Generalized Policy Elimination: an efficient algorithm for Nonparametric Contextual Bandits

Aurélien F. Bibaut, Antoine Chambaz, Mark J. van der Laan

Abstract

We propose the Generalized Policy Elimination (GPE) algorithm, an oracle-efficient contextual bandit (CB) algorithm inspired by the Policy Elimination algorithm of Dudik et al. [2011]. We prove the first regret optimality guarantee theorem for an oracle-efficient[1] and CB algorithm competing against a nonparametric class with infinite VC-dimension. Specifically, we show that GPE is regret-optimal (up to logarithmic factors) for policy classes with integrable entropy.

For classes with larger entropy, we show that the core techniques used to analyze GPE can be used to design an ε-greedy algorithm with regret bound matching that of the best algorithms to date. We illustrate the applicability of our algorithms and theorems with examples of large nonparametric policy classes, for which the relevant optimization oracles can be efficiently implemented.

1 INTRODUCTION

In the contextual bandit (CB) feedback model, an agent (the learner) sequentially observes a vector of covariates (the context), chooses an action among finitely many options, then receives a reward associated to the context and the chosen action. A CB algorithm is a procedure carried out by the learner, whose goal is to maximize the reward collected over time. Known as policies, functions that map any context to an action or to a distribution over actions play a key role in the CB literature. In particular, the performance of a CB algorithm is typically measured by the gap between the collected reward and the reward that would have been collected had the best policy in a certain class Π been exploited. This gap is the so-called *regret against policy class* Π. The class Π is called the *comparison class*.

The CB framework applies naturally to settings such as online recommender systems, mobile health and clinical trials, to name a few. Although the regret is defined relative to a given policy class, the goal in most settings is arguably to maximize the (expected cumulative) reward in an absolute sense. It is thus desirable to compete against large nonparametric policy classes, which are more likely to contain a policy close to the best measurable policy.

The complexity of a nonparametric class of functions can be measured by its covering numbers. The ϵ-covering number $N(\epsilon, \mathcal{F}, L_r(P))$ of a class $\mathcal{F}$ is the number of balls of radius $\epsilon > 0$ in $L_r(P)$ norm ($r \geq 1$) needed to cover $\mathcal{F}$. The ϵ-covering entropy is defined as $\log N(\epsilon, \mathcal{F}, L_r(P))$. Upper bounds on the covering entropy are well known for many classes of functions. For instance, the ϵ-covering entropy of a p-dimensional parametric class is $O(p \log(1/\epsilon))$ for all $r \geq 1$. In contrast, the ϵ-covering entropy of the class $\{f : [0,1]^d \to \mathbb{R} : \forall x, y, |f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)| \leq M\|x - y\|^{\alpha - \lfloor \alpha \rfloor}\}$[2]of d-variate Hölder functions is $O(\epsilon^{-d/\alpha})$ for $r = \infty$ (hence all $r \geq 1$) [van der Vaart and Wellner, 1996, Theorem 2.7.1]. Another popular measure of complexity is the Vapnik-Chervonenkis (VC) dimension. Since the ϵ-covering entropy of a class of VC dimension V is $O(rV \log(1/\epsilon))$ for all $r \geq 1$ [van der Vaart and Wellner, 1996, Theorem 2.6.7], the complexity of a class with finite VC dimension is essentially the same as that of a

[1] So as to dispel any possible confusion early on, we mean a CB algorithm is oracle-efficient if, over T rounds, the number of calls to optimization oracles it makes is polynomial in T, rather than exponential in T. This is in that sense that oracle-efficiency in meant in articles such as Dudik et al. [2011], Agarwal et al. [2014], Dudík et al. [2017]. We are not claiming that our proposed methods are efficient in the sense that they would have a reasonable runtime in practice.

[2] $\lfloor \alpha \rfloor$ is the integer part; $f^{(m)}$ is the m-th derivative.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

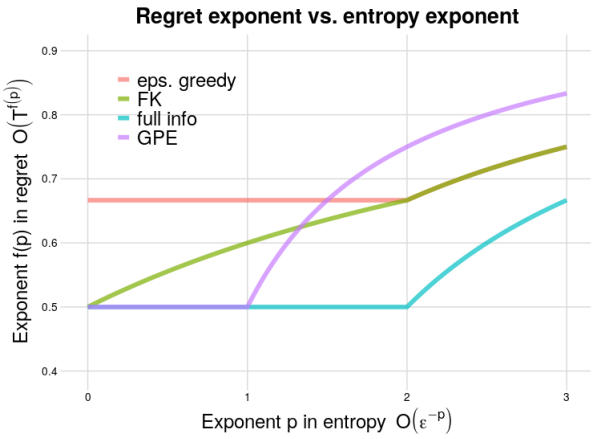

Figure 1: Exponent in regret upper bound (up to logarithmic factors) as a function of the exponent in the (supremum norm) covering entropy. *FK* is the theoretical upper bound of Foster and Krishnamurthy [2018]. *Full info* is the bound achieved by Empirical Risk Minimizers under full information feedback.

parametric class.

We will consider classes Π of policies with either a *polynomial* or a *logarithmic* covering entropy, for which $\log N(\epsilon, \Pi, L_r(P))$ is either $O(\epsilon^{-p})$ for some $p > 0$ or $O(\log(1/\epsilon))$. The former are much bigger than the latter.

Efficient CB algorithms competing against classes of functions with polynomial covering entropy have been proposed [e.g. by Cesa-Bianchi et al., 2017, Foster and Krishnamurthy, 2018]. However, these algorithm are not regret-optimal in a minimax sense. In parallel, Dudik et al. [2011], Agarwal et al. [2014] have proposed efficient algorithms which are regret-optimal for finite policy classes, or for policy classes with finite VC dimension. Thus there seems to be a gap: as of today, no efficient algorithm has been proven to be regret-optimal for comparison classes with polynomial entropy (or with infinite VC dimension). In this article, we partially bridge this gap. We provide the first efficient algorithm to be regret-optimal (up to some logarithmic factors) for comparison classes with integrable entropy (that is, $\log N(\epsilon, \Pi, L_r(P)) = O(\epsilon^{-p})$ for $p \in (0, 1)$). Our main algorithm, that we name Generalized Policy Elimination (GPE) algorithm, is derived from the Policy Elimination algorithm of Dudik et al. [2011].

1.1 PREVIOUS WORK

Many contributions have been made to the area of nonparametric contextual bandits. Among others, one way to classify them is according to whether they rely on some version of the exponential weights algorithm, on optimization oracles, or on a discretization of the covariates space.

