# Online Nonparametric Learning, Chaining, and the Role of Partial Feedback

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, Sébastien Gerchinovitz

JMLR: Workshop and Conference Proceedings vol 65:1–36, 2017

# Online Nonparametric Learning, Chaining, and the Role of Partial Feedback

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

**Pierre Gaillard** PIERRE@GAILLARD.ME
*INRIA, 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 characterize the minimax regret up to log factors by proving an upper bound matching a previously known lower bound. In a partial feedback model motivated by second-price auctions, we prove upper bounds for Lipschitz and semi-Lipschitz losses that improve 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.

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 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 $\mathcal{O}\big(T^{(d-1)/d}\big)$ lower bound, leaving open the problem of closing the gap between upper and lower bound. As noted elsewhere —see, e.g., (Slivkins, 2014)— the approach of Hazan and Megiddo (2007) can be 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.

**Our contributions.** We essentially close the gap in the full information model, by showing that the minimax regret rate for Lipschitz policies and Lipschitz losses is indeed $T^{(d-1)/d}$ (excluding logarithmic factors in $T$ and polynomial factors in the metric dimension of the action space). Moreover, motivated by a problem in online advertising where the action space is the $[0,1]$ interval, we 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 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 attained by an 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.

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

| 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 $d \geqslant 2$. 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 apply chaining to a hierarchical covering of the policy space. Despite we are not the first ones to use chaining 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 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 $\varepsilon$. For any $0 < \varepsilon \leqslant 1$, we use $\mathcal{Y}_\varepsilon$ to denote any $\varepsilon$-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 $(x_1, \ell_1), (x_2, \ell_2), \ldots$ of contexts $x_t \in \mathcal{X}$ and loss functions $\ell_t : \mathcal{Y} \to [0,1]$, for every round $t = 1, 2, \ldots$

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 $\ell_t$ 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,

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

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 (see Footnote 2) 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 $\ell_t$ according to law $p_t$. Finally, the set of all probability distributions over a finite set of $K$ elements is denoted by $\Delta(K)$.

## 2. Related work

Contextual online learning generalizes online convex optimization (Hazan, 2015) to nonconvex losses. Besides the paper by Hazan and Megiddo (2007), which we already mentioned in the introduction, works about nonparametric online learning in full information include (Vovk, 2007; Gaillard and Gerchinovitz, 2015) for the square loss, and (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; 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. Cesa-Bianchi and Lugosi (1999) designed an algorithm based on chaining to prove regret bounds in online learning with linear losses and full information. Other notable examples of algorithmic chaining are the stochastic bandit algorithms of Contal et al. (2015) and Contal and Vayatis (2016). The 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

2. Assuming a unit Lipschitz constant is without loss of generality because our algorithms are oblivious to it. More detailed regret bounds, which optimize the dependence on the unknown Lipschitz constant, are proven in (Bubeck et al., 2011) in a stochastic and context-free bandit setting.

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. Deriving regret bounds that scale with the effective range of the losses is however not always possible, as shown by Gerchinovitz and Lattimore (2016) in the nonstochastic $K$-armed bandit setting. Indeed, chaining does not seem to help in online nonparametric learning when the feedback is bandit. However, as we show in this paper, chaining helps improving the regret with one-sided full information feedback, while with full information feedback chaining delivers regret bounds that match (up to log factors) the lower bound of Hazan and Megiddo (2007).

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 $\varepsilon$ in the context space, where each ball hosts an instance of the Exp3 algorithm of Auer et al. (2002).[3] At each round $t$, if a new incoming context $x_t \in \mathcal{X}$ is not contained 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 $K_\varepsilon$ is at most $C_{\mathcal{Y}}\,\varepsilon^{-p}$. For completeness, the proof of the following theorem is provided in

**Algorithm 1:** ContextualExp3 (for bandit feedback)

**Input:** Ball radius $\varepsilon > 0$, $\varepsilon$-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 $\varepsilon$ 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**

Appendix B.

3. 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.

**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 $\varepsilon = (\ln T)^{\frac{2}{p+d+2}} T^{-\frac{1}{p+d+2}}$, then its regret $\mathrm{Reg}_T(\mathcal{F})$ with respect to the set $\mathcal{F}$ of 1-Lipschitz[4] 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 $\ell_t$ 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. In fact, we are unaware of a lower bound matching the upper bound in Theorem 1 when $\mathcal{F}$ is the class of (global) Lipschitz functions.

Note that the dependence on $p$ in the bound of Theorem 1 can be greatly improved in the special case of convex Lipschitz losses. 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 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 $\ell_t\big(\widehat{y}_t\big)$, 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)

4. It is not hard to see that the bound of Theorem 1 also holds when the functions in $\mathcal{F}$ satisfy a weaker piecewise Lipschitz condition. In fact, this weaker Lipschitz condition also works for Theorem 3 in Section 4.1 (but not for the results using chaining through hierarchical covering in subsequent sections).

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 $\widetilde{\mathcal{O}}\big(\sqrt{T}\big)$ 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, in Theorem 3 (Appendix C) we prove the first $\widetilde{\mathcal{O}}\big(\sqrt{T}\big)$ regret bound for the seller in nonstochastic and noncontextual second-price auctions. 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 $g\big(\widehat{y}, b(1), b(2)\big) = \max\big\{\widehat{y}, b(2)\big\}\mathbb{I}_{\widehat{y}\le b(1)}$. Note that the revenue only depends on the reserve price $\widehat{y}$ and on the two highest bids $b(1) \ge b(2)$, 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 $g_t\big(\widehat{y}_t\big) = g\big(\widehat{y}_t, b_t(1), b_t(2)\big)$, 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 $\ell_t\big(\widehat{y}_t\big) = 1 - g_t\big(\widehat{y}_t\big)$ instead of revenues, see Figure 1 for a pictorial representation.

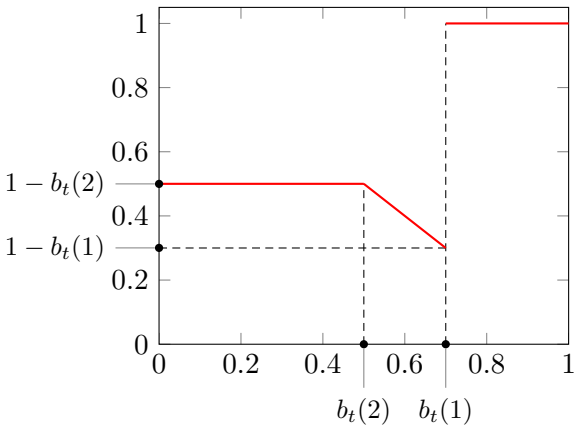


Figure 1: The loss function $\ell_t(\widehat{y}_t) = 1 - \max\{\widehat{y}_t, b_t(2)\}\mathbb{I}_{\widehat{y}_t \le b_t(1)}$ when $b_t(1) = 0.7$ and $b_t(2) = 0.5$.

Remarkably, the loss functions $\ell_t$ 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 prove 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 3 in Appendix C), a variant of ContextualExp3 (Algorithm 1), where each ball hosts an instance of Exp3-RTB, instead of Exp3. The proof is given in Appendix D.

**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 $\varepsilon = T^{-1/(d+2)}$ 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 $\mathrm{Reg}_T(\mathcal{F}) = \widetilde{\mathcal{O}}\left(T^{\frac{d+1}{d+2}}\right)$, 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 $\varepsilon$ 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 $(x_t, \ell_t)$ 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

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

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 a class $\mathcal{F}$ of functions $f : \mathcal{X} \to [0,1]$ and introduce the sup norm

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

We denote by $\mathcal{N}_\infty\big(\mathcal{F}, \varepsilon\big)$ the cardinality of the smallest $\varepsilon$-cover of $\mathcal{F}$ w.r.t. the sup norm. 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

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

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 $v \in \mathcal{T}_\mathcal{F}$.

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 Algorithm 4 in Appendix E) 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

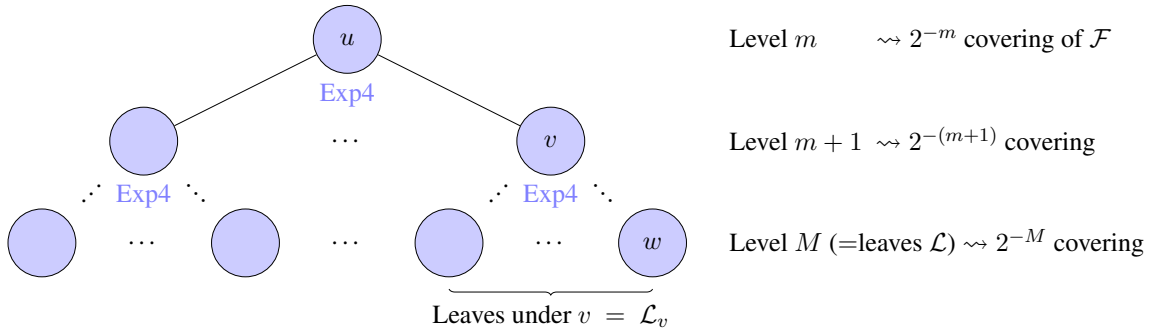


Figure 2: Hierarchical covering of the function space (used in Algorithm 2).

$q_t(v,\cdot) \in \Delta(|\mathcal{C}_v|)$ 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.

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 in Appendix F.

**Theorem 4** *Fix any class $\mathcal{F}$ of functions $f : \mathcal{X} \to [0,1]$ and any sequence $(x_1,\ell_1),(x_2,\ell_2),\dots$ 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 $m = 0,\dots,M-1$ is given by (11) in Appendix E, with $E = 2^{-m+3}$, $N = |\mathcal{C}_v|$, and $\widetilde{V}_{t-1}$ being the cumulative variance of $\widetilde{\ell}_s$ according to $q_s(v,\cdot)$ up to time $s = t-1$. Then for all $T \geq 1$ 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 $\gamma$ for the different choices of $d$.*

**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**

### 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 $f : [0,1]^d \to [0,1]$. 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 $\mathcal{T}_{\mathcal{F}}^{\star}$ and of our algorithm HierExp4$^{\star}$, as well as the proof of the following regret bound, can be found in Appendix G. 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 $f : [0,1]^d \to [0,1]$. Consider $T \geqslant 3$ and any sequence $(x_1,\ell_1),\ldots,(x_T,\ell_T)$ of contexts $x_t \in [0,1]^d$ and 1-Lipschitz loss functions $\ell_t : [0,1] \to [0,1]$. Assume HierExp4$^{\star}$ (Algorithm 5) is run with one-sided full information feedback using tree $\mathcal{T}_{\mathcal{F}}^{\star}$ of depth $M = \lceil \log_2(1/\gamma) \rceil$, 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* $a = (1 + \log_2 3)/(d + 2/3)$.

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 our analysis of Exp3-RTB in Appendix C.

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

In this section we apply the machinery developed in Section 4.2 to a 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 characterize, up to logarithmic factors, the minimax regret rate by proving a regret bound of order $\widetilde{O}\big(T^{(d-1)/d}\big)$ for all classes of Lipschitz functions, where $d$ is the dimension of the context space and $p$ is the dimension of the action space. This matches the lower bound of Hazan and Megiddo (2007), while generalizing 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, $\ell_t$. See Algorithm 6 in Appendix H for a definition of HierHedge. The same appendix 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* $\ell_t : \mathcal{X} \to [0, 1]$. *Assume HierHedge 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 (11) in Appendix E, with* $E = 2^{-m+3}$, $N = |\mathcal{C}_v|$, *and* $\widetilde{V}_{t-1}$ *being the cumulative variance of* $\widetilde{\ell}_s$ *according to* $q_s(v, \cdot)$ *up to time* $s = t - 1$. *Then for all* $T \geq 1$ *the regret satisfies*

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

*In particular, if $d \geqslant 3$ and $\mathcal{F}$ is the set of Lipschitz functions $f : [0,1]^d \to [0,1]^p$, where $[0,1]^d$ is endowed with the norm $\|x - x'\|_\infty$, 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.*

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.

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

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, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

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.

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.

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.

## Appendix A. A Useful Lemma

The following is a standard result in the analysis of Hedge with variable learning rate —see, e.g., the analysis of (Bubeck and Cesa-Bianchi, 2012, Theorem 3.2). Recall that, when run on a loss sequence $\ell_1, \ell_2, \dots$ Hedge computes distributions $p_1, p_2, \dots$ where $p_1$ is uniform and $p_t(i)$ is proportional to $\exp\big(-\eta_t \sum_{s=1}^{t-1} \ell_s(i)\big)$ for all actions $i$.

