Efficient Algorithms for Adversarial Contextual Learning

Vasilis Syrgkanis VASY@MICROSOFT.COM
Microsoft Research, 641 Avenue of the Americas, New York, NY 10011 USA

Akshay Krishnamurthy AKSHAYKR@CS.CMU.EDU
Microsoft Research, 641 Avenue of the Americas, New York, NY 10011 USA

Robert E. Schapire SCHAPIRE@MICROSOFT.COM
Microsoft Research, 641 Avenue of the Americas, New York, NY 10011 USA

Abstract

We provide the first oracle efficient sublinear regret algorithms for adversarial versions of the contextual bandit problem. In this problem, the learner repeatedly makes an action on the basis of a context and receives reward for the chosen action, with the goal of achieving reward competitive with a large class of policies. We analyze two settings: i) in the transductive setting the learner knows the set of contexts a priori, ii) in the small separator setting, there exists a small set of contexts such that any two policies behave differently on one of the contexts in the set. Our algorithms fall into the Follow-The-Perturbed-Leader family (Kalai & Vempala, 2005) and achieve regret $O(T^{3/4}\sqrt{K\log(N)})$ in the transductive setting and $O(T^{2/3}d^{3/4}K\sqrt{\log(N)})$ in the separator setting, where T is the number of rounds, K is the number of actions, N is the number of baseline policies, and d is the size of the separator. We actually solve the more general adversarial contextual semi-bandit linear optimization problem, whilst in the full information setting we address the even more general contextual combinatorial optimization. We provide several extensions and implications of our algorithms, such as switching regret and efficient learning with predictable sequences.

1. Introduction

We study contextual online learning, a powerful framework that encompasses a wide range of sequential decision making problems. Here, on every round, the learner receives contextual information which can be used as an aid in selecting an action. In the full-information version of the problem, the learner then observes the loss that would have been suffered for each of the possible actions, while in the more challenging bandit version, only the loss that was actually incurred (i.e. for the chosen action) is observed. The goal is to achieve low loss over several rounds. The contextual bandit problem is of particular practical relevance, with applications to personalized recommendations, clinical trials, and targeted advertising.

Algorithms for contextual learning, such as Hedge (Freund & Schapire, 1997; Cesa-Bianchi et al., 1997) and Exp4 (Auer et al., 1995), are well-known to have remarkable theoretical properties, being effective even in adversarial, non-stochastic environments and capable of performing almost as well as the best among an exponentially large family of *policies*, or rules for choosing actions. However, the space requirements and running time of these algorithms are generally linear in the number of policies, which is far too expensive for a the many applications that call for an extremely large policy space. In this paper, we address this gap between the statistical and computational complexity of contextual online learning in an adversarial setting.

As an approach to solving online learning problems, we posit that the corresponding batch version is solvable. In other words, we assume access to a certain optimization oracle for solving a batch-learning problem. Concrete instances of such an oracle include empirical risk minimization procedures for supervised learning, algorithms for the shortest paths problem, and dynamic programming.

Such an oracle is central to the Follow-the-Perturbed-Leader algorithms of Kalai & Vempala (2005), although these algorithms are not generally efficient since they require separately "perturbing" each policy in the entire

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

space. Oracles of this kind have also been used in designing efficient contextual bandit algorithms (Agarwal et al., 2014; Langford & Zhang, 2008; Dudík et al., 2011); however, these require a much more benign setting in which contexts and losses are chosen randomly and independently rather than by an adversary.

In this paper, for a wide range of problems, we present computationally efficient algorithms for contextual online learning in an adversarial setting, assuming oracle access. We give results for both the full-information and bandit settings. To the best of our knowledge, these results are the first of their kind at this level of generality.

Overview of results. We begin in Section 2 with a new and general Follow-the-Perturbed-Leader algorithm in the style of Kalai & Vempala (2005). This algorithm *only* accesses the policy class using the optimization oracle.

We then apply these results in Section 3 to two settings. The first is a *transductive setting* (Ben-David et al., 1997) in which the learner knows the arriving contexts a priori, or, less stringently, knows only the set, but not necessarily the actual sequence or multiplicity with which each context arrives. In the second, *small-separator* setting, we assume that the policy space admits the existence of a small set of contexts, called a *separator*, such that any two policies differ on at least one context from the set. The size of the smallest separator for a particular policy class can be viewed as a new measure of complexity, different from the VC dimension, and potentially of independent interest.

We analyze our algorithm for a generalized online learning setting called *online combinatorial optimization*, which includes as special cases transductive contextual experts, online shortest-path routing, online linear optimization (Kalai & Vempala, 2005), and online submodular minimization (Hazan & Kale, 2012).

In Section 4, we extend our results to the bandit setting, or in fact, to the more general semi-bandit setting, using a technique of Neu & Bartók (2013). Among our main results, we obtain regret bounds for the adversarial contextual bandit problem of $O(T^{3/4}\sqrt{K\log(N)})$ in the transductive setting, and $O(T^{2/3}d^{3/4}K\sqrt{\log(N)})$ in the small-separator setting, where T is the number of time steps, K the number of actions, N the size of the policy space, and d the size of the separator. Being sublinear in T, these bounds imply the learner's performance will eventually be almost as good as the best policy, although they are worse than the generally optimal dependence on T of $O(\sqrt{T})$, obtained by many of the algorithms mentioned above. On the other hand, these preceding algorithms are computationally intractable when the policy space is gigantic, while ours runs in polynomial time in T, K, d and $\log(N)$, assuming access to an optimization oracle. Improving these bounds with an efficient algorithm remains an open problem.

