take advantage of these results by instantiating SqAlg within SquareCB to derive end-to-end regret guarantees for various classes—some new, some old.

[1] This can be achieved through Vovk's aggregating algorithm (Vovk, 1995).

[2] For example, for linear classes, regret can be made to scale only with $\log K$ (Chu et al., 2011).

Low-dimensional linear classes. We first consider the familiar LinUCB setting, where

$$\mathcal{F} = \left\{(x,a) \mapsto \langle \theta, x_a\rangle \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \le 1\right\}, \qquad (9)$$

and $x = (x_a)_{a\in\mathcal{A}}$, where $x_a \in \mathbb{R}^d$ has $\|x_a\|_2 \le 1$. Here LinUCB obtains $\mathrm{Reg}_{\mathsf{CB}}(T) \le \mathcal{O}(\sqrt{dT\log^3(KT)})$ (Chu et al., 2011). By choosing SqAlg to be the Vovk-Azoury-Warmuth forecaster, which has $\mathrm{Reg}_{\mathsf{Sq}}(T) \le d\log(T/d)$ (Vovk, 1998; Azoury & Warmuth, 2001), SquareCB has $\mathrm{Reg}_{\mathsf{CB}}(T) \le \mathcal{O}(\sqrt{dKT\log(T/d)})$.[3] While this has worse dependence on K (square root rather than logarithmic), the resulting algorithm works when contexts are chosen by an adaptive adversary, whereas LinUCB requires an oblivious adversary. It would be interesting to understand whether such a tradeoff is optimal. We also remark that—ignoring dependence on K—the algorithm precisely matches a recently established lower bound of $\Omega(\sqrt{dT\log(T/d)})$ for this setting (Li et al., 2019).

High-dimensional linear classes and Banach spaces. In the same setting as above, by choosing SqAlg to be Online Gradient Descent, we obtain $\mathrm{Reg}_{\mathsf{Sq}}(T) \le \mathcal{O}(\sqrt{T})$, and consequently $\mathrm{Reg}_{\mathsf{CB}}(T) \le \mathcal{O}(K^{1/2}\cdot T^{3/4})$. This rate is interesting because it has worse dependence on the time-horizon T, but is completely *dimension-independent*, and the algorithm runs in linear time, which is considerably faster than LinUCB ($\mathcal{O}(d^2)$ per step). This result generalizes the BW algorithm of Abe et al. (2003), who gave the same bound for the setting where rewards are binary, and showed that $T^{3/4}$ is optimal when d is large. We believe this tradeoff between dimension dependence and T dependence has been somewhat overlooked and merits further investigation, especially as it pertains to practical algorithms.

For a more general version of this result, we let $(\mathfrak{B}, \|\cdot\|)$ be a separable Banach space and take

$$\mathcal{F} = \{(x,a) \mapsto \langle \theta, x_a\rangle \mid \theta \in \mathfrak{B}, \|\theta\| \le 1\},$$

where x_a to belongs to the dual space $(\mathfrak{B}^\star, \|\cdot\|_\star)$ and has $\|x_a\|_\star \le 1$. For this setting, whenever $\mathfrak{B}$ is $(2, D)$-uniformly convex, Online Mirror Descent can be configured to have $\mathrm{Reg}_{\mathsf{Sq}}(T) \le \sqrt{T/D}$ (Srebro et al., 2011), and SquareCB consequently has $\mathrm{Reg}_{\mathsf{CB}}(T) \le \mathcal{O}(K^{1/2}\cdot T^{3/4}D^{-1/4})$. This leads to linear time algorithms with nearly dimension-free rates for, e.g., ℓ_1- and nuclear norm-constrained linear classes.

Kernels. Suppose that $\mathcal{F}$ is a reproducing kernel Hilbert space with RKHS norm $\|\cdot\|_{\mathcal{H}}$ and kernel $\mathcal{K}$. Let $\|f\|_{\mathcal{H}} \le 1$ for all $f \in \mathcal{H}$, and assume $\mathcal{K}(x_a, x_a) \le 1$ for all $x \in \mathcal{X}$. A simple observation is that, since Online Gradient Descent kernelizes, the $\mathcal{O}(T^{3/4})$ regret bound from the previous example immediately extends to this setting. This appear to be a new result; Previous work on kernel-based contextual bandits (Valko et al., 2013) gives regret bounds of the form $\sqrt{d_{\text{eff}}T}$, assuming that the effective dimension d_{eff} of the empirical design matrix is bounded. Again there is a tradeoff, since our result requires no assumptions on the data beyond bounded RKHS norm, but has worse (albeit optimal under these assumptions) dependence on the time horizon.

