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 in 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 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 much more challenging bandit version, only the loss that was actually incurred for the chosen action is observed. 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 at each step. 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 great many applications which call for an extremely large policy space. In this paper, we address this gap between the statistical promise and computational challenge of algorithms for 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 an associated 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

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 by proposing and analyzing in Section 2 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 set of 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 study these for a generalized online learning problem 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, assuming access to an optimization oracle. Improving these bounds without sacrificing computational efficiency 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).

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 previous 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 in the present paper.

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.

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

[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.

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 utility 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.

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.

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

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,

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

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 foreseeing 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, at each iteration, the learning algorithm picks a binary

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

action vector $a^t \in \mathcal{A} \subseteq \{0,1\}^K$, from some feasibility set $\mathcal{A}$. We will interchangeably use a^t both as a vector and as the set $\{j \in [K] : a^t(j) = 1\}$. 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.

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 $N = |\Pi| \leq d^K$, 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, demonstrating a significant difference between the two settings. See Corollary 15 and Theorem 16 in the Appendix E for details on these arguments. ■

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 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

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

$O(1/\tau)$ 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.

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 an *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:*

- *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*.*

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

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. regret that compares to 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.

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 $\pi \in \Pi$, 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.

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

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

is computable in $O(Tk)$ time, with $O(T^2)$ calls to the oracle over the original space, M. This process can be amor-

tized so that solving a sequence of T problems in the augmented space requires $O(T^2)$ calls to M in total.

Proof. Oracle $\tilde{M}$ must compute the best sequence of policies $\pi^1, \ldots, \pi^T$, such that $\pi^t \neq \pi^{t-1}$ at most k times. Let $R(t, q)$ denote the loss of the optimal sequence of policies up to time-step t and with at most q switches. Then it is easy to see that:

$$R(t,q) = \min_{\tau \leq t} R(\tau, q-1) + \mathcal{L}\left(M(y^{\tau+1:t}), y^{\tau+1:t}\right), \quad (11)$$

i.e. compute the best sequence of policies up till some time step $\tau \leq t$ with at most $q-1$ switches and then augment it with the optimal fixed policy for the period $(\tau+1, t)$. Then take the best over possible times $\tau \leq t$.

This can be implemented by first invoking oracle M for every possible period $[\tau_1, \tau_2]$. Then filling up iteratively all the entries $R(t, q)$. For $q = 0$, the problem $R(t, 0)$ corresponds to exactly the original oracle problem M, hence for each t, we can solve the problem $R(t, 0)$. Computing all values of $R(t, q)$ then takes time Tk in total. ■

Example 5. (Efficient switching regret for non-contextual problems) When the original space has no contexts, our result above implies the first efficient sub-linear switching regret algorithm for online linear optimization. In this case, the transductivity assumption is trivially satisfied as there is no contextual information, and our the instance of CONTEXT-FTPL runs on a sequence of contexts that just encode time. One concrete example where linear optimization with both positive and negative weights is polynomially solvable is the online shortest path problem on a directed acyclic graph. Our result implies a fully efficient, sublinear switching regret algorithm for the online shortest-path problem on a DAG, and our algorithm performs t shortest-path computations at the tth iteration. The result also covers other examples, such as online matroid optimization. ■

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 utility 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 is general enough to allow 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:

$$\pi^t = M(\{z\} \cup y^{1:t-1} \cup (x^t, Q^t)) \quad (12)$$

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 can define the error between the predictor and the actual sequence:

$$\mathcal{E}^t = \mathbb{E}\left[\|f^t - Q^t\|_*^2\right] \quad (13)$$

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:*

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

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

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

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, $m = 1$ 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. We give a refined analysis for this algorithm that guarantees sublinear regret for all of these problems. All of our results hold against adaptive adversaries, both with full and partial feedback.

While our algorithms achieve sublinear regret in all problems we consider, we do not always match the regret bounds attainable by inefficient alternatives. 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.

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.

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.

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)*, pp. 3066–3074, 2013a.

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

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.

Supplementary material for "Efficient Algorithms for Adversarial Contextual Learning"

A. Omitted Proofs from Section 2

A.1. Proof of Theorem 1

We prove the theorem by analyzing a slightly modified algorithm, that only draws the perturbation once at the beginning of the learning process but is otherwise identical. The bulk of the proof is devoted to bounding this modified algorithm's regret against oblivious adversaries, i.e. an adversary that chooses the outcomes $y^{1:T}$ before the learning process begins. We use this regret bound along with a reduction due to Hutter and Poland (Hutter & Poland, 2005) (see their Lemma 12) to obtain a regret bound for Algorithm 1 against adaptive adversaries. We provide a proof of this reduction in Appendix A.2 and proceed here with the analysis of the modified algorithm.

To bound the regret of the modified algorithm, consider letting the algorithm observe y^t ahead of time, so that at each time step t, the algorithm plays $\pi^{t+1} = M\left(\{z\} \cup y^{1:t}\right)$. Notice trivially that the regret of the modified algorithm is,

$$\textsc{Regret} = \sum_{t=1}^{T} \ell(\pi^t, y^t) - \min_{\pi\in\Pi} \ell(\pi^*, y^t) = \sum_{t=1}^{T} \ell(\pi^t, y^t) - \ell(\pi^{t+1}, y^t) + \sum_{t=1}^{T} \ell(\pi^{t+1}, y^t) - \min_{\pi\in\Pi} \ell(\pi^*, y^t)$$

The first sum here is precisely the STABILITY term in the bound, so we must show that the second sum is bounded by ERROR. This is proved by induction in the following lemma.

Lemma 7 (Be-the-leader with fixed sample perturbations)**.** *For any realization of the sample sequence $\{z\}$ and for any policy π^*:*

$$\sum_{t=1}^{T} \left(\ell(\pi^{t+1}, y^t) - \ell(\pi^*, y^t)\right) \le \max_{\pi\in\Pi} \sum_{z^\tau\in\{z\}} \ell(\pi, z^\tau) - \min_{\pi\in\Pi} \sum_{z^\tau\in\{z\}} \ell(\pi, z^\tau) \tag{14}$$

Proof. Denote with k the length of sequence $\{z\}$. Consider the sequence $\{z\} \cup y^{1:T}$ and let $a^1 = M(\{z\})$. We will show that for any policy π^*:

$$\sum_{\tau=1}^{k} \ell(\pi^1, z^\tau) + \sum_{t=1}^{T} \ell(\pi^{t+1}, y^t) \le \sum_{\tau=1}^{k} \ell(\pi^*, z^\tau) + \sum_{t=1}^{T} \ell(\pi^*, y^t) \tag{15}$$

For $T = 0$, the latter trivially holds by the definition of a^1. Suppose it holds for some T, we will show that it holds for $T + 1$. Since the induction hypothesis holds for any π^*, applying it for a^{T+2}, i.e.,:

$$\begin{aligned}\sum_{\tau=1}^{k} \ell(\pi^1, z^\tau) + \sum_{t=1}^{T+1} \ell(\pi^{t+1}, y^t) \le & \sum_{\tau=1}^{k} \ell(\pi^{T+2}, z^\tau) + \sum_{t=1}^{T} \ell(\pi^{T+2}, y^t) + \ell(\pi^{T+2}, y^{T+1}) \\ = & \sum_{\tau=1}^{k} \ell(\pi^{T+2}, z^\tau) + \sum_{t=1}^{T+1} \ell(\pi^{T+2}, y^t)\end{aligned}$$

By definition of a^{T+2} the latter is at most: $\sum_{\tau=1}^{k} \ell(\pi^*, z^\tau) + \sum_{t=1}^{T+1} \ell(\pi^*, y^t)$ for any π^*. Which proves the induction step. Thus, by re-arranging Equation (15) we get:

$$\sum_{t=1}^{T} \left(\ell(\pi^{t+1}, y^t) - \ell(\pi^*, y^t)\right) \le \sum_{\tau=1}^{k} \left(\ell(\pi^*, z^\tau) - \ell(\pi^1, z^\tau)\right) \le \max_{\pi\in\Pi} \sum_{\tau=1}^{k} \ell(\pi, z^\tau) - \min_{\pi\in\Pi} \sum_{\tau=1}^{k} \ell(\pi, z^\tau)$$

■

Thus the regret of the modified algorithm against an oblivious adversary is bounded by STABILITY + ERROR. By applying the reduction of Hutter and Poland (Hutter & Poland, 2005) (see Appendix A.2 for a proof sketch), the regret of Algorithm 1 is bounded is bounded in the same way.

