Proceedings of Machine Learning Research vol 65:1–17, 2017

Algorithmic Chaining and the Role of Partial Feedback in Online Nonparametric Learning

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano, Milano, Italy

Pierre Gaillard PIERRE.GAILLARD@INRIA.FR
INRIA - Sierra Project-team, Département d'Informatique de l'École Normale Supérieure, Paris, France

Claudio Gentile CLAUDIO.GENTILE@UNINSUBRIA.IT
Università degli Studi dell'Insubria, Varese, Italy

Sébastien Gerchinovitz SEBASTIEN.GERCHINOVITZ@MATH.UNIV-TOULOUSE.FR
Université Toulouse III - Paul Sabatier, Toulouse, France

Abstract

We investigate contextual online learning with nonparametric (Lipschitz) comparison classes under different assumptions on losses and feedback information. For full information feedback and Lipschitz losses, we design the first explicit algorithm achieving the minimax regret rate (up to log factors). In a partial feedback model motivated by second-price auctions, we obtain algorithms for Lipschitz and semi-Lipschitz losses with regret bounds improving on the known bounds for standard bandit feedback. Our analysis combines novel results for contextual second-price auctions with a novel algorithmic approach based on chaining. When the context space is Euclidean, our chaining approach is efficient and delivers an even better regret bound.

Keywords: online learning, nonparametric, chaining, bandits.

1. Introduction

In online learning (Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2011; Hazan, 2015) an agent (or learner) interacts with an unknown and arbitrary environment in a sequence of rounds. At each round, the learner chooses an action from a given action space and incurs the loss associated with the chosen action. The loss functions, which are different in each round, are fixed by the environment at the beginning of the interaction. After choosing an action, the learner observes some feedback, which can be used to reduce his loss in subsequent rounds. A variety of different feedback models are discussed in the literature. The most common feedback model is full information, also known as prediction with expert advice, where the learner gets access to the entire loss function at the end of each round. Another common feedback model is bandit information, where the learner just observes the loss assigned to the action chosen in the current round. Feedback models in between full and bandit information are also possible, and can be used to describe many interesting online learning applications —see e.g., (Alon et al., 2014, 2015). The performance of an online learner is measured using a notion of regret, which is typically defined as the amount by which the learner's cumulative loss exceeds the cumulative loss of the best fixed action in hindsight.

Extended abstract. Full version appears as ArXiv:1702.08211,v2.

Online contextual learning is a generalization of online learning where the loss functions generated by the environment are paired with contexts from a given context space. On each round, before choosing an action, the learner observes the current context. In the presence of contextual information, the learner's regret is no longer defined against the best action in hindsight, but rather against the best policy (i.e., mapping from the context space to the action space) in a given reference class of policies. In agreement with the online learning framework, online contextual learning is nonstochastic. Namely, regret bounds must hold for arbitrary sequences of contexts and losses.

In order to capture complex environments, the reference class of policies should be as large as possible. In this work, we focus on nonparametric classes of policies, such as classes containing policies that are Lipschitz with respect to metrics defined on the context and action spaces. The best possible (minimax) growth rate of the regret, as a function of the number T of rounds, is then determined by the interplay among the richness of the policy class, the constraints on the loss functions (e.g., Lipschitz, convex, etc.), and the type of feedback information (full, bandit, or in between). Whereas most of the previous works study online nonparametric learning with convex losses, in this paper we investigate nonparametric regret rates for general Lipschitz losses (in fact, some of our results apply to an even larger class of loss functions).

In the full information setting, a very general yet simple algorithmic approach to online nonparametric learning with convex and Lipschitz losses was introduced by Hazan and Megiddo (2007). For any reference class of Lipschitz policies, they proved a $\widetilde{\mathcal{O}}\big(T^{(d+1)/(d+2)}\big)$ upper bound[1] on the regret for any context space of metric dimension d, where the $\widetilde{\mathcal{O}}$ notation hides logarithmic factors in T. In the same work, they also proved a $\Omega\big(T^{(d-1)/d}\big)$ lower bound. The gap between the upper and lower bound was closed by Rakhlin et al. (2015) for arbitrary Lipschitz (not necessarily convex) losses, showing that $T^{(d-1)/d}$ is indeed the minimax rate for full information. Yet, since their approach is nonconstructive, they did not give an explicit algorithm achieving this bound.

As noted elsewhere —see, e.g., (Slivkins, 2014)— the approach of Hazan and Megiddo (2007) can be also adapted to prove a $\widetilde{\mathcal{O}}\big(T^{(d+p+1)/(d+p+2)}\big)$ upper bound on the regret against any class of Lipschitz policies in the bandit information setting with Lipschitz losses, where p is the metric dimension of the action space. The lower bound $\Omega\big(T^{(p+1)/(p+2)}\big)$ proven for $d = 0$ (Bubeck et al., 2011a; Kleinberg et al., 2008) rules out the possibility of improving the dependence on p in the upper bound.

Our contributions. In the full information model, we show the first explicit algorithm achieving the minimax regret rate $\widetilde{\mathcal{O}}\big(T^{(d-1)/d}\big)$ for Lipschitz policies and Lipschitz losses (excluding logarithmic factors in T and polynomial factors in the metric dimension of the action space). When the context space is $[0,1]^d$, our algorithm can be implemented efficiently (i.e., with a running time polynomial in T).

Motivated by a problem in online advertising where the action space is the $[0,1]$ interval, we also study a "one-sided" full information model in which the loss of each action greater than or equal to the chosen action is available to the learner after each round. For this feedback model, which lies between full and bandit information, we prove a regret bound for Lipschitz policies and Lipschitz losses of order $\widetilde{\mathcal{O}}\big(T^{d/(d+1)}\big)$, which is larger than the minimax regret for full information but smaller than the upper bound for bandit information when $p = 1$. For the special case when the

1. This bound has a polynomial dependence on the metric dimension of the action space, which is absorbed by the asymptotic notation.

context space is $[0,1]^d$, we use a specialized approach offering the double advantage of an improved $\widetilde{\mathcal{O}}\big(T^{(d-1/3)/(d+2/3)}\big)$ regret bound which is also attained by a time-efficient algorithm.

We then study a concrete application for minimizing the seller's regret in contextual second-price auctions with reserve price, a setting where the loss function is not Lipschitz but only semi-Lipschitz. When the feedback after each auction is the seller's revenue together with the highest bid for the current auction, we prove a $\widetilde{\mathcal{O}}\big(T^{(d+1)/(d+2)}\big)$ regret bound against Lipschitz policies (in this setting, a policy maps contexts to reserve prices for the seller). As a by-product, we show the first $\widetilde{\mathcal{O}}\big(\sqrt{T}\big)$ regret bound on the seller's revenue in context-free second-price auctions under the same feedback model as above. Table 1 summarizes our results.

Feedback model	**Loss functions**	**Upper bound**	
Bandit	Lipschitz	$T^{\frac{d+2}{d+3}}$	(Theorem 1)
	Convex	$T^{\frac{d+1}{d+2}}$	(Corollary 2)
One-sided full information	Semi-Lipschitz	$T^{\frac{d+1}{d+2}}$	(Theorem 3)
	Lipschitz	$T^{\frac{d-1/3}{d+2/3}}$	(Theorem 6)
Full information	Lipschitz	$T^{\frac{d-1}{d}}$	(Theorem 7)

Table 1: Some regret bounds obtained in this paper. The rates are up to logarithmic factors for Lipschitz policies $f : [0,1]^d \to [0,1]$ with $d \geqslant 2$. All upper bounds are constructive (i.e., achieved by explict algorithms). The only matching lower bound is the one for full information feedback due to Hazan and Megiddo (2007).