**Lemma 8** *The sequence $p_1, p_2, \dots$ of probability distributions computed by Hedge when run on $K$ actions with learning rates $\eta_1 \geqslant \eta_2 \geqslant \cdots > 0$ and losses $\ell_t(k) \geqslant 0$ satisfies*

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\ell_t(i) - \min_{k=1,\dots,K} \sum_{t=1}^{T} \ell_t(k) \leqslant \frac{\ln K}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} p_t(i)\ell_t(i)^2 \,.$$

## Appendix B. Proof of Theorem 1

**Proof** Let $N_t$ be the number of balls created by the ContextualExp algorithm after $t$ rounds, $B_s$ be the $s$-th ball so created, being $\bar{x}_s$ its center ($\bar{x}_s$ is some past context), and $T_s$ be the subsequence of rounds $t$ such that $x_t$ is handled by the $s$-th ball. Notice that $N_t$ and $T_s$ are deterministic quantities, since the $x_t$'s are generated obliviously. Since $f$ is 1-Lipschitz and $\ell_t$ is also 1-Lipschitz, for all $x_t$ handled by $B_s$, we can write $\big|\ell_t\big(f(x_t)\big) - \ell_t\big(f(\bar{x}_s)\big)\big| \leq \varepsilon$ . Now fix any 1-Lipschitz policy $f$. For each $s = 1, \dots, N_T$, there exists $\bar{y}_s \in \mathcal{Y}_\varepsilon$ such that $\rho_{\mathcal{Y}}\big(\bar{y}_s, f(\bar{x}_s)\big) \leq \varepsilon$. Then we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\widehat{y}_t)\right] - \sum_{t=1}^{T} \ell_t\big(f(x_t)\big) \qquad & (4)\\
&= \sum_{s=1}^{N_T} \sum_{t \in T_s} \Big(\mathbb{E}\big[\ell_t(\widehat{y}_t)\big] - \ell_t\big(f(x_t)\big)\Big)\\
&= \sum_{s=1}^{N_T} \sum_{t \in T_s} \Big(\mathbb{E}\big[\ell_t(\widehat{y}_t)\big] - \ell_t\big(\bar{y}_s\big) + \ell_t\big(\bar{y}_s\big) - \ell_t(f(\bar{x}_s)) + \ell_t(f(\bar{x}_s)) - \ell_t(f(x_t))\Big)\\
&\leq \sum_{s=1}^{N_T} \sum_{t \in T_s} \Big(\mathbb{E}\big[\ell_t(\widehat{y}_t)\big] - \ell_t\big(\bar{y}_s\big)\Big) + 2T\varepsilon \,. \qquad (5)
\end{aligned}$$

We now apply to each $s = 1, \dots, N_T$ the standard regret bound of Exp3 with learning rate $\eta$ (e.g., (Bubeck and Cesa-Bianchi, 2012)) w.r.t. the best action $\bar{y}_s$. This yields

$$\sum_{t \in T_s} \Big(\mathbb{E}\big[\ell_t(\widehat{y}_t)\big] - \ell_t\big(\bar{y}_s\big)\Big) \leqslant \frac{\ln K_\varepsilon}{\eta} + \frac{\eta}{2} \left|T_s\right| K_\varepsilon \,,$$

implying

$$\sum_{s=1}^{N_T} \sum_{t \in T_s} \Big( \mathbb{E}\big[\ell_t(\widehat{y}_t)\big] - \ell_t(\bar{y}_s) \Big) \leqslant \frac{N_T \ln K_\varepsilon}{\eta} + \frac{\eta}{2} T K_\varepsilon \,.$$

Combining with (5), setting $\eta = \sqrt{\frac{2N_T \ln K_\varepsilon}{T K_\varepsilon}}$, recalling that $K_\varepsilon \leqslant C_{\mathcal{Y}} \varepsilon^{-p}$ and observing that, by the way balls are constructed, $N_T$ can never exceed the size of the smallest $\varepsilon/2$-cover of $\mathcal{X}$ (that is, $N_T \leqslant C_{\mathcal{X}} (\varepsilon/2)^{-d}$), we obtain

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant \sqrt{2\,T\,N_T\,K_\varepsilon\,\ln K_\varepsilon} + 2\,T\,\varepsilon = \mathcal{O}\left( \sqrt{T\,\varepsilon^{-(d+p)}\,\ln\frac{1}{\varepsilon}} + T\,\varepsilon \right) .$$

Setting $\varepsilon = \left( \left(\frac{p+d}{2}\right) \frac{\ln T}{T^{1/2}} \right)^{\frac{2}{p+d+2}}$ gives $\mathrm{Reg}_T(\mathcal{F}) = \widetilde{\mathcal{O}}\left( T^{\frac{p+d+1}{p+d+2}} \right)$ as claimed. ■

## Appendix C. The Exp3-RTB algorithm for reserve-price optimization

Cesa-Bianchi et al. (2015) showed that a regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$ can be achieved with high probability for the problem of regret minimization in second-price auctions with i.i.d. bids, when the feedback received after each auction is the seller's revenue. In this appendix, we show that the same regret rate can be obtained even when the sequence of bids is nonstochastic, provided the feedback also includes the highest bid. We use this result in order to upper bound the contextual regret in Section 4.1.

We consider a setting slightly more general than second-price auctions. Fix any unknown sequence $\ell_1, \ell_2, \ldots$ of loss functions $\ell_t : [0,1] \to [0,1]$ satisfying the semi-Lipschitz condition,

$$\ell_t(y+\delta) \geq \ell_t(y) - \delta \qquad \text{for all } 0 \leqslant y \leqslant y+\delta \leqslant 1. \tag{6}$$

In each auction instance $t = 1, 2, \ldots$, the learner selects a reserve price $\widehat{y}_t \in \mathcal{Y} = [0,1]$ and suffers the loss $\ell_t(\widehat{y}_t)$. The learner's feedback is $\ell_t(y)$ for all $y \geqslant \widehat{y}_t$ (i.e., the one-sided full information feedback). As explained in Section 4.1, this setting includes online revenue maximization in second-price auctions as a special case when the learner's feedback includes both revenue and highest bid. The learner's regret is defined by

$$\mathrm{Reg}_T \triangleq \mathbb{E}\left[ \sum_{t=1}^{T} \ell_t(\widehat{y}_t) \right] - \inf_{0 \leqslant y \leqslant 1} \sum_{t=1}^{T} \ell_t(y) \,,$$

where the expectation is with respect to the randomness in the predictions $\widehat{y}_t$. We introduce the Exp3-RTB algorithm (Algorithm 3 below), a variant of Exp3 (Auer et al., 2002) exploiting the richer feedback $\big\{\ell_t(y) \,:\, y \geqslant \widehat{y}_t\big\}$. The algorithm uses a discretization of the action space $[0,1]$ in $K = \lceil 1/\gamma \rceil$ actions $y_k := (k-1)\gamma$ for $k = 1, \ldots, K$.

**Theorem 9** *In the one-sided full information feedback, the Exp3-RTB algorithm tuned with $\gamma > 0$ satisfies*

$$\mathrm{Reg}_T \leqslant \gamma T \left( 2 + \frac{1}{4} \ln \frac{e}{\gamma} \right) + \frac{2 \ln \lceil 1/\gamma \rceil}{\gamma} \,.$$

*In particular, $\gamma = T^{-1/2}$ gives $\mathrm{Reg}_T = \widetilde{\mathcal{O}}(\sqrt{T})$.*

**Algorithm 3:** Exp3-RTB (for one-sided full information feedback)

**Input** : Exploration parameter $\gamma > 0$.

**Initialization**: Set learning rate $\eta = \gamma/2$ and uniform distribution $p_1$ over $\{1, \dots, K\}$ where $K = \lceil 1/\gamma \rceil$;

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

1. compute distribution $q_t(k) = (1-\gamma)p_t(k) + \gamma \mathbb{I}_{k=1}$ for $k = 1, \dots, K$;
2. draw $I_t \sim q_t$ and choose $\widehat{y}_t = y_{I_t} = (I_t - 1)\gamma$;
3. for each $k = 1, \dots, K$, compute the estimated loss

$$\widehat{\ell}_t(k) = \frac{\ell_t(y_k)}{\sum_{j=1}^{k} q_t(j)} \mathbb{I}_{I_t \leqslant k}$$

4. for each $k = 1, \dots, K$, compute the new probability assignment

$$p_{t+1}(k) = \frac{\exp\big(-\eta \sum_{s=1}^{t} \widehat{\ell}_s(k)\big)}{\sum_{j=1}^{K} \exp\Big(-\eta \sum_{s=1}^{t} \widehat{\ell}_s(j)\Big)} \,.$$

**end**

**Proof** The proof follows the same lines as the regret analysis of Exp3 in (Auer et al., 2002). The key change is a tighter control of the variance term allowed by the richer feedback.

Pick any reserve price $y_k = (k-1)\gamma$. We first control the regret associated with actions drawn from $p_t$ (the regret associated with $q_t$ will be studied as a direct consequence). More precisely, since the estimated losses $\widehat{\ell}_t(j)$ are nonnegative, we can apply Lemma 8 to get

$$\sum_{t=1}^{T} p_t \cdot \widehat{\ell}_t - \sum_{t=1}^{T} \widehat{\ell}_t(k) \leqslant \frac{\eta}{2} \sum_{t=1}^{T} \sum_{j=1}^{K} p_t(j) \widehat{\ell}_t(j)^2 + \frac{\ln K}{\eta} \,. \tag{7}$$

Writing $\mathbb{E}_{t-1}[\,\cdot\,]$ for the expectation conditioned on $I_1, \dots, I_{t-1}$, we note that

$$\mathbb{E}_{t-1}\big[\widehat{\ell}_t(j)\big] = \ell_t(y_j) \quad \text{and} \quad \mathbb{E}_{t-1}\big[p_t(j)\widehat{\ell}_t(j)^2\big] = \frac{p_t(j)\ell_t(y_j)^2}{\sum_{i=1}^{j} q_t(i)} \leqslant \frac{q_t(j)}{(1-\gamma)\sum_{i=1}^{j} q_t(i)} \,,$$

where we used the definition of $q_t$ and the fact that $|\ell_t(y_j)| \leqslant 1$ by assumption. Therefore, taking expectation on both sides of (7) entails, by the tower rule for expectations,

$$\mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t\right] - \sum_{t=1}^{T} \ell_t(y_k) \leqslant \frac{\eta}{2(1-\gamma)} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{K} \frac{q_t(j)}{\sum_{i=1}^{j} q_t(i)}\right] + \frac{\ln K}{\eta} \,.$$

Setting $s_t(j) \triangleq \sum_{i=1}^{j} q_t(i)$, we can upper bound the sum with an integral,

$$\sum_{j=1}^{K} \frac{q_t(j)}{\sum_{i=1}^{j} q_t(i)} = 1 + \sum_{j=2}^{K} \frac{s_t(j) - s_t(j-1)}{s_t(j)} = 1 + \sum_{j=2}^{K} \int_{s_t(j-1)}^{s_t(j)} \frac{\mathrm{d}x}{s_t(j)}$$
$$\leqslant 1 + \sum_{j=2}^{K} \int_{s_t(j-1)}^{s_t(j)} \frac{\mathrm{d}x}{x} = 1 + \int_{q_t(1)}^{1} \frac{\mathrm{d}x}{x} \leqslant 1 - \ln q_t(1) \leqslant 1 + \ln \frac{1}{\gamma} \,,$$

where we used $q_t(1) \geqslant \gamma$. Therefore, substituting into the previous bound, we get

$$\mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t\right] - \sum_{t=1}^{T} \ell_t(y_k) \leqslant \frac{\eta T \ln(e/\gamma)}{2(1-\gamma)} + \frac{\ln K}{\eta} \,. \tag{8}$$

We now control the regret of the predictions $\widehat{y}_t = y_{I_t}$, where $I_t$ is drawn from $q_t = (1-\gamma)p_t + \gamma\delta_1$. By the tower rule,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\widehat{y}_t)\right] - \sum_{t=1}^{T} \ell_t(y_k) = \mathbb{E}\left[\sum_{t=1}^{T} \big((1-\gamma)p_t \cdot \ell_t + \gamma \ell_t(y_1)\big)\right] - \sum_{t=1}^{T} \ell_t(y_k)$$
$$\leqslant (1-\gamma)\,\mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t - \sum_{t=1}^{T} \ell_t(y_k)\right] + \gamma T$$
$$\leqslant \frac{\eta T \ln(e/\gamma)}{2} + \frac{\ln K}{\eta} + \gamma T \,, \tag{9}$$

where the last inequality is by (8).

To conclude the proof, we upper bound the regret against any fixed $y \in [0,1]$. Since there exists $k \in \{1,\dots,K\}$ such that $y \in [y_k, y_k + \gamma]$, and since each $\ell_t$ satisfies the semi-Lipschitz condition (6), we have $\ell_t(y) \geqslant \ell_t(y_k) - \gamma$. This gives

$$\min_{k=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(y_k)\right] \leqslant \min_{0 \leqslant y \leqslant 1} \sum_{t=1}^{T} \ell_t(y) + \gamma\, T \,.$$

Replacing the last inequality into (9), and recalling that $K = \lceil 1/\gamma \rceil$ and $\eta = \frac{\gamma}{2}$, finally yields

$$\mathrm{Reg}_T \leqslant \frac{\gamma T}{4} \ln \frac{e}{\gamma} + \frac{2 \ln \lceil 1/\gamma \rceil}{\gamma} + 2\gamma T \,.$$

Choosing $\gamma \approx T^{-1/2}$ concludes the proof. ■

## Appendix D. Proof of Theorem 3

**Proof** The proof is very similar to that of Theorem 1. We only highlight the main differences. Let $f \in \mathcal{F}$ be any 1-Lipschitz function from $\mathcal{X}$ to $[0,1]$. For all $s = 1,\dots,N_T$, we define the constant approximation of $f$ in the $s$-th ball by $y_s^{\min} = \min_{t \in T_s} f(x_t)$. Since $f$ is 1-Lipschitz and

the balls have radius $\varepsilon$, we have $\max_{t,t' \in T_s} \left| f(x_t) - f(x_{t'}) \right| \leq 2\varepsilon$. Hence, for all $t \in T_s$, by the semi-Lipschitz property (6),

$$\ell_t\big(f(x_t)\big) \geq \ell_t(y_s^{\min}) - 2\varepsilon \,. \tag{10}$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\ell_t(\widehat{y}_t)] - \sum_{t=1}^{T} \ell_t\big(f(x_t)\big) &= \sum_{s=1}^{N_T} \sum_{t \in T_s} \Big( \mathbb{E}[\ell_t(\widehat{y}_t)] - \ell_t\big(f(x_t)\big) \Big) \\
&= \sum_{s=1}^{N_T} \underbrace{\sum_{t \in T_s} \Big( \mathbb{E}[\ell_t(\widehat{y}_t)] - \ell_t(y_s^{\min}) \Big)}_{R_s} + \sum_{s=1}^{N_T} \sum_{t \in T_s} \Big( \ell_t(y_s^{\min}) - \ell_t\big(f(x_t)\big) \Big) \\
&\leq \gamma T \left( 2 + \frac{1}{4} \ln \frac{e}{\gamma} \right) + \frac{2 N_T \ln \lceil 1/\gamma \rceil}{\gamma} + 2\varepsilon T \,.
\end{aligned}$$

Each term $R_s$ is the regret suffered by the $s$-th instance of Exp3-RTB against the constant value $y_s^{\min}$, and so we bound it using Theorem 9 in Appendix C and then sum over $s$ recalling that $\sum_s T_s = T$. The other double sum is bounded by $2T\varepsilon$ using (10). We bound $N_T$ as in Theorem 1 obtaining

$$\mathrm{Reg}_T(\mathcal{F}) \lesssim \gamma T \left( 1 + \ln \frac{1}{\gamma} \right) + \frac{\varepsilon^{-d}}{\gamma} \ln \frac{1}{\gamma} + \varepsilon T \,.$$

Finally, choosing $\varepsilon = \gamma = T^{-1/(d+2)} \leqslant 1$ gives

$$\mathrm{Reg}_T(\mathcal{F}) = \widetilde{\mathcal{O}}\left( T^{\frac{d+1}{d+2}} \right)$$

concluding the proof. ■

## Appendix E. Exp4 regret bound scaling with the range of the losses

In this section we revisit the Exp4 algorithm (Bubeck and Cesa-Bianchi, 2012, Figure 4.1) and prove a regret bound for the one-sided full information model that scales with the range of the losses. In view of the application to chaining, we formulate this result in a setting similar to (Maillard and Munos, 2011). Namely, when the action $I_t$ played at time $t$ is not necessarily drawn from the distribution prescribed by Exp4.