In Section 5, we give an efficient algorithm when regret is measured in comparison to a competitor that is allowed to switch from one policy to another a bounded number of times. Here, we show that the optimization oracle can be efficiently implemented given an oracle for the original policy class. Specifically, this leads to a fully efficient algorithm for the online switching shortest path problem in directed acyclic graphs.

Finally, Section 6 shows how "path length" regret bounds can be derived in the style of Rakhlin & Sridharan (2013b). Such bounds have various applications, for instance, in obtaining better bounds for playing repeated games (Rakhlin & Sridharan, 2013a; Syrgkanis et al., 2015). Our results easily extend to infinite policy classes with bounded Natarajan dimension and more generally to classes with bounded Laplacian complexity: $\mathcal{L}(\Pi) = \sup_{x_{1:T}} \mathbb{E}[\sup_{\pi\in\Pi} \sum_{t=1}^{T} L_t(\pi(x_t))]$, with L_t a vector of independent Laplace distributions in each coordinate.

Other related work. Contextual, transductive online learning using an optimization oracle was previously studied by Kakade & Kalai (2005), whose work was later extended and improved by Cesa-Bianchi & Shamir (2011) using a generalization of a technique from Cesa-Bianchi et al. (1997). However, these results are for binary classification or other convex losses defined on one-dimensional predictions and outcomes; as such, they are special cases of the much more general setting we consider here.

Awerbuch & Kleinberg (2008) present an efficient algorithm for the online shortest paths problem. This can be viewed as solving an adversarial bandit problem with a very particular optimization oracle over an exponentially large but highly structured space of "policies" corresponding to paths in a graph. However, their setting is clearly far more restrictive and structured than ours is.

Concurrently with our work, Rakhlin & Sridharan (2016) also obtain sublinear regret guarantees for an oracle-based adversarial contextual bandit algorithm, albeit non-combinatorial. Their algorithm achieves $O(T^{3/4}\sqrt{K}(\log(N))^{1/4})$ in two settings: a hybrid stochastic-adversarial setting, where the contexts are drawn i.i.d. from a distribution that the learner knows how to sample from, and a fully transductive setting, where the contexts and their multiplicities are known a priori. Their algorithm is of the random playout (Cesa-Bianchi & Shamir, 2011) style and is based on a minimax analysis.

2. Online Learning with Oracles

We start by analyzing the family of Follow-the-Perturbed-Leader algorithms in a very general online learning setting. Parts of this generic formulation follow the recent formulation of Daskalakis & Syrgkanis (2015), but we present a more refined analysis which is essential for our contextual learning result in the next sections. The main theorem of this section is essentially a generalization of Theorem 1.1 of Kalai & Vempala (2005).

Consider an online learning problem where at each time-step an adversary picks an outcome $y^t \in \mathcal{Y}$ and the algorithm picks a policy $\pi^t \in \Pi$ from some policy space Π.[1] The algorithm receives a loss: $\ell(\pi^t, y^t)$, which could be positive or negative. At the end of each iteration the algorithm observes the realized outcome y^t. We will denote with $y^{1:t}$ a sequence of outcomes $\{y^1, y^2, \ldots, y^t\}$. Moreover, we denote with:

$$\mathcal{L}(\pi, y^{1:t}) = \sum_{\tau=1}^{t} \ell(\pi, y^\tau), \tag{1}$$

the cumulative loss of a fixed policy $\pi \in \Pi$ for a sequence of choices $y^{1:t}$ of the adversary. The goal of the learning algorithm is to achieve loss that is competitive with the best fixed policy in hindsight. As the algorithms we consider will be randomized, we will analyze the expected regret,

$$\textsc{Regret} = \sup_{\pi^\star \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell(\pi^t, y^t) - \sum_{t=1}^{T} \ell(\pi^\star, y^t)\right], \tag{2}$$

which is the worst case difference between the cumulative loss of the learner and the loss of any fixed policy $\pi \in \Pi$.

We consider adversaries that are *adaptive*, which means that they can choose the outcome y^t at time t, using knowledge of the entire history of interaction. The only knowledge not available to an adaptive adversary is any randomness used by the learning algorithm at time t. In contrast, an *oblivious* adversary is one that picks the sequence of outcomes $y^{1:T}$ before the start of the learning process.

To develop computationally efficient algorithms that compete with large sets of policies Π, we assume that we are given oracle access to the following optimization problem.

Definition 1 (Optimization oracle)**.** Given outcomes $y^{1:t}$ compute the fixed optimal policy for this sequence:

$$M\left(y^{1:t}\right) = \operatorname{argmin}_{\pi \in \Pi} \mathcal{L}(\pi, y^{1:t}). \tag{3}$$

We will also assume that the oracle performs consistent deterministic tie-breaking: i.e. whenever two policies are tied, then it always outputs the same policy.

[1] We refer to the choice of the learner as a policy, for uniformity of notation with subsequent sections, where the learner will choose some policy that maps contexts to actions.