Exponential weights-based algorithms. The exponential weights algorithm has a long history in adversarial online learning, dating back to the seminal articles of Vovk [1990] and Littlestone and Warmuth [1994]. The Exp3 algorithm of Auer et al. [2002b] is the first instance of exponential weigthts for the adversarial multi-armed bandit problem. The Exp4 algorithm of Auer et al. [2002a] extends it to the contextual bandit setting. Infinite policy classes can be handled by running a version of the Exp4 algorithm on an ε-cover of the policy class. While the Exp4 algorithm enjoys optimal (in a minimax sense) regret guarantees, it requires maintaining a set of weights over all elements of the cover, and is thus intractable for most nonparametric classes, because their covering numbers typically grow exponentially in $1/\epsilon$. Cesa-Bianchi et al. [2017] proposed the first cover-based efficient online learning algorithm. Their algorithm relies on a hierarchical cover obtained by the celebrated chaining device of Dudley [1967]. It achieves the minimax regret under the full information feedback model but not under the bandit feedback model, although it yields rate improvements over past works for large nonparametric policy classes. Cesa-Bianchi et al. [2017]'s regret bounds are expressed in terms of an entropy integral. An alternative approach to nonparametric adversarial online learning is that of Chatterji et al. [2019], who proposed an efficient exponential-weights algorithm for a reproducing kernel Hilbert-space (RKHS) comparison class. They characterized the regret in terms of the eigen-decay of the kernel. They obtained optimal regret if the kernel has exponential eigen-decay.

Oracle efficient algorithms. The first oracle-based CB algorithm is the epoch-greedy algorithm of Langford and Zhang [2008]. Epoch-greedy allows to turn any supervised learning algorithm into a CB algorithm, making it practical and efficient (in terms of the number of calls to a supervised classification subroutine). Its regret can be characterized in a straighforward manner as a function of the sample complexity of the supervised learning algorithm, but is suboptimal. Dudik et al. [2011] introduced RandomizedUCB, the first regret-optimal efficient CB algorithm. Agarwal et al. [2014] improved on their work by requiring fewer calls to the oracle. [Foster et al., 2018] pointed out that the aforementioned algorithms rely on cost-sensitive classification oracles, which are in general intractable (even though for some relatively natural classes there exist efficient algorithms). Foster et al. [2018] proposed regret-optimal, regression oracles-based algorithms, motivated by the fact that regression

oracles can in general be implement efficiently. Another way to make tractable these oracles is, in the case of cost-sensitive classification oracles, to use surrogate losses, as studied by Foster and Krishnamurthy [2018]. They gave regret upper bounds (see Figure 1) and a nonconstructive proof of the existence of an algorithm that achieves them. They also proposed an epoch greedy-style algorithm that achieves the best regret guarantees to date for entropy $\log N(\epsilon, \Pi)$ of order ϵ^{-p} for some $p > 2$. The caveat of the surrogate loss-based approach is that guarantees are either in terms of so-called *margin-based regret*, or can be expressed in terms of the usual regret, but under the so-called realizability assumption. We refer the interested reader to Foster and Krishnamurthy [2018] for further details.

Covariate space discretization-based algorithms. A third way to design nonparametric CB algorithms consists in discretizing the context space into bins and running multi-armed bandit algorithms in each bin. This approach was pioneered by Rigollet and Zeevi [2010] and extended by Perchet and Rigollet [2013]. They take a relatively different perspective from the previously mentioned works, in the sense that the comparison class is defined in an implicit fashion: they assume that the expected reward of each action is a smooth (Hölder) function of the context, and they compete against the policy defined by the argmax over actions of the expected reward. Their regret guarantees are optimal in a minimax sense.

1.2 OUR CONTRIBUTIONS

Primary contribution. In this article, we introduce the Generalized Policy Elimination algorithm, derived from the Policy Elimination algorithm of Dudik et al. [2011]. GPE is an oracle-efficient algorithm, of which the regret can be bounded in terms of the metric entropy of the policy class. In particular we show that if the entropy is integrable, then GPE has optimal regret, up to logarithmic factors. The key enabler of our results is a new maximal inequality for martingale processes (Theorem 5 in appendix C), inspired by [van de Geer, 2000, van Handel, 2011]. Although our regret upper bounds for GPE are no longer optimal for policy classes with non-integrable entropy, we show that we can use the same type of martingale process techniques to design an ε-greedy type algorithm that matches the current best upper bounds.

Comparison to previous work. Earlier works on regret-optimal oracle-efficient algorithms [Dudik et al., 2011, Agarwal et al., 2014, Foster et al., 2018, for instance] have in common that the regret analysis holds for a finite number of policies or for policy classes with finite VC dimension. GPE is the first oracle-efficient algorithm for which are proven regret optimality guarantees against a truly nonparametric policy classes (that is, larger than VC). We refer the reader to appendix A for additional comparisons with previous articles.

Secondary contributions. In addition to the nonparametric extension of policy elimination and analysis of ε-greedy in terms of (bracketing) entropy[3], we introduce several ideas that, to the best of our knowledge, have not appeared so far in the literature. In particular, we demonstrate the possibility of doing what we call *direct policy optimization* over a nonparametric policy class, that is of directly finding a maximizer $\widehat{\pi}$ of $\pi \mapsto \widehat{\mathcal{V}}(\pi)$ over some nonparametric Π where $\widehat{\mathcal{V}}(\pi)$ estimates the value $\mathcal{V}(\pi)$ of policy π. As far as we know, no example has been given yet of a nonparametric class Π for which $\widehat{\pi}$ can be efficiently computed, although some articles postulate the availability of $\widehat{\pi}$ [Luedtke and Chambaz, 2019, Athey and Wager, 2017]. Here, we exhibit several rich classes for which direct policy optimization can be efficiently implemented. Another secondary contribution is the first formal regret bounds for the ε-greedy algorithm, which follows from the same type of arguments as in the analysis of GPE. We were relatively surprised to see that unlike the epoch-greedy algorithm, the ε-greedy algorithm has not been formally analyzed yet, to the best of our knowledge. This may be due to the fact that doing so requires martingale process theory, which has only recently started to receive attention in the CB literature.

1.3 SETTING

For each $m \geq 1$, denote $[m] \doteq \{1, \ldots, m\}$.

At time $t \geq 1$, the learner observes context $W_t \in \mathcal{W} \doteq [0,1]^d$, chooses an action $A_t \in [K]$, $K \geq 2$, and receives the outcome/reward $Y_t \in \{0,1\}$. We suppose that the contexts are i.i.d. and the rewards are conditionally independent given actions and contexts, with fixed conditional distributions across time points. We denote O_t the triple (W_t, A_t, Y_t), and P the distribution[4] of the infinite sequence $O_1, O_2, \ldots, O_t, \ldots$. Moreover, let $O^{\mathrm{ref}} \doteq (W^{\mathrm{ref}}, A^{\mathrm{ref}}, Y^{\mathrm{ref}})$ be a random variable such that $W^{\mathrm{ref}} \sim W_1$, $A^{\mathrm{ref}}|W^{\mathrm{ref}} \sim \mathrm{Unif}([K])$, $Y^{\mathrm{ref}}|A^{\mathrm{ref}}, W^{\mathrm{ref}} \sim Y_1|A_1, W_1$. We denote F_t the filtration induced by $O_1, \ldots, O_t$.

