ORACLE-EFFICIENT AND PARAMETER-FREE AGNOSTIC SMOOTHED ONLINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical online learning deals with scenarios where data arrive sequentially and decisions must be made in an online fashion, often under adversarial conditions, and is attractive due to its robustness relative to more classical learning that relies on assuming independence. Unfortunately, due to the minimal assumptions it makes about the data-generating process, it often suffers from poor statistical and computational efficiency. Recently, smoothed online learning has emerged as a promising framework that interpolates between the fully adversarial and fully stochastic settings, making the assumption that the conditional law of each covariate has a density at most $1/\sigma$ with respect to some fixed base measure μ. Prior work has investigated this setting either under the assumption that μ is known to the learner or under the assumption that the labels are perfectly predicted by a fixed hypothesis; in both cases, smoothed online learning admits statistically and computationally efficient algorithms. Recent work has demonstrated that even absent these conditions, smoothed online learning with unknown μ can be *statistically* efficient, but the algorithm studied was computationally intractable, leaving open the key question of whether it is possible to achieve both statistical and computational efficiency simultaneously when μ is unknown. In this work we resolve this question by giving the first oracle-efficient algorithm that achieves sublinear regret in the agnostic setting without knowledge of μ. We introduce a new, parameter-free algorithm based on Gaussian Follow-The-Perturbed-Leader that does not require prior knowledge of the smoothing parameter σ and is capable of achieving regret $\widetilde{O}(d\sqrt{T/\sigma})$ for binary classes of VC dimension d with a single call to an optimization oracle per round. En route to establishing the regret bound, we introduce several new techniques that may be of independent interest.

1 INTRODUCTION

Learning from independent data without a priori knowledge of the generating process is the archetypal paradigm in machine learning (Valiant, 1984): a learner that observes a training sample drawn from an *unknown* distribution can generalize to new examples from the same distribution. Its main appeal is that it removes the need to model nature's distribution, which may be highly complex and need not belong to a known parametric family. Beyond the statistical benefits of assuming independence, there are computational advantages: near-optimal learning can be achieved with a simple algorithm, empirical risk minimization (ERM), which simply optimizes its predictions on historical data (Vapnik & Chervonenkis, 1971; Blumer et al., 1989).

In many modern applications, however, the independence assumption is difficult to justify, let alone verify. Data often arrive sequentially, and the data-generating process may even react to the learner's own predictions. Many modern data modalities, such as text, video, and state trajectories of physical systems (e.g., in robotics and control), exhibit strong temporal dependencies. In interactive decision making, such as bandits and reinforcement learning, the learner's own past decisions determine which data it collects (Lattimore & Szepesvári, 2020; Foster et al., 2021). Financial markets are another example: prices are strongly dependent over time, and the predictions of trading systems influence the actions of market participants, which in turn shape the data those systems observe. Deployed recommendation systems exhibit similar feedback loops: the system's recommendations influence users' behavior, which in turn shapes the data the system collects (Chaney et al., 2018; Jiang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2019). In such settings, guarantees premised on independence need not hold, and algorithms developed assuming independence may not learn, which motivates a theory of learning that makes minimal assumptions on how the data are generated.

Online learning (Littlestone, 1988) provides a promise of such a theory: it dispenses with the independence assumption altogether and permits learning even from dependent, possibly adversarial data. Formally, online learning is a repeated game between a learner and nature: in each round t, the learner observes a covariate X_t, predicts a label $\hat{Y}_t$, observes the true label Y_t, and suffers loss $\ell(\hat{Y}_t, Y_t)$. The learner's goal is to minimize its *regret* with respect to a fixed comparator class $\mathcal{F}$:

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(X_t), Y_t).$$

Importantly, the data-generating process may be arbitrary; in particular, it may adapt to the learner's past actions. This generality, however, comes at both a statistical and a computational price.