Algorithm 1 Follow the perturbed leader with fake sample perturbations - FTPL.

for each time step t **do**

Draw a random sequence of outcomes $\{z\} = (z^1, \ldots, \ldots, z^k)$ independently, based on some time-independent distribution over sequences. Both the length of the sequence and the outcome $z^i \in \mathcal{Y}$ at each iteration of the sequence can be random

Denote with $\{z\} \cup y^{1:t-1}$ the augmented sequence where we append the extra outcome samples $\{z\}$ at the beginning of sequence $y^{1:t-1}$

Invoke oracle M and play policy:

$$\pi^t = M\left(\{z\} \cup y^{1:t-1}\right). \tag{4}$$

end for

In this generic setting, we define a new family of Follow-The-Perturbed-Leader (FTPL) algorithms where the perturbation takes the form of extra samples of outcomes (see Algorithm 1). In each round, the learning algorithm draws a random sequence of outcomes independently, and appends this sequence to the outcomes experienced during the learning process. The algorithm invokes the oracle on this augmented outcome sequence, and plays the resulting policy.

Perturbed Leader Regret Analysis. We give a general theorem on the regret of a perturbed leader algorithm with sample perturbations. In the sections that follow we will give instances of this analysis in specific settings.

Theorem 1. *For a distribution over sample sequences $\{z\}$ and a sequence of adversarially and adaptively chosen outcomes $y^{1:T}$, define:*

$$\textsc{Stability} = \sum_{t=1}^{T} \mathbb{E}_{\{z\}}\left[\ell(\pi^t, y^t) - \ell(\pi^{t+1}, y^t)\right]$$

$$\textsc{Error} = \mathbb{E}_{\{z\}}\left[\max_{\pi \in \Pi} \sum_{z^\tau \in \{z\}} \ell(\pi, z^\tau)\right] - \mathbb{E}_{\{z\}}\left[\min_{\pi \in \Pi} \sum_{z^\tau \in \{z\}} \ell(\pi, z^\tau)\right],$$

where π^t is defined in Equation (4). Then the expected regret of Algorithm 1 is upper bounded by,

$$\textsc{Regret} \leq \textsc{Stability} + \textsc{Error}. \tag{5}$$

This theorem shows that any FTPL-variant where the perturbation can be described as a random sequence of outcomes has regret bounded by the two terms STABILITY and

ERROR. Below we will instantiate this theorem to obtain concrete regret bounds for several problems.

The proof of the theorem is based on a well-known "be-the-leader" argument. We first show that if we included the tth loss vector in the oracle call at round t, we would have regret bounded by ERROR, and then we show that the difference between our algorithm and this prescient one is bounded by STABILITY. See Appendix A for the proof.

3. Adversarial Contextual Learning

Our first specialization of the general setting is to *contextual online combinatorial optimization*. In this learning setting, the action space is a feasibility set $\mathcal{A} \subseteq \{0,1\}^K$ and we use $a \in \mathcal{A}$ both as a binary vector and as the set $\{j \in [K] : a(j) = 1\}$. The adversary picks a outcome $y^t = (x^t, f^t)$ where x^t belongs to some context space $\mathcal{X}$ and $f^t : \mathcal{A} \to \mathbb{R}$ is a cost function that maps each feasible action vector $a \in \mathcal{A}$ to a cost $f^t(a)$. The goal of the learning algorithm is to achieve low regret relative to a set of policies $\Pi \subset (\mathcal{X} \to \mathcal{A})$ that map contexts to feasible action vectors. At each iteration the algorithm picks a policy π^t and incurs a cost $\ell(\pi^t, y^t) = f^t(\pi^t(x^t))$. In this section, we consider the full-information problem, where after each round, the entire loss function f^t is revealed to the learner. Online versions of a number of important learning tasks, including cost-sensitive classification, multi-label prediction, online linear optimization (Kalai & Vempala, 2005) and online submodular minimization (Hazan & Kale, 2012) are all special cases of the contextual online combinatorial optimization problem, as we will see below.

Contextual Follow the Perturbed Leader. We will analyze the performance of an instantiation of the FTPL algorithm in this setting. To specialize the algorithm, we need only specify the distribution from which the sequence of fake outcomes $\{z\}$ is drawn at each time-step. This distribution is parameterized by a subset of contexts $X \subseteq \mathcal{X}$, with $|X| = d$ and a noise parameter ϵ. We draw the sequence $\{z\}$ as follows: for each context $x \in X$, we add the fake sample $z_x = (x, f_x)$ where f_x is a linear loss function based on a loss vector $\ell_x \in \mathbb{R}^K$, meaning that $f_x(a) = \langle a, \ell_x \rangle$. Each coordinate of the loss vector ℓ_x is drawn from a independent Laplace distribution with parameter ϵ, i.e. for each coordinate $j \in [K]$ the density of $\ell_x(j)$ at q is $f(q) = \frac{\epsilon}{2}\exp\{-\epsilon|q|\}$. The latter distribution has mean 0 and variance $\frac{2}{\epsilon^2}$. Using this distribution for fake samples gives an instantiation of Algorithm 1, which we refer to as CONTEXT-FTPL(X, ϵ) (see Algorithm 2).

We analyze CONTEXT-FTPL(X, ϵ) in two settings: the *transductive setting* and the *small separator* setting.

Definition 2. In the *transductive setting*, at the beginning of the learning process, the adversary reveals to the learner the set of contexts that will arrive, although the ordering and multiplicity need not be revealed.

Algorithm 2 Contextual Follow the Perturbed Leader Algorithm - CONTEXT-FTPL(X, ϵ).

