Zeroth-Order Non-Convex Learning via Hierarchical Dual Averaging

Amélie Héliou [1] **Matthieu Martin** [1] **Panayotis Mertikopoulos** [2 1] **Thibaud Rahier** [1]

Abstract

We propose a hierarchical version of dual averaging for zeroth-order online non-convex optimization – i.e., learning processes where, at each stage, the optimizer is facing an unknown non-convex loss function and only receives the incurred loss as feedback. The proposed class of policies relies on the construction of an online model that aggregates loss information as it arrives, and it consists of two principal components: (a) a regularizer adapted to the *Fisher information metric* (as opposed to the metric norm of the ambient space); and (b) a principled exploration of the problem's state space based on an adapted hierarchical schedule. This construction enables sharper control of the model's bias and variance, and allows us to derive tight bounds for both the learner's static and dynamic regret – i.e., the regret incurred against the best dynamic policy in hindsight over the horizon of play.

1. Introduction

Zeroth-order – or *derivative-free* – optimization concerns the problem of optimizing a given function without access to its gradient, stochastic or otherwise. Its study dates back at least to Rosenbrock (1960), and it has recently attracted significant interest in machine learning and artificial intelligence due to the prohibitive cost of automatic differentiation in very large neural nets and language models.

A popular approach to zeroth-order optimization involves sampling the function to be optimized at several nearby points, using the observed values to reconstruct the gradient of the function, and then employing a standard, first-order method (Conn et al., 2009). This approach allows the optimizer to approximate the gradient of the function to arbitrary precision (at least, if enough queries are made). However, it also requires that the problem's objective remain stationary during the query process.

Motivated by applications to online ad auctions and recommender systems, our paper concerns the case where this stationarity assumption breaks down – the *zeroth-order online optimization* (ZOO) setting. Specifically, we consider an adversarial ZOO problem that unfolds as follows:

1. At each stage $t = 1, 2, \dots,$ the optimizer selects an action x_t from a compact convex subset $\mathcal{K}$ of $\mathbb{R}^d$.
2. Simultaneously, an adversary selects a reward function $u_t \colon \mathcal{K} \to \mathbb{R}$, often assumed to take values in $[0, 1]$.
3. The optimizer receives $u_t(x_t)$ as a reward, and the process repeats.

The learner's performance after T stages is measured here by their regret, viz. $R_T = \sum_{t=1}^{T} [u_t(x) - u_t(x_t)]$, and the learner's goal is to minimize the growth rate of R_T.

Since each individual u_t may be encountered once – and only once – it is no longer possible to perform multiple queries per function. On that account, the simultaneous perturbation stochastic approximation (SPSA) estimator of Spall (1992) has been studied extensively as a viable alternative to multiple-point query methods for online optimization. In particular, using a variant of the SPSA scheme, Flaxman et al. (2005) showed that it is possible to achieve $\mathcal{O}(T^{3/4})$ regret if the payoff functions encountered are concave. The corresponding lower bound is $\Omega(T^{1/2})$, and it was only recently achieved by the kernel-based method of Bubeck & Eldan (2016) and Bubeck et al. (2017).

When venturing beyond problems with a convex structure, the situation is significantly more complicated. The most widely studied case is the "Lipschitz bandit" – or, sometimes, "Hölder bandit" – framework where each u_t is a random realization of a parametric model of the form $u_t(x) = \hat{u}(x; \xi_t)$ with Lipschitz continuous mean $u(x) = \mathbb{E}_\xi[\hat{u}(x; \xi)]$, cf. Agrawal (1995). In this case, the lower bound for the regret is $\Omega(T^{\frac{d+1}{d+2}})$, and several algorithms have been proposed to achieve it, typically by combining an intelligent discretization of the problem's search region with a deterministic UCB-type policy (Bubeck et al., 2011; Kleinberg et al., 2008; Slivkins, 2019).

On the other hand, in an adversarial setting, an informed adversary can always impose $\Omega(T)$ regret to any *determin-*

[1] Criteo AI Lab [2] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. Correspondence to: Panayotis Mertikopoulos <panayotis.mertikopoulos@imag.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

istic decision algorithm employed by the learner, cf. Hazan et al. (2017); Shalev-Shwartz (2011); Suggala & Netrapalli (2020). This makes the algorithms designed for Lipschitz bandits ill-suited for the framework at hand, and necessitates a different approach. In this direction, Krichene et al. (2015) showed that, if each payoff function u_t is revealed to the learner after playing, it is possible to achieve $\mathcal{O}(T^{1/2})$ regret. Similar bounds were obtained more recently by Agarwal et al. (2019) and Suggala & Netrapalli (2020), who examined the *"follow the perturbed leader"* (FTPL) algorithm of Kalai & Vempala (2005) assuming access to an offline optimization oracle; however, the knowledge of u_t is still implicitly required in these works (as input to an optimization or sampling oracle, depending on the context).

More recently, Héliou et al. (2020) proposed a general dual averaging framework for online non-convex learning with imperfect feedback, including the bona fide, adversarial ZOO case. Specifically, by using a "kernel smoothing" method in the spirit of Bubeck et al. (2017), Héliou et al. (2020) proposed a ZOO method achieving *a*) a suboptimal $\mathcal{O}(T^{\frac{d+2}{d+3}})$ regret bound; and *b*) a commensurate $\mathcal{O}(T^{\frac{d+3}{d+4}} V_T^{\frac{1}{d+4}})$ bound for the learner's *dynamic* regret, with $V_T = \sum_{t=1}^{T} \|u_{t+1} - u_t\|_\infty$ denoting the *total variation* of the payoff functions encountered (a common dynamic regret benchmark introduced by Besbes et al., 2015). However, the kernel method employed by Héliou et al. (2020) is difficult to implement because the kernel's support function may grow exponentially in both T and d.

Our contributions. In this paper, we take a different approach that fuses the dual averaging framework of Krichene et al. (2015) with a hierarchical exploration scheme in the spirit of Bubeck et al. (2011) and Kleinberg et al. (2008; 2019). Specifically, we propose a flexible, anytime *hierarchical dual averaging* (HDA) method with the following desirable properties: *(i)* it enjoys a min-max optimal $\mathcal{O}(T^{\frac{d+1}{d+2}})$ static regret bound; *(ii)* it guarantees at most $\mathcal{O}(T^{\frac{d+2}{d+3}} V_T^{\frac{1}{d+3}})$ dynamic regret. In this way, our paper closes the optimality gap in the regret analysis of Héliou et al. (2020), and it answers in the positive the authors' conjecture that it is possible to achieve $\mathcal{O}(T^{\frac{d+2}{d+3}} V_T^{\frac{1}{d+3}})$ dynamic regret in adversarial ZOO problems.

As far as we are aware, HDA is the first algorithm in the literature enjoying this dynamic regret guarantee. Moreover, in contrast to the CAB algorithm of Kleinberg (2004), we should stress that HDA *does not* require a restart schedule or a doubling trick. From a practical viewpoint, this is particularly important because the doubling trick leads to sharp performance drops when the algorithm periodically restarts from scratch – an unpleasant property, which is one of the main reasons that doubling methods are rarely employed by practitioners (Bubeck et al., 2011).

Our analysis relies on two principal components: *a*) a logarithmic scheduler for controlling the hierarchical exploration of the problem's state space; and *b*) a regularization framework adapted to the Fisher information metric on the learner's mixed strategies. The first of these components marks a crucial point of departure from the hierarchical approach of Bubeck et al. (2011) and Kleinberg et al. (2019) since, instead of increasing the granularity of our search "pointwise", we do so "dimension-wise" (but at a slower pace). As for the second component, the use of the Fisher information metric allows us to drop the reliance of dual averaging on a global norm that is not adapted to the geometry of the problem at hand, and it allows us to bring into play a wide range of regularizers that were previously unexplored in the literature – such as the Burg entropy. This is a crucial difference with existing results on dual averaging, and it allows for much finer control of the learning process as it unfolds – precisely because the information content of the learner's policy is not ignored in the process.

Upon completion of our paper, we discovered a very recent preprint by Podimata & Slivkins (2021) that proposes an adversarial zooming algorithm. The authors achieve a static $\mathcal{O}(T^{\frac{d+1}{d+2}})$ regret bound in high probability (but do not provide any dynamic regret guarantees). Their algorithm uses an explicit exploration term, plus a confidence term in the per-round sampling uncertainty. Their splitting rule splits only one-by-one cover set into 2^d sub-covers, which might be more difficult to implement in practice.

2. Setup and preliminaries

2.1. The model

We assume throughout that $\mathcal{K}$ is a compact convex subset of an ambient real space $\mathbb{R}^d$ endowed with an abstract norm $\|\cdot\|$ and a reference measure λ (typically the ordinary Lebesgue measure). As for the payoff functions encountered by the learner, we will make the following blanket assumption:

Assumption 1. The stream of payoff functions $u_t \colon \mathcal{K} \to \mathbb{R}$, $t = 1, 2, \dots$, is *uniformly bounded Lipschitz*, i.e., there exist nonnegative constants $R, L \geq 0$ such that

1. $0 \leq u_t(x) \leq R$ for all $x \in \mathcal{K}$.
2. $|u_t(x') - u_t(x)| \leq L\|x' - x\|$ for all $x, x' \in \mathcal{K}$.

To avoid exploitable, deterministic strategies, we will assume that the learner has access to an unobservable randomizer that can be used to choose an action $x \in \mathcal{K}$ by means of a probability distribution on $\mathcal{K}$ – that is, a *mixed strategy*. Of course, in complete generality, the space of all mixed strategies is impractical to work with because it contains probability distributions that cannot be described in closed form (let alone have a "sampling-friendly" structure). For this reason, we will focus on *simple strategies*, i.e., probability distributions with a piecewise constant density.

Definition 1. A mixed strategy on $\mathcal{K}$ is called *simple* if it admits a density function of the form $q = \sum_{i=1}^{m} \alpha_i \, \mathbb{1}_{\mathcal{S}_i}$ for a collection of weights $\alpha_i > 0$, $i = 1, \dots, m$, and mutually disjoint λ-measurable subsets $\mathcal{S}_i$ of $\mathcal{K}$ ($\mathcal{S}_i \cap \mathcal{S}_j = \varnothing$ for $i \neq j$) such that $\int_{\mathcal{K}} q = \sum_i \alpha_i \lambda_i(\mathcal{S}_i) = 1$. The space of simple strategies on $\mathcal{K}$ will be denoted by $\mathcal{Q}(\mathcal{K})$, and the expectation of a function $f : \mathcal{K} \to \mathbb{R}$ under q will be written as $\langle f, q \rangle \coloneqq \mathbb{E}_{x \sim q}[f(x)] = \sum_{i=1}^{m} \alpha_i \int_{\mathcal{S}_i} f(x) \, d\lambda(x)$

Owing to their decomposable structure, simple strategies are relatively easy to sample from, and they can approximate general distributions on $\mathcal{K}$ to arbitrary precision – formally, they are dense in the weak topology of (regular) probability measures on $\mathcal{K}$ (Folland, 1999, Chap. 2). On the other hand, this "universal approximation" guarantee comes at the cost of an increased number of supporting sets $\mathcal{S}_i$, $i = 1, \dots, m$. In particular, there is no "free lunch": when m grows large, sampling from a simple strategy can become computationally expensive – if not intractable – so we will pay particular attention to the support of such strategies.

Remark 1. To facilitate sampling, we will also consider strategies of the form $q = \sum_{i=1}^{m} \alpha_i \psi_{\mathcal{S}_i}$ where $\psi_{\mathcal{S}}$ is supported on $\mathcal{S}$ and can be sampled cheaply – e.g., $\psi_{\mathcal{S}}$ could be a suitably weighted Dirac distribution on a specific point of $\mathcal{S}$. Strategies of this type are not *stricto sensu* "simple", but our results will also cover this case, cf. Section 4.

2.2. Regret: static and dynamic

Going back to the learner's sequence of play, we will assume that, at each stage $t = 1, 2, \dots$, the learner picks an action $x_t \in \mathcal{K}$ based on a simple strategy $q_t \in \mathcal{Q}$, and receives the reward $u_t(x_t)$. The *regret* of the policy q_t against a *benchmark action* $x \in \mathcal{K}$ is then defined as the difference between the player's mean cumulative payoff under q_t and x over a horizon of T rounds. Formally, we have

$$\mathrm{Reg}_x(T) \coloneqq \sum_{t=1}^{T} \mathbb{E}_{x_t \sim q_t}[u_t(x) - u_t(x_t)]. \qquad (1)$$

Moreover, letting $x^* \in \arg\max_{x \in \mathcal{K}} \sum_{t=1}^{T} u_t(x)$ be the "best fixed action in hindsight" over the horizon T, we also define the learner's *static regret* as

$$\mathrm{Reg}(T) \coloneqq \mathrm{Reg}_{x^*}(T) = \max_{x \in \mathcal{K}} \mathrm{Reg}_x(T). \qquad (2)$$

Finally, to relax the requirement of using a "fixed" action as a comparator, we will also consider the learner's *dynamic regret*, defined here as

$$\mathrm{DynReg}(T) \coloneqq \sum_{t=1}^{T} \max_{x \in \mathcal{K}} \mathbb{E}_{x_t \sim q_t}[u_t(x) - u_t(x_t)], \qquad (3)$$

i.e., as the difference between the player's mean cumulative payoff and that of the best sequence of actions $x_t^* \in \arg\min_x u_t(x)$ over the horizon of play T.

In regard to its static counterpart, the agent's dynamic regret is considerably more ambitious, and achieving sublinear dynamic regret is not always possible; we examine this issue in detail in Section 5.

In both cases, it should also be clear that there is no simple strategy that can match the exact performance of the "best" action (x^* or x_t^*, depending on the context). For example, consider the static optimization problem $u_t(x) = 1 - x^2/2$ with $x \in \mathcal{K} = [-1, 1]$: then, any simple strategy $q \in \mathcal{Q}$ would yield a payoff strictly less than 1 at each round because it is sampling with probability 1 points other than 0. Nevertheless, the following lemma shows that the propagated error on the regret can be made arbitrarily small:

Lemma 1. *Let $\mathcal{U}$ be a neighborhood of $x \in \mathcal{K}$. Then, for every simple strategy $q \in \mathcal{Q}$ supported on $\mathcal{U}$, we have*

$$\mathrm{Reg}_x(T) \leq L \operatorname{diam}(\mathcal{U}) T + \sum_{t=1}^{T} \langle u_t, q - q_t \rangle \qquad (4)$$

Proof. By Assumption 1, we have $u_t(x) \leq u_t(x') + L\|x - x'\| \leq u_t(x') + L \operatorname{diam}(\mathcal{U})$ for all $x' \in \mathcal{U}$. Hence, letting $x' \sim q$ and expectations on both sides, we get $u_t(x) \leq \langle u_t, q \rangle + L \operatorname{diam}(\mathcal{U})$. Our claim then follows by summing over t and invoking the definition of the regret. ■

Remark 2. We note here that the bound (4) does not need the full capacity of the Lipschitz continuity framework; in fact, it continues to hold under much less restrictive notions, such as the weak one-sided continuity condition of Bubeck et al. (2011). Nevertheless, in the sequel we will maintain the assumption of Lipschitz continuity for simplicity.

3. Dual averaging with an explicit cover

To build some intuition for the analysis to come, we begin by adapting the *dual averaging* (DA) algorithm of Nesterov (2009) to the (infinite) space of simple strategies with an explicit cover. This will allow us to introduce the relevant notions that we will need in the sequel, namely the *range* of an estimator and the *Fisher information metric*.

3.1. Basic setup

Let $\mathcal{P} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ be a measurable partition of $\mathcal{K}$ with nontrivial covering sets, i.e., $\lambda(\mathcal{S}) > 0$ and $\mathcal{S} \cap \mathcal{S}' = \varnothing$ for all $\mathcal{S}, \mathcal{S}' \in \mathcal{P}$ with $\mathcal{S} \neq \mathcal{S}'$. In particular, this implies that every point $x \in \mathcal{K}$ belongs to a unique element of $\mathcal{P}$, denoted below by $\mathcal{S}_x$. Since the elements of $\mathcal{P}$ cover $\mathcal{K}$ in an unambiguous way, we will refer to $\mathcal{P}$ as an *explicit cover* of $\mathcal{K}$. This cover will be assumed fixed throughout this section.

In terms of sampling actions from $\mathcal{K}$, the above also gives rise to a set of *simple strategies* supported on $\mathcal{P}$, namely

$$\mathcal{Q}_{\mathcal{P}} = \{\textstyle\sum_{\mathcal{S}} \alpha_{\mathcal{S}} \mathbb{1}_{\mathcal{S}} : \alpha_{\mathcal{S}} \geq 0, \sum_{\mathcal{S}} \alpha_{\mathcal{S}} \lambda(\mathcal{S}) = 1\} \tag{5}$$

Geometrically, it will be convenient to interpret $\mathcal{Q}_{\mathcal{P}}$ as a simplex embedded in the space of all test functions $\phi\colon \mathcal{K} \to \mathbb{R}$ that are piecewise constant on the covering sets of $\mathcal{P}$. Since such functions may be viewed equivalently as functions $\phi\colon \mathcal{P} \to \mathbb{R}$, we will denote this function space by $\mathbb{R}^{\mathcal{P}}$.

Moving forward, we will assume that the learner is sampling from $\mathcal{K}$ with simple strategies taken from $\mathcal{Q}_{\mathcal{P}}$, and we will write $q_{\mathcal{S}} \coloneqq \mathbb{P}_{x\sim q}(x \in \mathcal{S}) = \int_{\mathcal{S}} q = \alpha_{\mathcal{S}} \lambda(\mathcal{S})$ for the probability of choosing an element of $\mathcal{S}$ under q. Accordingly, our non-convex learning framework may be encoded in more concrete terms as follows: (i) at each stage $t = 1, 2, \dots$, the adversary chooses (but does not reveal) a payoff function $u_t\colon \mathcal{K} \to [0, R]$; (*ii*) the learner selects an action $x_t \in \mathcal{K}$ based on some simple strategy X_t supported on $\mathcal{P}$; and (*iii*) the corresponding reward $u_t(x_t)$ is received by the learner and the process repeats.

As an algorithmic template for learning in this setting, we will consider an adaptation of the classical dual averaging algorithm of Nesterov (2009). Specifically, we will focus on an online policy that we call *dual averaging with an explicit cover* (DAX), and which is defined recursively as

$$\begin{aligned} S_{t+1} &= S_t + \hat{u}_t \\ x_{t+1} \sim X_{t+1} &= Q(\eta_{t+1} S_{t+1}) \end{aligned} \tag{DAX}$$

where

1. $\hat{u}_t \in \mathbb{R}^{\mathcal{P}}$ is an *estimate* – or *model* – of the otherwise unobserved payoff function u_t of stage t.
2. $S_t \in \mathbb{R}^{\mathcal{P}}$ is an auxiliary *scoring function* that aggregates previous payoff models – so $S_t(x)$ indicates the learner's propensity of choosing $x \in \mathcal{K}$ at stage t.
3. $\eta_t > 0$ is a "learning rate" parameter that adjusts the sharpness of the learning process.
4. $Q\colon \mathbb{R}^{\mathcal{P}} \to \mathcal{Q}_{\mathcal{P}}$ is a *choice map* that transforms scoring functions $S_t \in \mathbb{R}^{\mathcal{P}}$ into simple strategies $X_t \in \mathcal{Q}_{\mathcal{P}}$.

Each component of the method is discussed in detail below. We also note that this method is often referred to as *"follow the regularized leader"* (FTRL), cf. Shalev-Shwartz (2011); Shalev-Shwartz & Singer (2006). Our choice of terminology follows Nesterov (2009) and Xiao (2010).

3.2. The choice map

We begin by detailing the method's "choice map" $Q\colon \mathbb{R}^{\mathcal{P}} \to \mathcal{Q}_{\mathcal{P}}$ which determines action choice probabilities based on the "score function" $S_t(x)$. With this in mind, we will focus on a class of "regularized strategies" that output at each stage a simple strategy $X_t \in \mathcal{Q}_{\mathcal{P}}$ that maximizes the learner's expected score minus a regularization penalty.