Learnability in online binary classification is characterized by the Littlestone dimension, which may be infinite even for simple classes such as linear threshold functions (Littlestone, 1988; Ben-David et al., 2009). The situation is further exacerbated by computational barriers. A natural computational framework is the *ERM oracle model*, designed to capture the efficacy of optimization in modern practice. When queried, an ERM oracle for $\mathcal{F}$ returns a predictor $f \in \mathcal{F}$ that minimizes the empirical loss on the given dataset. This oracle forms the basic computational primitive of statistical learning and is a practically meaningful abstraction, since ERM is routinely solved, at least approximately, with off-the-shelf tools such as stochastic gradient descent. Ideally, then, one would like online learning algorithms that are *oracle-efficient*, i.e., that run efficiently when the optimization oracle is presumed to be efficient and of bounded cost. Without further assumptions in the sequential setting, however, Hazan & Koren (2016) showed that this goal is unattainable: even for a finite comparator class, access to an ERM oracle does not make online learning computationally tractable. In other words, in the absence of independence, generalization and learning do not reduce to optimization.

In order to circumvent these barriers, there has been substantial effort in understanding smoothed online learning, which provides a way to interpolate between the stochastic and adversarial regimes (Rakhlin et al., 2011; Haghtalab et al., 2020; Block et al., 2022; 2024b; Blanchard, 2025; Blanchard et al., 2026). Intuitively, smoothness prevents the adversary from concentrating probability mass on small sets, while still permitting strong temporal dependencies. Formally, letting p_t denote the law of X_t conditional on the history, we require that there exist a single probability measure μ, referred to as the *base measure*, such that in every round t,

$$\left\| \frac{\mathrm{d}p_t}{\mathrm{d}\mu} \right\|_\infty \leq \frac{1}{\sigma},$$

where the smoothness parameter $\sigma \in (0, 1]$ controls the power of the adversary. At $\sigma = 1$, the covariates are drawn independently from μ in every round; as $\sigma \downarrow 0$, the constraint becomes increasingly permissive. This relaxation removes both statistical and computational barriers: an extensive literature shows that VC classes are learnable in the smoothed setting for every $\sigma > 0$ and, moreover, efficiently so in the ERM-oracle model (Haghtalab et al., 2020; Block et al., 2022; Haghtalab et al., 2022). Further, from a statistical perspective, Blanchard et al. (2026) showed that a technical generalization of the smoothed model essentially captures learnability for general sequences of distributions.

However, in contrast to the classical theory of learning, most efficient algorithms in this setting assume that μ *is known to the learner*. This leaves a gap between online and statistical learning. We have lost precisely what makes the latter appealing, the lack of the need to explicitly model nature. Conceptually, knowledge of μ amounts to an additional modeling assumption on nature; indeed, it is unclear how one would justify a particular choice of μ in practice. Technically, however, knowledge of μ simplifies the algorithmic landscape, since it allows the learner to generate fresh reference samples from the base measure; intuitively, these samples provide *coverage* over the space of covariates it may receive in the future. When μ is unknown, the learner sees only the adaptively chosen covariates $X_1, \ldots, X_t$, none of which need be distributed according to μ.

Two notable exceptions study the setting of an unknown base measure. Block et al. (2024b) showed that ERM itself achieves sublinear regret in the *realizable* setting, where the data are labeled by some

fixed function $f^\star \in \mathcal{F}$, i.e., $Y_t = f^\star(X_t)$.[1] More recently, building on the techniques of Block et al. (2024b), Blanchard (2025) gave an agnostic algorithm with regret $\widetilde{O}(\sqrt{dT/\sigma})$ for binary classes of VC dimension d, which is optimal up to logarithmic factors. Their algorithm, however, constructs recursive covers whose size can be exponential in the complexity of the class, and is therefore not oracle-efficient. Thus, while the statistical picture is well understood, the computational question remains open: *Is oracle-efficient agnostic learning possible when the base measure is unknown?*