Generalized linear models. Let $\sigma : \mathbb{R} \to [0,1]$ be a fixed non-decreasing 1-Lipschitz link function, and let

$$\mathcal{F} = \left\{(x,a) \mapsto \sigma(\langle \theta, x_a\rangle) \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \le 1\right\},$$

where we again take $\|x_a\|_2 \le 1$. For this setting, under the realizability assumption, the GLMtron algorithm (Kakade et al., 2011) satisfies Assumption 2b, in the sense that is has

$$\sum_{t=1}^{T}(\widehat{y}_t - \sigma(\langle \theta^\star, x_{a_t}\rangle))^2 \le \mathcal{O}(\sqrt{T}),$$

where $f^\star(x,a) = \sigma(\langle\theta^\star, x_a\rangle)$; see Proposition 2 in Appendix B.2 for details. This leads to a dimension-free regret bound $\mathrm{Reg}_{\mathsf{CB}}(T) \le \mathcal{O}(T^{3/4})$, similar to the linear setting. If we have a lower bound on the link function derivative (i.e., $\sigma' \ge c_\sigma > 0$), then a second-order variant of GLMtron (Proposition 3) satisfies Assumption 2b with $\mathrm{Reg}_{\mathsf{Sq}}(T) = \mathcal{O}(d\log T/c_\sigma^2)$. Plugging this into SquareCB gives regret $\mathcal{O}(\sqrt{dKT\log T/c_\sigma^2})$. This matches the dependence on d and T in previous results for generalized linear contextual bandits with finite actions (Li et al., 2017), but unlike these results the algorithm does not require stochastic contexts, and requires no assumptions on the design matrix $\frac{1}{T}\sum_{t=1}^{T} x_{t,a_t}x_{t,a_t}^\top$ or its population analogue.

2.4. Minimax Perspective

The analysis of SquareCB is interesting because the reduction from square loss regret to contextual bandit regret completely ignores the structure of the function class $\mathcal{F}$. At a high level, the proof proceeds by showing that the probability selection strategy ensures that

$$\sum_{t=1}^{T}\mathbb{E}_{a\sim p_t}[f^\star(x_t,a) - f^\star(x_t,\pi^\star(x_t))]$$
$$\le \frac{2KT}{\gamma} + \frac{\gamma}{4}\sum_{t=1}^{T}\mathbb{E}_{a\sim p_t}\Big[(\widehat{y}_{t,a} - f^\star(x_t,a))^2\Big]. \qquad (10)$$

at which point we can bound the right-hand side by using the regret bound for SqAlg. In fact, the probability selection strategy in SquareCB actually gives a stronger guarantee

[3] In order satisfy the condition that predictions $\widehat{y}_t$ are bounded, we must use a variant of Vovk-Azoury-Warmuth with projection onto the ℓ_2 ball. This can easily be achieved using, e.g., the analysis in Orabona et al. (2015).

than (10). Consider the following *per-round minimax* problem, whose value $\mathsf{Val}(\gamma)$ is given by

$$\max_{\widehat{y}\in[0,1]^K}\min_{p\in\Delta_K}\max_{f^\star\in[0,1]^K}\max_{a^\star}\mathop{\mathbb{E}}_{a\sim p}\left[f^\star_a-f^\star_{a^\star}-\frac{\gamma}{4}(\widehat{y}_a-f^\star_a)^2\right]. \tag{11}$$

If $\mathsf{Val}(\gamma)\leq c$, we can interpret this as saying, "For every choice of $\widehat{y}$, there exists an action distribution such that regardless of the value of $f^\star$, the immediate regret with respect to $f^\star$ is bounded by the squared prediction error of $\widehat{y}$, plus a constant c." The probability selection rule used in SquareCB with parameter γ certifies that $\mathsf{Val}(\gamma)\leq\frac{2K}{\gamma}$. The takeaway is that, at the level of the reduction, $f^\star(x_t,a)$ might as well be chosen adversarially at each round rather than realized by a specific function $f^\star\in\mathcal{F}$ chosen a-priori. We are hopeful that this per-round minimax approach to reductions will be more broadly useful, and indeed our extension to infinite actions in Section 5.2 uses similar per-round reasoning. To close the section, we give a lower bound on the minimax value which shows that the action selection strategy used in SquareCB is near-optimal for the minimax problem (11).