Exp4 (see Algorithm 4) operates in a bandit setting with $K$ actions and $N$ experts. We assume a total order $1 < \cdots < K$ on the action set $\mathcal{K} = \{1, \ldots, K\}$. At each time step $t = 1, 2, \ldots$ the advice $\xi_t(j, \cdot)$ of expert $j$ is a probability distribution over the $K$ actions. The learner combines the expert advice using convex coefficients $q_t \in \Delta(N)$. These coefficients are computed by Hedge based on the expert losses $\widetilde{\ell}_t(j) = \xi_t(j, \cdot) \cdot \widehat{\ell}_t$ for $j = 1, \ldots, N$, where $\widehat{\ell}_t$ is a vector of suitably defined loss estimates. The distribution prescribed by Exp4 is $p_t = \sum_{j=1}^{N} q_t(j) \xi_t(j, \cdot) \in \Delta(K)$, but the learner's play at time $t$ is $I_t \sim p_t^*$ for some other distribution $p_t^* \in \Delta(K)$. The feedback at time $t$ is the vector of losses $\ell_t(i)$ for all $i \geqslant I_t$.

**Algorithm 4:** Exp4 with unspecified sampling distributions $p_t^*$ (for one-sided full info feedback)

**Input** : Learning rate sequence $\eta_1 \geqslant \eta_2 \geqslant \cdots > 0$.
**Initialization**: Set $q_1$ to the uniform distribution over $\{1, \ldots, N\}$;
**for** $t = 1, 2, \ldots$ **do**

1. get expert advice $\xi_t(1, \cdot), \ldots, \xi_t(N, \cdot) \in \Delta(K)$;
2. compute distribution $p_t^*$ over $\{1, \ldots, K\}$;
3. draw $I_t \sim p_t^*$ and observe $\ell_t(i)$ for all $i \geqslant I_t$;
4. compute loss estimates $\widehat{\ell}_t(i)$ for $i = 1, \ldots, K$;
5. compute expert losses $\widetilde{\ell}_t(j) = \xi_t(j, \cdot) \cdot \widehat{\ell}_t$ for $j = 1, \ldots, N$;
6. compute the new probability assignment

$$q_{t+1}(j) = \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} \widetilde{\ell}_s(j)\right)}{\sum_{k=1}^{N} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} \widetilde{\ell}_s(k)\right)} \qquad \text{for each } j = 1, \ldots, N.$$

**end**

In this setting, the goal is to minimize the regret with respect to the performance of the best expert,

$$\max_{j=1,\ldots,N} \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t - \sum_{t=1}^{T} \xi_t(j, \cdot) \cdot \ell_t\right]$$

in expectation with respect to the random draw of $I_1, \ldots, I_T$, where $\xi_t(j, \cdot) \cdot \ell_t$ is the expected loss of expert $j$ at time $t$. In view of our chaining application, we defined the quantities $\xi_t(j, i)$ as random variables because the expert advice at time $t$ might depend on the past realizations of $I_1, \ldots, I_{t-1}$. For all $k \in \mathcal{K}$ let

$$P_t^*(k) = \sum_{i=1}^{k} p_t^*(i) \,.$$

Also, let $\mathcal{K}_t \equiv \{k \in \mathcal{K} : (\exists j)\, \xi_t(j, k) > 0\}$ (in chaining, $\mathcal{K}_t$ is typically small compared to $\mathcal{K}$). The next result shows that, when Exp4 is applied to arbitrary real-valued losses, there is a way of choosing the learning rate sequence so that the regret scales with a quantity smaller than the largest loss value. More specifically, the regret scales with a known bound $E$ on the size of the effective range of the losses

$$\max_{t=1,\ldots,T} \max_{k,k' \in \mathcal{K}_t} \left|\ell_t(k) - \ell_t(k')\right| \leqslant E \,.$$

The rich feedback structure $(\ell_t(k), k \geqslant I_t)$ is crucial to get our result, since it enables us to use $\ell_t\big(\max \mathcal{K}_t\big)$ in the definition of the loss estimates $\widehat{\ell}_t(k)$ below. Indeed, as explained in Section 4.2, $\widehat{\ell}_t(k)$ has to be explicitly computed only for those actions $k$ such that $I_t \leqslant k$ and $\xi_t(j, k) > 0$ for some $j$, i.e., $k \in \mathcal{K}_t$. In this case, $I_t \leqslant k \leqslant \max \mathcal{K}_t$ so that $\ell_t\big(\max \mathcal{K}_t\big)$ is observed.

**Theorem 10** *Let $1 < \cdots < K$ be a total order on the action set $\mathcal{K} = \{1, \ldots, K\}$. Suppose Exp4 (Algorithm 4) is run with one-sided full information feedback on an arbitrary loss sequence*

$\ell_1, \ell_2, \dots$ *with loss estimates*

$$\widehat{\ell}_t(k) = \frac{\ell_t(k) - \ell_t\big(\max \mathcal{K}_t\big)}{P_t^*(k)} \, \mathbb{I}_{I_t \leqslant k} \qquad \text{for } k = 1, \dots, K$$

*and adaptive learning rate*

$$\eta_t = \min\left\{ \frac{\gamma}{2E}, \sqrt{\frac{2\big(\sqrt{2}-1\big)(\ln N)}{(e-2)\widetilde{V}_{t-1}}} \right\} \qquad \text{for } t \geqslant 2 \tag{11}$$

*for some parameter* $E > 0$ *and where* $\widetilde{V}_{t-1} = \sum_{s=1}^{t-1} \sum_{j=1}^{N} q_s(j)\big(\widetilde{\ell}_s(j) - q_s \cdot \widetilde{\ell}_s\big)^2$ *is the cumulative variance of Hedge up to time* $t-1$*. If* $p_t^*(1) \geqslant \gamma$ *and*

$$\max_{t=1,\dots,T} \max_{k,k' \in \mathcal{K}_t} \big|\ell_t(k) - \ell_t(k')\big| \leqslant E$$

*almost surely for all* $t \geqslant 1$*, then, for all* $T \geqslant 1$*,*

$$\max_{j=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t - \sum_{t=1}^{T} \xi_t(j,\cdot) \cdot \ell_t\right] \leqslant 4E\sqrt{\frac{T \ln N}{\gamma}} + \frac{E}{\gamma}(4 \ln N + 1)\,.$$

Note that the above regret bound does not depend on the number $K$ of actions, but instead on a lower bound $\gamma$ on the probability of observing the smallest action.

**Proof** From the definition of $\mathcal{K}_t$ and because $P_t^*(k) \geqslant p_t^*(1) \geqslant \gamma$,

$$\begin{aligned}
\max_{i,j=1,\dots,N} \big|\widetilde{\ell}_t(i) - \widetilde{\ell}_t(j)\big| &= \max_{i,j=1,\dots,N} \big|\xi_t(i,\cdot) \cdot \widehat{\ell}_t - \xi_t(j,\cdot) \cdot \widehat{\ell}_t\big| \\
&\leqslant \max_{k,k' \in \mathcal{K}_t} \big|\widehat{\ell}_t(k) - \widehat{\ell}_t(k')\big| \\
&\leqslant \frac{2E}{\min_{k \in \mathcal{K}} P_t^*(k)} \leqslant \frac{2E}{\gamma}\,.
\end{aligned}$$

Since $\widetilde{E} = (2E)/\gamma$ is an upper bound on the size of the range of the losses $\widetilde{\ell}_t(j)$, we can use the bound of (Cesa-Bianchi et al., 2007, Theorem 5), which applies to Hedge run on arbitrary real-valued losses with the learning rate (11). This gives us

$$\sum_{t=1}^{T} q_t \cdot \widetilde{\ell}_t \leqslant \min_{j=1,\dots,N} \sum_{t=1}^{T} \widetilde{\ell}_t(j) + 4\sqrt{(\ln N)\widetilde{V}_T} + 2\widetilde{E} \ln N + \frac{\widetilde{E}}{2}\,. \tag{12}$$

Note that $q_t \cdot \widetilde{\ell}_t = p_t \cdot \widehat{\ell}_t$. Moreover, $K_t = \max \mathcal{K}_t$ is measurable with respect to $(I_1, \dots, I_{t-1})$, implying $\mathbb{E}_{t-1}\big[\widehat{\ell}_t(k)\big] = \ell_t(k) - \ell_t(K_t)$. Therefore, taking the expectation on both sides of (12) with respect to the random draw of $I_1, \dots, I_T$, and using Jensen together with $\widetilde{V}_T \leq \sum_{t=1}^{T} \sum_{j=1}^{N} q_t(j) \cdot \widetilde{\ell}_t(j)^2$, we get

$$\begin{aligned}
&\max_{j=1,\dots,N} \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \ell_t - \sum_{t=1}^{T} \xi_t(j,\cdot) \cdot \ell_t\right] \\
&\qquad \leqslant 4\sqrt{(\ln N) \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{N} q_t(j)\Big(\xi_t(j,\cdot) \cdot \widehat{\ell}_t\Big)^2\right]} + \frac{4E}{\gamma} \ln N + \frac{E}{\gamma}\,.
\end{aligned} \tag{13}$$

The variance term inside the square root can be upper bounded as follows. Using Jensen again,

$$\begin{aligned}\sum_{j=1}^{N} q_t(j)\Big(\xi_t(j,\cdot)\cdot\widehat{\ell}_t\Big)^2 &\leqslant \sum_{j=1}^{N} q_t(j)\sum_{k=1}^{K}\xi_t(j,k)\widehat{\ell}_t(k)^2 \\ &= \sum_{j=1}^{N} q_t(j)\sum_{k=1}^{K}\xi_t(j,k)\left(\frac{\ell_t(k)-\ell_t(K_t)}{P_t^*(k)}\right)^2 \mathbb{I}_{I_t\leqslant k} \\ &\leqslant \frac{E^2}{\gamma}\sum_{j=1}^{N} q_t(j)\sum_{k=1}^{K}\frac{\xi_t(j,k)}{P_t^*(k)}\,\mathbb{I}_{I_t\leqslant k}\end{aligned}$$

where the last inequality is because $\big|\ell_t(k)-\ell_t(K_t)\big|\leqslant E$ when $\xi_t(j,k)>0$ and because $P_t^*(k)\geqslant p_t^*(1)\geqslant\gamma$. Therefore, recalling $P_t^*(k)=\mathbb{P}_{t-1}(I_t\leqslant k)$,

$$\mathbb{E}_{t-1}\left[\sum_{j=1}^{N} q_t(j)\big(\xi_t(j,k)\cdot\widehat{\ell}_t\big)^2\right]\leqslant\frac{E^2}{\gamma}\sum_{j=1}^{N} q_t(j)\sum_{k=1}^{K}\frac{\xi_t(j,k)}{P_t^*(k)}\mathbb{P}_{t-1}(I_t\leqslant k)=\frac{E^2}{\gamma}\,.$$

Substituting the last bound in (13) concludes the proof. ■

Next we extend the previous result to penalized loss estimates, which is useful in Section 4.3 to control the variance terms all along the covering tree.