Main results. We answer this question in the affirmative, giving the first oracle-efficient agnostic learnability guarantee without knowledge of the base measure. Surprisingly, this guarantee is achieved by a natural algorithm: follow-the-perturbed-leader (FTPL) with Gaussian perturbations. It is well known since Hannan (1957) and Cover (1967) that ERM can fail on agnostic data even when the covariate domain has size 1 (which forces $\sigma = 1$). Intuitively, this failure stems from a lack of *stability*: ERM is sensitive to individual data points, and its solution can change drastically in response to a single label flip. A standard way to stabilize ERM is to regularize it, either with an explicit potential, as in follow-the-regularized-leader (FTRL) (Shalev-Shwartz & Singer, 2007; Abernethy et al., 2008; McMahan, 2017), or with a random perturbation of the objective, as in FTPL (Hannan, 1957; Kalai & Vempala, 2005; Hutter & Poland, 2005; Devroye et al., 2013). Motivated by the algorithm in Block et al. (2022), we show that natural Gaussian perturbations suffice to achieve sublinear regret. In particular, for binary classification, we show that FTPL with Gaussian perturbations attains

$$\mathbb{E}\,\mathrm{Reg}_T \le \widetilde{O}\left(d\sqrt{\frac{T}{\sigma}}\right) \tag{1}$$

simultaneously for all μ and σ. This guarantee has optimal dependence on T and σ, and is statistically suboptimal only by a factor of $\sqrt{d}$ (Blanchard, 2025). We now define our Gaussian FTPL learner formally. Let $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a comparator class and let $\ell(\hat{y}, y) = -y\hat{y}$ be the loss. In round t, the algorithm samples a fresh Gaussian vector $G^{(t)} = (G_1^{(t)}, \dots, G_{t-1}^{(t)}) \sim \mathcal{N}(0, I_{t-1})$ and computes

$$\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \left[\ell(f(X_s), Y_s) + G_s^{(t)} f(X_s)\right]. \tag{2}$$

We use a weighted ERM oracle (Section 2 gives the precise model). This algorithm is completely parameter-free: it requires no knowledge of the base measure μ, the smoothness parameter σ, or the horizon T. It thus addresses precisely the limitation of prior work discussed above, achieving oracle-efficient learning adaptively for every μ and σ.

Our algorithm and guarantee extend to classes $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ and convex 1-Lipschitz losses by adding an averaging step in each round (Section 3). In terms of the fat-shattering dimension $\mathrm{fat}_\varepsilon(\mathcal{F})$ (Section 2), the resulting bound is $\widetilde{O}(d\sqrt{T/\sigma})$ for parametric classes with $\mathrm{fat}_\varepsilon(\mathcal{F}) \lesssim d\log(1/\varepsilon)$, and $\widetilde{O}_\alpha(T^{(2\alpha+1)/(2\alpha+2)}/\sqrt{\sigma})$ for nonparametric classes with $\mathrm{fat}_\varepsilon(\mathcal{F}) \lesssim \varepsilon^{-\alpha}$, matching the leading exponents of T and $1/\sigma$ in Blanchard (2025, Theorem 4).

Technical overview. We sketch the analysis for the binary case, which already contains the main ideas for the general real-valued setting. A central tool in the analysis of FTPL algorithms is stability (Kalai & Vempala, 2005): intuitively, if the perturbations have controlled magnitude and the algorithm's predictions are stable, a regret bound follows. However, we find it difficult to analyze the stability of Eq. (2) directly; instead, we are able to show that the predictions of Eq. (2) are close to predictions of an auxiliary algorithm that is itself stable.

In more detail, our proof proceeds in two steps. We first compare the predictions of our FTPL forecaster to those of an auxiliary algorithm: exponential weights over a cover $\mathcal{G}$ of $\mathcal{F}$ (Vovk, 1990; Littlestone & Warmuth, 1994; Freund & Schapire, 1997). The auxiliary algorithm need not be efficient, as it is used only in the analysis. Formally, its prediction $\tilde{f}_t \in \mathrm{conv}(\mathcal{F})$ is defined as follows:

$$\tilde{f}_t := \sum_{f \in \mathcal{G}} w_t(f \mid G^{(t)}) f, \quad \text{with} \quad w_t(f \mid G^{(t)}) \propto \exp\left(-\sum_{s<t}[\ell(f(X_s), Y_s) + G_s^{(t)} f(X_s)]\right). \tag{3}$$

[1] More generally, their result applies to well-specified regression.

