Proceedings of Machine Learning Research vol :1–25, 2026 39th Annual Conference on Learning Theory

Online Convex Optimization with Sublinear Noisy Probes

Simone Di Gregorio SIMONE.DIGREGORIO@UNIROMA1.IT
Department of Computer, Control and Management Engineering Antonio Ruberti
Sapienza University of Rome, Rome, Italy.

Anupam Gupta ANUPAM.G@NYU.EDU
Department of Computer Science
New York University, New York, USA.

Stefano Leonardi LEONARDI@DIAG.UNIROMA1.IT
Department of Computer, Control and Management Engineering Antonio Ruberti
Sapienza University of Rome, Rome, Italy.

Matteo Russo MATTEO.RUSSO@EPFL.CH
Institute of Mathematics
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study Online Convex Optimization (OCO) over a convex set $K \subseteq \mathbb{R}^d$, where in each round t the learner selects $x_t \in K$ and then observes a convex loss $f_t : K \to [0,1]$, with the goal of minimizing regret to the best fixed decision in hindsight. We introduce a unified probing model that generalizes two recent lines of work: sublinear *best-expert* queries in the experts setting (Russo et al., 2024), and pairwise (comparison-based) feedback available every round in OCO (Bhaskara et al., 2023b). In our framework, the learner has a budget of $k \leq T$ *pairwise probes*; on a probed round it may query two points and learn which one has smaller loss.

Our main result shows that even a *sublinear and noisy* probe budget can provably improve worst-case regret in the full feedback OCO regime. With k δ-noisy pairwise probes, we obtain

$$\textsc{Reg}_T \leq O\left(\min\left\{\sqrt{dT\ln T},\ \frac{dT\ln T}{k|1-2\delta|}\right\}\right),$$

which is tight (up to logarithmic factors in T) across T, k and δ. Specifically regarding the noise parameter $\delta \in [0,1]$, the regret guarantee smoothly degrades as the oracle response approaches a coin flip, i.e., δ is close to $1/2$. When applying the same techniques to a finite K for the prediction with d experts setting, the resulting rates are instead completely tight in all parameters, including d. Our analysis gives a streamlined treatment of pairwise probing in OCO by quantifying the benefit of probing via a variance reduction effect, combined with a second-order (variance-based) analysis of Continuous Exponential Weights (Bubeck, 2011; de Rooij et al., 2014).

Keywords: Online Convex Optimization, Prediction with Expert Advice, Noisy Probes.

1. Introduction

Online convex optimization (OCO) is a general framework for sequential decision-making in which, over a horizon of T rounds, a learner repeatedly selects a decision x_t from a convex set $K \subseteq \mathbb{R}^d$, incurs an adversarially chosen convex loss $f_t(x_t)$, and receives feedback at the end of each round (Shalev-Shwartz, 2012; Hazan, 2016; Orabona, 2019). The learner's performance is measured by *regret*, defined as the gap between its cumulative loss and that of the best fixed comparator $x^\star \in K$ in hindsight:

$$\textsc{Reg}_T = \sum_{t=1}^{T} f_t(x_t) \; - \; \inf_{x \in K} \sum_{t=1}^{T} f_t(x).$$

In the *full-information* variant, the entire loss function f_t (or equivalently a subgradient oracle at every point) is revealed after each round; classical methods such as Online Gradient Descent and Follow-the-Regularized-Leader achieve $\Theta(\sqrt{T})$ regret (up to problem-dependent geometry factors). In the bandit version, only $f_t(x_t)$ is revealed for the played point x_t.

The OCO framework subsumes Online linear optimization (OLO), where losses are linear and expressed as $f_t(x) = f_t^\top x$ for vector $f_t \in \mathbb{R}^d$. As a further special case, we have the classical online learning with finitely many actions/experts: taking Δ^{d-1} (the $d-1$-dimensional simplex) and linear losses, the regret against the best vertex recovers the usual notion of regret in the experts problem. In this discrete specialization, full feedback reveals all coordinates of f_t, while bandit feedback reveals only $f_t(x_t)$ for the played expert x_t.

Motivated by the recent surge of algorithms augmented by external predictions, hints, or side-information (Mitzenmacher and Vassilvitskii, 2020), we study a model in which the learner may occasionally obtain limited *advance* information about the losses *before* committing to its decision on that round. Most relevant to our setting are Bhaskara et al. (2023b) and Russo et al. (2024), who show that even very restricted query access can dramatically reduce regret. In Bhaskara et al. (2023b), the learner is allowed *pairwise comparison probes*: at the beginning of each round t, before choosing the decision x_t, the learner may choose two candidate decisions and learn which one will incur smaller instantaneous loss in this round, without revealing the loss values themselves. They obtain time-independent regret bounds for online linear optimization and improved rates for OCO under additional curvature assumptions. In the discrete bandits setting, they show that pairwise comparisons can yield regret of order $O(d \ln T)$ in the stochastic case. Complementarily, Russo et al. (2024) allow *best-expert probes* (which they call queries) in the experts special case only, which reveal the identity of the minimizing expert in the current round t (before the decision x_t is played); however, only $k \le T$ such queries can be made. In this setting, they show a regret bound of $O\Big(\min\Big\{\sqrt{T \ln d}, {}^{T \ln d}/_{k}\Big\}\Big)$.

These results raise the following natural question:

If the learner is able to issue only a sublinear number $k \ll T$ of (possibly noisy) <u>pairwise</u> probes, how much can the OCO regret be improved with respect to the base $\sqrt{T}$ rate?

To address this question, we first formalize an OCO-with-probing model that unifies both discrete and continuous settings, and then state our results.

1.1. Our Model

We present a probing-augmented online optimization protocol. The OCO and prediction with expert advice settings arise as two specializations.

1.1.1. ONLINE CONVEX OPTIMIZATION WITH NOISY PROBING

Let $K \subset \mathbb{R}^d$ be either a compact convex set or a finite set. The interaction proceeds for T rounds against an *oblivious adversary* who pre-commits to a sequence of loss functions bounded in $[0,1]$: $\{f_t : K \to [0,1]\}_{t=1}^T$. At each round $t \in \{1,\dots,T\}$, the learner picks $x_t \in K$, incurs loss $f_t(x_t)$, and then receives feedback $f_t(x)$ for all $x \in K$. Moreover, for appropriate functions f and g we define $\langle f, g\rangle = \int_K f(x)g(x)\mu(dx)$, where the base measure μ is explicit or clear from the context.

Probing budget. In addition to the feedback, the learner may issue *probes* in at most k rounds. Let $Q \subseteq \{1,\dots,T\}$ denote the set of probed rounds, decided (possibly randomly and adaptively) by the learner, with $|Q| \le k$. If $t \in Q$, before selecting x_t, the learner may specify a (possibly randomized) finite set of C candidate points $P_t = \{y_{t,1},\dots,y_{t,C}\} \subseteq K$, and the probe returns the identity of the candidate with minimal instantaneous loss in P_t, possibly corrupted by noise. In this paper we focus on the least powerful nontrivial probes, namely pairwise comparisons with $C = 2$.

To model noise, we assume that the probe returns the candidate with the *maximal* (instead of minimal) loss with probability δ, but our results clearly apply to less challenging noise: for example, the oracle could return, with probability δ, any of the two probed points, which is better than returning the $\arg\max$, and our regret bounds still hold. Moreover, any probe model with general $C > 2$ can be reduced to the pairwise case by constructing P_t as multiset made up of $y_{t,1}$ repeated $C-1$ times and $y_{t,2}$ once: our upper bounds thus extend for general $C > 2$.

Definition 1 (Noisy comparison probe) *Fix $\delta \in [0,1]$ and an integer $C \ge 1$. A δ-*noisy (comparison) probe *at round t is a (possibly randomized) set $P_t = \{y_{t,1},\dots,y_{t,C}\} \subseteq K$ together with an outcome:*

$$\hat{y}_t = \begin{cases} y_t^\star \in \arg\min_{y\in P_t} f_t(y), & \textit{with probability } 1-\delta, \\ y_t^- \in \arg\max_{y\in P_t} f_t(y), & \textit{with probability } \delta. \end{cases}$$

Whenever the minimizer or maximizer is not unique, ties are broken uniformly at random. When $\delta = 0$, we call the probe noiseless. *Moreover, a probe is* global *when $P_t = K$ and $\delta = 0$, that is the probe outcome is $y_t^\star \in \arg\min_{y\in K} f_t(y)$ with probability 1, assuming the minimum is achieved.*[1]

Protocol. Formally, first an oblivious adversary pre-commits to a sequence of convex loss functions $f_1,\dots,f_T$ bounded in $[0,1]$, and then, in each round $t = 1,\dots,T$:

1. The learner decides if $t \in Q$: if so, it chooses $P_t \subseteq K$ and observes $\hat{y}_t \in P_t$.
2. The learner chooses $x_t \in K$ (possibly as a randomized function of $\hat{y}_t$ and the past).
3. The learner incurs loss $f_t(x_t)$ and observes full feedback $\{f_t(x)\}_{x\in K}$.

1. If the minimum is not achieved, one may model the notion of global probe by letting it return any point in a sublevel set characterized by a function value arbitrary close to the minimum.

Regret. For any point $x \in K$, define the cumulative loss $L_T(x) = \sum_{t=1}^T f_t(x)$. Similarly, let $L_T(\textsc{Alg}) = \sum_{t=1}^T f_t(x_t)$ be the cumulative loss of a learner $\textsc{Alg}$ whose decisions over rounds are $x_1, \ldots, x_T$, each belonging to K. In turn, these points are sampled from the probability measures induced by the densities $p_1, \ldots, p_T$ set by the learner $\textsc{Alg}$ across rounds. In the following, for any time $t > 0$, we abuse notation and write $x \sim p_t$ for x being sampled from the probability measure induced by p_t and the base measure μ. Moreover, let $v_t = \mathbb{V}_{x \sim p_t}[f_t(x)]$ be the variance of the function at time t according to p_t, and let $V_t = \sum_{\tau \le t} v_\tau$ be the cumulative variance up until time t. The (worst-case) expected regret of an algorithm $\textsc{Alg}$ is

$$\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg})\right] = \sup_{\boldsymbol{f}} \left\{ \mathbb{E}\left[\sum_{t=1}^T f_t(x_t)\right] - \inf_{x \in K} \sum_{t=1}^T f_t(x) \right\},$$

where the expectation is with respect to the learner's internal randomness and the probe noise, and the supremum ranges over all obliviously chosen sequences $\boldsymbol{f} = \{f_t\}_{t=1}^T$ of admissible losses bounded in $[0, 1]$; for online convex optimization, we restrict to convex losses. Our goal is sublinear regret, i.e., $\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg})\right] = o(T)$, while leveraging at most k noisy comparison probes.