Input: parameter ϵ, set of contexts X, policies Π.
for each time step t **do**
  Draw a sequence $\{z\} = (z_1, \ldots, z_d)$ of d fake samples.
  The context associated with sample z_x is equal to x and each coordinate of the loss vector ℓ_x is drawn i.i.d. from a Laplace(ϵ)
  Pick and play according to policy

$$\pi^t = M(\{z\} \cup y^{1:t-1}) \tag{6}$$

end for

Definition 3. In the *small separator* setting, there exists a set $X \subset \mathcal{X}$ such that for any two distinct policies $\pi, \pi' \in \Pi$, there exists $x \in X$ such that $\pi(x) \neq \pi'(x)$.

In the transductive setting, the set X that we use in CONTEXT-FTPL(X, ϵ) is precisely this set of contexts that will arrive, which by assumption is available to the learning algorithm. In this small separator setting, the set X used by CONTEXT-FTPL is the separating set. This enables non-transductive learning, but one must be able to compute a small separator prior to learning. Below we will see examples where this is possible.

We now turn to bounding the regret of CONTEXT-FTPL(X, ϵ). Let $d = |X|$ be the number of contexts that are used in the definition of the noise distribution, let $N = |\Pi|$, and let m denote the maximum number of non-zero coordinates that any policy can choose on any context, i.e. $m = \max_{a \in \mathcal{A}} \|a\|_1$. Even though at times we might constrain the sequence of loss functions that the adversary can pick (e.g. linear non-negative losses), we will assume that *the oracle M can handle at least linear loss functions with both positive and negative coordinates*. Our main result is:

Theorem 2 (Complete Information Regret)**.** CONTEXT-FTPL(X, ϵ) *achieves regret against any adaptively and adversarially chosen sequence of contexts and loss functions:*

1. *In the* transductive *setting:*

$$\text{REGRET} \leq 4\epsilon K \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|f^t\|_*^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

2. *In the* transductive *setting, when loss functions are* linear and non-negative*, i.e.* $f^t(a) = \langle a, \ell^t \rangle$ *with*

$\ell^t \in \mathbb{R}^K_{\geq 0}$:

$$\text{REGRET} \leq \epsilon \cdot \sum_{t=1}^{T} \mathbb{E}\left[\langle \pi^t(x^t), \ell^t \rangle^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

3. *In the* small separator *setting:*

$$\text{REGRET} \leq 4\epsilon K d \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|f^t\|_*^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

where $\|f^t\|_* = \max_{a \in \mathcal{A}} |f^t(a)|$.

When ϵ *is set optimally, loss functions are in* $[0,1]$*, and loss vectors are in* $[0,1]^K$*, these give regret:*[2] $O\left((dm)^{1/4}\sqrt{KT\log(N)}\right)$ *in the first setting,* $O\left(d^{1/4}m^{5/4}\sqrt{T\log(N)}\right)$ *in the second and* $O\left(m^{1/4}d^{3/4}\sqrt{KT\log(N)}\right)$ *in the third.*

To prove the theorem we separately upper bound the STABILITY and the ERROR terms and then Theorem 2 follows from Theorem 1. One key step is a refined ERROR analysis that leverages the symmetry of the Laplace distribution to obtain a bound with dependence $\sqrt{d}$ rather than d. This is possible only if the perturbation is centered about zero, and therefore does not apply to other FPTL variants that use non-negative distributions such as exponential or uniform (Kalai & Vempala, 2005). Due to lack of space we defer proof details to Appendix B.

This general theorem has implications for many specific settings that have been extensively studied in the literature. We turn now to some examples.

Example 1. (Transductive Contextual Experts) The contextual experts problem is the online version of cost-sensitive multiclass classification, and the full-information version of the widely-studied contextual bandit problem. The setting is as above, but $\mathcal{A}$ corresponds to sets with cardinality 1, meaning that $m = 1$ in our formulation. As a result, CONTEXT-FTPL can be applied as is, and the second claim in Theorem 2 shows that the algorithm has regret at most $O\left(d^{1/4}\sqrt{T\log(N)}\right)$ if at most d contexts arrive. In the worst case this bound is $O(T^{3/4}\sqrt{\log(N)})$, since the adversary can choose at most T contexts. To our knowledge, this is the first fully oracle-efficient algorithm for online adversarial cost-sensitive multiclass classification, albeit in the transductive setting.

This result can easily be lifted to infinite policy classes that have small Natarajan Dimension (a multi-class analog of VC-dimension), since such classes behave like finite ones once the set of contexts is fixed. Thus, in the transductive setting, Theorem 2 can be applied along with the analog of the Sauer-Shelah lemma, leading to a sublinear regret bound for classes with finite Natarajan dimension. On the other hand, in the non-transductive case it is possible to construct examples where achieving sublinear regret against a VC class is information-theoretically hard (Ben-David et al., 2009), demonstrating a significant difference between the two settings. See Corollary 15 and Theorem 16 in the Appendix F for details. ■

Example 2. (Non-contextual Shortest Path Routing and Linear Optimization) For the case when the linear optimization corresponds to computing the shortest (s,t)-path in a DAG, then K and m equal to the number of edges and the problem can be solved in poly-time even when edge costs are negative. More generally, CONTEXT-FTPL can also be applied to non-contextual problems, which is a special case where $d = 1$. In such a case, CONTEXT-FTPL reduces to the classical FTPL algorithm with Laplace instead of Exponential noise, and Theorem 2 matches existing results for online linear optimization (Kalai & Vempala, 2005). In particular, for problems without context, CONTEXT-FTPL has regret that scales with $\sqrt{T}$. ■