A.2. From adaptive to oblivious adversaries

We will utilize a generic reduction provided in Lemma 12 of (Hutter & Poland, 2005), which states that given that in Algorithm 1 we draw independent randomization at each iteration, it suffices to provide a regret bound only for oblivious adversaries, i.e., the adversary picks a fixed sequence $y^{1:T}$ ahead of time without observing the policies of the player. Moreover, for any such fixed sequence of an oblivious adversary, the expected utility of the algorithm can be easily shown to be equal to the expected utility if we draw a single random sequence $\{z\}$ ahead of time and use the same random vector all the time.

The proof is as follows: by linearity of expectation and the fact that each sequence $\{z\}^t$ drawn at each time-step t is identically distributed:

$$\begin{aligned}\mathbb{E}_{\{z\}^1,\ldots,\{z\}^t}\left[\sum_{t=1}^{T} u(M(\{z\}^t \cup y^{1:t-1}), y^t)\right] &= \sum_{t=1}^{T} \mathbb{E}_{\{z\}^t}\left[u(M(\{z\}^t \cup y^{1:t-1}), y^t)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\{z\}^1}\left[u(M(\{z\}^1 \cup y^{1:t-1}), y^t)\right] \\ &= \mathbb{E}_{\{z\}^1}\left[\sum_{t=1}^{T} u(M(\{z\}^1 \cup y^{1:t-1}), y^t)\right]\end{aligned}$$

The latter is equivalent to the expected reward if we draw a single random sequence $\{z\}$ ahead of time and use the same random vector all the time. Thus it is sufficient to upper bound the regret of this modified algorithm, which draws randomness only once.

Thus it is sufficient to upper bound the regret of this modified algorithm, which draws randomness only once.

B. Omitted Proofs from Section 3

B.1. Bounding the Laplacian Error

The upper bound on the ERROR term is identical in all settings, since it only depends on the input noise distribution, which is the same for all variants and for which it does not matter whether X is the set of contexts that will arrive or a separator. In subsequent sections we will upper bound the stability of the algorithm in each setting.

Lemma 8 (Laplacian Error Bound)**.** *Let $\{z\}$ denote a sample from the random sequence of fake samples used by* CONTEXT-FTPL(X, ϵ)*. Then:*

$$\text{ERROR} = \mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle\right] - \mathbb{E}_{\{z\}}\left[\min_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle\right] \leq \frac{10}{\epsilon}\sqrt{dm}\log(N) \tag{16}$$

Proof. First we start by observing that each random variable $\ell_x(j)$ is distributed i.i.d. according to a Laplace(ϵ) distribution. Since a Laplace distribution is symmetric around 0, we get that $\ell_x(j)$ and $-\ell_x(j)$ are distributed identically. Thus we can write:

$$\mathbb{E}_{\{z\}}\left[\min_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle\right] = \mathbb{E}_{\{z\}}\left[\min_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), -\ell_x\rangle\right] = -\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle\right]$$

Hence we get:

$$\text{ERROR} = 2\cdot\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle\right] \tag{17}$$

We now bound the latter expectation via a moment generating function approach. For any $\lambda \geq 0$:

$$\begin{aligned}\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right] &= \frac{1}{\lambda}\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\lambda\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right]\\ &= \frac{1}{\lambda}\log\left\{\exp\left\{\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\lambda\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right]\right\}\right\}\end{aligned}$$

By convexity and monotonicity of the exponential function:

$$\begin{aligned}\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right] &\leq \frac{1}{\lambda}\log\left\{\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\exp\left\{\lambda\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right\}\right]\right\}\\ &\leq \frac{1}{\lambda}\log\left\{\sum_{\pi\in\Pi}\mathbb{E}_{\{z\}}\left[\exp\left\{\lambda\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right\}\right]\right\}\\ &\leq \frac{1}{\lambda}\log\left\{\sum_{\pi\in\Pi}\prod_{x\in X}\mathbb{E}\left[\exp\left\{\lambda\langle\pi(x),\ell_x\rangle\right\}\right]\right\}\\ &= \frac{1}{\lambda}\log\left\{\sum_{\pi\in\Pi}\prod_{x\in X}\mathbb{E}\left[\exp\left\{\lambda\sum_{j:\pi(x)(j)=1}\ell_x(j)\right\}\right]\right\}\\ &= \frac{1}{\lambda}\log\left\{\sum_{\pi\in\Pi}\prod_{x\in X}\prod_{j:\pi(x)(j)=1}\mathbb{E}\left[\exp\left\{\lambda\ell_x(j)\right\}\right]\right\}\end{aligned}$$

For any $j \in [K]$ and $x \in X$, $\ell_x(j)$ is a Laplace(ϵ) random variable. Hence, the quantity $\mathbb{E}\left[\exp\{\lambda\ell_x(j)\}\right]$ is the moment generating function of the Laplacian distribution evaluated at λ, which is equal to $\frac{1}{1-\frac{\lambda^2}{\epsilon^2}}$ provided that $\lambda < \epsilon$. Since $\sup_{x,\pi}|\{j\in[K]:\pi(x)(j)\}| \leq m$, we get:

$$\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right] \leq \frac{1}{\lambda}\log\left\{N\left(\frac{1}{1-\frac{\lambda^2}{\epsilon^2}}\right)^{dm}\right\} = \frac{1}{\lambda}\log(N) + \frac{dm}{\lambda}\log\left(\frac{1}{1-\frac{\lambda^2}{\epsilon^2}}\right)$$

By simple calculus, it is easy to derive that $\frac{1}{1-x} \leq e^{2x}$ for any $x \leq \frac{1}{4}$.[3] Thus as long as we pick $\lambda \leq \frac{\epsilon}{2}$, we get:

$$\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right] \leq \frac{1}{\lambda}\log(N) + \frac{dm}{\lambda}\log\left(\exp\left\{\frac{\lambda^2}{\epsilon^2}\right\}\right) = \frac{1}{\lambda}\log(N) + \frac{2dm\lambda}{\epsilon^2}$$

Picking $\lambda = \frac{\epsilon}{2\sqrt{dm}}$ and since $N \geq 2$:

$$\mathbb{E}_{\{z\}}\left[\max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x),\ell_x\rangle\right] \leq \frac{2\sqrt{dm}\log(N)}{\epsilon} + \frac{\sqrt{2dm}}{\epsilon} \leq \frac{5\sqrt{dm}\log(N)}{\epsilon}$$

■

B.2. Bounding Stability: Transductive Setting

We now turn to bounding the stability in the transductive combinatorial optimization setting. Combining the following lemma with the error bound in Lemma 8 and applying Theorem 1 proves the first claim of Theorem 2.