**Theorem 11** *Let $1<\cdots<K$ be a total order on the action set $\mathcal{K}=\{1,\dots,K\}$. Let $E,F>0$. Consider any penalty $\mathrm{pen}_t\in\mathbb{R}^K$ measurable with respect to $(I_1,\dots,I_{t-1})$ at time $t$. Suppose Exp4 (Algorithm 4) is run with one-sided full information feedback on an arbitrary loss sequence $\ell_t\in\mathbb{R}^K$, $t\geqslant 1$, with loss estimates*

$$\widehat{\ell}_t(k)=\frac{\ell_t(k)-\ell_t\big(\max\mathcal{K}_t\big)+E}{P_t^*(k)}\,\mathbb{I}_{I_t\leqslant k}+\mathrm{pen}_t(k)+F\qquad\text{for } k=1,\dots,K$$

*and constant learning rate $\eta>0$. If we have, for all $t\geqslant 1$, almost surely,*

$$\max_{t=1,\dots,T}\max_{k,k'\in\mathcal{K}_t}\big|\ell_t(k)-\ell_t(k')\big|\leqslant E\qquad\text{and}\qquad\max_{t=1,\dots,T}\max_{k\in\mathcal{K}_t}\big|\,\mathrm{pen}_t(k)\big|\leqslant F\,,$$

*then, for all $T\geqslant 1$,*

$$\begin{aligned}&\max_{j=1,\dots,K}\mathbb{E}\left[\sum_{t=1}^{T} p_t\cdot\big(\ell_t+\mathrm{pen}_t\big)-\sum_{t=1}^{T}\xi_t(j,\cdot)\cdot\big(\ell_t+\mathrm{pen}_t\big)\right]\\ &\qquad\leqslant\frac{\ln N}{\eta}+4\eta TF^2+4\eta E^2\sum_{t=1}^{T}\mathbb{E}\left[\sum_{k=1}^{K}\frac{p_t(k)}{P_t^*(k)}\right].\end{aligned}$$

**Proof** From the definition of $\mathcal{K}_t$ and because $E$ and $F$ are upper bounds on the losses and penalties associated with actions in $\mathcal{K}_t$,

$$\min_{j=1,\dots,N}\widetilde{\ell}_t(j)=\min_{j=1,\dots,N}\xi_t(j,\cdot)\cdot\widehat{\ell}_t\geqslant\min_{k\in\mathcal{K}_t}\widehat{\ell}_t(k)\geqslant 0\,.$$

We can thus use the regret bound for Hedge with weight vectors $q_t$, constant learning rate $\eta$, and nonnegative losses $\widetilde{\ell}_t(k)$ (see Lemma 8 in Appendix A):

$$\sum_{t=1}^{T} q_t \cdot \widetilde{\ell}_t \leqslant \min_{j=1,\dots,N} \sum_{t=1}^{T} \widetilde{\ell}_t(j) + \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{j=1}^{N} q_t(j) \widetilde{\ell}_t(j)^2 \,. \tag{14}$$

Note that $q_t \cdot \widetilde{\ell}_t = p_t \cdot \widehat{\ell}_t$. Moreover, $K_t = \max \mathcal{K}_t$ and $\mathrm{pen}_t$ are measurable with respect to $(I_1, \dots, I_{t-1})$, implying $\mathbb{E}_{t-1}\big[\widehat{\ell}_t(k)\big] = \ell_t(k) - \ell_t(K_t) + E + \mathrm{pen}_t(k) + F$. Therefore, taking the expectation on both sides of (14) with respect to the random draw of $I_1, \dots, I_T$, we get

$$\begin{aligned} &\max_{j=1,\dots,N} \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \big(\ell_t + \mathrm{pen}_t\big) - \sum_{t=1}^{T} \xi_t(j,\cdot) \cdot \big(\ell_t + \mathrm{pen}_t\big)\right] \\ &\qquad \leqslant \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{N} q_t(j) \Big(\xi_t(j,\cdot) \cdot \widehat{\ell}_t\Big)^2\right] . \end{aligned} \tag{15}$$

The variance term can be upper bounded as follows. Using Jensen's inequality,

$$\begin{aligned} \sum_{j=1}^{N} q_t(j) \Big(\xi_t(j,\cdot) \cdot \widehat{\ell}_t\Big)^2 &\leqslant \sum_{j=1}^{N} q_t(j) \sum_{k=1}^{K} \xi_t(j,k) \widehat{\ell}_t(k)^2 \\ &= \sum_{k=1}^{K} \sum_{j=1}^{N} q_t(j) \xi_t(j,k) \left(\frac{\ell_t(k) - \ell_t(K_t) + E}{P_t^*(k)} \mathbb{I}_{I_t \leqslant k} + \mathrm{pen}_t(k) + F\right)^2 \\ &\leqslant 8F^2 + 8E^2 \sum_{k=1}^{K} \frac{p_t(k)}{P_t^*(k)^2} \mathbb{I}_{I_t \leqslant k} \,, \end{aligned}$$

where the last inequality is because $\big|\ell_t(k) - \ell_t(K_t)\big| \leqslant E$ and $\big|\,\mathrm{pen}_t(k)\big| \leqslant F$ when $\xi_t(j,k) > 0$ and because $(a+b+c+d)^2 \leqslant 4\big(a^2+b^2+c^2+d^2\big)$. Therefore, recalling $P_t^*(k) = \mathbb{P}_{t-1}(I_t \leqslant k)$,

$$\mathbb{E}_{t-1}\left[\sum_{j=1}^{N} q_t(j) \big(\xi_t(j,k) \cdot \widehat{\ell}_t\big)^2\right] \leqslant 8F^2 + 8E^2 \sum_{k=1}^{K} \frac{p_t(k)}{P_t^*(k)} \,.$$

Substituting the last bound in (15) concludes the proof. ∎

Note that when $\mathrm{pen}_t \equiv 0$ and $p_t^*(1) \geqslant \gamma$ almost surely for all $t \geqslant 1$, Theorem 11 above used with $\eta = (2E)^{-1}\sqrt{\gamma \ln(N)/T}$ yields a regret bound of $4E\sqrt{T \ln(N)/\gamma}$, similarly to Theorem 10.

We use yet another corollary in Section 4.3. It follows directly from the choice of $\mathrm{pen}_t(k) = -\alpha/P_t^*(k)$ and $F = \alpha/\gamma$ in Theorem 11.

**Corollary 12** *Let $1 < \cdots < K$ be a total order on the action set $\mathcal{K} = \{1, \dots, K\}$. Let $E, \alpha, \gamma > 0$ be three parameters. Suppose Exp4 (Algorithm 4) is run on an arbitrary loss sequence $\ell_t \in \mathbb{R}^K$, $t \geqslant 1$, with loss estimates*

$$\widehat{\ell}_t(k) = \frac{\ell_t(k) - \ell_t\big(\max \mathcal{K}_t\big) + E}{P_t^*(k)} \mathbb{I}_{I_t \leqslant k} - \frac{\alpha}{P_t^*(k)} + \frac{\alpha}{\gamma} \qquad \text{for } k = 1, \dots, K$$

*and constant learning rate* $\eta > 0$. *If we have, for all* $t \geqslant 1$, *almost surely,*

$$\max_{t=1,\dots,T} \max_{k,k' \in \mathcal{K}_t} \left|\ell_t(k) - \ell_t(k')\right| \leqslant E \qquad \textit{and} \qquad p_t^*(1) \geqslant \gamma\,,$$

*then, for all* $T \geqslant 1$,

$$\max_{j=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} p_t(k)\left(\ell_t(k) - \frac{\alpha + 4\eta E^2}{P_t^*(k)}\right) - \sum_{t=1}^{T}\sum_{k=1}^{K}\xi_t(j,k)\left(\ell_t(k) - \frac{\alpha}{P_t^*(k)}\right)\right] \leqslant \frac{\ln N}{\eta} + \frac{4\eta T\alpha^2}{\gamma^2}\,.$$

## Appendix F. Missing Proofs from Section 4.2

We prove Theorem 4 and Corollary 5. As we said in the main text, the key contribution of chaining is that it allows us to sum up local regret bounds scaling as the range of the local losses. This is possible because of the richer feedback structure induced by the total order on the actions. We first state a lemma indicating that the losses associated with neighboring nodes are indeed close to one another. Recall that $M$ is the depth of $\mathcal{T}_{\mathcal{F}}$.

**Lemma 13** *Let* $v \in \mathcal{T}$ *be any node at level* $m \in \{0, 1, \dots, M-1\}$. *Then all leaves* $w, w' \in \mathcal{L}_v$ *satisfy* $\left\|f_w - f_{w'}\right\|_\infty \leqslant 2^{-m+2}$. *Therefore,* $\left|\ell_t(y_{i_t(w)}) - \ell_t(y_{i_t(w')})\right| \leqslant 2^{-m+3}$ *for all* $t \geqslant 1$.

**Proof** Consider a path $v = v_m \to v_{m+1} \to \cdots \to v_M = w$ joining $v$ to leaf $w$ in the tree. For each $k = m, \dots, M-1$, since $\mathcal{F}_k$ is a $(2^{-k})$-covering of $\mathcal{F}$ in sup norm, we have $\|f_{v_k} - f_{v_{k+1}}\|_\infty \leqslant 2^{-k}$. Therefore,

$$\|f_v - f_w\|_\infty = \left\|\sum_{k=m}^{M-1}(f_{v_k} - f_{v_{k+1}})\right\|_\infty \leqslant \sum_{k=m}^{M-1}\|f_{v_k} - f_{v_{k+1}}\|_\infty \leqslant \sum_{k=m}^{M-1} 2^{-k} \leqslant 2^{-m+1}\,.$$

Therefore, since $w' \in \mathcal{L}_v$ as well, $\|f_w - f_{w'}\|_\infty = \|f_v - f_w\|_\infty + \|f_v - f_{w'}\|_\infty \leqslant 2^{-m+2}$, which proves the first inequality. Now, since $\ell_t$ is 1-Lipschitz,

$$\begin{aligned}\left|\ell_t(y_{i_t(w)}) - \ell_t(y_{i_t(w')})\right| &\leqslant \left|y_{i_t(w)}, y_{i_t(w')}\right| \leqslant \left(2^{-M+1} + \|f_w - f_{w'}\|_\infty\right) \\ &\leqslant \left(2^{-M+1} + 2^{-m+2}\right) \leqslant 2^{-m+3}\end{aligned}$$

where the second inequality uses the definition of $i_t(v)$ in Section 4.2 and the fact that $\mathcal{K}$ is a $(2^{-M})$-covering of $[0,1]$. This concludes the proof. ∎

We are now ready to prove Theorem 4 from the main text.

**Proof (of Theorem 4)** Each node $v$ at level $m = 0, \dots, M-1$ is running an instance of the variant of Exp4 described in Algorithm 4 (Appendix E) over expert set $\mathcal{C}_v$, where the advice of $w \in \mathcal{C}_v$ is $\xi_t(w, \cdot) = p_t(w, \cdot)$, and effective action set $\mathcal{K}_t = \{i : (\exists w \in \mathcal{C}_v)\ p_t(w,i) > 0\} = \mathcal{K}_t(v)$. Note that, for any $w \in \mathcal{C}_v$, the distribution $p_t(w, \cdot)$ is a mixture of actions in $\mathcal{L}_w \subseteq \mathcal{L}_v$, so that $\mathcal{K}_t(v) \subseteq \{i_t(w') : w' \in \mathcal{L}_v\}$. By Lemma 13, the losses of these actions belong to a range of size $E = 2^{-m+3}$. Since $p_t^*(1) \geqslant \gamma$ by definition, we are in position to apply Theorem 10 in Appendix E with $N = |\mathcal{C}_v| \leq N_{m+1}$, and obtain the bound

$$\max_{w \in \mathcal{C}_v} \mathbb{E}\left[\sum_{t=1}^{T} p_t(v,\cdot)\cdot\ell_t - \sum_{t=1}^{T} p_t(w,\cdot)\cdot\ell_t\right] \leqslant 2^{-m+5}\sqrt{\frac{T\ln N_{m+1}}{\gamma}} + \frac{2^{-m+3}}{\gamma}(4\ln N_{m+1} + 1)$$

(for simplicity, we use $\ell_t$ to denote the vector of elements $\ell_t(y_i)$ for $i = 1, \dots, K$). Now consider the path $v_0 \to v_1 \to \cdots \to v_M = w^*$ from the root $v_0$ to the leaf $v_M = w^*$ minimizing $\ell_1(y_{i_1(w)}) + \cdots + \ell_T(y_{i_T(w)})$ over $w \in \mathcal{L}$. Recalling that $p_t(w, i) = \mathbb{I}_{i=i_t(w)}$ for any leaf $w$, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} p_t(v_0, \cdot) \cdot \ell_t\right] &- \min_{w \in \mathcal{L}} \sum_{t=1}^{T} \ell_t(y_{i_t(w)}) \\
&= \sum_{m=0}^{M-1} \mathbb{E}\left[\sum_{t=1}^{T} p_t(v_m, \cdot) \cdot \ell_t - \sum_{t=1}^{T} p_t(v_{m+1}, \cdot) \cdot \ell_t\right] \\
&\leqslant 2^5 \sum_{m=0}^{M-1} 2^{-m} \left(\sqrt{\frac{T \ln N_{m+1}}{\gamma}} + \frac{1}{\gamma}(\ln N_{m+1} + 1)\right) . \qquad (16)
\end{aligned}$$

Now, recalling that $I_t \sim p_t^*$ and $p_t^*(i)\ell_t(y_i) = (1-\gamma)p_t(v_0, i)\ell_t(y_i) + \gamma\, \mathbb{I}_{i=1}$ we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} p_t^* \cdot \ell_t\right] &\leqslant \min_{w \in \mathcal{L}} \left\{\sum_{t=1}^{T} \ell_t\big(y_{i_t(w)}\big) + \gamma \sum_{t=1}^{T} \Big(\ell_t(y_1) - \ell_t\big(y_{i_t(w)}\big)\Big)\right\} \\
&+ 2^5 \sum_{m=0}^{M-1} 2^{-m} \left(\sqrt{\frac{T \ln N_{m+1}}{\gamma}} + \frac{1}{\gamma}(\ln N_{m+1} + 1)\right) .
\end{aligned}$$