[3] We recall the definition of bracketing entropy further down. Bracketing entropy in $\|\cdot\|_\infty$ norm is dominated by the more widely-known metric entropy in $\|\cdot\|_\infty$ norm, so that our results can be read with the latter in mind without much loss of generality.

[4] P is partly a fact of nature, through the marginal distribution of context and the conditional distributions of reward given context and action, and the result of the learner's decisions.

Generically denoted f or π, a policy is a mapping from $\mathcal{W} \times [K]$ to $\mathbb{R}_+$ such that, for all $w \in \mathcal{W}$, $\sum_{a\in[K]} f(a, w) = 1$. Thus, a policy can be viewed as mapping a context to a distribution over actions. We say the learner is carrying out policy π at time t if, for all $a \in [K]$, $w \in \mathcal{W}$, $P[A_t = a|W_t = w] = \pi(a, w)$. Owing to statistics terminology, we also call *design* the policy carried out at a given time point. The value $\mathcal{V}(\pi)$ of π writes as

$$\mathcal{V}(\pi) \doteq E_P \left[\sum_{a\in[K]} E_P[Y|A = a, W]\pi(a|W) \right].$$

For any two policies f and g, we denote

$$V(g, f) \doteq E_P \left[\sum_{a\in[K]} \frac{f(a|W)}{g(a|W)} \right]. \tag{1}$$

We call $V(g, f)$ the importance sampling (IS) ratio of f and g. The IS ratio drives the variance of IS estimators of $\mathcal{V}(f)$ had the data been collected under policy g.

2 GENERALIZED POLICY ELIMINATION

Introduced by Dudik et al. [2011], the policy elimination algorithm relies on the following key fact. Let g_{ref} be the uniform distribution over actions used as a reference design/policy:

$$\forall (a, w) \in [K] \times \mathcal{W},\ g_{\text{ref}}(a, w) \doteq K^{-1}.$$

Proposition 1. *Let $\delta > 0$. For all compact and convex set $\mathcal{F}$ of policies, there exists a policy $g \in \mathcal{F}$ such that*

$$\sup_{f\in\mathcal{F}} V(\delta g_{\text{ref}} + (1 - \delta)g, f) \leq 2K. \tag{2}$$

We refer to their article for a proof of this result. Proposition 1 has an important consequence for exploration. Suppose that at time t we have a set of candidate policies $\mathcal{F}_t$, and that the designs $g_1, ..., g_t$ satisfy (2) with $\mathcal{F}_t$ substituted for $\mathcal{F}$. We can then estimate the value of candidate policies with error uniformly small over $\mathcal{F}_t$. This in turn has an important implication for exploitation: we can eliminate from $\mathcal{F}_t$ all the policies that have value below some well-chosen threshold, yielding a new policy set $\mathcal{F}_{t+1}$, and choose the next exploration policy g_{t+1} in $\mathcal{F}_{t+1}$. This reasoning suggested to Dudik et al. [2011] their policy elimination algorithm: (1) initialize the set of candidate policies to the entire policy class, (2) choose an exploration policy that ensures small value estimation error uniformly over candidate policies, (3) eliminate low value policies, (4) repeat steps (2) and (3). We present formally our version of the policy algorithm as algorithm 1 below.

In this section, we show that under an entropy condition, and if we have access to a certain optimization oracle, our GPE algorithm is efficient and beats existing regret upper bounds in some nonparametric settings. Our contribution here is chiefly to extend the regret analysis of Dudik et al. [2011] to classes of functions characterized by their metric entropy in $L_\infty(P)$ norm. This requires us to prove a new chaining-based maximal inequality for martingale processes (Theorem 6 in appendix C). On the computational side, our algorithm relies on having access to slightly more powerful oracles than that of Dudik et al. [2011]. We present them in subsection 2.2 and give several examples where these oracles can be implemented efficiently.

We now formally state our GPE algorithm. Consider a policy class $\mathcal{F}$. For any policy f, any $o = (w, a, y) \in \mathcal{W} \times [K] \times \{0, 1\}$, define the policy loss and its IS-weighted counterpart

$$\ell(f)(o) \doteq f(a, w)(1 - y),$$
$$\ell_\tau(f)(o) \doteq \frac{g_{\text{ref}}(a, w)}{g_\tau(a, w)} f(a, w)(1 - y),$$

the corresponding risk $R(f) \doteq E[\ell(f)(O^{\text{ref}})] = E_P[\ell_\tau(f)(O_\tau)]$ and its empirical counterpart $\widehat{R}_t(f) \doteq t^{-1} \sum_{\tau=1}^{t} \ell_\tau(f)(O_\tau)$.

Algorithm 1 Generalized Policy Elimination

Inputs: policy class $\mathcal{F}$, $\epsilon > 0$, sequences $(\delta_t)_{t\geq 1}$, $(x_t)_{t\geq 1}$.
Initialize $\mathcal{F}_1$ as $\mathcal{F}$.
for $t \geq 1$ **do**
Find $\widetilde{g}_t \in \mathcal{F}_t$ such that, for all $f \in \mathcal{F}_t$,

$$\frac{1}{t-1} \sum_{\tau=1}^{t-1} \frac{f(a|W_\tau)}{(\delta_t g_{\text{ref}} + (1 - \delta_t)\widetilde{g}_t)(a|W_\tau)} \leq 2K. \tag{3}$$

Define $g_t = \delta_t g_{\text{ref}} + (1 - \delta_t)\widetilde{g}_t$.
Observe context W_t, sample action $A_t \sim g_t(\cdot|W_t)$, collect reward Y_t.
Define $\mathcal{F}_{t+1}$ as

$$\left\{ f \in \mathcal{F}_t : \widehat{R}_t(f) \leq \min_{f\in\mathcal{F}_t} \widehat{R}_t(f) + x_t \right\}. \tag{4}$$

end for

2.1 REGRET ANALYSIS

Our regret analysis relies on the following assumption.