Example 3. (Online sub-modular minimization) A special case of our setting is the online-submodular minimization problem studied in previous work (Hazan & Kale, 2012; Jegelka & Bilmes, 2011). As above, this is a non-contextual online combinatorial optimization problem, where the loss function f^t presented at each round is submodular. Here, CONTEXT-FTPL reduces to the strongly polynomial algorithm of Hazan & Kale (2012), although our noise follows a Laplace instead of Uniform distribution. A straightforward application of the first claim of Theorem 2 shows that CONTEXT-FTPL achieves regret at most $O(KH\sqrt{T\log(K)})$ if the losses are bounded in $[-H,H]$, and a slightly refined analysis of the error terms gives $O(KH\sqrt{T})$ regret. This matches the FTPL analysis of Hazan & Kale (2012), although they also develop an algorithm based on online convex optimization that achieves $O(H\sqrt{KT})$ regret. ■

Example 4. (Contextual Experts with linear policy classes) The third clause of Theorem 2 gives strong guarantees for the non-transductive contextual experts problem, provided one can construct a small separating set of contexts. Often this is possible, and we provide some examples here.

1. For binary classification where the policies are boolean disjunctions (conjunctions) over n binary variables, the set of 1-sparse ($n-1$-sparse) boolean vectors form a separator of size n. This is easy to see as two disjunctions must disagree on at least one variable, so they will make different predictions on

[2] Observe that when loss vectors are in $[0,1]^K$, then the linear loss function is actually in $[0,m]$ not in $[0,1]$.

the vector that is non-zero only in that component. Note that the size of the small separator is independent of the time horizon T and logarithmic in the number of policies. Thus, Theorem 2 shows that CONTEXT-FTPL suffers at most $O(\sqrt{T}\log(N))$ regret since $d = \log(N), m = 1$ and $K = 2$.

2. For binary classification in n dimensions, consider a discretization of linear classifiers defined as follows, the separating hyperplane of each classifier is defined by choosing the intercept with each axis from one of $O(1/\tau)$ values (possibly including something denoting no intercept). Then a small separator includes, for each axis, one point between each pair in the discretization, for a total of $O(n/\tau)$ points. This follows since any two distinct classifiers have different intercepts for at least one axis, and our small separator has one point between these two different intercepts, leading to different predictions. Note that the number of classifiers in the discretization is $O(\tau^{-n})$. Here Theorem 2 shows that CONTEXT-FTPL suffers at most $O(\frac{n\sqrt{T}}{\tau^{3/4}}(\log(\frac{1}{\tau}))^{1/4})$ regret since $N = O(\tau^{-n}), d = \frac{n}{\tau}, m = 1$ and $K = 2$. This bound has a undesireable polynomial dependence on the discretization resolution τ but avoids exponential dimension dependence. Note that competing with the set of all linear classifiers (without discretization) is impossible because the class has infinite Littlestone dimension (Ben-David et al., 2009) (See also Theorem 16 in Appendix F).

Thus we believe that the smallest separator size for a policy class can be viewed as a new complexity measure, which may be of independent interest. ■

4. Linear Losses and Semi-Bandit Feedback

In this section, we consider contextual learning with semi-bandit feedback and linear non-negative losses. At each round t of this learning problem, the adversary chooses a non-negative vector $\ell^t \in \mathbb{R}^K_{\geq 0}$ and sets the loss function to $f^t(a) = \langle a, \ell^t \rangle$. The learner chooses an action $a^t \in \mathcal{A} \subset \{0,1\}^K$ accumulates loss $f^t(a^t)$ and observes $\ell^t(j)$ for each $j \in a^t$. In other words, the learner observes the coefficients for only the elements in the set that he picked. Notice that if $\mathcal{A}$ is the one-sparse vectors, then this setting is equivalent to the well-studied contextual bandit problem (Langford & Zhang, 2008).

Semi-bandit algorithm. Our semi-bandit algorithm proceeds as follows: At each iteration it makes a call to CONTEXT-FTPL(ϵ), which returns a policy π^t and implies a chosen action $a^t = \pi^t(x^t)$. The algorithm plays the action a^t, observes the coordinates of the loss $\{\ell^t(j)\}_{j \in a^t}$ and proceeds to construct a *proxy loss vector* $\hat{\ell}^t$, which it passes to the instance of CONTEXT-FTPL, before proceeding to the next round.

To describe the construction of $\hat{\ell}^t$, let $p^t(\pi) = \Pr[\pi^t = \pi | \mathcal{H}^{t-1}]$ denote the probability that CONTEXT-FTPL returns policy π at time-step t conditioned on the past history (observed losses and contexts, chosen actions, current iteration's context, internal randomness etc., which we denote with $\mathcal{H}^{t-1}$). For any element $j \in [K]$, let:

$$q^t(j) = \sum_{\pi \in \Pi : j \in \pi(x^t)} p^t(\pi) \tag{7}$$

denote the probability that element j is included in the action chosen by CONTEXT-FTPL(X, ϵ) at time-step t.

Typical semi-bandit algorithms aim to construct proxy loss vectors by dividing the observed coordinates of the loss by the probabilities $q^t(j)$ and setting other coordinates to zero, which is the well-known inverse propensity scoring mechanism (Horvitz & Thompson, 1952). Unfortunately, in our case, the probabilities $q^t(j)$ stem from randomness fed into the oracle, so that they are implicit maintained and therefore must be approximated.