1.2. Our Results and Technical Challenges

Our main contribution is to show that even a *sublinear* number of pairwise comparison probes can substantially reduce regret. We unify the two probing paradigms of Bhaskara et al. (2023b); Russo et al. (2024) into a single framework in which the learner is granted a budget of $k \le T$ *pairwise* (possibly noisy) probes over a horizon of T rounds, and we analyze the resulting regret tradeoffs for full feedback OCO and prediction with experts.

Regret guarantees. With k pairwise δ-noisy probes, we prove that one can achieve

$$\textsc{Reg}_T \le O\left(\min\left\{\sqrt{dT \ln T},\ \frac{dT \ln T}{k\,|1 - 2\delta|}\right\}\right), \tag{1}$$

for general convex losses bounded in $[0, 1]$ (see Theorem 2). In the special case of experts (i.e., linear losses over the vertices of the simplex Δ^{d-1}), we can replace a factor of $d \ln T$ by $\ln d$ for both terms of the $\min$ in (1) to get an improved bound of

$$\textsc{Reg}_T \le O\left(\min\left\{\sqrt{T \ln d},\ \frac{T \ln d}{k\,|1 - 2\delta|}\right\}\right). \tag{2}$$

A key takeaway is that the regret bounds of (2) match those obtained in Russo et al. (2024) under *best-expert* probes, where in each queried round the learner effectively "looks ahead" and learns the globally optimal action $x_t^\star$. In the special case of experts *only*, they leverage a crucial property of best-expert probes: on a probed round, the learner can guarantee the minimum possible loss, since it knows the expert with lowest loss at that round. This yields a non-positive (and potentially negative) contribution to regret during probing rounds. Combined with a standard Hedge bound on the remaining rounds, this leads to worst-case regret $O\left(\frac{T \ln d}{k}\right)$ once $k \ge \Omega(\sqrt{T \ln d})$.

Making this argument with pairwise probes runs into a basic obstacle even just for the special case of experts: to benefit, one must compare against (or identify) a near-best expert with sufficiently large probability. One approach to doing this is adapting the arguments from Russo et al. (2024) and mix a uniformly random probe with probes from the distribution of the Hedge algorithm, to avoid hurting the baseline performance. In Appendix B, we carry over this analysis, attaining suboptimal rates even in the noiseless case. In contrast, our algorithms manage to get the (optimal) bounds in Equation (2) using only (noisy) pairwise probes, moreover getting the rate in Equation (1) for general bounded convex losses over an arbitrary convex body K!

Our approach and the main technical insight. Our main theorem (Theorem 2) simultaneously handles (i) general convex losses over general convex bodies, (ii) sublinear probe budgets, and (iii) noisy probes, recovering essentially tight (in terms of T, k and δ) bounds on the regret rate. The algorithm is a direct and adaptive probe-augmented version of *continuous* exponential weights: it maintains a density p_t over K and samples points accordingly (e.g., Bubeck (2011)). This route avoids differential privacy machinery and dispenses with curvature/gradient assumptions (Bhaskara et al., 2023b).

For noiseless probing, the key insight is that a pairwise comparison drawn from p_t yields an advantage that can be measured by the *variance* of the losses under p_t. Concretely, on a probing round t, comparing two i.i.d. samples from p_t and playing the better one decreases the expected loss by $\mathbb{V}_{x \sim p_t}[f_t(x)]$, the variance of the function according to the density p_t:

$$\mathbb{E}_{p_t}[f_t(x_t) \mid t \in Q] \leq \mathbb{E}_{p_t}[f_t(x_t)] - \mathbb{V}_{p_t}[f_t(x_t)] \,.$$

We then couple this with a variance-based analysis of continuous exponential weights (following the less standard "second-order" style guarantees of de Rooij et al. (2014)), which relates the baseline regret to the same variance term. This alignment is what enables the T/k-type improvement without an additional factor in d, nor making assumptions about gradient norms or curvature.

Noisy probes and learning when to trust them. Noisy probes introduce a further challenge: the noise level δ is unknown, so the learner must also infer whether to trust the probe outcome. We cast this as a lightweight meta-learning layer that, on probed rounds, decides whether to *follow* the probe's suggestion or *invert* it. Instantiating this meta-problem with an adaptive two-action learner allows us to recover the bound in (1) with the correct dependence on $|1 - 2\delta|$.

Lower bounds. Finally, we show that the rates in (2) are tight in all parameters for prediction with expert advice, and those in (1) are tight in T, k and δ (up to logarithmic factors in T) for general convex losses over arbitrary convex sets; we give these results in Appendix A.

1.3. On the Power of the Adversary and the Definition of Noisy Probes

The protocol above requires the adversary to fix the loss f_t before the learner decides whether to probe at time t. It is crucial that the adversarial loss construction is not coupled with the probing decision. If the adversary could anticipate a probe, she could set $f_t \equiv 0$ for that round, rendering the probe useless. This would degrade the optimal regret to $\Theta(\sqrt{T - k})$, preventing the faster $\tilde{\Theta}(T/k)$ rates we achieve. This constraint distinguishes our work from the "adversarially noisy" probes considered in the literature (Bhaskara et al., 2020, 2021a, 2023a, 2021b, 2023b). In those settings, the adversary effectively knows when probing occurs. For instance, Bhaskara et al. (2023b) derive a lower bound of $\Omega(\sqrt{T - k})$ by constructing an instance where the adversary sets losses to $(0, 0)$ specifically during noiseless probes (and $(0, 1)$ vs. $(1, 0)$ each with probability $1/2$ otherwise). This contradicts our results because their model assumes the adversary has access to the learner's internal randomization (knowing *when* a probe happens), whereas our model does not.

2. Related Work

The model of learning-augmented algorithms (Mitzenmacher and Vassilvitskii, 2020) has been most recently considered in the context of Online Linear Optimization (OLO) over a convex polytope (Shalev-Shwartz, 2012). Here, the learner has access to a query mechanism that returns vectors correlated with the true loss vectors. When the optimization domain is the d-dimensional unit sphere and the algorithm has full feedback, this additional information enables a reduction in regret from $O(\sqrt{T})$ to $O(\ln T)$, even if the learner only receives such hints on $O(\sqrt{T})$ rounds (Bhaskara et al., 2020, 2021a,b). Interestingly, this is not possible with bandit feedback, and a $\Omega(\sqrt{T})$ regret is unavoidable even if the hint vectors are well correlated with the true ones (Bhaskara et al., 2023a). The same work shows that active queries to specific points at which to evaluate the loss function are enough to guarantee regret rates of the form $O(d^{3/2} \ln T)$, where d is the dimension of the ambient space. Their techniques do not directly apply to our setting, since their probes are defined differently from ours and those from (Bhaskara et al., 2023b). Moreover, in subsequent work, Bacchiocchi et al. (2026) show that under bandit feedback, even in the experts setting and with access to best-arm probes, the worst-case regret remains $\Omega(\sqrt{T-k})$.

Regarding pairwise probes for OCO in particular, the closest work to ours is Bhaskara et al. (2023b), who obtain a $\Theta(Hd + Gd^{3/2})$ regret bound in the full feedback setting when probing is available every round ($k = T$), under uniform bounds G, H on gradient norms and curvature. Their analysis relies on a differential-privacy-based reduction to Be-The-Regularized-Leader (BTRL), and the dependence on curvature and gradient bounds appears intrinsic to their arguments, which we avoid using our techniques.

Techniques developed in this work, as well as in prior ones, aim to adapt to structure in the loss sequence during non-probing rounds, leading to regret guarantees that depend on the "easiness" of the instance. These forms of easiness include stochastic losses or small effective loss ranges. In full-information settings, such structure can be leveraged by algorithms like Hedge to achieve regret depending on the easiness of the instance, without degrading worst-case performance (Cesa-Bianchi et al., 2007; Gaillard et al., 2014; Koolen and van Erven, 2015; Luo and Schapire, 2015). Another relevant direction is the predictable sequences framework, which focuses on scenarios where the loss sequence follows a regular pattern or is correlated with past observations (Rakhlin and Sridharan, 2013). Algorithms in this line adjust their learning dynamics to exploit such predictability and achieve improved regret when it exists (Steinhardt and Liang, 2014; Wei and Luo, 2018). Limited supervision has also been studied through the abstention learning framework, where learners may choose not to predict and instead defer to an oracle, thereby trading off prediction effort with access to external feedback (Li et al., 2011; Sayedi et al., 2010; Zhang and Chaudhuri, 2016; Cortes et al., 2018; Neu and Zhivotovskiy, 2020; Gangrade et al., 2021).

Our work is also connected to stochastic probing, where the learner acts under uncertainty and can selectively probe the environment to obtain partial feedback (Gupta and Nagarajan, 2013; Gupta et al., 2017; Singla, 2018). This model has applications in problems like Pandora's Box (Weitzman, 1979; Beyhaghi and Kleinberg, 2019; Beyhaghi and Cai, 2023), online matching (Singla, 2018), and submodular optimization (Patton et al., 2023). A central challenge in these settings is deciding when and what to probe to optimize long-term performance under uncertainty, and recent algorithms address this by carefully balancing exploration with exploitation (Agarwal et al., 2024).

3. Main Algorithm and Regret Guarantees

In this section, we show how a sublinear number of (noisy) probes enables a significant drop in regret rates. To do so, we recall a generic algorithmic template based on Continuous Exponential Weights with different priors p_1 over the domain K. This template has been used by, e.g., de Rooij et al. (2014) for prediction with expert advice, and by Bubeck (2011) for OCO: we augment it with a META-LEARNER and k pairwise comparison probes issued uniformly at random throughout the time horizon (see below). Different instantiations of Algorithm 1 guarantee regret bounds for general convex classes and stronger bounds for the experts setting (see Section 3.5).

Meta-Learner. To handle the noisy case with unknown δ, the idea is to run a meta-learner on the *"follow the oracle"* and the *"invert the oracle"* actions, succinctly denoted as F and I respectively. Essentially, on probing rounds, we do what the META-LEARNER says: if the routine META-LEARNER returns F, the algorithm follows the probe and plays $\hat{y}_t$, otherwise it plays the other probed point. After receiving feedback, the META-LEARNER is updated based on the value of f_t in the probed points.