Proposition 1. For any $\gamma\geq 2$, we have $\mathsf{Val}(\gamma)\geq\frac{(1-1/K)}{\gamma}$.

3. Optimality and Universality

In light of Theorem 1, a natural question is whether one can always instantiate SquareCB such that its regret is optimal for the class $\mathcal{F}$ under consideration. More broadly, we seek to understand the minimax rate for the RichCB setting where $\mathcal{F}$ is a large, potentially nonparametric function class and the problem is realizable. In this section we first prove a lower bound on minimax regret achievable for any function class $\mathcal{F}$. We then show that SquareCB is *universal*, in the sense that there always exist a choice for SqAlg that achieves the lower bound (up to dependence on the number of actions, which is not our focus).

For technical reasons, we make two simplifying assumptions in this section. First, we focus on the setting where (x_t,ℓ_t) are drawn i.i.d. from a joint distribution μ. Second, we assume that the regression function class $\mathcal{F}$ *tensorizes*: There is a base function class $\mathcal{G}\subseteq(\mathcal{X}\to[0,1])$ such that $\mathcal{F}=\mathcal{G}^K$, in the sense $\mathcal{F}$ consists of functions of the form $f(x,a)=g_a(x)$, where $g_a\in\mathcal{G}$.

Our upper and lower bounds are stated in terms of the *metric entropy* of the base class $\mathcal{G}$. For a sample set $S=\{x_1,\ldots,x_n\}$, let $\mathcal{N}_2(\mathcal{G},\varepsilon,S)$ denote the size of the smallest set $\mathcal{G}'$ such that

$$\forall g\in\mathcal{G},\ \ \exists g'\in\mathcal{G}'\ \text{ s.t. }\left(\frac{1}{n}\sum_{t=1}^{n}(g(x_t)-g'(x_t))^2\right)^{1/2}\leq\varepsilon.$$

The *empirical entropy* of $\mathcal{G}$ is then defined as $\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon)=\sup_{n\geq1,S\in\mathcal{X}^n}\log\mathcal{N}_2(\mathcal{G},\varepsilon,S)$. Empirical entropy is a fundamental quantity in statistical learning that is both necessary and sufficient for learnability, as well as polynomially related to other standard complexity measures such as (local) Rademacher complexity and fat-shattering dimension (Rakhlin & Sridharan, 2012). We give concrete examples in the sequel, but for now we make the following assumption.

Assumption 3. *Contexts and losses are drawn i.i.d. from a joint distribution μ, and there exists a constant $p>0$ such that for all $\varepsilon>0$, the empirical entropy for $\mathcal{G}$ grows as*

$$\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon)\lesssim\varepsilon^{-p}.$$

Our upper and lower bounds characterize the optimal regret for RichCBs as a function of the growth rate parameter $p>0$ in Assumption 3. We first state the lower bound.

Theorem 2 (Lower bound)**.** *Let $\mathcal{G}$ be any function class for which $\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon)=\Theta(\varepsilon^{-p})$ for some $p>0$. Then there exists a slightly modified class $\mathcal{G}'$ with $\mathcal{H}^{\mathrm{iid}}(\mathcal{G}',\varepsilon)=\widetilde{\Theta}(\varepsilon^{-p})$ for which the corresponding function class $\mathcal{F}$ (with $K=2$) is such that any algorithm must have*

$$\mathbb{E}[\mathrm{Reg}_{\mathsf{CB}}(T)]\geq\widetilde{\Omega}\Big(T^{\frac{1+p}{2+p}}\Big), \tag{12}$$

on some realizable instance for $\mathcal{F}$.

We now show that SquareCB can always be instantiated to match the lower bound (12) in terms of dependence on T.