Specifically, we will consider choice maps of the form

$$Q(y) = \operatorname*{arg\,max}_{q \in \mathcal{Q}_{\mathcal{P}}} \{\langle y, q \rangle - h(q)\} \quad \text{for all } y \in \mathbb{R}^{\mathcal{P}}, \tag{6}$$

where the *regularizer* $h\colon \mathcal{Q}_{\mathcal{P}} \to \mathbb{R}$ is assumed to be continuous and strictly convex on $\mathcal{Q}_{\mathcal{P}}$. To streamline our presentation, we will further assume that h is *decomposable*, i.e., it can be written as $h(q) = \sum_{\mathcal{S}\in\mathcal{P}} \theta(q_{\mathcal{S}})$ for some strictly convex, C^2-smooth function $\theta\colon (0, 1] \to \mathbb{R}$. Two widely used examples are as follows:

Example 1 (Negentropy)**.** Consider the *entropic kernel* $\theta(x) = x \log x$ with the continuity convention $0 \log 0 = 0$. Then, by a standard calculation, the associated choice map is given by the logit choice model

$$\Lambda(y) = \frac{\exp(y)}{\int_{\mathcal{K}} \exp(y)}, \tag{7}$$

where $y \equiv y(x)$ is an arbitrary piecewise constant function on $\mathcal{P}$.

Example 2 (Log-barrier)**.** Another important example is the *log-barrier* (or *Burg entropy*) kernel $\theta(x) = -\log x$. In this case, the associated choice map does not admit a closed form expression, but it can be calculated by a binary search algorithm in logarithmic time.[1] The induced algorithm has deep links to Karmarkar's "affine scaling" method for linear programming (Karmarkar, 1990; Vanderbei et al., 1986), cf. Alvarez et al. (2004) and references therein.

3.3. Estimators

The second basic ingredient of (DAX) is the estimate $\hat{u}_t$ of the learner's payoff function u_t at time t. Since we are working with a fixed cover $\mathcal{P}$ of $\mathcal{K}$, the estimator $\hat{u}_t$ may

[1] This is done by noting that any solution of the defining maximization problem (6) would have to satisfy the first-order optimality condition $\sum_{\mathcal{S}\in\mathcal{P}} (\xi - y_{\mathcal{S}})^{-1} = 1$ for some $\xi > \max_{\mathcal{S}} y_{\mathcal{S}}$ (in which region the function being searched is strictly decreasing).

not exceed the cover's granularity, which is why we require $\hat{u}_t$ to be piecewise constant on $\mathcal{P}$ – i.e., $\hat{u}_t \in \mathbb{R}^{\mathcal{P}}$.

Overall, we will measure the quality of $\hat{u}_t$ as an estimator by means of the corresponding error process $Z_t = \hat{u}_t - u_t$ which is assumed to capture all sources of uncertainty and lack of precision in the learner's estimation process. To differentiate further between random (zero-mean) and systematic (nonzero-mean) errors, we will decompose Z_t as

$$Z_t = U_t + b_t, \tag{8}$$

where $b_t = \mathbb{E}[Z_t \mid \mathcal{F}_t]$ denotes the *bias* of the estimator, and $U_t = Z_t - b_t$ the inherent *random noise* (so $\mathbb{E}[U_t \mid \mathcal{F}_t] = 0$ for all t). In terms of measurability, these processes are all conditioned on the history $\mathcal{F}_t := \mathcal{F}(X_1, \dots, X_t)$ of the learner's policy up to – and including – stage t. Thus, in terms of the sequence of events described earlier, X_t is $\mathcal{F}_t$-measurable (by definition), but x_t, Z_t, U_t and b_t are not.

For concreteness, we provide some examples below:

Example 3 (Importance weighted estimator)**.** Motivated by the literature on multi-armed bandits (Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2020; Slivkins, 2019), a natural way to reconstruct u_t is via the *importance weighted estimator*

$$\hat{u}_t(x) = R - \frac{R - u_t(x_t)}{X_{\mathcal{S}_t,t}} \mathbb{1}(x \in \mathcal{S}_t), \tag{IWE}$$

where $\mathcal{S}_t := \mathcal{S}_{x_t}$ denotes the element of $\mathcal{P}$ containing the sampled action x_t, and R is one upper bound of the learner's rewards. This particular formulation of (IWE) is known as "loss-based"; other normalizations are possible but this is the most widely used one when considering sampling policies based on exponential weights algorithms (Slivkins, 2019).

Example 4 (Importance weighted estimator with explicit exploration)**.** One shortfall of (IWE) is that it requires knowledge of the upper bound R for the learner's rewards. When this is not known, a suitable alternative is to introduce an *explicit exploration* parameter $\varepsilon_t > 0$ in the learner's sampling strategy X_t. This means that the learner now chooses an action $x_t \in \mathcal{P}$ according to the perturbed strategy $\hat{X}_t = (1 - \varepsilon_t) X_t + \varepsilon_t \, \mathrm{unif}_{\mathcal{P}}$, where $\mathrm{unif}_{\mathcal{P}} = |\mathcal{P}|^{-1} \sum_{\mathcal{S} \in \mathcal{P}} \lambda(\mathcal{S})^{-1} \mathbb{1}_{\mathcal{S}}$ denotes the uniform distribution on $\mathcal{P}$. The *importance weighted estimator with explicit exploration* is then defined as

$$\hat{u}_t(x) = \frac{u_t(x_t)}{\hat{X}_{\mathcal{S}_t,t}} \mathbb{1}(x \in \mathcal{S}_t) \tag{IWE^3}$$

with $\mathcal{S}_t := \mathcal{S}_{x_t}$ as above. In contrast to (IWE), the estimator (IWE^3) has bias and variance bounded respectively as $\mathbb{E}[b_t] = \mathcal{O}(\varepsilon_t)$ and $\mathbb{E}[U_{\mathcal{S},t}^2] = \mathcal{O}(1/\varepsilon_t)$, i.e., both can be controlled by tuning ε_t. This provides additional flexibility relative to (IWE), but the introduction of the explicit exploration parameter ε_t often ends up having a negative impact on the regret (Slivkins, 2019), an important disadvantage.

Other estimators have also been used in the literature, such as *implicit* exploration and its variants (Kocák et al., 2014). For posterity, we only note here that the set of possible values $\mathcal{R} := \bigcup_t \mathrm{im}(\hat{u}_t) \subseteq \mathbb{R}^{\mathcal{P}}$ attained by an estimator will play an important role in the sequel. When the estimator is understood from the context, we will refer to this image set as the *range* of the estimator. In the examples above, we have:

1. For (IWE): $\mathcal{R} = (-\infty, R]^{\mathcal{P}}$.
2. For (IWE^3): $\mathcal{R} = \mathbb{R}_+^{\mathcal{P}}$.

We will return to this point in the next section.

3.4. Strong convexity and the Fisher metric

Deriving explicit regret guarantees for dual averaging methods is typically contingent on the method's regularizer being *strongly convex* (Bubeck & Cesa-Bianchi, 2012; Shalev-Shwartz, 2011). Formally, strong convexity posits that there exists some $K > 0$ such that, for all $q, q' \in \mathcal{Q}$ and all $s \in [0, 1]$, we have

$$\begin{aligned} h(sq + (1-s)q') &\leq sh(q) + (1-s)h(q') \\ &\quad - \frac{K}{2} s(1-s) \|q - q'\|^2 \end{aligned} \tag{9}$$

In the above, $\|\cdot\|$ denotes an arbitrary reference norm on $\mathbb{R}^{\mathcal{P}}$, usually taken to be the Euclidean norm $\|\cdot\|_2$ or the Manhattan L^1 norm $\|\cdot\|_1$. However, in our case, seeing as we are comparing *probability distributions*, an arbitrary reference norm does not seem particularly adapted to the problem at hand.

Instead, when dealing with probability distributions, it is common to measure the distance of q' relative to q via the *Fisher information metric*, which is typically used to compute the informational difference between probability distributions. In our context, the Fisher metric is defined for all $q, q' \in \mathcal{Q}_{\mathcal{P}}$ with $q \ll q'$ as

$$\|q' - q\|_q^2 = \int_{\mathcal{K}} \left[\frac{d(q' - q)}{dq}\right]^2 dq = \sum_{\mathcal{S} \in \mathcal{P}} \frac{(q'_{\mathcal{S}} - q_{\mathcal{S}})^2}{q_{\mathcal{S}}}. \tag{10}$$

We will then posit the following strong convexity requirement relative to the Fisher metric

$$\begin{aligned} h(sq + (1-s)q') &\leq sh(q) + (1-s)h(q') \\ &\quad - \frac{K}{2} s(1-s) \|q - q'\|_q^2. \end{aligned} \tag{11}$$

Since this is a non-standard requirement, we immediately proceed with an example.

Example 5. The Burg entropy $h(x) = -\sum_{\mathcal{S}\in\mathcal{P}} \log q_{\mathcal{S}}$ is 1-strongly convex relative to the Fisher metric. Indeed, since h is smooth, the strong convexity requirement for h with $K = 1$ can be rewritten as $D_{\mathrm{IS}}(q', q) \geq \frac{1}{2}\sum_{\mathcal{S}\in\mathcal{P}}(q'_{\mathcal{S}} - q_{\mathcal{S}})^2/q_{\mathcal{S}}$ where $D_{\mathrm{IS}}(q', q) = \sum_{\mathcal{S}\in\mathcal{P}}[q'_{\mathcal{S}}/q_{\mathcal{S}} - \log(q'_{\mathcal{S}}/q_{\mathcal{S}}) - 1]$ denotes the Itakura–Saito distance on $\mathcal{Q}_{\mathcal{P}}$. Our claim then follows from Antonakopoulos et al. (2020, Ex. 4).

The key implication of Fisher strong convexity for our analysis is the following characterization:

Lemma 2. *Let $h^*(y) = \max_{q\in\mathcal{Q}_{\mathcal{P}}}\{\langle y, q\rangle - h(q)\}$ be the convex conjugate of h. The following are equivalent:*

1. *h satisfies (11).*
2. *h^* is $(1/K)$-Lipschitz smooth relative to the dual Fisher norm $\|y\|^2_{q,*} = \sum_{\mathcal{S}\in\mathcal{P}} q_{\mathcal{S}} y^2_{\mathcal{S}}$ on $\mathbb{R}^{\mathcal{P}}$; specifically, for all $y, v \in \mathbb{R}^{\mathcal{P}}$, we have*

$$h^*(y + v) \leq h^*(y) + \langle v, \chi\rangle + \frac{1}{2K}\|v\|^2_{\chi,*}, \qquad (12)$$

where $\chi = Q(y)$.

Lemma 2 mirrors the well-known equivalence between strong convexity in the primal and Lipschitz smoothness in the dual (Bubeck & Cesa-Bianchi, 2012; Shalev-Shwartz, 2011). However, we must stress here that the norms in (11) are *not* global, but *strategy-dependent* – in effect, they comprise a Riemannian metric on the set of simple strategies $\mathcal{Q}_{\mathcal{P}}$. This is a crucial difference with the standard analysis of dual averaging, and it allows for much finer control of the learning process as it unfolds – precisely because the base distribution $\chi = Q(y)$ is not ignored in the process.

We close this section by noting that the entropic regularizer of (1) *does not* satisfy (11); we provide an explicit discussion of this point in the supplement. However, as we also show in the supplement, it *does* satisfy the Lipschitz smoothness requirement (12) for all $v \in \mathbb{R}^{\mathcal{P}}$ that are "upper-bounded", i.e., $\sup_{\mathcal{S}\in\mathcal{P}} v_{\mathcal{S}} \leq M$ for some $M \in \mathbb{R}$. From an algorithmic viewpoint, this relaxation of (11) will play a pivotal role in the sequel, so we encode it as follows:

Definition 2. Let $\mathcal{R}$ be a nonempty convex subset of $\mathbb{R}^{\mathcal{P}}$. We say that h is K-*tame* relative to $\mathcal{R}$ if (12) holds for all $y \in \mathbb{R}^{\mathcal{P}}$ and all $v \in \mathcal{R}$.

Clearly, by Lemma 2, any regularizer satisfying (11) is tame relative to any subset of $\mathbb{R}^{\mathcal{P}}$ (including $\mathbb{R}^{\mathcal{P}}$ itself). By contrast, as we mentioned above, the entropic regularizer of Example 1 is 1-tame over the region $\mathcal{R} = \{y \in \mathbb{R}^{\mathcal{P}} : y_{\mathcal{S}} \leq 1\}$, but *it is not tame* over all of $\mathbb{R}^{\mathcal{P}}$. In the analysis to come, we will see that this property introduces an intricate interplay between the two principal components of (DAX), namely the choice of regularizer h and the estimator $\hat{u}$. Unless explicitly mentioned otherwise, in the rest of this section we will assume that $\mathcal{R}$ is fixed and h is K-tame relative to $\mathcal{R}$.

3.5. Regret analysis

The key element in our analysis will be to control the "divergence" between a scoring function S_t and a comparator strategy $q \in \mathcal{Q}_{\mathcal{P}}$. Because these two elements live in different spaces, we introduce below the *Fenchel coupling*

$$F(q, y) = h(q) + h^*(y) - \langle y, q\rangle, \qquad (13)$$

for all $q \in \mathcal{Q}_{\mathcal{P}}$, $y \in \mathbb{R}^{\mathcal{P}}$. Clearly, by the Fenchel-Young inequality, we have $F(q, y) \geq 0$ with equality if and only if $Q(y) = q$. More to the point, as we show in the supplement, the Fenchel coupling enjoys the following growth property:

Lemma 3. *For all $y \in \mathbb{R}^{\mathcal{P}}$ and all $v \in \mathcal{R}$, we have*

$$F(q, y + v) = F(q, y) + \langle v, \chi - q\rangle + F(\chi, y + v) \qquad (14a)$$
$$\leq F(q, y) + \langle v, \chi - q\rangle + \frac{1}{2K}\|v\|^2_{\chi,*} \qquad (14b)$$

where $\chi = Q(y)$.

Using (14), we will analyze the regret properties of (DAX) via the η_t*-deflated coupling*

$$E_t = \frac{1}{\eta_t} F(q, \eta_t S_t). \qquad (15)$$

Doing so leads to the following result:

Lemma 4. *Suppose that (DAX) is run with an estimator with range $\mathcal{R}$. For all $t = 1, 2, \dots,$ we have*

$$E_{t+1} \leq E_t + \langle \hat{u}_t, X_t - q\rangle + \left(\eta_{t+1}^{-1} - \eta_t^{-1}\right)[h(q) - \min h] + \eta_t^{-1} F(X_t, \eta_t S_{t+1}). \qquad (16)$$

If, in addition, h is K-tame relative to $\mathcal{R}$, the last term in (16) is bounded as

$$\eta_t^{-1} F(X_t, \eta_t S_{t+1}) \leq \eta_t/(2K)\|\hat{u}_t\|^2_t, \qquad (17)$$

where $\|\cdot\|_t$ is the dual Fisher norm $\|v\|_t \coloneqq \|v\|_{X_t,*}$.

Thus, telescoping Lemma 4, we obtain the regret bound below.

Proposition 1. *The regret incurred by the learner relative to $q \in \mathcal{Q}_{\mathcal{P}}$ over the interval $\mathcal{T} = \{t_1, \dots, t_2 - 1\}$ is bounded as*

$$\mathrm{Reg}_q(\mathcal{T}) \leq E_{t_1} - E_{t_2} + \left(\eta_{t_2}^{-1} - \eta_{t_1}^{-1}\right)[h(q) - \min h] + \sum_{t\in\mathcal{T}}\langle Z_t, X_t - q\rangle + \frac{1}{2K}\sum_{t\in\mathcal{T}}\eta_t\|\hat{u}_t\|^2_t. \qquad (18)$$

We are finally in a position to state our main regret guarantees for (DAX). For generality, we state our result with a generic estimator $\hat{u}_t$ enjoying the following bounds:

$$a)\quad \textit{Bias:} \qquad |\langle b_t, q\rangle| \leq \mu_t \tag{19a}$$

$$b)\quad \textit{Mean square:} \qquad \mathbb{E}[\|\hat{u}_t\|_t^2 \,|\, \mathcal{F}_t] \leq M_t^2 \tag{19b}$$

for all $t = 1, 2, \dots$, and all $q \in \mathcal{Q}_{\mathcal{P}}$. We stress here that the use of the Fisher metric in (19) is *crucial*: for example, the IWE estimator satisfies (19b) with $M_t = \mathcal{O}(R^2|\mathcal{P}|)$ (where $|\mathcal{P}|$ is the size of the underlying partition) but it does not satisfy this bound for *any* global norm. Again, the reason for this is that the dual Finsler norm can be considerably smaller than any other global norm, depending on the information content of X_t.

This feature plays an essential role in deriving the algorithm's regret below.

Theorem 1. *Suppose that* (DAX) *is run with assumptions as in Proposition 1. Then the learner's regret is bounded as*

$$\mathbb{E}[\mathrm{Reg}_q(T)] \leq E_1 - E_{T+1} + \left(\eta_{T+1}^{-1} - \eta_1^{-1}\right)[h(q) - \min h] + 2\sum_{t=1}^{T}\mu_t + \frac{1}{2K}\sum_{t=1}^{T}\eta_t M_t^2. \tag{20}$$

This theorem is proved in the supplement and constitutes the main ingredient for our results on the hierarchical dual averaging that we present in the next section.

4. Hierarchical dual averaging

In this section, we proceed to define the mechanism that we will use to recursively "zoom-in" on different regions of the state space. This hierarchical approach is inspired by earlier works by Bubeck et al. (2011), but with the crucial difference that we do not zoom in "pointwise" but "dimension-wise". We explain all this in detail below.

4.1. The splitting mechanism

As in the case of Bubeck et al. (2011) and Kleinberg et al. (2008; 2019), the basic element of our construction is an infinite "tree of coverings", each of whose levels $\sigma = 1, \dots$ defines a successively finer cover $\mathcal{P}_\sigma$ of $\mathcal{K}$ (i.e., $\mathcal{P}_\sigma \subseteq \mathcal{P}_{\sigma+1}$ for all $\sigma = 1, \dots$). However, in contrast to these previous works, we do not consider *binary* trees, but *dyadic* ones; specifically, each cover $\mathcal{P}_\sigma = \{\mathcal{S}_{\sigma,i}\}_{i\leq 2^\sigma}$ is defined inductively as follows: (i) $\mathcal{P}_0 = \{\mathcal{S}_{0,1}\} = \{\mathcal{K}\}$; (ii) at specific stages of the learning process (that we define later), a *splitting event* occurs, and each leaf[2] of the current cover is split into 2 sub-leaves as detailed below (refer also to Figs. 1 and 2 for intuition in the case $d = 2$). For a given node $\mathcal{S}_{\sigma,i}$, we define $\mathcal{S}_{\sigma+1,2i-1}$ and $\mathcal{S}_{\sigma+1,2i}$ as the two subsets obtained from splitting the leaf $\mathcal{S}_{\sigma,i}$ in 2 equally sized leaves along the dimension of largest diameter (with ties broken deterministically). We then have $\mathcal{S}_{\sigma+1,2i} \cup \mathcal{S}_{\sigma+1,2i-1} = \mathcal{S}_{\sigma,i}$, $\mathcal{S}_{\sigma+1,2i} \cap \mathcal{S}_{\sigma+1,2i-1} = \varnothing$ and $\lambda(\mathcal{S}_{\sigma+1,2i}) = \lambda(\mathcal{S}_{\sigma+1,2i-1}) = \lambda(\mathcal{S}_{\sigma,i})/2$.

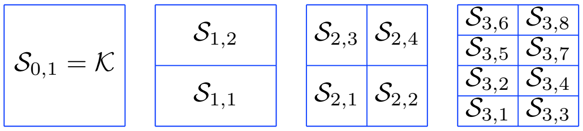


Figure 1: Example of the 3 first *splitting events* for $\mathcal{K} = [0, 1]^2$

In the sequel, for any cover $\mathcal{P}$, we write $\mathcal{P}^+$ for its *successor* cover, i.e., the cover after a splitting event on $\mathcal{P}$.

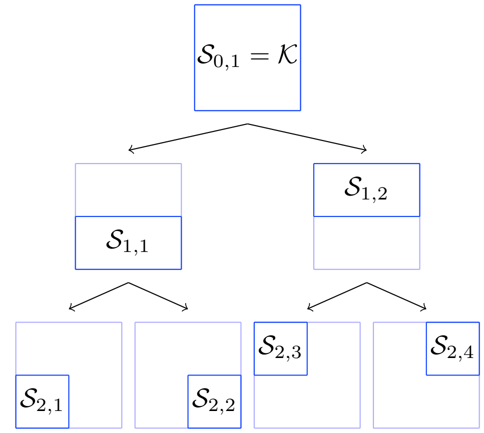


Figure 2: Example of a covering tree for the cube $\mathcal{K} = [0, 1]^2$

A crucial information for the sequel is the diameter of the leaves $\mathcal{S}_{\sigma,i}$ of a given cover $\mathcal{P}$. For a fixed integer $\sigma \geq 0$ and for any $i \leq 2^\sigma$, we have the following bound on the diameter of the leaves of a tree of height σ $\mathrm{diam}(\mathcal{S}_{\sigma,i}) \leq \frac{\mathrm{diam}(\mathcal{K})}{2^{\lfloor \sigma/d \rfloor}}$.