In order to prove our results, we approximate the action space using a finite covering (finite coverability is a necessary condition for our results to hold). This allows us to use the many existing algorithms for experts (full information feedback) and bandits when the action space is finite, such as Hedge (Freund and Schapire, 1997) and Exp3/Exp4 (Auer et al., 2002). The simplest of our algorithms, adapted from Hazan and Megiddo (2007), incrementally covers the context space with balls of fixed radius. Each ball hosts an instance of an online learning algorithm which predicts in all rounds when the context falls into the ball. New balls are adaptively created when new contexts are observed which fall outside the existing balls (see Algorithm 1 for an example). We use this simple construction to prove the regret bound for contextual second-price auctions, a setting where losses are not Lipschitz. In order to exploit the additional structure provided by Lipschitz losses, we resort to more sophisticated constructions based on chaining (Dudley, 1967). In particular, inspired by previous works in this area (especially the work of Gaillard and Gerchinovitz, 2015), we design a chaining-inspired algorithm applied to a hierarchical covering of the policy space. Despite we are not the first ones to use chaining algorithmically in online learning, our idea of constructing a hierarchy of online learners, where each node uses its children as experts, is novel in this context as far as we know. Finally, the time-efficient algorithm achieving the improved regret bound is derived from a different (and more involved) chaining algorithm based on wavelet-like approximation techniques.

Setting and main definitions. We assume the context space $\mathcal{X}$ is a metric space $(\mathcal{X}, \rho_{\mathcal{X}})$ of finite metric dimension d and the action space $\mathcal{Y}$ is a metric space $(\mathcal{Y}, \rho_{\mathcal{Y}})$ of finite metric dimension p.

Hence, there exist $C_{\mathcal{X}}, C_{\mathcal{Y}} > 0$ such that, for all $0 < \varepsilon \leqslant 1$, $\mathcal{X}$ and $\mathcal{Y}$ can be covered, respectively, with at most $C_{\mathcal{X}}\varepsilon^{-d}$ and at most $C_{\mathcal{Y}}\varepsilon^{-p}$ balls of radius ε. For any $0 < \varepsilon \leqslant 1$, we use $\mathcal{Y}_\varepsilon$ to denote any ε-covering of $\mathcal{Y}$ of size $K_\varepsilon \leqslant C_{\mathcal{Y}}\varepsilon^{-p}$. Finally, we assume that $\mathcal{Y}$ has diameter bounded by 1 with respect to metric $\rho_{\mathcal{Y}}$.

We consider the following online learning protocol with oblivious adversary and loss functions $\ell_t : \mathcal{Y} \to [0,1]$. Given an unknown sequence $(x_1, \ell_1), (x_2, \ell_2), \dots$ of contexts $x_t \in \mathcal{X}$ and loss functions $\ell_t : \mathcal{Y} \to [0,1]$, for every round $t = 1, 2, \dots$:

1. The environment reveals context $x_t \in \mathcal{X}$;
2. The learner selects an action $\widehat{y}_t \in \mathcal{Y}$ and incurs loss $\ell_t(\widehat{y}_t)$;
3. The learner obtains feedback from the environment.

Loss functions ℓ_t satisfy the 1-Lipschitz[2] condition $\left|\ell_t(y) - \ell_t(y')\right| \leqslant \rho_{\mathcal{Y}}(y, y')$ for all $y, y' \in \mathcal{Y}$. However, we occasionally consider losses satisfying a weaker semi-Lipschitz condition.

We study three different types of feedback: bandit feedback (the learner only observes the loss $\ell_t(\widehat{y}_t)$ of the selected action $\widehat{y}_t$), full information feedback (the learner can compute $\ell_t(y)$ for any $y \in \mathcal{Y}$), and one-sided full information feedback ($\mathcal{Y} \equiv [0,1]$, and the learner can compute $\ell_t(y)$ if and only if $y \geqslant \widehat{y}_t$). Given a reference class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ of policies, the learner's goal is to minimize the regret against the best policy in the class,

$$\mathrm{Reg}_T(\mathcal{F}) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\widehat{y}_t)\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_t\big(f(x_t)\big)\,,$$

where the expectation is with respect to the learner's internal randomization. We derive regret bounds for the competitor class $\mathcal{F}$ made up of all bounded functions $f : \mathcal{X} \to \mathcal{Y}$ that are 1-Lipschitz[3] w.r.t. $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$. Namely, $\rho_{\mathcal{Y}}\big(f(x), f(x')\big) \leqslant \rho_{\mathcal{X}}(x, x')$ for all $f \in \mathcal{F}$ and all $x, x' \in \mathcal{X}$. We occasionally use the dot product notation $p_t \cdot \ell_t$ to indicate the expectation of ℓ_t according to law p_t. Finally, the set of all probability distributions over a finite set of K elements is denoted by $\Delta(K)$.

Organization of the paper. The rest of the paper is organized as follows. In Section 2, we give an overview of the related literature. Then, starting from the subsequent section, our results are presented in the order dictated by the amount of feedback available to the learner, from bandit feedback (Section 3) to one-sided full information feedback (Section 4) to full information feedback (Section 5).

2. Related Work

Contextual online learning generalizes online convex optimization (Hazan, 2015) to nonconvex losses, nonparametric policies, and partial feedback. Papers about nonparametric online learning in full information include (Vovk, 2007; Gaillard and Gerchinovitz, 2015) for the square loss, and (Hazan and Megiddo, 2007; Rakhlin and Sridharan, 2015) for general convex losses. In the bandit feedback model, earlier work on context-free bandits on metric spaces includes (Kleinberg, 2004;

2. Assuming a unit Lipschitz constant is without loss of generality because our algorithms are oblivious to it.
3. Almost all our algorithms are oblivious to the Lipschitz constant L_f of f and yield similar regret bounds (as a function of T) whatever L_f. Our algorithm HierExp4* of Section 4.3 is only guaranteed to work when $L_f = 1$, but a similar regret bound can be achieved for an arbitrary known L_f via a simple modification. See also (Bubeck et al., 2011b) when L_f is unknown (with regret bounds optimized in L_f) in a stochastic and context-free bandit setting.

Kleinberg et al., 2008). The paper (Auer et al., 2002) introduces the Exp4 algorithm for nonstochastic contextual bandits when both the action space and the policy space are finite, and policies are maps from contexts to distributions over actions. Moreover, rather than observing the current context, the learner sees the output of each policy for that context. In the contextual bandit model of Maillard and Munos (2011), context space and action space are finite, and the learner observes the current context while competing against the best policy among all functions mapping contexts to actions. Finally, a nonparametric bandit setting related to ours was studied by Slivkins (2014). We refer the reader to the discussion after Theorem 1 for connections with our results.

Chaining (Dudley, 1967) is a powerful technique to obtain tail bounds on the suprema of stochastic processes. In nonparametric online learning with full information feedback, chaining was used constructively by Cesa-Bianchi and Lugosi (1999) to design an algorithm for linear losses, and nonconstructively by Rakhlin et al. (2015) to derive minimax rates for Lipschitz losses. Other notable examples of chaining are the stochastic bandit algorithms of Contal et al. (2015) and Contal and Vayatis (2016). The constructive algorithmic chaining technique developed in this work is inspired by the nonparametric analysis of the full information setting of Gaillard and Gerchinovitz (2015). However, their multi-variable EG algorithm heavily relies on convexity of losses and requires access to loss gradients. In order to cope with nonconvex losses and lack of gradient information, we develop a novel chaining approach based on a tree of hierarchical coverings of the policy class, where each internal tree node hosts a bandit algorithm.