Theorem 3 (Universality of SquareCB)**.** *Whenever Assumption 3 holds, there exists a choice for the base regret minimization algorithm* SqAlg *such that with probability at least* $1-\delta$, SquareCB *has*

$$\mathrm{Reg}_{\mathsf{CB}}(T)\leq\widetilde{\mathcal{O}}\Big((KT)^{\frac{1+p}{2+p}}+\sqrt{K^2T\log(\delta^{-1})}\Big).$$

The idea behind the proof of Theorem 3 is to choose SqAlg to run Vovk's aggregating algorithm over an empirical cover for $\mathcal{G}$. The main difficulty is that we must find a cover that is close on the distribution μ, which the algorithm has no prior knowledge of. To get around this issue, the algorithm continually refines a cover based on data collected so far.

Examples. Let us make matters slightly more concrete and show how to extract some familiar regret bounds from Theorem 2 and Theorem 3. First, for linear classes (9) (specifically, the tensorized variants), one has $\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon)\propto d\log(1/\varepsilon)\wedge\varepsilon^{-2}$ (Zhang, 2002), and hence the theorems recover the $\sqrt{dT}$ and $T^{3/4}$ regret bounds for linear classes described in the previous section.

Slivkins (2011) derives fairly general results for nonparametric contextual bandits. As one example, their results imply that when $\mathcal{G}$ is the set of all 1-Lipcshitz functions over $[0,1]^d$, the optimal regret is $T^{\frac{1+d}{2+d}}$. Since such classes have $\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon)\propto\varepsilon^{-d}$, our theorems recover this result.

Similarly, for Hölder-smooth functions of order β, we have $p = d/\beta$ which yields the rate $T^{\frac{d+\beta}{d+2\beta}}$ (Rigollet & Zeevi, 2010).

As a final example, Bartlett et al. (2017) show that neural networks with appropriately bounded spectral norm and $\ell_{2,1}$ norm have $\mathcal{H}^{\mathrm{iid}}(\mathcal{G},\varepsilon) \propto \varepsilon^{-2}$. Our theorems imply that $\widetilde{\Theta}(T^{3/4})$ is optimal for such models.

Discussion. The assumptions made in this section (tensorization, stochastic contexts) can be relaxed, but we do not have a complete picture of the optimal regret for all values of p in this case. For adversarial contexts and without the tensorization assumption, if $\mathcal{F}$ has bounded *sequential* metric entropy then Theorem 1 of Rakhlin & Sridharan (2014) implies that there exists a choice for SqAlg such that $\mathrm{Reg}_{\mathsf{Sq}}(T) \leq T^{1-\frac{2}{2+p}}$ and thus SquareCB has $\mathrm{Reg}_{\mathsf{CB}}(T) \leq \mathcal{O}(T^{\frac{1+p}{2+p}})$ as in Theorem 3, but only for $p \leq 2$. On the other hand, for stochastic contexts it is also possible to show that a variant of the algorithm in Theorem 3 based on slightly different concentration arguments matches the regret bound $\mathcal{O}(T^{\frac{1+p}{2+p}})$ without the tensorization assumption, but only for $p \geq 1$. Resolving the optimal dependence on K seems challenging and likely requires more refined complexity measures; see also Daniely et al. (2015b).

Previous works have given regret bounds for infinite policy classes that depend on the complexity (e.g., VC dimension) of the policy class (Beygelzimer et al., 2011; Foster & Krishnamurthy, 2018). These guarantees are somewhat different than the ones we provide here, which depend on the complexity of the regression function class $\mathcal{F}$ rather than the class of policies it induces (but require realizability).

4. On Gap-Dependent Regret Bounds

In this section we give some negative results regarding instance-dependent regret bounds for RichCBs. Since Theorem 1 recovers the usual $\widetilde{\mathcal{O}}(\sqrt{KT})$ bound for multi-armed bandits, a natural question is whether the algorithm can recover *instance-dependent* regret bounds of the form $\mathcal{O}(\frac{K\log T}{\Delta})$ when there is a gap Δ between the best and second-best action. More ambitiously, can the algorithm achieve similar instance-dependent regret bounds for rich function classes $\mathcal{F}$?

To address this question, we assume Bayes regressor $f^\star$ enjoys a gap between the optimal and second-best action for every context. The following definition is adapted from Dani et al. (2008).