4.2. The hierarchical dual averaging algorithm

As a prelude to the definition of the hierarchical dual averaging algorithm, we introduce the following notions: for all $t = 1, 2, \dots$, (i) $\mathcal{P}_t$ will denote the current cover at time t; (ii) we will write σ_t for the number of splitting events made prior to time t (so σ_t is also the height of the tree $\mathcal{P}_t$); and (iii) $m_t = 2^{\sigma_t}$ will denote the number of leaves of $\mathcal{P}_t$. Moreover, a *splitting schedule* is an increasing sequence of integers $\mathcal{T}_{\text{split}} = \{t_1, t_2, \dots\}$ such that we perform a splitting event at each round $t \in \mathcal{T}_{\text{split}}$. For convenience, we will rather manipulate *scheduler sequences* $\{v_t\}_{t\geq 1}$, i.e., increasing real sequences that are uniquely mapped to a splitting schedule by $\mathcal{T}_{\text{split}}(v) = \{t \geq 1$ such that $\lfloor v_t \rfloor =$

[2] In a slight overload, we also write $\mathcal{P}$ for the tree inducing the cover, and therefore refer to its components as *leaves*

$\lfloor v_{t-1} \rfloor + 1\}$. In the sequel and when the context is non ambiguous we may use the term *splitting schedule* to refer to its associated *scheduler sequence*. We note that for all t, these definitions imply $\lfloor v_t \rfloor = \sigma_t$, and that in the light of the relation stated in the previous subsection we have, for any $\mathcal{S} \in \mathcal{P}_t$, $\operatorname{diam}(\mathcal{S}) \leq 2 \operatorname{diam}(\mathcal{K}) m_t^{-1/d}$ and $\lambda(\mathcal{S}) = m_t^{-1} \lambda(\mathcal{K})$.

We are now in a position to define our learning algorithm in detail. Its components are threefold: (i) a sequence of estimators $\hat{u}_t$ with range $\mathcal{R}$; (ii) a regularizer that is K-tame relative to $\mathcal{R}$; and (iii) a splitting schedule $\mathcal{T}_{\text{split}}(v)$ as above. Then, the *hierarchical dual averaging* (HDA) is defined via the recursion

$$\begin{aligned} S_{t+1} &= S_t + \hat{u}_t \\ x_{t+1} \sim X_{t+1} &= Q^{\mathcal{P}_t}(\eta_{t+1} S_{t+1}) \\ \mathcal{P}_{t+1} &= \begin{cases} \mathcal{P}_t^+ & \text{if } t \in \mathcal{T}_{\text{split}}(v) \\ \mathcal{P}_t & \text{otherwise.} \end{cases} \end{aligned} \tag{HDA}$$

where $Q^{\mathcal{P}}$ denotes the choice map induced by h for a given cover $\mathcal{P}$ of $\mathcal{K}$ (by convention, we take $\mathcal{P}_0 = \{\mathcal{K}\}$), η_t is a variable learning rate sequence, and we implicitly treat $\hat{u}_t$ and S_t as elements $\mathbb{R}^{\mathcal{P}_t}$ and X_t as an element of $\mathcal{Q}_{\mathcal{P}_t}$.

By construction, (HDA) comprises a succession of applications of the DAX sub-routine to sequences of successive rounds during which the underlying partition $\mathcal{P}_t$ stays the same (i.e., in between two successive splitting events). An important special case is the specific instance of (HDA) obtained by the entropic kernel $\theta(x) = x \log x$ (cf. Example 1) and the estimator (IWE); we will refer to this instance as the *hierarchical exponential weights* (HEW) algorithm.

5. Analysis and results

In this section, we leverage the regret guarantees established in Section 3 for (DAX) to derive a template regret bound for (HDA). We then use this result to derive both static and dynamic regret bounds for the specific case of HEW.

5.1. Static regret guarantees

Our template regret guarantee for HDA is as follows.

Theorem 2. *The HDA algorithm enjoys the regret bound*

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}_x(T)] \leq &\frac{\phi(m_T) + C_\theta \log_2(m_T)}{\eta_{T+1}} \\ &+ 2L \operatorname{diam}(\mathcal{K}) \sum\nolimits_{t=1}^{T} m_t^{-1/d} \\ &+ 2 \sum\nolimits_{t=1}^{T} \mu_t + \frac{1}{2K} \sum\nolimits_{t=1}^{T} \eta_t M_t^2, \end{aligned} \tag{21}$$

where m_t is the number of sets in the partition $\mathcal{P}_t$, $\phi(z) = z\theta(1/z)$ for all $z > 0$ and C_θ is a constant depending only on θ. In particular, if (HDA) *is run with learning rate $\eta_t \propto 1/t^{\varrho}$, $\varrho \in (0, 1)$, a logarithmic splitting schedule $v_t = p \log_2(t)$ and a sequence of estimators $\hat{u}_t$ such that $\mu_t = \mathcal{O}(1/t^{\beta})$ and $M_t^2 = \mathcal{O}(t^{2\mu})$ for some $\beta, \mu \geq 0$, then*

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathcal{O}(\phi(T^{-p})T^{\varrho} + T^{1-p/d} + T^{1-\beta} + T^{1+2\mu-\varrho}). \tag{22}$$

This general template bound can be used to derive tight regret bounds for particular instances of (i) estimator sequence $\{\hat{u}_t\}_t$ with range $\mathcal{R}$, (ii) regularizer h which is K-*tame* relative to $\mathcal{R}$ and (iii) splitting schedule $\{v_t\}_t$, as we show in the following corollary for the case of the HEW algorithm.

Corollary 1. *If HEW is run with learning rate $\eta_t \propto t^{-\varrho}$ and the logarithmic splitting schedule $v_t = p \log_2(t)$, the learner enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathcal{O}(T^{\varrho} + T^{1-p/d} + T^{1+p-\varrho}). \tag{23}$$

In particular, if the algorithm is run with $\varrho = (d+1)/(d+2)$ and $p = d/(d+2)$ we obtain the bound

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathcal{O}(T^{\frac{d+1}{d+2}}). \tag{24}$$

5.2. Dynamic regret guarantees

We now show guarantees of HDA in terms of the dynamic regret introduced in (3). We would like to stress that the expected dynamic regret of an algorithm cannot be bounded without any restriction on the sequence of payoffs (Shalev-Shwartz, 2011). For this reason, dynamic regret guarantees are often stated in terms of the *variation* of the payoff functions $\{u_t\}_t$, defined as follows (Besbes et al., 2015)

$$V_T := \sum\nolimits_{t=1}^{T} \|u_{t+1} - u_t\|_\infty, \tag{25}$$

with the convention $u_{t+1} = u_t$ for $t = T$. We then have:

Theorem 3. *Suppose that* (HDA) *is run with the negentropy kernel, and assumptions as in Theorem 2. Then:*

$$\begin{aligned} &\mathbb{E}[\mathrm{DynReg}(T)] \\ &= \mathcal{O}(T^{1+2\mu-\varrho} + T^{1-\beta} + T^{1-p/d} + T^{2\varrho-2\mu} V_T). \end{aligned} \tag{26}$$

Finally, with judiciously chosen parameters, our template bound yields the following improvement over previous dynamic regret bounds in the literature:

Corollary 2. *Suppose that HEW is run with a splitting schedule $v_t = p \log_2(t)$ and a learning rate $\eta_t \propto 1/t^{\varrho}$. Then:*

$$\mathbb{E}[\mathrm{DynReg}(T)] = \mathcal{O}(T^{1+p-\varrho} + T^{1-p/d} + T^{2\varrho-p} V_T). \tag{27}$$

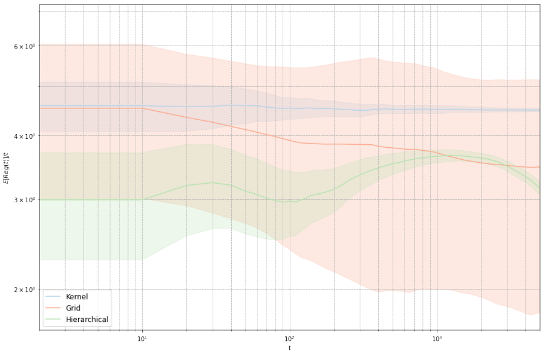


Figure 3: Expected average regret, averaged on 92 realizations for each algorithm (solid line). The variance is presented (shaded area) up to a standard deviation from the mean.

Hence, if $V_T = \mathcal{O}(T^{\nu})$ *for some* $\nu < 1$*, setting* $\varrho = (1-\nu)(d+1)/(d+3)$ *and* $p = (1-\nu)d/(d+3)$ *delivers*

$$\mathbb{E}[\mathrm{DynReg}(T)] = \mathcal{O}(T^{(d+2)/(d+3)} V_T^{1/(d+3)}). \tag{28}$$

This result was conjectured by Héliou et al. (2020) and, as far as we are aware, this is the first time it is achieved.

5.3. Numerical experiments

For illustration purposes, Fig. 3 provides some numerical experiments on different no-regret policies discussed in the rest of our paper. Specifically, we compared 3 strategies, "Grid", "Kernel" and "Hierarchical". The "Hierarchical" method is as outlined in Section 4 with parameters described below. The "Grid" method involves partitioning the search space into a grid of a given mesh-size (a hyperparameter), and then treats the problem as a finite-armed bandit, applying the EXP3 algorithm (Auer et al., 2002). Finally, the "Kernel" is based on (Héliou et al., 2020), using a squared-kernel based estimate. The adversarial function is analytic and randomly drawn, with known maximum. We present the full details of our experiments in the supplement.

References

Agarwal, N., Gonen, A., and Hazan, E. Learning in non-convex games with an optimization oracle. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

Agrawal, R. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, November 1995.

Alvarez, F., Bolte, J., and Brahic, O. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *ICLR '20: Proceedings of the 2020 International Conference on Learning Representations*, 2020.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47: 235–256, 2002.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

Berge, C. *Topological Spaces*. Dover, New York, 1997.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, October 2015.

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

Bubeck, S. and Eldan, R. Multi-scale exploration of convex functions and bandit convex optimization. In *COLT '16: Proceedings of the 29th Annual Conference on Learning Theory*, 2016.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. $\mathcal{X}$-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

Bubeck, S., Lee, Y. T., and Eldan, R. Kernel-based methods for bandit convex optimization. In *STOC '17: Proceedings of the 49th annual ACM SIGACT symposium on the Theory of Computing*, 2017.

Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

Conn, A. R., Scheinberg, K., and Vicente, L. N. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA '05: Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Folland, G. B. *Real Analysis*. Wiley-Interscience, 2 edition, 1999.

Hazan, E., Singh, K., and Zhang, C. Efficient regret minimization in non-convex games. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

Héliou, A., Martin, M., Mertikopoulos, P., and Rahier, T. Online non-convex optimization with imperfect feedback. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Kalai, A. T. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, October 2005.

Karmarkar, N. Riemannian geometry underlying interior point methods for linear programming. In *Mathematical Developments Arising from Linear Programming*, number 114 in Contemporary Mathematics. American Mathematical Society, 1990.

Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS' 04: Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

Kleinberg, R. D., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *STOC '08: Proceedings of the 40th annual ACM symposium on the Theory of Computing*, 2008.

Kleinberg, R. D., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *Journal of the ACM*, 66(4), May 2019.

Kocák, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

Krichene, W., Balandat, M., Tomlin, C., and Bayen, A. The Hedge algorithm on a continuum. In *ICML '15: Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Podimata, C. and Slivkins, A. Adaptive discretization for adversarial Lipschitz bandits. https://arxiv.org/abs/2006.12367, 2021.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Rosenbrock, H. H. An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3(3):175–184, 1960.

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

Shalev-Shwartz, S. and Singer, Y. Convex repeated games and Fenchel duality. In *NIPS' 06: Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, pp. 1265–1272. MIT Press, 2006.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, November 2019.

Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, 37(3):332–341, March 1992.

Suggala, A. S. and Netrapalli, P. Online non-convex learning: Following the perturbed leader is optimal. In *ALT '20: Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Vanderbei, R. J., Meketon, M. S., and Freedman, B. A. A modification of Karmarkar's linear programming algorithm. *Algorithmica*, 1(1):395–407, November 1986.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.

A. Fisher regularizers and their properties

Our goal here is to formally state and prove some basic properties for the regularizer functions that underlie the definition of (DAX). These properties are relatively well-known in the literature in the case where h is strongly convex relative to a global, reference norm; however, the use of the Fisher information metric introduces a number of complications that necessitate a more careful treatment.

We begin by recalling the basic setup of (DAX), as formalized in Section 3 for a fixed cover $\mathcal{P}$ of $\mathcal{K}$. In particular, we will write $\mathbb{R}^{\mathcal{P}}$ for the space of piecewise constant functions on $\mathcal{P}$ and $\mathcal{Q}_{\mathcal{P}}$ for the space of probability distributions supported on $\mathcal{P}$. Then, given $z \in \mathbb{R}^{\mathcal{P}}$ and $y \in \mathbb{R}^{\mathcal{P}}$, we define respectively the primal and dual Fisher norm relative to $q \in \mathcal{Q}_{\mathcal{P}}$ as

$$\|z\|_q^2 = \int_{\mathcal{K}} \frac{z(x)^2}{q(x)}\, d\lambda(x) = \sum_{\mathcal{S}\in\mathcal{P}} \frac{z_{\mathcal{S}}^2}{q_{\mathcal{S}}}, \tag{A.1a}$$

$$\|y\|_{q,*}^2 = \int_{\mathcal{K}} q(x) y(x)^2\, d\lambda(x) = \sum_{\mathcal{S}\in\mathcal{P}} q_{\mathcal{S}} y_{\mathcal{S}}^2. \tag{A.1b}$$

We then have the following basic lemma.

Lemma A.1. *With notation as above, we have*

1. $\|y\|_{q,*} = \max\{|\langle y, z\rangle| : \|z\|_q = 1\}$.
2. $\|\cdot\|_q \geq \|\cdot\|_1$ *and* $\|\cdot\|_{q,*} \leq \|\cdot\|_\infty$.

Proof. For the first part of our claim, an application of the Cauchy-Schwarz inequality gives

$$|\langle y, z\rangle| = \left|\sum_{\mathcal{S}\in\mathcal{P}} z_{\mathcal{S}} y_{\mathcal{S}}\right| = \left|\sum_{\mathcal{S}\in\mathcal{P}} \frac{z_{\mathcal{S}}}{\sqrt{q_{\mathcal{S}}}} \cdot \sqrt{q_{\mathcal{S}}} y_{\mathcal{S}}\right| \leq \|z\|_q \cdot \|y\|_{q,*}. \tag{A.2}$$

Since equality is attained when $z_{\mathcal{S}} \propto q_{\mathcal{S}} y_{\mathcal{S}}$, maximizing over the Fisher unit sphere $\|z\|_q = 1$ yields the desired result.

For the second part of our claim, a second application of the Cauchy-Schwarz inequality readily gives

$$\sum_{\mathcal{S}\in\mathcal{P}} \frac{z_{\mathcal{S}}^2}{q_{\mathcal{S}}} = \sum_{\mathcal{S}\in\mathcal{P}} q_{\mathcal{S}} \cdot \sum_{\mathcal{S}\in\mathcal{P}} \frac{z_{\mathcal{S}}^2}{q_{\mathcal{S}}} \geq \left(\sum_{\mathcal{S}\in\mathcal{P}} \sqrt{q_{\mathcal{S}}} \frac{|z_{\mathcal{S}}|}{\sqrt{q_{\mathcal{S}}}}\right)^2 = \|z\|_1^2, \tag{A.3}$$

i.e., $\|\cdot\|_q \geq \|\cdot\|_1$, as claimed. The inequality $\|\cdot\|_{q,*} \leq \|\cdot\|_\infty$ then follows by taking duals. ■

To proceed, recall that the convex conjugate h^* of h is defined as

$$h^*(y) = \sup_{q\in\mathcal{Q}_{\mathcal{P}}} \{\langle y, q\rangle - h(q)\}. \tag{A.4}$$

Since h is assumed strongly convex relative to the Fisher information metric, Lemma A.1 shows that it is also strongly convex relative to the ℓ_1-norm on $\mathcal{Q}_{\mathcal{P}}$. As a result, the supremum in (A.4) is always attained, and $h^*(y)$ is finite for all $y \in \mathbb{R}^{\mathcal{P}}$ (Bauschke & Combettes, 2017). Moreover, by standard results in convex analysis (Rockafellar, 1970, Chap. 26), it follows that h^* is differentiable on $\mathbb{R}^{\mathcal{P}}$; finally, by Danskin's theorem (Berge, 1997, Chap. 4), its gradient satisfies the identity

$$\nabla h^*(y) = \underset{q\in\mathcal{Q}_{\mathcal{P}}}{\arg\max}\{\langle y, q\rangle - h(q)\}. \tag{A.5}$$

Thus, recalling the definition (6) of the choice map $Q\colon \mathbb{R}^{\mathcal{P}} \to \mathcal{Q}_{\mathcal{P}}$, we get the equivalent expression

$$Q(y) = \nabla h^*(y). \tag{A.6}$$

For convenience and concision, any regularizer as above will be referred to as a *Fisher regularizer* on $\mathcal{Q}_{\mathcal{P}}$.

With this background in hand, we proceed to prove some auxiliary results and estimates that are used throughout the analysis of Sections 3 and 5. The first concerns the basic primal-dual properties of the choice map Q.

Lemma A.2. *Let h be a Fisher regularizer on $\mathcal{Q}_{\mathcal{P}}$. Then, for all $\chi \in \operatorname{dom}\partial h$ and all $y, v \in \mathbb{R}^{\mathcal{P}}$, we have:*

$$a)\quad \chi = Q(y) \iff y \in \partial h(\chi). \tag{A.7a}$$

$$b)\quad \chi^+ = Q(\nabla h(\chi) + v) \iff \nabla h(\chi) + v \in \partial h(\chi^+) \tag{A.7b}$$

Finally, if $\chi = Q(y)$ and $q \in \mathcal{Q}_{\mathcal{P}}$, we have

$$\langle \nabla h(\chi), \chi - q\rangle \leq \langle y, \chi - q\rangle. \tag{A.8}$$

Remark. Note that (A.7b) directly implies that $\partial h(q^+) \neq \varnothing$, i.e., $q^+ \in \operatorname{dom}\partial h$ for all $v \in \mathbb{R}^{\mathcal{P}}$. An immediate consequence of this is that the update rule $q^+ = Q(y+v)$ is *well-posed* for all $y \in \mathbb{R}^{\mathcal{P}}$, $v \in \mathcal{R}$, i.e., it can be iterated in perpetuity.

Proof of Lemma A.2. To prove (A.7a), note that χ solves (A.5) if and only if $y - \partial h(\chi) \ni 0$, i.e., if and only if $y \in \partial h(\chi)$. Eq. (A.7b) is then obtained in the same manner.

For the inequality (A.8), it suffices to show it holds for all $q \in \operatorname{ri}\mathcal{Q}_{\mathcal{P}}$ (by continuity). To do so, let

$$\phi(t) = h(\chi + t(q-\chi)) - [h(\chi) + \langle y, \chi + t(q-\chi)\rangle]. \tag{A.9}$$

Since h is strongly convex relative to the Fisher metric and $y \in \partial h(\chi)$ by (A.7a), it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Moreover, note that $\psi(t) = \langle \nabla h(\chi + t(q-\chi)) - y, q - \chi\rangle$ is a continuous selection of subderivatives of ϕ. Since ϕ and ψ are both continuous on $[0,1]$, it follows that ϕ is continuously differentiable and $\phi' = \psi$ on $[0,1]$. Thus, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0,1]$, we conclude that $\phi'(0) = \langle \nabla h(\chi) - y, q - \chi\rangle \geq 0$, from which our claim follows. ■

We now proceed to prove the basic properties of h and h^* relative to the primal and dual Fisher norms respectively. For convenience, we restate the relevant result below.

