Journal of Machine Learning Research 21 (2020) 1-45 Submitted 8/19; Revised 6/20; Published 7/20

Contextual Bandits with Continuous Actions: Smoothing, Zooming, and Adapting

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research
New York, NY 10011

John Langford JCL@MICROSOFT.COM
Microsoft Research
New York, NY 10011

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research
New York, NY 10011

Chicheng Zhang CHICHENGZ@CS.ARIZONA.EDU
University of Arizona
Tucson, AZ 85721

Editor: Manfred Warmuth

Abstract

We study contextual bandit learning with an abstract policy class and continuous action space. We obtain two qualitatively different regret bounds: one competes with a smoothed version of the policy class under no continuity assumptions, while the other requires standard Lipschitz assumptions. Both bounds exhibit data-dependent “zooming” behavior and, with no tuning, yield improved guarantees for benign problems. We also study adapting to unknown smoothness parameters, establishing a price-of-adaptivity and deriving optimal adaptive algorithms that require no additional information.

Keywords: Contextual bandits, nonparametric learning

1. Introduction

We consider contextual bandits, a setting in which a learner repeatedly makes an action on the basis of contextual information and observes a loss for the action, with the goal of minimizing cumulative loss over a series of rounds. Contextual bandit learning has received much attention, and has seen substantial success in practice (e.g., Auer et al., 2002; Langford and Zhang, 2007; Agarwal et al., 2014, 2017a). This line of work mostly considers small, finite action spaces, yet in many real-world problems actions are chosen from an interval, so the action space is continuous and infinite. Therefore, we ask:

> *How can we learn to make decisions from continuous action spaces,*
> *using (only) bandit feedback?*

We could assume that nearby actions have similar losses, for example that the losses are Lipschitz continuous as a function of the action (following Agrawal, 1995, and a long line of subsequent work). Then we could discretize the action space and apply generic contextual bandit techniques (Kleinberg, 2004) or more refined "zooming" approaches (Kleinberg et al., 2019; Bubeck et al., 2011a; Slivkins, 2014) that are specialized to the Lipschitz structure.

However, this approach has several drawbacks. A global Lipschitz assumption is crude and limiting; actual problems exhibit more complex loss structures where smoothness varies with location, often with discontinuities. Second, prior works incorporating context — including the zooming approaches — employ a nonparametric benchmark set of policies, which yields a poor dependence on the context dimension and prevents application beyond low-dimensional context spaces. Finally, existing algorithms require knowledge of the Lipschitz constant or other pertinent parameters, which are typically unknown.

Here we show that it is possible to avoid all of these drawbacks with a conceptually new approach, resulting in a more robust solution for managing continuous action spaces. The key idea is to *smooth* the actions: each action a is mapped to a well-behaved distribution over actions. When the action space is the interval $[0,1]$, this distribution can be a uniform distribution over a narrow band around a: an interval $[a-h, a+h]$, where $h>0$ is a given *bandwidth* parameter. Rather than restrict the loss function, we posit a different, "smoothed" benchmark. This approach leads to provable guarantees with no assumptions on the loss function, since the loss for smoothed actions is always well-behaved. Essentially, we may focus on estimation considerations while ignoring approximation issues. We recover prior results that assume a small Lipschitz constant, but the guarantees are meaningful in much broader scenarios.

Our algorithms work with any competitor policy set Π of mappings from context to actions, which we smooth as above. We measure performance by comparing the learner's loss to the loss of the best smoothed policy, and our guarantees scale with $\log|\Pi|$, regardless of the dimensionality of the context space. Compared with prior work, this recovers some known worst-case results that can only accommodate nonparametric policy sets (Slivkins, 2014; Cesa-Bianchi et al., 2017), but, more importantly, our results accommodate *parametric* policy sets that scale to high-dimensional context spaces. Further, we are able to exploit benign structure in the policy set and the instance to obtain better regret rates.

We also design algorithms that require no knowledge of problem parameters. Particularly, our algorithm works for all bandwidths h at once, and is *optimally adaptive*, matching lower bounds that we prove here. We accomplish this with a unified algorithmic approach.

Our contributions, specialized to the interval $[0,1]$ action space for clarity, are:

1. We define a new notion of *smoothed regret* where policies map contexts to distributions over actions. These distributions are parametrized by a bandwidth h governing the spread. We show that the optimal worst-case regret bound with bandwidth h is $\Theta(\sqrt{T/h \log|\Pi|})$, which requires no smoothness assumptions on the losses (first row of Table 1).

2. We obtain instance-dependent guarantees in terms of a *smoothing coefficient*, which can yield much faster rates in favorable instances (second row of Table 1).

3. We obtain an adaptive algorithm with $\sqrt{T}/h$ regret bound for all bandwidths h simultaneously. Further we show this to be optimal, demonstrating a price of adaptivity (third row of Table 1).

Type	Setting	Params	Regret Bound	Status	Sec.
Smoothed	Worst-case	$h \in (0,1]$	$\Theta\left(\sqrt{T/h}\right)$	New	4.1
Smoothed	Instance-dependent	$h \in (0,1]$	$O\left(\min_\epsilon T\epsilon + \theta_h(\epsilon)\right)$	New	4.1
Smoothed	Adaptive: $h \in (0,1]$	None	$\Theta\left(\sqrt{T}/h\right)$	New	4.2
Lipschitz	Worst-case	$L \geq 1$	$\Theta\left(T^{2/3}L^{1/3}\right)$	"Old"	5.1
Lipschitz	Instance-dependent	$L \geq 1$	$O\left(\min_\epsilon TL\epsilon + \psi_L(\epsilon)/L\right)$	New	5.1
Lipschitz	Adaptive: $L \geq 1$	None	$\Theta(T^{2/3}\sqrt{L})$	New	5.2

Table 1: A summary of results for stochastic contextual bandits, specialized to action space $[0,1]$. For notation, T is the number of rounds, h is the smoothing bandwidth, and $\theta_h(\epsilon) \leq 1/(h\epsilon)$ is the *smoothing coefficient*. For the Lipschitz results, L is the Lipschitz constant and $\psi_L(\epsilon) \leq 1/\epsilon^2$ is the *policy zooming coefficient*. All algorithms take T and Π as additional inputs. Logarithmic dependence on $|\Pi|$ and T is suppressed in all upper bounds.

We obtain analogous results when the losses are L-Lipschitz (see rows 3-6 of Table 1). First, we obtain an instance-dependent result with improved regret rates when near-optimal arms are confined to a relatively small region of the action space. We capture the improvements via a new quantity called the *policy zooming coefficient*, generalizing the *zooming dimension* from prior work on the non-contextual case. Our regret bounds generalize and improve those from prior work on "zooming" in Lipschitz bandits, whereby the algorithm gradually "zooms in" on more promising regions of the action space. Second, we design an algorithm that adapts to an unknown L and obtain matching lower bounds, thus demonstrating the "price of adaptivity" in the Lipschitz case.

Our results hold in much more general settings: for higher-dimensional and (almost) arbitrary action spaces and arbitrary smoothing distributions. Our results also apply to the non-contextual case, where we obtain several new guarantees.

Our algorithms are not computationally efficient, with running times that scale polynomially in $|\Pi|$. The significance lies is in the new conceptual approach and the regret bounds. However, our algorithms *are* computationally efficient in the non-contextual case.

Our techniques. Our core conceptual contribution is the new definition of smoothed regret for continuous-action contextual bandits, which, as we have mentioned, offers many advantages over previous discretization based approaches. While many of our results are based on adapting techniques from prior work to the smoothing framework, there are many technical challenges that we pause now to highlight.

Our instance dependent guarantees are based on the `PolicyElimination` algorithm of Dudik et al. (2011), which was originally designed for discrete action stochastic contextual bandits. Here we provide a refined analysis of this algorithm, showing that it adapts to the effective size of the action space, which informally corresponds to the number of actions selected by the near-optimal policies. To obtain this adaptivity property, we crucially use the median-of-means technique to avoid an unfavorable range dependence in our estimates of the expected loss of each policy. We believe these robust estimation techniques will be broadly useful in other bandit settings. Indeed, since the preliminary version of this paper, robust estimators have been successfully used by Wei et al. (2020) to incorporate loss predictors into contextual bandit algorithms.

Our adaptive algorithms are based on aggregating instances of EXP4 (Auer et al., 2002) using the Corral algorithm of Agarwal et al. (2017b). The key challenge here is that Corral can only aggregate over a finite number of base algorithm, but we would like our final bound to hold for all bandwidths h taking continuous values. We address this with a discretization argument, using smoothing to show that a single instance of EXP4 obtains the desired guarantee for a small interval of h values, which then allows us to use Corral with a finite number of base algorithms.

Roadmap. For the majority of the paper, we focus on the setting where the action space is the unit interval, which simplifies the discussion while preserving all of the key ideas. The setup and key definitions are described in Section 3. Assumption-free results for smoothed regret are developed in Section 4 and results for Lipschitz problems are developed in Section 5. General theorems extending beyond the unit interval action space are presented in Section 6, where we also introduce the necessary additional definitions. The algorithms are analyzed in Section 7 and Section 8. The lower bounds are presented in Section 9. We close the paper with some future directions.

2. Related work

With small, discrete action spaces, contextual bandit learning is quite mature, with rich theoretical results and successful deployments in practice. To handle large or infinite action spaces, two high-level approaches exist (see books Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020, for surveys and background). The parametric approach, including work on linear or combinatorial bandits, posits that the loss is a parametric function of the action, e.g., a linear function. The nonparametric approach, which is closer to our results, typically makes much weaker continuity assumptions.[1]

Bandits with Lipschitz assumptions were introduced in Agrawal (1995), and optimally solved in the worst case by Kleinberg (2004). Kleinberg et al. (2008, 2019); Bubeck et al. (2011a) achieve data-dependent regret bounds via "zooming" algorithms which gradually "zoom in" on the more promising regions of the action space. Kleinberg et al. (2008, 2019); Kleinberg and Slivkins (2010) consider regret rates with instance-dependent constant, analogous to the well-known $\log(t)$ instance-dependent rates for finitely many arms, and use zooming algorithms to characterize the corresponding worst-case optimal regret rates for any given metric space. Further work focused on relaxing the smoothness assumptions and adapting to unknown smoothness parameters, as well as extensions to contextual bandits (see Ch. 4 Slivkins, 2019, for a more comprehensive background).

Several papers relax global smoothness assumptions with various local definitions (Auer et al., 2007; Kleinberg et al., 2008, 2019; Bubeck et al., 2011a; Slivkins, 2011; Valko et al., 2013; Minsker, 2013; Grill et al., 2015; Shang et al., 2019). While the assumptions and results vary, our smoothing-based approach can be used in many of these settings. More importantly, in contrast with these approaches, our guarantees remain meaningful even in pathological instances, for example when the global optimum is a discontinuity as in the top panel of Figure 1 (See Example 2).

[1] However, we emphasize that for smoothed regret, we make no assumptions on the loss.

While most of this literature focuses on the non-contextual version, three papers consider contextual settings, albeit only with fixed policy sets Π. Lu et al. (2010) and Slivkins (2014) posit that the mean loss function is Lipschitz in both context x and action a and the learner must compete with the best mapping from $\mathcal{X}$ to $\mathcal{A}$. While Lu et al. (2010) focus on worst-case regret bounds, the algorithm and guarantees in Slivkins (2014) exhibit "zooming" behavior in the action space, which is qualitatively similar to ours. However, his regret bound also has a zooming-dependence on the context dimension, whereas our regret bound applies to arbitrary policy sets and defines packing numbers via expectation over contexts rather than supremum. Cesa-Bianchi et al. (2017) competes with policies that are themselves Lipschitz (w.r.t. a given metric on contexts). We can recover their result via Corollary 7 and a suitable discretized policy set.

Turning to adaptivity, Bubeck et al. (2011b) develops an algorithm that adapts to the Lipschitz constant in the non-contextual setting given a bound on the second derivative. Locatelli and Carpentier (2018) obtain optimal adaptive algorithms, but require knowledge of either the value of the minimum, or a sharp bound on the achievable regret. Slivkins (2011); Bull (2015) achieve optimal regret bounds in terms of the zooming dimension, but their regret bounds depend on a certain "quality parameter." A line of work studying the non-contextual setting (Valko et al., 2013; Grill et al., 2015; Shang et al., 2019), establishes adaptive guarantees when performance is measured in terms of optimization error, which is the difference between the best action selected and the globally optimal action. However, these results do not translate to our performance measure, cumulative regret. Moreover, all of the above results concern the stochastic setting, while our optimally adaptive guarantees carry through to the adversarial setting. Locatelli and Carpentier (2018) also obtain lower bounds against adapting to the smoothness exponent, and we build on their construction for our lower bounds.

A parallel line of work on Bayesian optimization, considers the related problem of maximizing either a sample from a Gaussian process, or a function with bounded norm in some Reproducing Kernel Hilbert Space (RKHS) (Srinivas et al., 2012). The conceptual difference with our work is that these results impose regularity assumptions on the problem, in the same vein as prior work with Lipschitz assumptions, while we make no assumptions and instead provide guarantees in terms of smoothed regret. On the more technical side, Krause and Ong (2011) consider a contextual Bayesian optimization setting where there is a kernel over the joint context-action space, which is analogous to the Lipschitz contextual bandits setting studied by Slivkins (2014). As mentioned above, these results consider a specific "nonparametric" policy set, while our results apply to arbitrary policy sets. Berkenkamp et al. (2019) establish adaptive guarantees for Bayesian optimization, but they obtain incomparable results using very different techniques from ours.

Finally, our smoothing-based importance weighted loss estimator (5) was analyzed by Kallus and Zhou (2018); Chen et al. (2016) in the related off-policy evaluation problem, but they do not consider the smoothed regret benchmark or the online setting, so the results are considerably different. We also use the median-of-means approach from robust statistics — specifically a result of Hsu and Sabato (2016) — to avoid an unfavorable range dependence in our loss estimator. This estimator has been used by Sen et al. (2018) for contextual bandits with discrete actions, but their results are incomparable to ours.

3. Smoothed regret

We work in a standard setup for stochastic contextual bandits. We have a context space $\mathcal{X}$, action space $\mathcal{A}$, a (possibly large but finite) policy set $\Pi : \mathcal{X} \to \mathcal{A}$, and a distribution $\mathcal{D}$ over context/loss pairs $\mathcal{X} \times \{\text{functions } \mathcal{A} \to [0,1]\}$. The protocol proceeds for T rounds where in each round t: (1) nature samples $(x_t, \ell_t) \sim \mathcal{D}$; (2) the learner observes x_t and chooses an action $a_t \in \mathcal{A}$; (3) the learner suffers loss $\ell_t(a_t)$, which is observed. For simplicity, we focus on the case when $\mathcal{D}_X$, the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$ is known.[2] The learner's goal is to minimize regret relative to the policy class.

Key new definitions. We depart from the standard setup by positing a *smoothing operator*

$$\texttt{Smooth}_h : \mathcal{A} \to \Delta(\mathcal{A}),$$

where $\Delta(\mathcal{A})$ is the set of probability distributions over $\mathcal{A}$ and $h \geq 0$ is the *bandwidth*: a parameter that determines the spread of the distribution.[3] Bandwidth $h = 0$ corresponds to the Dirac distribution. Each action a then maps to the *smoothed action* $\texttt{Smooth}_h(a)$, and each policy $\pi \in \Pi$ maps to a randomized *smoothed policy* $\texttt{Smooth}_h(\pi) : x \mapsto \texttt{Smooth}_h(\pi(x))$. We compete with the *smoothed policy class*

$$\Pi_h := \{\texttt{Smooth}_h(\pi) : \ \pi \in \Pi\}.$$

We then define the *smoothed loss* of a given policy $\pi \in \Pi$ and the *benchmark* optimal loss as

$$\lambda_h(\pi) := \mathop{\mathbb{E}}_{(x,\ell)\sim\mathcal{D}} \ \mathop{\mathbb{E}}_{a\sim\texttt{Smooth}_h(\pi(x))} [\, \ell(a) \,], \quad \text{and} \quad \texttt{Bench}(\Pi_h) := \inf_{\pi\in\Pi} \lambda_h(\pi) = \inf_{\pi\in\Pi_h} \lambda_0(\pi). \tag{1}$$

Note that there is a duality between smoothing the policy class and smoothing the loss function, as $\lambda_h(\pi) = \lambda_0(\pi_h)$. We are interested in *smoothed regret*, which compares the learner's total loss against the benchmark:

$$\texttt{Regret}(T, \Pi_h) := \mathbb{E}\left[\textstyle\sum_{t=1}^{T} \ell_t(a_t) \right] - T \cdot \texttt{Bench}(\Pi_h).$$

Our regret bounds work for an arbitrary policy set Π, leaving the choice of Π to the practitioner. For comparison, a standard benchmark for contextual bandits is $\texttt{Bench}(\Pi)$, the best policy in the original policy class Π, and one is interested in $\texttt{Regret}(T, \Pi)$.

For the first several sections of the paper, we posit that the actions set is a unit interval: $\mathcal{A} := [0,1]$, endowed with a metric $\rho(a, a') := |a - a'|$. $\texttt{Smooth}_h(a)$ is defined as a uniform distribution over the closed ball $\texttt{B}_h(a) := \{a' \in \mathcal{A} : \ \rho(a, a') \leq h\} = [a - h, a + h] \cap [0,1]$. Let ν denote the Lebesgue measure, which corresponds to the uniform distribution over $[0,1]$. As notation, $\texttt{Smooth}_{\pi,h}(a|x)$ is the probability density, w.r.t., ν, for $\texttt{Smooth}_h(\pi(x))$ at action a. In Section 6 we present results that apply to a more general setting where the action space $\mathcal{A}$ is embedded in some ambient space and the smoothing operator is given by a probability kernel. However, all of the key ideas appear in the case of the unit interval.

For some intuition, the bandwidth h governs a bias-variance tradeoff inherent in the continuous-action setting: for small h the smoothed loss $\lambda_h(\pi)$ closely approximates the true loss $\lambda_0(\pi)$, but small h also admits worse smoothed regret guarantees.

[2] We mention how this can be relaxed in the next section.

[3] The term *bandwidth* here is in line with the nonparametric statistics literature.

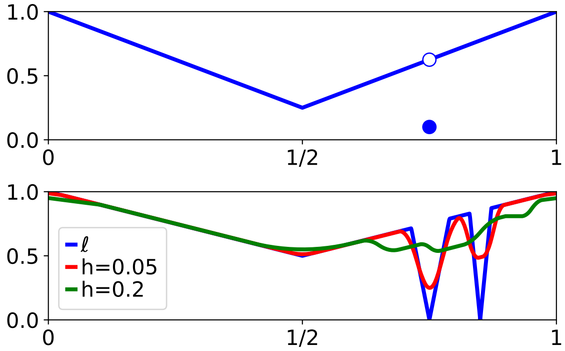

Figure 1: The discontinuous function in Example 2. Smoothed regret provides a meaningful guarantee, competing with $a_h^\star = 1/2$.

The loss function (in blue) has large Lipschitz constant and "needles" that are hard to find. Smoothing with small bandwidth does not change the optimum while a large bandwidth does.