Here, the META-LEARNER is simply AdaHedge (see Theorem 5 in Cesa-Bianchi et al. (2007) and Theorem 6 in de Rooij et al. (2014)) run on the two actions F, I and for rounds $Q \subseteq [T]$ only (see pseudocode in Algorithm 1 for convenience). AdaHedge is a standard Hedge routine with a time-adaptive learning rate, which allows getting variance-based regret bounds. The losses for F and I are simply the values of f_t obtained by either following or inverting the oracle.

Algorithm 1: Probe-Augmented Continuous Exponential Weights with Noisy Probes

Input: Sequence of measurable functions f_t, domain $K \subseteq \mathbb{R}^d$, prior density p_1 with respect to base measure μ over K, probe budget $k \leq T$, parameter Λ

Let $Z_1 = \int_K p_1(x)\mu(dx) = 1$, $V_0 = 0$. Uniformly sample k out of T rounds, obtaining Q

for $t = 1, \dots, T$ **do**
- **if** $t \in Q$ **then**
  - Probe 2 points $y_{t,1}, y_{t,2} \sim p_t$ independently and observe $\hat{y}_t$
  - **if** META-LEARNER$(\cdot) = F$ **then**
    - Play $\hat{y}_t$
  - **else**
    - Play $y_{t,1}$ if $y_{t,1} \neq \hat{y}_t$ else play $y_{t,2}$
- **else**
  - Play $x_t \sim p_t$;
- Observe loss function $f_t(x)$ for all $x \in K$;
- Set $V_t = V_{t-1} + v_t$ and $\eta_{t+1} = \min\left(\frac{1}{2}, \sqrt{\frac{\Lambda}{V_t+1}}\right)$
- Update density for all $x \in K$ as

$$p_{t+1}(x) = \frac{p_1(x) \cdot \exp\left(-\eta_{t+1} L_t(x)\right)}{Z_{t+1}} \quad \text{where} \quad Z_{t+1} = \int_K p_1(x) \cdot \exp\left(-\eta_{t+1} L_t(x)\right) \mu(dx)$$

- Update META-LEARNER using $\{f_t(y_{t,1}), f_t(y_{t,2})\}$ **if** $t \in Q$

Theorem 2 *Consider the problem of online convex optimization with full feedback and k pairwise comparison probes: $K \subset \mathbb{R}^d$ is convex and compact, with non-empty interior. Then, for an unknown* *noise parameter $\delta \in [0,1]$ and for any sequence of convex functions $\{f_t\}_{t=1}^T$ bounded in $[0,1]$, when instantiated with p_1 being the uniform density over K and $\Lambda = d \ln T$, Algorithm 1 has regret*

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG}_k)\right] \leq O\left(\min\left(\sqrt{dT\ln T}, \frac{dT\ln T}{k|1-2\delta|}\right)\right).$$

In order to prove the above theorem, let us first write the following decomposition that holds for any fixed density q over K and any fixed point $x^\star \in K$:

$$\mathbb{E}\left[\sum_{t=1}^T f_t(x_t)\right] - \sum_{t=1}^T f_t(x^\star) = \underbrace{\mathbb{E}\left[\sum_{t=1}^T f_t(x_t)\right] - \sum_{t=1}^T \langle q, f_t\rangle}_{\text{(A)}} + \underbrace{\sum_{t=1}^T \langle q, f_t\rangle - \sum_{t=1}^T f_t(x^\star)}_{\text{(B)}}. \tag{3}$$

We bound term (A) and (B) separately; convexity of f_t's is used only for bounding (B).

3.1. Loss during non-probing rounds

In this subsection, we compare the performance of Algorithm 1 without considering probes against any distribution q over the domain K. We assume that $q \ll p_1$, i.e. q is absolutely continuous w.r.t. p_1, where by this we mean that the property is satisfied by the two underlying measures. The analysis is similar in spirit to the one present in Cesa-Bianchi et al. (2007); de Rooij et al. (2014). Before proceeding, we bound the average loss using the log-partition function and a variance term.

Lemma 3 *It holds that*

$$\langle p_t, f_t\rangle \leq \frac{1}{\eta_t}\ln Z_t - \frac{1}{\eta_{t+1}}\ln Z_{t+1} + \frac{\eta_t}{2(1-\eta_t)}\cdot v_t$$

Proof For this proof, we recall the following bound (e.g. Wainwright (2019, Proposition 2.10)):

Proposition 4 (Bernstein Subexponential Tail Bound) *For a bounded random variable X with $|X - \mathbb{E}[X]| \leq 1$, it holds that:*

$$\ln \mathbb{E}\left[\exp(-\eta X)\right] \leq -\eta \cdot \mathbb{E}[X] + \frac{\eta^2}{2(1-|\eta|)}\cdot \mathbb{V}[X] \qquad \forall\, \eta : |\eta| < 1.$$

To use the above, we define an intermediate normalizer, which uses the same learning rate η_t but after observing f_t; conventionally, in what follows, we let $\eta_1 = 1/2$:

$$\begin{aligned}\widetilde{Z}_{t+1} = \int_K p_1(x)\cdot \exp\left(-\eta_t L_t(x)\right)\mu(dx) &= Z_t \cdot \int_K p_t(x)\cdot \exp\left(-\eta_t f_t(x)\right)\mu(dx) \\ &= Z_t \cdot \mathbb{E}_{x_t \sim p_t}\left[\exp\left(-\eta_t f_t(x_t)\right)\right].\end{aligned}$$

Applying Proposition 4 to the bounded random variable $f_t(x_t)$ with density p_t, we then obtain:

$$\begin{aligned}\ln\left(\frac{\widetilde{Z}_{t+1}}{Z_t}\right) = \ln \mathbb{E}_{x_t\sim p_t}\left[\exp\left(-\eta_t f_t(x_t)\right)\right] &\leq -\eta_t \cdot \mathbb{E}_{x_t\sim p_t}\left[f_t(x_t)\right] + \frac{\eta_t^2}{2(1-\eta_t)}\cdot \mathbb{V}_{x_t\sim p_t}\left[f_t(x_t)\right] \\ &= -\eta_t \cdot \langle p_t, f_t\rangle + \frac{\eta_t^2}{2(1-\eta_t)}\cdot v_t.\end{aligned}$$

Hence, using that $\eta_t \leq 1$ and rearranging:

$$\langle p_t, f_t \rangle \leq \frac{1}{\eta_t} \ln \left(\frac{Z_t}{\widetilde{Z}_{t+1}} \right) + \frac{\eta_t}{2(1-\eta_t)} \cdot v_t.$$

Let us now write

$$\begin{aligned} Z_{t+1} = \int_K p_1(x) \cdot \exp\left(-\eta_{t+1} L_t(x)\right) \mu(dx) &= \int_K p_1(x) \cdot \exp\left(-\eta_t L_t(x)\right)^{\eta_{t+1}/\eta_t} \mu(dx) \\ &\leq \left(\int_K p_1(x) \cdot \exp\left(-\eta_t L_t(x)\right) \mu(dx) \right)^{\eta_{t+1}/\eta_t} \\ &= (\widetilde{Z}_{t+1})^{\eta_{t+1}/\eta_t}, \end{aligned}$$

where the inequality follows by Jensen's inequality since the function $y \mapsto y^{\eta_{t+1}/\eta_t}$ is concave in y as $\eta_{t+1} \leq \eta_t$. Combining the above two displays, we have:

$$\langle p_t, f_t \rangle \leq \frac{1}{\eta_t} \ln \left(\frac{Z_t}{\widetilde{Z}_{t+1}} \right) + \frac{\eta_t}{2(1-\eta_t)} \cdot v_t \leq \frac{1}{\eta_t} \ln Z_t - \frac{1}{\eta_{t+1}} \ln Z_{t+1} + \frac{\eta_t}{2(1-\eta_t)} \cdot v_t,$$

as desired. ■

Lemma 5 *Let q be an absolutely continuous density with respect to p_1. If $\Lambda \geq \max(1, \mathrm{KL}(q||p_1))$, it holds that Algorithm 1 without probes, referred to as* ALG_0*, satisfies:*

$$\sum_{t=1}^{T} \langle p_t - q, f_t \rangle \leq 2\Lambda + 5\sqrt{\Lambda(1 + V_T)}.$$

Proof First, we use Lemma 3 and obtain via a telescopic sum throughout t:

$$\sum_{t=1}^{T} \langle p_t, f_t \rangle \leq \frac{1}{\eta_1} \ln Z_1 - \frac{1}{\eta_{T+1}} \ln Z_{T+1} + \sum_{t=1}^{T} \frac{\eta_t}{2(1-\eta_t)} \cdot v_t = -\frac{1}{\eta_{T+1}} \ln Z_{T+1} + \sum_{t=1}^{T} \frac{\eta_t}{2(1-\eta_t)} \cdot v_t,$$

where we have observed that $Z_1 = \int_K p_1(x)\mu(dx) = 1$ by definition, so that $\ln Z_1 = 0$. We next bound the two terms separately. Now, for all $q \ll p_1$, we have the following:

$$\begin{aligned} Z_{T+1} = \int_{x \in K} p_1(x) \exp\left(-\eta_{T+1} L_T(x)\right) \mu(dx) &\geq \int_{x \in K:\, q(x)>0} q(x) \frac{p_1(x)}{q(x)} \exp\left(-\eta_{T+1} L_T(x)\right) \mu(dx) \\ &= \mathbb{E}_{x \sim q} \left[\exp\left(-\eta_{T+1} L_T(x) - \ln\left(\frac{q(x)}{p_1(x)} \right) \right) \right] \\ &\geq \exp\left(\mathbb{E}_{x \sim q} \left[-\eta_{T+1} L_T(x) - \ln\left(\frac{q(x)}{p_1(x)} \right) \right] \right), \end{aligned}$$

where the first and third inequalities hold because the integrand is positive and because of Jensen's inequality, respectively. The expectation is well-defined since $q \ll p_1$. We thus have:

$$-\frac{1}{\eta_{T+1}} \ln Z_{T+1} \leq \frac{1}{\eta_{T+1}} \left(\mathbb{E}_{x \sim q} \left[\eta_{T+1} L_T(x) \right] + \mathrm{KL}(q||p_1) \right) = \sum_{t=1}^{T} \langle q, f_t \rangle + \frac{\mathrm{KL}(q||p_1)}{\eta_{T+1}}.$$

Second, choosing $\eta_t = \min\left(\frac{1}{2}, \sqrt{\frac{\Lambda}{1+V_{t-1}}}\right)$, we have that $\frac{\eta_t}{2(1-\eta_t)} \leq \eta_t$ and also (e.g., Lemma 14 in Gaillard et al. (2014)):