Assumption 1 (Entropy condition)**.** *There exist $c > 0$, $p > 0$ such that, for all $\epsilon > 0$, $\log N(\epsilon, \mathcal{F}, L_\infty(P)) \leq c\epsilon^{-p}$.*

Defining $\mathcal{F}_{t+1} \subset \mathcal{F}_t$ as (4), the policy elimination step, consists in removing from $\mathcal{F}_t$ all the policies that are known to be suboptimal with high probability. The threshold x_t thus plays the role of the width of a uniform-over-$\mathcal{F}_t$ confidence interval. Set $\epsilon > 0$ arbitrarily. We will show that the following choice of $(\delta_\tau)_{\tau \geq 1}$ and $(x_\tau)_{\tau \geq 1}$ ensures that the confidence intervals hold with probability $1 - 6\epsilon$, uniformly both in time and over the successive $\mathcal{F}_\tau$'s: for all $\tau \geq 1$, $\delta_\tau \doteq \tau^{-(1/2 \wedge 1/(2p))}$ and

$$x_\tau \doteq x_\tau(\epsilon) \doteq \sqrt{v_\tau(\epsilon)} \Bigg\{ \frac{c_1}{\tau^{\frac{1}{2} \wedge \frac{1}{2p}}} + \frac{c_2 + c_5 \sqrt{v_\tau(\epsilon)}}{\sqrt{\tau}} \times \sqrt{\log\left(\frac{\tau(\tau+1)}{\epsilon}\right)} + \frac{1}{\tau \delta_\tau} \left(c_3 + c_7 \log\left(\frac{\tau(\tau+1)}{\epsilon}\right)\right) \Bigg\}$$

— defined in appendix D, $v_\tau(\epsilon)$ is a high probability upper bound on $\sup_{f \in \mathcal{F}_\tau} \mathrm{Var}_P(\ell_\tau(f)(O_\tau)|F_{\tau-1})$. It is constructed as follows. It can be shown that the conditional variance of $\ell_\tau(f)(O_\tau)$ given $F_{\tau-1}$ is driven by the expected IS ratio $E_P[\sum_{a \in [K]} f(a,W)/g_\tau(a,W)|F_{\tau-1}]$. Step 3 ensures that the empirical mean over past observations of the IS ratio is no greater than $2K$, uniformly over $\mathcal{F}_\tau$. The gap $(v_\tau(\epsilon) - 2K)$ is a bound on the supremum over $\mathcal{F}_\tau$ of the deviation between empirical IS ratios and the true IS ratios.

We now state our regret theorem for algorithm 1. Let $f^* \doteq \arg\min_{f \in \mathcal{F}}$ be the optimal policy in $\mathcal{F}$.

Theorem 1 (High probability regret bound for policy elimination)**.** *Consider algorithm 1. Suppose that Assumption 1 is met. Then, with probability at least $1 - 7\epsilon$, for all $t \geq 1$,*

$$\sum_{\tau=1}^{t} (\mathcal{V}(f^*) - Y_\tau) \leq \sqrt{t \log\left(\frac{1}{\epsilon}\right)} + 2\sum_{\tau=1}^{t} x_\tau(\epsilon) + \sum_{\tau=1}^{t} \delta_\tau = \begin{cases} O\left(\sqrt{t}\left(\log(\frac{t}{\epsilon})\right)^{3/2}\right) & \text{if } p \in (0,1) \\ O\left(t^{\frac{p-1/2}{p}}\left(\log(\frac{t}{\epsilon})\right)^{3/2}\right) & \text{if } p > 1 \end{cases}.$$

The proof of Theorem 1, presented in appendix D, hinges on the three following facts.

1. Controlling the supremum w.r.t. $f \in \mathcal{F}_\tau$ of the empirical estimate of the IS ratio (see (3) in the first step of the loop in algorithm 1) allows to control the supremum w.r.t. f of the true IS ratio $V(g_\tau, f)$.

2. With the specification of $(x_t)_{t \geq 1}$ and $(\delta_t)_{t \geq 1}$ sketched above we can guarantee that, with probability at least $1 - 3\epsilon$, $f^* \in \mathcal{F}_t \subset \ldots \subset \mathcal{F}_1$.

3. If $f^* \in \mathcal{F}_t$ then we can prove that, with probability at least $1 - 5\epsilon$, for all $\tau \in [t]$,

$$R(\widetilde{g}_\tau) - R(f^*) \leq 2x_\tau(\epsilon).$$

This in turn yields a high probability bound on the cumulative regret of algorithm 1.

2.2 AN EFFICIENT ALGORITHM FOR THE EXPLORATION POLICY SEARCH STEP

We show that the exploration policy search step can be performed in $O(\mathrm{poly}(t))$ calls to two optimization oracles that we define below. The explicit algorithm and proof of the claim are presented in appendix F.

Definition 1 (Linearly Constrained Least-Squares Oracle)**.** *We call Linearly Constrained Least-Squares Oracle (LCLSO) over $\mathcal{F}$ a routine that, for any $t \geq 1$, $q \geq 1$, vector $w \in \mathbb{R}^{Kt}$, sequence of vectors $W_1, ..., W_t \in \mathcal{W}$, set of vectors $u_1, ..., u_q \in \mathbb{R}^{Kt}$, and scalars $b_1, ..., b_q$, returns, if there exists one, a solution to*

$$\min_{f \in \mathcal{F}} \sum_{\substack{a \in [K] \\ \tau \in [t]}} (w(a,\tau) - f(a, W_\tau))^2 \text{ subject to}$$

$$\forall m \in [q], \sum_{\substack{a \in [K] \\ \tau \in [t]}} u_m(a,\tau) f(a, W_\tau) \leq b_\tau.$$

Definition 2 (Linearly Constrained Cost-Sensitive Classification Oracle)**.** *We call Linearly Constrained Cost-Sensitive Classification Oracle (LCCSCO) over $\mathcal{F}$ a routine that, for any $t \geq 1$, $q \geq 1$, vector $C \in (\mathbb{R}_+)^{Kt}$, set of vectors $W_1, ..., W_t \in \mathcal{W}$, set of vectors $u_1, ..., u_q \in \mathbb{R}^{Kt}$, and set of scalars $b_1, ..., b_q \in \mathbb{R}$ returns, if there exists one, a solution to*

$$\min_{f \in \mathcal{F}} \sum_{\substack{a \in [K] \\ \tau \in [t]}} C(a,\tau) f(a, W_\tau) \text{ subject to}$$

$$\forall m \in [q], \sum_{\substack{a \in [K] \\ \tau \in [t]}} u_m(a,\tau) f(a, W_\tau) \leq b_\tau.$$

The following theorem is our main result on the computational tractability of the policy search step.

Theorem 2 (Computational cost of exploration policy search)**.** *For every $t \geq 1$, exploration policy search at time t can be performed in $O((Kt)^2 \log t)$ calls to both LCLSO and LCCSCO.*

The proof of Theorem 2 builds upon the analysis of Dudik et al. [2011]. Like them, we use the famed ellipsoid algorithm as the core component. The general idea is as follows. We show that the exploration policy search step (3) boils down to finding a point $w \in \mathbb{R}^{Kt}$ that belongs to a certain convex set $\mathcal{U}$, and to identifying a $\widetilde{g}_t \in \mathcal{F}_t$ such that $\sum_{a,\tau}(f(a, W_\tau) - w(a,\tau))^2 \leq \Delta$ for a certain $\Delta > 0$. In section F.1, we identify $\mathcal{U}$ and Δ. In section F.2, we demonstrate how to find a point in $\mathcal{U}$ with the ellipsoid algorithm.