Now clearly $\big|\ell_t(y_1) - \ell_t\big(y_{i_t(w)}\big)\big| \leqslant 1$. 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 $\big|\ell_t\big(y_{i_t(w)}\big) - \ell_t\big(f(x_t)\big)\big| \leqslant \big|y_{i_t(w)} - f(x_t)\big| \leqslant 1\big|y_{i_t(w)} - f_w(x_t) + f_w(x_t) - f(x_t)\big| \leqslant 2^{1-M}$ by definition of $i_t(w)$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} p_t^* \cdot \ell_t\right] &- \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_t\big(f(x_t)\big) \\
&\leqslant (2^{1-M} + \gamma)T + 2^5 \sum_{m=0}^{M-1} 2^{-m} \left(\sqrt{\frac{T \ln N_{m+1}}{\gamma}} + \frac{1}{\gamma}(\ln N_{m+1} + 1)\right) . \qquad (17)
\end{aligned}$$

We now use $N_{m+1} = \mathcal{N}_\infty(\mathcal{F}, 2^{-(m+1)})$, and follow the standard chaining approach approximating the sums by integrals,

$$\begin{aligned}
\sum_{m=0}^{M-1} 2^{-m} \sqrt{\ln N_{m+1}} &= 4 \sum_{m=0}^{M-1} \left(2^{-(m+1)} - 2^{-(m+2)}\right) \sqrt{\ln \mathcal{N}_\infty(\mathcal{F}, 2^{-(m+1)})} \\
&\leqslant 4 \sum_{m=0}^{M-1} \int_{2^{-(m+2)}}^{2^{-(m+1)}} \sqrt{\ln \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} \mathrm{d}\varepsilon \leqslant 4 \int_{\gamma/2}^{1/2} \sqrt{\ln \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} \mathrm{d}\varepsilon \qquad (18)
\end{aligned}$$

where the second inequality is by monotonicity of $\varepsilon \mapsto \ln \mathcal{N}_\infty(\mathcal{F}, \varepsilon)$ and the last inequality follows from $\gamma \leqslant 2^{-M}$ due to $M = \lfloor \log_2(1/\gamma) \rfloor$. Similarly,

$$\sum_{m=0}^{M-1} 2^{-m} \ln N_{m+1} \leqslant 4 \int_{\gamma/2}^{1/2} \ln \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \mathrm{d}\varepsilon . \qquad (19)$$

We conclude the proof by substituting (18) and (19) into (17), and using $2^{-M} \leqslant 2\gamma$. ■

**Proof (of Corollary 5)** The metric entropy satisfies $\ln \mathcal{N}_\infty(\mathcal{F}, \varepsilon) = \mathcal{O}\big(\varepsilon^{-d}\big)$. Therefore, substituting into the regret bound of Theorem 4 and computing the integrals, the regret satisfies

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant 5T\gamma + \begin{cases} \mathcal{O}\Big(\sqrt{\frac{T}{\gamma}} + \frac{1}{\gamma}\ln\frac{1}{\gamma}\Big) & \text{if } d = 1 \\ \mathcal{O}\Big(\sqrt{\frac{T}{\gamma}}\ln\frac{1}{\gamma} + \gamma^{-2}\Big) & \text{if } d = 2 \\ \mathcal{O}\big(\sqrt{T}\gamma^{(1-d)/2} + \gamma^{-d}\big) & \text{if } d \geqslant 3. \end{cases}$$

Optimizing $\gamma$ for the different choices of $d$ concludes the proof. ■

## Appendix G. Algorithm HierExp4$^\star$ and proof of Theorem 6

### G.1. Algorithm HierExp4$^\star$

We first construct the HierExp4$^\star$ algorithm used in Theorem 6. It is a variant of HierExp4 based on a special hierarchical covering of $\mathcal{F}$ that we first define below. Recall that $\mathcal{F}$ is the set of all 1-Lipschitz functions from $[0,1]^d$ to $[0,1]$.

In the sequel, we use a dyadic discretization of the input space $[0,1]^d$. For each depth $m = 0, 1, \ldots$ we define a partition of $[0,1]^d$ with $2^{md}$ equal cubes of width $2^{-m}$. Each cube $\mathcal{X}_m(\sigma_{1:m}) \subseteq [0,1]^d$ at depth $m$ is indexed by $\sigma_{1:m} = (\sigma_1, \ldots, \sigma_m) \in \{1, \ldots, 2^d\}^m$. Note that the partitions at different depths are nested. We can thus represent them via a tree $\mathcal{P}$ whose root $\emptyset$ is labeled by $[0,1]^d$ and whose each node $\sigma_{1:m}$ at depth $m$ is labeled with the cube $\mathcal{X}_m(\sigma_{1:m})$. The children of node $\sigma_{1:m}$ are the nodes at depth $m+1$ of the form $(\sigma_1, \ldots, \sigma_m, \sigma_{m+1})$ with $\sigma_{m+1} \in \{1, \ldots, 2^d\}$. They correspond to sub-cubes of $\mathcal{X}_m(\sigma_{1:m})$.

**Wavelet-like approximation of $\mathcal{F}$.** Using the above dyadic discretization $\mathcal{P}$, we approximate any $f \in \mathcal{F}$ with piecewise-constant functions $f_M : [0,1]^d \to [-1/2, 3/2]$ of the form

$$f_M(x) = \frac{1}{2} + \sum_{m=1}^{M} \sum_{\sigma_{1:m} \in \{1,\ldots,2^d\}^m} 2^{-m} c_m(\sigma_{1:m})\, \mathbb{I}_{x \in \mathcal{X}_m(\sigma_{1:m})} \quad \text{where } c_m(\sigma_{1:m}) \in \big\{-1, 0, 1\big\}. \tag{20}$$

As shown in the next lemma, the functions $f_M$ form a $\big(2^{-M}\big)$-covering of $\mathcal{F}$ in the sup norm $\|g\|_\infty = \sup_{x \in [0,1]^d} \big|g(x)\big|$. We use the following important property: for any given $x \in [0,1]^d$, there exists a unique $\sigma_{1:M} \in \{1, \ldots, 2^d\}^M$ such that $x \in \mathcal{X}_m(\sigma_{1:m})$ for all $m = 1, \ldots, M$, so that

$$f_M(x) = \frac{1}{2} + \sum_{m=1}^{M} 2^{-m} c_m(\sigma_{1:m})\,. \tag{21}$$

**Lemma 14** *Let $f \in \mathcal{F}$ and $M \geqslant 1$. There exist coefficients $c_m(\sigma_{1:m}) \in \big\{-1, 0, 1\big\}$ such that $f_M$ defined in (20) satisfies $\|f_M - f\|_\infty \leqslant 2^{-M}$.*

**Proof** We denote the center of each cube $\mathcal{X}_m(\sigma_{1:m})$ by $x_m(\sigma_{1:m})$ and prove by induction on $M \geqslant 1$ that there exist coefficients $c_m(\sigma_{1:m})$, $m = 0, \dots, M$, such that

$$\left| f_M\big(x_M(\sigma_{1:M})\big) - f\big(x_M(\sigma_{1:M})\big) \right| \leqslant 2^{-(M+1)} \,. \tag{22}$$

This directly yields the conclusion, since every $x \in \mathcal{X}_M(\sigma_{1:M})$ is $2^{-(M+1)}$-close to the center $x_M(\sigma_{1:M})$ in sup norm and $f$ is 1-Lipschitz, which by (22) and $f_M(x) = f_M\big(x_M(\sigma_{1:M})\big)$ implies

$$\begin{aligned} \left| f_M(x) - f(x) \right| &\leqslant \left| f_M\big(x_M(\sigma_{1:M})\big) - f\big(x_M(\sigma_{1:M})\big) \right| + \left| f\big(x_M(\sigma_{1:M})\big) - f(x) \right| \\ &\leqslant 2^{-(M+1)} + 2^{-(M+1)} = 2^{-M} \,. \end{aligned}$$

We now carry out the induction. For $M = 1$ and $\sigma_1 \in \{1, \dots, 2^d\}$, we set

$$c_1(\sigma_1) \in \operatorname*{arg\,min}_{c \in \{-1,0,1\}} \left| \frac{1}{2} + \frac{c}{2} - f\big(x_1(\sigma_1)\big) \right| \,,$$

which corresponds to projecting the value of $f$ at the center $x_1(\sigma_1)$ onto the coarse grid $\{0, 1/2, 1\}$. Therefore $\left|1/2 + c_1(\sigma_1)/2 - f(x_1(\sigma_1))\right| \leqslant 1/4$, which implies (22) by (21).

Now, let $M \geqslant 2$ and assume that there exist coefficients $c_m(\sigma_{1:m})$, $m = 0, \dots, M-1$, such that (22) holds true at level $M-1$. We complete these coefficients at level $M$ as follows: for all $\sigma_{1:M} \in \{1, \dots, 2^d\}^M$, we set

$$c_M(\sigma_{1:M}) \in \operatorname*{arg\,min}_{c \in \{-1,0,1\}} \left| f_{M-1}\big(x_M(\sigma_{1:M})\big) + \frac{c}{2^M} - f\big(x_M(\sigma_{1:M})\big) \right| \,.$$

Note that by (21) and $x_M(\sigma_{1:M}) \in \mathcal{X}_M(\sigma_{1:M}) \subset \mathcal{X}_{M-1}(\sigma_{1:(M-1)})$, we have

$$f_M\big(x_M(\sigma_{1:M})\big) = f_{M-1}\big(x_M(\sigma_{1:M})\big) + \frac{c_M(\sigma_{1:M})}{2^M} \,. \tag{23}$$

Therefore, the definition of $c_M(\sigma_{1:M})$ above corresponds to making $f_M\big(x_M(\sigma_{1:M})\big)$ as close to $f\big(x_M(\sigma_{1:M})\big)$ as possible. Now, note that $f_{M-1}$ is constant over $\mathcal{X}_{M-1}(\sigma_{1:(M-1)})$, so that the difference $\Delta_{M-1} \triangleq f_{M-1}\big(x_M(\sigma_{1:M})\big) - f\big(x_M(\sigma_{1:M})\big)$ satisfies

$$\begin{aligned} &\left|\Delta_{M-1}\right| = \left| f_{M-1}\big(x_{M-1}(\sigma_{1:(M-1)})\big) - f\big(x_M(\sigma_{1:M})\big) \right| \\ &\leqslant \left| f_{M-1}\big(x_{M-1}(\sigma_{1:(M-1)})\big) - f\big(x_{M-1}(\sigma_{1:(M-1)})\big) \right| + \left| f\big(x_{M-1}(\sigma_{1:(M-1)})\big) - f\big(x_M(\sigma_{1:M})\big) \right| \\ &\overset{(22)}{\leqslant} 2^{-M} + \left\| x_{M-1}(\sigma_{1:(M-1)}) - x_M(\sigma_{1:M}) \right\|_\infty \\ &= 2^{-M} + 2^{-(M+1)} \,, \end{aligned}$$

where the last inequality is by (22) at level $M-1$, and where the last equality is by comparison of two cube centers at subsequent depths.

To conclude the proof, we note that the bound $\left|\Delta_{M-1}\right| \leqslant 2^{-M} + 2^{-(M+1)}$ implies the existence of $c_M(\sigma_{1:M}) \in \{-1, 0, 1\}$ such that $\left| f_M\big(x_M(\sigma_{1:M})\big) - f\big(x_M(\sigma_{1:M})\big) \right| \leqslant 2^{-(M+1)}$, as required to finish the induction. We can indeed consider three possible cases. If $\left|\Delta_{M-1}\right| \leqslant 2^{-(M+1)}$, then setting $c_M(\sigma_{1:M}) = 0$ concludes the proof (by (23)). If $\Delta_{M-1} > 2^{-(M+1)}$, then $\Delta_{M-1}$ lies in the interval $\left[2^{-(M+1)}, 2^{-M} + 2^{-(M+1)}\right]$, so that setting $c_M(\sigma_{1:M}) = -1$ also concludes the proof (using (23) again). Similarly we set $c_M(\sigma_{1:M}) = 1$ when $\Delta_{M-1} < -2^{-(M+1)}$. ■

We are now ready to define our hierarchical covering of $\mathcal{F}$.

**Construction of $\mathcal{T}_{\mathcal{F}}^{\star}$.** We build a tree $\mathcal{T}_{\mathcal{F}}^{\star}$ such that the value of any function $f_M$ at any point $x \in [0,1]^d$ —as given by (21)— can be computed by following a path from the root to a leaf of $\mathcal{T}_{\mathcal{F}}^{\star}$. Unlike Section 4.2, this tree is not labeled by functions but instead by either values in $\mathbb{R}$ or cubes $\mathcal{X}_m(\sigma_{1:m}) \subseteq [0,1]^d$, as illustrated in Figure 3. More precisely, our tree $\mathcal{T}_{\mathcal{F}}^{\star}$ is composed of two types of nodes:

- *Nodes at depths* $m$. The root $v_0 = \emptyset$ (at depth $m = 0$) is labeled by $1/2$. Each node at depth $m = 1, 2, \dots$ is indexed by a tuple $(\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m)$ with $\sigma_k \in \{1, \dots, 2^d\}$ and $c_k \in \{-1, 0, 1\}$; the node is labeled by $1/2 + \sum_{k=1}^{m} 2^{-k} c_k$, which corresponds to (21).
- *Nodes at depths* $m + 1/2$. Each node at depth $m + 1/2$, $m \geqslant 0$, is indexed by a tuple $(\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m, \sigma_{m+1})$ with $\sigma_k \in \{1, \dots, 2^d\}$ and $c_k \in \{-1, 0, 1\}$; the node is labeled by the cube $\mathcal{X}_m(\sigma_{1:m+1})$.