In our nonstochastic online setting, chaining yields improved rates when the regret is decomposed into a sum of local regrets, each one scaling with the range of the local losses. However, deriving regret bounds that scale with the effective range of the losses is not always possible, as shown by Gerchinovitz and Lattimore (2016) in the nonstochastic K-armed bandit setting. This result suggests that chaining might not be useful in online nonparametric learning when the feedback is bandit. However, as we show in this paper, algorithmic chaining does help improving the regret when the feedback is one-sided full information or full information. In full information, chaining-based algorithms deliver regret bounds that match (up to log factors) the nonconstructive bounds of (Rakhlin et al., 2015).

In a different but interesting research thread on contextual bandits, the learner is confronted with the best within a finite (but large) class of policies over finitely-many actions, and is assumed to have access to this policy class through an optimization oracle for the offline full information problem. Relevant references include (Agarwal et al., 2014; Rakhlin and Sridharan, 2016; Syrgkanis et al., 2016). The main concern is to devise (oracle-based) algorithms with small regret and requiring as few calls to the optimization oracle as possible.

3. Warmup: Nonparametric Bandits

As a simple warmup exercise, we prove a known result —see e.g., (Slivkins, 2014). Namely, a regret bound for contextual bandits with Lipschitz policies and Lipschitz losses. ContextualExp3 (Algorithm 1) is a bandit version of the algorithm by Hazan and Megiddo (2007) and maintains a set of balls of fixed radius ε in the context space, where each ball hosts an instance of the Exp3 algorithm of Auer et al. (2002).[4] At each round t, if a new incoming context $x_t \in \mathcal{X}$ is not contained

4. Instead of Exp3 we could use INF (Audibert and Bubeck, 2010), which enjoys a minimax optimal regret bound up to constant factors. This would avoid a polylog factor in T in the bound. Since we do not optimize for polylog factors anyway, we opted for the better known algorithm.

in any existing ball, then a new ball centered at x_t is created, and a fresh instance of Exp3 is allocated to handle x_t. Otherwise, the Exp3 instance associated with the closest context so far w.r.t. $\rho_{\mathcal{X}}$ is used to handle x_t. Each allocated Exp3 instance operates on the discretized action space $\mathcal{Y}_\varepsilon$ whose size K_ε is at most $C_{\mathcal{Y}}\,\varepsilon^{-p}$. The proof of the following theorem is provided in (Cesa-Bianchi et al., 2017).

Algorithm 1: ContextualExp3 (for bandit feedback)

Input: Ball radius $\varepsilon > 0$, ε-covering $\mathcal{Y}_\varepsilon$ of $\mathcal{Y}$ such that $|\mathcal{Y}_\varepsilon| \le C_{\mathcal{Y}}\,\varepsilon^{-p}$.
for $t = 1, 2, \dots$ **do**
1. Get context $x_t \in \mathcal{X}$;
2. If x_t does not belong to any existing ball, then create a new ball of radius ε centered on x_t, and allocate a fresh instance of Exp3;
3. Let the active Exp3 instance be the instance allocated to the existing ball whose center x_s is closest to x_t;
4. Draw an action $\widehat{y}_t$ using the active Exp3 instance;
5. Get $\ell_t(\widehat{y}_t)$ and use it to update the active Exp3 instance.

end

Theorem 1 *Fix any any sequence $(x_1, \ell_1), (x_2, \ell_2), \dots$ of contexts $x_t \in \mathcal{X}$ and 1-Lipschitz loss functions $\ell_t : \mathcal{Y} \to [0,1]$. If ContextualExp3 is run in the bandit feedback model with parameter*[5] *$\varepsilon = (\ln T)^{\frac{2}{p+d+2}}\, T^{-\frac{1}{p+d+2}}$, then its regret $\mathrm{Reg}_T(\mathcal{F})$ with respect to the set $\mathcal{F}$ of 1-Lipschitz functions $f : \mathcal{X} \to \mathcal{Y}$ satisfies $\mathrm{Reg}_T(\mathcal{F}) = \widetilde{\mathcal{O}}\big(T^{\frac{p+d+1}{p+d+2}}\big)$, where the $\widetilde{\mathcal{O}}$ notation hides factors polynomial in $C_{\mathcal{X}}$ and $C_{\mathcal{Y}}$, and $\ln T$ factors.*

A lower bound matching up to log factors the upper bound of Theorem 1 is contained in (Slivkins, 2014) —see also (Lu et al., 2010) for earlier results in the same setting. However, our setting and his are subtly different: the adversary of Slivkins (2014) uses more general Lipschitz losses which, translated into our context, imply that the Lipschitz assumption is required to hold only for the composite function $\ell_t(f(\cdot))$, rather than the two functions ℓ_t and f separately. Hence, being the adversary less constrained (and the comparison class wider), the lower bound contained in (Slivkins, 2014) does not seem to apply to our setting.

While we are unaware of a lower bound matching the upper bound in Theorem 1 when $\mathcal{F}$ is the class of (global) Lipschitz functions and $d \geqslant 1$, in the noncontextual case ($d = 0$), the lower bound $\Omega\big(T^{(p+1)/(p+2)}\big)$ proven by Bubeck et al. (2011a); Kleinberg et al. (2008) shows that improvements on the dependence on p are generally impossible. Yet, the dependence on p in the bound of Theorem 1 can be greatly improved in the special case when the Lipschitz losses are also convex. Assume $\mathcal{Y}$ is a convex and compact subset of $\mathbb{R}^p$. Then we use the same approach as in Theorem 1, where the Exp3 algorithm hosted at each ball is replaced by an instance of the algorithm by Bubeck et al. (2016), run on the non-discretized action space $\mathcal{Y}$. The regret of the algorithm that replaces Exp3 is bounded by $\mathrm{poly}(p, \ln T)\sqrt{T}$. This immediately gives the following corollary.

Corollary 2 *Fix any any sequence $(x_1, \ell_1), (x_2, \ell_2), \dots$ of contexts $x_t \in \mathcal{X}$ and convex loss functions $\ell_t : \mathcal{Y} \to [0,1]$, where $\mathcal{Y}$ is a convex and compact subset of $\mathbb{R}^p$. Then there exists an algorithm*

5. Here and throughout, T is assumed to be large enough so as to ensure $\varepsilon \leqslant 1$.

for the bandit feedback model whose regret with respect to the set $\mathcal{F}$ of 1-Lipschitz functions satisfies $\mathrm{Reg}_T(\mathcal{F}) \leqslant \mathrm{poly}(p, \ln T) T^{(d+1)/(d+2)}$, where poly *is a polynomial function of its arguments.*

4. One-Sided Full Information Feedback

In this section we show that better nonparametric rates can be achieved in the one-sided full information setting, where the feedback is larger than the standard bandit feedback but smaller than the full information feedback. More precisely, we consider the same setting as in Section 3 in the special case when the action space $\mathcal{Y}$ is $[0,1]$. We further assume that, after each play $\widehat{y}_t \in \mathcal{Y}$, the learner can compute the loss $\ell_t(y)$ of any number of actions $y \geqslant \widehat{y}_t$. This in contrast to observing only $\ell_t(\widehat{y}_t)$, as in the standard bandit setting. We start with an important special case: maximizing the seller's revenue in a sequence of repeated second-price auctions. In Section 4.2, we use the chaining technique to design a general algorithm for arbitrary Lipschitz losses in the one-sided full information model. An efficient variant of this algorithm is obtained using a more involved construction in Section 4.3.