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\eta_t}{2(1-\eta_t)} \cdot v_t \leq 1 + \sum_{t=2}^{T} \eta_t v_t \leq 1 + \sqrt{\Lambda} \cdot \sum_{t=2}^{T} \frac{v_t}{\sqrt{1+V_{t-1}}} \\
&= 1 + \sqrt{\Lambda} \cdot \sum_{t=2}^{T} \frac{v_t}{\sqrt{1+V_t}} + \sqrt{\Lambda} \cdot \sum_{t=2}^{T} v_t \cdot \left(\frac{1}{\sqrt{1+V_{t-1}}} - \frac{1}{\sqrt{1+V_t}}\right) \\
&\leq 1 + \sqrt{\Lambda} \cdot \sum_{t=2}^{T} \frac{v_t}{\sqrt{1+V_t}} + \sqrt{\Lambda} \cdot \sum_{t=2}^{T} \left(\frac{1}{\sqrt{1+V_{t-1}}} - \frac{1}{\sqrt{1+V_t}}\right) \\
&\leq 1 + \sqrt{\Lambda}\left(1 + \sum_{t=2}^{T} \frac{v_t}{\sqrt{1+V_t}}\right) \leq 1 + \sqrt{\Lambda}\left(1 + \int_{V_1}^{V_T} \frac{dx}{\sqrt{1+x}}\right) \\
&\leq 1 + 3\sqrt{\Lambda(1+V_T)}.
\end{aligned}$$

Combining the derivations above, we obtain, using our assumption on $\Lambda \geq \max(1, \mathrm{KL}(q||p_1))$:

$$\sum_{t=1}^{T} \langle p_t - q, f_t \rangle \leq \frac{\mathrm{KL}(q||p_1)}{\eta_{T+1}} + 1 + 3\sqrt{\Lambda(1+V_T)} \leq 2\Lambda + 5\sqrt{\Lambda(1+V_T)},$$

as desired. ■

3.2. Loss during probing rounds

We decompose the loss for probing rounds in two components: (i) the loss assuming δ is known (Lemma 6)—corresponding to the loss suffered by the superior action between F and I; and (ii) the excess regret incurred relative to that superior action (Lemma 7).

In the subsequent lemmas, we denote the instantaneous losses of actions F and I by $\ell_t(F)$ and $\ell_t(I)$, and their cumulative losses over the rounds in Q by $L_Q(F)$ and $L_Q(I)$, respectively. These losses are random variables contingent on the sampling of the outer algorithm and the Bernoulli noise variables governing the probe responses (Definition 1). To avoid conditioning on Q before it is needed, we extend the definition of $\ell_t(F)$ and $\ell_t(I)$ to rounds $t \notin Q$ by considering the necessary Bernoulli noise and pairs $y_{t,1}, y_{t,2}$, which are completely excluded from the learning protocol.

Lemma 6 *The following holds:*

$$\begin{aligned}
\mathbb{E}_{p_t}\left[\ell_t(F)\right] &\leq \langle p_t, f_t \rangle - (1-2\delta)v_t \quad \text{if } \delta \leq 1/2 \\
\mathbb{E}_{p_t}\left[\ell_t(I)\right] &\leq \langle p_t, f_t \rangle + (1-2\delta)v_t \quad \text{if } \delta > 1/2.
\end{aligned}$$

Therefore, accounting for the choice of $Q \subseteq [T]$, the expected cumulative loss accumulated by the better of the two actions F and I is:

$$\min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} \leq \frac{k}{T} \cdot \left(\sum_{t=1}^{T} \langle p_t, f_t \rangle - |1-2\delta| \cdot V_T\right).$$

Proof First, assume $\delta \leq 1/2$. By definition, we have that the action *"follow the oracle"* incurs loss:

$$\begin{aligned}\mathbb{E}_{p_t}\left[\ell_t(F)\right] &= \int_{K\times K} p_t(u)p_t(y)\cdot(1-\delta)\cdot\min(f_t(u),f_t(y)) \\ &\qquad + \delta\cdot\max(f_t(u),f_t(y))\,\mu(du)\,\mu(dy) \\ &= \int_{K\times K} p_t(u)p_t(y)\cdot\frac{f_t(u)+f_t(y)-(1-2\delta)\cdot|f_t(u)-f_t(y)|}{2}\,\mu(du)\,\mu(dy) \\ &\leq \langle p_t, f_t\rangle - (1-2\delta)\cdot\int_{K\times K} p_t(u)p_t(y)\cdot\frac{(f_t(u)-f_t(y))^2}{2}\,\mu(du)\,\mu(dy) \\ &= \langle p_t, f_t\rangle - (1-2\delta)\cdot\left(\int_K p_t(u)f_t^2(u)\,\mu(du) - \langle p_t, f_t\rangle^2\right) \\ &= \langle p_t, f_t\rangle - (1-2\delta)v_t.\end{aligned}$$

Summing over $t\in Q$, we have $\mathbb{E}\left[L_Q(F)\,|\,Q\right] \leq \sum_{t\in Q}\langle p_t, f_t\rangle - (1-2\delta)\cdot\sum_{t\in Q} v_t$. Similarly, if $\delta > 1/2$, we have, via a derivation almost identical to the one above: $\mathbb{E}\left[L_Q(I)\,|\,Q\right] \leq \sum_{t\in Q}\langle p_t, f_t\rangle + (1-2\delta)\sum_{t\in Q} v_t$, thus getting a bound for the action *"invert the oracle"*. Hence, taking another expectation over Q and combining the two cases depending on the sign of $1-2\delta$, we get:

$$\min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} \leq \frac{k}{T}\cdot\left(\sum_{t=1}^{T}\langle p_t, f_t\rangle - |1-2\delta|\cdot\sum_{t=1}^{T} v_t\right),$$

which concludes the proof. ■

Lemma 7 *The* META-LEARNER *has expected cumulative loss bounded by*

$$\mathbb{E}\left[L_Q(\text{META-LEARNER})\right] \leq \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + \sqrt{\frac{2k}{T}V_T} + 3.$$

Proof First, recall that the META-LEARNER runs Hedge with adaptive learning rate (i.e., AdaHedge from de Rooij et al. (2014) run on the two actions F, I and for rounds $Q\subseteq[T]$ only). For any $t\in Q$, let $\tilde{p}_{t,F}, \tilde{p}_{t,I}$ be the probabilities maintained by the META-LEARNER over F, I. Similarly, let $\tilde{v}_t$ be the instantaneous variance induced by $\tilde{p}_{t,F}, \tilde{p}_{t,I}$ of the META-LEARNER, i.e., $\tilde{v}_t = \tilde{p}_{t,F}\tilde{p}_{t,I}(\ell_t(F)-\ell_t(I))^2$, and let $\tilde{V}_Q = \sum_{t\in Q}\tilde{v}_t$. Observe that

$$\tilde{v}_t = \tilde{p}_{t,F}\tilde{p}_{t,I}(\ell_t(F)-\ell_t(I))^2 \leq \frac{1}{4}(\ell_t(F)-\ell_t(I))^2.$$

Also note that $|\ell_t(F)-\ell_t(I)| = |f_t(y_{t,1})-f_t(y_{t,2})| = \Delta_t$, which means that $\tilde{V}_Q \leq \frac{1}{4}\sum_{t\in Q}\Delta_t^2$. In particular, since $y_{t,1}$ and $y_{t,2}$ are i.i.d. and taking the difference outputs a centered random variable, $\mathbb{E}\left[\sum_{t\in Q}\Delta_t^2\,|\,Q\right] = \sum_{t=1}^{T} 2v_t\mathbb{I}\{t\in Q\}$, where v_t is the variance at t of the original learning task.

In addition, Theorem 6 in (de Rooij et al., 2014) gives a regret of the META-LEARNER (AdaHedge) with respect to the better of F, I of at most, letting $\mathcal{G}$ be the sigma-algebra generated by Q, the noise variables and the sampling from p_t:

$$\mathbb{E}\left[L_Q(\text{META-LEARNER}) - \min\{L_Q(F), L_Q(I)\}\,|\,\mathcal{G}\right] \leq 2\sqrt{\tilde{V}_Q\ln 2} + \frac{4\ln 2}{3} + 2 \leq 2\sqrt{\tilde{V}_Q} + 3.$$

Combining the above and taking expectations, we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[L_Q(\text{META-LEARNER}) \,|\, \mathcal{G}\right]\right] &\leq \mathbb{E}\left[\min\{L_Q(F), L_Q(I)\}\right] + 2\mathbb{E}\left[\sqrt{\tilde{V}_Q}\right] + 3 \\
&\leq \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + 2\sqrt{\mathbb{E}\left[\tilde{V}_Q\right]} + 3 \\
&\leq \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + \sqrt{\mathbb{E}\left[\mathbb{E}\left[\sum_{t\in Q} \Delta_t^2 \,\middle|\, Q\right]\right]} + 3 \\
&= \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + \sqrt{\sum_{t=1}^{T} 2v_t \mathbb{P}\left[t \in Q\right]} + 3 \\
&= \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + \sqrt{\frac{2k}{T} V_T} + 3,
\end{aligned}$$

where the second inequality holds by using Jensen's inequality twice, and the second-to-last equality by using standard properties of conditional expectations. ∎

3.3. Proving the regret bound

We can finally prove the claimed regret bound from Theorem 2, and we start by bounding term (A) in Equation (3) using the results from the previous two subsections.