Lemma 2. *Let $h^*(y) = \max_{q\in\mathcal{Q}_{\mathcal{P}}}\{\langle y, q\rangle - h(q)\}$ be the convex conjugate of h. The following are equivalent:*

1. *h satisfies (11).*
2. *h^* is $(1/K)$-Lipschitz smooth relative to the dual Fisher norm $\|y\|_{q,*}^2 = \sum_{\mathcal{S}\in\mathcal{P}} q_{\mathcal{S}} y_{\mathcal{S}}^2$ on $\mathbb{R}^{\mathcal{P}}$; specifically, for all $y, v \in \mathbb{R}^{\mathcal{P}}$, we have*

$$h^*(y+v) \leq h^*(y) + \langle v, \chi\rangle + \frac{1}{2K}\|v\|_{\chi,*}^2, \tag{12}$$

where $\chi = Q(y)$.

Proof of Lemma 2. We begin with the direct implication "(1) $\implies$ (2)". For convenience, let $y^+ = y + v$, and set $\chi = Q(y)$, $\chi^+ = Q(y^+)$. We then have:

$$\begin{aligned} h^*(y^+) - h^*(y) - \langle v, \chi\rangle &= h^*(y^+) - \langle y^+, \chi\rangle + h(\chi) \\ &= \langle y^+, \chi^+\rangle - h(\chi^+) - \langle y^+, \chi\rangle + h(\chi) \\ &= h(\chi) - h(\chi^+) - \langle y^+, \chi - \chi^+\rangle. \end{aligned} \tag{A.10}$$

However, by Lemma A.2, we also have $y \in \partial h(\chi)$ and $y^+ \in \partial h(\chi^+)$. Hence, by the strong convexity of h relative to the Fisher information metric, we readily get

$$h(\chi^+) \geq h(\chi) + \langle y, \chi^+ - \chi\rangle + \frac{K}{2}\|\chi^+ - \chi\|_\chi^2. \tag{A.11}$$

Therefore, substituting (A.11) into (A.10) and rearranging, we obtain

$$\begin{aligned} h^*(y^+) - h^*(y) - \langle v, \chi\rangle &= h(\chi) - h(\chi^+) - \langle y^+, \chi - \chi^+\rangle \\ &\leq \langle y, \chi - \chi^+\rangle - \frac{K}{2}\|\chi^+ - \chi\|_\chi^2 - \langle y^+, \chi - \chi^+\rangle \\ &= \langle y^+ - y, \chi^+ - \chi\rangle - \frac{K}{2}\|\chi^+ - \chi\|_\chi^2 \\ &\leq \frac{K}{2}\|\chi^+ - \chi\|_\chi^2 + \frac{1}{2K}\|y^+ - y\|_{\chi,*}^2 - \frac{K}{2}\|\chi^+ - \chi\|_\chi^2 = \frac{1}{2K}\|y^+ - y\|_{\chi,*}^2 \end{aligned} \tag{A.12}$$

where, in the last line, we used Lemma A.1 to apply the Fenchel–Young inequality to the convex function $\phi(\cdot) = (K/2)\|\cdot\|_{\chi}^2$ and its conjugate $\phi^*(\cdot) = 1/(2K)\|\cdot\|_{\chi,*}^2$. Our claim then follows by a trivial rearrangement of (A.12).

For the converse direction "(2) $\implies$ (1)", fix some $\chi, \chi^+ \in \operatorname{ri} \mathcal{Q}_{\mathcal{P}}$, and let $y \in \partial h(\chi)$, $y^+ \in \partial h(\chi^+)$. Then, reversing (A.10) gives

$$h(\chi^+) - h(\chi) - \langle y, \chi^+ - \chi \rangle = h^*(y) - h^*(y^+) + \langle y^+ - y, \chi^+ \rangle. \tag{A.13}$$

However, by the Lipschitz smoothness of h^*, we have

$$h^*(y^+) \leq h^*(y) + \langle \chi, y^+ - y \rangle + \frac{1}{2K} \|y^+ - y\|_{\chi,*}^2 \tag{A.14}$$

and hence

$$\begin{aligned} h(\chi) - h(\chi^+) + \langle y, \chi^+ - \chi \rangle &= h^*(y^+) - h^*(y) - \langle y^+ - y, \chi^+ \rangle \\ &= h^*(y^+) - h^*(y) - \langle y^+ - y, \chi \rangle - \langle y^+ - y, \chi^+ - \chi \rangle \\ &\leq \frac{1}{2K} \|y^+ - y\|_{\chi,*}^2 - \frac{1}{2K} \|y^+ - y\|_{\chi,*}^2 - \frac{K}{2} \|\chi^+ - \chi\|_{\chi}^2 = -\frac{K}{2} \|\chi^+ - \chi\|_{\chi}^2, \end{aligned} \tag{A.15}$$

where we used the Fenchel–Young inequality as above. Our claim then follows by rearranging. ■

We now proceed to establish some of the basic properties for the Fenchel coupling

$$F(q, y) = h(q) + h^*(y) - \langle y, q \rangle. \tag{A.16}$$

The first property we present is a primal-dual analogue of the so-called "three-point identity" that is commonly used in the theory of Bregman functions (Chen & Teboulle, 1993).

Lemma A.3. *With notation as above, we have:*

$$F(q, y^+) = F(q, y) + F(\chi, y^+) + \langle y^+ - y, \chi - q \rangle. \tag{A.17}$$

Proof. By alternating the dual point of comparison in the definition of the Fenchel coupling, we have:

$$F(q, y^+) = h(q) + h^*(y^+) - \langle y^+, q \rangle \tag{A.18a}$$

$$F(q, y) = h(q) + h^*(y) - \langle y, q \rangle. \tag{A.18b}$$

Then, by subtracting (A.18a) from (A.18b), we get:

$$\begin{aligned} F(q, y^+) - F(q, y) &= h(q) + h^*(y^+) - \langle y^+, q \rangle - h(q) - h^*(y) + \langle y, q \rangle \\ &= h^*(y^+) - h^*(y) - \langle y^+ - y, q \rangle \\ &= h^*(y^+) - \langle y, Q(y) \rangle + h(Q(y)) - \langle y^+ - y, q \rangle \\ &= h^*(y^+) - \langle y, \chi \rangle + h(\chi) - \langle y^+ - y, q \rangle \\ &= h^*(y^+) + \langle y^+ - y, \chi \rangle - \langle y^+, \chi \rangle + h(\chi) - \langle y^+ - y, q \rangle \\ &= F(\chi, y^+) + \langle y^+ - y, \chi - q \rangle. \end{aligned}$$

■

We are now in a position to prove Lemma 3, which we restate below for convenience:

Lemma 3. *For all $y \in \mathbb{R}^{\mathcal{P}}$ and all $v \in \mathcal{R}$, we have*

$$F(q, y + v) = F(q, y) + \langle v, \chi - q \rangle + F(\chi, y + v) \tag{14a}$$

$$\leq F(q, y) + \langle v, \chi - q \rangle + \frac{1}{2K} \|v\|_{\chi,*}^2 \tag{14b}$$

where $\chi = Q(y)$.

Proof of Lemma 3. Let $y^+ = y + v$. Then, by the three-point identity (A.17), we readily get

$$F(q, y^+) = F(q, y) + \langle v, \chi - q \rangle + F(\chi, y^+) \tag{A.19}$$

so we are left to show that $F(\chi, y^+) \leq 1/(2K)\|y^+ - y\|^2_{\chi,*}$. To that end, Lemma 2 yields

$$\begin{aligned} F(\chi, y^+) &= h(\chi) + h^*(y^+) - \langle y^+, \chi \rangle \\ &\leq h(\chi) + h^*(y) + \langle y^+ - y, \chi \rangle + \frac{1}{2K}\|y^+ - y\|^2_{\chi,*} - \langle y^+, \chi \rangle \\ &= h(\chi) + h^*(y) - \langle y, \chi \rangle + \frac{1}{2K}\|y^+ - y\|^2_{\chi,*} \\ &= \frac{1}{2K}\|y^+ - y\|^2_{\chi,*} \end{aligned} \tag{A.20}$$

where we used the fact that $\chi = Q(y)$, so $h(\chi) + h^*(y) - \langle y, \chi \rangle = 0$. ■

We close this section by discussing the properties of the negentropy regularizer $\theta(x) = x \log x$. Regarding the strong convexity of this regularizer relative to the Fisher information metric, we would need θ to satisfy the condition

$$\theta(q) \geq \theta(x) + \theta'(x)(q - x) + \frac{K}{2}\frac{(q - x)^2}{x} \tag{A.21}$$

for some $K > 0$ and for all $q, x \in (0, 1)$. Rearranging the above inequality, and recalling the definition of the Kullback-Leibler divergence $D_{\mathrm{KL}}(q, x) = q \log(q/x)$, this requirement boils down to

$$D_{\mathrm{KL}}(q, x) \geq (q - x) + \frac{K}{2}\frac{(q - x)^2}{x} \tag{A.22}$$

for some $K > 0$ and for all $q, x \in (0, 1)$. However, for any *fixed* $q \in (0, 1)$, the right-hand side of the above equation exhibits an $\Omega(1/x)$ singularity as $x \to 0^+$, while the left-hand side grows as $\mathcal{O}(\log x)$. As a result, we conclude that the negentropy regularizer is *not* strongly convex relative to the Fisher information metric.

On the other hand, as we show below, the entropy is *tame* relative to the estimation region $\mathcal{R} = (-\infty, R]^d$. To see this, note that $h^*(y) = \log \sum_{\mathcal{S} \in \mathcal{P}} e^{y_{\mathcal{S}}}$, so

$$h^*(y + v) - h^*(y) = \log \frac{\sum_{\mathcal{S} \in \mathcal{P}} \exp(y_{\mathcal{S}} + v_{\mathcal{S}})}{\sum_{\mathcal{S} \in \mathcal{P}} \exp(y_{\mathcal{S}})} = \log \sum_{\mathcal{S} \in \mathcal{P}} \chi_{\mathcal{S}} \exp(v_{\mathcal{S}}). \tag{A.23}$$

Now, if $v \in \mathcal{R}$, we have $v_{\mathcal{S}} \leq R$ for all $\mathcal{S} \in \mathcal{P}$, so there exists some $K > 0$ such that $\exp(v_{\mathcal{S}}) \leq 1 + v_{\mathcal{S}} + v_{\mathcal{S}}^2/(2K)$ for all $\mathcal{S} \in \mathcal{P}$. Then, plugging this estimate into (A.23), we conclude that

$$\begin{aligned} h^*(y + v) - h^*(y) &\leq \log \sum_{\mathcal{S} \in \mathcal{P}} \chi_{\mathcal{S}}(1 + v_{\mathcal{S}} + \frac{v_{\mathcal{S}}^2}{2K}) = \log\left(1 + \langle v, \chi \rangle + \frac{1}{2K}\sum_{\mathcal{S} \in \mathcal{P}} \chi_{\mathcal{S}} v_{\mathcal{S}}^2\right) \\ &\leq 1 + \langle v, \chi \rangle + \frac{1}{2K}\sum_{\mathcal{S} \in \mathcal{P}} \chi_{\mathcal{S}} v_{\mathcal{S}}^2. \end{aligned} \tag{A.24}$$

The specific value of K is $1/2$ if $R = 1$; for general R, the value of K can be estimated by backsolving the equation $1 + R + R^2/(2K) = \exp(R)$.

B. Regret guarantees for dual averaging with an explicit cover

In this appendix, our aim is to prove the rest of the results presented in Section 3 for (DAX). We begin with the algorithm's template bound for the η-deflated Fenchel coupling $E_t = \frac{1}{\eta_t} F(q, \eta_t S_t)$ as defined in (15); for convenience, we restate the relevant result below.

Lemma 4. *Suppose that* (DAX) *is run with an estimator with range $\mathcal{R}$. For all $t = 1, 2, \dots$, we have*

$$E_{t+1} \leq E_t + \langle \hat{u}_t, X_t - q \rangle + \left(\eta_{t+1}^{-1} - \eta_t^{-1}\right)[h(q) - \min h] \\ + \eta_t^{-1} F(X_t, \eta_t S_{t+1}). \tag{16}$$

If, in addition, h is K-tame relative to $\mathcal{R}$, the last term in (16) *is bounded as*

$$\eta_t^{-1} F(X_t, \eta_t S_{t+1}) \leq \eta_t/(2K) \|\hat{u}_t\|_t^2, \tag{17}$$

where $\|\cdot\|_t$ is the dual Fisher norm $\|v\|_t \coloneqq \|v\|_{X_t,*}$.

Proof of Lemma 4. Our proof follows the general structure of the proof of Héliou et al. (2020, Lemma 2); however, the use of the Fisher information metric instead of a global norm introduces a number of subtleties that require special care.

We begin by rewriting the difference $E_{t+1} - E_t$ as

$$E_{t+1} - E_t = \frac{1}{\eta_{t+1}} F(q, \eta_{t+1} S_{t+1}) - \frac{1}{\eta_t} F(q, \eta_t S_t) = \frac{1}{\eta_{t+1}} F(q, \eta_{t+1} S_{t+1}) - \frac{1}{\eta_t} F(q, \eta_t S_{t+1}) \tag{B.1a}$$

$$+ \frac{1}{\eta_t} F(q, \eta_t S_{t+1}) - \frac{1}{\eta_t} F(q, \eta_t S_t). \tag{B.1b}$$

We will proceed to bound each of these terms separately.

Beginning with the latter, the first part of Lemma 3 allows us to rewrite (B.1b) as

$$\text{(B.1b)} = \frac{1}{\eta_t}[F(q, \eta_t S_t + \eta_t \hat{u}_t) - F(q, \eta_t S_t)] = \frac{1}{\eta_t}[F(X_t, \eta_t S_{t+1}) + \langle \eta_t \hat{u}_t, X_t - q \rangle] \\ = \frac{F(X_t, \eta_t S_{t+1})}{\eta_t} + \langle \hat{u}_t, X_t - q \rangle \tag{B.2}$$

where we used the fact that $X_t = Q(\eta_t S_t)$ by the definition of (DAX). As for the term (B.1a), we readily have

$$\text{(B.1a)} = \left[\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right] h(q) + \frac{1}{\eta_{t+1}} h^*(\eta_{t+1} S_{t+1}) - \frac{1}{\eta_t} h^*(\eta_t S_{t+1}) \tag{B.3}$$

by the definition (13) of the Fenchel coupling. We will proceed to bound this term by studying the function $\varphi(\eta) = \eta^{-1}[h^*(\eta y) + \min h]$ as a function of η for a *fixed* $y \in \mathbb{R}^{\mathcal{P}}$. To that end, using Lemma A.2 to differentiate φ gives

$$\varphi'(\eta) = \frac{1}{\eta}\langle y, Q(\eta y) \rangle - \frac{1}{\eta^2}[h^*(\eta y) + \min h] = \frac{1}{\eta^2}[h(Q(\eta y)) - \min h] \geq 0, \tag{B.4}$$

where we used the fact that $\langle \eta y, Q(\eta y) \rangle - h^*(\eta y) = h(Q(\eta y))$. Thus, with $\eta_{t+1} \leq \eta_t$ for all $t = 1, 2, \dots$, we conclude that $\varphi(\eta_t) \geq \varphi(\eta_{t+1})$, and hence:

$$\frac{1}{\eta_{t+1}} h^*(\eta_{t+1} S_{t+1}) - \frac{1}{\eta_t} h^*(\eta_t S_{t+1}) \leq \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right] \min h. \tag{B.5}$$

Thus, recombining everything in (B.1), we obtain (16), as claimed.

Finally, for (17), recall that the first part of Lemma 3 is valid independently of the strong convexity modulus of h relative to the Fisher metric. Thus, by invoking the assumption that h is K-tame relative to $\mathcal{R}$, we get

$$\begin{aligned} F(X_t, \eta_t S_{t+1}) &= h(X_t) + h^*(\eta_t S_{t+1}) - \langle \eta_t S_{t+1}, X_t \rangle \\ &= h(X_t) + h^*(\eta_t S_t + \eta_t \hat{u}_t) - \eta_t \langle S_{t+1}, X_t \rangle \\ &\leq h(X_t) + h^*(\eta_t S_t) + \eta_t \langle \hat{u}_t, X_t \rangle + \frac{\eta_t^2}{2K} \|\hat{u}_t\|_{X_t,*}^2 - \langle \eta_t S_{t+1}, X_t \rangle \\ &= h(X_t) + h^*(\eta_t S_t) - \langle \eta_t S_t, X_t \rangle + \frac{\eta_t^2}{2K} \|\hat{u}_t\|_t^2 \\ &= \frac{\eta_t^2}{2K} \|\hat{u}_t\|_t^2 \end{aligned} \tag{B.6}$$

where, in the last line, we used the fact that $X_t = Q(\eta_t S_t)$, so $h(X_t) + h^*(\eta_t S_t) - \langle \eta_t S_t, X_t \rangle = 0$. Thus, dividing both sides of the above inequality by η_t yields the desired result. ■

We are now in a position to prove our template regret bounds for (DAX); for completeness, we restate them both below.

Proposition 1. *The regret incurred by the learner relative to $q \in \mathcal{Q}_{\mathcal{P}}$ over the interval $\mathcal{T} = \{t_1, \dots, t_2 - 1\}$ is bounded as*

$$\begin{aligned} \operatorname{Reg}_q(\mathcal{T}) \leq E_{t_1} - E_{t_2} + \left(\eta_{t_2}^{-1} - \eta_{t_1}^{-1}\right)[h(q) - \min h] \\ + \sum_{t \in \mathcal{T}} \langle Z_t, X_t - q \rangle + \frac{1}{2K} \sum_{t \in \mathcal{T}} \eta_t \|\hat{u}_t\|_t^2. \end{aligned} \tag{18}$$

Theorem 1. *Suppose that (DAX) is run with assumptions as in Proposition 1. Then the learner's regret is bounded as*

$$\begin{aligned} \mathbb{E}[\operatorname{Reg}_q(T)] \leq E_1 - E_{T+1} + \left(\eta_{T+1}^{-1} - \eta_1^{-1}\right)[h(q) - \min h] \\ + 2 \sum_{t=1}^{T} \mu_t + \frac{1}{2K} \sum_{t=1}^{T} \eta_t M_t^2. \end{aligned} \tag{20}$$

Proof of Proposition 1. Substituting $\hat{u}_t \leftarrow u_t + Z_t$ in (17) and rearranging, we get

$$\langle u_t, q - X_t \rangle = E_t - E_{t+1} + \langle Z_t, X_t - q \rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)[h(q) - \min h] + \frac{\eta_t}{2K} \|\hat{u}_t\|_t^2. \tag{B.7}$$

Our claim then follows by summing the above over $t \in \mathcal{T} = \{t_1, \dots, t_2 - 1\}$. ■

Proof of Theorem 1. Simply set $t_1 \leftarrow 1$, $t_2 \leftarrow T$ in (18) and take expectations. ■

C. Regret guarantees for hierarchical dual averaging

C.1. Static regret guarantees

In the first part of this appendix, our aim is to prove the regret guarantees of (HDA) against static comparators, as presented in Theorem 2 below.

Theorem 2. *The HDA algorithm enjoys the regret bound*

$$\begin{aligned} \mathbb{E}[\operatorname{Reg}_x(T)] \leq \frac{\phi(m_T) + C_\theta \log_2(m_T)}{\eta_{T+1}} \\ + 2L \operatorname{diam}(\mathcal{K}) \sum\nolimits_{t=1}^{T} m_t^{-1/d} \\ + 2 \sum\nolimits_{t=1}^{T} \mu_t + \frac{1}{2K} \sum\nolimits_{t=1}^{T} \eta_t M_t^2, \end{aligned} \tag{21}$$

where m_t is the number of sets in the partition $\mathcal{P}_t$, $\phi(z) = z\theta(1/z)$ for all $z > 0$ and C_θ is a constant depending only on θ. In particular, if (HDA) is run with learning rate $\eta_t \propto 1/t^{\varrho}$, $\varrho \in (0,1)$, a logarithmic splitting schedule $v_t = p \log_2(t)$ and a sequence of estimators $\hat{u}_t$ such that $\mu_t = \mathcal{O}(1/t^{\beta})$ and $M_t^2 = \mathcal{O}(t^{2\mu})$ for some $\beta, \mu \geq 0$, then

$$\mathbb{E}[\operatorname{Reg}(T)] = \mathcal{O}(\phi(T^{-p})T^{\varrho} + T^{1-p/d} + T^{1-\beta} + T^{1+2\mu-\varrho}). \tag{22}$$

Overview. Our proof hinges on applying Proposition 1 to bound the regret of (HDA) on each time window during which the algorithm maintains a constant cover of $\mathcal{K}$. Aggregating these bounds provides a regret guarantee for (HDA) over the entire horizon time of play; however, since the algorithm is *not* restarted at each window, joining the resulting window-by-window bounds ends up being fairly delicate. The main dificulties (and associated contributing terms in the regret) are as follows:

1. A comparator for a given time frame may not be admissible for a previous time frame because the granularity of an antecedent cover may not suffice to include the comparator in question. This propagates a "resolution error" that becomes smaller when the cover gets finer, but larger when the window gets longer.
2. At every change of window, the algorithm retains the same probability distribution over $\mathcal{K}$ (to avoid restart-forget effects). However, this introduces a "splitting residue" term in the regret because of the necessary correction in the learner's scores when the resolution of the cover increases.