4.1. Nonparametric second-price auctions

In online advertising, publishers sell their online ad space to advertisers through second-price auctions managed by ad exchanges. For each impression (ad display) created on the publisher's website, the ad exchange runs an auction on the fly. Empirical evidence (Ostrovsky and Schwarz, 2011) shows that an informed choice of the seller's reserve price, disqualifying any bid below it, can indeed have a significant impact on the revenue of the seller. Regret minimization in second-price auctions was studied by Cesa-Bianchi et al. (2015) in a non-contextual setting. They showed that, when buyers draw their bids i.i.d. from the same unknown distribution on $[0,1]$, there exists an efficient strategy for setting reserve prices such that the seller's regret is bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ with high probability with respect to the bid distribution. Here we extend those results to a nonstochastic and nonparametric contextual setting with nonstochastic bids, and prove a regret bound of order $T^{(d+1)/(d+2)}$ where d is the context space dimension. This improves on the bound $T^{(d+2)/(d+3)}$ of Theorem 1 when $p = 1$. As a byproduct, taking $d = 0$, this proves the first $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound for the seller in nonstochastic and noncontextual second-price auctions —see also (Cesa-Bianchi et al., 2017, Theorem 3). Unlike (Cesa-Bianchi et al., 2015), where the feedback after each auction was "strictly bandit" (i.e., just the seller's revenue), here we assume the seller is also observing the highest bid together with the revenue. This richer feedback, which is key to proving our results, is made available by some ad exchanges such as AppNexus.

The seller's revenue in a second-price auction is computed as follows: if the reserve price $\widehat{y}$ is not larger than the second-highest bid $b(2)$, then the item is sold to the highest bidder and the seller's revenue is $b(2)$. If $\widehat{y}$ is between $b(2)$ and the highest bid $b(1)$, then the item is sold to the highest bidder but the seller's revenue is the reserve price. Finally, if $\widehat{y}$ is bigger than $b(1)$, then the item is not sold and the seller's revenue is zero. Formally, the seller's revenue is $g\big(\widehat{y}, b(1), b(2)\big) = \max\big\{\widehat{y}, b(2)\big\}\mathbb{I}_{\widehat{y} \leq b(1)}$. Note that the revenue only depends on the reserve price $\widehat{y}$ and on the two highest bids $b(1) \geq b(2)$, which —by assumption— belong all to the unit interval $[0,1]$.

In the online contextual version of the problem, unknown sequences of contexts $x_1, x_2, \ldots \in \mathcal{X}$ and bids are fixed beforehand (in the case of online advertising, the context could be public information about the targeted customers). At the beginning of each auction $t = 1, 2, \ldots$, the seller observes context x_t and computes a reserve price $\widehat{y}_t \in [0,1]$. Then, bids $b_t(1), b_t(2)$ are collected

by the auctioneer, and the seller (which is not the same as the auctioneer) observes his revenue $g_t(\widehat{y}_t) = g(\widehat{y}_t, b_t(1), b_t(2))$, together with the highest bid $b_t(1)$. Crucially, knowing $g_t(\widehat{y}_t)$ and $b_t(1)$ allows to compute $g_t(y)$ for all $y \geqslant \widehat{y}_t$. For technical reasons, we use losses $\ell_t(\widehat{y}_t) = 1 - g_t(\widehat{y}_t)$ instead of revenues, see Figure 1 for a pictorial representation.

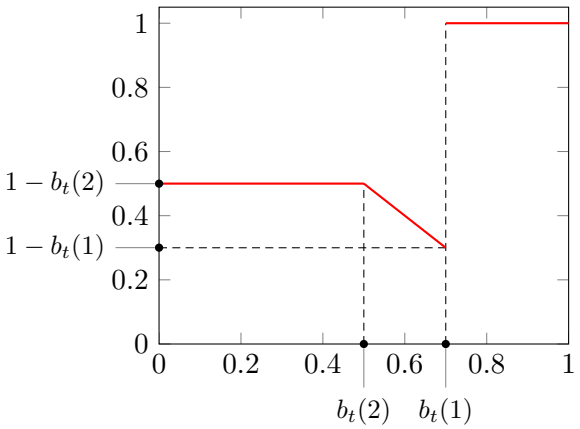

Figure 1: The loss function $\ell_t(\widehat{y}_t) = 1 - \max\{\widehat{y}_t, b_t(2)\}\mathbb{I}_{\widehat{y}_t \leq b_t(1)}$ when $b_t(1) = 0.7$ and $b_t(2) = 0.5$.

Remarkably, the loss functions ℓ_t are not Lipschitz and not even continuous, and so this problem falls outside the scope of standard results for contextual bandits. Instead, the losses only satisfy the semi-Lipschitz property $\ell_t(y + \delta) \geq \ell_t(y) - \delta$ for all $0 \leqslant y \leqslant y + \delta \leqslant 1$. We now state a bound on the regret $\mathrm{Reg}_T(\mathcal{F})$ with respect to any class $\mathcal{F}$ of Lipschitz functions $f : \mathcal{X} \to [0, 1]$. The algorithm that achieves this bound is ContextualRTB (where RTB stands for Real Time Bidding — see Algorithm Exp3-RTB in (Cesa-Bianchi et al., 2017)) , a variant of ContextualExp3 (Algorithm 1), where each ball hosts an instance of Exp3-RTB, instead of Exp3. The proof is given in the full version of the paper, (Cesa-Bianchi et al., 2017).

Theorem 3 *Fix any sequence of contexts $x_t \in \mathcal{X}$ and bid pairs $0 \leqslant b_t(2) \leqslant b_t(1) \leqslant 1$ for $t \geqslant 1$. If ContextualRTB is run with parameter $\varepsilon = T^{-\frac{1}{d+2}}$ and Exp3-RTB is tuned with parameter $\gamma = \varepsilon$, then the regret with respect to any class of 1-Lipschitz functions $f : \mathcal{X} \to [0, 1]$ satisfies $\mathrm{Reg}_T(\mathcal{F}) = \widetilde{\mathcal{O}}\big(T^{\frac{d+1}{d+2}}\big)$, where d is the dimension of $\mathcal{X}$ and the $\widetilde{\mathcal{O}}$ notation hides constants and $\ln T$ factors.*

ContextualRTB and ContextualExp3 of Section 3 can be modified so to avoid knowing the horizon T and so that the dimension d of the context space is replaced in the bound by the (unknown, and possibly much smaller) dimension of the set of contexts actually occurring in the sequence chosen by the adversary. This modification involves using a time-varying radius ε and a doubling trick to check when the current guess for the dimension is violated by the current number of balls. The omitted proof of this statement goes along the lines of the proof in (De Rosa et al., 2015, Theorem 1).