Our first technical contribution reduces the regret of $\hat{f}_t$ to that of $\tilde{f}_t$. Specifically, letting $\mathrm{Reg}^{\mathrm{EW}}$ denote the regret of the learner in Eq. (3), we show that, as long as $\mathcal{G}$ approximates $\mathcal{F}$ sufficiently well (see Section 4.2 for details),

$$\mathbb{E}\,\mathrm{Reg}_T - \mathbb{E}\,\mathrm{Reg}_T^{\mathrm{EW}} \le \widetilde{O}\left(d\sqrt{\frac{T}{\sigma}}\right). \tag{4}$$

The proof of this reduction has two steps. First, we compare the predictions of $\hat{f}_t$ and $\tilde{f}_t$ on the historical covariates $X_{1:t-1}$, showing that for every $1 \le t \le T$,

$$\mathbb{E}\sum_{s=1}^{t-1}(\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_s) - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_s))^2 \lesssim d^2\log^2(2T/\sigma). \tag{5}$$

This follows by expressing the prediction vectors of $\hat{f}_t$ and $\tilde{f}_t$ as gradients of certain potentials and using Stein's lemma (Stein, 1981; Abernethy et al., 2014) to relate differences in predictions to differences in potentials. Next, we use an improved version of the *surprise lemma* of Block et al. (2024b) to translate Eq. (5) into closeness of predictions on the current covariate X_t. Specifically, adapting arguments from Xie et al. (2022); Block et al. (2024b) to convex hulls of VC classes, we prove that for any sequence of predictable functions $h_t \in \mathrm{conv}(\pm\mathcal{F})$,

$$\mathbb{E}\sum_{t=1}^{T}|h_t(X_t)| \lesssim \sqrt{\frac{\log(2T)}{\sigma}\,\mathbb{E}\sum_{t=1}^{T}\sum_{s<t}h_t(X_s)^2} + \sqrt{\frac{dT}{\sigma}}\log^2(2T/\sigma).$$

Intuitively, this shows that the value of a predictable test function h_t at a new covariate X_t is controlled by its cumulative second moments on past covariates; informally, values of $|h_t(X_t)|$ that are surprisingly large relative to the history cannot occur too often. Applied to differences $h_t = (\mathbb{E}_{G^{(t)}}\,\hat{f}_t - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t)/2$ and combined with Eq. (5), this shows that predictions $\hat{f}_t(X_t)$ and $\tilde{f}_t(X_t)$ are close on average, which proves Eq. (4) since the loss is 1-Lipschitz.

Thus, we have reduced our problem to bounding the regret of the exponential weights algorithm in Eq. (3). Here, a central difficulty is that it uses a constant learning rate, whereas the usual worst-case bounds require $\eta \propto 1/\sqrt{T}$ and are vacuous for $\eta \propto 1$ (Cesa-Bianchi & Lugosi, 2006). It turns out, however, that the Gaussian perturbations themselves provide the necessary smoothing of the softmax objective. In particular, we show that the predictions of exponential weights with Gaussian perturbations are *stable*, in the sense that they have low cumulative variance:

$$\sum_{t=1}^{T}\mathbb{E}\,\mathrm{Var}_{f\sim w_t(\cdot\mid G^{(t)})}(f(X_t)) \le \widetilde{O}\left(\sqrt{\frac{T(\log|\mathcal{G}|+d)}{\sigma}}\right).$$

Combining this with the standard variance-based analysis of exponential weights (Cesa-Bianchi et al., 2007; de Rooij et al., 2014) yields a corresponding regret bound:

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{EW}} \lesssim d\log(2T/\sigma) + \sum_{t=1}^{T}\mathbb{E}\,\mathrm{Var}_{f\sim w_t(\cdot\mid G^{(t)})}(f(X_t)) \le \widetilde{O}\left(\sqrt{\frac{T(\log|\mathcal{G}|+d)}{\sigma}}\right).$$

By controlling the size of the cover (Proposition 4.2) and using Eq. (4), we obtain Eq. (1).