The covering hierarchy. We begin by detailing how the algorithm unfolds window-by-window. Referring to Section 4 for the relevant definitions, consider a splitting schedule $\mathcal{T}_{\text{split}} = \{t_j\}_{1 \leq j \leq \sigma_T}$ where we recall that σ_t is the number of splitting events which occurred before round t. For every $j \in \{1, \dots, \sigma_T\}$, we define $\mathcal{T}_j = \{t_j, \dots, t_{j+1} - 1\}$, the time window between the j-th and the $(j+1)$-th splitting event. By convention, we denote $t_{\sigma_T+1} = T+1$: the last time window is therefore $\mathcal{T}_{\sigma_T} = \{t_{\sigma_T}, \dots, T\}$ and is a priori "incomplete" since the $(\sigma_T + 1)$-th splitting time has not been reached yet at time T.

Now, during each window $\mathcal{T}_j$, the underlying partition $\mathcal{P}_j$ contains $m_{t_j} = 2^j$ components and is *fixed* throughout this window. At each time $t_j \in \mathcal{T}_{\text{split}}$, a splitting event is performed on $\mathcal{P}_{j-1}$, in order to obtain $\mathcal{P}_j$ by splitting in two each set in $\mathcal{P}_{j-1}$ such that $\mathcal{P}_j = \mathcal{P}_{j-1}^+$, as described in (HDA). Then, for a *fixed* point $x \in \mathcal{K}$ and all $j = 1, 2, \dots$, we define the corresponding *approximate identity at* x to be the simple strategy $q_j^x \in \mathcal{Q}_{\mathcal{P}_j}$ such that

$$q_j^x(\mathcal{S}) = \mathbb{1}(x \in \mathcal{S}) \quad \text{for all } \mathcal{S} \in \mathcal{P}_j \tag{C.1}$$

i.e., q_j^x is the best approximation of δ_x among simple strategies of $\mathcal{Q}_{\mathcal{P}_j}$. In the following, we will write $\mathcal{S}_j^x$ for the support of q_j^x, i.e., for the unique covering element of $\mathcal{P}_j$ containing x.

With this background in hand, Lemma 1 yields

$$\operatorname{Reg}_x(\mathcal{T}_j) \leq \operatorname{Reg}_{q_j^x}(\mathcal{T}_j) + L \operatorname{diam}(\mathcal{S}_j^x)|\mathcal{T}_j| \tag{C.2}$$

where, by definition, $|\mathcal{T}_j| = t_{j+1} - t_j$. In turn, this allows us to bound $\operatorname{diam}(\mathcal{S}_j^x)$ with respect to m_{t_j}, the number of sets in the partition $\mathcal{P}_j$.

Lemma C.1. *If $\mathcal{P}$ is the partition of $\mathcal{K}$ after σ splitting events (and therefore containing $m = 2^\sigma$ covering sets), then, for all $\mathcal{S} \in \mathcal{P}$, we have*

$$\operatorname{diam}(\mathcal{S}) \leq 2 \operatorname{diam}(\mathcal{K}) m^{-1/d} \tag{C.3}$$

where diam *is defined with respect to the ambient norm of $\mathcal{K} \subseteq \mathbb{R}^d$.*

Proof. Let $\mathcal{P}$ be the partition of $\mathcal{K}$ after σ splitting events, which contains $m = 2^\sigma$ covering sets.

For any set $\mathcal{S} \in \mathcal{P}$, we have the following bound for $\operatorname{diam}(\mathcal{S})$ (see for example (Bubeck et al., 2011)):

$$\operatorname{diam}(\mathcal{S}) \leq \frac{\operatorname{diam}(\mathcal{K})}{2^{\lfloor \sigma/d \rfloor}},$$

We may now write the following sequence of inequalities using the definition of $\lfloor . \rfloor$ and the fact that $x \mapsto 2^x$ is increasing

$$\begin{aligned} \lfloor \sigma/d \rfloor &> \sigma/d - 1 \\ 2^{\lfloor \sigma/d \rfloor} &> \frac{1}{2} 2^{\sigma/d} \\ 2^{-\lfloor \sigma/d \rfloor} &< 2 \times 2^{-\sigma/d} \\ \operatorname{diam}(\mathcal{K}) 2^{-\lfloor \sigma/d \rfloor} &< 2 \operatorname{diam}(\mathcal{K}) 2^{-\sigma/d}. \end{aligned}$$

Now, using the fact that $m = 2^\sigma$, we finally get that for any $\mathcal{S} \in \mathcal{P}$,

$$\operatorname{diam}(\mathcal{S}) \leq 2 \operatorname{diam}(\mathcal{K}) 2^{-1/d}.$$

■

Aggregating cover bounds. To proceed, injecting the estimate of Lemma C.1 into (C.2) delivers

$$\operatorname{Reg}_x(\mathcal{T}_j) \leq \operatorname{Reg}_{q_j^x}(\mathcal{T}_j) + 2L \operatorname{diam}(\mathcal{K}) |\mathcal{T}_j| m_{t_j}^{-1/d}. \tag{C.4}$$

and hence, by Proposition 1 applied to $q_j^x \in \mathcal{Q}_{\mathcal{P}_j}$, we get

$$\begin{aligned} \operatorname{Reg}_x(\mathcal{T}_j) \leq E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j} + \left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right) \left[h^{\mathcal{P}_j}(q_j^x) - \min h^{\mathcal{P}_j}\right] \\ + \sum_{t \in \mathcal{T}_j} \langle Z_t, X_t - q_j^x \rangle + \frac{1}{2K} \sum_{t \in \mathcal{T}_j} \eta_t \|\hat{u}_t\|_t^2 + 2L \operatorname{diam}(\mathcal{K}) |\mathcal{T}_j| m_{t_j}^{-1/d} \end{aligned} \tag{C.5}$$

Noting that $\min h^{\mathcal{P}_j} = \phi(m_{t_j})$ where $\phi(z) = z\theta(1/z)$ for all $z > 0$, and that $h(q_j^x) = 0$ we can write

$$h^{\mathcal{P}_j}(q_j^x) - \min h^{\mathcal{P}_j} = \phi(m_{t_j}) \tag{C.6}$$

leading in turn to the expression

$$\begin{aligned}\operatorname{Reg}_x(\mathcal{T}_j) \leq E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j} + \left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right)\phi(m_{t_j}) \\ + \sum_{t\in\mathcal{T}_j} \langle Z_t, X_t - q_j^x \rangle + \frac{1}{2K}\sum_{t\in\mathcal{T}_j} \eta_t \|\hat{u}_t\|_t^2 + 2L\operatorname{diam}(\mathcal{K})|\mathcal{T}_j| m_{t_j}^{-1/d}\end{aligned} \tag{C.7}$$

where we have made an explicit reference to the underlying partition in the exponent of the E and h terms. As indicated by the presence of the term $\sum_{j=2}^{\sigma_T}\left[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\right]$, this subtlety is crucial for the algorithm's regret, as it accounts for the cost of descending to a cover with higher granularity.

To make this precise, note that the regret incurred by (HDA) over T stages can be decomposed as

$$\operatorname{Reg}_x(T) = \sum_{j=1}^{\sigma_T} \operatorname{Reg}_x(\mathcal{T}_j) \tag{C.8}$$

where each $\operatorname{Reg}_x(\mathcal{T}_j)$ corresponds to the regret incurred by (HDA) on a fixed partition – i.e., the regret induced by (DAX) over the said partition, assuming the algorithm was initialized at the last state of the previous window (since the algorithm does not restart). Then, combining (C.7) and (C.8), we get

$$\begin{aligned}\operatorname{Reg}_x(T) &= \sum_{j=1}^{\sigma_T} \operatorname{Reg}_x(\mathcal{T}_j) \\ &\leq \sum_{j=1}^{\sigma_T}\left[E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j}\right] + \sum_{j=1}^{\sigma_T}\phi(m_{t_j})\left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right) \\ &\qquad + \sum_{j=1}^{\sigma_T}\sum_{t\in\mathcal{T}_j}\langle Z_t, X_t - q_j^x\rangle + \frac{1}{2K}\sum_{j=1}^{\sigma_T}\sum_{t\in\mathcal{T}_j}\eta_t\|\hat{u}_t\|_t^2 + 2L\operatorname{diam}(\mathcal{K})\sum_{j=1}^{\sigma_T}|\mathcal{T}_j| m_{t_j}^{-1/d} \\ &= E_{t_1}^{\mathcal{P}_1} - E_{t_{\sigma_T}}^{\mathcal{P}_{\sigma_T+1}} + \sum_{j=2}^{\sigma_T}\left[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\right] + \phi(m_T)\eta_{T+1}^{-1} - \phi(1)\eta_1^{-1} + \eta_T^{-1}\sum_{j=2}^{\sigma_T}\left(\phi(m_{t_{j-1}}) - \phi(m_{t_j})\right) \\ &\qquad + \sum_{j=1}^{\sigma_T}\sum_{t\in\mathcal{T}_j}\langle Z_t, X_t - q_j^x\rangle + \frac{1}{2K}\sum_{t=1}^{T}\eta_t\|\hat{u}_t\|_t^2 + 2L\operatorname{diam}(\mathcal{K})\sum_{t=1}^{T} m_t^{-1/d}.\end{aligned} \tag{C.9}$$

where we used the fact that η_t is nonincreasing. Thus, noting that $E_{t_{\sigma_T+1}}^{\mathcal{P}_{\sigma_T}} = E_{T+1}^{\mathcal{P}_{\sigma_T}} \geq 0$ and $E_1^{\mathcal{P}_1} = \eta_1^{-1}\left[h^{\mathcal{P}_1}(q_1^x) + h^*(0)\right] = \phi(1)/\eta_1$, we get the following bound for the regret incurred by (HDA):

$$\begin{aligned}\operatorname{Reg}_x(T) \leq \sum_{j=2}^{\sigma_T}\left[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\right] + \eta_T^{-1}\sum_{j=2}^{\sigma_T}\left(\phi(m_{t_{j-1}}) - \phi(m_{t_j})\right) + \phi(m_T)\eta_{t_{T+1}}^{-1} \\ + \sum_{j=1}^{\sigma_T}\sum_{t\in\mathcal{T}_j}\langle Z_t, X_t - q_j^x\rangle + \frac{1}{2K}\sum_{t=1}^{T}\eta_t\|\hat{u}_t\|_t^2 + 2L\operatorname{diam}(\mathcal{K})\sum_{t=1}^{T} m_t^{-1/d}.\end{aligned} \tag{C.10}$$

Controlling the growth of each term in the above will be the main focus of our analysis in the sequel.

The splitting residue. Somewhat surprisingly, the first two terms of (C.10) turn out to be the most challenging ones to control. Because both terms are due to the algorithm's hierarchical splitting schedule, we will refer collectively to the sum

$$\epsilon_T = \sum_{j=2}^{\sigma_T}\left[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\right] + \eta_T^{-1}\sum_{j=2}^{\sigma_T}\left(\phi(m_{t_{j-1}}) - \phi(m_{t_j})\right), \tag{C.11}$$

as the algorithm's *splitting residue*.

To analyze this term, given a regularization kernel θ and a partition $\mathcal{P}$ of $\mathcal{K}$, let $h^{\mathcal{P}}$ be the corresponding (decomposable) regularizer induced by θ on $\mathcal{Q}_{\mathcal{P}}$, and write $Q^{\mathcal{P}}$ for the associated choice map $Q^{\mathcal{P}}: \mathbb{R}^{\mathcal{P}} \to \mathcal{Q}_{\mathcal{P}}$. Moreover, given a simple strategy $X \in \mathcal{Q}_{\mathcal{P}}$ and recalling that $\mathcal{P}^+$ denotes the successor of $\mathcal{P}$ after a splitting event, we will write $X^+ \in \mathcal{Q}_{\mathcal{P}+}$ for the mixed strategy on $\mathcal{P}^+$ such that the canonical cast of X and X^+ as distributions (with piecewise constant densities) on $\mathcal{K}$ are the same. Finally, for $S \in \mathbb{R}^{\mathcal{P}}$ such that $Q^{\mathcal{P}}(S) = q$, we will write S^+ for any piecewise constant function in $\mathbb{R}^{\mathcal{P}^+}$ such that $Q^{\mathcal{P}^+}(S^+) = q^+$.[3]

With all this in hand, our next result provides an an inverse-rate proportional upper bound for the *splitting residue* term ϵ_T.

Lemma C.2. *Let $K_\theta = \sup_{a\in(0,1]}[\theta(a) - 2\theta(a/2)]/a$ and $K'_\theta = \sup_S \inf_{S^+} \|S^+ - S\|_\infty$. Then*

$$\sum_{j=2}^{\sigma_T} \Big[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\Big] \le (K_\theta + 2K'_\theta)\frac{\sigma_T - 1}{\eta_T} \tag{C.12a}$$

$$\frac{1}{\eta_T}\sum_{j=2}^{\sigma_T}\big(\phi(m_{t_{j-1}}) - \phi(m_{t_j})\big) \le K_\theta \frac{\sigma_T - 1}{\eta_T} \tag{C.12b}$$

Remark 3. In the above, S and S^+ are viewed as piecewise constant functions of $\mathcal{K}$, with respective covers $\mathcal{P}$ and $\mathcal{P}^+$. As an example, the negentropy regularizer $\theta(x) = x\log x$ has $K_\theta = K'_\theta = \log 2$.

Proof of Lemma C.2. The series of calculations required to prove the bounds (C.12) is quite intreicate and needs a fair amount of groundwork. First, for a given $j \in \{1, \dots, \sigma_T\}$ we will use a "$-$" exponent to refer to quantities that *would have existed if there had not been a splitting event at time t_j*, and a "$+$" exponent to refer to quantities that are derived in a scenario where there *is indeed a splitting event happening at t_j*. For more concreteness, let $y_{t_j-1} \in \mathbb{R}^{\mathcal{P}_j}$ be the score function at time t_j. Then $y^-_{t_j}$ is such that $y^-_{t_j} = y_{t_j-1} + \hat{u}_{t_j}$, i.e., it is still an element of $\mathbb{R}^{\mathcal{P}_{j-1}}$, and correspond of what to the score at time t_j is no splitting event happens, then we have $X^-_{t_j} = Q^{\mathcal{P}_{j-1}}(\eta_{t_j} y^-_{t_j})$ is the corresponding probability distribution on $\mathcal{P}_{j-1}$. On the contrary, $y^+_{t_j}$ is an element of $\mathbb{R}^{\mathcal{P}_j}$ and is such that $X^+_{t_j} = Q^{\mathcal{P}_j}(\eta_{t_j} y^+_{t_j})$ is *consistent* with $X^-_{t_j}$, i.e., their cast as densities of $\mathcal{K}$ are *the same*.

These distinctions are subtle, but essential to grasp the meaning of the splitting residue ϵ_T. To streamline the proof, let us decompose this terms into two terms as follows:

$$\epsilon_T = \underbrace{\sum_{j=2}^{\sigma_T}\Big[E_{t_j}^{\mathcal{P}_j} - E_{t_j}^{\mathcal{P}_{j-1}}\Big]}_{A_T} + \underbrace{\frac{1}{\eta_T}\sum_{j=2}^{\sigma_T}\big(\phi(m_{t_{j-1}}) - \phi(m_{t_j})\big)}_{B_T} \tag{C.13}$$

We bound each of these terms individually below.

Step 1: Bounding A_T. Let $j \in \{1, \dots, \sigma_T\}$. Using the definition of the energy and the notations introduced above we have:

$$E_{t_j}^{\mathcal{P}_{j-1}} = \eta_{t_j}^{-1}\Big[h^{\mathcal{P}_{j-1}}(q^x_{j-1}) + h^{*\mathcal{P}_{j-1}}(\eta_{t_j} y^-_{t_j}) - \Big\langle \eta_{t_j} y^-_{t_j}, q^x_{j-1}\Big\rangle^{\mathcal{P}_{j-1}}\Big] \tag{C.14}$$

We may drop the explicit reference to the underlying partition in exponents of h, h^* and $\langle .,.\rangle$, since there respective arguments now explicitly belong to $\mathcal{Q}_{\mathcal{P}_{j-1}}$ and $\mathbb{R}^{\mathcal{P}_{j-1}}$. Using the fact that $X^-_{t_j} = Q(\eta_{t_j} y^-_{t_j})$, we can write that

$$h^*(\eta_{t_j} y^-_{t_j}) = \Big\langle \eta_{t_j} y^-_{t_j}, X^-_{t_j}\Big\rangle - h(X^-_{t_j}). \tag{C.15}$$

Injecting this in (C.14), and proceeding similarly for $E_{t_j}^{\mathcal{P}_j}$ finally gives

$$E_{t_j}^{\mathcal{P}_{j-1}} = \Big\langle y^-_{t_j}, X^-_{t_j} - q^x_{j-1}\Big\rangle - \eta_{t_j}^{-1} h(X^-_{t_j}) \tag{C.16}$$

$$E_{t_j}^{\mathcal{P}_j} = \Big\langle y^+_{t_j}, X^+_{t_j} - q^x_j\Big\rangle - \eta_{t_j}^{-1} h(X^+_{t_j}) \tag{C.17}$$

[3] Any two such functions will only differ by a constant. This constant plays no role in our analysis, so we will ignore it in the sequel.