4.2. Chaining the bandits

We now show that whenever the richer feedback structure —i.e., the learner can compute the loss $\ell_t(y)$ of any number of actions $y \geqslant \widehat{y}_t$— is combined with Lipschitz losses (rather than just semi-

Lipschitz), then an improved regret bound $T^{d/(d+1)}$ can be derived. The key technical idea enabling this improvement is the application of the chaining technique to a hierarchical covering of the policy space (as opposed to the flat covering of the context space used in both Section 3 and Section 4.1). We start with a computationally inefficient algorithm that works for arbitrary policy classes $\mathcal{F}$ (not only Lipschitz) and is easier to understand. In Section 4.3 we derive an efficient variant for $\mathcal{F}$ that are Lipschitz. In this case we obtain even better regret bounds via a penalization trick.

A way of understanding the chaining approach is to view the hierarchical covering of the policy class $\mathcal{F}$ as a tree whose nodes are functions in $\mathcal{F}$, and where the nodes at each depth m define a (2^{-m})-covering of $\mathcal{F}$. The tree represents any function $f^* \in \mathcal{F}$ (e.g., the function with the smallest cumulative loss) by a unique path (or chain) $f_0 \to f_1 \to \cdots \to f_M \to f^*$, where f_0 is the root and f_M is the function best approximating f^* in the cover at the largest available depth M. By relying on this representation, we control the regret against any function in $\mathcal{F}$ by running an instance of an online bandit algorithm A on each node of the tree. The instance A_f at node f uses the predictions of the instances running on the nodes that are children of f as expert advice. The action drawn by instance A_0 running on the root node is the output of the tree. For any given sequence of pairs (x_t, ℓ_t) of contexts and losses, the regret against f^* with path $f_0 \to f_1 \to \cdots \to f_M \to f^*$ can then be written (ignoring some constants) as

$$\sum_{t=1}^{T} \Big(\mathbb{E}\left[\ell_t\big(A_0(x_t)\big)\right] - \ell_t\big(f^*(x_t)\big) \Big) \leqslant \sum_{m=0}^{M-1} \mathbb{E}\left[\sum_{t=1}^{T} \Big(\ell_t\big(A_m(x_t)\big) - \ell_t\big(A_{m+1}(x_t)\big) \Big) \right] + 2^{-M} T$$

where A_m is the instance running on node f_m for $m = 0, \dots, M-1$ and $A_M \equiv f_M$. The last term $2^{-M}T$ accounts for the cost of approximating f^* with the closest function f_M in a (2^{-M})-cover of $\mathcal{F}$ under suitable Lipschitz assumptions. The outer sum in the right-hand side of the above display can be viewed as a sum of M regrets, where the m-th term in the sum is the regret of A_m against the instances running on the children of the node hosting A_m. Since we face an expert learning problem in a partial information setting, the Exp4 algorithm of Auer et al. (2002) is a natural choice for the learner A. However, a first issue to consider is that we are using A_0 to draw actions in the bandit problem, and so the other Exp4 instances receive loss estimates that are based on the distribution used by A_0 rather than being based on their own distributions. A second issue is that our regret decomposition crucially relies on the fact that each instance A_m only competes (in the sense of regret) against functions f at the leaves of the subtree rooted at the node where A_m runs. By construction, these functions at the leaves are roughly (2^{-m})-close to each other and —by Lipschitzness— so are their losses. As a consequence, the regret of A_m should scale with the true loss range 2^{-m}. Via an appropriate modification of the original Exp4 algorithm, we manage to address both these issues. In particular, in order to make the regret dependent on the loss range, we heavily rely on the one-sided full information model assumed in this section. Finally, the hierarchical covering requires that losses be Lipschitz, rather than just semi-Lipschitz as in the application of Subsection 4.1, which uses a simpler flat covering.

Fix any class $\mathcal{F}$ of functions $f : \mathcal{X} \to [0,1]$. Let us introduce the sup norm

$$\|f - g\|_\infty = \sup_{x \in \mathcal{X}} \big| f(x) - g(x) \big| \,. \tag{1}$$

We denote by $\mathcal{N}_\infty\big(\mathcal{F}, \varepsilon\big)$ the cardinality of the smallest ε-cover of $\mathcal{F}$ w.r.t. the sup norm. Throughout this section, our only assumption on $\mathcal{F}$ is that $\mathcal{N}_\infty\big(\mathcal{F}, \varepsilon\big) < +\infty$ for all $\varepsilon > 0$ (this property is

known as *total boundedness*). In Section 4.3 we will focus on the special case $\mathcal{F} = \left\{f : [0,1]^d \to [0,1] : f \text{ is 1-Lipschitz}\right\}$ to derive an efficient version of our algorithm.

We now define a tree $\mathcal{T}_\mathcal{F}$ of depth M, whose nodes are labeled by functions in the class $\mathcal{F}$, so that functions corresponding to nodes with a close common ancestor are close to one another according to the sup norm (1). For all $m = 0, 1, \dots, M$, let $\mathcal{F}_m$ be a (2^{-m})-covering of $\mathcal{F}$ in sup norm with minimal cardinality $N_m = \mathcal{N}_\infty(\mathcal{F}, 2^{-m})$. Since the diameter of $(\mathcal{F}, \|\cdot\|_\infty)$ is bounded by 1, we have $N_0 = 1$ and $\mathcal{F}_0 = \{f_0\}$ for some $f_0 \in \mathcal{F}$. For each $m = 0, 1, \dots, M$ and for every $f_v \in \mathcal{F}_m$ we have a node v in $\mathcal{T}_\mathcal{F}$ at depth m. The parent of a node w at depth $m+1$ is some node v at depth m such that

$$v \in \operatorname*{arg\,min}_{v' \,:\, \mathrm{depth}(v') = m} \|f_{v'} - f_w\|_\infty \qquad \text{(ties broken arbitrarily)}$$

and we say that w is a child of v. Let $\mathcal{L}$ be the set of all the leaves of $\mathcal{T}_\mathcal{F}$, $\mathcal{L}_v$ be the set of all the leaves under $v \in \mathcal{T}_\mathcal{F}$ (i.e., the leaves of the subtree rooted at v), and $\mathcal{C}_v$ be the set of children of $v \in \mathcal{T}_\mathcal{F}$.

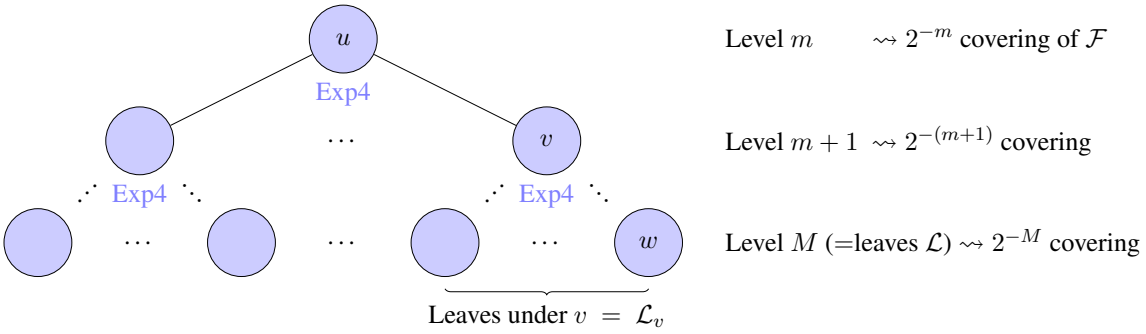

Figure 2: Hierarchical covering of the function space (used in Algorithm 2).