Lemma 8 *Let q be a probability density function on the domain K, such that $q \ll p_1$. Then, for an unknown noise parameter $\delta \geq 0$ and for any sequence of measurable functions f_t bounded in $[0, 1]$, Algorithm 1 guarantees, provided $\Lambda \geq \max(1, \mathrm{KL}(q||p_1))$:*

$$\mathbb{E}\left[L_T(\text{ALG}_k)\right] - \sum_{t=1}^{T} \langle q, f_t \rangle \leq \min\left(12\sqrt{\Lambda T}, \frac{25\Lambda T}{k|1-2\delta|}\right) + 4.$$

Proof Let us decompose the loss suffered by ALG_k into probing and non-probing rounds:

$$\mathbb{E}\left[L_T(\text{ALG}_k)\right] = \mathbb{E}\left[L_{T\setminus Q}(\text{ALG}_k)\right] + \mathbb{E}\left[L_Q(\text{ALG}_k)\right].$$

In particular, note that the loss suffered by ALG_k during the probing rounds is equal to that of the META-LEARNER. Since we have the same type of update to p_t in both probing and non-probing rounds, we have that $\mathbb{E}\left[L_{T\setminus Q}(\text{ALG}_k) \,|\, Q\right] = \sum_{t=1}^{T} \mathbb{I}\left\{t \notin Q\right\} \mathbb{E}\left[f_t(x_t)\right]$, and since Q is uniform:

$$\mathbb{E}\left[L_{T\setminus Q}(\text{ALG}_k)\right] = \left(1 - \frac{k}{T}\right) \sum_{t=1}^{T} \langle p_t, f_t \rangle.$$

By Lemma 6 and Lemma 7, it also holds that:

$$\begin{aligned}
\mathbb{E}\left[L_Q(\text{ALG}_k)\right] = \mathbb{E}\left[L_Q(\text{META-LEARNER})\right] &\leq \min\{\mathbb{E}\left[L_Q(F)\right], \mathbb{E}\left[L_Q(I)\right]\} + \sqrt{\frac{2k}{T} V_T} + 3 \\
&\leq \frac{k}{T} \cdot \left(\sum_{t=1}^{T} \langle p_t, f_t \rangle - |1-2\delta| \cdot V_T\right) + \sqrt{\frac{2k}{T} V_T} + 3.
\end{aligned}$$

Therefore, combining the two above displays, we get

$$\begin{aligned}\mathbb{E}\left[L_T(\textsc{Alg}_k)\right] &\leq \sum_{t=1}^{T} \langle p_t, f_t \rangle + \sqrt{\frac{2k}{T} V_T} + 3 - \frac{k|1-2\delta|}{T} \cdot V_T \\ &\leq \sum_{t=1}^{T} \langle q, f_t \rangle + 2\Lambda + \underbrace{5\sqrt{\Lambda(1+V_T)} + \sqrt{\frac{2k}{T} V_T}}_{\leq 7\sqrt{\Lambda(1+V_T)}} + 3 - \frac{k|1-2\delta|}{T} \cdot V_T \\ &\leq \sum_{t=1}^{T} \langle q, f_t \rangle + \min\left(10\sqrt{\Lambda T}, \frac{23\Lambda T}{k|1-2\delta|}\right) + 4 + 2\Lambda,\end{aligned}$$

where the second inequality holds by Lemma 5, while the third because $V_T \leq T$ and the maximizing V_T is $V_T = \frac{49T^2\Lambda}{4k^2(1-2\delta)^2} - 1$. To get the final bound for the first argument of the minimum, notice that $\Lambda \leq \sqrt{\Lambda T}$ if $\Lambda \leq T$, while if $\Lambda > T$, then the loss is trivially bounded by $T < \sqrt{\Lambda T}$. The bound for the second argument follows since $T/k|1-2\delta| \geq 1$, with the convention of setting it to ∞ when $\delta = 1/2$. ■

We now prove Theorem 2 by removing the dependency of the previous analysis on q.

Proof [of Theorem 2] First, define $S(x^\star, r) = (1-r)x^\star + rK$, for a generic $x^\star \in K$. Now fix any $t \in [T]$ and let q be the uniform density over $S(x^\star, r)$, i.e. $q(x) = \frac{1}{\mathrm{Vol}(S(x^\star,r))}\mathbb{I}\{x \in S(x^\star, r)\}$: the base measure in Algorithm 1 is thus in this case the Lebesgue measure. By definition of Minkowski sum, for any $y \in S(x^\star, r)$ there exists $x \in K$ such that $y = (1-r)x^\star + rx$. By convexity of f_t,

$$f_t(y) \leq (1-r)f_t(x^\star) + r f_t(x) \leq (1-r)f_t(x^\star) + r,$$

since $f_t(x) \leq 1$. Rearranging then gives $f_t(y) - f_t(x^\star) \leq r$ for all $y \in S(x^\star, r)$. This addresses the (B) term in Equation (3).

To address the (A) term, notice that $q \ll p_1$ and $\mathrm{KL}(q||p_1) = d\ln(1/r) = \Lambda$, since:

$$\mathrm{KL}(q||p_1) = \ln\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(S(x^\star, r))}\right) = \ln\left(\frac{\mathrm{vol}(K)}{r^d \mathrm{vol}(K)}\right) = d\ln\left(\frac{1}{r}\right).$$

Plugging the equality in the bound from Lemma 8, then yields:

$$\mathbb{E}\left[L_T(\textsc{Alg}_k)\right] - \sum_{t=1}^{T} f_t(x^\star) \leq \min\left(12\sqrt{dT\ln(1/r)}, \frac{25dT\ln(1/r)}{k|1-2\delta|}\right) + 4 + rT,$$

and setting $r = 1/T$ gives the claimed bound. ■

3.4. Prediction with Expert Advice

In the prediction with expert advice setting, Equation (3) and Lemma 8 imply a regret bound, with the proof relying on the fact that the convexity of domains and functions was only used when proving Theorem 2. We thus have to appropriately switch to the discrete setting: the domain is finite, where the experts can be seen as the d vertices of $\Delta^{d-1} = \{x \in \mathbb{R}^d \mid x_i \geq 0 \,\forall\, i, \sum_{i=1}^{d} x_i = 1\}$, the $(d-1)$-dimensional simplex. Recall that its set of vertices is $\{e_1, \ldots, e_d\}$.[2]

2. We stress that we are using this notation for finite sets of points to match the notation used in the lower bounds in Appendix A.

Theorem 9 *Consider the problem of prediction with expert advice with full feedback and k pairwise comparison probes, with the domain $K = \{e_1, \ldots, e_d\}$. Then, for an unknown noise parameter $\delta \geq 0$ and for any sequence of functions $\{f_t\}_{t=1}^T$ in $[0, 1]$, when instantiated with p_1 being the mass function of the uniform distribution over $\{e_1, \ldots, e_d\}$ and $\Lambda = \ln d$, Algorithm 1 has regret*

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG}_k)\right] \leq O\left(\min\left(\sqrt{T \ln d}, \frac{T \ln d}{k|1-2\delta|}\right)\right).$$

Proof For this result, the base measure of Algorithm 1 is the counting measure over $\{e_1, \ldots, e_d\}$, while p_1 is the probability mass function of the uniform distribution over the same set. Consider the mass function $q(x) = \mathbb{I}\{x = x^\star\}$, thus corresponding to a Dirac on $x^\star$, here the minimizing expert, so that $\mathbb{E}_{x\sim q}[f_t(x)] = f_t(x^\star)$, and term (B) in Equation (3) is null. Clearly, $q \ll p_1$. For term (A), we have:

$$\mathrm{KL}(q||p_1) = \sum_{x \in \{e_1, \ldots, e_d\}} q(x) \cdot \ln\left(\frac{q(x)}{p_1(x)}\right) = \ln d.$$

Therefore, since we set $\Lambda = \ln d$, Lemma 8 gives:

$$\mathbb{E}\left[L_T(\mathrm{ALG}_k)\right] - \sum_{t=1}^T f_t(x^\star) \leq \min\left(12\sqrt{T \ln d}, \frac{25 T \ln d}{k|1-2\delta|}\right) + 4,$$

which was to be shown. ■

4. Concluding Remarks

In this work, we analyze Online Convex Optimization (OCO) with a sublinear budget of pairwise probes. This generalizes Bhaskara et al. (2023b), which requires pairwise probes at *every* step, as well as Russo et al. (2024), which focuses on the *stronger* best-expert probes. Our main finding is that pairwise probes are as informative as best-expert probes and that our algorithms are robust to noise. In Theorem 2, we establish the following regret rates, which are tight in T, k and δ up to logarithmic factors in T, and for experts completely tight also in d:

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG}_k)\right] \leq O\left(\min\left(\sqrt{dT \ln T}, \frac{dT \ln T}{k|1-2\delta|}\right)\right).$$

Among the remaining open questions, one concerns understanding the benefit of probes in the special case of exp-concave functions, as opposed to general convex ones. Another relates to the trade-off between computational efficiency and improved regret: our main CEW algorithm requires computing an integral over a convex set, which is hard in general. It would therefore be interesting to develop efficient algorithms that achieve comparable guarantees.

Acknowledgments

A.G. was supported in part by NSF awards CCF-2422926 and CCF-2608359. S.D.G. and S.L. were supported in part by the PNRR MUR project IR0000013-SoBigData.it project, by the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society) and by the FAIR (Future Artificial Intelligence Research) project PE0000013, funded by the NextGenerationEU program within the PNRR- PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence). S.D.G. was also supported in part by the Institute for Complex Systems (Italian National Research Council).

References

Arpit Agarwal, Rohan Ghuge, and Viswanath Nagarajan. Semi-bandit learning for monotone stochastic optimization. In *FOCS*, pages 1260–1274. IEEE, 2024.

Francesco Bacchiocchi, Matteo Castiglioni, Alberto Marchesi, and Francesco Emanuele Stradi. Multi-armed bandits with best-action queries. *CoRR*, abs/2605.08287, 2026.

Hedyeh Beyhaghi and Linda Cai. Pandora's problem with nonobligatory inspection: Optimal structure and a PTAS. In *STOC*, pages 803–816. ACM, 2023.

Hedyeh Beyhaghi and Robert Kleinberg. Pandora's problem with nonobligatory inspection. In *EC*, pages 131–132. ACM, 2019.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online learning with imperfect hints. In *ICML*, Proceedings of Machine Learning Research, pages 822–831. PMLR, 2020.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Power of hints for online learning with movement costs. In *AISTATS*, Proceedings of Machine Learning Research, pages 2818–2826. PMLR, 2021a.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Logarithmic regret from sublinear hints. In *NeurIPS*, pages 28222–28232, 2021b.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Bandit online linear optimization with hints and queries. In *ICML*, Proceedings of Machine Learning Research, pages 2313–2336. PMLR, 2023a.

Aditya Bhaskara, Sreenivas Gollapudi, Sungjin Im, Kostas Kollias, and Kamesh Munagala. Online learning and bandits with queried hints. In *ITCS*, volume 251 of *LIPIcs*, pages 16:1–16:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023b.

Sebastien Bubeck. Introduction to online optimization. In *Lecture Notes Princeton University*, 2011.

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

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1067–1075. PMLR, 2018.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *J. Mach. Learn. Res.*, 15(1):1281–1316, 2014.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 176–196. JMLR.org, 2014.

Aditya Gangrade, Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. Online selective classification with limited feedback. In *NeurIPS*, pages 14529–14541, 2021.

Robert D. Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Anupam Gupta and Viswanath Nagarajan. A stochastic probing problem with applications. In *IPCO*, volume 7801 of *Lecture Notes in Computer Science*, pages 205–216. Springer, 2013.