Example 1 *The well-studied non-contextual version of the problem fits into our framework as follows: there is only one context $\mathcal{X} := \{x_0\}$ and policies are in one-to-one correspondence with actions: $\Pi := \{\, x_0 \mapsto a : a \in \mathcal{A} \,\}$. A problem instance is characterized by the expected loss function $\lambda_0(a) := \mathbb{E}[\ell(a)]$ and the smoothed benchmark is simply $\texttt{Bench}(\Pi_h) := \inf_{a\in\mathcal{A}} \lambda_h(a)$.*

Smoothing the policy class enables meaningful guarantees in much more general settings than prior work assuming global continuity (e.g., Lipschitzness). Our results require no smoothness assumptions on the loss function, in the spirit of the assumption-free analyses typical in the online learning literature. Our smoothed regret guarantees can be translated to standard regret bounds under significantly weaker assumptions than global smoothness; for example smoothness around the actions taken by the optimal policy suffices. Moreover, the guarantees remain meaningful even when the expected loss function has discontinuities, as demonstrated by the following example.

Example 2 *Consider a family of non-contextual settings with expected loss function*

$$\lambda_0(a) = (1/4 + 1.5\,\rho(a, 1/2)) \cdot \mathbf{1}_{\{a \neq a'\}} + 1/10 \cdot \mathbf{1}_{\{a = a'\}}, \qquad a' \in [0,1]$$

(see Figure 1). The optimal action $a^\star = a'$ cannot be found in finitely many rounds due to the discontinuity, so any algorithm is doomed to linear regret. However, the smoothed loss function λ_h for any $h > 0$ essentially ignores the discontinuity (and is minimized at $a_h^* = 1/2$). Accordingly, as we shall prove, it admits algorithms with sublinear smoothed regret.

While the above example is pathological, discontinuous loss functions are common in applications. One generic example is, when the algorithm controls the system parameters in a computer or a data center, even a small change can make a large difference when resources are close to saturation. For a more mathematically concrete example, consider the well-studied dynamic pricing problem (Kleinberg and Leighton, 2003), where the algorithm is a seller with an infinite inventory of identical goods. In each round the algorithm sets a price $p_t \in [0,1]$ for an item, a buyer arrives with value $v_t \in [0,1]$, and purchases the item if only if $p_t \le v_t$. The algorithm's goal is to maximize[4] the total revenue, $\sum_{t=1}^{T} p_t \cdot \mathbf{1}_{\{p_t \le v_t\}}$. So, we have a discontinuity at $v_t = p_t$, even though the payoffs are 1-Lipschitz everywhere

[4] To reformulate the problem in terms of losses, posit $\ell(p_t, v_t) = v_t - p_t \cdot \mathbf{1}_{\{p_t \le v_t\}}$.

else. More complex discontinuity structures can arise if the algorithm is selling multiple products at once, as the buyers can switch from one product to another.

The bottom panel of Figure 1 provides further intuition for the $\texttt{Smooth}_h$ operator.

Adversarial losses. Some of our results carry over as is to the adversarial setting in which the context-loss pairs are chosen by an adaptive adversary. The benchmark is redefined as

$$\texttt{Bench}(\Pi_h) := \tfrac{1}{T}\ \inf_{\pi\in\Pi_h} \mathbb{E}\left[\textstyle\sum_{t\in[T]} \ell_t(\pi(x_t))\right].$$

where the expectation accounts for any randomness. We will always explicitly specify which results apply to this setting.

Additional notation. We use $\mathbb{E}_{x\sim\mathcal{D}_X}[\,\cdot\,]$ to denote expectation over the marginal distribution over contexts. We use the standard big-Oh notation and use the notation $g=\tilde{O}(f)$ to denote that $g=O(f\cdot\mathrm{polylog}(f))$.

4. Smoothed regret guarantees

In this section we obtain smoothed-regret guarantees without imposing any continuity assumptions on the problem.

4.1. Instance-dependent and worst-case guarantees

Our first result is an instance-dependent smoothed regret bound for a given bandwidth $h\geq 0$.

An important part of the contribution is setting up the definitions. Recall the definition of the smoothed loss $\lambda_h(\cdot)$ and optimal smooth loss $\texttt{Bench}(\Pi_h)$ from (1). The *version space* of ϵ-optimal policies (according to the smoothed loss) is

$$\Pi_{h,\epsilon} := \{\,\pi\in\Pi:\ \lambda_h(\pi)\leq \texttt{Bench}(\Pi_h)+\epsilon\,\}.$$

For a given context $x\in\mathcal{X}$, a policy subset $\Pi'\subset\Pi$ maps to an action set $\Pi'(x):=\{\,\pi(x):\ \pi\in\Pi'\,\}$. We are interested in $\Pi_{h,\epsilon}(x)$, the subset of actions chosen by the ϵ-optimal policies on context x, and specifically the expected packing number of this set:

$$M_h(\epsilon,\delta) := \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathcal{N}_\delta\left(\,\Pi_{h,\epsilon}(x)\,\right)\right], \tag{2}$$

where $\mathcal{N}_\delta(A)$ is the δ-packing number of subset $A\subset\mathcal{A}$ in the ambient metric space $(\mathcal{A},\rho)$.[5] The *smoothing coefficient* $\theta_h:\mathbb{R}\to\mathbb{R}$ measures how the packing numbers $M_h(12\epsilon,h)$ shrink with ϵ:

$$\theta_h(\epsilon_0) := \sup_{\epsilon\geq\epsilon_0} M_h(12\epsilon,h)/\epsilon. \tag{3}$$

For the unit interval, observe that $\theta_h(\epsilon_0)\leq(h\epsilon_0)^{-1}$ always, but in favorable cases we might expect $\theta_h(\epsilon_0)\leq\max\{1/h, 1/\epsilon_0\}$, as demonstrated by the following example. Note that the constant 12 is not fundamental, but is consistent with prior work on instance-dependent guarantees for continuous action spaces (Slivkins, 2014).

[5] A subset S of a set A is a δ-packing if any two points in S are at a distance of at least δ. The δ-packing number of a set A is the maximum cardinality of a δ-packing of A.

Example 3 (Small smoothing coefficient) *Consider a non-contextual problem, where the expected loss function is $\lambda(a) := \mathbb{E}[\ell(a) \mid x_0] = |\, a - a^\star \,|$ for some $a^\star \in [2h, 1-2h]$. Then $M_h(\epsilon, h) \leq O(\max\{\, 1, \epsilon/h \,\})$. Consequently, $\theta_h(\epsilon_0) \leq O(\max\{\, 1/h, 1/\epsilon_0 \,\})$. (See Section 10 for a derivation.)*

Our first result is in terms of this smoothing coefficient.

Theorem 1 *For any given bandwidth $h > 0$, in the stochastic setting,* `SmoothPolicyElimination` *(Algorithm 1) with parameter h achieves*

$$\texttt{Regret}(T, \Pi_h) \leq O\left(\inf_{\epsilon_0 > 0} \{\, T\epsilon_0 + \theta_h(\epsilon_0)\ \log(|\Pi| T)\ \log(1/\epsilon_0) \,\} \right).$$

Since $\theta_h(\epsilon_0) \leq (h\epsilon_0)^{-1}$, we obtain a worst case guarantee as a corollary.

Corollary 2 *Fix any bandwidth $h > 0$, in the stochastic setting,* `SmoothPolicyElimination` *with parameter h achieves*

$$\texttt{Regret}(T, \Pi_h) \leq \tilde{O}\left(\sqrt{T/h \log|\Pi|} \right).$$

Contrasting with the standard $\Theta(\sqrt{T|\mathcal{A}|\log|\Pi|})$ regret bound for finite action spaces, we see that the $1/h$ term can be viewed as the effective number of actions.

In fact, this worst case bound can also be achieved by a simple variation of `EXP4` (Auer et al., 2002), which can operate in the adversarial version of our problem and actually achieves $O(\sqrt{T/h \log|\Pi|})$ regret, eliminating the logarithmic dependence on T. The pseudocode for this algorithm is displayed in Algorithm 2.

Theorem 3 *In the adversarial setting,* `ContinuousEXP4` *with policy set $\Xi = \Pi_h$ and learning rate $\eta = \sqrt{\frac{2h\ \ln|\Xi|}{T}}$ achieves* $\texttt{Regret}(T, \Pi_h) \leq O\left(\sqrt{T/h \log|\Pi|} \right)$.

Both algorithms are not computationally efficient in general, as the per-round running time scales as $|\Pi|$. For the non-contextual case, one can take $|\Pi| = T/h$, see Section 6.2(c).

Remarks. It is not hard to show a $\Omega(\sqrt{T/h \log|\Pi|})$ lower bound on smoothed regret. Specifically, every K arm contextual bandit instance can be reduced to a continuous action instance with bandwidth $h = 1/(2K)$ by using piecewise constant loss functions and by mapping actions $a \in \{1, \ldots, K\}$ to $h \cdot (2a - 1)$. Thus, we may embed the lower bound construction for contextual bandits with finite action space into our setup to verify that Corollary 2 is optimal up to logarithmic factors (and Theorem 3 is optimal up to constants).

While not technically very difficult, the worst-case bound showcases the power and generality of the new definition. In particular, we obtain meaningful guarantees for discontinuous losses as in Example 2. As we will see in the next section, under global smoothness assumptions, we can also obtain a bound on the more-standard quantity $\texttt{Regret}(T, \Pi)$.

Turning to the instance-specific bound in Theorem 1, we obtain a more-refined dependence on the effective number of actions $1/h$, which can be thought of as a “gap-dependent” bound. In the most favorable setting, we have $\theta_h(\epsilon_0) = \max\{\, 1/h, 1/\epsilon_0 \,\}$ which yields $\texttt{Regret}(T, \Pi_h) \leq$

$\tilde{O}\left(\sqrt{T \log |\Pi|} + \frac{1}{h} \log |\Pi| \right)$, eliminating the dependence on h in the leading term (Recall that Example 3 has this favorable behavior). Further, via the correspondence with the finite action setting, we also obtain a new instance-dependent bound for standard stochastic contextual bandits, which improves on prior worst case results by adapting to the effective size of the action space (Dudik et al., 2011; Agarwal et al., 2014). This result for the finite-action setting follows from our more general theorem statement, given in Section 6.

We also note that, while smoothing induces a Lipschitz loss function, a naïve application of a Lipschitz bandits algorithm yields a suboptimal regret rate. For example, in the non-contextual version, the smoothed loss function is $\lambda_h : a \mapsto \mathbb{E}_{\mathcal{D}} \mathbb{E}_{a' \sim \texttt{Smooth}_h} [\, \ell(a') \,]$, is $1/h$-Lipschitz, so we may apply a Lipschitz bandits algorithm in a black box fashion.[6] However, this reduction gives a smoothed regret bound of $O(T^{2/3} h^{-1/3})$, which is suboptimal when compared with our $\tilde{O}(\sqrt{T/h})$ result. Our guarantees exploit additional information sharing between actions enabled by the smoothing operator, in particular the fact that when we choose a particular action, we learn about all smoothed actions in an interval of size h.

Finally, we remark that Algorithm 1 actually achieves a high probability regret bound, which we have simplified to the stated expected regret bound.

The algorithm. The algorithm is an adaptation of `PolicyElimination` from Dudik et al. (2011), with pseudocode displayed in Algorithm 1. It is epoch based, maintaining a version space of good policies, denoted $\Pi^{(m)}$ in the m^{th} epoch, and pruning it over time by eliminating the provably suboptimal policies. In the m^{th} epoch, the algorithm computes a distribution Q_m over $\Pi^{(m)}$ by solving a convex program (4). The objective function is related to the variance of the loss estimator we use, and so Q_m ensures high-quality loss estimates for all policies in $\Pi^{(m)}$. We use Q_m to select actions at each round in the epoch by sampling $\pi \sim Q_m$ and playing $\texttt{Smooth}_h(\pi(x))$ on context x. To compute $\Pi^{(m+1)}$ for the next epoch, we use importance weighting to form single-sample unbiased estimates for $\lambda_h(\pi)$ in (5), and we aggregate these via a median-of-means approach (see e.g., Hsu and Sabato (2016)). $\Pi^{(m+1)}$ is then defined as the set of policies with low empirical regret measured via the median-of-means estimator. Naïvely, the running time is $\text{poly}(T, |\Pi|)$.[7]

The key changes over `PolicyElimination` are as follows. First, we write (4) as an optimization problem rather than a feasibility problem, which allows for instance-dependent improvements in our loss estimates. Second, our importance weighting crucially exploits smoothing for low variance. Finally, we employ the median-of-means estimator to eliminate an unfavorable range dependence with importance weighting. The immediate consequence of this estimator is that we can eliminate the need for uniform exploration, which appears in prior literature on contextual bandits with finite action spaces (e.g. Dudik et al., 2011; Agarwal et al., 2014). Perhaps more interestingly, the median-of-means estimator is unnecessary for Corollary 2 and for prior results with finite action spaces, but it is crucial for obtaining our instance-dependent bound, since we need the error of our loss estimator to scale with the characteristic volume $V_m := \mathbb{E}_{x \sim \mathcal{D}}\, \nu \left(\bigcup_{\pi \in \Pi^{(m)}} \texttt{B}_h(\pi(x)) \right)$.

[6] Formally, when the Lipschitz bandits algorithm recommends action a'_t, we sample $a_t \sim \texttt{Smooth}_h(a'_t)$, observe $\ell_t(a_t)$ — which has expectation $\lambda_h(a'_t)$ — and pass this value back to the algorithm.

[7] For the non-contextual case, the algorithm simplifies and the running time becomes $\text{poly}(T)$, see also Section 6.2(c).

Algorithm 1 `SmoothPolicyElimination`

```
Parameters: Bandwidth $h > 0$, policy set $\Pi$, number of rounds $T$.
Initialize: $\Pi^{(1)} = \Pi$, Batches $\delta_T = 5\lceil \log(T|\Pi|\log_2(T)) \rceil$, Radii $r_m = 2^{-m}, m = 1, 2, \ldots$.
for each epoch $m = 1, 2, \ldots$ do
    // Before the epoch: compute distribution $Q_m$ over policy set $\Pi^{(m)}$.
    Set $V_m \leftarrow \mathbb{E}_{x\sim\mathcal{D}}\, \nu\left(\bigcup_{\pi\in\Pi^{(m)}} \mathtt{B}_h(\pi(x))\right)$        // characteristic volume of $\Pi^{(m)}$
    Set batch size $\tilde{n}_m = \frac{320 V_m}{r_m^2 h}$, epoch length $n_m = \tilde{n}_m \delta_T$.
    Find distribution $Q_m$ over policy set $\Pi^{(m)}$ which minimizes
        $$\max_{\text{policies } \pi\in\Pi^{(m)}} \ \mathop{\mathbb{E}}_{\text{context } x\sim\mathcal{D}_X} \ \mathop{\mathbb{E}}_{\text{action } a\sim\mathtt{Smooth}_h(\pi(x))} \left[\frac{1}{q_m(a \mid x)}\right], \tag{4}$$
        where density $q_m(a \mid x) := \mathop{\mathbb{E}}_{\pi\sim Q_m} \mathtt{Smooth}_{\pi,h}(a|x)$.
    for each round $t$ in epoch $m$ do
        Observe context $x_t$, sample action $a_t$ from density $q_m(\cdot \mid x_t)$, observe loss $\ell_t(a_t)$.
    end for
    // After the epoch: update the policy set.
    for each batch $i = 1, 2, \ldots, \delta_T$ do
        Define $S_{i,m}$ as the indices of the $(i-1)\tilde{n}_m + 1, \ldots, i\tilde{n}_m^{\text{th}}$ examples in epoch $m$.
        Estimate $\lambda_h(\pi)$ with $\hat{L}_m^i(\pi) = \frac{1}{\tilde{n}_m}\sum_{t\in S_{i,m}} \hat{\ell}_{t,h}(\pi)$ for each policy $\pi \in \Pi^{(m)}$ where
            $$\hat{\ell}_{t,h}(\pi) := \frac{\mathtt{Smooth}_{\pi,h}(a_t|x_t)\ \ell_t(a_t)}{q_m(a_t|x_t)}. \tag{5}$$
    end for
    Estimate the loss $\hat{L}_m(\pi) = \text{median}\left(\hat{L}_m^1(\pi), \hat{L}_m^2(\pi), \ldots, \hat{L}_m^{\delta_T}(\pi)\right)$.
    $\Pi^{(m+1)} = \left\{\pi \in \Pi^{(m)} : \hat{L}_m(\pi) \le \min_{\pi'\in\Pi^{(m)}} \hat{L}_m(\pi') + 3\, r_m\right\}$.
end for
```

As we have described the algorithm, it requires knowledge of the marginal distribution over $\mathcal{X}$, which appears in the computation of V_m and in the optimization problem. Both of these can be replaced with empirical counterparts, and since the random variables are non-negative, via Bernstein's inequality, the approximation only affects the regret bound in the constant factors. This argument has been used in several prior contextual bandit results (Dudik et al., 2011; Agarwal et al., 2014; Krishnamurthy et al., 2016), and so we omit the details here.

For the proof, we first use convex duality to upper bound the value of (4) in terms of the characteristic volume V_m, refining Dudik et al. (2011). As the objective divided by h bounds the variance of the importance weighted estimate in (5), we may use Chebyshev and Chernoff bounds to control the error of the median-of-means estimator in terms of V_m, h, and n_m. Our setting of n_m then implies that $\Pi^{(m+1)} \subset \Pi_{h,12r_{m+1}}$. Two crucial facts follow: (1) the instantaneous regret in epoch $m+1$ is related to r_{m+1} and (2) V_{m+1}, which determines the length of the epoch, is related to the packing number $M_h(12r_{m+1}, h)$. Roughly speaking,

Algorithm 2 `ContinuousEXP4`: `EXP4` with continuous sampling

```
Parameters: Collection of randomized policies $\Xi$, learning rate $\eta > 0$.
// $\xi(\cdot \mid x_t)$ is the probability density for policy $\xi$ given context $x_t$.
Initialization: weights $W_1(\xi) \leftarrow 1$ for all policies $\xi \in \Xi$.
for $t = 1, \ldots, T$ do
    Sample policy $\xi_t \propto W_t$, sample action $a_t$ from $\xi_t(\cdot \mid x_t)$.
    // $p_t(\cdot \mid x_t)$ is the probability density for action $a_t$ given context $x_t$.
    Observe loss $\ell_t(a_t)$ and define
        $$\hat{\ell}_t(\xi) := \frac{\xi(a_t \mid x_t)}{p_t(a_t \mid x_t)} \cdot \ell_t(a_t).$$
    Update weights: $W_{t+1}(\xi) \leftarrow W_t(\xi) \cdot \exp(-\eta \hat{\ell}_t(\xi))$.
end for
```

this shows that the regret in epoch m is $n_m r_m \lesssim M_h(12r_m, h)/r_m$, which we can easily relate to the smoothing coefficient.

4.2. One algorithm for all h

`SmoothPolicyElimination` guarantees a refined regret bound against $\texttt{Bench}(\Pi_h)$ for a given $h > 0$. Yet choosing the bandwidth in practice seems challenging: since $\texttt{Bench}(\Pi_h)$ is unknown and not monotone in general, there is no a priori way to choose h to minimize the benchmark plus the regret. As such, we seek algorithms that can achieve a smoothed regret bound simultaneously for all bandwidths h, a guarantee we call *uniformly-smoothed.* This is achieved by our next result.