We therefore construct $\hat{\ell}^t$ through a geometric sampling scheme due to Neu & Bartók (2013). For each $j \in \pi^t(x^t)$, we repeatedly invoke the current execution of the CONTEXT-FTPL algorithm with fresh noise, until it returns a policy that includes j in its action for context x^t. The process is repeated at most L times for each $j \in \pi^t(x^t)$ and the number of invocations is denoted $J^t(j)$. The vector $\hat{\ell}^t$ that is returned to the full feedback algorithm is zero for all $j \notin \pi^t(x^t)$, and for each $j \in \pi^t(x^t)$ it is $\hat{\ell}^t(j) = J^t(j) \cdot \ell^t(j)$.

By Lemma 1 of Neu & Bartók (2013), this process yields a proxy loss vector $\hat{\ell}^t$ that satisfies,

$$\mathbb{E}\left[\hat{\ell}^t(j) \mid \mathcal{H}^{t-1}\right] = \left(1 - \left(1 - q^t(j)\right)^L\right) \ell^t(j). \tag{8}$$

The semi-bandit algorithm feeds this proxy loss vector to the CONTEXT-FTPL instance and proceeds to the next round.

The formal description of the complete bandit algorithm is given in Algorithm 3 and we refer to it as CONTEXT-SEMI-BANDIT-FTPL(X, ϵ, L). We bound its regret in the transductive and small separator setting.

Theorem 3. *The expected regret of* CONTEXT-SEMI-BANDIT-FTPL(X, ϵ, L) *in the semi-bandit setting against any adaptively and adversarially chosen sequence of contexts and linear non-negative losses, with* $\|\ell^t\|_* \leq 1$*, is at most:*

Algorithm 3 Contextual Semi-Bandit Algorithm - CONTEXT-SEMI-BANDIT-FTPL(X, ϵ, L).

Input: parameter ϵ, M, set of contexts X, policies Π.
Let D denote a distribution over a sequence of d samples, $\{z\} = (z_1, \ldots, z_d)$, where the context associated with sample z_x is equal to x and each coordinate of the loss vector ℓ_x is drawn i.i.d. from a Laplace(ϵ)
for each time-step t **do**
  Draw a sequence $\{z\}^t$ from distribution D.
  Pick and play according to policy

$$\pi^t = M(\{z\} \cup (x^{1:t-1}, \hat{\ell}^{1:t-1})) \tag{9}$$

  Observe loss $\ell^t(j)$ for each $j \in \pi^t(x^t)$
  Set $\hat{\ell}^t(j) = 0$ for any $j \notin \pi^t(x^t)$
  Set $\hat{\ell}^t(j) = J^t(j) \cdot \ell^t(j)$, for each $j \in \pi^t(x^t)$, where $J^t(j)$ is computed by the following geometric sampling process:
  for each element $j \in \pi^t(x^t)$ **do**
    for each iteration $i = 1, \ldots, L$ **do**
      Draw a sequence $\{y\}^i$ from distribution D.
      Compute $\pi^i = M(\{y\}^i \cup (x^{1:t-1}, \hat{\ell}^{1:t-1}))$
      If $j \in \pi^i(x^t)$ then stop and return $J^t(j) = i$
    end for
  end for
  If process finished without setting $J^t(j)$, then set $J^t(j) = L$
end for

- *In the transductive setting:*

$$\text{REGRET} \leq 2\epsilon mKT + \frac{10}{\epsilon}\sqrt{dm}\log(N) + \frac{KT}{eL}$$

- *In the small separator setting:*

$$\text{REGRET} \leq 8\epsilon K^2 dLmT + \frac{10}{\epsilon}\sqrt{dm}\log(N) + \frac{KT}{eL}$$

For $L = \sqrt{KT}$ and optimal ϵ, the regret is $O\left(d^{1/4}m^{3/4}\sqrt{KT\log(N)}\right)$ in the first setting. For $L = T^{1/3}$ and optimal ϵ, the regret is $O\left((md)^{3/4}KT^{2/3}\sqrt{\log(N)}\right)$ in the second setting. Moreover, each iteration of the algorithm requires mL oracle calls and otherwise runs in polynomial time in d, K.

This is our main result for adversarial variants of the contextual bandit problem. In the most well-studied setting, i.e. contextual bandits, we have $m = 1$, so our regret bound is $O(d^{1/4}\sqrt{KT\log(N)})$ in the transductive setting and $O(d^{3/4}KT^{2/3}\sqrt{\log(N)})$ in the small separator setting. Since for the transductive case $d \leq T$ and for the small-separator case d can be independent of T (see discussion above), this implies sublinear regret for adversarial contextual bandits in either setting. To our knowledge this is the first oracle-efficient sublinear regret algorithm for variants of the contextual bandit problem. However, as we mentioned before, neither regret bound matches the optimal $O(\sqrt{KT\log(N)})$ rate for this problem, which can be achieved by computationally intractable algorithms. An interesting open question is to develop computationally efficient, statistically optimal contextual bandit algorithms.

5. Switching Policy Regret

In this section we analyze switching regret for the contextual linear optimization setting, i.e. when competing with the best sequence of policies that switches at most k times. Such a notion of regret was first analyzed by Herbster & Warmuth (1998) and several algorithms, that are not computationally efficient for large policy spaces, have been designed since then (e.g. (Luo & Schapire, 2015)). Our results provide the first computationally efficient switching regret algorithms assuming offline oracle access. Note that (György et al., 2012) study a similar setting but assume access to an online oracle that can achieve low regret against the best fixed policy. The offline oracle we consider is significantly weaker.