Anupam Gupta, Viswanath Nagarajan, and Sahil Singla. Adaptivity gaps for stochastic probing: Submodular and XOS functions. In *SODA*, pages 1688–1702. SIAM, 2017.

Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3-4):157–325, 2016.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1155–1175. JMLR.org, 2015.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Mach. Learn.*, 82(3):399–443, 2011.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adanormalhedge. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1286–1304. JMLR.org, 2015.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. In *Beyond the Worst-Case Analysis of Algorithms*, pages 646–662. Cambridge University Press, 2020.

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 3030–3048. PMLR, 2020.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

Kalen Patton, Matteo Russo, and Sahil Singla. Submodular norms with applications to online facility location and stochastic probing. In *APPROX/RANDOM*, LIPIcs, pages 23:1–23:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 993–1019. JMLR.org, 2013.

Matteo Russo, Andrea Celli, Riccardo Colini-Baldeschi, Federico Fusco, Daniel Haimovich, Dima Karamshuk, Stefano Leonardi, and Niek Tax. Online learning with sublinear best-action queries. In *NeurIPS*, 2024.

Amin Sayedi, Morteza Zadimoghaddam, and Avrim Blum. Trading off mistakes and don't-know predictions. In *NIPS*, pages 2092–2100. Curran Associates, Inc., 2010.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, 2012.

I. G. Shevtsova. On the absolute constants in the Berry-Esseen inequality and its structural and nonuniform improvements. *Inform. Primen.*, 7(1):124–125, 2013.

I. G. Shevtsova. On the absolute constants in the berry-esseen-type inequalities. *Doklady Mathematics*, 89(3):378–381, May 2014.

Sahil Singla. The price of information in combinatorial optimization. In *SODA*, pages 2523–2532. SIAM, 2018.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*, JMLR Workshop and Conference Proceedings, pages 1593–1601. JMLR.org, 2014.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *COLT*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

Martin L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979.

Chicheng Zhang and Kamalika Chaudhuri. The extended littlestone's dimension for learning with mistakes and abstentions. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 1584–1616. JMLR.org, 2016.

Appendix A. Lower Bounds

In this appendix, we recall and (slightly) generalize a result of (see Russo et al. (2024, Section 3)), which shows that the regret rates from Theorem 2 are tight in T, k and δ, up to logarithmic factors in T. Their result can indeed be extended to work for linear (and, thus, general convex) functions. In particular, these lower bounds hold for a learner that has access to the more powerful *global probes*, i.e., it observes the global minimum of the function at that round, and the probe is noiseless. Before stating and proving the result, we stress that it shows complete tightness (also in d) of the rate in Theorem 9, since the proof below reduces to a lower bound on the problem of prediction with expert advice on the vertices of Δ^{d-1}.

Theorem 10 *There exists a family of instances defined as linear functions bounded in* $[0, 1]$ *over the* $(d-1)$*-dimensional simplex* Δ^{d-1} *such that for* $k \leq O\left(\frac{T \ln^{3/2} d}{d}\right)$ *and* $T \geq O\left(\frac{d^2}{\ln^2 d}\right)$*:*[3]

(1) Any algorithm equipped with k *global probes must suffer regret*

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG})\right] \geq \Omega\left(\min\left\{\sqrt{T \ln d}, \frac{T \ln d}{k}\right\}\right);$$

(2) Any algorithm equipped with k *pairwise comparison* δ*-noisy probes must suffer regret*

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG})\right] \geq \Omega\left(\min\left\{\sqrt{T \ln d}, \frac{T \ln d}{k|1-2\delta|}\right\}\right).$$

Proof The proof uses Yao's Minimax principle, so we need to prove a hardness result for a deterministic algorithm against a randomized sequence of losses. The family of instances is described as the following linear function $f_t(x) = f_t^\top x$ defined on the $(d-1)$-dimensional simplex Δ^{d-1}: each coordinate of the vector f_t is 1 with probability ξ independently from all other coordinates and across rounds.

Let us first observe that $\min_{x \in \Delta^{d-1}} \sum_{t=1}^T f_t^\top x \leq \min_{x \in \{e_1, \ldots, e_d\}} \sum_{t=1}^T f_t^\top x$ since the vertices are part of the simplex. In addition, since $\sum_{i=1}^d x_i = 1$ for all $x \in \Delta^{d-1}$ and the vertices $\{e_i\}_{i=1}^d$ form the canonical basis in $\mathbb{R}^d$, we have

$$\begin{aligned}
\min_{x \in \Delta^{d-1}} \sum_{t=1}^T f_t^\top x &= \min_{x \in \Delta^{d-1}} \sum_{t=1}^T \sum_{i=1}^d f_{ti} x_i = \min_{x \in \Delta^{d-1}} \sum_{i=1}^d \left(\sum_{t=1}^T f_{ti}\right) x_i \\
&\geq \min_{x \in \Delta^{d-1}} \sum_{i=1}^d \left(\min_{j=1}^d \sum_{t=1}^T f_{tj}\right) x_i = \min_{j=1}^d \sum_{t=1}^T f_{tj} = \min_{x \in \{e_1, \ldots, e_d\}} \sum_{t=1}^T f_t^\top x.
\end{aligned}$$

Therefore, $\min_{x \in \Delta^{d-1}} \sum_{t=1}^T f_t^\top x = \min_{x \in \{e_1, \ldots, e_d\}} \sum_{t=1}^T f_t^\top x$ and we can restrict our attention to vertices $\{e_1, \ldots, e_d\}$ of the simplex, as the minimal cumulative loss is achieved at one of them.

By the above construction, for each $x \in \{e_1, \ldots, e_d\}$, it holds that $\sum_{t=1}^T f_t^\top x \sim \mathrm{Bin}(T, \xi)$, which means that $\mathbb{E}\left[\sum_{t=1}^T f_t^\top x\right] = T\xi$. Hence, the expected minimal loss among vertices is, by

3. We require this to satisfy the sufficient condition in Equation (5) needed to apply Lemma 12: in particular, $T \geq {}^{200d^2}/_{\ln d}$ and $k \leq {}^{T \ln^{3/2} d}/_{700d}$ suffices.

Lemma 12 (below) with d sufficiently large (so that $2\ln d - 3\ln\ln d \geq \ln d$):

$$\mathbb{E}\left[\min_{x\in\{e_1,\dots,e_d\}} \sum_{t=1}^{T} f_t^\top x\right] \leq T\xi - \frac{1}{25}\sqrt{T\xi(1-\xi)\ln d}$$

Furthermore, by linearity and independence of f_t from the past, we have that for all $x_t \in \Delta^{d-1}$:

$$\mathbb{E}\left[f_t^\top x_t\right] = \mathbb{E}\left[\sum_{i=1}^{d} f_{ti}x_{ti}\right] = \sum_{i=1}^{d}\mathbb{E}\left[f_{ti}\right]x_{ti} = \xi\cdot\sum_{i=1}^{d}x_{ti} = \xi,$$

since $\sum_{i=1}^{d} x_{ti} = 1$. Therefore the expected cumulative loss of the algorithm on the $T-k$ non-probing rounds is exactly $(T-k)\xi$.

We can now proceed with the proof of (1): the algorithm's expected loss during probing rounds is $k\xi^d$ because we need all vertices to have loss 1 for the minimum to be 1. Thus, the overall expected loss of the algorithm reads

$$(T-k)\xi + k\xi^d = T\xi - k\xi(1-\xi^{d-1}).$$

The regret reads

$$\begin{aligned}\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg})\right] &\geq \frac{1}{25}\left(\sqrt{T\xi(1-\xi)\ln d} - 25k\xi(1-\xi^{d-1})\right)\\ &\geq \frac{1}{5\cdot 10^3}\min\left\{\sqrt{T\ln d}, \frac{T\ln d}{k}\right\},\end{aligned}$$

where we have chosen $\xi = \frac{1}{2}$ for $k \leq \frac{1}{50}\sqrt{T\ln d}$ and $\xi = \frac{T\ln d}{5\cdot 10^3 k^2} < \frac{1}{2}$ otherwise.

We conclude with the proof of (2): at any time step, let us condition on the past randomness of the loss functions and let us assume that, given this conditioning, the deterministic learner chooses to probe: fix the two queried points. Since the current loss vector is independent of the past, its coordinates are still i.i.d. $\mathrm{Bernoulli}(\xi)$. Now, the pairwise probe outcome is a single noisy binary signal about the current loss vector and a $\mathrm{Bernoulli}(1/2)$ random bit handling tie breaking. Moreover, since the learner is allowed to know δ, if $\delta > 1/2$ it may invert the probe outcome. Thus the effective probability that this binary signal is incorrect is $\min\{\delta, 1-\delta\}$.

By Lemma 11 applied with $\rho = \min\{\delta, 1-\delta\}$, even if the learner is allowed to play an arbitrary point of Δ^{d-1} after observing this signal, its expected loss on the probed round is at least

$$2\min\{\delta, 1-\delta\}\cdot\xi + (1-2\min\{\delta,1-\delta\})\cdot\xi^2 = \xi - |1-2\delta|\xi(1-\xi).$$

Therefore, if the learner uses $k' \leq k$ probes, then

$$\mathbb{E}\left[L_T(\textsc{Alg})\right] \geq (T-k')\xi + k'\left(\xi - |1-2\delta|\xi(1-\xi)\right) \geq T\xi - k|1-2\delta|\xi(1-\xi).$$

Consequently,

$$\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg})\right] \geq \frac{1}{25}\left(\sqrt{T\xi(1-\xi)\ln d} - 25k|1-2\delta|\xi(1-\xi)\right).$$

We split into two cases: first, if $|1-2\delta| \le \sqrt{T \ln d}/50k$, we choose $\xi = 1/2$, and then

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG})\right] \ge \frac{1}{25}\left(\frac{1}{2}\sqrt{T\ln d} - \frac{25}{4}k|1-2\delta|\right) \ge \frac{3}{200}\sqrt{T\ln d}.$$

Otherwise, $|1-2\delta| > \sqrt{T \ln d}/50k$, and we choose $\xi \in (0, 1/2]$ such that

$$\xi(1-\xi) = \frac{T\ln d}{10^4 k^2 |1-2\delta|^2},$$

which is feasible because the case assumption implies the right-hand side is at most $1/4$. With this choice,

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG})\right] \ge \frac{1}{25}\left(\frac{T\ln d}{100k|1-2\delta|} - \frac{25T\ln d}{10^4 k|1-2\delta|}\right) \ge \frac{T\ln d}{4\cdot 10^3 k|1-2\delta|}.$$

Combining the two cases yields

$$\mathbb{E}\left[\mathrm{REG}_T(\mathrm{ALG})\right] \ge \min\left\{\frac{3}{200}\sqrt{T\ln d}, \frac{T\ln d}{4\cdot 10^3 k|1-2\delta|}\right\},$$

with the convention that the second term is ∞ when $\delta = 1/2$. ■

As promised in the earlier proof, we are left to show that one noisy binary signal cannot reduce Bernoulli loss too much. We then also have to bound the expected minimum of i.i.d. binomials.