3 FINITE SAMPLE GUARANTEES FOR ε-GREEDY

In this section, we give regret guarantees for two variants of the ε-greedy algorithm competing against a policy class characterized by bracketing entropy, denoted thereon $\log N_{[]}$, and defined in the appendix[5]. Corresponding to two choices of an input argument ϕ, the two variants of algorithm 2 differ in whether they optimize w.r.t. the policy either an estimate of its value or an estimate of its hinge loss-based risk.

We formalize this as follows. We consider a class $\mathcal{F}_0$ of real-valued functions over $\mathcal{W}$ and derive from it two classes $\mathcal{F}^{\mathrm{Id}}$ and $\mathcal{F}^{\mathrm{hinge}}$ defined as

$$\mathcal{F}^{\mathrm{Id}} \doteq \big\{(a,w) \mapsto f_a(w) : f_1, \ldots, f_K \in \mathcal{F}_0, \\ \forall w \in \mathcal{W}, (f_1(w), ..., f_K(w)) \in \Delta(K)\big\}, \quad (5)$$

where $\Delta(K)$ is the K-dimensional probability simplex, and

$$\mathcal{F}^{\mathrm{hinge}} \doteq \big\{(a,w) \mapsto f_a(w) : f_1, \ldots f_K \in \mathcal{F}_0, \\ \forall w \in \mathcal{W}, \textstyle\sum_{a\in[K]} f_a(w) = 0\big\}. \quad (6)$$

Let ϕ^{Id} be the identity mapping and ϕ^{hinge} be the hinge mapping $x \mapsto \max(0, 1 + x)$, both over $\mathbb{R}$. Following exisiting terminology [Foster and Krishnamurthy, 2018, for instance], an element of $\mathcal{F}$ is called a regressor. Each regressor f is mapped to a policy π through a *policy mapping*, either $\widetilde{\pi}^{\mathrm{Id}}$ if $f \in \mathcal{F}^{\mathrm{Id}}$ or $\widetilde{\pi}^{\mathrm{hinge}}$ if $f \in \mathcal{F}^{\mathrm{hinge}}$ where, for all $(a,w) \in [K] \times \mathcal{W}$,

$$\widetilde{\pi}^{\mathrm{Id}}(f)(a,w) = f(a,w),$$
$$\widetilde{\pi}^{\mathrm{hinge}}(f)(a,w) = \mathbf{1}\{a = \operatorname*{arg\,max}_{a' \in [K]} f(a',w)\}.$$

For ϕ set either to ϕ^{Id} or ϕ^{hinge}, for any $f : [K] \times \mathcal{W} \to \mathbb{R}$, for every $o = (w, a, y) \in \mathcal{W} \times [K] \times \{0, 1\}$ and each $\tau \geq 1$, define

$$\ell^{\phi}(f)(o) \doteq \phi(f(a,w))(1-y),$$

$$\ell^{\phi}_{\tau}(f) \doteq \frac{g_{\mathrm{ref}}(a,w)}{g_\tau(a,w)}\phi(f(a,w))(1-y),$$

the corresponding ϕ-risk $R^{\phi}(f) \doteq E[\ell^{\phi}(f)(O^{\mathrm{ref}})] = E_P[\ell^{\phi}_{\tau}(f)(O_\tau)]$ and its empirical counterpart $\widehat{R}_t(f) \doteq t^{-1}\sum_{\tau=1}^{t} \ell^{\phi}_{\tau}(f)(O_\tau)$. Finally, the *risk* of any policy π is defined as $R(\pi) \doteq R^{\phi}(\pi)$ with $\phi = \phi^{\mathrm{Id}}$ and the *hinge-risk* of any regressor $f \in \mathcal{F}^{\mathrm{hinge}}$ is defined as $R^{\mathrm{hinge}}(f) \doteq R^{\phi}(f)$ with $\phi = \phi^{\mathrm{hinge}}$.

We can now present the ε-greedy algorithm.

Algorithm 2 ε-greedy.

Input: convex surrogate ϕ, regressor class $\mathcal{F}$, policy mapping $\widetilde{\pi}$, sequence $(\delta_t)_{t\geq 1}$.
Initialize $\widehat{\pi}_0$ as g_{ref}
for $t \geq 1$ **do**
Define policy as mixture between g_{ref} and $\widehat{\pi}_{t-1}$:

$$g_t = \delta_t g_{\mathrm{ref}} + (1 - \delta_t)\widehat{\pi}_{t-1}$$

Observe context W_t, sample action $A_t \sim g_t(\cdot|W_t)$, collect reward Y_t.
Compute optimal empirical regressor

$$\widehat{f}_t = \operatorname*{arg\,min}_{f\in\mathcal{F}} \frac{1}{t}\sum_{\tau=1}^{t} \ell^{\phi}_{\tau}(f)(O_\tau). \quad (7)$$

Compute optimal policy estimator $\widehat{\pi}_t = \widetilde{\pi}(\widehat{f}_t)$.
end for

We consider two instantiations of the algorithm: one corresponding to $(\phi^{\mathrm{Id}}, \mathcal{F}^{\mathrm{Id}}, \widetilde{\pi}^{\mathrm{Id}})$ and called *direct policy optimization*, the other corresponding to $(\phi^{\mathrm{hinge}}, \mathcal{F}^{\mathrm{hinge}}, \widetilde{\pi}^{\mathrm{hinge}})$ and called *hinge-risk optimization*.

Regret decomposition. Denote π^*_Π the optimal policy in $\Pi \doteq \widetilde{\pi}(\mathcal{F})$ and π^* any[6] optimal measurable policy. The key idea in the regret analysis of the ε-greedy algorithm is the following elementary decomposition (details in appendix E): $Y_t - R(\pi^*) =$

$$\underbrace{Y_t - E_P[Y_t|F_{t-1}]}_{\text{reward noise}} + \underbrace{\delta_t(R(g_{\mathrm{ref}}) - R(\pi^*))}_{\text{exploration cost}} \\ + (1-\delta_t)\underbrace{(R(\widehat{\pi}_{t-1}) - R(\pi^*))}_{\text{exploitation cost}}. \quad (8)$$

Control of the exploitation cost. In the direct policy optimization case, we can give exploitation cost guarantees under no assumption other than an entropy condition

[5] It is known that $\log N(\epsilon, \mathcal{F}, L_r(P))$ is smaller than $\log N_{[]}(2\epsilon, \mathcal{F}, L_r(P))$ for all $\epsilon > 0$.