[3] Consider the function $f(x) = (1-x)e^{2x} - 1$. Then $f(0) = 0$ and $f'(x) = e^{2x}(1-2x)$, which is ≥ 0 for $0 \leq x \leq 1/2$.

Lemma 9 (Transductive Stability)**.** *For all $t \in [T]$ and for any sequence $y^{1:t}$ of contexts $x^{1:t}$ and loss functions $f^{1:t}$ with $f^i : \{0,1\}^K \to \mathbb{R}^K$, the stability of* CONTEXT-FTPL(X, ϵ) *is upper bounded by:*

$$\mathbb{E}_{\{z\}}\left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t))\right] \leq 4\epsilon K \cdot \|f^t\|_*^2$$

Proof. By the definition of $\|f^t\|_*$:

$$\mathbb{E}_{\{z\}}\left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t))\right] \leq 2\|f^t\|_* \Pr\left[\pi^t(x^t) \neq \pi^{t+1}(x^t)\right]$$

Now observe that:

$$\Pr\left[\pi^t(x^t) \neq \pi^{t+1}(x^t)\right] \leq \sum_{j \in K}\left(\Pr[j \in \pi^t(x^t), j \notin \pi^{t+1}(x^t)] + \Pr[j \notin \pi^t(x^t), j \in \pi^t(x^t)]\right)$$

We bound the probability $\Pr[j \in \pi^t(x^t), j \notin \pi^{t+1}(x^t)]$. We condition on all random variables of $\{z\}$ except for the random variable $\ell_{x^t}(j)$, i.e. the random loss placed at coordinate j on the sample associated with context x^t. Denote the event corresponding to an assignment of all these other random variables as $\mathcal{E}_{-x^t j}$. Let $\ell_{x^t j}$ denote a loss vector which is $\ell_{x^t}(j)$ on the j-th coordinate and zero otherwise. Also let:

$$\Phi(\pi) = \sum_{\tau=1}^{t-1} f^\tau(\pi(x^\tau)) + \sum_{x \in X - \{x^t\}} \langle \pi(x), \ell_x \rangle + \langle \pi(x^t), \ell_{x^t} - \ell_{x^t j} \rangle \tag{18}$$

Let $\pi^* = \operatorname{argmin}_{\pi \in \Pi : j \in \pi(x^t)} \Phi(\pi)$ and $\tilde{\pi} = \min_{\pi \in \Pi : j \notin \pi(x^t)} \Phi(\pi)$. The event that $\{j \in \pi^t(x^t)\}$ happens only if:

$$\Phi(\pi^*) + \ell_{x^t}(j) \leq \Phi(\tilde{\pi}) \tag{19}$$

Let and $\nu = \Phi(\tilde{\pi}) - \Phi(\pi^*)$. Thus $j \in \pi^t(x^t)$ only if:

$$\ell_{x^t}(j) \leq \nu \tag{20}$$

Now if:

$$\ell_{x^t}(j) < \nu - 2\|f^t\|_* \tag{21}$$

then it is easy to see that $\{j \in \pi^{t+1}(x^t)\}$, since an extra loss of $f^t(a) \in [0,1]$ cannot push j out of the optimal solution. More elaborately, for any other policy $\pi \in \Pi$, such that $j \notin \pi(x^t)$, the loss of π^* including time-step t is bounded as:

$$\begin{aligned}
\Phi(\pi^*) + \ell_{x^t}(j) + f^t(\pi^*(x^t)) &< \Phi(\pi) - 2\|f^t\|_* + f^t(\pi^*(x^t)) \\
&< \Phi(\pi) - \|f^t\|_* \\
&< \Phi(\pi) + f^t(\pi(x^t))
\end{aligned}$$

Thus any policy π, such that $j \notin \pi(x^t)$ is suboptimal after seeing the loss at time-step t. Thus

$$\Pr[j \in \pi^t(x^t), j \notin \pi^{t+1}(x^t) \mid \mathcal{E}_{-x^t j}] \leq \Pr[\ell_{x^t}(j) \in [\nu - 2\|f^t\|_*, \nu] \mid \mathcal{E}_{-x^t j}]$$

Since all other random variables are independent of $\ell_{x^t}(j)$ and $\ell_{x^t}(j)$ is a Laplacian with parameter ϵ:

$$\begin{aligned}
\Pr[\ell_{x^t}(j) \in [\nu - 2\|f^t\|_*, \nu] \mid \mathcal{E}_{-x^t j}] &= \Pr[\ell_{x^t}(j) \in [\nu - 2\|f^t\|_*, \nu]] \\
&= \frac{\epsilon}{2} \int_{\nu - 2\|f^t\|_*}^{\nu} e^{-\epsilon|z|} dz \leq \frac{\epsilon}{2} \int_{\nu - 2\|f^t\|_*}^{\nu} dz \leq \epsilon \|f^t\|_*
\end{aligned}$$

Similarly it follows that that: $\Pr[j \notin \pi^t(x^t)$ and $j \in \pi^{t+1}(x^t)] \leq \epsilon\|f^t\|_*$. To sum we get that:

$$\mathbb{E}_{\{z\}}\left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t))\right] \leq 2\|f^t\|_* \Pr\left[\pi^t(x^t) \neq \pi^{t+1}(x^t)\right] \leq 4\epsilon K \|f^t\|_*^2$$

■

B.3. Bounding Stability: Transductive Setting with Linear Losses

In the transductive setting with linear losses, we provide a significantly more refined stability bound, which enables applications to partial information or bandit settings. As before, combining this stability bound with the error bound in Lemma 8 and applying Theorem 1 gives the second claim of Theorem 2.

Lemma 10 (Multiplicative Stability)**.** *For any sequence $y^{1:T}$ for all $t \in [T]$ of contexts and non-negative linear loss functions, the stability of* CONTEXT-FTPL(X, ϵ) *in the transductive setting, is upper bounded by:*

$$\mathbb{E}_{\{z\}}\left[\langle \pi^t(x^t), \ell^t\rangle - \langle \pi^{t+1}(x^t), \ell^t\rangle\right] \leq \epsilon \cdot \mathbb{E}\left[\langle \pi^t(x^t), \ell^t\rangle^2\right]$$

Proof. To prove the result we first must introduce some additional terminology. For a sequence of parameters $y^{1:t}$, let $\phi^t \in \mathbb{R}^{dK}$ be a vector with $\phi^t_{x,j} = \sum_{\tau \leq t: x^\tau = x} \ell^\tau(j)$. The component of this vector corresponding to context $x \in X$ and coordinate $j \in [K]$ is the cumulative loss associated with that coordinate on the subset of time points when context x appeared. Note that this vector ϕ^t is a sufficient statistic, since for any fixed policy π:

$$\sum_{\tau=1}^{t} \ell(\pi, y^\tau) = \sum_{x \in X} \sum_{\tau \leq t: x^\tau = x} \langle \pi(x), \ell^\tau\rangle = \sum_{x \in X} \langle \pi(x), \phi^t_x\rangle \tag{22}$$

where $\phi^t_x = \sum_{\tau \leq t: x^\tau = x} \ell^\tau$.

We denote with $z \in \mathbb{R}^{dK}$ the sufficient statistic that corresponds to the fake sample sequence $\{z\}$ and with ϕ^t the sufficient statistics for the parameter sequence $y^{1:t}$. Observe that the sufficient statistic for the augmented sequence $\{z\} \cup y^{1:t}$ is simply $z + \phi^t$. For any sequence of parameters $y^{1:T}$ we will be denoting with $\phi^{1:T}$ the sequence of $d \cdot K$ dimensional cumulative loss vectors. We will also overload notation and denote with $M(\phi^t) = M(y^{1:t})$ the best policy on a sequence $y^{1:t}$ with statistics ϕ^t.