We connect nodes in the following natural way. Nodes $v = (\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m)$ at depth $m$ are connected to all nodes at depth $m + 1/2$ of the form $(v, \sigma_{m+1})$, with $\sigma_{m+1} \in \{1, \dots, 2^d\}$. Nodes $w = (\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m, \sigma_{m+1})$ at depth $m + 1/2$ are connected to all nodes at depth $m + 1$ of the form $(w, c_{m+1})$, with $c_{m+1} \in \{-1, 0, 1\}$.

We stop the tree at depth $M \geqslant 1$. As previously, we denote by $\mathcal{L}$ the set of all the leaves of $\mathcal{T}_{\mathcal{F}}^{\star}$, by $\mathcal{L}_v$ the set of all the leaves under $v \in \mathcal{T}_{\mathcal{F}}^{\star}$, and by $\mathcal{C}_v$ the set of children of $v \in \mathcal{T}_{\mathcal{F}}^{\star}$.

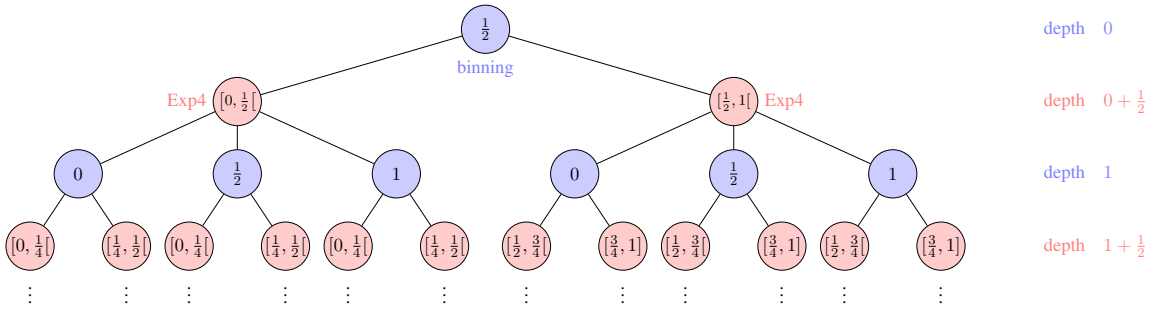


Figure 3: The structure of the efficient chaining tree in dimension $d = 1$.

**Algorithm HierExp4$^{\star}$.** Our new bandit algorithm HierExp4$^{\star}$ (Algorithm 5) is an efficient modification of HierExp4 based on the tree $\mathcal{T}_{\mathcal{F}}^{\star}$ defined above. HierExp4$^{\star}$ alternates between "binning nodes" at depths $m$ and "Exp4 nodes" at depths $m + 1/2$ (see Figure 3). We discretize the action space $\mathcal{Y} = [0,1]$ with $\mathcal{K} = \{k2^{-M} : k = 1, \dots, 2^M\}$ of cardinality $K = 2^M$. Note that, after projecting onto $[2^{-M}, 1]$, the set of labels of all the leaves $v \in \mathcal{L}$ exactly corresponds to this discretization, i.e.,

$$\left\{ \min\left( \max\left( \frac{1}{2} + \sum_{m=1}^{M} 2^{-m} c_m,\ \frac{1}{2^M} \right),\ 1 \right) : c_m \in \{-1, 0, 1\} \right\} = \left\{ \frac{k}{2^M} : k = 1, \dots, 2^M \right\} = \mathcal{K}\,.$$

At every round $t$, after observing context $x_t \in [0,1]^d$, the algorithm activates the root $v_0$ and all nodes $\big(\sigma_1, c_1, \dots, \sigma_m(, c_m)\big)$ such that $x_t \in \mathcal{X}_m(\sigma_{1:m})$; the other nodes are asleep during that round. In particular, there are $3^M$ activated leaves $v = (\sigma_1, c_1, \dots, \sigma_M, c_M) \in \mathcal{L}$. Each such leaf

recommends the action $i_t(v) \in \{1, \dots, K\}$ matching its label after projection onto $\left[2^{-M}, 1\right]$, i.e., such that

$$\frac{i_t(v)}{2^M} = \min\left(\max\left(\frac{1}{2} + \sum_{m=1}^{M} 2^{-m} c_m,\ \frac{1}{2^M}\right),\ 1\right). \tag{24}$$

Each node $v \in \mathcal{T}_{\mathcal{F}}^{\star}$ maintains a probability distribution $p_t(v, \cdot) \in \Delta(K)$ over actions in $\mathcal{K}$, which is only used and updated when $v$ is activated. The activated leaves $v \in \mathcal{L}$ correspond to unit masses $p_t(v, i) = \mathbb{I}_{i=i_t(v)}$. As mentioned earlier, the internal nodes $v \in \mathcal{T}_{\mathcal{F}}^{\star} \setminus \mathcal{L}$ are of two types:

- Each internal node $v = (\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m)$ at depth $m = 0, \dots, M-1$ is a "binning node". When activated, it identifies its only child $w = (v, \sigma_{m+1})$ handling $x_t$ (i.e., such that $x_t \in \mathcal{X}_{m+1}(\sigma_{1:(m+1)})$) and outputs $p_t(v, \cdot) = p_t(w, \cdot)$. We denote by $\mathcal{T}_{\text{bin}} \subset \mathcal{T}_{\mathcal{F}}^{\star} \setminus \mathcal{L}$ the set of "binning nodes".

- Each internal node $v = (\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m, \sigma_{m+1})$ at depth $m + 1/2$ is an "Exp4 node". It runs an instance of Exp4 using the children of $v$ as experts. The variant of Exp4 we use is Algorithm 4 (see Appendix E) applied to penalized loss estimates (cf. (2)), which adapts to the range of the losses and allows for a careful control of the variance terms along the tree. Let $\text{Exp4}_v$ be the instance of the Exp4 variant run at node $v$. On all the rounds $t$ when $v$ is activated, this instance updates a distribution $q_t(v, \cdot) \in \Delta(|\mathcal{C}_v|)$ over experts in $\mathcal{C}_v$ and outputs $p_t(v, i) = \sum_{w \in \mathcal{C}_v} q_t(v, w) p_t(w, i)$ for all actions $i = 1, \dots, K$. We denote by $\mathcal{T}_{\text{Exp4}} \subset \mathcal{T}_{\mathcal{F}}^{\star} \setminus \mathcal{L}$ the set of "Exp4 nodes".

Note that the root $v_0$ of $\mathcal{T}_{\mathcal{F}}^{\star}$ is activated at all rounds $t$. The prediction of HierExp4$^\star$ at time $t$ is $\widehat{y}_t = I_t 2^{-M} \in \mathcal{K}$, where $I_t$ is drawn according to a mixture of $p_t(v_0, \cdot)$ and a unit mass on the minimal action $i = 1$. A pseudo-code for HierExp4$^\star$ can be found in Algorithm 5.

### G.2. Proof of Theorem 6

As for the proof of Theorem 4, we start by stating a lemma indicating that the losses associated with neighboring leaves are close to one another.

**Lemma 15** *Let $v \in \mathcal{T}_{\text{Exp4}}$, with depth $m + 1/2$, $m \in \{0, \dots, M-1\}$. Then, all leaves $w, w' \in \mathcal{L}_v$ in the subtree rooted at $v$ satisfy*

$$\left| \ell_t\left(\frac{i_t(w)}{2^M}\right) - \ell_t\left(\frac{i_t(w')}{2^M}\right) \right| \leqslant 2^{1-m}$$

*at all rounds $t \geqslant 1$ when $w$ and $w'$ are both activated.*

**Proof** Denote $w = (\sigma_1, c_1, \dots, \sigma_M, c_M)$ and $w' = (\sigma'_1, c'_1, \dots, \sigma'_M, c'_M)$. First, we remark that since $w$ and $w'$ have the common ancestor $v$ at level $m$, we have $\sigma_i = \sigma'_i$ and $c_i = c'_i$ for all $i \leqslant m$, as well as $\sigma_{m+1} = \sigma'_{m+1}$ (recall that $v \in \mathcal{T}_{\text{Exp4}}$). Thus,

$$\left| \sum_{j=1}^{M} 2^{-j} c_j - \sum_{j=1}^{M} 2^{-j} c'_j \right| \leqslant \sum_{j=m+1}^{M} 2^{-j} |c_j - c'_j| \leqslant \sum_{j=m+1}^{M} 2^{-j+1} \leqslant 2^{1-m}.$$

**Algorithm 5:** HierExp4$^\star$ (for the one-sided full information feedback)

**Input** : Tree $\mathcal{T}_{\mathcal{F}}^\star$ with root $v_0$ and leaves $\mathcal{L}$, exploration parameter $\gamma \in (0,1)$, penalization parameters $\alpha_0, \dots, \alpha_{M-1} > 0$, learning rates $\eta_0, \dots, \eta_{M-1} > 0$.

**Initialization**: Set $q_1(v,\cdot)$ to the uniform distribution in $\Delta(|\mathcal{C}_v|)$ for every $v \in \mathcal{T}_{\text{Exp4}}$.

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

1. Get context $x_t \in \mathcal{X}$;
2. Activate $v_0$ and all nodes $v = \big(\sigma_1, c_1, \dots, \sigma_m(, c_m)\big) \in \mathcal{T}_{\mathcal{F}}^\star$ such that $x_t \in \mathcal{X}_m(\sigma_{1:m})$;
3. Set $p_t(v,i) = \mathbb{I}_{i=i_t(v)}$ for all $i \in \mathcal{K}$ and all activated leaves $v \in \mathcal{L}$;
4. Set $p_t(v,i) = q_t(v,\cdot) \cdot p_t(\cdot,i)$ for all $i \in \mathcal{K}$ and all activated nodes $v \in \mathcal{T}_{\text{Exp4}}$;
5. Set $p_t(v,\cdot) = p_t(w,\cdot)$ for all activated nodes $v \in \mathcal{T}_{\text{bin}}$, where $w \in \mathcal{C}_v$ is the unique activated child of $v$;
6. Draw $I_t \sim p_t^*$ and play $\widehat{y}_t = I_t 2^{-M}$, where $p_t^*(i) = (1-\gamma)p_t(v_0,i) + \gamma \mathbb{I}_{i=1}$ for all $i \in \mathcal{K}$;
7. Observe $\ell_t(y)$ for all $y \geqslant \widehat{y}_t$;
8. For all $m = 0, \dots, M-1$ and all activated nodes $v \in \mathcal{T}_{\text{Exp4}}$ at level $m + 1/2$, with $v = (\sigma_1, c_1, \sigma_2, c_2, \dots, \sigma_m, c_m, \sigma_{m+1})$, compute the loss estimate for each $i \in \mathcal{K}_t(v)$,

$$\widehat{\ell}_t(v,i) = \frac{\ell_t(i/2^M) - \ell_t\big(j_t(v)/2^M\big) + 2^{1-m}}{\sum_{k=1}^i p_t^*(k)} \mathbb{I}_{I_t \leqslant i} - \frac{\alpha_m}{\sum_{k=1}^i p_t^*(k)} + \frac{\alpha_m}{\gamma}\,, \tag{25}$$

where $\mathcal{K}_t(v) = \{i : (\exists w \in \mathcal{C}_v)\ p_t(w,i) > 0\}$ and $j_t(v) = \max \mathcal{K}_t(v)$.

Then, 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

$$q_{t+1}(v,w) = \frac{\exp\left(-\eta_m \sum_{s=1}^t \widetilde{\ell}_s(v,w) \mathbb{I}_{x_s \in \mathcal{X}_{m+1}(\sigma_{1:(m+1)})}\right)}{\sum_{w' \in \mathcal{C}_v} \exp\left(-\eta_m \sum_{s=1}^t \widetilde{\ell}_s(v,w') \mathbb{I}_{x_s \in \mathcal{X}_{m+1}(\sigma_{1:(m+1)})}\right)}\,. \tag{26}$$

**end**

Therefore, by definition (24) of $i_t(w)$ and $i_t(w')$, and since projecting two real numbers onto $\big[2^{-M}, 1\big]$ can only reduce their distance, we get, when $w$ and $w'$ are both activated,

$$\left|\frac{i_t(w)}{2^M} - \frac{i_t(w')}{2^M}\right| \leqslant \left|\frac{1}{2} + \sum_{j=1}^M 2^{-j} c_j - \frac{1}{2} - \sum_{j=1}^M 2^{-j} c'_j\right| \leqslant 2^{1-m}\,,$$

which implies the result since $\ell_t$ is 1-Lipschitz. ∎

We are now ready to prove Theorem 6 from the main text.

**Proof (of Theorem 6)** Let $f \in \mathcal{F}$. By Lemma 14, we can fix $f_M : [0,1]^d \to \big[-1/2, 3/2\big]$ of the form (20), with coefficients $c_m(\sigma_{1:m}) \in \big\{-1, 0, 1\big\}$, and such that $\|f_M - f\|_\infty \leqslant 2^{-M}$.

In the proof of Theorem 4, we rewrite the regret with respect to $f_M$ as the sum of $M$ regrets along the path joining the root of $\mathcal{T}_{\mathcal{F}}$ to the leaf corresponding to $f_M$. Next, we proceed similarly except that now $f_M$ corresponds to a $(2^d)$-ary subtree of $\mathcal{T}_{\mathcal{F}}^\star$ indexed by $\sigma_{1:M} \in \{1, \dots, 2^d\}^M$. The

$2^{dM}$ leaves $\big(\sigma_1, c_1(\sigma_1), \sigma_2, c_2(\sigma_{1:2}), \dots, \sigma_M, c_M(\sigma_{1:M})\big)$ are labeled with the values of $f_M$ (before projecting onto $\big[2^{-M}, 1\big]$) over the cubes $\mathcal{X}_M(\sigma_{1:M})$.