Definition 1 (Uniform gap). *A contextual bandit instance is said to have* uniform gap Δ *if for all $x \in \mathcal{X}$,*

$$f^\star(x,a) - f^\star(x,\pi^\star(x)) > \Delta \quad \forall a \neq \pi^\star(x).$$

We would like to understand whether Theorem 1 can be improved when the uniform gap condition holds. For example, is it possible to select the learning rate γ such that SquareCB has

$$\mathrm{Reg}_{\mathsf{CB}}(T) \leq \frac{K\mathrm{Reg}_{\mathsf{Sq}}(T)}{\Delta}\cdot \mathrm{polylog}(T)? \tag{13}$$

As a special case, such a regret bound would recover the $\widetilde{\mathcal{O}}(\frac{K}{\Delta})$-type regret bound for multi-armed bandits by choosing SqAlg with the exponential weights strategy. More generally, for any finite class $\mathcal{F}$, the hypothesized bound (13) would imply a regret bound of

$$\mathrm{Reg}_{\mathsf{CB}}(T) \leq \widetilde{\mathcal{O}}\left(\frac{K\log|\mathcal{F}|}{\Delta}\right) \tag{14}$$

by taking SqAlg to be Vovk's aggregating algorithm, which has $\mathrm{Reg}_{\mathsf{Sq}}(T) = \log|\mathcal{F}|$. Here we give an information-theoretic lower bound which shows that such a regret bound is not possible, not just for SquareCB but for *any* contextual bandit algorithm.

Theorem 4. *For every T, there exists a function class $\mathcal{F}$ with two arms and $|\mathcal{F}| \leq \sqrt{2T}$ such that for any (potentially randomized) contextual bandit algorithm, there exists a realizable and noiseless contextual bandit instance with uniform gap $\Delta = \frac{1}{4}$ on which*

$$\mathbb{E}[\mathrm{Reg}_{\mathsf{CB}}(T)] \geq \frac{1}{16}\sqrt{T}.$$

The function class in Theorem 4 has $|\mathcal{F}| = \mathcal{O}(\sqrt{T})$, and all instances considered in the theorem have constant gap. For such setups, the hypothesized regret bound (14) would give $\mathrm{Reg}_{\mathsf{CB}}(T) \leq \widetilde{\mathcal{O}}(1)$. Hence, Theorem 4 rules out (14) and (13), and in fact rules out any regret bound of the form $\mathrm{Reg}_{\mathsf{CB}}(T) \leq \widetilde{\mathcal{O}}\left(\frac{K\log|\mathcal{F}|}{\Delta}\cdot T^{1/2-\varepsilon}\right)$ for constant ε.

In essence, the theorem shows that to obtain instance-dependent regret guarantees, one can at best hope for regret bounds that scale with $\frac{|\mathcal{F}|}{\Delta}$ rather than $\frac{\log|\mathcal{F}|}{\Delta}$. In other words, instance-dependent regret is at odds with learning from rich function classes (RichCBs), where regret scaling with $|\mathcal{F}|$ is unacceptable. It is known that for linear function classes $\mathcal{F}$, and more broadly function classes that satisfy certain structural assumptions such as bounded *eluder dimension* (Russo & Van Roy, 2013), gap-dependent logarithmic regret bounds are achievable through variants of UCB and Thompson sampling. However, bounded eluder dimension is a rather strong assumption which is essentially only known to hold for linear models, generalized linear models, and classes for which the domain size $|\mathcal{X}|$ is bounded.[4] Theorem 4 shows that such assumptions are qualitatively required for instance-dependent logarithmic regret guarantees.

[4] There is no contradiction with Theorem 4, as the construction in the theorem scales $|\mathcal{X}|$ as $\sqrt{T}$

Langford & Zhang (2008) consider a different gap notion we refer to as a *policy gap* which, in the stochastic setting, posits that $L(\pi^\star) < L(\pi) - \Delta_{\text{policy}}$, where $L(\pi) = \mathbb{E}_{x,\ell}\,\ell(\pi(x))$. For instances with policy gap Δ_{policy}, they show that the Epoch-Greedy algorithm achieves regret $\text{poly}(\log|\Pi|, \log T, \Delta_{\text{policy}}^{-2})$. There is no contradiction between this result and Theorem 4, as the construction in the theorem has policy gap $\frac{1}{\sqrt{T}}$.

5. Extensions

5.1. Misspecified Models

In practice, the realizability assumption (Assumption 1) may be restrictive. In this section we show that the performance of SquareCB gracefully degrades when the assumption fails to hold, so long as the learning rate is changed appropriately. We consider the following relaxed notion of realizability.