Consider a specific sequence $y^{1:T}$ and a specific time step t. Define, for each $\pi \in \Pi$, a sparse tuple $y^t_\pi = (x^t, \ell^t_\pi)$ where $\ell^t_\pi(j) = \ell^t(j)$ if $\pi(x^t)(j) = 1$ and zero otherwise, i.e. we zero out coordinates of the true loss vector that were not picked by the policy π. Moreover, define with ϕ^t_π the sufficient statistic of the sequence $\phi(y^{1:t-1} \cup y^t_\pi)$ for each π. We define $1 + |\Pi|$ distributions over $|\Pi|$, via their probability density functions, as follows:

$$\begin{aligned} p^t(\pi) &= \Pr[M(z + \phi^{t-1}) = \pi] \\ \forall \pi^* \in \Pi: p^{t+1}_{\pi^*}(\pi) &= \Pr\left[M(z + \phi^t_{\pi^*}) = \pi\right] \end{aligned}$$

At the end of this proof, we will show that $p^{t+1}_\pi(\pi) \leq p^{t+1}(\pi)$. Moreover, we denote for convenience:

$$\begin{aligned} \texttt{FTPL}^t &= \mathbb{E}_z[\langle \pi^t(x^t), \ell^t\rangle] = \mathbb{E}_{\pi \sim p^t}\left[\langle \pi(x^t), \ell^t\rangle\right] \\ \texttt{BTPL}^t &= \mathbb{E}_z[\langle \pi^{t+1}(x^t), \ell^t\rangle] = \mathbb{E}_{\pi \sim p^{t+1}}\left[\langle \pi(x^t), \ell^t\rangle\right] \end{aligned}$$

We will construct a mapping $\mu_\pi : \mathbb{R}^{dK} \to \mathbb{R}^{dK}$ such that for any $z \in \mathbb{R}^{dK}$,

$$M(z + \phi^t_\pi) = M(\mu_\pi(z) + \phi^{t-1})$$

Notice that $\mu_\pi(z) = z + \phi^t_\pi - \phi^{t-1}$. Now,

$$\begin{aligned} p^t(\pi) &= \int_z \mathbf{1}[\pi = M(z + \phi^{t-1})] f(z) dz \\ &= \int_z \mathbf{1}[\pi = M(\mu_\pi(z) + \phi^{t-1})] f(\mu_\pi(z)) dz \\ &= \int_z \mathbf{1}[\pi = M(z + \phi^t_\pi)] f(\mu_\pi(z)) dz \end{aligned}$$

Now observe that for any $z \in \mathbb{R}^{dK}$:

$$\begin{aligned} f(\mu_\pi(z)) &= \exp\{-\epsilon\left(\|z + \phi^t_\pi - \phi^{t-1}\|_1 - \|z\|_1\right)\} f(z) \\ &\leq \exp\{-\epsilon\left(\|z + \phi^t_\pi - \phi^{t-1}\|_1 - \|z + \phi^t_\pi - \phi^{t-1}\|_1 - \|\phi^{t-1} - \phi^t_\pi\|_1\right)\} f(z) \\ &\leq \exp\{\epsilon \|\phi^t_\pi - \phi^{t-1}\|_1\} f(z) \\ &= \exp\{\epsilon \langle \pi(x^t), \ell^t\rangle\} f(z) \end{aligned}$$

Substituting in this bound, we have,

$$p^t(\pi) \leq \exp\{\epsilon\langle\pi(x^t),\ell^t\rangle\}\cdot p_\pi^{t+1}(\pi) \leq \exp\{\epsilon\langle\pi(x^t),\ell^t\rangle\}\cdot p^{t+1}(\pi)$$

Re-arranging and lower bounding $\exp\{-x\} \geq (1-x)$:

$$p^{t+1}(\pi) \geq \exp\{-\epsilon\langle\pi(x^t),\ell^t\rangle\}\cdot p^t(\pi) \geq (1-\epsilon\langle\pi(x^t),\ell^t\rangle)\cdot p^t(\pi) \tag{23}$$

Using the definition of $\texttt{FTPL}^t$ and $\texttt{BTPL}^t$, this gives,

$$\begin{aligned}\texttt{BTPL}^t &= \sum_\pi p^{t+1}(\pi)\langle\pi(x^t),\ell^t\rangle \geq \sum_\pi (1-\epsilon\langle\pi(x^t),\ell^t\rangle)p^t(\pi)\langle\pi(x^t),\ell^t\rangle \\ &= \texttt{FTPL}^t - \epsilon\sum_\pi p^t(\pi)\langle\pi(x^t),\ell^t\rangle^2 \\ &= \texttt{FTPL}^t - \epsilon\mathbb{E}\left[\langle\pi(x^t),\ell^t\rangle^2\right]\end{aligned}$$

We will finish the proof by showing that $p_\pi^{t+1}(\pi) \leq p^{t+1}(\pi)$ for all $\pi\in\Pi$. For succinctness we drop the dependence on t. Notice that for any other policy $\pi'\neq\pi$

$$\mathcal{L}(\pi, z+\phi_\pi^t) \leq \mathcal{L}(\pi', z+\phi_\pi^t) \Rightarrow \mathcal{L}(\pi, z+\phi^t) \leq \mathcal{L}(\pi', z+\phi^t).$$

And similarly for strict inequalities. This follows since the loss of π remains unchanged, but the loss of π' can only go up, since $\ell_\pi^t(j) \leq \ell^t(j)$ (as losses are non-negative). For simplicity assume that π always wins in case of ties, though the argument goes through if we assume a deterministic tie-breaking rule based on some global ordering of policies. Thus,

$$p^{t+1}(\pi) = \mathbf{P}\left[\bigcap_{\pi'}\mathcal{L}(\pi, z+\phi^t) \leq \mathcal{L}(\pi', z+\phi^t)\right] \leq \mathbf{P}\left[\bigcap_{\pi'}\mathcal{L}(\pi, z+\phi_\pi^t) \leq \mathcal{L}(\pi', z+\phi_\pi^t)\right] = p_\pi^{t+1}(\pi)$$

as claimed. ■

B.4. Bounding Stability: Small Separator Setting

Finally, we prove the third claim in Theorem 2. This involves a new stability bound for the small separator setting.

Lemma 11 (Stability for small separator)**.** *For any $t\in[T]$ and any sequence $y^{1:t}$ of contexts $x^{1:t}$ and losses $f^{1:t}$ with $f^i:\{0,1\}^K\to\mathbb{R}^K$, the stability of* CONTEXT-FTPL(ϵ)*, when X is a separator, is upper bounded by:*

$$\mathbb{E}_{\{z\}}\left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t))\right] \leq 4\epsilon K d\cdot\|f^t\|_*^2$$

Proof. By the definition of $\|f^t\|_*$:

$$\mathbb{E}_{\{z\}}\left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t))\right] \leq 2\|f^t\|_*\Pr\left[\pi^t(x^t)\neq\pi^{t+1}(x^t)\right] \leq 2\|f^t\|_*\Pr[\pi^t\neq\pi^{t+1}]$$

Since X is a separator, $\pi^t\neq\pi^{t+1}$ if and only if there exists a context $x\in X$, such that $\pi^t(x)\neq\pi^{t+1}(x)$. Otherwise the two policies are identical. Thus we have by two applications of the union bound:

$$\begin{aligned}\Pr[\pi^t\neq\pi^{t+1}] &\leq \sum_{x\in X}\Pr[\pi^t(x)\neq\pi^{t+1}(x)] \\ &\leq \sum_{x\in X}\sum_{j\in K}\left(\Pr[j\in\pi^t(x), j\notin\pi^{t+1}(x)] + \Pr[j\notin\pi^t(x), j\in\pi^{t+1}(x)]\right)\end{aligned}$$