We control the regret of HierExp4$^\star$ with respect to $f_M$ by starting from the root $v_0$ of $\mathcal{T}_{\mathcal{F}}^\star$ and by progressively moving down the tree. We write $\mathbb{P}_{t-1}$ and $\mathbb{E}_{t-1}$ for conditioning on $I_1, \dots, I_{t-1}$, and set $P_t^*(k) \triangleq \sum_{i=1}^k p_t^*(i) = \mathbb{P}_{t-1}(I_t \leqslant k)$. In the sequel, we repeatedly use Corollary 12 with parameters $E_m \triangleq 2^{1-m}$, $\alpha_m$ and $\gamma$ to control the regret on each "Exp4 node". This corollary applies to loss estimates penalized with $\mathrm{pen}_t(k) = -\alpha_m / P_t^*(k)$ for "Exp4 nodes" at depth $m + 1/2$, with $m = 0, \dots, M-1$.

Since the root $v_0$ is a "binning node", we have $p_t(v_0, k) = p_t(\sigma_1, k)$ at all rounds $t$ such that $x_t \in \mathcal{X}_1(\sigma_1)$. Therefore, writing $\ell_t(k)$ instead of $\ell_t\big(k/2^M\big)$,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \sum_{k=1}^K p_t(v_0,k)\left(\ell_t(k) - \frac{\alpha_0 + 4\eta_0 E_0^2}{P_t^*(k)}\right)\right] \\
&= \sum_{\sigma_1 \in \{1,\dots,2^d\}} \mathbb{E}\left[\sum_{t:\, x_t \in \mathcal{X}_1(\sigma_1)} \sum_{k=1}^K p_t(\sigma_1,k)\left(\ell_t(k) - \frac{\alpha_0 + 4\eta_0 E_0^2}{P_t^*(k)}\right)\right] \\
&\leqslant \sum_{\sigma_1 \in \{1,\dots,2^d\}} \left(\min_{c_1 \in \{-1,0,1\}} \mathbb{E}\left[\sum_{t:\, x_t \in \mathcal{X}_1(\sigma_1)} \sum_{k=1}^K p_t\big((\sigma_1,c_1),k\big)\left(\ell_t(k) - \frac{\alpha_0}{P_t^*(k)}\right)\right]\right. \\
&\qquad\qquad \left. + \frac{\ln 3}{\eta_0} + \frac{4\eta_0 T(\sigma_1)\alpha_0^2}{\gamma^2}\right) \qquad (27) \\
&\leqslant \sum_{\sigma_1} \mathbb{E}\left[\sum_{t:\, x_t \in \mathcal{X}_1(\sigma_1)} \sum_{k=1}^K p_t\big([\sigma_1, c_1(\sigma_1)],k\big)\left(\ell_t(k) - \frac{\alpha_0}{P_t^*(k)}\right)\right] + \frac{2^d \ln 3}{\eta_0} + \frac{4\eta_0 T \alpha_0^2}{\gamma^2}\,, \qquad (28)
\end{aligned}
$$

where (27) follows from Corollary 12 with parameters $E_0 = 2$, $\alpha_0$ and $\gamma$, and where $T(\sigma_1)$ is the number of rounds $t$ such that $x_t \in \mathcal{X}_1(\sigma_1)$. Indeed, the probability distributions $p_t\big((\sigma_1, c_1), \cdot\big)$ are mixtures of actions $i_t(w)$ associated with activated leaves $w \in \mathcal{L}_{\sigma_1}$, whose losses are at most at distance 2 by Lemma 15 with $m = 0$. The last inequality (28) above is obtained by choosing $c_1 = c_1(\sigma_1)$.

So far we have showed how to move from depth $m = 0$ to depth 1. Repeating the same argument at all subsequent depths $m = 1, \dots, M-1$, and noting that $\alpha_{m-1} = \alpha_m + 4\eta_m 2^{2-2m} =$

$\alpha_m + 4\eta_m E_m^2$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} p_t(v_0,k)\left(\ell_t(k) - \frac{\alpha_0 + 4\eta_0 E_0^2}{P_t^*(k)}\right)\right] \\
&\leqslant \sum_{\sigma_{1:M}} \mathbb{E}\left[\sum_{t:\, x_t \in \mathcal{X}_1(\sigma_{1:M})} \sum_{k=1}^{K} p_t\big([\sigma_1, c_1(\sigma_1), \ldots, \sigma_M, c_M(\sigma_{1:M})], k\big)\left(\ell_t(k) - \frac{\alpha_{M-1}}{P_t^*(k)}\right)\right] \\
&\qquad + \sum_{m=0}^{M-1}\left(\frac{2^{(m+1)d}\ln 3}{\eta_m} + \frac{4\eta_m T \alpha_m^2}{\gamma^2}\right) \\
&\leqslant \sum_{\sigma_{1:M}}\left(\sum_{t:\, x_t \in \mathcal{X}_1(\sigma_{1:M})} \ell_t\big(i_t(v)\big)\right) + \sum_{m=0}^{M-1}\left(\frac{2^{(m+1)d}\ln 3}{\eta_m} + \frac{4\eta_m T \alpha_m^2}{\gamma^2}\right), \qquad (29)
\end{aligned}$$

where in the last inequality $v$ denotes the leaf $\big(\sigma_1, c_1(\sigma_1), \ldots, \sigma_M, c_M(\sigma_{1:M})\big)$, whose probability distribution $p_t(v,\cdot)$ is the unit mass on its action $i_t(v)$.

We define $\widetilde{f}_M(x) \triangleq \min\big(\max\big(f_M(x), 2^{-M}\big), 1\big)$ to be the projection of $f_M(x)$ onto $\big[2^{-M}, 1\big]$, and, similarly, $\widetilde{f}(x) \triangleq \min\big(\max\big(f(x), 2^{-M}\big), 1\big)$. Since projection can only reduce the distance between two points, $\big|\widetilde{f}_M(x) - \widetilde{f}(x)\big| \leqslant \big|f_M(x) - f(x)\big| \leqslant \|f_M - f\|_\infty \leqslant 2^{-M}$. But, noting that $\widetilde{f}(x)$ is $2^{-M}$-close to $f(x) \in [0,1]$, this yields $\big|\widetilde{f}_M(x) - f(x)\big| \leqslant \big|\widetilde{f}_M(x) - \widetilde{f}(x)\big| + \big|\widetilde{f}(x) - f(x)\big| \leqslant 2^{1-M}$. Therefore, since $\ell_t$ is 1-Lipschitz, and using both (21) and the definition (24) of $i_t(v)$, we get that, when the leaf $v = \big(\sigma_1, c_1(\sigma_1), \ldots, \sigma_M, c_M(\sigma_{1:M})\big)$ is activated,

$$\ell_t\big(i_t(v)\big) = \ell_t\big(\widetilde{f}_M(x_t)\big) \leqslant \ell_t\big(f(x_t)\big) + \big|\widetilde{f}_M(x_t) - f(x_t)\big| \leqslant \ell_t\big(f(x_t)\big) + 2^{1-M}\,.$$

We plug the last bound into (29) and multiply the resulting inequality by $1-\gamma$. Then, using $\mathbb{E}_{t-1}\big[\ell_t\big(\widehat{y}_t\big)\big] = (1-\gamma)p_t(v_0,\cdot)\cdot\ell_t + \gamma\ell_t(1)$, together with $\ell_t(y) \in [0,1]$, we obtain

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{F}) \leqslant \sum_{m=0}^{M-1}\left(\frac{2^{(m+1)d}\ln 3}{\eta_m} + \frac{4\eta_m T\alpha_m^2}{\gamma^2}\right) + (1-\gamma)\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} p_t(v_0,k)\frac{\alpha_0 + 4\eta_0 E_0^2}{P_t^*(k)}\right] \\
+ \gamma T + 2^{1-M}T\,. \qquad (30)
\end{aligned}$$

Now, we use that by definition $p_t^*(k) = (1-\gamma)p_t(v_0,k) + \gamma\mathbb{I}_{k=1} \geqslant (1-\gamma)p_t(v_0,k)$. Therefore, similarly to the analysis of Exp3-RTB, we can control the variance term

$$\begin{aligned}
(1-\gamma)\sum_{k=1}^{K}\frac{p_t(v_0,k)}{P_t^*(k)} &\leqslant \sum_{k=1}^{K}\frac{p_t^*(k)}{P_t^*(k)} = 1 + \sum_{k=2}^{K}\frac{P_t^*(k) - P_t^*(k-1)}{P_t^*(k)} = 1 + \sum_{k=2}^{K}\int_{P_t^*(k-1)}^{P_t^*(k)}\frac{\mathrm{d}x}{P_t^*(k)} \\
&\leqslant 1 + \sum_{k=2}^{K}\int_{P_t^*(k-1)}^{P_t^*(k)}\frac{\mathrm{d}x}{x} = 1 + \int_{P_t^*(1)}^{1}\frac{\mathrm{d}x}{x} \leqslant 1 - \ln P_t^*(1) \leqslant 1 + \ln\frac{1}{\gamma} = \ln\frac{e}{\gamma}\,,
\end{aligned}$$

where we used $P_t^*(1) = p_t^*(1) \geqslant \gamma$. Hence, substituting into the previous bound (30) and recalling that $E_0 = 2$, we get

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant \sum_{m=0}^{M-1}\left(\frac{2^{(m+1)d}\ln 3}{\eta_m} + \frac{4\eta_m T\alpha_m^2}{\gamma^2}\right) + (\alpha_0 + 16\eta_0)T\ln\left(\frac{e}{\gamma}\right) + \gamma T + 2^{1-M}T\,. \quad (31)$$

**Optimization of the parameters.** In the sequel, we show that the prescribed choices of $M$, $\gamma$, $(\alpha_m)$, and $(\eta_m)$ lead to the stated regret bound. Before that, we explain these particular choices by approximately optimizing (31), which will lead to (32) and (33). For the sake of readability, we will sometimes write $f \lesssim g$ instead of $f = \mathcal{O}(g)$. First, recall that $\alpha_{m-1} = \alpha_m + 4\eta_m E_m^2 = \alpha_m + 2^{4-2m}\eta_m$. We thus make the approximation $\alpha_m \approx 2^{4-2m}\eta_m$, and start by optimizing in $\eta_m$ the terms inside the sum appearing in (31). To do so, we equalize the terms to get the equality

$$\frac{2^{(m+1)d}\ln 3}{\eta_m} = \frac{4\eta_m T\alpha_m^2}{\gamma^2}$$

which we approximate, using the previous remark, by

$$\frac{2^{md}}{\eta_m} \approx \frac{T2^{-4m}\eta_m^3}{\gamma^2} \quad \text{hence our choice} \quad \eta_m = c2^{m(\frac{d}{4}+1)}\gamma^{\frac{1}{2}}T^{-\frac{1}{4}}\,, \tag{32}$$

where $c > 0$ will be optimized by the analysis. Now, from these values of $\eta_m$, we can compute $\alpha_m$ by choosing $\alpha_M = 0$ and using the recursion $\alpha_{m-1} = \alpha_m + 2^{4-2m}\eta_m$. This yields

$$\alpha_m = \sum_{j=m+1}^{M} 2^{4-2j}\eta_j = c2^4\gamma^{\frac{1}{2}}T^{-\frac{1}{4}} \sum_{j=m+1}^{M} 2^{j(\frac{d}{4}-1)}\,. \tag{33}$$

We now upper bound (31) by distinguishing between the two cases $1 \leqslant d \leqslant 4$ and $d > 4$.

- If $d \leqslant 4$, then the terms inside the sum in (33) are non-increasing, so that

$$\alpha_m \leqslant c2^4 M_1\gamma^{\frac{1}{2}}T^{-\frac{1}{4}}2^{m(\frac{d}{4}-1)}\,, \tag{34}$$

  where $M_1 = M$ if $2 \leqslant d \leqslant 4$ and $M_1 = 2 \geqslant 1/\big(2^{1-d/4}-1\big)$ if $d = 1$. The sum in the right-hand side of (31) then becomes, by substituting the definition of $\eta_m$ (see (32)) and the above upper bound on $\alpha_m$,

$$\begin{aligned}
\sum_{m=0}^{M-1}\left(\frac{2^{(m+1)d}\ln 3}{\eta_m} + \frac{4\eta_m T\alpha_m^2}{\gamma^2}\right) &\leqslant \sum_{m=0}^{M-1}\left(\frac{2^{m(\frac{3}{4}d-1)+d}T^{\frac{1}{4}}\ln 3}{c\gamma^{\frac{1}{2}}} + \frac{2^{10}c^3M_1^2 2^{m(\frac{3d}{4}-1)}T^{\frac{1}{4}}}{\gamma^{\frac{1}{2}}}\right)\\
&= \Big(\frac{2^d\ln 3}{c} + 2^{10}M_1^2c^3\Big)T^{\frac{1}{4}}\gamma^{-\frac{1}{2}}\sum_{m=0}^{M-1}2^{m(\frac{3}{4}d-1)}\\
&\overset{(*)}{\leqslant} 2^7 M_1^{\frac{1}{2}}T^{\frac{1}{4}}\gamma^{-\frac{1}{2}}\sum_{m=0}^{M-1}2^{m(\frac{3}{4}d-1)}\\
&\leqslant \begin{cases} 2^{11}T^{\frac{1}{4}}\gamma^{-\frac{1}{2}} & \text{if } d=1\\ 2^7 M^{\frac{3}{2}}T^{\frac{1}{4}}\gamma^{-\frac{1}{2}}2^{M(\frac{3}{4}d-1)} & \text{if } 2\leqslant d\leqslant 4\end{cases}
\end{aligned}$$

  where inequality $(*)$ is by using $d \leqslant 4$ and choosing $c = 2^{-5/4}M_1^{-1/2}$, while the last inequality is because $\sum_{m=0}^{\infty}2^{-m/4} \leqslant 2^3$ for $d = 1$. Plugging this inequality into (31), upper bounding $\alpha_0$ using (34) and $\eta_0$ using (32), and recalling that $M = \lceil\log_2(1/\gamma)\rceil$, we get

  - case $d = 1$: choosing $\gamma = T^{-1/2}/\ln(T)$,

$$\mathrm{Reg}_T(\mathcal{F}) = \mathcal{O}\big(\sqrt{T\ln T}\big)\,. \tag{35}$$

– case $2 \leqslant d \leqslant 4$: choosing $\gamma = T^{-1/(d+2/3)}$,

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{F}) &\lesssim \ln(1/\gamma)^{\frac{3}{2}} T^{\frac{1}{4}} \gamma^{-\frac{1}{2}} \gamma^{-(\frac{3}{4}d-1)} + \ln(1/\gamma)^2 \gamma^{\frac{1}{2}} T^{\frac{3}{4}} + \gamma T \\ &\lesssim (\ln T)^{\frac{3}{2}} T^{\frac{d-1/3}{d+2/3}} + (\ln T)^2 T^{\frac{3d/4}{d+2/3}} + T^{\frac{d-1/3}{d+2/3}} \lesssim (\ln T)^{\frac{3}{2}} T^{\frac{d-1/3}{d+2/3}}\,, \end{aligned} \tag{36}$$

where the last inequality is because $3d/4 < d - 1/3$ for $d \geqslant 2$.