[6] There may exist more than one.

on $\mathcal{F}$. In the hinge-risk optimization case, we need a so-called realizability assumption. Denote $\mathbb{R}^K_{=0} \doteq \{x \in \mathbb{R}^K : \sum_{a\in[K]} x_a = 0\}$.

Assumption 2 (Hinge-realizability)**.** *Let*

$$f^* \doteq \operatorname*{arg\,min}_{f:[K]\times\mathcal{W}\to\mathbb{R}^K_{=0}} R^{\mathrm{hinge}}(f)$$

be the minimizer over all measurable regressors of the hinge-risk. We say that a regressor class $\mathcal{F}^{\mathrm{hinge}}$ satisfies the hinge-realizability assumption for the hinge-risk if $f^* \in \mathcal{F}^{\mathrm{hinge}}$.

Imported from the theory of classification calibration, Assumption 2 allows us to bound the risk of a policy $R(\widetilde{\pi}^{\mathrm{hinge}}(f))$ in terms of the hinge-risk of the regressor f. The proof relies on the following result:

Lemma 1 (Hinge-calibration)**.** *Consider a regressor class $\mathcal{F}^{\mathrm{hinge}}$. Let*

$$\pi^* \in \operatorname*{arg\,min}_{\pi:[K]\times\mathcal{W}\to\Delta(K)} R(\pi)$$

be an optimal measurable policy. It holds that $R(\pi^*) = R(\widetilde{\pi}^{\mathrm{hinge}}(f^*))$ and, for all $f \in \mathcal{F}^{\mathrm{hinge}}$,

$$R(\widetilde{\pi}^{\mathrm{hinge}}(f)) - R(\pi^*) \le R^{\mathrm{hinge}}(f) - R^{\mathrm{hinge}}(f^*).$$

We refer the reader to Bartlett et al. [2006], Ávila Pires and Szepesvári [2016] for proofs, respectively when $K = 2$ and when $K \ge 2$. Under Assumption 2, Lemma 1 teaches us that we can bound the exploitation cost in terms of the excess hinge-risk $R^{\mathrm{hinge}}(f) - \min_{f'\in\mathcal{F}^{\mathrm{hinge}}} R^{\mathrm{hinge}}(f')$, a quantity that we can bound by standard arguments from the theory of empirical risk minimization. The fondamental building block of our exploitation cost analysis is therefore the following finite sample deviation bound for the empirical ϕ-risk minimizer.

Theorem 3 (ϕ-risk exponential deviation bound for the ε-greedy algorithm)**.** *Let ϕ and $\mathcal{F}$ be either ϕ^{Id} and $\mathcal{F}^{\mathrm{Id}}$ or ϕ^{hinge} and $\mathcal{F}^{\mathrm{hinge}}$. Suppose that $g_1, \ldots, g_t$ is a sequence of policies such that, for all $\tau \in [t]$, g_τ is $F_{\tau-1}$-measurable. Suppose that there exist $B, \delta > 0$ such that*

$$\sup_{f_1,f_2\in\mathcal{F}} \sup_{a\in[K],w\in\mathcal{W}} |\phi(f_1(a,w)) - \phi(f_2(a,w))| \le B,$$
$$\min_{\tau\in[t]} g(A_\tau, W_\tau) \ge \delta \text{ a.s.}$$

Define $f^*_{\mathcal{F}} \doteq \arg\min_{f\in\mathcal{F}} R^\phi(f)$, the $\mathcal{F}$-specific optimal regressor of the ϕ-risk, and let $\widehat{f}_t$ be the empirical ϕ-risk minimizer (7)*. Then, for all $x > 0$ and $\alpha \in (0, B)$,*

$$P\Bigg[R^\phi(\widehat{f}_t) - R^\phi(f^*_{\mathcal{F}}) \ge H_t\left(\alpha, \delta, B^2K/\delta, B\right) + 160B\sqrt{Kx/\delta t} + 3B/\delta t x\Bigg] \le 2e^{-x},$$

with $H_t(\alpha, \delta, v, B) \doteq \alpha + 160\sqrt{v/t}$

$$\times \int_{\alpha/2}^{B} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))}d\epsilon + \frac{3B}{\delta t}\log 2.$$

As a direct corollary, we can express rates of convergence for the ϕ-risk in terms of the bracketing entropy rate.

Corollary 1. *Suppose that $\log(1+N_{[]}(\epsilon, \mathcal{F}, L_2(P))) = O(\epsilon^{-p})$ for some $p \in (0, 1)$. Then*

$$R^\phi(\widehat{f}_t) - R^\phi(f^*_{\mathcal{F}}) = O_P\left((\delta t)^{-\left(\frac{1}{2}\wedge\frac{1}{p}\right)}\right).$$

Control of the regret. The cumulative reward noise $\sum_{\tau=1}^t (Y_\tau - E_P[Y_\tau|F_{\tau-1}])$ can be bounded by the Azuma-Hoeffding inequality. From (16) and Corollary 1, δ_t controls the trade off between the exploration and exploitation costs. We must therefore choose a δ_t that minimizes the total of these two which, from the above, scales as $O(\delta_t + (t\delta_t)^{-(\frac{1}{2}\wedge\frac{1}{p})})$. The optimal choice is $\delta_t \propto t^{-(\frac{1}{3}\wedge\frac{1}{p+1})}$. The following theorem formalizes the regret guarantees under the form of a high-probability bound.

Theorem 4 (High probability regret bound for ε-greedy.)**.** *Suppose that the bracketing entropy of the regressor class $\mathcal{F}$ satisfies $\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P)) = O(\epsilon^{-p})$ for some $p > 0$. Set $\delta_t = t^{-(\frac{1}{3}\vee\frac{p}{p+1})}$ for all $t \ge 1$. Suppose that*

- *either $\phi = \phi^{\mathrm{Id}}$, $\mathcal{F}$ is of the form $\mathcal{F}^{\mathrm{Id}}$, $\widetilde{\pi} = \widetilde{\pi}^{\mathrm{Id}}$,*
- *or $\phi = \phi^{\mathrm{hinge}}$, $\mathcal{F}$ is of the form $\mathcal{F}^{\mathrm{hinge}}$, $\widetilde{\pi} = \widetilde{\pi}^{\mathrm{hinge}}$, and $\mathcal{F}$ satisfies Assumption 2.*

Then, with probability $1 - \epsilon$,

$$\sum_{\tau=1}^t (\mathcal{V}(\pi^*) - Y_\tau) \le \sqrt{t\log(2/\epsilon)} + t^{\frac{p}{p+1}}\sqrt{\log(2t(t+1)/\epsilon)}.$$