Theorem 4 *Consider the adversarial setting. For each parameter $\beta \in [0, 1]$, there exists an algorithm that guarantees*

$$\forall h \in (0, 1] : \texttt{Regret}(T, \Pi_h) \leq \tilde{O}\left(T^{\frac{1}{1+\beta}} h^{-\beta}\right) \cdot (\log |\Pi|)^{\frac{\beta}{1+\beta}}.$$

For the non-contextual setting, it achieves a uniformly-smoothed regret of $\tilde{O}\left(T^{\frac{1}{1+\beta}} h^{-\beta}\right)$. Moreover, for the non-contextual stochastic setting, there exist positive constants c and T_0, such that for any algorithm and any $T \geq T_0$, there exists $h \in (0, 1]$ and a problem instance, such that on this instance,

$$\texttt{Regret}(T, \Pi_h) \geq c \cdot T^{\frac{1}{1+\beta}} h^{-\beta}.$$

Remarks. The theorem provides a family of upper and lower bounds, one for each $\beta \in [0, 1]$. As two examples, taking $\beta = 1$ we obtain regret rate $\tilde{O}(\sqrt{T}/h)$ as listed in the third row of Table 1, while $\beta = 1/2$ yields $\tilde{O}(T^{2/3}/\sqrt{h})$. These bounds are incomparable in general and so the result establishes a Pareto frontier of exponent pairs. In the non-contextual setting, all pairs are optimal, and, in particular, the $\sqrt{T/h}$ rate from Corollary 2 is not achievable uniformly. More generally, the optimal uniformly-smoothed regret bounds are very different from those for a fixed bandwidth.

Note that while β is a parameter to the algorithm, it simply governs where on the Pareto frontier the algorithm lies, and is not based on any property of the problem.

The algorithm. The algorithm, Corral+EXP4, is an instantiation of Corral (Agarwal et al., 2017b), which can be used to run many sub-algorithms in parallel. Corral maintains a master distribution over sub-algorithms, and in each round it samples a sub-algorithm and chooses the action the sub-algorithm recommends. Corral sends an importance weighted loss (weighted by the master distribution) to all the sub-algorithms and it updates the master distribution using online mirror descent with the log-barrier mirror map.

For the sub-algorithms we use our variant of EXP4. Each sub-algorithm instance operates with a different bandwidth scale, and if run in isolation achieves the optimal non-adaptive smoothed regret for that bandwidth. Aggregating these sub-algorithms with Corral yields the uniformly-smoothed guarantee. Note that here and elsewhere, Corral results in a worse overall regret than the best individual sub-algorithm, but in our setting it nevertheless achieves all Pareto-optimal uniformly-smoothed guarantees. We describe Corral+EXP4 formally in Section 8.2.

The proof for the upper bound involves a more refined analysis for EXP4 than required for Theorem 3. First, we discretize bandwidths to multiples of $1/T^2$ and show that, for any $i \in \mathbb{N}$, a single instance of EXP4 using discretized bandwidths can compete with all $h \in [2^{-i}, 2^{-i+1}]$ simultaneously, without using Corral. Second, we show that EXP4 is stable in the sense that, in randomized environments, the regret scales linearly with the standard deviation of the losses and that this standard deviation need not be known a priori.[8] Stability is crucial for aggregating with Corral as the master's importance weighting induces high-variance randomized losses for each sub-algorithm. We finish the proof by applying the guarantee for Corral (Agarwal et al., 2017b) with $\log(T)$ instances of EXP4 as sub-algorithms, one for each bandwidth scale $[2^{-i}, 2^{-i+1}]$. For each $\beta \in [0, 1]$, we use a weakening of the EXP4 regret guarantee, essentially that $\min\left\{\sqrt{T/h}, T\right\} \le T^{\frac{1}{1+\beta}} h^{-\frac{\beta}{1+\beta}}$ for all $\beta \in [0, 1]$.

The lower bound is inspired by a construction of Locatelli and Carpentier (2018). We show that if an algorithm, Alg, has small regret against $\mathtt{Bench}(\Pi_{1/4})$, then it must suffer large regret against $\mathtt{Bench}(\Pi_h)$ for $h \ll 1/4$. The intuition is that the $1/4$-smoothed regret bound prevents Alg from sufficiently exploring. Specifically, we construct one instance where small losses occur in a subinterval $I_0 \subset [0, 1]$ of length $1/4$ and another that is identical on I_0 but where even smaller losses occur in a subinterval I_1 of width $h \ll 1/4$. Since Alg has low $1/4$-smoothed regret it cannot afford to explore to find I_1. In comparison with Locatelli and Carpentier (2018), the details of the construction are somewhat different, since they focus on adaptivity to unknown smoothness exponent, while we are adapting to bandwidth h (and later to unknown Lipschitz constant).

5. Lipschitz regret guarantees

Our results and techniques for smoothed regret project onto the well-studied Lipschitz contextual bandits problem: each of the three results in Section 4 has a “twin” for the

[8] This property was shown by Agarwal et al. (2017b), but our variant of EXP4 is necessarily slightly different. Nevertheless, the proof is quite similar.

Lipschitz version. We posit a Lipschitz condition on the expected loss $\lambda(\cdot \mid x) := \mathbb{E}[\ell(\cdot) \mid x]$:

$$\forall x \in \mathcal{X},\ a, a' \in \mathcal{A} : \quad \left| \lambda(a \mid x) - \lambda(a' \mid x) \right| \leq L \cdot \rho(a, a'), \quad L \geq 1. \tag{6}$$

We assume that $L \geq 1$ to avoid the pathological situation where Lipschitzness restricts the effective loss range. If the Lipschitz constant is less than 1, we set $L = 1$ in our results.

A version of the standard *uniform discretization* approach applies, even for the adversarial setting. Here, we uniformly discretize the action space and the policies (if needed), and we run `EXP4`. Standard arguments yield the following regret bound:

$$\texttt{Regret}(T, \Pi) \leq \tilde{O}\left(T^{2/3}\ \left(L \log |\Pi| \right)^{1/3} \right). \tag{7}$$

This result appears in prior work on the non-contextual case and is known to be optimal (Kleinberg, 2004; Kleinberg et al., 2019; Bubeck et al., 2011a), although the generalization to an arbitrary policy set Π is new. Interestingly, the worst-case regret bounds in Slivkins (2014); Cesa-Bianchi et al. (2017) — on Lipschitz contextual bandits with a metric on contexts or context-arm pairs and with specific policy sets Π, respectively — can be obtained from this uniform discretization approach. Equation (7) is the point of departure for several results presented below.

The key observation enabling results for the Lipschitz version is as follows:

Lemma 5 *If $f : \mathcal{A} \to [0, 1]$ is L-Lipschitz, then $\left| \mathbb{E}_{a' \sim \texttt{Smooth}_h(a)} f(a') - f(a) \right| \leq Lh$.*

In particular if $\lambda(\cdot \mid x)$ is L-Lipschitz, we have $\texttt{Bench}(\Pi_h) \leq \texttt{Bench}(\Pi) + Lh$, which allows us to easily obtain results for the Lipschitz version by way of smoothed regret.

5.1. Instance-dependent and worst-case guarantees

In correspondence with Theorem 1, our first result here is an instance-dependent regret bound. We recover the optimal worst-case regret bound for the Lipschitz setting, but we obtain an improvement when actions taken by near-optimal policies tend to lie in a relative small region of the action space. Specializing, we recover several state-of-the-art instance-dependent regret bounds from prior work. Our algorithm is a minor modification of `SmoothPolicyElimination` (Algorithm 1), which we denote `SmoothPolicyElimination.L` and spell out later in this section.

We reuse the packing numbers $M_h(\epsilon, \delta)$ defined in (2), but the instance-dependent complexity is slightly different. Instead of the smoothing coefficient $\theta_h(\epsilon_0)$, we use the *policy zooming coefficient*:

$$\psi_L(\epsilon_0) := \sup_{\epsilon \geq \epsilon_0} M_0(12L\epsilon, \epsilon)/\epsilon. \tag{8}$$

The main differences over the smoothing coefficient are that version space of good policies is based on the unsmoothed loss $\lambda_0(\pi)$, and we are using the ϵ- rather than h-packing number for a fixed bandwidth h. For intuition, we always have $\psi_L(\epsilon_0) \leq O(\epsilon_0^{-2})$ but a favorable instance might have $\psi_L(\epsilon_0) \leq O(\epsilon_0^{-1})$ which yields improved rates.

Theorem 6 *In the stochastic setting, Algorithm* `SmoothPolicyElimination.L` *with parameter L achieves regret bound*

$$\texttt{Regret}(T, \Pi) \leq O\left(\inf_{\epsilon_0 > 0}\left\{ TL\epsilon_0 + \frac{\psi_L(\epsilon_0)}{L} \cdot \log(|\Pi| T) \log(1/\epsilon_0) \right\}\right). \tag{9}$$

Since $\psi_L(\epsilon_0) \leq O(\epsilon_0^{-2})$, we obtain the following worst-case bound, which is known to be optimal up to $\log(T)$ factors.

Corollary 7 *In the stochastic setting, algorithm* `SmoothPolicyElimination.L` *with parameter L achieves the regret bound in* (7)*.*

The worst-case result is in correspondence with Corollary 2. It recovers the worst-case regret bound from prior work focusing on the non-contextual version (Kleinberg, 2004; Bubeck et al., 2011b). This regret bound can also be achieved by `ContinuousEXP4` as a simple corollary of Theorem 3 (see Corollary 19 in Section 6).

The result can also be applied to a nonparametric policy set in the setting of Cesa-Bianchi et al. (2017). Here we assume $\mathcal{X}$ is a p-dimensional metric space and the policy set is all 1-Lipschitz mappings from $\mathcal{X} \to \mathcal{A}$. By a suitable discretization, Corollary 7 yields $\tilde{O}\left(T^{\frac{p+2}{p+3}}\right)$ regret, which matches their result (since the interval is a 1-dimensional action space).

The advantage of Theorem 6 is its instance-dependence. Since the packing number $M_0(\cdot, \cdot)$ is always at least 1, the most favorable instances have $\psi_L(\epsilon_0) = O(\epsilon_0^{-1})$. In this case, Theorem 6 gives the much faster $\tilde{O}(\sqrt{T \log |\Pi|})$ regret rate. The next example demonstrates such favorable behavior.

Example 4 *Let $\mathbf{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$. Consider an instance where $\mathcal{X} := \mathbf{S}^{d-1}$, $\mathcal{A} := [-1, 1]$ and where the policy class Π is a finite subset of* linear *policies $\left\{ \pi_w : w \in \mathbf{S}^{d-1} \right\}$ where $\pi_w : x \mapsto \langle w, x \rangle$. The marginal distribution over contexts is uniform over $\mathbf{S}^{d-1}$ and the expected losses satisfy*

$$\forall x \in \mathcal{X} : \mathbb{E}\left[\, \ell(a) \mid x \,\right] = f(a - \pi_{w^\star}(x)), \tag{10}$$

where $\pi_{w^\star} \in \Pi$ is some fixed policy, f is L-Lipschitz and satisfies $f(z) - f(0) \geq L_0 \,|\, z \,|$ for all z in $\mathbb{R}$. By construction, $\mathbb{E}[\ell(a) \mid x]$ is L-Lipschitz in a, for all x. This instance has $M_0(L\epsilon, \epsilon) = O(L/L_0 \cdot \sqrt{d})$, and $\psi_L(\epsilon) = O(\frac{L}{L_0 \epsilon} \cdot \sqrt{d})$. (See Section 10 for a derivation.)

Instance-dependent bounds from prior work are often stated in terms of a packing number growth rate, called the *zooming dimension*. Our bound can also be stated in this way, so as to facilitate comparisons. With *zooming constant* $\gamma > 0$ the zooming dimension is defined as

$$z := \inf\left\{ d > 0 : M_0(12L\epsilon, \epsilon) \leq \gamma \cdot \epsilon^{-d}, \ \forall \epsilon \in (0, 1) \right\}. \tag{11}$$

It is easy to see that $\psi_L(\epsilon_0) \leq \gamma \cdot \epsilon_0^{-z-1}$, and so Theorem 6 may be further simplified to

$$\texttt{Regret}(T, \Pi) \leq O\left(L^{\frac{z}{2+z}} T^{\frac{1+z}{2+z}} \right) \cdot \left(\gamma \log(T\,|\Pi|) \right)^{\frac{1}{2+z}}. \tag{12}$$

This result agrees with prior zooming results in the non-contextual setting (Kleinberg et al., 2019; Bubeck et al., 2011a). In the contextual setting, our result is in general incomparable with the "contextual zooming algorithm" of Slivkins (2014), which scales with a different quantity called the *contextual zooming dimension.* Formally the contextual zooming dimension measures the growth of the ϵ-packing numbers of the set $\{(x, a) : \mathbb{E}[\ell(a) \mid x] - \min_{a' \in \mathcal{A}} \mathbb{E}[\ell(a') \mid x] \leq \epsilon\}$ as a function of ϵ. This definition, and our zooming dimension are conceptually similar, as both measure the size of certain near-optimal sets, but they are generally incomparable. Informally, the definition in Slivkins (2014) is adapted to a Lipschitz structure on the context space, which does not naturally accommodate arbitrary policy sets Π as we do, and our zooming dimension involves the "expected context" rather than the "worst context." In more detail:

1. Slivkins (2014) needs to assume a metric structure on $\mathcal{X} \times \mathcal{A}$, whereas we only assume a metric structure on $\mathcal{A}$. In addition, Slivkins (2014)'s contextual zooming dimension is at worst the covering dimension of $\mathcal{X} \times \mathcal{A}$, whereas our notion of zooming dimension is at worst the covering dimension of $\mathcal{A}$. On the other hand, our bound scales with $\log |\Pi|$ while his does not.

2. Aside from the metric structure, Slivkins (2014)'s contextual zooming dimension is only dependent on the conditional distribution of loss given context $\mathcal{D}(\ell|x)$. In contrast, our notion is dependent on the policy class Π, along with $\mathcal{D}$, the joint distribution of (x, ℓ), which admits policy class and distribution specific upper bounds.

3. Finally, Slivkins (2014) considers a setting where contexts are adversarially chosen, and so his contextual zooming dimension considers pessimistic context arrivals. Our definition involves an expectation over contexts, which may be more favorable.

The algorithm. The algorithm is almost identical to `SmoothPolicyElimination`. The main difference is that instead of a fixed bandwidth h across all epochs, we use $h_m = 2^{-m}$ in the m^{th} epoch. We also set the radius parameter $r_m = L2^{-m}$ which is slightly different from before. We call this algorithm `SmoothPolicyElimination.L`, to highlight the differences.

At a technical level, the main difference with the Lipschitz setting is that we must carefully balance bias and variance in loss estimates. This is not an issue for smoothed regret since we have unbiased estimators for $\lambda_h(\pi)$, but not for $\lambda_0(\pi)$. We do this by decreasing the bandwidth geometrically over epochs, but the rest of the algorithm, and much of the analysis are unchanged.

5.2. Optimal Lipschitz-Adaptivity

We now present the corresponding result to Theorem 4. We consider *Lipschitz-adaptive* algorithms: those that *do not know* any information about the problem, apart from T and Π, and yet achieve regret bounds in terms of T, L, and $|\Pi|$ *only.* In particular, the algorithm does not know L.

Theorem 8 *Consider the adversarial setting. For each $\beta \in [0, 1]$,* `Corral+EXP4` *(with parameter β) is Lipschitz-adaptive with*

$$\texttt{Regret}(T, \Pi) \leq \tilde{O}\left(T^{1-a}\, L^b\, \left(\log |\Pi|\right)^a\right), \text{ where } \quad a = \frac{\beta}{1 + 2\beta} \text{ and } b = \frac{\beta}{1 + \beta}.$$

For the non-contextual version it achieves a regret $\tilde{O}\left(T^{1-a}\, L^{b}\right)$ without knowing the Lipschitz constant L. Moreover, for the non-contextual stochastic version, there exist positive constants c and T_0, such that for any algorithm and any $T \geq T_0$, there exists $L \geq 1$ and a problem instance with L-Lipschitz losses, such that on that instance

$$\texttt{Regret}(T, \Pi) \geq c \cdot T^{1-a} L^{b}.$$

Remarks. As in Theorem 4, we obtain a family of upper and lower bounds, one for each $\beta \in [0, 1]$, which make up a Pareto frontier. With $\beta = 1$ an optimal Lipschitz-adaptive rate is $T^{2/3}\sqrt{L}$ which is worse than the $T^{2/3}L^{1/3}$ non-adaptive rate from Corollary 7. Note that it is easy to obtain the worse adaptive rate of $\tilde{O}\left(LT^{2/3}\right)$ simply by guessing that the Lipschitz constant is 1 in our variant of `EXP4`.

Several prior works develop adaptive algorithms that either require knowledge of unknown problem parameters, or yield regret bounds that, in addition to T and L, scale with such parameters (Slivkins, 2011; Bubeck et al., 2011b; Bull, 2015; Locatelli and Carpentier, 2018). These algorithms are not Lipschitz adaptive, contrasting with our algorithm that requires no additional knowledge or assumptions. However, this dependence on other parameters allows these prior results to side-step our lower bound and achieve faster rates.

Note that Lipschitz-adaptivity is qualitatively quite different from the uniformly-smoothed adaptivity studied in Theorem 4. With Lipschitz-adaptivity there is a single fixed benchmark policy class and we simply seek a guarantee against that class, albeit in an environment with unknown smoothness parameter. However, for Theorem 4 we are effectively competing with infinitely many policy sets simultaneously (Π_h for each $h \in (0, 1]$) and we seek a regret bound against all of them. Somewhat surprisingly, both settings demonstrate a similar price-of-adaptivity and the optimally adaptive algorithms are nearly identical.

The algorithm. The algorithm, `Corral`+`EXP4`, is again `Corral` with our variant of `EXP4` as the sub-algorithms. The only difference is in how we set the learning rate for the master algorithm.

6. Our results in a general setup

All results discussed so far are special cases of a more general set of results that we now present. While all of the key ideas appear in the special case of the unit interval, the following results demonstrate the generality of our approach. As we have already made many of the essential remarks, the discussion here is somewhat terse.

We generalize in two directions. First, all results extend to higher-dimensional action spaces. Formally, $\mathcal{A}$ can be an arbitrary convex subset of the d-dimensional unit cube $[0, 1]^d$, equipped with p-norm $\rho(a, a') := \|a - a'\|_p$, for any $p \geq 1$. As before, $\texttt{Smooth}_h(a)$ is a uniform distribution over the closed ball of radius h. The instance-dependent regret bounds carry over as is, and zooming dimension now takes values in $[0, d]$ depending on the problem instance. Regret bounds in the worst-case corollaries are modified so as to accommodate the dependence on d. In Corollary 2, the dependence on h is replaced with h^d, and there is a matching lower bound. In Corollary 7, the dependence on T becomes $\tilde{O}(T^{(d+1)/(d+2)})$, which is known to be optimal. The smoothness-adaptive regret bounds are modified similarly.