Our new bandit algorithm HierExp4 (Algorithm 2 below) is a hierarchical composition of instances of Exp4 on the tree $\mathcal{T}_\mathcal{F}$ constructed above (see Figure 2). Let $K = 2^M$ and $\mathcal{K} = \{y_1, \dots, y_K\}$, where $y_k = 2^{-M}(k-1)$ for $k = 1, \dots, 2^M$, be our discretization of the action space $\mathcal{Y} = [0,1]$. At every round t, after observing context $x_t \in \mathcal{X}$, each leaf $v \in \mathcal{L}$ recommends the best approximation of $f_v(x_t)$ in $\mathcal{K}$, $i_t(v) \in \arg\min_{i=1,\dots,K} \left|y_i - f_v(x_t)\right|$. Therefore, the leaves $v \in \mathcal{L}$ correspond to deterministic strategies $t \mapsto i_t(v)$, and we will find it convenient to view a set of leaves $\mathcal{L}$ as the set of actions played by those leaves at time t. Each internal node $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$ runs an instance of Exp4 using the children of v as experts. More precisely, we use a variant of Exp4 (see Cesa-Bianchi et al. (2017)) which adapts to the effective range of the losses. Let $\mathrm{Exp4}_v$ be the instance of the Exp4 variant run on node v. At each time t, this instance updates a distribution $q_t(v, \cdot) \in \Delta(|\mathcal{C}_v|)$ over experts in $\mathcal{C}_v$ and a distribution $p_t(v, \cdot) \in \Delta(K)$ over actions in $\mathcal{K}$ defined by $p_t(v,i) = \sum_{w \in \mathcal{C}_v} q_t(v,w) p_t(w,i)$.

Let v_0 be the root of $\mathcal{T}_\mathcal{F}$. The prediction of HierExp4 at time t is $\widehat{y}_t = y_{I_t} \in \mathcal{K}$, where I_t is drawn according to a mixture of $p_t(v_0, \cdot)$ and a unit mass on the minimal action $y_1 \in \mathcal{K}$.

For each $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$, let $\mathcal{K}_t(v) = \{i : (\exists w \in \mathcal{C}_v)\ p_t(w,i) > 0\}$ and $j_t(v) = \max \mathcal{K}_t(v)$. Note that $\widehat{\ell}_t(v,i)$ in (2) has to be explicitly computed only for those actions i such that $i \geqslant I_t$ and $i \in \mathcal{K}_t(v)$. This is because $\widehat{\ell}_t(v,i)$ is needed for the computation of $\widetilde{\ell}_t(v,w)$ only when $p_t(w,i) > 0$.

Therefore, whenever $\widehat{\ell}_t(v,i)$ has to be computed for some i, then $I_t \leqslant i \leqslant \max \mathcal{K}_t(v) = j_t(v)$, so that $\ell_t\big(y_{j_t(v)}\big)$ is observed and $\widehat{\ell}_t(v,i)$ is well defined.

Algorithm 2: HierExp4 (for one-sided full information feedback)

Input : Tree $\mathcal{T}_\mathcal{F}$ with root v_0 and leaves $\mathcal{L}$, exploration parameter $\gamma \in (0,1)$, learning rate sequences $\eta_1(v) \geqslant \eta_2(v) \geqslant \cdots > 0$ for $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$.

Initialization: Set $q_1(v,\cdot)$ to the uniform distribution in $\Delta(|\mathcal{C}_v|)$ for every $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$.

for $t = 1, 2, \ldots$ **do**

1. Get context $x_t \in \mathcal{X}$;
2. Set $p_t(v,i) = \mathbb{I}_{i=i_t(v)}$ for all $i \in \mathcal{K}$ and for all $v \in \mathcal{L}$;
3. Set $p_t(v,i) = q_t(v,\cdot) \cdot p_t(\cdot,i)$ for all $i \in \mathcal{K}$ and for all $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$;
4. Draw $I_t \sim p_t^*$ and play $\widehat{y}_t = y_{I_t}$, where $p_t^*(i) = (1-\gamma)p_t(v_0,i) + \gamma \mathbb{I}_{i=1}$ for all $i \in \mathcal{K}$;
5. Observe $\ell_t(y)$ for all $y \geqslant y_{I_t}$;
6. For every $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$ and for every $i \in \mathcal{K}_t(v)$ compute

$$\widehat{\ell}_t(v,i) = \frac{\ell_t(y_i) - \ell_t\big(y_{j_t(v)}\big)}{\sum_{k=1}^{i} p_t^*(k)} \mathbb{I}_{I_t \leqslant i}\,, \tag{2}$$

   where $\mathcal{K}_t(v) = \{i : (\exists w \in \mathcal{C}_v)\ p_t(w,i) > 0\}$ and $j_t(v) = \max \mathcal{K}_t(v)$.
7. For each $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$ and for each $w \in \mathcal{C}_v$ compute the expert loss $\widetilde{\ell}_t(v,w) = p_t(w,\cdot) \cdot \widehat{\ell}_t(v,\cdot)$ and perform the update

$$q_{t+1}(v,w) = \frac{\exp\Big(-\eta_{t+1}(v) \sum_{s=1}^{t} \widetilde{\ell}_s(v,w)\Big)}{\sum_{w' \in \mathcal{C}_v} \exp\Big(-\eta_{t+1}(v) \sum_{s=1}^{t} \widetilde{\ell}_s(v,w')\Big)} \tag{3}$$

end

Next, we show that the regret of HierExp4 is at most of the order of $T^{d/(d+1)}$, which improves on the rate $T^{(d+1)/(d+2)}$ obtained in Section 4.1 without using chaining. The required proofs are contained in the full version of the paper (Cesa-Bianchi et al., 2017).

Theorem 4 *Fix any class $\mathcal{F}$ of functions $f : \mathcal{X} \to [0,1]$ and any sequence $(x_1,\ell_1),(x_2,\ell_2),\ldots$ of contexts $x_t \in \mathcal{X}$ and 1-Lipschitz loss functions $\ell_t : [0,1] \to [0,1]$. Assume the HierExp4 (Algorithm 2) is run with one-sided full information feedback using tree $\mathcal{T}_\mathcal{F}$ of depth $M = \lfloor \ln_2(1/\gamma) \rfloor$. Moreover, the learning rate $\eta_t(v)$ used at each node v at depth $m = 0,\ldots,M-1$ is given by*

$$\eta_t(v) = \min\left\{\gamma 2^{m-4}, \sqrt{\frac{2\big(\sqrt{2}-1\big)\ln|\mathcal{C}_v|}{(e-2)\widetilde{V}_{t-1}(v)}}\right\}, \tag{4}$$

where $\widetilde{V}_{t-1}(v)$ is the cumulative variance of $\widetilde{\ell}_s(v,\cdot)$ according to $q_s(v,\cdot)$ up to time $s = t-1$. Then for all $T \geqslant 1$ the regret satisfies

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant 5\gamma T + 2^7 \int_{\gamma/2}^{1/2} \left(\sqrt{\frac{T}{\gamma}\ln \mathcal{N}_\infty(\mathcal{F},\varepsilon)} + \frac{1}{\gamma}\Big(\ln \mathcal{N}_\infty(\mathcal{F},\varepsilon) + 1\Big)\right) \mathrm{d}\varepsilon\,.$$

In particular, if $\mathcal{X} \equiv [0,1]^d$ is endowed with the sup norm $\rho_{\mathcal{X}}(x,x') = \|x - x'\|_\infty$, then the set $\mathcal{F}$ of all 1-Lipschitz functions from $\mathcal{X}$ to $[0,1]$ satisfies $\ln \mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \varepsilon^{-d}$. Theorem 4 thus entails the following corollary.