4 EXAMPLES OF POLICY CLASSES

4.1 A NONPARAMETRIC ADDITIVE MODEL

We say that $a(\epsilon) = \widetilde{O}(b(\epsilon))$ if there exists $c > 0$ such that $a(\epsilon) = O(b(\epsilon)\log^c(1/\epsilon))$. We present a policy class that has entropy $\widetilde{O}(\epsilon^{-1})$, and over which the two optimization oracles presented in Definitions 1 and 2 reduce

to linear programs. Let $\mathbb{D}([0,1])$ be the set of càdlàg functions and let the variation norm $\|\cdot\|_v$ be given, for all $h \in \mathbb{D}([0,1])$, by

$$\|h\|_v \doteq \sup_{m\geq 2} \sup_{x_1,\ldots,x_m} \sum_{i=1}^{m-1} |h(x_{i+1}) - h(x_i)|$$

where the right-hand side supremum is over the subdivisions of $[0,1]$, that is over $\{(x_1,\ldots,x_m) : 0 \leq x_1 \leq \ldots \leq x_m \leq 1\}$. Set $C, M > 0$ then introduce

$$\mathcal{H} \doteq \{h \in \mathbb{D}([0,1]) : \|h\|_v \leq M\}$$

and the additive nonparametric additive model derived from it by setting $\mathcal{F}_0 \doteq$

$$\Big\{(a,w) \mapsto \sum_{l=1}^{d} \alpha_{a,l} h_l(w_l) : |\alpha_{a,l}| \leq C, h_{a,l} \in \mathcal{H}\Big\}.$$

Let $\mathcal{F} = \mathcal{F}^{\mathrm{Id}}$ derived from $\mathcal{F}_0$ as in (5).

The following lemma formally bounds the entropy of the policy class.

Lemma 2. *There exists $\epsilon_0 \in (0,1)$ such that, for all $\epsilon \in (0,\epsilon_0)$,*

$$\begin{aligned}\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq & K \log N_{[]}(\epsilon, \mathcal{F}_0, \|\cdot\|_\infty) \\ \leq & K c_0 \epsilon^{-1} \log(1/\epsilon).\end{aligned}$$

for some $c_0 > 0$ depending on (C, d, M).

We now state a result that shows that LCLSO and LCCSCO reduce to linear programs over $\mathcal{F}$. We first need to state a definition.

Definition 3 (Grid induced by a set of points)**.** *Consider d subdivisions of $[0,1]$ of the form*

$$\begin{aligned}0 = & w_{1,1} \leq w_{1,2} \leq \ldots \leq w_{1,q_1} = 1, \\ & \vdots \\ 0 = & w_{d,1} \leq w_{1,2} \leq \ldots \leq w_{d,q_d} = 1.\end{aligned}$$

The rectangular grid induced by these d subdivisions *is the set of points $(w_{1,i_1}, w_{2,i_2}, \ldots, w_{i,i_d})$ with $i_1 \in [q_1], ..., i_d \in [q_d]$. We call a* rectangular grid *any rectangular grid induced by some set of d subdivisions of $[0,1]$.*

Consider a set of points $w_1, \ldots, w_n \in [0,1]^d$. A minimal grid induced by $w_1, \ldots w_n$ is any rectangular grid that contains $w_1, \ldots w_n$ and that is of minimal cardinality. We denote by $G(w_1, \ldots, w_n)$ a minimal rectangular grid induced by $w_1, \ldots w_n$ chosen arbitrarily.

Lemma 3. *Let $w_0 = \mathbf{0}, w_1, \ldots, w_t \in [0,1]^d$. For all $l \in [d]$, let $\widetilde{\mathcal{H}}_{l,t} \doteq \widetilde{\mathcal{H}}_{l,t}(w_{0,l}, \ldots, w_{t,l}) \doteq$*

$$\Big\{x \mapsto \sum_{\tau=0}^{t} \beta_\tau 1\{x \geq w_{\tau,l}\} : \beta_\tau \in \mathbb{R}, \sum_{\tau=0}^{t} |\beta_\tau| \leq M\Big\}$$

and $\widetilde{\mathcal{F}}_{0,t} \doteq$

$$\Big\{(a,w) \mapsto \sum_{l=1}^{d} \alpha_{a,l} \widetilde{h}_{a,l}(w_l) : |\alpha_{a,l}| \leq B, \widetilde{h}_{a,l} \in \mathcal{H}_{l,t}\Big\}.$$

Let $(u_{a,\tau})_{a\in[K],\tau\in[t]}$ be a vector in $\mathbb{R}^{Kt}$. Let $\widetilde{f}^*$ be a solution to the following optimization problem $(\mathcal{P}_2)$:

$$\max_{\widetilde{f}\in\widetilde{\mathcal{F}}_{0,t}} \sum_{a\in[K]} \sum_{\tau=1}^{t} u_{a,\tau} \widetilde{f}(a, W_\tau)$$

$$\text{s.t. } \forall a \in [K],\ \forall w \in \mathcal{G}(w_0,\ldots,w_t),\ \widetilde{f}(a,w) \geq 0, \quad (9)$$

$$\forall w \in \mathcal{G}(w_0,\ldots,w_t),\ \sum_{a\in[K]} \widetilde{f}(a,w) = 1. \quad (10)$$

Then, $\widetilde{f}$ is a solution to the following optimization problem $(\mathcal{P}_1)$:

$$\max_{f\in\mathcal{F}_0} \sum_{a\in[K]} \sum_{\tau=1}^{t} u_{a,\tau} f(a, W_\tau)$$

$$\text{s.t. } \forall a \in [K], \forall w \in [0,1]^d, f(a,w) \geq 0, \quad (11)$$

$$\forall w \in [0,1]^d,\ \sum_{a\in[K]} f(a,w) = 1. \quad (12)$$

4.2 CÀDLÀG POLICIES WITH BOUNDED SECTIONAL VARIATION NORM

The class of d-variate càdlàg functions with bounded sectional variation norm is a nonparametric function class with bracketing entropy bounded by $O(\epsilon^{-1}\log(1/\epsilon)^{2(d-1)})$, over which empirical risk minimization takes the form of a LASSO problem. It has received attention recently in the nonparametric statistics literature [van der Laan, 2016, Fang et al., 2019, Bibaut and van der Laan, 2019]. Empirical risk minimizers over this class of functions have been termed Highly Adaptive Lasso estimators by van der Laan [2016]. The experimental study of Benkeser and van der Laan [2016] suggests that Highly Adaptive Lasso estimators are competitive against supervised learning algorithms such as Gradient Boosting Machines and Random Forests.