Assumption 4. *There exists a regressor $f^\star \in \mathcal{F}$ such that for all t, $f^\star(x,a) = \mathbb{E}[\ell_t(a) \mid x_t = x] + \varepsilon_t(x,a)$, where $|\varepsilon_t(x,a)| \le \varepsilon$.*

The main theorem for this section shows that when Assumption 4 holds, the performance of SquareCB degrades by an additive $\varepsilon \cdot \sqrt{KT}$ factor. We state the result in terms of Assumption 2a, since this assumption is typically easier to satisfy than Assumption 2b when the model is misspecified.

Theorem 5. *Suppose the adversary satisfies Assumption 4 and* SqAlg *satisfies Assumption 2a. Then* SquareCB *with $\gamma = 2\sqrt{KT/(\text{Reg}_{\mathsf{Sq}}(T) + 2\varepsilon^2 T)}$ and $\mu = K$ ensures that*

$$\overline{\text{Reg}}_{\mathsf{CB}}(T) \le 2\sqrt{KT \cdot \text{Reg}_{\mathsf{Sq}}(T)} + \varepsilon \cdot 5\sqrt{K},$$

where $\overline{\text{Reg}}_{\mathsf{CB}}(T) := \sup_\pi \mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(\pi(x_t))\right]$.

An extension of Theorem 5 for adaptive adversaries is given in Appendix E.1.

Examples and discussion. Regret bounds for misspecified linear contextual bandits have recently gathered interest (Van Roy & Dong, 2019; Lattimore & Szepesvari, 2019; Neu & Olkhovskaya, 2020) due to their connection to reinforcement learning with misspecified linear feature maps (Du et al., 2020). Consider again the LinUCB-type setting where $\mathcal{F} = \left\{(x,a) \mapsto \langle\theta, x_a\rangle \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \le 1\right\}$, so that $x_t = (x_{t,a})_{a\in\mathcal{A}}$ is a finite collection of contexts that varies from round to round. By instantiating SquareCB with the Vovk-Azoury-Warmuth forecaster, which has $\text{Reg}_{\mathsf{Sq}}(T) \le d\log(T/d)$ (Vovk, 1998; Azoury & Warmuth, 2001) and appealing to Theorem 5, we get an efficient algorithm with regret

$$\widetilde{\mathcal{O}}(\sqrt{dKT} + \varepsilon\sqrt{KT}). \tag{15}$$

Previous algorithms with similar guarantees either apply only to non-contextual linear bandits (Lattimore & Szepesvari, 2019), or attain sub-optimal regret and require additional assumptions when specialized to this setting (Neu & Olkhovskaya, 2020).[5] Interestingly, as remarked by Lattimore & Szepesvari (2019), the lower bounds of Du et al. (2020) imply that when $K \gg d$, the "price" of ε-misspecification must grow as $\Omega(\varepsilon\sqrt{d}T)$. On the other hand, our result shows that the price can be improved to $\mathcal{O}(\varepsilon\sqrt{KT})$ in the small-K regime.

Theorem 5 shows that we can be robust to misspecification efficiently whenever online regression is possible efficiently. More broadly, these theorems give the first result we are aware of that considers ε-misspecification for arbitrary classes $\mathcal{F}$. The theorems imply that $\mathcal{O}(\varepsilon\sqrt{KT})$ bounds the price of misspecification for general classes; the complexity of $\mathcal{F}$ is only reflected in $\text{Reg}_{\mathsf{Sq}}(T)$.

5.2. Infinite Actions

While the finite-action setting in which SquareCB works is arguably the most basic and fundamental contextual bandit setting, it is desirable to have algorithms that work for large or infinite sets of actions. As a proof of concept, we extend SquareCB to an infinite-action setting where the action space $\mathcal{A}$ is $d_{\mathcal{A}}$-dimensional unit ℓ_2 ball $\mathbb{B}_{d_{\mathcal{A}}}$. The result is deferred to Appendix E.2 for space.

6. Discussion

We have presented the first optimal reduction from contextual bandits to online square loss reduction. Conceptually, we showed that *online* oracles are a powerful primitive for designing contextual bandit algorithms, both in terms of computational and statistical efficiency. Beyond our algorithmic contribution, we have shed light on the fundamental limits of algorithms for RichCBs, including minimax rates and gap-dependent regret guarantees. We are hopeful that our techniques will find broader use, and that our results will inspire further research on provably efficient contextual bandits with flexible function approximation. Going forward, some natural questions are:

- *Reinforcement learning.* Can the SquareCB strategy be adapted to give regret bounds for reinforcement learning or continuous control with unknown dynamics and function approximation?