Now using the fact that $\{\eta_t\}_t$ is decreasing, we can bound A_T as

$$A_T \leq \eta_T^{-1} \underbrace{\sum_{j=2}^{\sigma_T} \Big[h(X_{t_j}^-) - h(X_{t_j}^+) \Big]}_{A_T^{(1)}} + \underbrace{\sum_{j=2}^{\sigma_T} \Big[\Big\langle y_{t_j}^+, X_{t_j}^+ - q_j^x \Big\rangle - \Big\langle y_{t_j}^-, X_{t_j}^- - q_{j-1}^x \Big\rangle \Big]}_{A_T^{(2)}} . \tag{C.18}$$

We recall that the k^{th} set of partition $\mathcal{P}_{j-1}$ is split into the (equally-sized) $2k^{th}$ and $(2k+1)^{th}$ sets of partition $\mathcal{P}_j$, and that we ensure distributions $X_{t_j}^+$ and $X_{t_j}^-$ have the same canonical cast as a distribution on $\mathcal{K}$, i.e., $X_{t_j,2k}^+ = X_{t_j,2k+1}^+ = X_{t_j,k}^-/2$. Now using the decomposability of regularizer h gives for any j,

$$\begin{aligned} h(X_{t_j}^-) - h(X_{t_j}^+) &= \sum_{k=1}^{2^{j-1}} \theta(X_{t_j,k}^-) - \sum_{k=1}^{2^j} \theta(X_{t_j,k}^+) \\ &= \sum_{k=1}^{2^{j-1}} \Big[\theta(X_{t_j,k}^-) - \theta(X_{t_j,2k}^+) - \theta(X_{t_j,2k+1}^+) \Big] \\ &= \sum_{k=1}^{2^{j-1}} \underbrace{\Big[\theta(X_{t_j,k}^-) - 2\theta(X_{t_j,k}^-/2) \Big]}_{\leq K_\theta X_{t_j,k}^-} \\ &\leq K_\theta \sum_{k=1}^{2^{j-1}} X_{t_j,k}^- \end{aligned} \tag{C.19}$$

where we used the fact that $X_{t_j,2k}^+ = X_{t_j,2k+1}^+ = X_{t_j,k}^-/2$ to go from line 2 to 3, and the definition of K_θ to go from line 3 to 4. Now using that $\sum_{k=1}^{2^{j-1}} X_{t_j,k}^- = 1$ and summing for $j \in \{2, \ldots, \sigma_T\}$ delivers

$$A_T^{(1)} \leq (\sigma_T - 1) K_\theta \eta_T^{-1} \tag{C.20}$$

Turning now to A_T^B we begin by writing explicitly the braket terms associated to partitions $\mathcal{P}_j$ and $\mathcal{P}_{j-1}$

$$\Big\langle y_j^-, X_{t_j}^- \Big\rangle = \sum_{k=1}^{2^{j-1}} y_{j,k}^- X_{t_j,k}^- \tag{C.21}$$

$$\Big\langle y_j^+, X_{t_j}^+ \Big\rangle = \sum_{k=1}^{2^j} y_{j,k}^+ X_{t_j,k}^+ \tag{C.22}$$

Therefore, their difference can be rewritten as:

$$\begin{aligned} \Big\langle y_j^+, X_{t_j}^+ \Big\rangle - \Big\langle y_j^-, X_{t_j}^- \Big\rangle &= \sum_{k=1}^{2^{j-1}} \Big[y_{j,2k}^+ X_{t_j,2k}^+ + y_{j,2k+1}^+ X_{t_j,2k+1}^+ - y_{j,k}^- X_{t_j,k}^- \Big] \\ &= \sum_{k=1}^{2^{j-1}} X_{t_j,k}^- \left[\frac{y_{j,2k}^+ + y_{j,2k+1}^+}{2} - y_{j,k}^- \right] \\ &= \frac{1}{2} \sum_{k=1}^{2^{j-1}} X_{t_j,k}^- \left[\Big(y_{j,2k}^+ - y_{j,k}^- \Big) + \Big(y_{j,2k+1}^+ - y_{j,k}^- \Big) \right] \end{aligned} \tag{C.23}$$

where we used the fact that $X_{t_j,2k}^+ = X_{t_j,2k+1}^+ = X_{t_j,k}^-/2$. Finally, the second assumption of Lemma C.2 states that there exists a constant K'_θ such that, if S, S^+ are such that $X = Q(S)$ and $X^+ = Q^+(S^+)$ where X and X^+ are consistent distributions on two successive partitions of $\mathcal{K}$, $\mathcal{P}$ and $\mathcal{P}^+$, then $\|S - S^+\|_\infty \le K'_\theta$. Applying this condition to $S = \eta_{t_j} y_{t_j}^-$ and $S^+ = \eta_{t_j} y_{t_j}^+$ readily gives

$$\|y_{t_j}^- - y_{t_j}^+\|_\infty \le \eta_{t_j}^{-1} K'_\theta,$$

where $y_{t_j}^- \in \mathbb{R}^{\mathcal{P}_{j-1}}$ and $y_{t_j}^+ \in \mathbb{R}^{\mathcal{P}_j}$ are cannonically casted as piecewise constant functions on $\mathcal{K}$. Injecting this inequality in (C.23), using the fact that $\sum_{k=1}^{2^{j-1}} X_{t_j,k}^- = 1$ and summing for $j \in \{2, \dots, \sigma_T\}$ gives

$$\sum_{j=2}^{\sigma_T} \Big[\Big\langle y_{t_j}^+, X_{t_j}^+ \Big\rangle - \Big\langle y_{t_j}^-, X_{t_j}^- \Big\rangle \Big] \le K'_\theta \sum_{j=2}^{\sigma_T} \eta_{t_j}^{-1} \le K'_\theta \eta_T^{-1} \tag{C.24}$$

To finally conclude on a bound for A_T^B, we just need to bound $\langle y_j^+, q_j^x \rangle - \langle y_j^-, q_{j-1}^x \rangle$. Remarking that for all $\mathcal{S} \in \mathcal{P}_j$, $q_{j,k}^x(\mathcal{S}) = \mathbb{1}_{x \in \mathcal{S}}$ and using a similar approach as before, it is straightforward to show that

$$|\langle y_j^+, q_j^x \rangle - \langle y_j^-, q_{j-1}^x \rangle| \le \eta_{t_j}^{-1} K'_\theta,$$

which combined with (C.24) finally gives

$$A_T^{(2)} \le 2K'_\theta (\sigma_T - 1) \eta_T^{-1}, \tag{C.25}$$

Showing (C.12a) comes down to summing up (C.20) (C.25).

Step 2: Bounding B_T. We now finish this showing (C.12b). The bound can be directly obtained from using previously introduced tools. Indeed, we can write

$$\begin{aligned}
\sum_{j=2}^{\sigma_T} \big[\phi(m_{t_{j-1}}) - \phi(m_{t_j})\big] &= \sum_{j=2}^{\sigma_T} \Big[m_{t_{j-1}} \theta(m_{t_{j-1}}^{-1}) - m_{t_j} \phi(m_{t_j}^{-1}) \Big] \\
&= \sum_{j=2}^{\sigma_T} m_{t_{j-1}} \underbrace{\Big[\theta(m_{t_{j-1}}^{-1}) - 2\phi(m_{t_{j-1}}^{-1}/2) \Big]}_{\le K_\theta m_{t_{j-1}}^{-1}} \\
&\le \sum_{j=2}^{\sigma_T} m_{t_{j-1}} K_\theta m_{t_{j-1}}^{-1} \\
&= K_\theta (\sigma_T - 1).
\end{aligned} \tag{C.26}$$

where we used the fact that $m_{t_j} = 2m_{t_{j-1}}$ to go from line 1 to 2, and the first assumption of θ to go from line 2 to 3. This directly delivers (C.12b) after multiplying by η_T^{-1}, and therefore completes the proof. ■

Putting everything together. We are finally in a position to derive our static regret guarantees for (HDA).

Proof of Theorem 2. Let $C_\theta = 2(K_\theta + K'_\theta)$. Then, plugging (C.12a) and (C.12b) into (C.10), we obtain

$$\operatorname{Reg}_x(T) \le \frac{C_\theta \sigma_T + \phi(m_T)}{\eta_{T+1}} + \sum_{j=1}^{\sigma_T} \sum_{t \in \mathcal{T}_j} \langle Z_t, X_t - q_j^x \rangle + \frac{1}{2K} \sum_{t=1}^{T} \eta_t \|\hat{u}_t\|_t^2 + 2L \operatorname{diam}(\mathcal{K}) \sum_{t=1}^{T} \frac{1}{m_t^{1/d}} \tag{C.27}$$

The bound (21) then follows by taking expectations in (C.27), using the bounds (19) for the estimator $\hat{u}_t$, and recalling that $\sigma_T = \log_2(m_T)$.

We now turn to the second part of Theorem 2, namely the expected regret bound (22). The main challenge here is that (22) bounds the algorithm's *expected regret* (and not the incurred *pseudo*-regret), so we cannot simply exchange the maximum and expectation operations. The obstacle to this is the term $\sum_{j=1}^{\sigma_T} \sum_{t \in \mathcal{T}_j} \langle Z_t, X_t - q_j^x \rangle$ in (C.27), which we will bound window-by-window below.

To do so, let

$$\tilde{R}_j(x) = \sum_{t \in \mathcal{T}_j} \langle Z_t, X_t - q_j^x \rangle \qquad j = 1, 2, \dots, \sigma_T, \tag{C.28}$$

and consider the auxiliary processes

$$\tilde{S}_{t+1} = \tilde{S}_t - Z_t, \quad \tilde{X}_{t+1} = Q^{\mathcal{P}_j}(\eta_{t+1}\tilde{S}_{t+1}), \qquad t = t_j, \dots, t_{j+1}-1, \tag{C.29}$$

with $\tilde{X}_{t_j} = X_{t_j}$. We then have

$$\tilde{R}_j(x) = \sum_{t\in\mathcal{T}_j} \left\langle Z_t, (X_t - \tilde{X}_t) + (\tilde{X}_t - q_j^x) \right\rangle = \sum_{t\in\mathcal{T}_j} \left\langle Z_t, X_t - \tilde{X}_t \right\rangle \tag{C.30a}$$

$$+ \sum_{t\in\mathcal{T}_j} \left\langle Z_t, \tilde{X}_t - q_j^x \right\rangle \tag{C.30b}$$

We now proceed to bound each of the above terms in expectation:

a) Since the term (C.30a) does not depend on x, we readily obtain

$$\mathbb{E}\left[\max_{x\in\mathcal{K}} \sum_{t\in\mathcal{T}_j} \left\langle Z_t, X_t - \tilde{X}_t \right\rangle\right] = \sum_{t\in\mathcal{T}_j} \mathbb{E}\Big[\mathbb{E}\Big[\left\langle Z_t, X_t - \tilde{X}_t \right\rangle \,\Big|\, \mathcal{F}_t\Big]\Big] = \sum_{t\in\mathcal{T}_j} \mathbb{E}\Big[\left\langle b_t, X_t - \tilde{X}_t \right\rangle\Big] \leq 2 \sum_{t\in\mathcal{T}_j} \mu_t \tag{C.31}$$

where the last line follows from (19a) and the fact that X_t and $\tilde{X}_t$ are both simple strategies supported on $\mathcal{P}_j$ (so $\langle b_t, X_t - \tilde{X}_t\rangle \leq |\langle b_t, X_t\rangle| + |\left\langle b_t, \tilde{X}_t\right\rangle| \leq 2\mu_t$).

b) For the term (C.30b), applying Proposition 1 to the sequence of "virtual" payoff functions $-Z_t$, $t = 1, 2, \dots$, we get

$$\sum_{t\in\mathcal{T}_j} \left\langle Z_t, \tilde{X}_t - q_j^x \right\rangle \leq E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j} + \left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right)\phi(m_{t_j}) + \frac{1}{2K} \sum_{t\in\mathcal{T}_j} \eta_t \|Z_t\|_t^2. \tag{C.32}$$

Since the right-hand side of this last equation does not depend on x, maximizing and taking expectations yields

$$\mathbb{E}\left[\max_{x\in\mathcal{K}} \sum_{t\in\mathcal{T}_j} \left\langle Z_t, \tilde{X}_t - q_j^x \right\rangle\right] \leq E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j} + \left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right)\phi(m_{t_j}) + \frac{1}{2K} \sum_{t\in\mathcal{T}_j} \eta_t \zeta_t^2, \tag{C.33}$$

where we set $\zeta_t^2 = \mathbb{E}\big[\|Z_t\|_t^2\big]$.

Therefore, taking expectations in (C.30) and plugging Eqs. (C.31) and (C.33) into the resulting expression, we obtain

$$\mathbb{E}\left[\max_{x\in\mathcal{K}} \tilde{R}_j(x)\right] \leq E_{t_j}^{\mathcal{P}_j} - E_{t_{j+1}}^{\mathcal{P}_j} + \left(\eta_{t_{j+1}}^{-1} - \eta_{t_j}^{-1}\right)\phi(m_{t_j}) + 2\sum_{t\in\mathcal{T}_j} \mu_t + \frac{1}{2K} \sum_{t\in\mathcal{T}_j} \eta_t \zeta_t^2 \tag{C.34}$$

and hence, using Lemma C.2 and working as in the case of (C.27), we get:

$$\mathbb{E}\left[\max_{x\in\mathcal{K}} \sum_{j=1}^{\sigma_T} \sum_{t\in\mathcal{T}_j} \langle Z_t, X_t - q_j^x\rangle\right] \leq \sum_{j=1}^{\sigma_T} \mathbb{E}\left[\max_{x\in\mathcal{K}} \tilde{R}_j(x)\right] \leq \frac{C_\theta(\sigma_T - 1) + \phi(m_T)}{\eta_{T+1}} + 2\sum_{t=1}^{T} \mu_t + \frac{1}{2K} \sum_{t=1}^{T} \eta_t \zeta_t^2. \tag{C.35}$$

Thus, going back to (C.27) and taking expectations, we get the expected regret bound

$$\mathbb{E}[\operatorname{Reg}(T)] \leq 2\frac{C_\theta(\sigma_T - 1) + \phi(m_T)}{\eta_{T+1}} + 2\sum_{t=1}^{T} \mu_t + \frac{1}{2K} \sum_{t=1}^{T} \eta_t(\zeta_t^2 + M_t^2) + 2L \operatorname{diam}(\mathcal{K}) \sum_{t=1}^{T} \frac{1}{m_t^{1/d}} \tag{C.36}$$

As a last step, since $Z_t = \hat{u}_t - u_t$, we readily get $\mathbb{E}\big[\|Z_t\|_t^2\big] \leq 2\,\mathbb{E}\big[\|u_t\|_t^2 + \|\hat{u}_t\|_t^2\big] \leq 2(R^2 + M_t^2)$ by Lemma A.1, Assumption 1, and the definition (19b) of M_t. The bound (22) then follows by a straightforward substitution. ■

To proceed with the proof of the specific regret bound for the HEW instantiation of (HDA), we will require a series of intermediate results to bound the bias and second moment of the estimator (IWE). These are as follows.

Lemma C.3. *Running HEW with any splitting schedule implying m_t components of the underlying partiton of $\mathcal{K}$ at time t, the bias and mean square of the* (IWE) *satisfy for all t:*

$$\begin{aligned} \mu_t &\leq 2L \operatorname{diam}(\mathcal{K}) m_t^{-1/d} \\ M_t^2 &\leq R^2(m_t+1) \end{aligned} \tag{C.37}$$

Proof. To streamline the proof, we first need to introduce some notation. Specifically, we will write $\mathcal{P}_t$ for the underlying partition at time t, and for any $x \in \mathcal{K}$, $\mathcal{S}_t^x$ denotes the component of $\mathcal{P}_t$ such that $x \in \mathcal{S}_t^x$. Let $x_t \in \mathcal{K}$ be the action played at time t ; to simplify the notations we use the convention introduced in the main text and denote $\mathcal{S}_t := \mathcal{S}_t^{x_t}$.

Moreover, we recall that $X_t \in \mathcal{Q}_{\mathcal{P}_t}$ designates the current mixed strategy at t. Specifically for any $\mathcal{S} \in \mathcal{P}_t$, $X_{\mathcal{S},t}$ denote the probability to pick an action $x_t \in \mathcal{S}$ at time t. In a slight abuse of notation, we overload X_t and also consider it refers to the corresponding *density function* defined on $\mathcal{K}$, i.e., for all $x \in \mathcal{K}$, we have

$$X_t(x) = \lambda(\mathcal{S}_t^x)^{-1} X_{\mathcal{S}_t^x,t}. \tag{C.38}$$

Finally, we recall the definition of the *importance weighted estimator* (IWE):

$$\hat{u}_t(x) = R - \frac{R - u_t(x_t)}{X_{\mathcal{S}_t,t}} \mathbb{1}(x \in \mathcal{S}_t), \tag{IWE}$$

Bounding μ_t in the setting of HEW. Recall first that

$$\textit{Bias:} \quad |\langle b_t, q\rangle| \leq \mu_t \tag{C.39}$$

for all $t = 1, 2, \dots$, and all $q \in \mathcal{Q}$.

Let $x \in \mathcal{K}$.
By definition $b_t(x) = u_t(x) - \mathbb{E}[\hat{u}_t(x)|\mathcal{F}_t]$. Using (IWE), a series of mechanical computations bring

$$\begin{aligned} \mathbb{E}[\hat{u}_t(x) \,|\, \mathcal{F}_t] &= \mathbb{E}\left[R - \frac{R - u_t(x_t)}{X_{\mathcal{S}_t,t}} \mathbb{1}(x \in \mathcal{S}_t) \,\middle|\, \mathcal{F}_t\right] \\ &= R - \int_{\mathcal{K}} \left(\frac{R - u_t(x')}{X_{\mathcal{S}_t^{x'},t}} \mathbb{1}(x \in \mathcal{S}_t^{x'}) \right) X_t(x') dx' \\ &= R - \int_{\mathcal{S}_t^x} \left(\frac{R - u_t(x')}{X_{\mathcal{S}_t^x,t}} \right) \underbrace{X_t(x)}_{\lambda(\mathcal{S}_t^x)^{-1} X_{\mathcal{S}_t^x,t}} dx' \\ &= R - R + \lambda(\mathcal{S}_t^x)^{-1} \int_{\mathcal{S}_t^x} u_t(x') dx' \end{aligned} \tag{C.40}$$

For any $\mathcal{S} \subset \mathcal{K}$ and for any measurable function $f : \mathcal{K} \to \mathbb{R}$ we denote $\bar{f}(\mathcal{S}) = \lambda(\mathcal{S})^{-1} \int_S f(x) dx$. We can therefore write

$$\mathbb{E}[\hat{u}_t(x) \,|\, \mathcal{F}_t] = \bar{u}_t(\mathcal{S}_t^x). \tag{C.41}$$

Therefore,

$$b_t(x) = u_t(x) - \bar{u}_t(\mathcal{S}_t^x), \tag{C.42}$$

and the fact that the stream of payoff functions u_t is uniformly Lipschitz directly delivers $b_t(x) \leq L \operatorname{diam}(\mathcal{S}_t^x)$. Using Lemma C.1 finally brings, for all $x \in \mathcal{K}$:

$$b_t(x) \leq 2L \operatorname{diam}(\mathcal{K}) m_t^{-1/d} \tag{C.43}$$

which, using $|\langle b_t, q\rangle| \leq \|b_t\|_\infty \|q\|_1 = \|b_t\|_\infty$ shows that

$$\mu_t \leq 2L \operatorname{diam}(\mathcal{K}) m_t^{-1/d} \tag{C.44}$$

Bounding M_t^2 in the setting of HEW. We recall the definition of M from (19b):

$$\textit{Mean square:} \quad \mathbb{E}\big[\|\hat{u}_t\|_t^2 \,\big|\, \mathcal{F}_t\big] \leq M_t^2 \tag{C.45}$$

To simplify the incoming computations, we denote $l_t : \mathcal{K} \to \mathbb{R}+$ the *loss* function such that for all $x \in \mathcal{K}$, $l_t(x) = R - u_t(x)$. Since $0 \leq u_t \leq R$, we also have $0 \leq l_t \leq R$. For any $\mathcal{S} \subset \mathcal{K}$, we also introduce $\delta_{\mathcal{S}} : \mathcal{K} \to \{0, 1\}$ the function such that for all $x \in \mathcal{K}$, $\delta_{\mathcal{S}}(x) = \mathbb{1}(x \in \mathcal{S})$.

With this in hand, we can proceed to the following rewriting of $\|\hat{u}_t\|_t^2$, which we recall is a random quantity given filtration $\mathcal{F}_t$ since it depends on the choice of the t^{th} action, x_t:

$$\begin{aligned}\|\hat{u}_t\|_t^2 = \langle \hat{u}_t^2, X_t \rangle &= \left\langle \left(R - \frac{l_t(x_t)}{X_{\mathcal{S}_t,t}} \delta_{\mathcal{S}_t}\right)^2, X_t \right\rangle \\ &= R^2 - 2R\frac{l_t(x_t)}{X_{\mathcal{S}_t,t}} \langle \delta_{\mathcal{S}_t}, X_t \rangle + \frac{l_t(x_t)^2}{X_{\mathcal{S}_t,t}^2} \langle \delta_{\mathcal{S}_t}^2, X_t \rangle \end{aligned} \tag{C.46}$$

For any $\mathcal{S} \subset \mathcal{K}$, $\delta_{\mathcal{S}}^2 = \delta_{\mathcal{S}}$, and simple computations give $\langle \delta_{\mathcal{S}_t}, X_t \rangle = X_{\mathcal{S}_t,t}$. This then delivers the following expression for $\|\hat{u}_t\|_t^2$:

$$\|\hat{u}_t\|_t^2 = R^2 - 2Rl_t(x_t) + \frac{l_t(x_t)^2}{X_{\mathcal{S}_t,t}}. \tag{C.47}$$

The aim of the proof is to bound the expectancy of (C.47) given filtration $\mathcal{F}_t$. We are primarily interested in the quantity $\mathbb{E}\Big[\frac{l_t(x_t)^2}{X_{\mathcal{S}_t,t}} \,\Big|\, \mathcal{F}_t\Big]$ which is the most complex to handle. We write

$$\begin{aligned}\mathbb{E}\left[\frac{l_t(x_t)^2}{X_{\mathcal{S}_t,t}} \,\middle|\, \mathcal{F}_t\right] &= \int_{\mathcal{K}} \frac{l_t(x')^2}{X_{\mathcal{S}_t^{x'},t}} X_t(x') dx' \\ &= \sum_{\mathcal{S} \in \mathcal{P}_t} \int_{\mathcal{S}} \frac{l_t(x')^2}{X_{\mathcal{S},t}} \underbrace{X_t(x')}_{\lambda(\mathcal{S})^{-1} X_{\mathcal{S},t}} dx' \\ &= \sum_{\mathcal{S} \in \mathcal{P}_t} \lambda(\mathcal{S})^{-1} \int_{\mathcal{S}} l_t(x')^2 dx' \\ &= \sum_{\mathcal{S} \in \mathcal{P}_t} \bar{l}_t^2(\mathcal{S}). \end{aligned} \tag{C.48}$$

Now using that $|l_t| \leq R$ we have that $\bar{l}_t^2(\mathcal{S}) \leq R^2$ for any $\mathcal{S} \in \mathcal{P}_t$ and therefore

$$\mathbb{E}\left[\frac{l_t(x_t)^2}{X_{\mathcal{S}_t,t}} \,\middle|\, \mathcal{F}_t\right] \leq |\mathcal{P}_t| R^2 = m_t R^2 \tag{C.49}$$

Then, remarking that $\mathbb{E}[l_t(x_t) \,|\, \mathcal{F}_t] \geq 0$ and combining (C.47) and (C.49) we finally get

$$\mathbb{E}\big[\|\hat{u}_t\|_t^2 \,\big|\, \mathcal{F}_t\big] \leq R^2(m_t + 1)$$

which delivers

$$M_t^2 \leq R^2(m_t + 1), \tag{C.50}$$

and concludes the proof. ■

To conclude, we now need to relate m_t, the number of sets in partition $\mathcal{P}_t$ at time t, and the chosen splitting schedule. In case of a logarithmic splitting schedule $v_t = p \log_2 t$, we present the following result giving upper and lower bound on m_t with respect to t and p.