Sectional variation norm. For a function $f : [0,1]^d \to \mathbb{R}$, and a non-empty subset s of $[d]$, we call the s-section of f and denote f_s the restriction of f to $\{x \in [0,1]^d : \forall i \in s, x_i = 0\}$. The sectional variation norm (svn) is defined based on the notion of Vitali variation. Defining the notion of Vitali variation in full generality requires introducing additional concepts. We thus relegate the full definition to appendix H, and present it in a particular case. The Vitali variation of an m-times continuously differentiable function $g : [0,1]^m \to \mathbb{R}$ is

defined as

$$V^{(m)}(g) \doteq \int_{[0,1]^m} \left| \frac{\partial^m g}{\partial x_1 \dots \partial x_m} \right|.$$

For arbitrary real-valued càdlàg functions g on $[0,1]^m$ (non necessarily m times continuously differentiable), the Vitali variation $V^{(m)}(g)$ is defined in appendix H. The svn of a function $f : [0,1]^d \to \mathbb{R}$ is defined as

$$\|f\|_v \doteq |f(0)| + \sum_{\emptyset \neq s \subset [d]} V^{(|s|)}(f_s),$$

that is the sum of its absolute value at the origin and the sum of the Vitali variation of its sections. Let $\mathbb{D}([0,1]^d)$ be the class of càdlàg functions with domain $[0,1]^d$ and, for some $M > 0$, let

$$\mathcal{F}_0 \doteq \left\{ f \in \mathbb{D}([0,1]^d) : \|f\|_v \leq M \right\} \tag{13}$$

be the class of càdlàg functions with svn smaller than M.

Entropy bound. The following result is taken from [Bibaut and van der Laan, 2019].

Lemma 4. *Consider $\mathcal{F}_0$ defined in* (13)*. Let P be a probability distribution over $[0,1]^d$ such that $\|\cdot\|_{P,2} \leq c_0 \|\cdot\|_{\mu,2}$, with μ the Lebesgue measure and $c_0 > 0$. Then there exist $c_1 > 0, \epsilon_0 \in (0,1)$ such that, for all $\epsilon \in (0, \epsilon_0)$ and all distributions P over $[0,1]^d$,*

$$\log N_{[]}(\epsilon, \mathcal{F}_0, L_2(P)) \leq c_1 M \epsilon^{-1} \log(M/\epsilon)^{2d-1}.$$

Representation of ERM. We show that empirical risk minimization (ERM) reduces to linear programming in both our direct policy and hinge-risk optimization settings.

Lemma 5 (Representation of the ERM in the direct policy optimization setting)**.** *Consider a class of policies of the form $\mathcal{F}^{\mathrm{Id}}$* (5) *derived from $\mathcal{F}_0$* (13)*. Let $\phi = \phi^{\mathrm{hinge}}$. Suppose we have observed (W_1, A_1, Y_1), …, (W_t, A_t, Y_t) and let $\widetilde{W}_1, \dots, \widetilde{W}_m$ be the elements of $G(W_1, \dots, W_t)$. Let $(\beta_j^a)_{a \in [K], j \in [m]}$ be a solution to*

$$\begin{aligned}
\min_{\beta \in \mathbb{R}^{Km}} \sum_{\tau=1}^{t} \sum_{a \in [K]} \Bigg\{ & \frac{1\{A_\tau = a\}}{g_\tau(A_\tau, W_\tau)} (1 - Y_\tau) \\
& \times \sum_{j=1}^{m} \beta_j^a 1\{W_\tau \geq \widetilde{W}_j\} \Bigg\} \\
\text{s.t. } \forall l \in [m],\ & \sum_{a \in [K]} \sum_{j=1}^{m} \beta_j^a 1\{\widetilde{W}_l \geq \widetilde{W}_j\} = 1, \\
\forall l \in [m], \forall a \in [K],\ & \sum_{j=1}^{m} \beta_j^a 1\{\widetilde{W}_l \geq \widetilde{W}_j\} \geq 0, \\
\forall a \in [K],\ & \sum_{j=1}^{m} |\beta_j^a| \leq M.
\end{aligned} \tag{14}$$

Then $f : (a, w) \mapsto \sum_{j=1}^{m} \beta_j^a 1\{w \geq \widetilde{W}_j\}$ is a solution to $\min_{f \in \mathcal{F}^{\mathrm{Id}}} \sum_{\tau=1}^{t} \ell_\tau^\phi(f)(O_\tau)$.

We present a similar result for the hinge-risk setting in appendix H. It is relatively easy to prove with the same techniques that ERM over $\mathcal{F}^{\mathrm{hinge}}$ also reduces to linear programming when $\mathcal{F}_0$ is an RKHS.

5 CONCLUSION

We present the first efficient CB algorithm that is regret-optimal against policy classes with *polynomial* entropy. We acknowledge that our algorithm might not be practical. It inherits some of the caveats of PE: (1) the probability of the regret bound is a pre-specified parameter, (2) if the algorithm eliminates the best policy, it never recovers.

We conjecture that regret optimality could be proven for classes with non-integrable entropy. The role of integrability is purely technical and due to our proof techniques.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Beijing, China, 2014. PMLR.

S. Athey and S. Wager. Efficient policy learning, 2017. arXiv preprint arXiv:1702.02896v5.

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

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

B. Ávila Pires and C. Szepesvári. Multiclass classification calibration functions, 2016. arXiv preprint arXiv:1609.06385v1.

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. Benkeser and M. J. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696, 2016.

A. F. Bibaut and M. J. van der Laan. Fast rates for empirical risk minimization over càdlà functions with bounded sectional variation norm, 2019. arXiv preprint arXiv:1907.09244v2.

N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 465–481, Amsterdam, Netherlands, 2017. PMLR.

N. Chatterji, A. Pacchiano, and P. Bartlett. Online learning with kernel losses. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 971–980, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI11, pages 169–178, Arlington, Virginia, USA, 2011. AUAI Press.

M. Dudík, N. Haghtalab, H. Luo, Schapire R. E., V. Syrgkanis, and J. Wortman Vaughan. Oracle-efficient online learning and auction design. pages 528–539. IEEE Computer Society, 2017.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation, 2019. arXiv preprint arXiv:1903.01395v2.

D. Foster, A. Agarwal, M. Dudik, H. Luo, and R. E. Schapire. Practical contextual bandits with regression oracles. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1539–1548, Stockholmsmssan, Stockholm Sweden, 2018. PMLR.

D. J. Foster and A. Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2621–2632. Curran Associates, Inc., 2018.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008.

N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2): 212–261, 1994.

A. R. Luedtke and A. Chambaz. Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré – Probabilité et Statistiques*, 0(0), 2019.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41(2):693–721, 04 2013.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In A. Tauman Kalai and M. Mohri, editors, *COLT*, pages 54–66, Haifa, Israel, 2010. Ominipress.

S. A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

M. J. van der Laan. A generally efficient TMLE. *The International Journal of Biostatistics*, 1(1), 2016.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

R. van Handel. On the minimal penalty for Markov order estimation. *Probability Theory and Related Fields*, 150:709–738, 2011.

V. G. Vovk. Aggregating strategies. In M. A. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990*, pages 371–386. Morgan Kaufmann, 1990.