We begin with the first claim: the uniform tie breaking convention after Definition 1 ensures that the probe outcome is only a noisy binary signal of the comparison and a $\mathrm{Bernoulli}(1/2)$ random bit which settles ties.

Lemma 11 *Let $A = (A_1, \dots, A_d)$, where the coordinates are independent* $\mathrm{Bernoulli}(\xi)$ *random variables, and let $U \sim \mathrm{Bernoulli}(1/2)$, independently of A. Let $S \in \{0,1\}$ be any binary random variable measurable with respect to (A, U). Consider $B \in \{0,1\}$ to be a noisy version of S defined as S with probability $1-\rho$ and $1-S$ otherwise, where the random choice above is independent of (A, U), and where $\rho \in [0, 1/2]$. Letting $z_B = f(B)$, for a function $f : \{0,1\} \to \Delta^{d-1}$, it holds that:*

$$\mathbb{E}\left[A^\top z_B\right] \ge 2\rho\xi + (1-2\rho)\xi^2.$$

Proof As for the proof of Theorem 10, we have that since the function f maps to the simplex Δ^{d-1}, it suffices to prove the lower bound for f mapping to $\{e_i\}_{i=1}^d \subseteq \Delta^{d-1}$. Let $f(0) = z_0$ and $f(1) = z_1$, for $z_0, z_1 \in \{e_i\}_{i=1}^d$. We want to lower bound the following:

$$\begin{aligned}\mathbb{E}\left[A^\top z_B\right] &= \mathbb{E}\left[\mathbb{I}\{B=0\}A^\top z_0 + \mathbb{I}\{B=1\}A^\top z_1\right] \\ &= (1-\rho)\mathbb{E}\left[\mathbb{I}\{S=0\}A^\top z_0 + \mathbb{I}\{S=1\}A^\top z_1\right] \\ &\quad + \rho\mathbb{E}\left[\mathbb{I}\{S=0\}A^\top z_1 + \mathbb{I}\{S=1\}A^\top z_0\right] \qquad (4)\end{aligned}$$

Disregarding the convex combination, summing the two expectations above gives $\mathbb{E}\left[A^\top(z_0+z_1)\right] = 2\xi$, since A is a Bernoulli vector and $z_0, z_1 \in \{e_i\}_{i=1}^d$. This means that:

$$\mathbb{E}\left[\mathbb{I}\{S=0\}\,A^\top z_1 + \mathbb{I}\{S=1\}\,A^\top z_0\right] = 2\xi - \mathbb{E}\left[\mathbb{I}\{S=0\}\,A^\top z_0 + \mathbb{I}\{S=1\}\,A^\top z_1\right].$$

Letting $\zeta = \mathbb{E}\left[\mathbb{I}\{S=0\}\,A^\top z_0 + \mathbb{I}\{S=1\}\,A^\top z_1\right]$ and plugging the above in Equation (4) we get: $\mathbb{E}\left[A^\top z_B\right] = (1-\rho)\zeta + \rho(2\xi - \zeta) = 2\rho\xi + \zeta(1-2\rho)$.[4] To prove the claim, it thus suffices to prove $\zeta \geq \xi^2$.

The first case is $z_0 = z_1$, and the lower bound follows simply because $\zeta = \mathbb{E}\left[A^\top z_0\right] = \xi \geq \xi^2$. Now suppose that $z_0 \neq z_1$: the event $\{A^\top z_0 = 1\} \cap \{A^\top z_1 = 1\}$ has probability ξ^2, since A is a vector of independent Bernoulli variables and the event happens if and only if the entries in A corresponding to the two non-null entries in z_0 and z_1 are non-null. Therefore,

$$\begin{aligned}
\zeta &= \mathbb{E}\left[\mathbb{I}\{S=0\}\,A^\top z_0 + \mathbb{I}\{S=1\}\,A^\top z_1\right] \\
&\geq \mathbb{E}\left[\mathbb{I}\left\{\{A^\top z_0 = 1\} \cap \{A^\top z_1 = 1\}\right\}\left(\mathbb{I}\{S=0\}\,A^\top z_0 + \mathbb{I}\{S=1\}\,A^\top z_1\right)\right] \\
&= \mathbb{P}\left[A^\top z_0 = 1 \wedge A^\top z_1 = 1\right] = \xi^2,
\end{aligned}$$

as desired. ■

We return to bounding the expected minimum of i.i.d. binomials from above. The following result is arguably folklore, but in the absence of a convenient source we provide a proof for completeness.

Lemma 12 *Let $d \geq 2$ be an integer, let $\xi \in (0,1)$ and $T \in \mathbb{N}$, and let $X_1, \ldots, X_d$ be i.i.d.* $\mathrm{Bin}(T,\xi)$. *Define*

$$\mu = \mathbb{E}[X_1] = T\xi, \qquad \sigma^2 = \mathbb{V}[X_1] = T\xi(1-\xi), \qquad \psi(s) = \mathbb{P}[X_1 \leq \mu - s] \quad (s \geq 0).$$

Let

$$u_d = \sqrt{2\ln d - 3\ln\ln d}, \qquad s_d = \sigma\, u_d.$$

Assume the following:

$$\frac{0.4690}{\sqrt{T\xi(1-\xi)}} \;\leq\; \frac{\ln d}{8\sqrt{\pi}\, d}. \tag{5}$$

Then

$$\mathbb{E}\left[\min_{1\leq i\leq d} X_i\right] \;\leq\; T\xi \;-\; s_d\Big(1 - d^{-1/(8\sqrt{\pi})}\Big). \tag{6}$$

In particular, since $d \geq 2$, we also have the weaker bound:

$$\mathbb{E}\left[\min_{1\leq i\leq d} X_i\right] \;\leq\; T\xi \;-\frac{1}{25}\sqrt{T\xi(1-\xi)}\;\sqrt{2\ln d - 3\ln\ln d}\,.$$

4. On the event that the coordinates corresponding to z_0 and z_1 are both equal to 1, the quantity inside the expectation defining ζ equals 1 regardless of whether $S=0$ or $S=1$. Thus the argument is insensitive to how ties are broken; it only uses that S is a binary random variable measurable with respect to (A, U).

Proof Let us first define

$$M = \max_{1 \le i \le d} (\mu - X_i) \ \ge\ 0,$$

so that then $\min_i X_i = \mu - M$, and hence

$$\mathbb{E}\left[\min_{1 \le i \le d} X_i\right] = \mu - \mathbb{E}\left[M\right]. \tag{7}$$

The proof proceeds in steps:

Step 1. We give a general lower bound on $\mathbb{E}\left[M\right]$ in terms of ψ: for any $s \ge 0$, by independence,

$$\mathbb{P}\left[M < s\right] = \mathbb{P}\left[\mu - X_i < s \ \forall i\right] = \prod_{i=1}^{d} \mathbb{P}\left[\mu - X_i < s\right] = (1 - \psi(s))^d.$$

Therefore

$$\mathbb{P}\left[M \ge s\right] = 1 - (1 - \psi(s))^d. \tag{8}$$

Using $\mathbb{E}\left[M\right] = \int_0^\infty \mathbb{P}\left[M \ge t\right] \, dt$, we obtain for any $s \ge 0$,

$$\mathbb{E}\left[M\right] \ \ge \int_0^s \mathbb{P}\left[M \ge t\right] dt \ \ge\ \int_0^s \mathbb{P}\left[M \ge s\right] \, dt \ =\ s\,\mathbb{P}\left[M \ge s\right] \ =\ s\big(1 - (1 - \psi(s))^d\big). \tag{9}$$

Combining Equations (7) and (9) gives, for any $s \ge 0$,

$$\mathbb{E}\left[\min_{1 \le i \le d} X_i\right] \ \le\ \mu \ -\ s\big(1 - (1 - \psi(s))^d\big). \tag{10}$$

Step 2. Next, we lower-bound $\psi(s_d)$ via Berry–Esseen and a Mills-ratio bound: let us expand $X_1 = \sum_{j=1}^{T} B_j$ where $B_j \sim \mathrm{Bern}(\xi)$ i.i.d.. Let $Y_j = B_j - \xi$, so that $S_T = \sum_{j=1}^{T} Y_j = X_1 - \mu$.

By the classical Berry–Esseen inequality with the upper bound $C_0 < 0.4690$ for i.i.d. summands with finite third moment (Shevtsova, 2013, 2014), one has

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left[\frac{S_T}{\sigma} \le x\right] - \Phi(x) \right| \ \le\ 0.4690 \cdot \frac{\mathbb{E}\left[|Y_1|^3\right]}{(\mathbb{E}\left[Y_1^2\right])^{3/2}} \cdot \frac{1}{\sqrt{T}}, \tag{11}$$

where Φ is the standard normal CDF. For centered Bernoulli variables, $\mathbb{E}\left[Y_1^2\right] = \xi(1 - \xi)$ and

$$\mathbb{E}\left[|Y_1|^3\right] = \xi(1 - \xi)^3 + (1 - \xi)\xi^3 = \xi(1 - \xi)\big(\xi^2 + (1 - \xi)^2\big) \ \le\ \xi(1 - \xi).$$

Hence $\mathbb{E}\left[|Y_1|^3\right] / (\mathbb{E}\left[Y_1^2\right])^{3/2} \le 1/\sqrt{\xi(1 - \xi)}$, and Equation (11) yields

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left[\frac{S_T}{\sigma} \le x\right] - \Phi(x) \right| \ \le\ \frac{0.4690}{\sqrt{T\xi(1 - \xi)}}. \tag{12}$$

Also notice that

$$\psi(s_d) = \mathbb{P}\left[X_1 \le \mu - s_d\right] = \mathbb{P}\left[\frac{S_T}{\sigma} \le -u_d\right],$$

since $s_d = \sigma u_d$ by definition. By Equation (12) we then have:

$$\psi(s_d) \;\geq\; \Phi(-u_d) \;-\; \frac{0.4690}{\sqrt{T\xi(1-\xi)}}. \tag{13}$$