- *Adaptivity.* Can the SquareCB strategy or a variant adapt to take advantage of easy or nice data?

[5] Neu & Olkhovskaya (2020) give an efficient algorithm for a similar but incomparable setting in which contexts are stochastic, but the (ε-approximately linear) Bayes regressor can vary adversarially from round to round. Their algorithm has regret $\widetilde{\mathcal{O}}((dK)^{1/3}T^{2/3} + \varepsilon\sqrt{d}T)$.

Acknowledgements

We thank Akshay Krishnamurthy and Haipeng Luo for helpful discussions. We acknowledge the support of ONR award #N00014-20-1-2336. DF acknowledges the support of NSF TRIPODS award #1740751.

References

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

Abbasi-Yadkori, Y., Pál, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9, 2012.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11. Morgan Kaufmann Publishers Inc., 1999.

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

Abernethy, J., Amin, K., Kearns, M., and Draief, M. Large-scale bandit problems and KWIK learning. In *International Conference on Machine Learning*, pp. 588–596, 2013.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. E. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2012.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646, 2014.

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

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

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

Azoury, K. S. and Warmuth, M. K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001.

Bartlett, P. L. and Long, P. M. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Bietti, A., Agarwal, A., and Langford, J. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

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

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 1405–1411, 2015a.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the erm principle. *The Journal of Machine Learning Research*, 16(1):2377–2404, 2015b.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *International Conference on Learning Representations (ICLR)*, 2020.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 169–178. AUAI Press, 2011.

Foster, D. J. and Krishnamurthy, A. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *Advances in Neural Information Processing Systems*, 2018.

Foster, D. J., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. E. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018.

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

Gerchinovitz, S. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pp. 927–935, 2011.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66 (4):1–77, 2019.

Klivans, A. R. and Sherstov, A. A. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Krishnamurthy, A., Agarwal, A., and Dudik, M. Contextual semibandits via supervised learning oracles. In *Advances In Neural Information Processing Systems*, pp. 2388–2396, 2016.

Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 3601–3610, 2019.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824, 2008.

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

Li, L., Littman, M. L., Walsh, T. J., and Strehl, A. L. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82(3):399–443, 2011.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2071–2080. JMLR. org, 2017.

Li, Y., Wang, Y., and Zhou, Y. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pp. 2173–2174, 2019.

Mendelson, S. and Vershynin, R. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1): 37–55, 2003.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. *Conference on Learning Theory (COLT)*, 2020.

Orabona, F., Crammer, K., and Cesa-Bianchi, N. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

Rakhlin, A. and Sridharan, K. Statistical learning and sequential prediction, 2012. Available at http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf.

Rakhlin, A. and Sridharan, K. Online nonparametric regression. In *Conference on Learning Theory*, 2014.

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

Rakhlin, A., Sridharan, K., and Tsybakov, A. B. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23 (2):789–824, 2017.

Rigollet, P. and Zeevi, A. Nonparametric bandits with covariates. *Conference on Learning Theory (COLT)*, pp. 54, 2010.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pp. 2256–2264, 2013.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Slivkins, A. Contextual bandits with similarity information. In *Conference on Learning Theory (COLT)*, 2011.

Srebro, N., Sridharan, K., and Tewari, A. On the universality of online mirror descent. In *Advances in neural information processing systems*, pp. 2645–2653, 2011.

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

Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 3135–3143, 2016b.

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

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 654–663, 2013.

Van Roy, B. and Dong, S. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Vaswani, S., Kveton, B., Wen, Z., Rao, A., Schmidt, M., and Abbasi-Yadkori, Y. New insights into bootstrapping for bandits. *arXiv preprint arXiv:1805.09793*, 2018.

Vovk, V. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 51–60. ACM, 1995.

Vovk, V. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pp. 364–370, Cambridge, MA, USA, 1998. MIT Press.

Vovk, V. Metric entropy in competitive on-line prediction. *arXiv preprint cs/0609045*, 2006.

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pp. 1564–1599, 1999.

Zhang, T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.