For this setting we will assume that *the learner knows the exact sequence $x^{1:T}$ of contexts ahead of time and not only the set of potential contexts.* The extension stems from the realization that we can simply think of time t as part of the context at time-step t. Thus now the contexts are of the form $\tilde{x}^t = (t, x^t)$. Moreover, policies in the augmented context space are now of the form: $\tilde{\pi}(\tilde{x}^t) = \pi_{I(t)}(x^t)$, where $I(t)$ is a selector which maps a time-step t to a policy index in $[N]$, with the constraint that the number of time-steps such that $I(t) \neq I(t-1)$ is at most k. If the original policy space Π was of size N, the new policy space, denoted $\tilde{\Pi}$, is of size $\tilde{N}$ at most $T^k N^k$, since there are at most T^k partitions of time into k consequetive intervals and each of the k intervals can be occupied by N possible policies. Moreover, in this augmented context space, the number of possible contexts, denoted $\tilde{X}$ is equal to $\tilde{d} = T$.

Thus if we run CONTEXT-FTPL(X, ϵ) on this augmented context and policy space, Theorem 2, bounds the regret against all policies in the augmented policy space $\tilde{\Pi}$. Since, regret against the augmented policy space, corresponds to switching regret against the original set of policies, the following corollary is immediate:

Corollary 4 (Contextual Switching Regret)**.** *In the transductive complete information setting,* CONTEXT-FTPL($\tilde{X}, \epsilon$) *applied to the augmented policy space $\tilde{\Pi}$, achieves k-switching regret against any adaptively and adversarially chosen sequence of contexts and*

losses at most: $O\left(m^{1/4}\sqrt{Kk\log(TN)}T^{3/4}\right)$ *for general loss functions in* $[0,1]$ *and* $O\left(\sqrt{k\log(TN)}m^{5/4}T^{3/4}\right)$ *for linear losses with loss vectors in* $[0,1]^K$.

It remains to show is that we can efficiently solve the offline optimization problem for the new policy space $\tilde{\Pi}$, if we have access to an optimization oracle for the original policy space Π. Then we can claim that CONTEXT-FTPL$(\tilde{X},\epsilon)$ in the augmented context and policy space is also an efficient algorithm. We show that the latter is true via a dynamic programming approach. The approach generalizes beyond contextual linear optimization settings. The proof of the Lemma is provided in the supplementary material.

Lemma 5. *The oracle* $\tilde{M}$ *in the augmented space,*

$$\tilde{M}(\tilde{y}^{1:T}) = \text{argmin}_{\tilde{\pi}\in\tilde{\Pi}} \sum_{\tau=1}^{T}\langle\tilde{\pi}(\tau,x_\tau),\ell^\tau\rangle \qquad (10)$$

is computable in $O(Tk)$ *time, with* $O(T^2)$ *calls to the oracle over the original space,* M*. This process can be amortized so that solving a sequence of* T *problems in the augmented space requires* $O(T^2)$ *calls to* M *in total.*

6. Efficient Path Length Regret Bounds

In this section we examine a variant of our CONTEXT-FTPL(ϵ) algorithm that is efficient and achieves regret that is upper bounded by structural properties of the loss sequence. Our algorithm is framed in terms of a generic predictor that the learner has access to and the regret is upper bounded by the deviation of the true loss vector from the predictor. For specific instances of the predictor this leads to path length bounds (Chiang et al., 2012) or variance based bounds (Hazan & Kale, 2010). Our approach allows for generalizations of variance and path length that can incorporate contextual information and can be viewed as an efficient version and a generalization of the results of Rakhlin & Sridharan (2013b) on learning with predictable sequences. Such results have also found applications in learning in game theoretic environments (Rakhlin & Sridharan, 2013a; Syrgkanis et al., 2015).

The algorithm is identical to CONTEXT-FTPL(ϵ) with the exception that now the policy that is used at time-step t is:

$$\pi^t = M(\{z\}\cup y^{1:t-1}\cup(x^t,Q^t)) \qquad (11)$$

where $Q^t \in \{0,1\}^K \to \mathbb{R}^K$ is an arbitrary loss function predictor, which can depend on the observed history up to time t. This predictor can be interpreted as partial side information that the learner has about the loss function that will arrive at time-step t. Given such a predictor we define the error between the predictor and the actual sequence:

$$\mathcal{E}^t = \mathbb{E}\left[\|f^t - Q^t\|_*^2\right] \qquad (12)$$

Theorem 6 (Predictor based regret bounds)**.** *The regret of* CONTEXT-FTPL(X,ϵ) *with predictors and complete information,*

1. *In the transductive setting is upper bounded by:*

$$\text{REGRET} \le 4\epsilon K\sum_{t=1}^{T}\mathcal{E}^t + \frac{10\sqrt{dm}\log(N)}{\epsilon}$$

2. *In the small separator setting is upper bounded by:*

$$\text{REGRET} \le 4\epsilon Kd\sum_{t=1}^{T}\mathcal{E}^t + \frac{10\sqrt{dm}\log(N)}{\epsilon}$$

Picking ϵ *optimally gives regret* $O\left((dm)^{1/4}\sqrt{K\log(N)\sum_{t=1}^{T}\mathcal{E}^t}\right)$ *in the first setting and* $O\left(m^{1/4}d^{3/4}\sqrt{K\log(N)\sum_{t=1}^{T}\mathcal{E}^t}\right)$ *in the second.*

Even without contexts, our result is the first efficient path length regret algorithm for online combinatorial optimization. For instance, for the case of non-contextual, online combinatorial optimization an instantiation of our algorithm achieves regret $O\left(m^{1/4}\sqrt{K\log(K)\sum_{t=1}^{T}\mathcal{E}^t}\right)$ against adaptive adversaries. For learning with expert advice, $m=1$ and K is number of experts, the results of Rakhlin & Sridharan (2013b) provide a non-efficient $O\left(\sqrt{\log(K)\sum_{t=1}^{T}\mathcal{E}^t}\right)$. Thus our bound incurs an extra cost of $\sqrt{K}$ in comparison. Removing this extra factor of $\sqrt{K}$ in an efficient manner is an interesting open question.