Lemma C.4. *In case of a logarithmic splitting schedule $v_t = p \log_2 t$, we have for every t:*

$$\frac{1}{2} t^p \leq m_t \leq t^p \tag{C.51}$$

Proof. Let $v_t = p \log_2 t$. By definition of the scheduler function v_t, this implies that at any time t, $\lfloor v_t \rfloor$ splitting events have occurred. Therefore, we have $\sigma_t = \lfloor v_t \rfloor = \lfloor p \log_2 t \rfloor$. Since $m_t = 2^{\sigma_t}$ by definition, we get

$$m_t = 2^{\lfloor p \log_2 t \rfloor}. \tag{C.52}$$

The result then follows directly from remarking that

$$p \log_2 t - 1 < \lfloor p \log_2 t \rfloor \leq p \log_2 t \tag{C.53}$$

and using the fact that $x \mapsto 2^x$ is an increasing function. ■

We are now in a position to prove our main regret guarantee for the HEW algorithm. For convenience, we restate the relevant result below.

Corollary 1. *If HEW is run with learning rate $\eta_t \propto t^{-\varrho}$ and the logarithmic splitting schedule $v_t = p \log_2(t)$, the learner enjoys the bound*

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathcal{O}(T^{\varrho} + T^{1-p/d} + T^{1+p-\varrho}). \tag{23}$$

In particular, if the algorithm is run with $\varrho = (d+1)/(d+2)$ and $p = d/(d+2)$ we obtain the bound

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathcal{O}(T^{\frac{d+1}{d+2}}). \tag{24}$$

Proof of Corollary 1. The idea of this proof consists in bounding the different terms on the right hand side of (2) in the case of HEW with learning rate $\eta_t \propto t^{\varrho}$ and a logarithmic splitting schedule $v_t = p \log_2 t$. With this in mind, combining Lemmas C.3 and C.4 we get

$$M_t^2 \leq R^2(m_t + 1) \qquad \leq R^2(t^p + 1) \qquad\qquad = \mathcal{O}(t^p) \tag{C.54a}$$

$$\mu_t \leq 2L\,\mathrm{diam}(\mathcal{K}) m_t^{-1/d} \leq 2L\,\mathrm{diam}(\mathcal{K}) \left(\frac{1}{2} t^p\right)^{-1/d} = \mathcal{O}(t^{-p/d}). \tag{C.54b}$$

The result stated in Corollary 1 directly follows from injecting this into (22). ■

C.2. Dynamic regret guarantees

We now turn to the dynamic regret guarantees of (HDA) as stated in Theorem 3 below.

Theorem 3. *Suppose that* (HDA) *is run with the negentropy kernel, and assumptions as in Theorem 2. Then:*

$$\begin{aligned} &\mathbb{E}[\mathrm{DynReg}(T)] \\ &\qquad = \mathcal{O}(T^{1+2\mu-\varrho} + T^{1-\beta} + T^{1-p/d} + T^{2\varrho-2\mu} V_T). \end{aligned} \tag{26}$$

Proof of Theorem 3. Our proof will be again based on a window-by-window analysis. However, instead of focusing on the windows of $\{1, \dots, T\}$ over which the cover of (HDA) remains constant and fixed, we will decompose the horizon of the process into N *virtual* batches, and we will compare the learner's static and dynamic regret over each such batch.[4] We will then harvest a bound for the aggregate dynamic regret over T stages following a comparison technique first introduced by Besbes et al. (2015).

To proceed, write the interval $\mathcal{T} = \{1, \dots, T\}$ as the union of N contiguous sub-intervals $\mathcal{T}_i$, $i = 1, \dots, N$, each of length Δ (with the possible exception of the final batch, which might be shorter). Formally, let $\Delta = \lceil T^{\gamma} \rceil$ for some constant $\gamma \in [0, 1]$ to be determined later; then the number of virtual batches is $N = \lceil T/\Delta \rceil = \Theta(T^{1-\gamma})$ and we have

$$\mathcal{T}_i = \{(i-1)\Delta + 1, \dots, i\Delta\} \qquad \text{for all } i = 1, \dots, N-1, \tag{C.55}$$

with $\mathcal{T}_N = \mathcal{T} \setminus \bigcup_{i=1}^{N-1} \Delta_i$ being excluded from the above enumeration as (possibly) smaller than the rest.

[4] A further important difference is that these virtual batches will all have the same length, in contrast to the windows of time between two consecutive splitting events.

Now, focusing on the i-th batch $\mathcal{T}_i$ of $\mathcal{T}$ and taking $x_t^* \in \arg\max_{x \in \mathcal{K}} u_t(x)$ and $x_i^* \in \arg\max_{x \in \mathcal{K}} \sum_{t \in \mathcal{T}_i} u_t(x)$, we get

$$u_t(x_t^*) - \langle u_t, X_t \rangle = u_t(x_i^*) - \langle u_t, X_t \rangle + u_t(x_t^*) - u_t(x_i^*) \tag{C.56}$$

We may then bound the dynamic regret incurred by (HDA) over the interval $\mathcal{T}_i$ as

$$\operatorname{DynReg}(\mathcal{T}_i) = \sum_{t \in \mathcal{T}_i} [u_t(x_t^*) - \langle u_t, X_t \rangle] + \sum_{t \in \mathcal{T}_i} [u_t(x_i^*) - u_t(x_i^*)] = \operatorname{Reg}(\mathcal{T}_i) + \sum_{t \in \mathcal{T}_i} [u_t(x_t^*) - u_t(x_i^*)]. \tag{C.57}$$

Moving forward, we will bound the difference $\sum_{t \in \mathcal{T}_i} [u_t(x_t^*) - u_t(x_i^*)]$ following a comparison technique originally due to Besbes et al. (2015, Prop. 2). To do so, let $\tau_i = \min \mathcal{T}_i$ denote the starting epoch of the i-th virtual batch, and let $x_{\tau_i}^*$ denote a maximizer of the first payoff function encountered in the batch $\mathcal{T}_i$. We then obtain by construction

$$\sum_{t \in \mathcal{T}_i} [u_t(x_t^*) - u_t(x_i^*)] \leq \sum_{t \in \mathcal{T}_i} \big[u_t(x_t^*) - u_t(x_{\tau_i}^*)\big] \leq \Delta \max_{t \in \mathcal{T}_i} \big[u_t(x_t^*) - u_t(x_{\tau_i}^*)\big] \leq 2\Delta V_i, \tag{C.58}$$

where we used the fact that $|\mathcal{T}_i| \leq \Delta$ for all $i = 1, \dots, N$ (this time *including* the last batch). Hence, by combining (C.58) and (C.57), we get

$$\operatorname{DynReg}(\mathcal{T}_i) \leq \operatorname{Reg}(\mathcal{T}_i) + 2\Delta V_i \tag{C.59}$$

Thus, finally, after summing over all batches and taking expectations, we obtain the static-to-dynamic comparison bound

$$\mathbb{E}[\operatorname{DynReg}(T)] \leq \sum_{i=1}^{N} \mathbb{E}[\operatorname{Reg}(\mathcal{T}_i)] + 2\Delta V_T. \tag{C.60}$$

We will proceed to bound $\mathbb{E}[\operatorname{DynReg}(T)]$ by bounding the “batch regret” $\sum_{i=1}^{N} \mathbb{E}[\operatorname{Reg}(\mathcal{T}_i)]$ and retroactively tuning the batch-size Δ.

To carry out this approach, Theorem 2 with $\theta(x) = x \log x$, $\eta_t \propto t^{-\varrho}$ and $\sigma_t = \lfloor p \log_2 t \rfloor$ readily yields

$$\mathbb{E}[\operatorname{Reg}(\mathcal{T}_i)] = \mathcal{O}\left((i\Delta)^{\varrho} + \sum_{t \in \mathcal{T}_i} t^{-p/d} + \sum_{t \in \mathcal{T}_i} t^{-\beta} + \sum_{t \in \mathcal{T}_i} t^{2\mu - \varrho} \right) \tag{C.61}$$

and hence, after summing over all batches:

$$\begin{aligned} \sum_{i=1}^{N} \mathbb{E}[\operatorname{Reg}(\mathcal{T}_i)] &= \mathcal{O}\left(\Delta^{\varrho} \sum_{i=1}^{N} i^{\varrho} + \sum_{t=1}^{T} t^{-p/d} + \sum_{t=1}^{T} t^{-\beta} + \sum_{t=1}^{T} t^{2\mu - \varrho} \right) \\ &= \mathcal{O}\Big(\Delta^{\varrho} N^{1+\varrho} + T^{1-p/d} + T^{1-\beta} + T^{1+2\mu-\varrho} \Big). \end{aligned} \tag{C.62}$$

Now, since $\Delta = \mathcal{O}(T^{\gamma})$ and $N = \mathcal{O}(T/\Delta) = \mathcal{O}(T^{1-\gamma})$, the first summand above can be bounded as

$$\Delta^{\varrho} N^{1+\varrho} = \mathcal{O}((N\Delta)^{\varrho}\, N) = \mathcal{O}(T^{\gamma\varrho} T^{(1-\gamma)(1+\varrho)}) = \mathcal{O}(T^{1+\varrho-\gamma}). \tag{C.63}$$

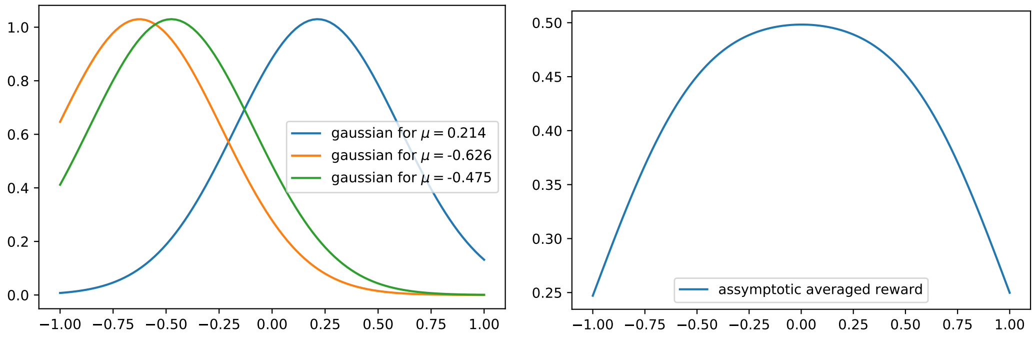
Thus, going back to (C.62) and (C.60), we get the dynamic regret bound

$$\mathbb{E}[\operatorname{DynReg}(T)] = \mathcal{O}\Big(T^{1+\varrho-\gamma} + T^{1-p/d} + T^{1-\beta} + T^{1+2\mu-\varrho} + T^{\gamma} V_T \Big). \tag{C.64}$$

To calibrate the above expression, the “virtual batch-size” exponent γ must be chosen such that $1 + \varrho - \gamma = 1 + 2\mu - \varrho$, i.e., $\gamma = 2\varrho - 2\mu$. This choice then yields the bound

$$\mathbb{E}[\operatorname{DynReg}(T)] = \mathcal{O}\Big(T^{1+2\mu-\varrho} + T^{1-p/d} + T^{1-\beta} + T^{2\varrho-2\mu} V_T \Big). \qquad \blacksquare$$

Finally, Corollary 2 simply follows by combining the dynamic regret guarantee of Theorem 3 with the bounds of Lemma C.3 for the IWE estimator.



(a) 3 realizations of such reward functions

(b) Assymptotic averaged reward

Figure 4: Gaussian adversarial reward

D. Numerical experiments

In this appendix, we present some supplementary numerical experiments for the HEW algorithm – dubbed `Hierarchical` in the sequel. Specifically, we ran different adversary models (reward design mechanisms) and compared the performance of `Hierarchical` with two baselines: *a*) a fixed-mesh policy – `Grid` – that employs an underlying the EXP3 algorithm structure (Auer et al., 2002) with rewards sampled at the grid points, as per the DAX template; and *b*) the `Kernel` policy proposed by Héliou et al. (2020), which plugs a kernel density estimate around the sampled action points, and combines it with an explicit exploration term, as per (IWE[3]).

About the adversary functions we choose 1 and 2 dimensional action sets ($\dim(\mathcal{K}) \in \{1, 2\}$) for:

- `Sine`, $\mathcal{K} = [0, 1]^d$: a linear combination of trigonometric terms with different frequencies and amplitudes, arbitrarily drawn, allowing us to know the best action to choose in hindsight (or instantaneously), in order to compute the instantaneous regret. However, we stress that this setting is more a stationary bandit than a proper adversarial one. For this first adversary, *the dynamic regret and the static regret coincide*. That is why we only display the static regret behavior hereafter.

- `Gauss` (gaussian reward with stochastic mean), $\mathcal{K} = [-1, 1]^d$: a stochastic bandit, with multinomial type reward (with fixed covariance), with mean μ randomly drawn (iid) round after round, following a uniform distribution on the action space $[-1, 1]^d$. We can compute the asymptotic averaged reward over a high number of rounds (used to know the best fixed action). We draw in Figure 4a some realization of the gaussian reward in 1 dimension and we display on Figure 4b its asymptotic mean, averaged over 10000 runs. This plot has been produced using simple Monte Carlo technique.

All numerical experiments were run on a machine with 48 CPUs (Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz), with 2 Threads per core, and 500Go of RAM. The horizon was set to $T = 10^5$, and we used the anytime version of every algorithm. We run the algorithm with 46 initial seeds, and then averaged the regret per round, divided by the current round (to exhibit the sub-linear behavior), over the 46 seeds. We add the 'moustache' box showing the confidence interval of the quantity $\mathrm{Reg}(t)/t$ for 2 specific round, namely at $t = 10^4$ and $t = 10^5$, computed empirically on the 46 seeds. We present different sets of hyperparameters for each algorithms, specifically:

- `Kernel`:
  - (γ_0, γ_r) if the learning rate is equal to $\gamma_t = \gamma_0 / t^{\gamma_r}$,
  - number of arms used to store an approximate of the functions defined on $\mathcal{K}$,
  - (w_0, w_r) if the windows of the squared kernel varies as $w_t = w_0 / t^{w_r}$,
  - $(\mathrm{ee}_0, \mathrm{ee}_r)$ if the explicit exploration equals $\mathrm{ee}_t = \mathrm{ee}_0 / t^{\mathrm{ee}_r}$.

- `Grid`
  - γ_0 if the learning rate is equal to $\gamma_t = \gamma_0 / t^{\frac{d+1}{d+2}}$,
  - number of arms used to discretize $\mathcal{K}$ in hindsight,

- Hierarchical
  - (γ_0, γ_r) if the learning rate is equal to $\gamma_t = \gamma_0/t^{\gamma_r}$

We would like to stress that the number of hyperparameters are not the same, and that the HEW algorithm enjoys a lower number of tunable hyperparameters.

On Fig. 5 we plot the mean regret for the Sine1D adversary, with different hyperparameters, over $T = 10^5$ iterations. We display the empirically distribution of such regret divided by the current round t on Fig. 6, to exhibit the sub-linear behavior. We process the same way on Fig. 7, and Fig. 8 for the Sine2D adversary, Fig. 9, Fig. 10 and Fig. 11 for the Gauss1D adversary and finally Fig. 12, Fig. 13 and Fig. 14 for the Gauss2D adversary.

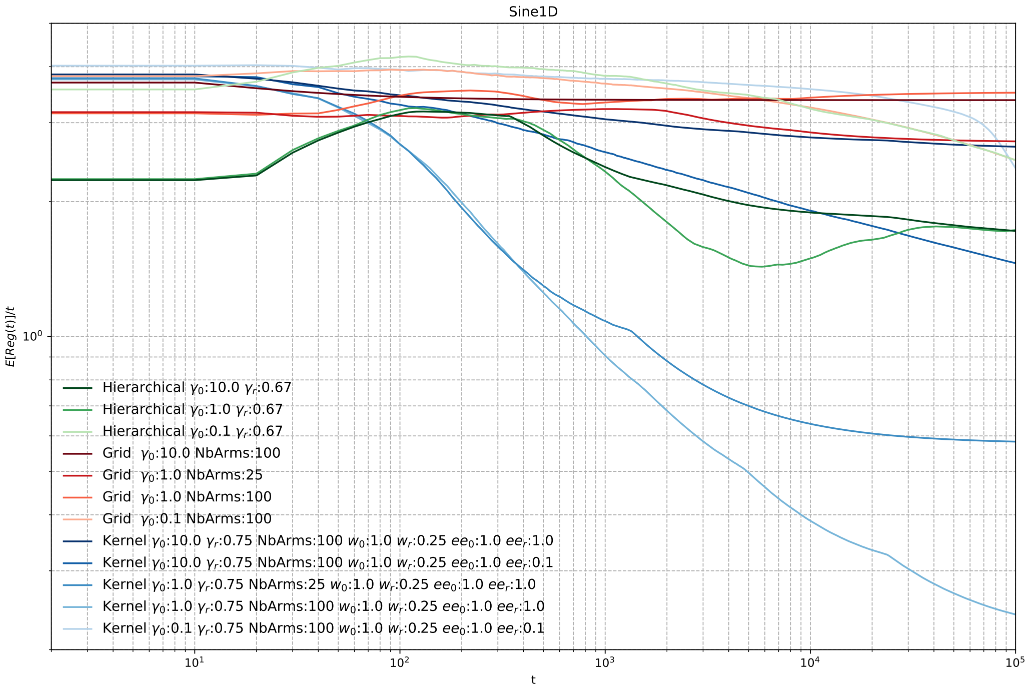


Figure 5: *Static* regret divided by t ($\mathrm{Reg}(t)/t$) in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Sine1D` adversary.

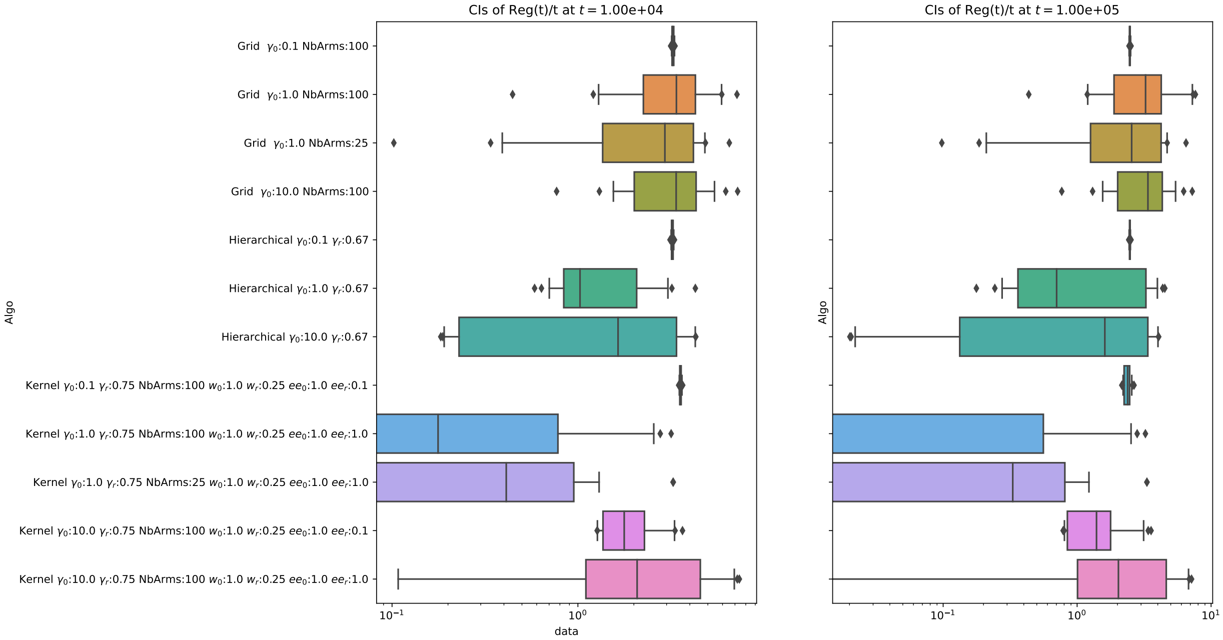


Figure 6: Distribution (moustache box are at CI of $[.05, .95]$) of the averaged static regret over t, averaged on 46 realizations for each algorithm at 2 different timestamps, against the `Sine1D` adversary.