Corollary 5 *Under the assumptions of Theorem 4, if $\mathcal{F}$ is the set of all 1-Lipschitz functions $f : [0,1]^d \to [0,1]$, then the regret of HierExp4 satisfies*

$$\mathrm{Reg}_T(\mathcal{F}) = \begin{cases} \mathcal{O}\big(T^{2/3}\big) & \text{if } d = 1 \\ \mathcal{O}\big(T^{2/3}(\ln T)^{2/3}\big) & \text{if } d = 2 \\ \mathcal{O}\big(T^{d/(d+1)}\big) & \text{if } d \geqslant 3 \end{cases}$$

where the last inequality is obtained by optimizing the choice of γ for the different choices of d.

Sketch of proof (of Theorem 4) As we said earlier, the key contribution of chaining is that it allows us to sum up local regret bounds scaling with the range of the local losses. We divide our proof into four parts, and sketch the main arguments below.

Part 1: small local ranges. By construction of the tree $\mathcal{T}_\mathcal{F}$, the losses associated with neighboring nodes are close to one another —see (Cesa-Bianchi et al., 2017, Lemma 13). This implies that, if $v \in \mathcal{T}_\mathcal{F}$ is a node at level $m \geqslant 0$, then the losses associated to its children $w, w' \in \mathcal{C}_v$ are close: $|p_t(w,\cdot)\cdot\ell_t - p_t(w',\cdot)\cdot\ell_t| \leqslant 2^{-m+3}$.

Part 2: apply a version of Exp4 that scales with the loss range. Now, by definition, each node $v \in \mathcal{T}_\mathcal{F}$ runs a version of Exp4 with $|\mathcal{C}_v| \leqslant N_{m+1}$ experts (its children), whose losses belong to a range of size $E_{m+1} := 2^{-m+3}$. In full generality Exp4 cannot scale with the range E_{m+1}, but here this is possible because of the richer feedback structure induced by the total order on the actions —see (Cesa-Bianchi et al., 2017, Theorem 10). We get the following regret bound for node v with respect to its children w:

$$\max_{w\in\mathcal{C}_v} \mathbb{E}\left[\sum_{t=1}^T p_t(v,\cdot)\cdot\ell_t - \sum_{t=1}^T p_t(w,\cdot)\cdot\ell_t\right] \lesssim E_{m+1}\sqrt{\frac{T\ln N_{m+1}}{\gamma}} + \cancel{E_{m+1}}\frac{\ln N_{m+1}}{\gamma}\,,$$

where $\gamma > 0$ is the exploration parameter. For simplicity, $\lesssim$ denotes an inequality up to constant factors; we also ignore the last additive term in this sketch.

Part 3: sum over a path to get the regret of the root. Now consider the path $v_0 \to v_1 \to \cdots \to v_M = w$ from the root v_0 to some leaf $v_M = w$. Recalling that $p_t(w,i) = \mathbb{I}_{i=i_t(w)}$ for any leaf w, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T p_t(v_0,\cdot)\cdot\ell_t\right] - \sum_{t=1}^T \ell_t(y_{i_t(w)}) &= \sum_{m=0}^{M-1}\mathbb{E}\left[\sum_{t=1}^T p_t(v_m,\cdot)\cdot\ell_t - \sum_{t=1}^T p_t(v_{m+1},\cdot)\cdot\ell_t\right] \\ &\lesssim \sum_{m=0}^{M-1} 2^{-m}\sqrt{\frac{T\ln N_{m+1}}{\gamma}}\,. \end{aligned} \tag{5}$$

Part 4: Comparing our prediction to that of the root v_0 and approximating $\mathcal{F}$ with $\mathcal{L}$. Relating our prediction $\widehat{y}_t$ with the root v_0, we have $\mathbb{E}[\ell_t(\widehat{y}_t)] = \mathbb{E}\big[(1-\gamma)p_t(v_0,\cdot)\cdot\ell_t + \gamma\,\ell_t(y_1)\big]$, which entails

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(\widehat{y}_t)\right] \leqslant \mathbb{E}\left[\sum_{t=1}^T p_t(v_0,\cdot)\cdot\ell_t\right] + \gamma T\,. \tag{6}$$

Moreover, because $\mathcal{L}$ is a (2^{-M})-covering of $\mathcal{F}$ and $\mathcal{K}$ is a (2^{-M})-covering of $[0,1]$, for any $f \in \mathcal{F}$ there exists $w \in \mathcal{L}$ such that $\big|\ell_t\big(y_{i_t(w)}\big) - \ell_t\big(f(x_t)\big)\big| \leqslant \big|y_{i_t(w)} - f(x_t)\big| \leqslant 2^{1-M}$ for all t (by definition of $i_t(w)$). Plugging the approximation $\ell_t\big(y_{i_t(w)}\big) \leqslant \ell_t\big(f(x_t)\big) + 2^{1-M}$ into (5), and combining with (6), we finally get

$$\mathrm{Reg}_T(\mathcal{F}) := \mathbb{E}\left[\sum_{t=1}^T \ell_t(\widehat{y}_t)\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell_t\big(f(x_t)\big) \lesssim (2^{-M} + \gamma)T + \sum_{m=0}^{M-1} 2^{-m}\sqrt{\frac{T\ln N_{m+1}}{\gamma}}\,.$$

The proof is concluded by using $M = \lfloor \ln_2(1/\gamma)\rfloor$, $N_{m+1} = \mathcal{N}_\infty(\mathcal{F}, 2^{-(m+1)})$, and approximating the last sum by an integral. ■

4.3. Efficient chaining

Though very general, HierExp4 (Algorithm 2) may be very inefficient. For example, when $\mathcal{F}$ is the set of all 1-Lipschitz functions from $[0,1]^d$ to $[0,1]$, a direct implementation of HierExp4 would require $\exp\big(\mathrm{poly}(T)\big)$ weight updates at every round. In this section we tackle the special case when $\mathcal{F}$ is the class of all 1-Lipschitz functions $f : [0,1]^d \to [0,1]$ w.r.t. the sup norm on $[0,1]^d$ (for simplicity). We construct an ad-hoc hierarchical covering of $\mathcal{F}$ and define a variant of HierExp4 whose running time at every round is polynomial in T. We rely on a well-known wavelet-like approximation technique which was used earlier —see, e.g., (Gaillard and Gerchinovitz, 2015)— for online nonparametric regression with full information feedback. However, we replace their multi-variable Exponentiated Gradient algorithm, which requires convex losses and gradient information, with a more involved chaining algorithm that still enjoys a polynomial running time. The definitions of our covering tree $\mathcal{T}^\star_{\mathcal{F}}$ and of our algorithm HierExp4*, as well as the proof of the following regret bound, can be found in the full version of the paper (Cesa-Bianchi et al., 2017). The exact value of c_T (depending at most logarithmically on T) is also provided there.