- If $d \geqslant 5$, then the terms inside the sum in (33) are exponentially increasing, so that

$$\alpha_m \leqslant c 2^4 \gamma^{\frac{1}{2}} T^{-\frac{1}{4}} 2^{M(\frac{d}{4}-1)} \frac{2^{\frac{d}{4}-1}}{2^{\frac{d}{4}-1}-1} \leqslant \frac{c 2^4}{1-2^{-1/4}} \gamma^{\frac{1}{2}} T^{-\frac{1}{4}} 2^{M(\frac{d}{4}-1)} \leqslant c 2^7 \gamma^{\frac{1}{2}} T^{-\frac{1}{4}} 2^{M(\frac{d}{4}-1)}\,. \tag{37}$$

Then, following the lines of the case $d \leqslant 4$, by substituting $\eta_m$ (see (32)) and $\alpha_m$ in the sum in the right-hand side of (31)

$$\begin{aligned}&\sum_{m=0}^{M-1} \left( \frac{2^{(m+1)d} \ln 3}{\eta_m} + \frac{4 \eta_m T \alpha_m^2}{\gamma^2} \right) \\ &\quad\leqslant \sum_{m=0}^{M-1} \left( \frac{2^{m(\frac{3}{4}d-1)+d} T^{\frac{1}{4}} \ln 3}{c \gamma^{\frac{1}{2}}} + \frac{2^{16} c^3 2^{m(\frac{d}{4}+1)+M(\frac{d}{2}-2)} T^{\frac{1}{4}}}{\gamma^{\frac{1}{2}}} \right) \\ &\quad= \frac{T^{\frac{1}{4}}}{\gamma^{\frac{1}{2}}} \left( \frac{2^d \ln 3}{c} \sum_{m=0}^{M-1} 2^{m(\frac{3}{4}d-1)} + 2^{16} c^3 2^{M(\frac{d}{2}-2)} \sum_{m=0}^{M-1} 2^{m(\frac{d}{4}+1)} \right) \\ &\quad\leqslant \frac{T^{\frac{1}{4}}}{\gamma^{\frac{1}{2}}} \left( \frac{2^{4+d}}{c} 2^{M(\frac{3}{4}d-1)} + 2^{16} c^3 2^{M(\frac{3d}{4}-1)} \right) \\ &\quad= \frac{T^{\frac{1}{4}}}{\gamma^{\frac{1}{2}}} \left( \frac{2^{4+d}}{c} + 2^{16} c^3 \right) 2^{M(\frac{3}{4}d-1)} \\ &\quad\leqslant T^{\frac{1}{4}} \gamma^{-\frac{1}{2}} 2^{M(\frac{3}{4}d-1)+8+\frac{3}{4}d}\end{aligned}$$

where in the last inequality we used $c = 2^{d/4-3}$. Plugging into the regret bound (31), upper bounding $\alpha_0$ using (37) and $\eta_0$ using (32), and recalling that $M = \lceil \log_2(1/\gamma) \rceil$, we get

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{F}) &\lesssim T^{\frac{1}{4}} \gamma^{\frac{1}{2}-\frac{3}{4}d} + \gamma^{\frac{3}{2}-\frac{d}{4}} T^{\frac{3}{4}} \ln(1/\gamma) + \gamma T \\ &\lesssim T^{\frac{d-1/3}{d+2/3}} + T^{\frac{d-1}{d+2/3}} \ln(T) + T^{\frac{d-1/3}{d+2/3}} \lesssim T^{\frac{d-1/3}{d+2/3}}\,, \end{aligned} \tag{38}$$

where the second inequality is by setting $\gamma = T^{-1(d+2/3)}$.

Putting together the three cases (35), (36), and (38) concludes the proof of the regret bound.

**Running time of HierExp4$^\star$.** We only address the case $d \geqslant 2$. First note that the total number of "Exp4 nodes" in $\mathcal{T}_{\mathcal{F}}^\star$ is

$$\begin{aligned}\left|\mathcal{T}_{\text{Exp4}}\right| &= \sum_{m=0}^{M-1} \left| \left\{ v = (\sigma_1, c_1, \sigma_2, c_2, \ldots, \sigma_m, c_m, \sigma_{m+1}) \,:\, \sigma_k \in \{1, \ldots, 2^d\}, c_k \in \{-1, 0, 1\} \right\} \right| \\ &= \sum_{m=0}^{M-1} 3^m 2^{d(m+1)} = \frac{1}{3} \frac{\left(3 \cdot 2^d\right)^{M+1} - 3 \cdot 2^d}{\left(3 \cdot 2^d\right) - 1} \leqslant \left(3 \cdot 2^d\right)^M .\end{aligned}$$

Similarly, $\left|\mathcal{T}_{\text{bin}}\right| \leqslant \left(3 \cdot 2^d\right)^M$ and $\left|\mathcal{L}\right| \leqslant \left(3 \cdot 2^d\right)^M$. Summing the three upper bounds, we get $\left|\mathcal{T}_{\mathcal{F}}^\star\right| \leqslant 3\left(3 \cdot 2^d\right)^M$. Therefore, using $M = \lceil \log_2(1/\gamma) \rceil$ and $\gamma = T^{-1/(d+2/3)}$, we obtain

$$M \leqslant 1 + \frac{\log_2 T}{d + 2/3} \qquad \text{so that} \qquad \left|\mathcal{T}_{\mathcal{F}}^\star\right| = \mathcal{O}\left(T^{\frac{d+\log_2 3}{d+2/3}}\right) .$$

The number of actions in $\mathcal{K}$ is $2^M = \mathcal{O}\left(T^{\frac{1}{d+2/3}}\right)$. The running time at every round $t$, which is at most proportional to the total number of tuples $(v, w, i)$ for $(v, i) \in \mathcal{T}_{\mathcal{F}}^\star \times \mathcal{K}$ and $w \in \mathcal{C}_v$, is at most of the order of $\left|\mathcal{T}_{\mathcal{F}}^\star\right| \times \left|\mathcal{K}\right| \times \max\{2^d, 3\}$. Therefore, the running time is at most of the order of $T^{\frac{d+1+\log_2 3}{d+2/3}} \leqslant T^{1.8}$ for all $d \geqslant 2$.

A tighter analysis however yields a smaller time complexity than the rate $T^{\frac{d+1+\log_2 3}{d+2/3}}$ derived above. This is because —from Algorithm 5— all elementary computations only need to be computed on *activated* "Exp4 nodes" nodes, on their (finite number of) children, and on actions in $\mathcal{K}$. Fix a round $t \geqslant 1$. From a computation similar to the one above, but noting that all activated nodes share the same coefficients $\sigma_m$ for all depths $m$, we can see that the total number of activated "Exp4 nodes" at round $t$ is at most of

$$\sum_{m=0}^{M-1} 3^m = \frac{3^M - 1}{2}$$

so that the running time per round is at most of the order of $3^M 2^M \lesssim T^{\frac{1+\log_2 3}{d+2/3}}$, which concludes the proof for $d \geqslant 2$. Similarly, in the case $d = 1$, the choice of $\gamma = T^{-1/2}/(\ln T)$ entails a running time at most of the order of $3^M 2^M = \mathcal{O}\left(\sqrt{T} \ln T\right)^{1+\log_2 3}\big) = o\Big(T^{\frac{1+\log_2 3}{1+2/3}}\Big)$, so that the stated running time also holds true for $d = 1$. ∎

## Appendix H. Algorithm HierHedge and proof of Theorem 7

The proof goes along the same lines as the proof of Theorem 4. Let $v$ be any internal node at depth $m = 0, \dots, M-1$. By construction of HierHedge, the distribution $p_t(v, \cdot) \in \Delta(K)$ is computed by Hedge and is supported on the set $\{i_t(w) : w \in \mathcal{L}_v\}$. By Lemma 13, the losses of any pair of actions in this set differ by at most $2^{-m+3}$. Hence, we may apply (Cesa-Bianchi et al., 2007, Theorem 5) with $N = |\mathcal{C}_v| \leqslant N_{m+1}$, $E = 2^{-m+3}$, and $\widehat{V}_t \leqslant E^2$ obtaining

$$\max_{w \in \mathcal{C}_v} \mathbb{E}\left[\sum_{t=1}^T p_t(v, \cdot) \cdot \ell_t - \sum_{t=1}^T p_t(w, \cdot) \cdot \ell_t\right] \leqslant 2^{-m+5}\left(\sqrt{T \ln N_{m+1}} + \ln N_{m+1} + 1\right) .$$

As in the proof of Theorem 4, we sum the above along a path $v_0 \to v_1 \to \cdots \to v_M = w$ for any leaf $w \in \mathcal{L}$ and get

$$\mathbb{E}\left[\sum_{t=1}^T p_t(v_0, \cdot) \cdot \ell_t\right] - \min_{w \in \mathcal{L}} \sum_{t=1}^T \ell_t(i_t(w)) \leqslant 2^5 \sum_{m=0}^{M-1} 2^{-m}\left(\sqrt{T \ln N_{m+1}} + \ln N_{m+1} + 1\right) .$$

**Algorithm 6:** HierHedge (for the full information feedback)

**Input** : Tree $\mathcal{T}_\mathcal{F}$ with root $v_0$ and leaves $\mathcal{L}$, learning rate sequences $\eta_2(v) \geqslant \eta_3(v) \geqslant \cdots > 0$ 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 $i = 1,\ldots,K$ and for all $v \in \mathcal{L}$;
3. Set $p_t(v,i) = q_t(v,\cdot) \cdot p_t(\cdot,i)$ for all $i = 1,\ldots,K$ and for all $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$;
4. Draw $I_t \sim p_t(v_0,\cdot)$;
5. Observe $\ell_t(i)$ for all $i = 1,\ldots,K$;
6. For each $v \in \mathcal{T}_\mathcal{F} \setminus \mathcal{L}$ and for each $w \in \mathcal{C}_v$ compute the expert loss $\widetilde{\ell}_t(v,w) = p_t(w,\cdot) \cdot \ell_t$ and perform the update

$$q_{t+1}(v,w) = \frac{\exp\Bigl(-\eta_{t+1}(v)\sum_{s=1}^t \widetilde{\ell}_s(v,w)\Bigr)}{\sum_{w' \in \mathcal{C}_v} \exp\Bigl(-\eta_{t+1}(v)\sum_{s=1}^t \widetilde{\ell}_s(v,w')\Bigr)} \tag{39}$$

**end**

Since $\mathcal{L}$ is a $(2^{-M})$-covering of $\mathcal{F}$, $2^{-M} \leq 2\varepsilon$, and $\mathcal{K}_\varepsilon$ is a $\varepsilon$-covering of $\mathcal{Y}$, we get

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t\bigl(y_{I_t}\bigr)\right] - \inf_{f\in\mathcal{F}} \sum_{t=1}^T \ell_t\bigl(f(x_t)\bigr) \leqslant 5\varepsilon T + 2^5 \sum_{m=0}^{M-1} 2^{-m}\left(\sqrt{T\ln N_{m+1}} + \ln N_{m+1} + 1\right) .$$

Overapproximating the sums with integrals similarly to (18) and (19),

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

If $\mathcal{F}$ is a set of Lipschitz functions $f : [0,1]^d \to [0,1]^p$, where $[0,1]^d$ is endowed with the norm $\|x - x'\|_\infty$, then $\ln \mathcal{N}_\infty(\mathcal{F},\varepsilon) = \mathcal{O}\bigl(p\varepsilon^{-d}\bigr)$ implying

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant 5T\varepsilon + \begin{cases} \mathcal{O}\bigl(\sqrt{pT} + p\ln(1/\varepsilon)\bigr) & \text{if } d = 1 \\ \mathcal{O}\bigl(\sqrt{pT}\ln(1/\varepsilon) + p\varepsilon^{1-d}\bigr) & \text{if } d = 2 \\ \mathcal{O}\bigl(\sqrt{pT}\varepsilon^{1-d/2} + p\varepsilon^{1-d}\bigr) & \text{if } d \geqslant 3. \end{cases}$$

Optimizing in $\varepsilon$, we obtain the stated result.