Second, we essentially allow *arbitrary* action spaces and smoothing operators. Formally, the action space $\mathcal{A}$ is endowed with a base metric ρ and a base measure ν. The smoothing distribution $\texttt{Smooth}(a)$ can be any distribution with a well-defined density with respect to ν, and the effective number of actions is the largest possible density value. We define bandwidth relative to the base metric. In particular, we can handle the unit cube $[0,1]^d$, endowed with a uniform measure and the p-norm, $p \geq 1$ as a base metric.

The proofs for all instance-dependent regret bounds are deferred to Section 7. The proofs for all "smoothness-adaptive" regret bounds can be found in Section 8.

6.1. General setup

For completeness, let us recap the basic setup of contextual bandits. There are two sets $\mathcal{X}, \mathcal{A}$, where $\mathcal{X}$ is an abstract *context space*, and $\mathcal{A}$ is an abstract *action space*. The following protocol continues over T rounds: at each round t, (i) nature chooses context $x_t \in \mathcal{X}$ and loss function $\ell_t \in (\mathcal{A} \to [0,1])$ and presents x_t to the learner, (ii) learner chooses action $a_t \in \mathcal{A}$, (iii) learner suffers loss $\ell_t(a_t)$, which is also observed. Performance of the learner is measured relative to a class of policies $\Pi : \mathcal{X} \to \mathcal{A}$ via the notion of regret

$$\texttt{Regret}(T, \Pi) := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\pi(x_t))\right].$$

We consider both adversarial and stochastic settings. In the *adversarial setting* the contexts and losses are chosen by an adaptive adversary, meaning that (x_t, ℓ_t) may be a randomized function of the entire history of interaction. In the *stochastic setting*, we assume $(x_t, \ell_t) \sim \mathcal{D}$ iid at each round t, for some unknown distribution $\mathcal{D}$, although we assume $\mathcal{D}_X$, the marginal distribution over $\mathcal{X}$, is known.

Base structure. Action space $\mathcal{A}$ is endowed with *base structure* $(\mathcal{A}, \rho, \nu)$, where ρ is a metric called the *base metric*, and ν is a probability measure called the *base measure*. The two are consistent, in the sense that ν is well-defined and strictly positive on the closed balls in $(\mathcal{A}, \rho)$ of strictly positive radius. This structure may have no bearing on the loss functions; it serves only to define and/or instantiate smoothed regret. Essentially, we smoothen relative to the base measure, and define bandwidth relative to the base metric.

The closed balls are denoted $\texttt{B}(a, r) := \{b \in \mathcal{A} : \rho(a, b) \leq r\}$, where $a \in \mathcal{A}$ is the center and $r \geq 0$ is the radius. For normalization, we assume that the metric space has diameter 1.

Smoothing kernel. We generalize the $\texttt{Smooth}$ operator to a *smoothing kernel*: a mapping $K : \mathcal{A} \to \Delta(\mathcal{A})$, the set of distributions over actions. For policy π, we use $K\pi : x \mapsto K(\pi(x))$ to denote the usual function composition. With $\Pi_K := \{\, K\pi : \pi \in \Pi \,\}$ as the smoothed policy class, smoothed regret is simply given by $\texttt{Regret}(T, \Pi_K)$.

We posit that distributions $K(a)$, $a \in \mathcal{A}$ are absolutely continuous with respect to the base measure ν, and represent them via their density functions $f_{K(a)}$. Formally, $f_{K(a)}$ is the Radon-Nikodym derivative of $K(a)$ relative to ν. As a convention, denote $(Ka)(a') := f_{K(a)}(a')$, $a' \in \mathcal{A}$. In words, it is the density of distribution $K(a)$ with respect to ν, evaluated at a'.

We derive (worst-case) bounds on $\texttt{Regret}(T, \Pi_K)$ for an arbitrary smoothing kernel K, without any assumptions on the loss functions. The regret bounds are in terms of the largest possible density assigned by K,

$$\kappa := \sup_{a, a' \in A} (Ka)(a'). \tag{13}$$

This quantity, called *kernel complexity*, serves as the effective number of actions.

We trade off κ against a suitably generalized notion of *bandwidth*:

$$\sup_{a,a' \in A:\ (Ka)(a')>0} \rho(a,a'). \tag{14}$$

In words, it is the largest distance that any action can be perturbed by.

The canonical example is the *rectangular kernel* K_h, where $h \in (0,1]$ is the *bandwidth*:

$$(K_h\, a)(a') = \frac{\mathbf{1}\{\, \rho(a,a') \leq h \,\}}{\nu(\mathtt{B}(a,h))}, \quad \forall a, a' \in \mathcal{A}. \tag{15}$$

In words, $K_h(a)$ puts uniform density on $\mathtt{B}(a,h)$, and zero density elsewhere.[9]

Discussion. In practice, we may have some freedom in choosing the base structure. The action space $\mathcal{A}$ may naturally admit a set system: e.g., the open/closed intervals when $\mathcal{A}$ is a unit cube, or the subtrees when $\mathcal{A}$ is the leaf set of a tree. Then, we may have some leeway in defining the base metric, e.g., it could be any p-norm, $p \geq 1$ when $\mathcal{A}$ is the unit cube, or any "exponential tree metric" $\rho(x,y) = \alpha^{\mathtt{depth}(\mathtt{LCA}(x,y))}$, $\alpha \in (0,1)$ when $\mathcal{A}$ is a leaf set.[10] Then, we may be able to tailor the base measure to the chosen base metric, so as to improve the kernel complexity (more on this below).

One can choose smoothing kernels other than the rectangular kernel. One fairly general formulation is the *f-symmetric kernel* K_f defined by

$$(K_f\, a)(a') \sim f(\rho(a,a')) \quad \forall a, a' \in \mathcal{A}, \tag{16}$$

for some function $f : [0,1] \to [0,\infty)$. In particular, the *triangular kernel* is the special case when $f(x) = \max(0, 1 - x/h)$, where $h > 0$ is the bandwidth. For more refined kernel complexity vs. bandwidth tradeoff, one could consider an averaged version of bandwidth:

$$\sup_{a \in A} \int \rho(a,\cdot)\, \mathtt{d}K(a). \tag{17}$$

That said, in our analysis the smoothing kernel is either arbitrary or rectangular.

Example: covering dimension. To instantiate kernel complexity, consider the notion of *covering dimension*. The formal definition is as follows:

Definition 9 *For a metric space $(\mathcal{A},\rho)$, the covering dimension with multiplier γ is the smallest number $d \geq 0$ such that for each $r \in (0,1]$, the metric space can be covered with $\gamma \cdot r^{-d}$ balls of radius r.*

This notion has been used to summarize the complexity of a metric space for Lipschitz bandits (Kleinberg, 2004). We can also use it to bound kernel complexity.

[9] If the action space $\mathcal{A}$ is a unit interval, the plot of the density function for $K_h(a)$ is a rectangle, hence the name *rectangular kernel*.

[10] $\mathtt{LCA}(x,y)$ is the least common ancestor of leaves x and y.

Claim 10 *Fix the base metric space $(\mathcal{A}, \rho)$ of covering dimension d with multiplier γ. Fix bandwidth $h > 0$. Then there exists a probability measure ν such that*

$$\nu(\mathtt{B}(a,h)) \geq (h/2)^d/\gamma \quad \textit{for each center } a \in \mathcal{A}. \tag{18}$$

With ν as the base measure, the rectangular kernel K_h has complexity $\kappa \leq \gamma \cdot (h/2)^{-d}$.

Proof By definition of the covering dimension, there is a collection $\mathcal{C}$ of at most $N = \gamma \cdot (h/2)^{-d}$ balls of radius $h/2$ whose union covers $\mathcal{A}$. Define probability measure ν as follows: pick a ball $B \in \mathcal{C}$ uniformly at random, then pick a point inside B according to an arbitrary fixed probability measure ν_B. Any ball $\mathtt{B}(a,h)$, $a \in \mathcal{A}$ contains some ball $B \in \mathcal{C}$, namely a ball in $\mathcal{C}$ that covers a. Hence, $\mathtt{B}(a,h) \geq \nu(B) \geq 1/N$. ■

Example: local uniformity. It may be desirable to ensure that the base structure is uniform, in the sense that balls of similar radius have a similar measure. A "local" version of this property can be stated as follows: for some number $d \geq 0$ called the *doubling dimension*,

$$\nu(\mathtt{B}(a,2r)) \leq 2^d \cdot \nu(\mathtt{B}(a,r)) \qquad \forall a \in \mathcal{A},\ r > 0. \tag{19}$$

Then the rectangular kernel K_h, $h > 0$ has complexity $\kappa \leq (h/2)^{-d}$, like in Claim 10.

By way of background, doubling dimension is a combinatorial notion of low-dimensionality, widely used in theoretical computer science.[11] It is a stronger notion than the covering dimension: it upper-bounds the covering dimension (with multiplier $\gamma = 2^d$). A canonical example is that any subset of $([0,1]^d, \ell_p)$, $d \in \mathbb{N}$, $p \geq 1$ has doubling dimension $d + O(1)$. However, there are examples that are provably very different (Gupta et al., 2003). When the metric space $(\mathcal{A}, \rho)$ is complete, a probability measure ν satisfying (19) exists if and only if the metric space satisfies a more basic property: any ball of radius r can be covered by a collection of 2^d balls of radius $r/2$ (Volberg and Konyagin, 1987; Wu, 1998; Luukkainen and Saksman, 1998). The latter property is typically used to define the doubling dimension. More background on doubling dimension can be found in (Slivkins, 2006, Chapter 2).

Global uniformity of the base structure. For our instance-dependent results in the stochastic setting, we require a "global" generalization of (19) which states that any two balls of a similar radius have a similar size. Formally:

Assumption 1 (instance-dependent results only)

$$\sup_{a,a' \in \mathcal{A},\ h \in (0,1/2]} \frac{\nu(\mathtt{B}(a,2h))}{\nu(\mathtt{B}(a',h))} \leq \alpha < \infty.$$

The effect of this assumption is that the α is a multiplier in the regret bounds. While α gives a direct bound on kernel complexity of the rectangular kernel K_h as $\kappa \leq (h/2)^{-\log \alpha}$, it

[11] Doubling dimension have been studied in many different contexts such as metric embeddings, traveling salesman and compact data structures, e.g., Gupta et al. (2003); Krauthgamer and Lee (2004); Krauthgamer et al. (2005); Talwar (2004); Kleinberg et al. (2009); Chan et al. (2009); Slivkins (2007); Mendel and Har-Peled (2005); Wong et al. (2005).

can be productive to bound κ using the covering dimension. Indeed, the latter is a much weaker property, in the sense that it is smaller than $\log \alpha$, and can be *much* smaller.

The canonical example is a finite subset of $[0,1]^d$ of near-uniform density, defined as follows. For a fixed scale $\epsilon > 0$, partition $[0,1]^d$ into axis-parallel hypercubes with side ϵ, called *ϵ-cells*. A subset $\mathcal{A} \subset [0,1]^d$ is *uniform-density* at scale ϵ if each ϵ-cell contain exactly one point in $\mathcal{A}$. Then Assumption 1 holds with $\alpha = O(1)^d$, with ℓ_∞ as the base metric and the uniform measure over $\mathcal{A}$ as the base measure.[12] Similar assumptions have been used in theoretical computer science literature on networks (Kleinberg, 2000; Kempe et al., 2005; Kempe and Kleinberg, 2002; Sarkar et al., 2010; Abraham et al., 2015).

6.2. Special cases

(a) *Unit interval.* Suppose action space $\mathcal{A} = [0,1]$ is endowed with base metric $\rho(a, a') := |\, a - a' \,|$, and the base measure is uniform over $\mathcal{A}$. Then the rectangular kernel (15) is precisely the $\texttt{Smooth}_h$ operator from Section 3. This example satisfies properties (18) and (19) (with $d = 1$) and Assumption 1 (with $\alpha = 4$, because of the edge effects). So, all results we presented in Section 4 and Section 5 follow from the general development.

(b) *Discretized unit interval.* Discretize the $[0,1]$ interval into M actions: $\mathcal{A} := \{i/M : i \in [M]\}$, with base metric/measure defined as above. The rectangular kernel K_h takes local averages across actions. Kernel complexity is $\kappa = {}^1\!/_h$, and we obtain bounds on smooth regret that are independent of the number of actions M.

(c) *Non-contextual setting.* The non-contextual setting can be embedded in ours by positing a single context $\mathcal{X} := \{x_0\}$ and policy set $\Pi : \{x_0 \mapsto a : a \in \mathcal{A}\}$. (When we state results for the non-contextual version, Π is *always* assumed to be this class.) Since our upper bounds typically scale with $\log |\Pi|$, they do not immediately yield meaningful guarantees when $|\mathcal{A}| = \infty$, but we can obtain meaningful results here via discretization.

For example, consider the basic setup in Section 3. Since the smoothed loss function $\lambda_h(\cdot)$ is $({}^1\!/_h)$-Lipschitz, we can discretize the action space uniformly with step $\sqrt{{}^h\!/_T}$, to ensure that $|\Pi| = \sqrt{{}^T\!/_h}$ and the discretization error — increase in regret due to the discretization — is at most $\sqrt{{}^T\!/_h}$.

(d) *Standard (non-smoothed) regret.* As a sanity check, let us recover a standard (non-smoothed) contextual bandit problem as a special case. Let $\mathcal{A}$ be a finite set of M actions, equipped with an identity metric $\rho(a, a') := \mathbf{1}\{a \neq a'\}$ and uniform base measure ν. Then with *identity kernel* $K : a \mapsto \delta_a$ we have $\Pi_K = \Pi$.

(e) *Fixed discretization.* Interestingly, we also recover regret bounds relative to a fixed discretization of the action space $\mathcal{A}$. Formally, let $\mathcal{A}_0$ be a finite subset of $\mathcal{A}$, let the base measure be the uniform distribution over $\mathcal{A}_0$, and define the smoothing kernel K to deterministically map each action a to the closest point in $\mathcal{A}_0$. Then the smoothed policy set Π_K is precisely the set of policies whose actions are discretized to $\mathcal{A}_0$. It is easy to see that the kernel complexity ("effective number of arms") is $|\mathcal{A}_0|$.

[12] To see this, observe that for every a in $\mathcal{A}$, $(\lfloor \frac{h}{2\epsilon} \rfloor + 1)^d \leq |\, \mathtt{B}(a,h) \cap \mathcal{A} \,| \leq (\frac{2h}{\epsilon} + 2)^d$.

6.3. Results for smoothed regret

Stochastic Setting: instance-dependent results

We focus on the rectangular kernel K_h. Our results are stated in terms of the *smoothing coefficient* $\theta_h(\epsilon_0)$, as defined in (3), with *smoothed loss* suitably redefined as

$$\lambda_h(\pi) := \mathop{\mathbb{E}}_{(x,\ell)\sim\mathcal{D}} \ \mathop{\mathbb{E}}_{a\sim K_h(\pi(x))} \left[\, \ell(a) \,\right], \quad \pi \in \Pi. \tag{20}$$

Generalizing `SmoothPolicyElimination` requires the following changes. Rather than use the `Smooth` operator, we use the kernel K_h in the variance constraint, action selection scheme, and importance weighted loss. We also update the batch size parameter $\tilde{n}_m := \frac{320\kappa_h V_m}{r_m^2}$.

Theorem 11 (generalizes Theorem 1) *Consider the stochastic setting with rectangular kernel K_h, under Assumption 1. Then* `SmoothPolicyElimination` *has*

$$\texttt{Regret}(T, \Pi_{K_h}) \leq O\left(\inf_{\epsilon_0 > 0} \left\{ T\epsilon_0 + \alpha\, \theta_h(\epsilon_0)\ \log(|\Pi| T)\ \log(1/\epsilon_0) \right\} \right).$$

While Assumption 1 enables better regret rates for benign instances, the algorithm can be analyzed without this assumption, and for arbitrary kernels. The following regret bound can be extracted from the proof of Theorem 11 without much difficulty:

Theorem 12 (generalizes Corollary 2) *Consider the stochastic setting with an arbitrary smoothing kernel of kernel complexity κ. Then* `SmoothPolicyElimination` *has*

$$\texttt{Regret}(T, \Pi_K) \leq \tilde{O}\left(\sqrt{T\kappa \log(|\Pi|)} \right).$$

Adversarial setting

We use a version of `EXP4` (Algorithm 2), as before. We handle an arbitrary smoothing kernel K, obtaining a regret bound in terms of its kernel complexity κ. We obtain Theorem 3 by specializing to the unit interval, as explained in Section 6.1.

Theorem 13 (generalizes Theorem 3) *Consider the adversarial setting with an arbitrary smoothing kernel of kernel complexity κ.* `ContinuousEXP4` *(Algorithm 2) with policy set $\Xi = \Pi_K$ and learning rate $\eta = \sqrt{\frac{2 \ln|\Xi|}{T\kappa}}$ admits smoothed regret*

$$\texttt{Regret}(T, \Pi_K) \leq O\left(\sqrt{T\kappa \log|\Pi|} \right).$$

Proof One subtle point in the proof is that we separate the base measure, the smoothing kernel, and the action sampling distribution. The details follow standard techniques.

From the analysis of algorithm Hedge (Freund and Schapire, 1997), we obtain

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\xi\sim P_t} \hat{\ell}_t(\xi) - \min_{\xi\in\Xi} \sum_{t=1}^{T} \hat{\ell}_t(\xi) \leq \frac{\eta}{2} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\xi\sim P_t} \hat{\ell}_t(\xi)^2 + \frac{\log|\Xi|}{\eta}, \tag{21}$$

where P_t is a distribution over policies proportional to the weights W_t in the algorithm.

Now, by standard importance weighting arguments we have (i) $\mathbb{E}_{\xi \sim P_t} \hat{\ell}_t(\xi) = \ell_t(a_t)$ and (ii) $\mathbb{E}_{a_t \sim p_t} \hat{\ell}_t(\xi) = \mathbb{E}_{a \sim \xi(\cdot | x_t)} \ell_t(a)$. For the variance term, we have

$$\mathop{\mathbb{E}}_{a_t,\xi} \hat{\ell}_t(\xi)^2 \leq \kappa \mathop{\mathbb{E}}_{a_t,\xi} \ell_t(a_t)^2 \frac{\xi(a_t \mid x_t)}{p_t(a_t \mid x_t)^2} = \kappa \int \ell_t^2(a) \frac{p_t(a \mid x_t)}{p_t(a \mid x_t)} d\lambda(a) = \kappa \, \| \, \ell_t \, \|_2^2 \leq \kappa \, \| \, \ell_t \, \|_\infty^2 \, .$$

Therefore, taking expectation over both sides of (21), we have

$$\mathbb{E} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\xi \sim P_t} \hat{\ell}_t(\xi) - \mathbb{E} \min_{\xi \in \Xi} \sum_{t=1}^{T} \hat{\ell}_t(\xi) \leq \mathbb{E} \sum_{t=1}^{T} \frac{\eta \kappa}{2} \|\ell_t\|_\infty^2 + \frac{\log |\Xi|}{\eta}. \tag{22}$$

Applying Jensen's inequality on the left hand side, using the fact that $\|\ell_t\|_\infty \leq 1$, and optimizing for η, we obtain the claimed regret bound. ■

Uniformly-smoothed regret

We consider an arbitrary finite family of smoothing kernels $K_1 \, , \; \ldots \; , K_M$. The goal is to obtain small smoothed regret with respect to each of these kernels.