We next lower-bound $\Phi(-u_d)$. To do this, we use the Mills-ratio lower bound (see Gordon (1941)): for all $u > 0$,

$$\Phi(-u) \;\geq\; \frac{1}{\sqrt{2\pi}} \cdot \frac{u}{1+u^2}\, e^{-u^2/2}. \tag{14}$$

Since $d \geq 2$, we have $u_d^2 = 2\ln d - 3\ln\ln d > 1$, so $u_d \geq 1$. Therefore $1 + u_d^2 \leq 2u_d^2$, hence $\frac{u_d}{1+u_d^2} \geq \frac{1}{2u_d}$, and Equation (14) gives

$$\Phi(-u_d) \;\geq\; \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2u_d}\, e^{-u_d^2/2}. \tag{15}$$

Moreover,

$$e^{-u_d^2/2} = e^{-(2\ln d - 3\ln\ln d)/2} = \frac{(\ln d)^{3/2}}{d},$$

and also $u_d^2 \leq 2\ln d$ implies $u_d \leq \sqrt{2\ln d}$, i.e. $\frac{1}{u_d} \geq \frac{1}{\sqrt{2\ln d}}$. Plugging these into Equation (15),

$$\Phi(-u_d) \;\geq\; \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{2\ln d}} \cdot \frac{(\ln d)^{3/2}}{d} = \frac{\ln d}{4\sqrt{\pi}\, d}.$$

Insert this into Equation (13) and use the assumption in Equation (5) to obtain

$$\psi(s_d) \;\geq\; \frac{\ln d}{4\sqrt{\pi}\, d} \;-\; \frac{\ln d}{8\sqrt{\pi}\, d} \;=\; \frac{\ln d}{8\sqrt{\pi}\, d}. \tag{16}$$

Step 3. We apply Equation (10) with $s = s_d$. Using $(1-x)^d \leq e^{-dx}$ for $x \in [0,1]$, we get

$$1 - (1-\psi(s_d))^d \;\geq\; 1 - e^{-d\psi(s_d)}.$$

With Equation (16), $d\psi(s_d) \geq \frac{\ln d}{8\sqrt{\pi}}$, hence

$$1 - e^{-d\psi(s_d)} \;\geq\; 1 - \exp\left(-\frac{\ln d}{8\sqrt{\pi}}\right) = 1 - d^{-1/(8\sqrt{\pi})}.$$

Substituting into Equation (10) yields

$$\mathbb{E}\left[\min_{1\leq i\leq d} X_i\right] \;\leq\; \mu - s_d\Big(1 - d^{-1/(8\sqrt{\pi})}\Big),$$

which is exactly Equation (6). ■

Appendix B. A Strawman Algorithm for Experts: a Uniform and a Hedge Probe

In this appendix, we focus on the case of experts. We show how a uniform probe together with a Hedge probe (Algorithm 2) can yield non-trivial regret rates in the noiseless case, albeit losing an additional factor d. This algorithm is a simple adaptation of Hedge (Freund and Schapire, 1997) (subsequently denoted as $\textsc{Alg}_0$), which, during probing rounds Q, probes one expert uniformly at random and one expert according to the Hedge distribution p_t at round $t \in Q$. We set the learning rate to $\eta = \max\left(\sqrt{\ln d / T}, k/dT+k\right)$.

Before stating the theorem, we first recall that, in prediction with expert advice, instances are described by linear functions $f_t(x) = f_t^\top x$, for some vector f_t, where $x \in \{e_1, \dots, e_d\}$ is one of the vertices of the $(d-1)$-dimensional simplex Δ^{d-1}, i.e., $K = \{e_1, \dots, e_d\}$: coordinate i of vector f_t therefore represents the loss of expert e_i.

Theorem 13 *Consider the problem of prediction with expert advice with full feedback and k pairwise comparison probes, with domain $K = \{e_1, \dots, e_d\}$. Then, for any sequence of functions $\{f_t\}_{t=1}^T$ in $[0, 1]$, Algorithm 2, denoted as $\textsc{Alg}_k$, guarantees regret*

$$\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg}_k)\right] \leq O\left(\min\left\{\sqrt{T \ln d}, \frac{T d \ln d}{k}\right\}\right).$$

To prove the theorem, we state a useful lemma from Russo et al. (2024), which bounds the regret of the probeless version $\textsc{Alg}_0$ of $\textsc{Alg}_k$; it directly follows from Lemma 2.1 in Russo et al. (2024) shifting losses, denoting $x_t^\star = \arg\min_{x \in \{e_i\}_{i=1}^d} f_t(x)$:

Lemma 14 *For the problem of prediction with expert advice with full feedback and domain $K = \{e_1, \dots, e_d\}$, it holds that the expected regret of $\textsc{Alg}_0$ with $\eta < 1$ is bounded above as*

$$\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg}_0)\right] \leq \frac{\ln d}{\eta(1-\eta)} + \frac{\eta}{1-\eta} \sum_{t=1}^{T} f_t^\top (x^\star - x_t^\star). \tag{17}$$

Algorithm 2: Uniform Probe Hedge

Input: Sequence of gradient vectors f_t, probe budget $k \leq T$
Sample k out of T rounds uniformly at random and denote this random set by Q
Set $\eta = \max\left(\sqrt{\frac{\ln d}{T}}, \frac{k}{dT+k}\right)$ **if** $T \geq 4 \ln d$ **else** $\eta = \frac{k}{dT+k}$
Initialize $w_1(x) = 1$ for all $x \in \{e_1, \dots, e_d\}$
for $t \in \{1, \dots, T\}$ **do**
  Let $W_t = \sum_{x \in \{e_1, \dots, e_d\}} w_t(x)$ and $p_t(x) = \frac{w_t(x)}{W_t}$
  if $t \in Q$ **then**
    Probe $y_1 \sim p_t$ and $y_2 \sim$ Unif$(\{e_1, \dots, e_d\})$
    Play point $y_t^\star = \arg\min\{f_t^\top y_1, f_t^\top y_2\}$
  else
    Select $x \sim p_t$
  Observe $f_t^\top x \ \forall x \in \{e_1, \dots, e_d\}$
  Update $w_{t+1}(x) = w_t(x) \cdot \exp\left(-\eta f_t^\top x\right) \ \forall x \in \{e_1, \dots, e_d\}$

Proof [of Theorem 13] First, we relate the regret of $\textsc{Alg}_k$ to that of $\textsc{Alg}_0$. To this end, observe that, since $\textsc{Alg}_k$ is provided with full feedback, then, the distribution p_t Hedge keeps over experts is not affected by earlier probes or earlier decisions taken by the algorithm. Let z_t denote the point/expert played by the algorithm in (probing or non-probing) round t. During non-probing rounds $t \in \overline{Q} = \{1, \dots, T\} \setminus Q$, the expected loss $\textsc{Alg}_k$ suffers is

$$\mathbb{E}\left[f_t^\top z_t \mid t \in \overline{Q}\right] = \sum_{x \in \{e_1, \dots, e_d\}} p_t(x) \cdot f_t^\top x.$$

Moreover, in rounds $t \in Q$, $\textsc{Alg}_k$ probes the expert of minimum loss with probability $1/d$. In the event that the expert of minimum loss is not probed, we have that, in expectation, the algorithm $\textsc{Alg}_k$ suffers at most the loss $\textsc{Alg}_0$ (Hedge) would have suffered. Hence, the expected instantaneous loss of $\textsc{Alg}_k$ at round $t \in Q$ is

$$\mathbb{E}\left[f_t^\top z_t \mid t \in Q\right] \le \left(1 - \frac{1}{d}\right) \sum_{x \in \{e_1, \dots, e_d\}} p_t(x) \cdot f_t^\top x + \frac{f_t^\top x_t^\star}{d}.$$

Summing over all t, by linearity of expectation, we have that the expected loss of the algorithm is:

$$\begin{aligned}
\mathbb{E}\left[L_T(\textsc{Alg}_k)\right] &= \mathbb{E}\left[\sum_{t=1}^T f_t^\top z_t\right] = \sum_{t=1}^T \mathbb{E}\left[f_t^\top z_t \cdot \mathbb{I}\left\{t \in \overline{Q}\right\}\right] + \sum_{t=1}^T \mathbb{E}\left[f_t^\top z_t \cdot \mathbb{I}\left\{t \in Q\right\}\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[f_t^\top z_t \mid t \in \overline{Q}\right] \cdot \mathbb{P}\left[t \in \overline{Q}\right] + \sum_{t=1}^T \mathbb{E}\left[f_t^\top z_t \mid t \in Q\right] \cdot \mathbb{P}\left[t \in Q\right] \\
&\le \left(1 - \frac{k}{T}\right) \sum_{t=1}^T \sum_{x \in \{e_1, \dots, e_d\}} p_t(x) \cdot f_t^\top x \\
&\qquad + \frac{k}{T} \sum_{t=1}^T \left(\left(1 - \frac{1}{d}\right) \sum_{x \in \{e_1, \dots, e_d\}} p_t(x) \cdot f_t^\top x + \frac{f_t^\top x_t^\star}{d}\right) \\
&= \left(1 - \frac{k}{dT}\right) \mathbb{E}\left[L_T(\textsc{Alg}_0)\right] + \frac{k}{dT} \sum_{t=1}^T f_t^\top x_t^\star.
\end{aligned}$$

Combining the above with Equation (17), we have

$$\begin{aligned}
\mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg}_k)\right] &\le \left(1 - \frac{k}{dT}\right) \mathbb{E}\left[\textsc{Reg}_T(\textsc{Alg}_0)\right] - \frac{k}{dT} \sum_{t=1}^T f_t^\top (x^\star - x_t^\star) \\
&\le \left(1 - \frac{k}{dT}\right) \frac{\ln d}{\eta(1-\eta)} + \left(\frac{\eta}{1-\eta} - \frac{k}{dT}\right) \sum_{t=1}^T f_t^\top (x^\star - x_t^\star) \\
&\le \min\left\{4\sqrt{T \ln d}, \frac{2Td\ln d}{k}\right\},
\end{aligned}$$

by assuming $T \ge 4 \ln d$ and choosing $\eta = \max\left(\sqrt{\frac{\ln d}{T}}, \frac{k}{dT+k}\right)$, since $\sum_{t=1}^T f_t^\top (x^\star - x_t^\star) \le T$. If $T < 4 \ln d$, the regret guarantee is satisfied since $\eta = \frac{k}{dT+k}$ and the regret is bounded by T. ■