We bound the probability $\Pr[j\in\pi^t(x), j\notin\pi^{t+1}(x)]$. We condition on all random variables of $\{z\}$ except for the random variable $\ell_x(j)$, i.e. the random loss placed at coordinate j on the sample associated with context x. Denote the event corresponding to an assignment of all these other random variables as $\mathcal{E}_{-xj}$. Let ℓ_{xj} denote a loss vector which is $\ell_x(j)$ on the j-th coordinate and zero otherwise. Also let:

$$\Phi(\pi) = \sum_{\tau=1}^{t-1} f^\tau(\pi(x^\tau)) + \sum_{x'\neq x}\langle\pi(x'),\ell_{x'}\rangle + \langle\pi(x),\ell_x-\ell_{xj}\rangle \tag{24}$$

Let $\pi^* = \operatorname{argmin}_{\pi \in \Pi: j \in \pi(x)} \Phi(\pi)$ and $\tilde{\pi} = \min_{\pi \in \Pi: j \notin \pi(x)} \Phi(\pi)$. The event that $\{j \in \pi^t(x)\}$ happens only if:

$$\Phi(\pi^*) + \ell_x(j) \leq \Phi(\tilde{\pi}) \tag{25}$$

Let and $\nu = \Phi(\tilde{\pi}) - \Phi(\pi^*)$. Thus $j \in \pi^t(x)$ only if:

$$\ell_x(j) \leq \nu \tag{26}$$

Now if:

$$\ell_x(j) < \nu - 2\|f^t\|_* \tag{27}$$

then it is easy to see that $\{j \in \pi^{t+1}(x)\}$, since an extra loss of $f^t(a) \leq \|f^t\|_*$ cannot push j out of the optimal solution. More elaborately, for any other policy $\pi \in \Pi$, such that $j \notin \pi(x)$, the loss of π^* including time-step t is bounded as:

$$\begin{aligned}
\Phi(\pi^*) + \ell_x(j) + f^t(\pi^*(x^t)) &< \Phi(\pi) - 2\|f^t\|_* + f^t(\pi^*(x^t)) \\
&< \Phi(\pi) - \|f^t\|_* \\
&< \Phi(\pi) + f^t(\pi(x^t))
\end{aligned}$$

Thus any policy π, such that $j \notin \pi(x)$ is suboptimal after seeing the loss at time-step t. Thus

$$\Pr[j \in \pi^t(x), j \notin \pi^{t+1}(x) \mid \mathcal{E}_{-xj}] \leq \Pr[\ell_x(j) \in [\nu - 2\|f^t\|_*, \nu] \mid \mathcal{E}_{-xj}]$$

Since all other random variables are independent of $\ell_x(j)$ and $\ell_x(j)$ is a Laplacian with parameter ϵ:

$$\begin{aligned}
\Pr[\ell_x(j) \in [\nu - 2\|f^t\|_*, \nu] \mid \mathcal{E}_{-xj}] &= \Pr[\ell_x(j) \in [\nu - 2\|f^t\|_*, \nu]] \\
&= \frac{\epsilon}{2} \int_{\nu - 2\|f^t\|_*}^{\nu} e^{-\epsilon|z|} dz \leq \frac{\epsilon}{2} \int_{\nu - 2\|f^t\|_*}^{\nu} dz \leq \epsilon \|f^t\|_*
\end{aligned}$$

Similarly it follows that that: $\Pr[j \notin \pi^t(x), j \in \pi^{t+1}(x)] \leq \epsilon \|f^t\|_*$. To sum we get that:

$$\mathbb{E}_{\{z\}} \left[f^t(\pi^t(x^t)) - f^t(\pi^{t+1}(x^t)) \right] \leq 2\|f^t\|_* \Pr\left[\pi^t \neq \pi^{t+1}\right] \leq 4\epsilon K d \cdot \|f^t\|_*^2$$

■

C. Omitted Proofs from Section 4

C.1. Proof of Theorem 3: Transductive Setting

Consider the expected loss of the bandit algorithm at time-step t, conditional on $\mathcal{H}^{t-1}$:

$$\mathbb{E}[\langle \pi^t(x^t), \ell^t \rangle \mid \mathcal{H}^{t-1}] = \sum_{j=1}^{K} q^t(j) \cdot \ell^t(j) \leq \sum_{j=1}^{K} q^t(j) \cdot \mathbb{E}\left[\hat{\ell}^t(j) \mid \mathcal{H}^{t-1}\right] + \sum_{j=1}^{K} \ell^t(j) q^t(j) \cdot (1 - q^t(j))^L \tag{28}$$

As was observed by (Neu & Bartók, 2013), the second quantity can be upper bounded by $\frac{K}{eL}\|\ell^t\|_*$, since $q(1-q)^L \leq qe^{-Lq} \leq \frac{1}{eL}$.

Now observe that: $\sum_{j \in K} q^t(j) \cdot \mathbb{E}\left[\hat{\ell}^t(j) \mid \mathcal{H}^{t-1}\right]$ is the expected loss of the full feedback algorithm on the sequence of losses it observed and conditional on the history of play. By the regret bound of CONTEXT-FTPL(X, ϵ), given in case 2 of Theorem 2, we have that for any policy π^*:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} q^t(j) \cdot \hat{\ell}^t(j)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi^*(x^t), \hat{\ell}^t \rangle\right] + \epsilon \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi \in \Pi} p^t(\pi) \langle \pi(x^t), \hat{\ell}^t \rangle^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

Using the fact that expected estimates $\hat{\ell}$ are upper bounded by true losses:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} q^t(j) \hat{\ell}^t(j)\right] \leq \min_{\pi^* \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi^*(x^t), \hat{\ell}^t \rangle\right] + \epsilon \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi \in \Pi} p^t(\pi) \langle \pi(x^t), \hat{\ell}^t \rangle^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

Combining the two upper bounds, we get that the expected regret of the bandit algorithm is upper bounded by:

$$\textsc{Regret} \leq \epsilon\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi}p^t(\pi)\langle\pi(x^t),\hat{\ell}^t\rangle^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N) + \frac{K}{eL}\sum_{t=1}^{T}\mathbb{E}\left[\|\ell^t\|_*\right]$$

Now observe that, by a simple norm inequality and re-grouping:

$$\sum_{\pi\in\Pi}p^t(\pi)\langle\pi(x^t),\hat{\ell}^t\rangle^2 = \sum_{\pi\in\Pi}p^t(\pi)\left(\sum_{j\in\pi(x^t)}\hat{\ell}^t(j)\right)^2 \leq m\sum_{\pi\in\Pi}p^t(\pi)\sum_{j\in\pi(x^t)}\hat{\ell}^t(j)^2 = m\sum_{j\in[K]}q^t(j)\hat{\ell}^t(j)^2$$

Thus we get:

$$\textsc{Regret} \leq \epsilon m\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j\in[K]}q^t(j)\hat{\ell}^t(j)^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N) + \frac{K}{eL}\sum_{t=1}^{T}\mathbb{E}\left[\|\ell^t\|_*\right]$$

Now we bound each of the terms in the first summation, conditional on any history of play:

$$\sum_{j\in[K]}q^t(j)\mathbb{E}\left[\hat{\ell}^t(j)^2 \mid \mathcal{H}^{t-1}\right] = \sum_{j\in[K]}q^t(j)q^t(j)\ell^t(j)^2\mathbb{E}\left[J^t(j)^2 \mid \mathcal{H}^{t-1}, j\in\pi^t(x^t)\right]$$