We start with a simple result in terms of the maximal kernel complexity. We use `ContinuousEXP4` (Algorithm 2) with policy set $\Xi = \cup_i ; \Pi_{K_i}$, the union of the smoothed policy classes. The analysis of Theorem 13 carries over verbatim.

Theorem 14 *Consider the adversarial setting with smoothing kernels $K_1 \, , \; \ldots \; , K_M$ defined on the same base structure. Suppose each kernel has complexity at most κ. Then* `ContinuousEXP4` *(Algorithm 2) with policy set $\Xi = \cup_{i=1}^{M} \Pi_{K_i}$ and learning rate $\eta = \sqrt{\frac{2 \ln |\Xi|}{T \kappa}}$ admits smoothed regret*

$$\texttt{Regret}(T, \Xi) \leq O\left(\sqrt{T \kappa \log |\Xi|} \right), \quad \textit{where } |\Xi| = M \cdot |\Pi|.$$

Our main result is more nuanced, obtaining improved smoothed regret relative to kernels of small kernel complexity.

Theorem 15 *Consider the adversarial setting with smoothing kernels $K_1 \, , \; \ldots \; , K_M$, whose their respective kernel complexities are $\kappa_1 \, , \; \ldots \; , \kappa_M$. Let $\kappa_\star$ and $\kappa_{\max}$ be, resp., the smallest and the largest kernel complexity. Algorithm* `Corral+EXP4` *with parameter $\beta \in [0, 1]$ guarantees the following for each $i \in [M]$:*

$$\texttt{Regret}(T, \Pi_{K_i}) \leq O\left(T^{\frac{1}{1+\beta}} \; (\, \kappa_i \log(|\Pi| M) \,)^{\frac{\beta}{1+\beta}} \right) \left(\min\left\{ M, \log \frac{\kappa_{\max}}{\kappa_\star} \right\} \right)^{\frac{1}{1+\beta}} \; \left(\frac{\kappa_i}{\kappa_\star} \right)^{\frac{\beta^2}{1+\beta}} .$$

Remark 16 *Each kernel K_i, $i \in [M]$ can have its own base structure $(\mathcal{A}, \rho_i, \nu_i)$.*

The setup above (specialized to action space $\mathcal{A} = [0,1]$) almost yields the upper bound in Theorem 4, except that we can only compete with a finite set of kernels. For example, choosing K_i as the rectangular kernel with bandwidth $h = 2^{-i}$ recovers a weaker version of the theorem. For the stronger version that competes with all bandwidths $h \in [0,1]$, we must exploit further structure. The next result achieves this, generalizing Theorem 4 to action space $\mathcal{A} = [0,1]^d$ with an arbitrary dimension d.

Theorem 17 *Consider the adversarial setting, with action space $\mathcal{A} = [0,1]^d$, $d \in \mathbb{N}$, uniform base measure ν, and base metric $\rho = \ell_\infty$. Theorem 4 extends, with h replaced by h^d.*

Compared to Theorem 17, the dependence on $T, |\Pi|$, and κ in the regret bound is unchanged. Indeed, for the rectangular kernel K_h in d dimensions, we have $\kappa = O(h^{-d})$ and of course $\kappa_\star = O(1)$ here. Therefore, the main improvement is that we have eliminated the dependence on the number of kernels, M. We also provide a refinement for the non-contextual version, eliminating the dependence on $\log|\Pi|$, which is infinite in this case.

Turning to the lower bound, observe that the lower bound in Theorem 4 is precisely the second claim here with $d = 1$. This result, coupled with the upper bound for the non-contextual version establishes the optimal uniformly-smoothed regret rate. It further implies a lower bound for competing with multiple arbitrary kernels. Specifically, there exist an action space, two kernels K_1 and K_2, and positive constants c, T_0, such that for any algorithm and any $T \geq T_0$, there exists an instance for which either

$$\texttt{Regret}(T, \Pi_{K_1}) \geq c \cdot T^{\frac{1}{1+\beta}} \kappa_1^{\frac{\beta}{1+\beta}} \quad \text{or} \quad \texttt{Regret}(T, \Pi_{K_2}) \geq c \cdot T^{\frac{1}{1+\beta}} \kappa_2^{\frac{\beta}{1+\beta}} (\kappa_2/\kappa_1)^{\frac{\beta^2}{1+\beta}}.$$

This confirms the near-optimality of Theorem 15.

6.4. Results for Lipschitz losses

Let us turn to Lipschitz contextual bandits, where we posit the Lipschitz condition (6) with Lipschitz constant $L \geq 1$. The uniform discretization approach applies to general action spaces, and yields a suitable generalization of regret bound (7). The latter is stated in terms of the covering dimension d (recall Definition 9):

$$\texttt{Regret}(T, \Pi) = \tilde{O}\left(T^{1-a}\ L^{1-2a}\ \left(\gamma \log|\Pi| \right)^a \right), \text{where} \quad a = \tfrac{1}{2+d}. \tag{23}$$

This regret bound is a departure point for several results presented below.

Stochastic setting: instance-dependent results

We make several minor modifications to $\texttt{SmoothPolicyElimination.L}$, as in Section 6.3. We use rectangular kernel K_h instead of the $\texttt{Smooth}$ operator. We also set the parameters as follows: recall from Section 5 that instead of using a single smoothing parameter throughout, $\texttt{SmoothPolicyElimination.L}$ uses bandwidth $h_m = 2^{-m}$ at epoch m. In addition, we set, $r_m = L2^{-m}$, $V_m := \mathbb{E}_{x \sim \mathcal{D}}\, \nu \left(\bigcup_{\pi \in \Pi^{(m)}} \mathtt{B}_{h_m}(\pi(x)) \right)$ and $\tilde{n}_m := \frac{320 \kappa_{h_m} V_m}{r_m^2}$.

Theorem 18 (generalizes Theorem 6) *Consider the stochastic setting under Assumption 1. Recall policy-zooming coefficient $\psi_L(\epsilon_0)$ and zooming dimension z (with constant γ),* *as defined in* (8) *and* (11)*. In this setting,* `SmoothPolicyElimination.L` *with parameter L achieves*

$$\begin{aligned}\texttt{Regret}(T,\Pi) &\leq O\left(\inf_{\epsilon_0>0} TL\epsilon_0 + \frac{\alpha\cdot\psi_L(\epsilon_0)}{L}\cdot\log(|\Pi|T)\log(1/\epsilon_0)\right)\\ &\leq \tilde{O}\left(T^{1-a}\, L^{1-2a}\, \left(\gamma\log|\Pi|\,\right)^a\right),\ \textit{where}\quad a=\tfrac{1}{2+z}.\end{aligned}$$

It is easy to see that the zooming dimension is upper-bounded by the covering dimension.

Adversarial setting

Our algorithm for smoothed regret in the adversarial setting — a suitably parameterized version of `EXP4` in Theorem 13 — yields meaningful guarantees for the Lipschitz setting, and in fact essentially recovers the optimal regret rate in (23).

Corollary 19 *Consider the adversarial setting. Let K_h, $h>0$ be a rectangular kernel, and let κ_h be its kernel complexity. Consider* `ContinuousEXP4` *(Algorithm 2) parametrized as in Theorem 13: policy set $\Xi=\Pi_{K_h}$ and and learning rate $\eta=\sqrt{\frac{2\,\ln|\Xi|}{T\kappa}}$. Then*

$$\texttt{Regret}(T,\Pi) \leq TLh + O\left(\,T\kappa_h\log|\Pi|\,\right). \tag{24}$$

We recover the regret bound Equation (23) in terms of the covering dimension d, up to the multiplicative factor of 2^d, for suitable choice of bandwidth $h=\Theta((\log|\Pi|/T)^{\frac{1}{d+2}}L^{\frac{-2}{d+2}})$.

This is an immediate consequence of Theorem 13 and the following simple fact:

Lemma 20 (generalizes Lemma 5) *Let K_h be a rectangular kernel, and $f:\mathcal{A}\to[0,1]$ be an L-Lipschitz function. Then $\left|\,\mathbb{E}_{a'\sim K_h(a)}\,f(a')-f(a)\,\right|\leq Lh$.*

Lipschitz-adaptivity

We extend Theorem 8 to higher dimension, specifically to metric space $([0,1]^d,\ \ell_\infty)$.

Theorem 21 *Consider the adversarial setting, with action space $\mathcal{A}=[0,1]^d$, $d\in\mathbb{N}$ and base metric $\rho=\ell_\infty$. Theorem 8 extends, with exponents $a=\frac{\beta}{1+d\beta+\beta}$ and $b=\frac{d\beta}{1+d\beta}$.*

7. Analysis: instance-dependent regret bounds

We prove both instance-dependent regret bounds: Theorem 11 for smoothed regret and Theorem 18 for Lipschitz losses. In fact, we present a joint proof for both results.

7.1. Auxiliary lemmas

We start by stating two auxiliary lemmas whose proofs are deferred to the end of this section. Recall that the marginal distribution over $\mathcal{X}$, denoted $\mathcal{D}_X$, is assumed to be known.

The first lemma provides a guarantee on the optimization problem (4). For a policy set $\Pi'\subset\Pi$, bandwidth h and context x, define $A(x;\Pi',h):=\bigcup_{\pi\in\Pi'}\mathtt{B}(\pi(x),h)=\bigcup_{a\in\Pi'(x)}\mathtt{B}(a,h)$

which is a subset of the action space. Similarly, let $V(\Pi', h) = \mathbb{E}_{x\sim\mathcal{D}_X}\, \nu(A(x;\Pi',h))$. Finally, for a distribution $Q \in \Delta(\Pi')$, bandwidth h, we define its induced action-selection density as

$$q(a \mid x) := \sum_{\pi\in\Pi'} Q(\pi)(K_h\pi(x))(a).$$

Note that this is the density over the action space of the action-selection distribution induced by Q on context x.

Lemma 22 *For any subset $\Pi' \subset \Pi$ with $|\Pi'| < \infty$, any bandwidth $h > 0$, and any data distribution $\mathcal{D}_X$, the program* (4) *is convex and we have*

$$\min_{Q\in\Delta(\Pi')} \max_{\pi\in\Pi'} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_X} \mathop{\mathbb{E}}_{a\sim K_h\pi(x)} \left[\frac{1}{q(a \mid x)}\right] \le V(\Pi', h).$$

Note that $V(\Pi', h) \le 1$, which yields a weaker, but more interpretable bound.

The following lemma gives a uniform deviation bound on $\hat{L}_m(\pi)$ and $\mathbb{E}_{(x,\ell)\sim\mathcal{D}}\,\langle K_{h_m}\pi(x), \ell\rangle$ in epoch m. Recall that in epoch m, the estimator $\hat{L}_m(\pi)$ is the median of several base estimators $\left\{ \hat{L}^i_m(\pi) \right\}_{i=1}^{I}$, where $I = \delta_T = 5\lceil\log(|\Pi|\log_2(T)/\delta)\rceil$ is the number of batches. In comparison to using the naive empirical mean estimator, this median-of-means estimator has the advantage that it avoids a dependency on the range of the individual losses, therefore admitting sharper concentration.

Lemma 23 (Concentration of median-of-means loss estimator) *Fix $\Pi' \subset \Pi$, $h \in (0,1)$, $\delta \in (0,1)$ and let $Q \in \Delta(\Pi')$ be the solution to* (4). *Let $I = 5\lceil\log(|\Pi|/\delta)\rceil$, $\tilde{n}$ be an integer, and $\{x_j, a_j, \ell_j(a_j)\}_{j=1}^n$ be a dataset of $n = I\tilde{n}$ samples, where $(x_j, \ell_j) \sim \mathcal{D}$ and $a_j \sim q(\cdot \mid x_j)$. Define*

$$\hat{L}(\pi) = \mathrm{median}(\hat{L}^1(\pi), \ldots, \hat{L}^I(\pi)),$$

where $\hat{L}^i(\pi) = \frac{1}{\tilde{n}} \sum_{j=(i-1)\tilde{n}+1}^{i\tilde{n}} \frac{K_h(\pi(x_j))(a_j)}{q(a_j|x_j)} \ell_j(a_j)$. Then with probability at least $1-\delta$, for all $\pi \in \Pi'$, we have

$$\left| \lambda_h(\pi) - \hat{L}(\pi) \right| \le \sqrt{\frac{80\kappa_h V(\Pi', h)}{n} \log(e|\Pi|/\delta)}.$$

7.2. Proof of Theorem 11 and Theorem 18

The proof proceeds inductively over the epochs and we will do both proofs simultaneously. In the proof of Theorem 11 we use $L(\pi) := \lambda_h(\pi)$, while for Theorem 18 we use $L(\pi) := \lambda_0(\pi) = \mathbb{E}\,\ell(\pi(x))$. In both cases $\pi^\star := \mathrm{argmin}_{\pi\in\Pi} L(\pi)$. For both proofs we use $L_m(\pi) := \lambda_{h_m}(\pi)$, noting that for Theorem 11, $L_m(\pi) = L(\pi)$. Recall the definitions of the "radii" r_m which are either 2^{-m} or $L2^{-m}$ depending on the theorem statement. In epoch m we prove two things, inductively:

1. $\pi^\star \in \Pi_{m+1}$ (assuming inductively that $\pi^\star \in \Pi_m$).
2. For all $\pi \in \Pi_{m+1}$ we have $L(\pi) \le L(\pi^\star) + 12r_{m+1}$.

Before proving these two claims, we first lower bound n_m which provides a bound on the number of epochs. Assuming $\pi^\star \in \Pi_m$, which we will soon prove, we have

$$n_m \geq \frac{\kappa_{h_m} V_m}{r_m^2} \geq \frac{\kappa_{h_m} \mathbb{E}_{x \sim \mathcal{D}_X} \nu(\mathtt{B}(\pi^\star(x), h_m))}{r_m^2} \geq \frac{1}{r_m^2} = 2^{2m}$$

The first inequality requires $\delta_T \geq 1$ (which follows since $\delta \leq 1/e$) while the third uses the fact that $\mathrm{supp}(K_{h_m}(a)) \subset \mathtt{B}(a, h_m)$ so that $\kappa_{h_m} \geq \sup_a \frac{1}{\nu(\mathtt{B}(a,h_m))}$. Hence we know that there are at most $m_T := \log_2(T)$ epochs. Applying Lemma 23 to all m_T epochs and taking a union bound, we have

$$\forall m \in [m_T], \forall \pi \in \Pi_m : \left| L_m(\pi) - \hat{L}_m(\pi) \right| \leq \sqrt{\frac{80\kappa_{h_m} V_m \delta_T}{n_m}}.$$

Here we are using the fact that $V_m = V(\Pi_m, h_m)$ where V_m is defined in the algorithm. Plugging in the choices for $n_m := \frac{320\kappa_{h_m} V_m \delta_T}{r_m^2}$ the above inequality simplifies to

$$\forall m \in [m_T], \forall \pi \in \Pi_m : \left| L_m(\pi) - \hat{L}_m(\pi) \right| \leq r_m/2. \tag{25}$$

Let us now prove the two inductive claims under the event that these inequalities hold, which occurs with probability at least $1 - \delta$. For the base case, since $\Pi_1 \leftarrow \Pi$ we clearly have $\pi^\star \in \Pi$. We also always have $L(\pi) \leq L(\pi^\star) + 2r_1$ since the losses are bounded in $[0, 1]$. For the inductive step, first we observe that for Theorem 11, $L(\pi) = L_m(\pi)$, and for Theorem 18, $|L(\pi) - L_m(\pi)| \leq Lh_m = r_m$. In conjunction with (25), in both cases, we have

$$\forall m \in [m_T], \forall \pi \in \Pi_m : \left| L(\pi) - \hat{L}_m \right| \leq 3r_m/2. \tag{26}$$

By the standard analysis of empirical risk minimization, for the first claim,

$$\hat{L}_m(\pi^\star) \leq L(\pi^\star) + 3r_m/2 = \min_{\pi \in \Pi_m} L(\pi) + 3r_m/2 \leq \min_{\pi \in \Pi_m} \hat{L}_m(\pi) + 3r_m.$$

which verifies that $\pi^\star \in \Pi_{m+1}$. For the second claim, for both Theorem 11 and Theorem 18, we have for all π in Π_{m+1},

$$L(\pi) \leq \hat{L}_m(\pi) + 3r_m/2 \leq \min_{\pi' \in \Pi_m} \hat{L}_m(\pi') + 9r_m/2 \leq L(\pi^\star) + 6r_m.$$

This proves the second claim since $r_m = 2r_{m+1}$.

For the final regret bound, define $\hat{m}_T$ to be the actual number of epochs. For each $m \in \mathbb{N}$, define $\hat{n}_m$ to be the actual number of rounds in each epoch, formally defined as follows: (1) for $m < \hat{m}_T$, $\hat{n}_m := n_m$, (2) for $m > \hat{m}_T$, $\hat{n}_m := 0$, and (3) $\hat{n}_{\hat{m}_T} = T - \sum_{m < \hat{m}_T} \hat{n}_m$. We have that $\hat{n}_m \leq n_m$ for all m and that $\sum_{m=1}^\infty \hat{n}_m = T$. Then, in the $1 - \delta$ good event, we can bound the regret of the algorithm as

$$\mathtt{Regret} \leq \sum_{m=1}^{\infty} \hat{n}_m \cdot 12r_m,$$

where we have used the fact that $\sum_{m=1}^\infty \hat{n}_m = T$.

We optimize the bound as follows: For any $\epsilon_0 > 0$, we first truncate the sum at epoch $m_{\epsilon_0} := \lceil \log \frac{1}{\epsilon_0} \rceil$. Using the fact that $r_m \le r_{m_{\epsilon_0}}$ for $m \ge m_{\epsilon_0}$, we can bound the regret in the later epochs simply by $T\epsilon_0$. For the earlier epochs we substitute the choice of $\hat{n}_m$. This gives

$$\sum_{m=1}^{\infty} 12\hat{n}_m r_m \le 12 \min_{\epsilon_0 > 0} \left(T\epsilon_0 + 320 \sum_{m \le m_{\epsilon_0} - 1} \frac{\kappa_{h_m} V_m \delta_T}{r_m} \right).$$

To simplify further, by our inductive hypothesis we know that

$$V_m \le V(\Pi_m, h_m) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \nu(A(x; \Pi_m, h_m)) \le \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \mathcal{N}_{h_m}(\Pi_m(x)) \cdot \sup_a \nu(\mathtt{B}(a, 2h_m)).$$

The final inequality is based on the fact that we can always cover $A(x; \Pi_m, h_m)$ by a union of balls of radius $2h_m$ with centers on a h_m-covering of Π_m, along with the fact that a maximum (therefore, maximal) δ-packing is a δ-covering. On the other hand we have $\kappa_{h_m} \le \sup_a \frac{1}{\nu(\mathtt{B}(a, h_m))}$, so that under Assumption 1 we have

$$\kappa_{h_m} V_m \le \alpha \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \mathcal{N}_{h_m}(\Pi_m(x))$$

Set $\mathcal{S} := \{2^{-i} : i \in \mathbb{N}\}$. For Theorem 11, using the definition of $M_h(\epsilon, \delta)$, and the fact that $\Pi_m \subseteq \Pi_{h, 12r_m}$, we have $\kappa_{h_m} V_m \le \alpha \, \mathbb{E}_{x \sim \mathcal{D}_X} \mathcal{N}_{h_m}(\Pi_m(x)) \le \alpha M_h(12r_m, h)$. Therefore, the bounds simplify to

$$\begin{aligned}
\mathtt{Regret}(T, \Pi_h) &\le 12 \min_{\epsilon_0 > 0} \left(T\epsilon_0 + 320\alpha \cdot \sum_{\epsilon \in \mathcal{S}, \epsilon \ge 2\epsilon_0} \frac{M_h(12\epsilon, h)\delta_T}{\epsilon} \right) \\
&\le 12 \min_{\epsilon_0 > 0} \left(T\epsilon_0 + 320\alpha \cdot \theta_h(\epsilon_0) \cdot \log(|\Pi| \log_2(T)/\delta) \cdot \log_2(1/\epsilon_0) \right),
\end{aligned}$$

where in the second inequality, we use the definition of $\theta_h(\epsilon)$, and the fact that there are $m_{\epsilon_0} - 1 \le \log_2(1/\epsilon_0)$ summands in the second term.