7. Discussion

In this work we give fully oracle efficient algorithms for adversarial online learning problems including contextual experts, contextual bandits, and problems involving linear optimization or switching experts. Our main algorithmic contribution is a new Follow-The-Perturbed-Leader style algorithm that adds perturbed low-dimensional statistics. Our analysis for this algorithm guarantees sublinear regret against adaptive adversaries for all of these problems.

While our algorithms achieve sublinear regret in all problems we consider, we do not always attain the optimal regret bounds. An interesting direction for future work is whether fully oracle-based algorithms can achieve optimal regret bounds in the settings we consider. Another interesting direction focuses on a deeper understanding of the small-separator condition and whether it enables efficient non-transductive learning in other settings. We look forward to studying these questions in future work.

Acknowledgements

We thank Jacob Abernethy and Alekh Agarwal for insightful formative discussions and Gergely Neu for providing detailed feedback on an early draft of this paper.

References

Agarwal, Alekh, Hsu, Daniel, Kale, Satyen, Langford, John, Li, Lihong, and Schapire, Robert E. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2014.

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. Gambling in a rigged casino: The adversarial multi-armed bandit pproblem. In *Foundations of Computer Science (FOCS)*, 1995.

Awerbuch, Baruch and Kleinberg, Robert. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 2008.

Ben-David, Shai, Cesa-Bianchi, Nicolo, Haussler, David, and Long, Philip M. Characterizations of learnability for classes of (0,..., n)-valued functions. *Journal of Computer and System Sciences*, 1995.

Ben-David, Shai, Kushilevitz, Eyal, and Mansour, Yishay. Online learning versus offline learning. *Machine Learning*, 1997.

Ben-David, Shai, Pál, Dávid, and Shalev-Shwartz, Shai. Agnostic online learning. In *COLT*, 2009.

Cesa-Bianchi, Nicolo and Shamir, Ohad. Efficient online learning via randomized rounding. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Cesa-Bianchi, Nicolo, Freund, Yoav, Haussler, David, Helmbold, David P, Schapire, Robert E, and Warmuth, Manfred K. How to use expert advice. *Journal of the ACM (JACM)*, 1997.

Chiang, Chao-Kai, Yang, Tianbao, Lee, Chia-Jung, Mahdavi, Mehrdad, Lu, Chi-Jen, Jin, Rong, and Zhu, Shenghuo. Online optimization with gradual variations. In *Conference on Learning Theory (COLT)*, 2012.

Daskalakis, Constantinos and Syrgkanis, Vasilis. Learning in auctions: Regret is hard, envy is easy. *arXiv:1511.01411*, 2015.

Dudík, Miroslav, Hsu, Daniel, Kale, Satyen, Karampatziakis, Nikos, Langford, John, Reyzin, Lev, and Zhang, Tong. Efficient optimal learning for contextual bandits. In *Uncertainty and Artificial Intelligence (UAI)*, 2011.

Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

György, András, Linder, Tamás, and Lugosi, Gábor. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 2012.

Haussler, David and Long, Philip M. A generalization of sauer's lemma. *Journal of Combinatorial Theory*, 1995.

Hazan, Elad and Kale, Satyen. Extracting certainty from uncertainty: regret bounded by variation in costs. *Machine Learning*, 2010.

Hazan, Elad and Kale, Satyen. Online submodular minimization. *Journal of Machine Learning Research (JMLR)*, 2012.

Herbster, Mark and Warmuth, Manfred K. Tracking the best expert. *Machine Learning*, 1998.

Horvitz, Daniel G and Thompson, Donovan J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association (JASA)*, 1952.

Hutter, Marcus and Poland, Jan. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research (JMLR)*, 2005.

Jegelka, Stefanie and Bilmes, Jeff A. Online submodular minimization for combinatorial structures. In *International Conference on Machine Learning (ICML)*, 2011.

Kakade, Sham M and Kalai, Adam. From batch to transductive online learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Kalai, Adam and Vempala, Santosh. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 2005.

Langford, John and Zhang, Tong. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Luo, Haipeng and Schapire, Robert E. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory (COLT)*, 2015.

Neu, Gergely and Bartók, Gábor. An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory (ALT)*, 2013.

Rakhlin, Alexander and Sridharan, Karthik. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems (NIPS)*, 2013a.

Rakhlin, Alexander and Sridharan, Karthik. Online learning with predictable sequences. In *Conference on Learning Theorem (COLT)*, 2013b.

Rakhlin, Alexander and Sridharan, Karthik. BISTRO: an efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2016.

Syrgkanis, Vasilis, Agarwal, Alekh, Luo, Haipeng, and Schapire, Robert E. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.