Each $J^t(j)$ conditional on $\mathcal{H}^{t-1}$ and $j\in\pi^t(x^t)$ is distributed according to a geometric distribution with mean $q^t(j)$ truncated at L. Hence, it is stochastically dominated by a geometric distribution with mean $q^t(j)$. By known properties, if X is a geometrically distributed random variable with mean q, then $\mathbb{E}[X^2] = Var(X)+(\mathbb{E}[X])^2 = \frac{1-q}{q^2}+\frac{1}{q^2} = \frac{2-q}{q^2} \leq \frac{2}{q^2}$. Thus we have:

$$\sum_{j\in[K]}q^t(j)\mathbb{E}\left[\hat{\ell}^t(j)^2 \mid \mathcal{H}^{t-1}\right] \leq \sum_{j\in[K]}q^t(j)^2\ell^t(j)^2\frac{2}{q^t(j)^2} = 2\sum_{j=1}^{K}\ell^t(j)^2 \leq 2K\|\ell^t\|_\infty^2$$

Combining all the above we get the theorem.

C.2. Proof of Theorem 3: Small Separator Setting

Consider the expected loss of the bandit algorithm at time-step t, conditional on $\mathcal{H}^{t-1}$:

$$\mathbb{E}[\langle\pi^t(x^t),\ell^t\rangle \mid \mathcal{H}^{t-1}] = \sum_{j=1}^{K}q^t(j)\cdot\ell^t(j) \leq \sum_{j=1}^{K}q^t(j)\cdot\mathbb{E}\left[\hat{\ell}^t(j) \mid \mathcal{H}^{t-1}\right] + \sum_{j=1}^{K}\ell^t(j)q^t(j)\cdot(1-q^t(j))^L \qquad (29)$$

As was observed by (Neu & Bartók, 2013), the second quantity can be upper bounded by $\frac{K}{eL}\|\ell^t\|_*$, since $q(1-q)^L \leq qe^{-Lq} \leq \frac{1}{eL}$.

Now observe that: $\sum_{j\in K}q^t(j)\cdot\mathbb{E}\left[\hat{\ell}^t(j) \mid \mathcal{H}^{t-1}\right]$ is the expected loss of the full feedback algorithm on the sequence of losses it observed and conditional on the history of play. By the regret bound of $\textsc{Context-FTPL}(X,\epsilon)$, given in case 3 of Theorem 2, we have that for any policy π^*:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{K}q^t(j)\cdot\hat{\ell}^t(j)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\langle\pi^*(x^t),\hat{\ell}^t\rangle\right] + 4\epsilon Kd\cdot\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{f}^t\|_*^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)\\ &\leq \sum_{t=1}^{T}\langle\pi^*(x^t),\hat{\ell}^t\rangle + 4\epsilon Kd\cdot\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{\ell}^t\|_1^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)\end{aligned}$$

Using the fact that expected estimates $\hat{\ell}$ are upper bounded by true losses:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{K}q^t(j)\hat{\ell}^t(j)\right] \leq \min_{\pi^*\in\Pi}\mathbb{E}\left[\sum_{t=1}^{T}\langle\pi^*(x^t),\hat{\ell}^t\rangle\right] + 4\epsilon Kd\cdot\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{\ell}^t\|_1^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N)$$

Combining the two upper bounds, we get that the expected regret of the semi-bandit algorithm is upper bounded by:

$$\textsc{Regret} \leq \ 4\epsilon K d \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|\hat{\ell}^t\|_1^2\right] + \frac{10}{\epsilon}\sqrt{dm}\log(N) + \frac{K}{eL}\sum_{t=1}^{T}\mathbb{E}\left[\|\ell^t\|_*\right]$$

Now we bound each of the terms in the first summation, conditional on any history of play:

$$\mathbb{E}\left[\|\hat{\ell}^t\|_1^2 \mid \mathcal{H}^{t-1}\right] \leq m\mathbb{E}\left[\|\hat{\ell}^t\|_2^2\right] = m\sum_{j\in[K]}\mathbb{E}\left[\hat{\ell}^t(j)^2\right] = m\sum_{j\in[K]} q^t(j)\ell^t(j)^2\mathbb{E}\left[J^t(j)^2 \mid \mathcal{H}^{t-1}, j\in\pi^t(x^t)\right]$$

Each $J^t(j)$ conditional on $\mathcal{H}^{t-1}$ and $j \in \pi^t(x^t)$ is distributed according to a geometric distribution with mean $q^t(j)$ truncated at L. Hence, it is stochastically dominated by a geometric distribution with mean $q^t(j)$. By known properties, if X is a geometrically distributed random variable with mean q, then $\mathbb{E}[X^2] = Var(X) + (\mathbb{E}[X])^2 = \frac{1-q}{q^2} + \frac{1}{q^2} = \frac{2-q}{q^2} \leq \frac{2}{q^2}$. Moreover, trivially $\mathbb{E}[X^2] \leq L^2$, since X is truncated at L. Thus we have:

$$\mathbb{E}\left[\|\hat{\ell}^t\|_1^2 \mid \mathcal{H}^{t-1}\right] \leq m\sum_{j\in[K]} q^t(j)\ell^t(j)^2\min\left\{\frac{2}{q^t(j)^2}, L^2\right\} \leq m\|\ell^t\|_*^2\sum_{j\in[K]}\min\left\{\frac{2}{q^t(j)}, q^t(j)L^2\right\}$$

Now observe that: $\min\left\{\frac{2}{q^t(j)}, q^t(j)L^2\right\} \leq 2L$, since either $\frac{1}{q^t(j)} \leq L$ or otherwise, $q^t(j)L^2 \leq \frac{1}{L}L \leq L$. Thus we get:

$$\mathbb{E}\left[\|\hat{\ell}^t\|_1^2 \mid \mathcal{H}^{t-1}\right] \leq 2LKm\|\ell^t\|_*^2$$

Combining all the above we get the theorem.

D. Omitted Proofs from Section 6

D.1. Proof of Theorem 6

Similar to the analysis in Section 3, the proof of the Theorem is broken apart in two main Lemmas. The first lemma is an analogue of Theorem 1 for algorithms that use a predictor. This lemma can be phrased in the general online learning setting analyzed in Section 2. The second Lemma is an anaolgue of our multiplicative stability Lemma 10.

Let

$$\rho^t = M(\{z\} \cup y^{1:t}) \tag{30}$$

denote the policy that would have been played at time-step t if the predictor was equal to the actual loss vector that occured at time-step t. Moreover, for succinctness we will denote with $a^t = \pi^t(x^t)$ and with $b^t = \rho^t(x^t)$.

Lemma 12 (Follow vs Be the Leader with Predictors)**.** *The regret of a player under the optimistic FTPL and with respect to any $\pi^* \in \Pi$ is upper bounded by:*

$$\textsc{Regret} \leq \sum_{t=1}^{T}\mathbb{E}\left[\Delta Q^t(a^t) - \Delta Q^t(b^t)\right] + \mathbb{E}[\textsc{Error}] \tag{31}$$

where $\Delta Q^t(a) = f^t(a) - Q^t(a)$ and $\textsc{Error} = \max_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle - \min_{\pi\in\Pi}\sum_{x\in X}\langle\pi(x), \ell_x\rangle$.