Likewise, for Theorem 18, we have

$$\begin{aligned}
\mathtt{Regret}(T, \Pi) &\le 12 \min_{\epsilon_0 > 0} \left(TL\epsilon_0 + 320\alpha \sum_{\epsilon \in \mathcal{S}, \epsilon \ge 2\epsilon_0} \frac{M_0(12L\epsilon, \epsilon)\delta_T}{L\epsilon} \right) \\
&\le 12 \min_{\epsilon_0 > 0} \left(TL\epsilon_0 + 320\alpha \cdot \psi_L(\epsilon_0)/L \cdot \log(|\Pi| \log_2(T)/\delta) \cdot \log_2(1/\epsilon_0) \right)
\end{aligned}$$

Both bounds are conditional on the good event, which happens with probability $1 - \delta$. In the bad event, the expected regret is at most T. Setting $\delta = 1/T$, the theorems follow.

7.3. Proofs for the lemmata

Proof of Lemma 22 The proof follows that of Lemma 1 of Dudik et al. (2011). We introduce the following notation: for a distribution P over a set of policies Π', bandwidth h, denote by its induced action-selection density as

$$p(a \mid x) := \sum_{\pi \in \Pi} P(\pi)(K_h \pi(x))(a).$$

Likewise, for a distribution Q over a set of policies Π', define

$$q(a \mid x) := \sum_{\pi \in \Pi} Q(\pi)(K_h \pi(x))(a).$$

Define $\mathbf{1}_{|\Pi'|}$ to be the $|\Pi'|$-dimensional vector that takes value 1 on all its entries; in addition, for policy π in Π', define e_π as the $|\Pi'|$-dimensional vector that takes value 1 on the entry that corresponds to policy π and takes value 0 everywhere else.

In addition, for Q and P in $\Delta(\Pi')$, define

$$f(Q, P) := \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \int \frac{p(a \mid x)}{q(a \mid x)} \mathbf{1}(a \in A(x; \Pi', h)) d\nu(a).$$

It suffices to show that $\min_{Q \in \Delta(\Pi')} \max_{\pi \in \Pi'} f(Q, e_\pi) \leq V(\Pi', h)$, as for any π in Π',

$$\begin{aligned} f(Q, e_\pi) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \int \frac{K_h \pi(x)(a)}{q(a \mid x)} \mathbf{1}(a \in A(x; \Pi', h)) d\nu(a) \\ &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \int \frac{K_h \pi(x)(a)}{q(a \mid x)} d\nu(a) \\ &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \mathop{\mathbb{E}}_{a \sim K_h \pi(x)} \frac{1}{q(a \mid x)}, \end{aligned}$$

where the first equality is from the fact that for all a, if $K_h \pi(x)(a) \neq 0$ then $a \notin A(x; \Pi', h)$.

Define $\mathbb{Q} := \{ Q \in \Delta(\Pi') : \max_{\pi \in \Pi'} f(Q, e_\pi) < \infty \}$. Observe that $\mathbb{Q}$ is a convex set. $\mathbb{Q}$ is nonempty, as any vector Q such that $Q_\pi > 0$ for all π in Π' (e.g. the uniform distribution over Π', $\frac{1}{|\Pi'|} \mathbf{1}_{|\Pi'|}$) is in $\mathbb{Q}$. With this notation,

$$\min_{Q \in \Delta(\Pi')} \max_{\pi \in \Pi'} f(Q, e_\pi) = \min_{Q \in \mathbb{Q}} \max_{\pi \in \Pi'} f(Q, e_\pi).$$

Now, note that

$$\min_{Q \in \mathbb{Q}} \max_{\pi \in \Pi'} f(Q, e_\pi) = \min_{Q \in \mathbb{Q}} \max_{P \in \Delta(\Pi')} \mathop{\mathbb{E}}_{\pi \sim P} f(Q, e_\pi) = \min_{Q \in \mathbb{Q}} \max_{P \in \Delta(\Pi')} f(Q, P).$$

where the first equality uses the fact that $f(Q, \cdot)$ is linear. Now, as $\Delta(\Pi')$ is compact and convex, $\mathbb{Q}$ is convex, $f(\cdot, P)$ is convex and continuous and $f(Q, \cdot)$ is concave and continuous, we may apply Sion's minimax theorem (Sion, 1958, Corollary 3.3), to obtain that the above is equal to

$$\max_{P \in \Delta(\Pi')} \min_{Q \in \mathbb{Q}} f(Q, P)$$

Now, given any P in $\Delta(\Pi')$, consider $P_\epsilon = (1 - \epsilon) P + \frac{\epsilon}{|\Pi'|} \mathbf{1}_{|\Pi'|}$. We have that P_ϵ is in $\mathbb{Q}$. Moreover,

$$\begin{aligned} f(P_\epsilon, P) &\leq \mathop{\mathbb{E}}_{\pi \sim P} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \int \frac{p(a \mid x)}{(1 - \epsilon) p(a \mid x)} \mathbf{1}(a \in A(x; \Pi', h)) d\nu(a) \\ &= \frac{1}{1 - \epsilon} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \nu(A(x; \Pi', h)) = \frac{1}{1 - \epsilon} V(\Pi', h). \end{aligned}$$

Letting $\epsilon \to 0$, this implies that for any P in $\Delta(\Pi')$, $\inf_{Q\in\mathbb{Q}} f(Q,P) \le V(\Pi',h)$. Therefore,

$$\max_{P\in\Delta(\Pi')} \min_{Q\in\mathbb{Q}} f(Q,P) \le V(\Pi',h).$$

The lemma follows. ■

Proof of Lemma 23 First, as we have seen, $\mathbb{E}\,\hat{\ell}_i(\pi(x_i)) = \lambda_h(\pi)$. Moreover,

$$\begin{aligned}\mathrm{Var}\left(\hat{\ell}_i(\pi(x_i))\right) &\le \mathbb{E}\left[\hat{\ell}_i(\pi(x_i))^2\right] = \mathop{\mathbb{E}}_{(x,\ell)\sim\mathcal{D}}\left[\int \frac{(K_h\pi(x))^2(a)\ell(a)^2}{q(a|x)} d\nu\right] \\ &\le \kappa_h V(\Pi',h) \le \kappa_h V(\Pi',h).\end{aligned}$$

where the penultimate inequality uses the fact that Q is the solution to (4), so it satisfies the guarantee in Lemma 22. Therefore, using Lemma 29 below, we have that for every $\pi \in \Pi'$, with probability at least $1 - \frac{\delta}{|\Pi|}$, the following holds:

$$\left|\bar{L}(\pi) - \hat{L}(\pi)\right| \le \sqrt{\frac{80\kappa_h V(\Pi',h)}{n}\log(e|\Pi|/\delta)}.$$

The lemma is concluded by taking a union bound over all π in Π'. ■

8. Analysis: smoothness-adaptive guarantees

We now turn to smoothness-adaptive guarantees: Theorem 15 and Theorem 17 for uniformly-smoothed regret bounds, and Theorem 21 for Lipschitz-adaptive regret bounds. We prove these results via a joint exposition, building on `Corral` algorithm from Agarwal et al. (2017b).

8.1. Stability of `ContinuousEXP4`

We start with a result on the *stability* of `ContinuousEXP4`. We consider a slightly modified protocol. The learner is now presented with randomized loss functions ℓ_t which are generated by importance weighting an original loss function $\bar{\ell}_t$ with some probability p_t set by the adversary. Formally $\ell_t = Q_t\bar{\ell}_t/p_t$ where $Q_t \sim \mathrm{Ber}(p_t)$ at each round t. Note that here, the losses presented to the learner are not guaranteed to be bounded, but we do have variance information, via p_t. The original losses $\bar{\ell}_t$ are bounded in $[0,1]$. Note further that p_t is revealed at the beginning of round t.

In this setup, Agarwal et al. (2017b) define the following notion of stability.

Definition 24 (See Agarwal et al. (2017b), Definitions 3 and 14) *An algorithm with policy class Ξ is called $(\beta, R(T))$-*stable*, if in the above protocol it achieves*

$$\mathbb{E}\sum_{t=1}^{T} \bar{\ell}_t(a_t) - \min_{\xi\in\Xi}\mathbb{E}\sum_{t=1}^{T}\left\langle \xi(x_t), \bar{\ell}_t\right\rangle \le \mathbb{E}[\rho]^\beta \cdot R(T), \tag{27}$$

where $\rho := \max_{t\in[T]} \frac{1}{p_t}$.

The definition here is slightly different than the one in Agarwal et al. (2017b, Definition 3), in that the right hand side has the term $\mathbb{E}[\rho]^\beta$ instead of $\mathbb{E}[\rho^\beta]$. This has no bearing on the analysis of `Corral`, but is important for our application, as we will see.

Agarwal et al. (2017b) shows that `EXP4` is $(1/2, \sqrt{KT\log|\Pi|})$-stable in the discrete action setting, where K is the number of actions. We provide a similar result here, replacing K with κ and establishing stability whenever the first parameter is in $[0, 1/2]$.

Theorem 25 *Algorithm 2 is* $\left(\frac{\beta}{1+\beta}, O\left(T^{\frac{1}{1+\beta}}(\kappa\log|\Xi|)^{\frac{\beta}{1+\beta}}\right)\right)$*-stable, for each* $\beta \in [0,1]$.

Proof For this proof only, we use $\xi(x_t)$ to denote the density for the action distribution of expert ξ on context x_t, with respect to ν. Thus the expected loss for expert ξ on round t is $\langle \xi(x_t), \ell_t\rangle$.

We first show a weaker form of stability. Suppose that $\hat{\rho} \geq \max_{t\in[T]} 1/p_t$ is provided to the algorithm ahead of time. Then following the analysis for `EXP4`, we have

$$\mathbb{E}\sum_{t=1}^T \ell_t(a_t) - \min_{\xi\in\Xi}\mathbb{E}\sum_{t=1}^T \langle\xi(x_t),\ell_t\rangle \leq \mathbb{E}\,\frac{\eta\kappa}{2}\sum_{t=1}^T \|\ell_t\|_\infty^2 + \frac{\log|\Xi|}{\eta}.$$

The key observation is that,

$$\mathbb{E}\sum_{t=1}^T \|\ell_t\|_\infty^2 \leq \mathbb{E}\sum_{t=1}^T \frac{Q_t}{p_t^2} = \sum_{t=1}^T \mathbb{E}\, 1/p_t \leq T\hat{\rho}$$

Therefore, with the choice of $\eta = \sqrt{\frac{2\log|\Xi|}{T\kappa\hat{\rho}}}$, and using the fact that the conditional expectation of ℓ_t is $\bar{\ell}_t$, we get

$$\mathbb{E}\sum_{t=1}^T \bar{\ell}_t(a_t) - \min_{\xi\in\Xi}\mathbb{E}\sum_{t=1}^T \left\langle\xi(x_t),\bar{\ell}_t\right\rangle \leq \sqrt{2\kappa T\log|\Xi|\cdot\hat{\rho}}. \tag{28}$$

This proves a weaker version of stability, where a bound on ρ is specified in advance. The stronger version is based on the "doubling trick" argument in Agarwal et al. (2017b, Theorem 15). We run `EXP4` with a guess for $\hat{\rho}$ and if we experience a round t where $1/p_t > \hat{\rho}$, we double our guess and restart the algorithm, always with learning rate $\eta = \sqrt{\frac{2\log|\Xi|}{T\kappa\hat{\rho}}}$. In their Theorem 15, they prove that if an algorithm is weakly stable in the sense of (28) then, with restarts, it is strongly stable according to Definition 24. In our setting, their result reveals that the restarting variant of `EXP4` guarantees

$$\mathbb{E}\sum_{t=1}^T \bar{\ell}_t(a_t) - \min_{\xi\in\Xi}\mathbb{E}\sum_{t=1}^T \left\langle\xi(x_t),\bar{\ell}_t\right\rangle \leq \frac{\sqrt{2}}{\sqrt{2}-1}\cdot\mathbb{E}[\rho]^{\frac{1}{2}}\cdot\sqrt{2\kappa T\log|\Xi|}$$

To obtain a stability guarantee for every β, since the regret is trivially at most T, we obtain

$$\mathbb{E}\sum_{t=1}^T \bar{\ell}_t(a_t) - \min_{\xi\in\Xi}\mathbb{E}\sum_{t=1}^T \left\langle\xi(x_t),\bar{\ell}_t\right\rangle \leq \min\left(T, c\,\mathbb{E}[\rho]^{\frac{1}{2}}\sqrt{\kappa T\log|\Xi|}\right) \leq cT^{\frac{1}{1+\beta}}(\mathbb{E}[\rho]\kappa\log|\Xi|)^{\frac{\beta}{1+\beta}},$$

where $c > 0$ is a universal constant. The second inequality is from the simple fact that $\min(A, B) \leq A^\gamma B^{(1-\gamma)}$ for $A, B > 0$, $\gamma\in[0,1]$. ■

8.2. Corral+EXP4 and its analysis

We first provide a formal description of Corral+EXP4 in the notation of abstract smoothing kernels. Given a family of smoothing kernels $\mathcal{K} = \{ K_1, \ldots, K_M \}$, we bucket the kernels according to their kernel complexity κ, $\mathcal{K}_b = \{ i \in [M] : \lceil \log \kappa_{K_i} \rceil = b \}$ for each $b \in \mathbb{N}$, and we initialize one instance of EXP4 with restarting for each bucket. Then we run Corral over these instances. We call $B = \{ \lceil \log \kappa_{K_i} \rceil : i \in \{ 1, \ldots, M \} \}$ the set of "active" indices.

Define $r := \frac{\max_{K \in \mathcal{K}} \kappa_K}{\min_{K \in \mathcal{K}} \kappa_K}$ and $\kappa_\star := \min_{K \in \mathcal{K}} \kappa_K$. Observe that $B \leq \min \{ M, \log r + 1 \}$. We have the following guarantee for Corral+EXP4.

Lemma 26 *Suppose* Corral+EXP4 *is run with learning rate η and horizon T. Then, for all $\beta \in [0, 1]$, it has the following regret guarantee simultaneously for all kernels K in $\mathcal{K}$:*

$$\texttt{Regret}(T, \Pi_K) \leq \tilde{O} \left(\frac{\min \{ M, \log r \}}{\eta} + T\eta + T \left(\eta \ln(|\Pi| M) \kappa_K \right)^\beta \right).$$

Proof This is almost a direct consequence of Agarwal et al. (2017b, Theorem 4). By the definition of $\mathcal{K}_b$ and Ξ_b, $\kappa_b := \max_{\xi \in \Xi_b} \max_{a,x} \xi(a|x) \leq 2^b$. In addition, $|\Xi_b| \leq |\Pi| \cdot M$. Since for all $K \in \mathcal{K}_b$ we have $\lceil \log \kappa_K \rceil = b$, therefore $\kappa_K \in (2^{b-1}, 2^b]$. By applying Theorem 25 we see that EXP4 with restarting has the stability guarantee when measuring regret against $\texttt{Bench}(\Pi_{K_i})$ for each $K_i \in \mathcal{K}_b$.

Now, by Theorem 4 of (Agarwal et al., 2017b), Corral ensures $\forall b \in [B], \forall K \in \mathcal{K}_B$:

$$\texttt{Regret}(T, \Pi_K) \leq \tilde{O} \left(\frac{B}{\eta} + T\eta - \frac{\mathbb{E}[\rho_b]}{\eta \log T} + T^{\frac{1}{1+\beta}} \left(\mathbb{E}[\rho_b] \kappa_K \log(|\Pi| M) \right)^{\frac{\beta}{1+\beta}} \right)$$

Optimizing over $\mathbb{E}[\rho_b]$ gives

$$\forall K \in \mathcal{K} : \texttt{Regret}(T, \Pi_K) \leq \tilde{O} \left(\frac{B}{\eta} + T\eta + T \left(\eta \kappa_K \log(|\Pi| M) \right)^\beta \right).$$

The result follows by observing that $B \leq \min \{ M, \log r \}$. ■

Proof of upper bound in Theorem 15 We simply run Corral+EXP4 with

$$\eta = \frac{B^{\frac{1}{1+\beta}}}{T^{\frac{1}{1+\beta}} (\ln(|\Pi| M) \kappa_\star)^{\frac{\beta}{1+\beta}}},$$

and apply Lemma 26. ■

Proof of upper bounds in Theorem 17 Recall that for Theorem 17 we are in the d-dimensional cube with uniform base measure and with ℓ_∞ metric. Our goal is to obtain a uniformly-smoothed regret guarantee for all bandwidths $h \in [0, 1]$, where we are using the rectangular kernel. This requires a bit more work.

First, set $D := d 2^{d+2} T^2$ and form the discretized set:

$$\mathcal{H} = \left\{ h \in \{ \tfrac{1}{D}, \tfrac{2}{D}, \ldots, 1 \} : 1 \leq \tfrac{1}{h^d} \leq 2^{\lceil \log_2 T \rceil + 1} \right\}.$$

We run `Corral` with kernel class $\mathcal{K} = \{K_h : h \in \mathcal{H}\}$ and we use `EXP4` with restarts as the sub-algorithms. As $|\mathcal{H}| \le d2^{d+2}T^2$, applying Theorem 15 gives

$$\forall h \in \mathcal{H} : \texttt{Regret}(T, \Pi_h) \le \tilde{O}\left(T^{\frac{1}{1+\beta}} h^{-d\beta} (\log|\Pi|)^{\frac{\beta}{1+\beta}} \right). \tag{29}$$

We now must lift (29) to all $h \in [0,1]$. We have the following lemma.