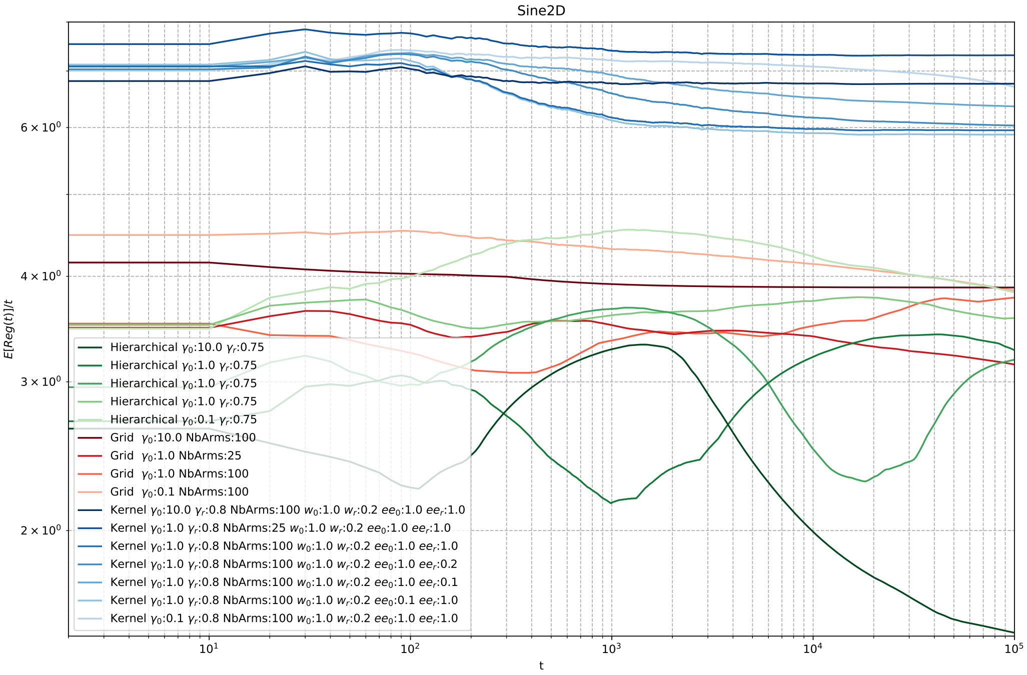


Figure 7: *Static* regret divided by t $(\mathrm{Reg}(t)/t)$ in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Sine2D` adversary.

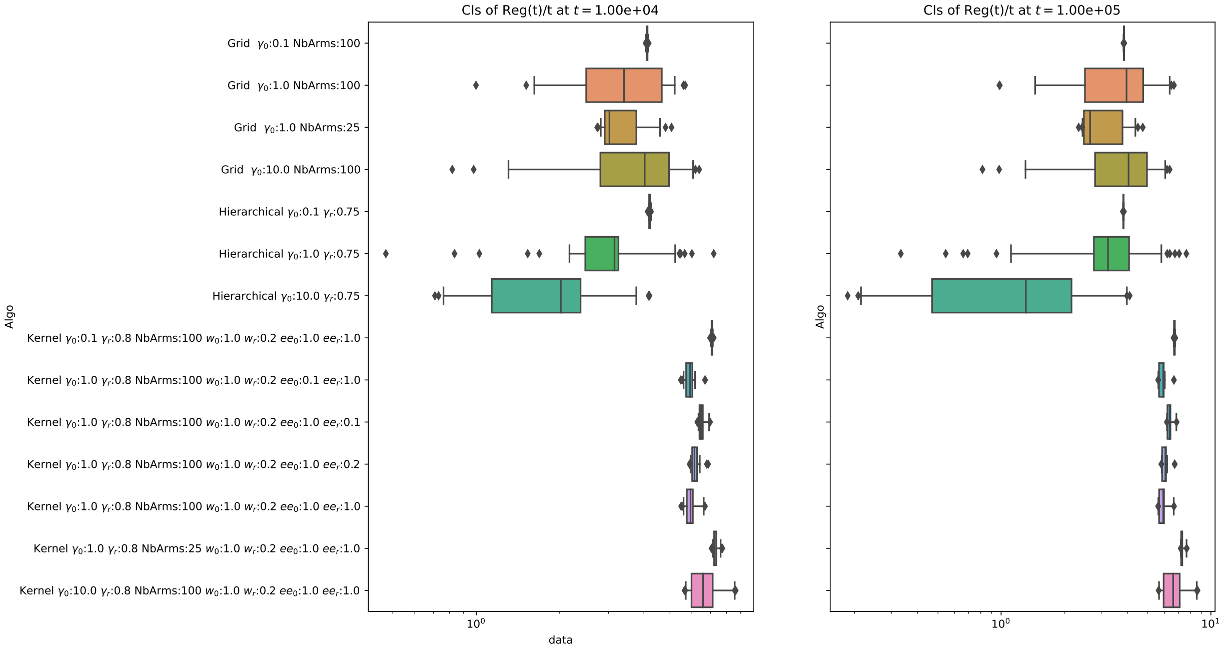


Figure 8: Distribution (moustache box are at CI of $[.05, .95]$) of the averaged static regret over t, averaged on 46 realizations for each algorithm at 2 different timestamps, against the `Sine2D` adversary.

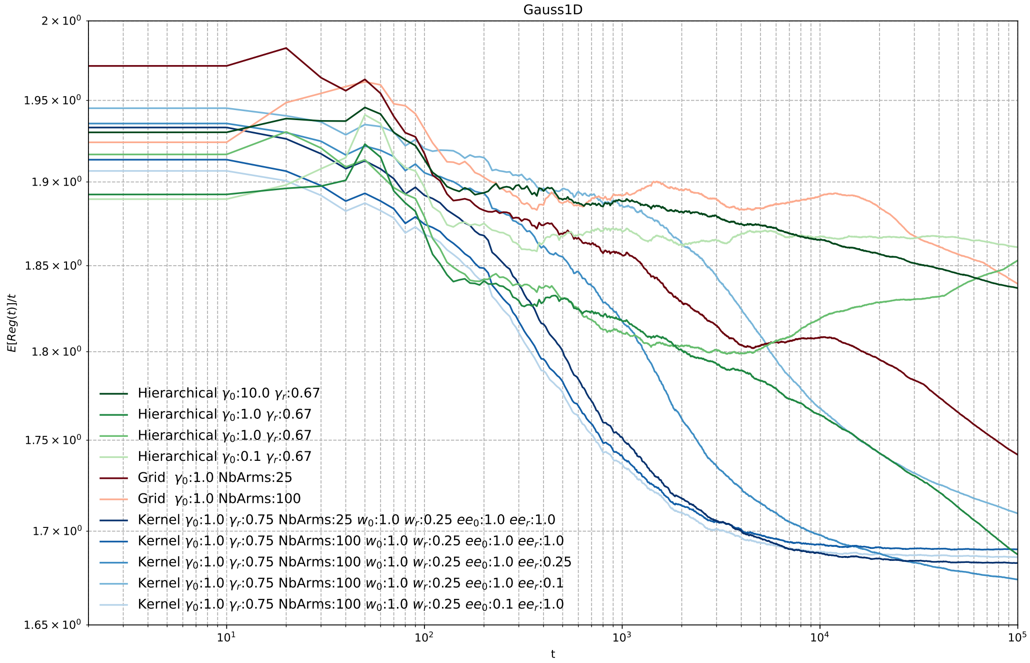


Figure 9: *Static* regret divided by t $(\mathrm{Reg}(t)/t)$ in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Gauss1D` adversary.

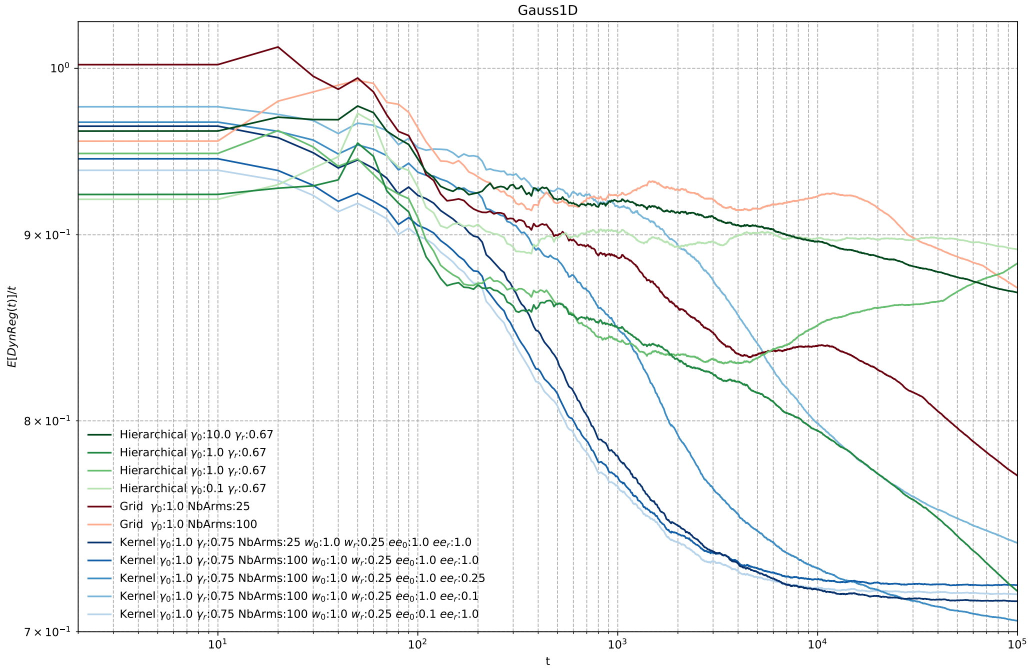


Figure 10: *Dynamic* regret divided by t $(\mathrm{DynReg}(t)/t)$ in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Gauss1D` adversary.

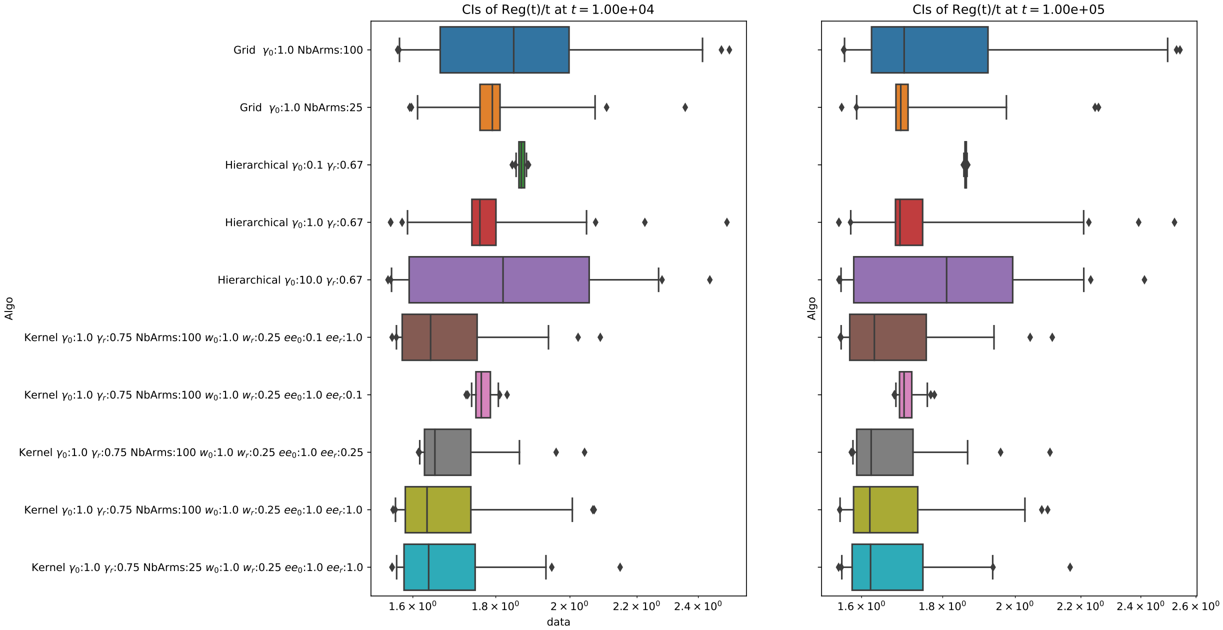


Figure 11: Distribution (moustache box are at CI of $[.05, .95]$) of the averaged static regret over t, averaged on 46 realizations for each algorithm at 2 different timestamps, against the `Gauss1D` adversary.

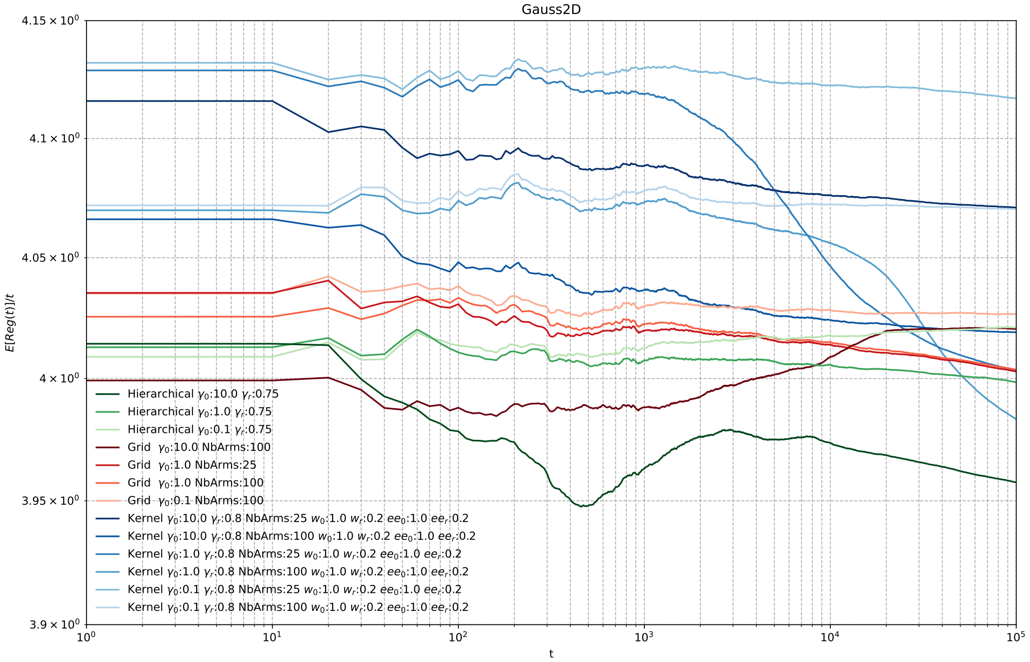


Figure 12: *Static* regret divided by t ($\mathrm{Reg}(t)/t$) in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Gauss2D` adversary.

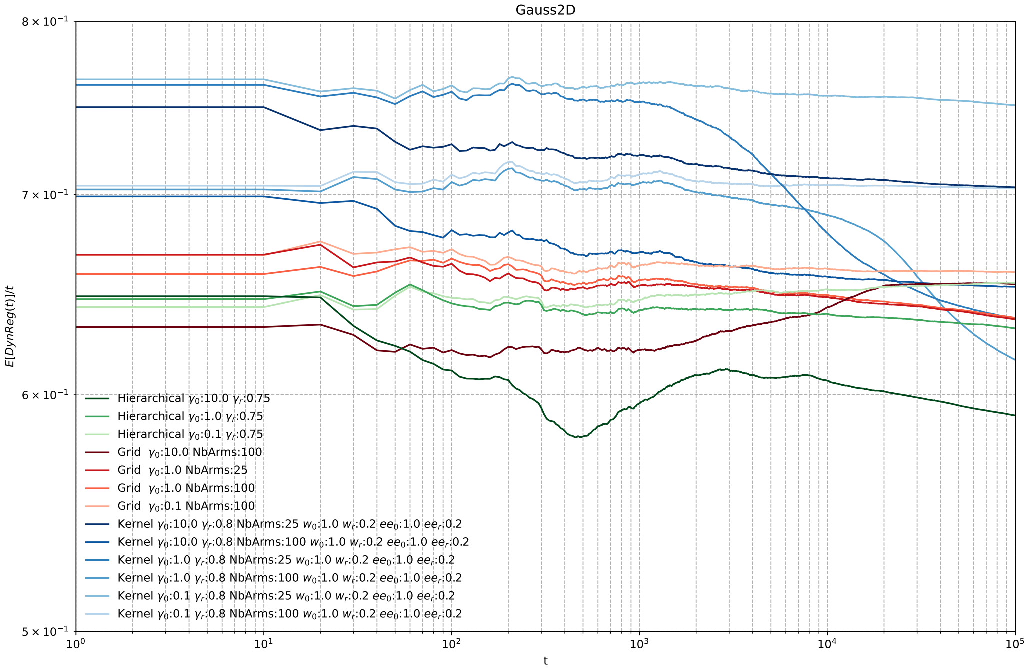


Figure 13: *Dynamic* regret divided by t $(\mathrm{DynReg}(t)/t)$ in log-log scale, averaged on 46 realizations for each algorithm (solid line), against the `Gauss2D` adversary.

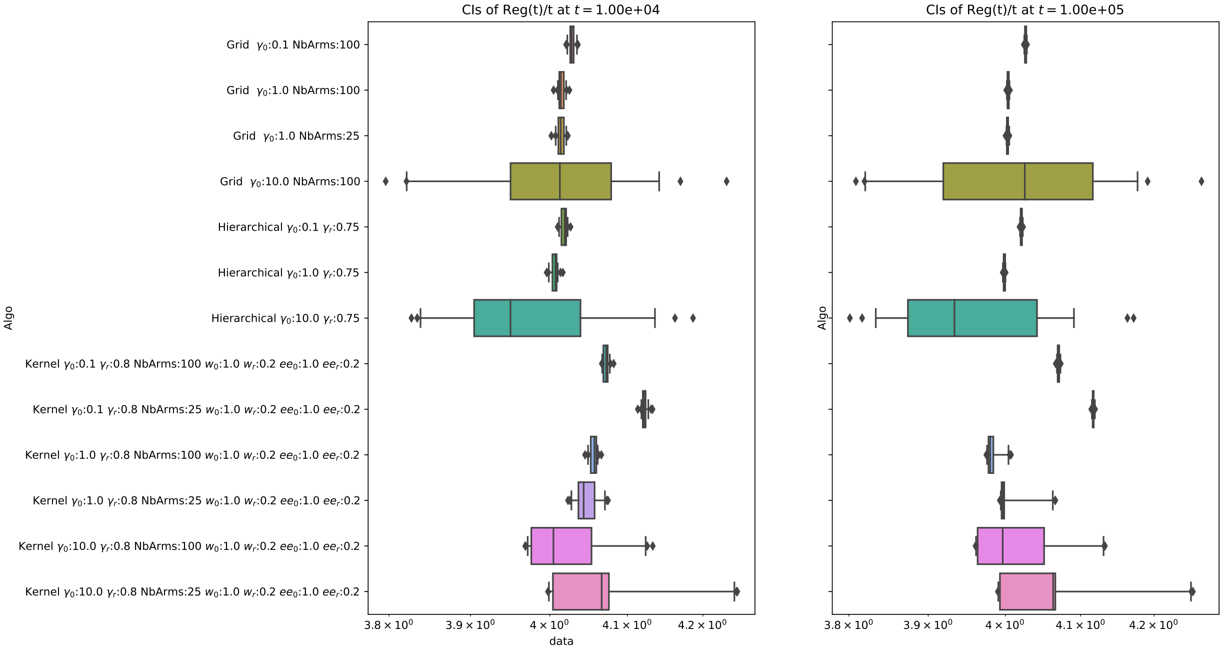


Figure 14: Distribution (moustache box are at CI of $[.05, .95]$) of the averaged static regret over t, averaged on 46 realizations for each algorithm at 2 different timestamps, against the `Gauss2D` adversary.