Proof. Consider the augmented sequence $(x^1, Q^1), (x^1, f^1 - Q^1), (x^2, Q^2), (x^2, f^2 - Q^2), \ldots$, where each observation (x^t, f^t) is replaced by two observations (x^t, Q^t) followed by $(x^t, f^t - Q^t)$. Observe that by linearity of the objective, the two observations cancell out each other at the end, to give the same effect as a single observation of (x^t, f^t). Moreover, the leader after observing (x^t, Q^t) is equal to a^t, whilst after observing $(x^t, f^t - Q^t)$ is equal to b^t. Thus by applying Lemma 7 to this augmented sequence we get:

$$\begin{aligned}\sum_{t=1}^{T}\left(Q^t(a^t) + f^t(b^t) - Q^t(b^t)\right) \leq \ & \sum_{t=1}^{T}\left(Q^t(\pi^*(x^t)) + f^t(\pi^*(x^t)) - Q^t(\pi^*(x^t))\right) + \textsc{Error} \\ = \ & \sum_{t=1}^{T} f^t(\pi^*(x^t)) + \textsc{Error}\end{aligned}$$

Let $\texttt{BTPL}^t_Q = Q^t(a^t) + f^t(b^t) - Q^t(b^t)$ and $\texttt{FTPL}^t = f^t(a^t)$. Then, observe that:

$$\texttt{FTPL}^t - \texttt{BTPL}^t_Q = f^t(a^t) - Q^t(a^t) - (f^t(b^t) - Q^t(b^t)) = \Delta Q^t(a^t) - \Delta Q^t(b^t) \tag{32}$$

Combining the two properties we get that for any policy π^*:

$$\begin{aligned}\sum_{t=1}^T \texttt{FTPL}^t \leq & \sum_{t=1}^T \left(\Delta Q^t(a^t) - \Delta Q^t(b^t)\right) + \sum_{t=1}^T \texttt{BTPL}^t_Q \\ \leq & \sum_{t=1}^T \left(\Delta Q^t(a^t) - \Delta Q^t(b^t)\right) + \sum_{t=1}^T f^t(\pi^*(x^t)) + \textsc{Error}\end{aligned}$$

Re-arranging and taking expectation concludes the proof. ■

Lemma 13 (Stability with Predictors)**.** *In the transductive setting:*

$$\mathbb{E}\left[\Delta Q^t(a^t) - \Delta Q^t(b^t)\right] \leq 4\epsilon K \|f^t - Q^t\|_*^2 \tag{33}$$

In the small separator setting:

$$\mathbb{E}\left[\Delta Q^t(a^t) - \Delta Q^t(b^t)\right] \leq 4\epsilon K d \|f^t - Q^t\|_*^2 \tag{34}$$

Proof. We prove the first part of the Lemma. The second follows along identical arguments. By the definition of $\|f^t - Q^t\|_* = \|\Delta Q^t\|_* = \max_{a\in\mathcal{A}} |\Delta Q^t(a)|$, we have:

$$\mathbb{E}_{\{z\}}\left[\Delta Q^t(a^t) - \Delta Q^t(b^t)\right] \leq 2\|\Delta Q^t\|_* \Pr\left[a^t \neq b^t\right]$$

Now observe that:

$$\Pr[a^t \neq b^t] \leq \sum_{j\in K} \left(\Pr[j \in a^t, j \notin b^t] + \Pr[j \notin a^t, j \in b^t]\right)$$

We bound the probability $\Pr[j \in a^t, j \notin b^t]$. We condition on all random variables of $\{z\}$ except for the random variable $\ell_{x^t}(j)$, i.e. the random loss placed at coordinate j on the sample associated with context x^t. Denote the event corresponding to an assignment of all these other random variables as $\mathcal{E}_{-x^t j}$. Let $\ell_{x^t j}$ denote a loss vector which is $\ell_{x^t}(j)$ on the j-th coordinate and zero otherwise. Also let:

$$\Phi(\pi) = \sum_{\tau=1}^{t-1} f^\tau(\pi(x^\tau)) + Q^t(\pi(x^t)) + \sum_{x\in X - \{x^t\}} \langle \pi(x), \ell_x\rangle + \langle \pi(x^t), \ell_{x^t} - \ell_{x^t j}\rangle \tag{35}$$

Let $\pi^* = \operatorname{argmin}_{\pi\in\Pi: j\in\pi(x^t)} \Phi(\pi)$ and $\tilde{\pi} = \min_{\pi\in\Pi: j\notin\pi(x^t)} \Phi(\pi)$. The event that $\{j \in a^t\}$ happens only if:

$$\Phi(\pi^*) + \ell_{x^t}(j) \leq \Phi(\tilde{\pi}) \tag{36}$$

Let and $\nu = \Phi(\tilde{\pi}) - \Phi(\pi^*)$. Thus $j \in a^t$ only if:

$$\ell_{x^t}(j) \leq \nu \tag{37}$$

Now if:

$$\ell_{x^t}(j) < \nu - 2\|\Delta Q^t\|_* \tag{38}$$

then it is easy to see that $\{j \in b^t\}$, since an extra loss of $f^t(a) - Q^t(a) \leq \|\Delta Q^t\|_*$ cannot push j out of the optimal solution. More elaborately, for any other policy $\pi \in \Pi$, such that $j \notin \pi(x^t)$, the loss of π^* including time-step t is bounded as:

$$\begin{aligned}\Phi(\pi^*) + \ell_{x^t}(j) + f^t(\pi^*(x^t)) - Q^t(\pi^*(x^t)) < & \Phi(\pi) - 2\|\Delta Q^t\|_* + f^t(\pi^*(x^t)) - Q^t(\pi^*(x^t)) \\ < & \Phi(\pi) - \|\Delta Q\|_* \\ < & \Phi(\pi) + f^t(\pi(x^t)) - Q^t(\pi(x^t))\end{aligned}$$

Thus any policy π, such that $j \notin \pi(x^t)$ is suboptimal after seeing the loss at time-step t. Thus

$$\Pr[j \in a^t, j \notin b^t \mid \mathcal{E}_{-x^t j}] \leq \Pr[\ell_{x^t}(j) \in [\nu - 2\|\Delta Q^t\|_*, \nu] \mid \mathcal{E}_{-x^t j}]$$

Since all other random variables are independent of $\ell_{x^t}(j)$ and $\ell_{x^t}(j)$ is a Laplacian with parameter ϵ:

$$\begin{aligned}\Pr[\ell_{x^t}(j) \in [\nu - 2\|\Delta Q^t\|_*, \nu] \mid \mathcal{E}_{-x^t j}] = \;& \Pr[\ell_{x^t}(j) \in [\nu - 2\|\Delta Q^t\|_*, \nu]] \\ = \;& \frac{\epsilon}{2}\int_{\nu-2\|\Delta Q^t\|_*}^{\nu} e^{-\epsilon|z|}dz \leq \frac{\epsilon}{2}\int_{\nu-2\|\Delta Q^t\|_*}^{\nu} dz \leq \epsilon \cdot \|\Delta Q^t\|_*\end{aligned}$$

Similarly it follows that that: $\Pr[j \notin \pi^t(x^t) \text{ and } j \in \pi^{t+1}(x^t)] \leq \epsilon \cdot \|\Delta Q^t\|_*$. To sum we get that:

$$\mathbb{E}_{\{z\}}\left[\Delta Q^t(a^t) - \Delta Q^t(b^t)\right] \leq \; 2\|\Delta Q^t\|_* \Pr\left[\pi^t(x^t) \neq \pi^{t+1}(x^t)\right] \leq 4\epsilon K \|\Delta Q^t\|_*^2$$

■

The expected error term is identical to the expected error that we upper bounded in Lemma 8, hence the same bound carries over. Combining the above Lemmas with this observation, yields Theorem 6.