Lemma 27 *For any loss $\ell : \mathcal{A} \to [0,1]$ and bandwidth $h \ge T^{-1/d}$, there exists $\hat{h} \in \mathcal{H}$ such that $\frac{1}{\hat{h}^d} \le \frac{2}{h^d}$ and $\sup_a \left\langle K_{\hat{h}}(a) - K_h(a), \ell_t \right\rangle \le \frac{1}{T}$.*

Applying this lemma allows us to obtain a smoothed regret bound for $h \notin \mathcal{H}$ by translating to $\hat{h} \in \mathcal{H}$, since the former benchmark is smaller by at most $O(1)$ while the latter has $\hat{h}^{-d} \le 2(h)^{-d}$. This yields Theorem 17.

For the non-contextual bound, instantiate each sub-algorithm with a policy set $\Pi' : \{x_0 \mapsto a : a \in \mathcal{A}'\}$ where $\mathcal{A}'$ is a ε-covering of $\mathcal{A}$, satisfying $|\mathcal{A}'| \le O(\epsilon^{-d})$. The above analysis carries through, and to translate to $a \notin \mathcal{A}'$ we require a different discretization lemma.

Lemma 28 *For $\rho(a, a') \le \varepsilon$ and $\ell : \mathcal{A} \to [0,1]$, we have $|\left\langle K_h(a) - K_h(a'), \ell \right\rangle| \le 4d\varepsilon h^{-d}$.*

The proofs of both lemmas are deferred to the end of this section.

To finish the proof set $\varepsilon = \frac{1}{4dT^2}$ and note that for $h < T^{-1/d}$ the desired guarantee is trivial. Thus for all $h \ge T^{-1/d}$ the cumulative approximation error introduced by discretization is at most 1 while the policy set Π' has $\ln|\Pi'| \le O(d \log dT)$. ■

Proof of upper bounds in Theorem 21 For a finite set of bandwidths $\mathcal{H}$ let us apply Lemma 26 with $\mathcal{K} = \{K_h : h \in \mathcal{H}\}$ to obtain

$$\forall h \in \mathcal{H} : \texttt{Regret}(T, \Pi_h) \le \tilde{O}\left(\frac{|\mathcal{H}|}{\eta} + T\eta + T\left(\eta \log(|\Pi||\mathcal{H}|) h^{-d} \right)^{\beta} \right)$$

Applying Lemma 20, we know that

$$\min_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{T} \left\langle K_h \pi(x_t), \ell_t \right\rangle \le \min_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{T} \ell_t(\pi(x_t)) + TLh,$$

and so we obtain

$$\texttt{Regret}(T, \Pi) \le \min_{h \in \mathcal{H}} TLh + \tilde{O}\left(\frac{|\mathcal{H}|}{\eta} + T\eta + T\left(\eta \log(|\Pi||\mathcal{H}|) h^{-d} \right)^{\beta} \right).$$

Define $\mathcal{L} = \{2^i : i \in \{1, 2, \ldots, \lceil \log_2(T) \rceil\}\}$ to be an exponentially spaced grid. If the true parameter $L \ge T$ then the bound is trivial, and otherwise $L \le \hat{L} \le 2L$ from some $\hat{L} \in \mathcal{L}$. We choose $\mathcal{H}$ of size $\lceil \log_2(T) \rceil$ to optimize the above bound for each $\hat{L} \in \mathcal{L}$. Specifically, set

$$\mathcal{H} = \left\{ h_i = (\eta \log(|\Pi| \log_2(T)))^{\frac{\beta}{d\beta+1}} 2^{\frac{-i}{d\beta+1}} : i \in [\lceil \log_2(T) \rceil] \right\}.$$

This yields

$$\begin{aligned}\texttt{Regret}(T,\Pi) &\le \min_{h\in\mathcal{H}} TLh + \tilde{O}\left(\frac{|\mathcal{H}|}{\eta} + T\eta + T\left(\eta\log(|\Pi||\mathcal{H}|)h^{-d}\right)^{\beta}\right).\\ &\le \min_{h\in\mathcal{H}} T\hat{L}h + \tilde{O}\left(\frac{|\mathcal{H}|}{\eta} + T\eta + T\left(\eta\log(|\Pi||\mathcal{H}|)h^{-d}\right)^{\beta}\right)\\ &\le \tilde{O}\left(T\hat{L}^{\frac{d\beta}{d\beta+1}}(\eta\log|\Pi|)^{\frac{\beta}{d\beta+1}} + \frac{1}{\eta} + T\eta\right)\\ &\le \tilde{O}\left(TL^{\frac{d\beta}{d\beta+1}}(\eta\log|\Pi|)^{\frac{\beta}{d\beta+1}} + \frac{1}{\eta} + T\eta\right).\end{aligned}$$

We finish the proof by tuning the master learning rate η while ignoring L. This gives

$$\eta = T^{\frac{-(d\beta+1)}{1+(d+1)\beta}}(\log|\Pi|)^{\frac{-\beta}{1+(d+1)\beta}},$$

and the overall regret bound is

$$\texttt{Regret}(T,\Pi) \le \tilde{O}\left(L^{\frac{d\beta}{1+d\beta}}T^{\frac{1+d\beta}{1+(d+1)\beta}}(\log|\Pi|)^{\frac{\beta}{1+(d+1)\beta}}\right).$$

As in the proof of Theorem 17, for the non-contextual case we discretize the action space to a minimal ε cover $\mathcal{A}'$ for $\mathcal{A}$. Choosing $\varepsilon = (4dT^2)^{-1}$ as in that proof suffices here as well. ■

We remark that Theorem 21 is not a direct corollary of Theorem 17. Rather we must start with Lemma 26 and first tune h to balance the sub-algorithm's regret with the TLh term. Then we tune the master's learning rate. In particular for fixed exponent β the master learning rates for Theorem 17 and Theorem 21 are different.

8.3. Proofs of the lemmata

We prove a few auxiliary lemmas used in the previous sections, namely Lemma 27, Lemma 28 and Lemma 29.

Lemma 29 (follows from Hsu and Sabato (2016)) *Suppose $\delta \in (0,1)$, $k = 5\lceil \ln\frac{1}{\delta}\rceil$, $\tilde{n}$ is an integer, and $n = k\tilde{n}$. In addition, $X_1,\ldots,X_n$ are iid random variables with mean μ and variance σ^2. Define*

$$\hat{\mu} := \text{median}\left\{\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} X_i, \frac{1}{\tilde{n}}\sum_{i=\tilde{n}+1}^{2\tilde{n}} X_i, \ldots, \frac{1}{\tilde{n}}\sum_{i=(k-1)\tilde{n}+1}^{k\tilde{n}} X_i\right\}.$$

Then with probability $1-\delta$,

$$|\hat{\mu}-\mu| \le \sigma\sqrt{\frac{40\ln\frac{e}{\delta}}{n}}.$$

Proof From the first part of Hsu and Sabato (2016, Proposition 5), taking $k = 5\lceil \ln\frac{1}{\delta}\rceil$, we have that with probability $1-\delta$,

$$|\hat{\mu}-\mu| \le \sigma\sqrt{\frac{8k}{n}}.$$

The proof is completed by noting that $k \le 5(1 + \ln \frac{1}{\delta}) = 5 \ln \frac{e}{\delta}$. ■

Proof of Lemma 27 Recall the definition of $\mathcal{H}$:

$$\mathcal{H} := \left\{ h \in \left\{ \frac{1}{D}, \frac{2}{D}, \ldots, 1 \right\} : 1 \le \frac{1}{h^d} \le 2^{\lceil \log T \rceil + 1} \right\}.$$

Set $h_D = \frac{\lfloor hD \rfloor}{D}$. Note that h_D is a multiple of $\frac{1}{D}$. In addition, we note that $h \ge T^{-1}$, and $h_D \ge h - \frac{1}{d2^{d+2}T^2} \ge h - \frac{1}{4dT^2} \ge h(1 - \frac{1}{4dT})$. Therefore, by Fact 30 below, $\frac{1}{h_D^d} \le \frac{1}{h^d}(\frac{1}{1-\frac{1}{4dT}})^d \le \frac{2}{h^d} \le 2T \le 2^{\lceil \log T \rceil + 1}$. Hence, h_D is in $\mathcal{H}$. Moreover, $\nu(\mathtt{B}(a,h)) \ge h^d$, and

$$\nu(\mathtt{B}(a,h)\Delta\mathtt{B}(a,h_D)) \le (2h)^d - (2h_D)^d \le (2h)^d(1 - (1 - \frac{1}{d2^{d+2}T})^d) \le \frac{(2h)^d}{2^d T} = \frac{h^d}{2T}.$$

Here Δ denotes the symmetric set difference. Therefore, applying Fact 31, we obtain

$$|\langle K_h(a) - K_{h_D}(a), \ell \rangle| \le \frac{2\nu(\mathtt{B}(a,h)\Delta\mathtt{B}(a,h_D))}{\max\{\nu(\mathtt{B}(a,h)), \nu(\mathtt{B}(a,h_D))\}} \le \frac{1}{T}.$$ ■

Proof of Lemma 28 Since we are using the ℓ_∞ distance and $\rho(a,a') \le \varepsilon$, we have that $\nu(\mathtt{B}(a,h)\Delta\mathtt{B}(a',h)) \le 2\,\|a - a'\|_1 \le 2d\varepsilon$. Applying Fact 31 we obtain

$$\left|\langle K_h(a) - K_h(a'), \ell \rangle\right| \le \frac{2\nu(\mathtt{B}(a,h)\Delta\mathtt{B}(a',h))}{\max\{\nu(\mathtt{B}(a,h)), \nu(\mathtt{B}(a',h))\}} \le 4d\varepsilon h^{-d}.$$ ■

Fact 30 *For $T, d \ge 1$, $\left(\frac{1}{1-\frac{1}{4dT}}\right)^d \le 1 + \frac{1}{T}$.*

Proof We use the following simple facts: for all x in $[0,1]$, $e^x \le 1 + 2x$ and $e^{-x} \le 1 - \frac{1}{2}x$. The proof is completed by noting that $\frac{1}{(1-\frac{1}{4dT})^d} \le e^{\frac{1}{2T}} \le 1 + \frac{1}{T}$. ■

Fact 31 *For sets S_1 and S_2, and a loss function $\ell : \mathcal{A} \to [0,1]$*

$$\left| \frac{\int_{S_1} \ell(a) d\nu(a)}{\nu(S_1)} - \frac{\int_{S_2} \ell(a) d\nu(a)}{\nu(S_2)} \right| \le \frac{2\nu(S_1 \Delta S_2)}{\max(\nu(S_1), \nu(S_2))}.$$

Proof

$$\begin{aligned}
&\left| \frac{\int_{S_1} \ell(a) d\nu(a)}{\nu(S_1)} - \frac{\int_{S_2} \ell(a) d\nu(a)}{\nu(S_2)} \right| \\
&= \left| \frac{\int_{S_1} \ell(a) d\nu(a) \cdot (\nu(S_2) - \nu(S_1)) + \nu(S_1) \cdot (\int_{S_1} \ell(a) d\nu(a) - \int_{S_2} \ell(a) d\nu(a))}{\nu(S_1)\nu(S_2)} \right| \\
&\le \frac{\nu(S_1) \cdot \nu(S_1 \Delta S_2) + \nu(S_1) \cdot \nu(S_1 \Delta S_2)}{\nu(S_1)\nu(S_2)} = \frac{2\nu(S_1 \Delta S_2)}{\nu(S_2)}
\end{aligned}$$

By symmetry, the above is also bounded by $\frac{2\nu(S_1 \Delta S_2)}{\nu(S_1)}$. The proof is completed by taking the smaller of the two upper bounds. ■

9. Lower bounds for smoothness-adaptive algorithms

In this section, we prove the lower bounds in Theorem 15 and Theorem 21, showing that the exponent combinations we achieve with `Corral` are optimal. We start with two lemmas that describe the constructions and contain the main technical argument. In the next subsection we prove the theorems.

9.1. The constructions

The following two lemmas are based on a construction due to Locatelli and Carpentier (2018). Their work concerns adapting to the smoothness exponent, while ours focuses on the smoothness constant. We also use a similar construction to show lower bounds against uniformly-smoothed algorithms.

We focus on the stochastic non-contextual setting, where we consider policy class $\Pi = \{ x_0 \mapsto a : a \in \mathcal{A} \}$, and at each time, a dummy context x_0 is shown. We use the shorthand $\texttt{Regret}(T, h)$ to denote $\texttt{Regret}(T, \Pi_h)$. We define Λ to be the set of all functions from $\mathcal{A}$ to $[0, 1]$. A function $\lambda \in \Lambda$ defines an instance where $\ell(a) \sim \mathrm{Ber}(\lambda(a))$ for all $a \in \mathcal{A}$.

Lemma 32 *Fix any $T \in \mathbb{N}$ and $h \in (0, 1/8]$. Suppose an algorithm* Alg *guarantees that for all instances $\lambda \in \Lambda$, $\texttt{Regret}(T, 1/4) \le R_S(1/4, T)$ where $R_S(1/4, T) \le \frac{\sqrt{T}}{20(8h)^d}$. Then there exists $\lambda \in \Lambda$ such that* Alg *has*

$$\texttt{Regret}(T, h) \ge \min\left\{ \frac{T}{40 \cdot 2^d}, \frac{T}{400(8h)^d R_S(1/4, T)} \right\}.$$

Proof We let $N = \lfloor 1/4h \rfloor^d$. Note that as $h \le 1/8$, $(1/8h)^d \le N \le (1/4h)^d$. We also define $\Delta = \min\left\{ \frac{N}{40 R_\mathrm{S}(1/4, T)}, 1/4 \right\} \in (0, 1/4]$. By our assumption that $R_\mathrm{S}(1/4, T) \le \frac{\sqrt{T}}{20(8h)^d}$, we have

$$R_\mathrm{S}(1/4, T) \le \min\left\{ \frac{N^2 T}{200 R_\mathrm{S}(1/4, T)}, \frac{NT}{20} \right\} = 0.2 N T \Delta. \tag{30}$$

For each tuple $(s_1, \ldots, s_d) \in [\lfloor 1/4h \rfloor]^d$, we define a point $c_{s_1, \ldots, s_d} = (h(2s_1 - 1), \ldots, h(2s_d - 1))$. There are N points in total, which we call $c_1, \ldots, c_N$. Define regions

$$H_i = \mathtt{B}(c_i, h), i = 1, \ldots, N,$$

which are disjoint subsets in $[0, 1/2]^d$. Finally, define $S = [1/2, 1]^d = \mathtt{B}(c_0, 1/4)$, where $c_0 = (3/4, \ldots, 3/4)$. We define several plausible loss functions $\phi_0, \ldots, \phi_N \in \Lambda$:

$$\phi_0(a) = \begin{cases} 1/2, & a \notin S \\ 1/2 - \Delta/2 & a \in S \end{cases} \qquad \text{and} \qquad \phi_i(a) = \begin{cases} 1/2, & a \notin (H_i \cup S) \\ 1/2 - \Delta & a \in H_i \\ 1/2 - \Delta/2 & a \in S \end{cases}$$

Note that $\mathbb{E}_{a \sim \mathtt{Smooth}_{1/4}(c_0)}\, \phi_0(a) = 1/2 - \Delta/2$, and $\mathbb{E}_{a \sim \mathtt{Smooth}_h(c_i)}\, \phi_i(a) = 1/2 - \Delta$.

The environments are parameterized by ϕ_i where losses are always Bernoulli with mean ϕ_i. Denote by $\mathbb{E}_i$ (resp. $\mathbb{P}_i$) the expectation (resp. probability) over the randomness of the algorithm, along with the randomness in environment ϕ_i.

Observe that under environment ϕ_0, for $h = 1/4$, we have $T \cdot \min_a \lambda_{1/4}(a) = T \cdot (1/2 - \Delta/2)$. Since Alg guarantees that $\texttt{Regret}(T, 1/4) \leq R_{\mathrm{S}}(1/4, T)$, we have

$$\mathbb{E}_0 \sum_{t=1}^{T} \phi_0(a_t) - T \cdot (1/2 - \Delta/2) \leq R_{\mathrm{S}}(1/4, T).$$

As for all a, $\phi_0(a) - (1/2 - \Delta/2) = \Delta/2 \mathbf{1}\left\{ a \notin S \right\}$, we get that

$$\sum_{t=1}^{T} \mathbb{E}_0 \mathbf{1}\left\{ a_t \notin S \right\} \leq \frac{2R(1/4, T)}{\Delta}.$$

Denote by $T_i = \sum_{t=1}^{T} \mathbf{1}\left\{ a_t \in H_i \right\}$ and observe that

$$\sum_{j=1}^{N} \mathbb{E}_0[T_j] \leq \mathbb{E}_0\left[\mathbf{1}\left\{ a_t \in \cup_{j=1}^{N} H_j \right\} \right] \leq \sum_{t=1}^{T} \mathbb{E}_0\left[\mathbf{1}\left\{ a_t \notin S \right\} \right] \leq \frac{2R(1/4, T)}{\Delta}.$$

By the pigeonhole principle, there exists at least one i such that

$$\mathbb{E}_0[T_i] \leq \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_0[T_j] \leq \frac{2R(1/4, T)}{N\Delta}. \tag{31}$$

Therefore, by Lemma 33 and the fact that $\Delta \leq 1/4$, we have

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i) \leq \mathbb{E}_0[T_i] \cdot (4\Delta^2) \leq \frac{8R(1/4, T)\Delta}{N}.$$

By the choice of $\Delta \leq \frac{N}{40R(1/4,T)}$, we have $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i) \leq 0.2$ and so Pinsker's inequality yields $\mathrm{d}_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) \leq \sqrt{1/2 \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)} \leq 0.4$. Therefore,

$$\mathbb{E}_i[T_i] \leq \mathbb{E}_0[T_i] + T \cdot d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_i) \leq \frac{2R_{\mathrm{S}}(1/4, T)}{N\Delta} + 0.4T \leq 0.8T.$$

where the first inequality is from the definition of the total variation distance and that $T_i \in [0, T]$ almost surely; the second inequality is by (31); the third inequality is by (30). Therefore, $\mathbb{E}_i[T_i] \leq 0.8T$, which implies that on ϕ_i

$$\begin{aligned} \texttt{Regret}(T, h) = \mathbb{E}_i \sum_{t=1}^{T} \phi_i(a_t) - (1/2 - \Delta) &\geq \frac{\Delta}{2} \cdot (T - \mathbb{E}_i[T_i]) \geq \frac{\Delta}{2} \cdot 0.2T \\ &\geq \min\left\{ \frac{T}{40 \cdot 2^d}, \frac{T}{400(8h)^d R_{\mathrm{S}}(1/4, T)} \right\}. \end{aligned}$$

■

Lemma 33 *For* $\Delta \in [0, \frac{1}{4}]$, $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i) \leq \mathbb{E}_0[T_i] \cdot (4\Delta^2)$.