Theorem 6 *Let $\mathcal{F}$ be the set of all 1-Lipschitz functions $f : [0,1]^d \to [0,1]$ w.r.t. the sup norm on $[0,1]^d$. Consider $T \geqslant 3$ and any sequence $(x_1,\ell_1),\dots,(x_T,\ell_T)$ of contexts $x_t \in [0,1]^d$ and 1-Lipschitz loss functions $\ell_t : [0,1] \to [0,1]$. Assume HierExp4* (Cesa-Bianchi et al., 2017) is run with one-sided full information feedback using tree $\mathcal{T}^\star_{\mathcal{F}}$ of depth $M = \lceil \ln_2(1/\gamma)\rceil$, exploration parameter $\gamma = T^{-1/2}(\ln T)^{-1}\mathbb{I}_{d=1} + T^{-1/(d+2/3)}\mathbb{I}_{d>1}$, learning rate $\eta_m = c_T 2^{m(\frac{d}{4}+1)}\gamma^{\frac{1}{2}}T^{-\frac{1}{4}}$, and penalization $\alpha_m = \sum_{j=m+1}^M 2^{4-2j}\eta_j$ for $m = 0,\dots,M-1$. Then the regret satisfies*

$$\mathrm{Reg}_T(\mathcal{F}) = \begin{cases} \mathcal{O}\big(\sqrt{T\ln T}\big) & \text{if } d = 1, \\ \mathcal{O}\big(T^{\frac{d-1/3}{d+2/3}}(\ln T)^{3/2}\big) & \text{if } d \geqslant 2. \end{cases}$$

Moreover, the running time at every round is $\mathcal{O}\big(T^a\big)$ with $a = (1+\ln_2 3)/(d+2/3)$.

The above result improves on Corollary 5 in two ways. First, as we said, the running time is now polynomial in T, contrary to what could be obtained via a direct implementation of HierExp4. Second, when $d \geqslant 2$, the regret bound is of order $T^{(d-1/3)/(d+2/3)}$, improving on the rate $T^{d/(d+1)}$ from Corollary 5. Remarkably, Theorem 6 also yields a regret of $\widetilde{\mathcal{O}}(\sqrt{T})$ for nonparametric bandits with one-sided full information feedback in dimension $d = 1$. The improvement on the rates compared to HierExp4 is possible because we use a variant of Exp4 with *penalized* loss estimates.

This allows for a careful hierarchical control of the variance terms inspired by the analysis of Exp3-RTB in (Cesa-Bianchi et al., 2017).

Note that the time complexity decreases as the dimension d increases. Indeed, when d increases the regret gets worse but, at the same time, the size of the discretized action space and the number of layers in our wavelet-like approximation can be both set to smaller values.

5. A Tight Bound for Full Information through an Explicit Algorithm

In this section we apply the machinery developed in Section 4.2 to the full information setting, where after each round t the learner can compute the loss $\ell_t(y)$ of any number of actions $y \in \mathcal{Y}$. We obtain the first explicit algorithm achieving, up to logarithmic factors, the minimax regret rate $T^{(d-1)/d}$ for all classes of Lipschitz functions, where d is the dimension of the context space. This achieves the same upper bound as the one proven by Rakhlin et al. (2015) in a nonconstructive manner, and matches the lower bound of Hazan and Megiddo (2007). Our approach generalizes the approach of Gaillard and Gerchinovitz (2015) to nonconvex Lipschitz losses. We consider a full information variant of HierExp4 (Algorithm 2, Section 4.2), where —using the same notation as in Section 4.2— the Exp4 instances running on the nodes of the tree $\mathcal{T}_\mathcal{F}$ are replaced by instances of Hedge —e.g., (Bubeck and Cesa-Bianchi, 2012). Note that, due to the full information assumption, the new algorithm, called HierHedge, observes losses at all leaves $v \in \mathcal{L}$. As a consequence, no exploration is needed and so we can set $\gamma = 0$. For the same reason, the estimated loss vectors defined in (2) can be replaced with the true loss vectors, ℓ_t. See (Cesa-Bianchi et al., 2017) for a definition of HierHedge. The latter also contains a proof of the next result.

Theorem 7 *Fix any class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$ and any sequence $(x_1, \ell_1), (x_2, \ell_2), \dots$ of contexts $x_t \in \mathcal{X}$ and 1-Lipschitz loss functions $\ell_t : \mathcal{Y} \to [0,1]$. Assume HierHedge (Cesa-Bianchi et al., 2017) is run with full information feedback on the tree $\mathcal{T}_\mathcal{F}$ of depth $M = \lfloor \ln_2(1/\varepsilon) \rfloor$ with action set $\mathcal{Y}_\varepsilon$ for $\varepsilon > 0$. Moreover, the learning rate $\eta_t(v)$ used at each node v at depth $m = 0, \dots, M-1$ is given by (4). Then for all $T \geqslant 1$ the regret satisfies*

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant 5\varepsilon T + 2^7 \int_{\varepsilon/2}^{1/2} \left(2\sqrt{T \ln \mathcal{N}_\infty(\mathcal{F}, x)} + \ln \mathcal{N}_\infty(\mathcal{F}, x) \right) \mathrm{d}x \,.$$

In particular, if $d \geqslant 3$ and $\mathcal{F}$ is the set of 1-Lipschitz functions $f : [0,1]^d \to [0,1]^p$, where $[0,1]^d$ and $[0,1]^p$ are endowed with their sup norms, the choice $\varepsilon = (p/T)^{1/d}$ yields $\mathrm{Reg}_T(\mathcal{F}) = \widetilde{O}\big(T^{(d-1)/d}\big)$, while for $1 \leqslant d \leqslant 2$ the regret is of order $\sqrt{pT}$, ignoring logarithmic factors.

When using the sup norms, the dimension p of the action space only appears as a multiplicative factor $p^{1/d}$ in the regret bound for Lipschitz functions. Note also that an efficient version of HierHedge for Lipschitz functions can be derived along the same lines as the construction in Section 4.3.

Acknowledgments

This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA).

References

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 (ICML)*, 2014.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *CoRR*, abs/1409.8428, 2014.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, pages 23–35, 2015.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011a.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the Lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011b.

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *arXiv preprint arXiv:1607.03084*, 2016.

Nicolò Cesa-Bianchi and Gábor Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1895, 1999.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2015.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic Chaining and the Role of Partial Feedback in Online Nonparametric Learning. Preprint, February 2017. URL https://arxiv.org/abs/1702.08211.

Emile Contal and Nicolas Vayatis. Stochastic process bandits: Upper confidence bounds algorithms via generic chaining. *arXiv preprint arXiv:1602.04976*, 2016.

Emile Contal, Cédric Malherbe, and Nicolas Vayatis. Optimization for gaussian processes via chaining. *arXiv preprint arXiv:1510.05576*, 2015.

Rocco De Rosa, Francesco Orabona, and Nicolò Cesa-Bianchi. The ABACOC algorithm: A novel approach for nonparametric classification of data streams. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 733–738, 2015.

Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

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*, 55(1):119–139, 1997.

Pierre Gaillard and Sebastien Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Proceedings of COLT'15*, volume 40, pages 764–796. JMLR: Workshop and Conference Proceedings, 2015.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems 29 (NIPS'16)*, pages 1198–1206, 2016.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2015.

Elad Hazan and Nimrod Megiddo. Online learning with prior knowledge. In *International Conference on Computational Learning Theory (COLT'07)*, pages 499–513. 2007.

Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS*, volume 17, pages 697–704, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *AISTATS*, pages 485–492, 2010.

Odalric-Ambrym Maillard and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In *AISTATS*, pages 570–578, 2011.

Michael Ostrovsky and Michael Schwarz. Reserve prices in Internet advertising auctions: a field experiment. In *ACM Conference on Electronic Commerce*, pages 59–60, 2011.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *CoRR*, abs/1501.06598, 2015.

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

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16:155–186, 2015.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Aleksandrs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2016.

Vladimir Vovk. Competing with wild prediction rules. *Machine Learning*, 69(2-3):193–212, 2007.