E. Infinite Policy Classes

In this section we focus on the contextual experts problem but consider infinite policy classes. Recall that in this setting, in each round t, the adversary picks a context $x^t \in \mathcal{X}$ and a loss function $\ell^t \in \mathbb{R}^K_{\geq 0}$, the learner, upon seeing the context x^t, chooses an action $a^t \in [K]$, and then suffers loss
$ell^t(a^t)$. We showed that as a simple consequence of Theorem 2, that when competing with a set of policies $\Pi \subset (\mathcal{X} \to [K])$ with $|\Pi| = N$ and against an adaptive adversary, CONTEXT-FTPL has regret at most $O(d^{1/4}\sqrt{T\log(N)})$ in the transductive setting and regret at most $O(d^{3/4}\sqrt{KT\log(N)})$ in the non-transductive setting with small separator.

Here we consider the situation where the policy class Π is infinite in size, but has small Natarajan dimension, which generalizes VC-dimension to multiclass problems. Specifically, we prove two results in this section: First we show that in the transductive case, CONTEXT-FTPL can achieve low regret relative to a policy class with bounded Natarajan dimension. Then we show that in the non-transductive case, it is hard in an information-theoretic sense to achieve sublinear regret relative to a policy class with constant Natarajan dimension. Together, these results show that finite Natarajan or VC dimension is sufficient for sublinear regret in the transductive setting, but it is *insufficient* for sublinear regret in the fully online setting.

Before proceeding with the two results, we must introduce the notion of Natarajan dimension, which requires some notation. For a class of functions $\mathcal{F}$ from $\mathcal{X} \to [K]$ and for a sequence $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define $\mathcal{F}_X = \{(f(x_1), \ldots, f(x_n)) \in [K]^n : f \in \mathcal{F}\}$ be the restriction of the functions to the points. Let Ψ be a family of mappings from $[K] \to \{0, 1, \star\}$. Let $\bar{\psi} = (\psi_1, \ldots, \psi_n) \in \Psi^n$ be a fixed sequence of such mappings and for a sequence $(s_1, \ldots, s_n) \in [K]^N$ define $\bar{\psi}(s) = (\psi_1(s_1), \ldots, \psi_1(s_n)) \in \{0, 1, \star\}^n$. We say a sequence $X \in \mathcal{X}^n$ is Ψ-shattered by $\mathcal{F}$ if there exists $\bar{\psi} \in \Psi^n$ such that:

$$\{0, 1\}^n \subseteq \{\bar{\psi}(s) : s \in \mathcal{F}_X\}$$

The Ψ-dimension of a function class $\mathcal{F}$ is the largest n such that there exist a sequence $X \in \mathcal{X}^n$ that is Ψ-shattered by $\mathcal{F}$. Notice that if $K = 2$ and Ψ contains only the identity map, then the Ψ-dimension is exactly the VC dimension.

The **Natarajan dimension** is the Ψ dimension for the class $\Psi_N = \{\psi_{N,i,j}, i, j \in [K], j \neq i\}$ where $\psi_{N,i,j}(a) = 1$ if $a = i$, $\psi_{N,i,j}(a) = 0$ if $a = j$ and $\psi_{N,i,j}(a) = \star$ otherwise. Notice that Natarajan dimension is a strict generalization of VC-dimension as Ψ_N contains only the identity map if $K = 2$. Thus our result also applies to VC-classes in the two-action case. The main property we will use about function classes with bounded Natarajan Dimension is the following analog of the Sauer-Shelah Lemma:

Lemma 14 (Sauer-Shelah for Natarajan Dimension (Haussler & Long, 1995; Ben-David et al., 1995))**.** *Suppose that $\mathcal{F}$ has Ψ_N dimension at most ν. Then for any set $X \in \mathcal{X}^n$, we have:*

$$|\mathcal{F}_X| \leq \left(\frac{ne(K+1)^2}{2\nu}\right)^{\nu}$$

Our positive result for transductive learning with a Natarajan class is the following regret bound for CONTEXT-FTPL,

Corollary 15. *Consider running* CONTEXT-FTPL(X, ϵ) *in the transductive contextual experts setting with a policy class* Π *with Natarajan dimension at most* ν*. Then the algorithm achieves regret against an adaptive and adversarially chosen sequence of contexts and loss functions,*

$$\epsilon \sum_{t=1}^{T} \mathbb{E}[\langle \pi^t(x^t), \ell^t \rangle^2] + \frac{10}{\epsilon} \sqrt{d\nu \log(K) \log\left(\frac{de(K+1)^2}{2\nu}\right)}.$$

When ϵ *is set optimally and losses are in* $[0,1]^K$*, this is* $O((d\nu \log(K) \log(dK/\nu))^{1/4}\sqrt{T})$.

Proof. The result is a consequence of the second clause of Theorem 2, using the additional fact that any sequence of contexts $X = (x_1, \ldots, x_d)$ induce a finite policy class $\Pi_X \subseteq [K]^d$. The fact that Π has Natarajan dimension at most ν means that $|\Pi_X| \leq \left(\frac{de(K+1)^2}{2\nu}\right)^{\nu}$ by Lemma 14. Therefore, once the d contexts are fixed, as they are in the transductive setting, we are back in the finite policy case and can apply Theorem 2 with N replaced by $|\Pi_X|$. ■

Thus we see that CONTEXT-FTPL has sublinear regret relative to policy classes with bounded Natarajan dimension, even against adaptive adversaries. The second result in this section shows that this result cannot be lifted to the non-transductive setting. Specifically, we prove the following theorem in the section, which shows that no algorithm, including inefficient ones, can achieve sublinear regret against a VC class in the non-transductive setting.

Theorem 16. *Consider an online binary classification problem in one dimension with* $\mathcal{F} \subset [0,1] \to \{0,1\}$ *denoting the set of all threshold functions. Then there is no learning algorithm that can guarantee* $o(T)$ *expected regret against an adaptive adversary. In particular, there exists a policy class of VC dimension one such that no learning algorithm can achieve sublinear regret against an adaptive adversary in the contextual experts problem.*

Proof. We define an adaptive adversary and argue that it ensures at least $1/2$ expected regret per round. While the adversary does not have access to the random coins of the learner, it can compute the probability that the learner would label any point as $\{0,1\}$. At round t, let $p_t(x)$ denote the probability that the learner would label a point $x \in [0,1]$ as 1, and note that this quantity is conditioned on the entire history of interaction. At each round t, the adversary will have played a set of points X_t^+ with positive label and X_t^- with negative label and she will maintain the invariant that $\min_{x \in X_t^+} x > \max_{x \in X_t^-} x$ for all t. At every time t, the adversary will play context $x_t \in (\max_{x \in X_t^-} x, \min_{x \in X_t^+} x)$. The adversary, knowing the learning algorithm, will compute $p_t(x_t)$ and assign label $y_t = 1$ if $p_t(x) < 1/2$ and 0 otherwise. The adversary will then update the sets $X_{t+1}^+ \leftarrow X_t^+ \cup \{x_t\}$ if $y_t = 1$ and $X_{t+1}^+ \leftarrow X_t^+$ otherwise. X_{t+1}^- is updated analogously.

Clearly this sequence of contexts maintains the appropriate invariant for the adversary, namely there is always an interval between the positive and negative examples in which he can pick a context. This implies that on the sequence, there is a threshold $f^\star \in \mathcal{F}$ that perfectly classifies the points, so its cumulative reward is T. Moreover, by the choice of label selected by the adversary, the expected reward of the learner at round t is at most $1/2$, which means the cumulative expected reward of the learner is at most $T/2$. Thus the regret of the learner is at least $T/2$. ■