Proof We abbreviate l_t as the outcome of $\ell_t(a_t)$. We have the following:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i) &= \sum_{a_1, l_1, \ldots, a_T, l_T} \mathbb{P}_0(a_1, l_1, \ldots, a_T, l_T) \log \frac{\mathbb{P}_0(a_1, l_1, \ldots, a_T, l_T)}{\mathbb{P}_i(a_1, l_1, \ldots, a_T, l_T)} \\
&= \mathbb{E}_0 \sum_{t=1}^{T} \log \frac{\mathbb{P}_0(l_t | a_t)}{\mathbb{P}_i(l_t | a_t)} = \mathbb{E}_0 \sum_{t=1}^{T} \mathbf{1}\left\{ a_t \in H_i \right\} \cdot \mathrm{KL}(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2 - \Delta)) \\
&= \mathbb{E}_0[T_i] \cdot (-\frac{1}{2} \log(1 - 4\Delta^2)) \le \mathbb{E}_0[T_i] \cdot (4\Delta^2)
\end{aligned}$$

where the last inequality uses the fact that $\log(1 - x/2) \ge -x$ for $x \in [0, 1]$. ■

For the next lemma, let $\Lambda(L)$ be the set of all L-Lipschitz mean loss functions.

Lemma 34 *Fix any $T \in \mathbb{N}$ and $L \ge 1$. Suppose an algorithm* Alg *guarantees that for all instances λ in $\Lambda(1)$, $\texttt{Regret}(T, 0) \le R_{Lip}(1, T)$ where $R_{Lip}(1, T) \le \frac{T}{40} L^d$. Then there exists a loss function $\lambda \in \Lambda(L)$ such that*

$$\texttt{Regret}(T, 0) \ge \min \left\{ \frac{T}{80}, \frac{T L^{\frac{d}{d+1}}}{3200 R_{Lip}(1, T)^{\frac{1}{d+1}}} \right\}.$$

Proof We let $\Delta = \min \left\{ \left(\frac{L^d}{40 \cdot R_{\mathrm{Lip}}(1,T) \cdot 8^d}\right)^{\frac{1}{d+1}}, 1/8 \right\} \in (0, 1/8]$, and $N = \lfloor L/4\Delta \rfloor^d$. As $L \ge 1$, $L/4\Delta \ge 2$. Therefore, $(L/8\Delta)^d \le N \le (L/4\Delta)^d$. Observe that by the choices of Δ and N:

$$\Delta \le \frac{(\frac{L}{8\Delta})^d}{40 R_{\mathrm{Lip}}(1, T)} \le \frac{N}{40 R_{\mathrm{Lip}}(1, T)}.$$

By our assumption that $R_{\mathrm{Lip}}(1, T) \le \frac{T}{40} L^d$, we have that

$$R_{\mathrm{Lip}}(1, T) \le 0.2T \cdot \frac{L^d}{8^d (1/8)^{d-1}} \le 0.2T \cdot \frac{L^d}{8^d \Delta^{d-1}} \le 0.2 N T \Delta,$$

where the first inequality is from that $R_{\mathrm{Lip}}(1, T) \le \frac{T}{40} L^d$; the second inequality is from the fact that $\Delta \le \frac{1}{8}$; the third inequality is from the fact that $N \ge (L/8\Delta)^d$.

For each tuple $(s_1, \ldots, s_d) \in [\lfloor L/4\Delta \rfloor]^d$, define point $c_{s_1, \ldots, s_d} = (\frac{\Delta}{L}(2s_1 - 1), \ldots, \frac{\Delta}{L}(2s_d - 1))$. There are N points in total which we call $c_1, \ldots, c_N$. Define regions

$$H_i = \mathtt{B}\left(c_i, \frac{\Delta}{L}\right), i = 1, \ldots, N,$$

which are disjoint subsets in $[0, 1/2]^d$. Finally, define $S = [1/2, 1]^d = \mathtt{B}(c_0, 1/4)$, where $c_0 = (3/4, \ldots, 3/4)$. We define several plausible loss functions $\phi_0 \in \Lambda(1)$, $\phi_1, \ldots, \phi_N \in \Lambda(L)$:

$$\phi_0(a) = \begin{cases} 1/2 - (\Delta/2 - ||a - c_0||_\infty)_+ & a \in S \\ 1/2 & \text{else} \end{cases}, \quad \phi_i(a) = \begin{cases} 1/2 - (\Delta - L||a - c_i||_\infty)_+ & a \in H_i \\ 1/2 - (\Delta/2 - ||a - c_0||_\infty)_+ & a \in S \\ 1/2 & \text{else} \end{cases}$$

Observe that ϕ_0 is 1-Lipschitz, and each ϕ_i is L-Lipschitz for $i \geq 1$.

Each mean loss function ϕ_i defines an environment where realized losses are Bernoulli random variables. Denote by $\mathbb{E}_i$ (resp. $\mathbb{P}_i$) the expectation (resp. probability) over the randomness of the algorithm, along with the randomness in environment ϕ_i.

As Alg guarantees $\texttt{Regret}(T, 0) \leq R_{\text{Lip}}(1, T)$ against all loss functions in $\Sigma(1)$, we have

$$\mathbb{E}_0 \sum_{t=1}^{T} \phi_0(a_t) - T\left({}^1\!/\!{}_2 - {}^\Delta\!/\!{}_2 \right) \leq R_{\text{Lip}}(1, T).$$

Denote by $T_i = \sum_{t=1}^{T} \mathbf{1}\left\{ a_t \in H_i \right\}$. Observe that the instantaneous regret for playing in any H_i is at least ${}^\Delta\!/\!{}_2$. Therefore, by pigeonhole principle, there exists at least one i such that

$$\mathbb{E}_0[T_i] \leq \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_0[T_j] = \frac{1}{N} \mathbb{E}_0 \sum_{j=1}^{N} T_j \leq \frac{2R_{\text{Lip}}(1, T)}{N\Delta}. \tag{32}$$

Following the exact same calculation as in the proof of Lemma 32 we get that $\mathbb{E}_i[T_i] \leq 0.8T$, which implies that on instance ϕ_i

$$\texttt{Regret}(T) \geq \mathbb{E}_i \sum_{t=1}^{T} \phi_i(a_t) - \left({}^1\!/\!{}_2 - \Delta \right) \geq 0.2T \cdot \frac{\Delta}{2} \geq \min\left\{ \frac{T}{80}, \frac{L^{\frac{d}{d+1}}}{3200 R_{\text{Lip}}(1, T)^{\frac{1}{d+1}}} \right\}. \quad \blacksquare$$

9.2. Proofs of the lower bounds

Proof of the lower bound in Theorem 17 We show that the lower bound statement holds for $T_0 = 2^{3d(1+\beta)}$ and $c = \frac{1}{80 \cdot 2^{d(\beta+3)}}$.

Fix $T \geq T_0$; let $h_1 = \frac{1}{4}$ and $h_2 = T^{\frac{-1}{d(\beta+1)}} \in (0, \frac{1}{8}]$. In addition, let $f(T, h) = c \cdot T^{\frac{1}{1+\beta}} h^{-d\beta} = \frac{T^{\frac{1}{1+\beta}} h^{-d\beta}}{80 \cdot 2^{d(\beta+3)}}$.

To finish the proof, we claim that for any algorithm Alg, one of the following must hold:

1. There exists some instance $\lambda \in \Lambda$, under which $\texttt{Regret}(T, \Pi_{h_1}) \geq f(T, h_1) = \frac{4^{d\beta} T^{\frac{1}{1+\beta}}}{80 \cdot 2^{d(\beta+3)}}$;
2. There exists some instance $\lambda \in \Lambda$, under which $\texttt{Regret}(T, \Pi_{h_2}) \geq f(T, h_2) = \frac{T}{80 \cdot 2^{d(\beta+3)}}$.

Indeed, suppose Alg is such that for all instances λ, $\texttt{Regret}(T, \Pi_{h_1}) < f(T, h_1)$. By our choice of h_2 and $T \geq T_0$, $f(T, h_1) \leq \frac{\sqrt{T}}{20 \cdot (8h_2)^d}$. Provided that this is satisfied, Lemma 32 with $R_{\text{S}}({}^1\!/\!{}_4, T) = f(T, h_1)$ gives that there is an instance λ', under which

$$\texttt{Regret}(T, \Pi_{K_2}) \geq \min\left\{ \frac{T}{40 \cdot 2^d}, \frac{1}{400 \cdot 8^d} T^{\frac{\beta}{1+\beta}} h_2^{-d} \right\} = \min\left\{ \frac{1}{40 \cdot 2^d}, \frac{1}{5 \cdot 2^{d\beta}} \right\} T > f(T, h_2),$$

proving the above claim. ■

Proof of the lower bound in Theorem 21 We show that the lower bound statement holds for $T_0 = 1$ and $c = \frac{1}{3200}$.

Fix $T \geq T_0$; We take $L_1 = 1$ and $L_2 = T^{\frac{1+d\beta}{d(1+(d+1)\beta)}}$. In addition, we let $g(T, L) = c \cdot T^{\frac{1+d\beta}{1+(d+1)\beta}} L^{\frac{d\beta}{1+d\beta}} = \frac{1}{3200} \cdot T^{\frac{1+d\beta}{1+(d+1)\beta}} L^{\frac{d\beta}{1+d\beta}}$.

To finish the proof, we claim that for any algorithm Alg, one of the following must hold:

1. There exists some instance $\lambda \in \Lambda(L_1)$, under which $\texttt{Regret}(T, \Pi) \geq g(T, L_1) = \frac{1}{3200} \cdot T^{\frac{1+d\beta}{1+(d+1)\beta}}$;

2. There exists some instance $\lambda \in \Lambda(L_2)$, under which $\texttt{Regret}(T, \Pi) \geq g(T, L_2) = \frac{T}{3200}$.

Indeed, suppose Alg is such that for all instances $\lambda \in \Lambda(L_1)$, $\texttt{Regret}(T, \Pi) < g(T, L_1)$. By our choice of L_2 and $T \geq T_0$, $g(T, L_1) \leq \frac{1}{40} L_2^d T$. Provided this is satisfied, Lemma 34 with $R_{\text{Lip}}(1, T) = g(T, L_1)$ gives that there is an instance $\lambda' \in \Lambda(L_2)$, under which

$$\texttt{Regret}(T, \Pi) \geq \min \left\{ \frac{T}{80}, \frac{1}{3200^{\frac{d}{d+1}}} \cdot T^{1 - \frac{1+d\beta}{(d+1)(1+(d+1)\beta)}} L_2^{\frac{d}{d+1}} \right\} > \frac{1}{3200} T = g(T, L_2),$$

proving the above claim. ■

10. Calculations for the examples

Calculation for Example 3. Straightforward computations reveal that (1) $\lambda_h(a^\star) = h/2$, (2) $\forall a \in [a^\star - h, a^\star + h]$ $\lambda_h(a) \leq \lambda_h(a^\star) + h/2$, and (3) $\forall a \notin [a^\star - h, a^\star + h]$, $\lambda_h(a) \geq \lambda_h(a^\star) + |a - a^\star|/2$. In particular, the third item follows from a Taylor expansion, since $\partial \lambda_h(a)/\partial a \geq 1/2$ for $a \geq a^\star + h$ (with a similar property for $a \leq a^\star - h$).

Therefore, if $\epsilon \leq h/2$, we have $\Pi_h(\epsilon) \subset \Pi_h(h/2) \subset [a^\star - h, a^\star + h]$, which implies that $M_h(\epsilon, h) \leq 1$. On the other hand, if $\epsilon > h/2$, we have $\Pi_h(\epsilon) \subset [a^\star - 2\epsilon, a^\star + 2\epsilon]$, implying that $M_h(\epsilon, h) \leq 4\epsilon/h$. Together we have that $M_h(\epsilon, h) \leq O(\max\{1, \epsilon/h\})$, and plugging into the definition of $\theta_h(\cdot)$ concludes the proof.

Calculation for Example 4. First, for all x and all w in $\mathbf{S}^{d-1}$, $\mathbb{E}[\ell(\pi_{w^\star}(x))|x] = f(0) \leq f(\langle w, x \rangle - \langle w^\star, x \rangle) = \mathbb{E}[\ell(\pi_w(x))|x]$, which implies that $\pi_{w^\star}$ is the optimal policy. Next, consider the expected regret of any policy π_w in Π. Using the properties of f, we have

$$\begin{aligned} \mathbb{E}[\ell(\pi_w(x))] - \mathbb{E}[\ell(\pi_{w^\star}(x))] &\geq L_0 \, \mathbb{E}[| \langle w^\star, x \rangle - \langle w, x \rangle |] \\ &= L_0 \|w^\star - w\|_2 \, \mathbb{E}[|x_1|] \geq \Omega(L_0/\sqrt{d}) \cdot \|w^\star - w\|_2. \end{aligned}$$

The equality follows from spherical symmetry, while the last inequality follows since the probability density function of x_1 is $p(x_1) = \frac{(1-x_1^2)^{\frac{d-3}{2}}}{\mathrm{B}(\frac{d-1}{2}, \frac{1}{2})}$ and thus $\mathbb{P}(|x_1| \geq \frac{1}{\sqrt{d}}) = \Omega(1)$.

This latter inequality implies that, for any $\pi_w \in \Pi_{0, L\epsilon}$, we have $\| w - w^\star \|_2 \leq O(L/L_0 \cdot \sqrt{d}\epsilon)$. Therefore, for any x we have

$$\Pi_{0, L\epsilon}(x) \subset \left[\langle w^\star, x \rangle - O(L/L_0 \cdot \sqrt{d}\epsilon), \langle w^\star, x \rangle + O(L/L_0 \cdot \sqrt{d}\epsilon) \right].$$

This implies that $M_0(L\epsilon, \epsilon) = \mathbb{E}_{x \sim \mathcal{D}_X} \left[\mathcal{N}_\epsilon(\Pi_{0, L\epsilon}(x)) \right] \leq O(L/L_0 \cdot \sqrt{d})$. This immediately implies that $\psi_L(\epsilon) = O(\frac{L}{L_0 \epsilon} \cdot \sqrt{d})$. Instantiating Theorem 18 yields the regret bound.

11. Conclusions

The main conceptual contribution of our paper is a new smoothing-based notion of regret that admits guarantees with no assumptions on the loss. Using this, we design new algorithms providing instance-dependent guarantees with optimal worst-case performance and Pareto-optimal adaptivity. This also yields new guarantees for non-contextual and Lipschitz bandits.

While our algorithms are computationally efficient in the low-dimensional non-contextual setting, they are not tractable in general since they require enumerating the policy set. Hence, the key open question is: Are there algorithms with similar statistical performance *and* fast running time?

Acknowledgments

We thank Wen Sun for insightful conversations during the initial stages of this work. Part of this work was done while CZ was a postdoc at Microsoft Research New York City.

References

Ittai Abraham, Shiri Chechik, David Kempe, and Aleksandrs Slivkins. Low-distortion inference of latent similarities from a multiplex social network. *SIAM Journal on Computing*, 2015. Expanded and revised version of the conference paper in *SODA 2013*.

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

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2017a.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 2017b.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Conference on Learning Theory*, 2007.

Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. No-regret bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 2019.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 2011a.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the lipschitz constant. In *Algorithmic Learning Theory*, 2011b.

Adam Bull. Adaptive-treed bandits. *Bernoulli*, 2015.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, 2017.

Hubert T-H. Chan, Kedar Dhamdhere, Anupam Gupta, Jon Kleinberg, and Aleksandrs Slivkins. Metric embeddings with relaxed guarantees. *SIAM Journal on Computing*, 2009. Preliminary version in *FOCS 2005*, merged with an independent effort by another research group.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 2016.

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

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

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, 2015.

Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low–distortion embeddings. In *Symposium on Foundations of Computer Science*, 2003.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 2016.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, 2018.

David Kempe and Jon Kleinberg. Protocols and impossibility results for gossip-based communication mechanisms. In *Symposium on Foundations of Computer Science*, 2002.

David Kempe, Jon Kleinberg, and Alan Demers. Spatial gossip and resource location protocols. *Journal of the ACM*, 2005. Preliminary version in *STOC 2001*.

Jon Kleinberg. The small-world phenomenon: an algorithmic perspective. In *Symposium on Theory of Computing*, 2000.

Jon Kleinberg, Aleksandrs Slivkins, and Tom Wexler. Triangulation and embedding using small sets of beacons. *Journal of the ACM*, 2009. Subsumes conference papers in *FOCS 2004* and *SODA 2005*.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, 2004.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Symposium on Foundations of Computer Science*, 2003.

Robert Kleinberg and Aleksandrs Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Symposium on Discrete Algorithms*, 2010.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Symposium on Theory of Computing*, 2008.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM*, 2019. Merged and revised version of conference papers in *STOC 2008* and *SODA 2010*. Also available at http://arxiv.org/abs/1312.1277.

Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, 2011.

Robert Krauthgamer and James R. Lee. Navigating nets: simple algorithms for proximity search. In *Symposium on Discrete Algorithms*, 2004.

Robert Krauthgamer, James Lee, Manor Mendel, and Assaf Naor. Measured descent: A new embedding method for finite metrics. *Geometric and Functional Analysis*, 2005. Preliminary version in *FOCS*, 2004.

Akshay Krishnamurthy, Alekh Agarwal, and Miro Dudik. Contextual semibandits via supervised learning oracles. In *Advances In Neural Information Processing Systems*, 2016.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. Versions available at `https://banditalgs.com/` since 2018.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, 2018.

Tyler Lu, Dávid Pál, and Martin Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Jouni Luukkainen and Eero Saksman. Every complete doubling metric space carries a doubling measure. *Proceedings of the American Mathematical Society*, 1998.

Manor Mendel and Sariel Har-Peled. Fast construction of nets in low dimensional metrics, and their applications. In *Symposium on Computational Geometry*, 2005.

Stanislav Minsker. Estimation of extreme values and associated level sets of a regression function via selective sampling. In *Conference on Learning Theory*, 2013.

Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W. Moore. Theoretical justification of popular link prediction heuristics. In *Conference on Learning Theory*, 2010.

Rajat Sen, Karthikeyan Shanmugam, and Sanjay Shakkottai. Contextual bandits with stochastic experts. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Xuedong Shang, Emilie Kaufmann, and Michal Valko. General parallel optimization a without metric. In *Algorithmic Learning Theory*, 2019.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 1958.

Aleksandrs Slivkins. *Embedding, Distance Estimation and Object Location in Networks*. PhD thesis, Cornell University, 2006. Available online at `http://research.microsoft.com/en-us/people/slivkins`.

Aleksandrs Slivkins. Distance estimation and object location via rings of neighbors. *Distributed Computing*, 2007. Special issue for *PODC 2005*. Preliminary version in *PODC 2005*.

Aleksandrs Slivkins. Multi-armed bandits on implicit metric spaces. In *Advances in Neural Information Processing Systems*, 2011.

Aleksandrs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 2014. Preliminary version in *COLT 2011*.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 2019. Also available at `https://arxiv.org/abs/1904.07272`.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 2012.

Kunal Talwar. Bypassing the embedding: Algorithms for low-dimensional metrics. In *Symposium on Theory of Computing*, 2004.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, 2013.

A. L. Volberg and S. V. Konyagin. On measures with the doubling condition. *Izvestiya Akademii Nauk SSSR*, 1987. In Russian; English translation in Mathematics of the USSR-Izvestiya, 1988.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, 2020.

Bernard Wong, Aleksandrs Slivkins, and Emin Gün Sirer. Meridian: A lightweight network location service without virtual coordinates. In *SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 2005. Full version is available at `http://research.microsoft.com/en-us/people/slivkins`.

Jang-Mei Wu. Hausdorff dimension and doubling measures on metric spaces. *Proceedings of the American Mathematical Society*, 1998.