Conventions and notation. We write $\log$ for the binary logarithm. For an integer n, we let $[n]$ denote the set $\{1,\dots,n\}$. We use standard non-asymptotic big-O notation: we write $f \le O(g)$ or $f \lesssim g$ to mean that $f \le Cg$ for a sufficiently large universal constant C. We use $\widetilde{O}$ to suppress logarithmic factors in d, T, and $1/\sigma$. We write $Z_{a:b}$ for the sequence $(Z_a, Z_{a+1}, \dots, Z_b)$. For a class $\mathcal{H}$ of functions, we write $\mathrm{conv}(\mathcal{H})$ to denote its convex hull and let $\mathrm{conv}(\pm\mathcal{H}) := \mathrm{conv}(\mathcal{H}\cup -\mathcal{H})$.

2 Preliminaries and Related Work

Online learning and smoothness. We reserve $\mathcal{X}$ to refer to the space of covariates, and d to refer to the VC dimension of the comparator class $\mathcal{F}$. For convenience, we assume binary labels are in

$\{\pm 1\}$. Consider a sequence of covariates X_t and labels Y_t, and let $\mathcal{H}_t = \sigma((X_s, Y_s, \hat{Y}_s)_{s\in[t]})$ be the σ-algebra induced by the learner's observations up to round t, and write $p_t = \mathrm{Law}(X_t \mid \mathcal{H}_{t-1})$. For a fixed probability measure μ and $\sigma \in (0, 1]$, the covariate process is σ-smooth with respect to μ if, almost surely, for all t, p_t is absolutely continuous w.r.t. μ and

$$\left\| \frac{\mathrm{d}p_t}{\mathrm{d}\mu} \right\|_\infty \leq \frac{1}{\sigma}.$$

Complexity measures. Let $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$. We say that $\mathcal{F}$ *shatters* a set $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if for every $\epsilon_{1:n} \in \{\pm 1\}^n$ there exists $f \in \mathcal{F}$ with $f(x_i) = \epsilon_i$ for all $i \in [n]$. The *VC dimension* of $\mathcal{F}$ is the largest integer d such that $\mathcal{F}$ shatters some set of size d (or ∞ if no such largest d exists) (Vapnik & Chervonenkis, 1971).

For real-valued classes $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ and a scale $\varepsilon > 0$, we say that $\mathcal{F}$ *ε-shatters* a set $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if there exist witnesses $s_{1:n} \in \mathbb{R}^n$ such that for every $\epsilon_{1:n} \in \{\pm 1\}^n$ there exists $f \in \mathcal{F}$ with

$$\epsilon_i \left(f(x_i) - s_i\right) \geq \varepsilon \qquad \text{for all } i \in [n].$$

The *fat-shattering dimension* of $\mathcal{F}$ at scale ε, denoted $\mathrm{fat}_\varepsilon(\mathcal{F})$, is the largest d such that $\mathcal{F}$ ε-shatters some set of size d (Kearns & Schapire, 1994; Alon et al., 1997). The map $\varepsilon \mapsto \mathrm{fat}_\varepsilon(\mathcal{F})$ is non-increasing, and for binary classes $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$, $\mathrm{fat}_\varepsilon(\mathcal{F})$ is equal to the VC dimension of the class for all $\varepsilon \in (0, 1]$.

ERM oracle model. Following Hazan & Koren (2016) and the literature on oracle-efficient smoothed online learning (Block et al., 2022; Haghtalab et al., 2022), we assume access to a weighted agnostic ERM oracle. In particular, for any n covariates $X_{1:n}$, a sequence $\ell_i \colon [-1, 1] \to [-1, 1]$ of convex functions, and weights $w_i \geq 0$, we assume we can compute a solution to the following optimization problem:

$$\arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} w_i \ell_i(f(X_i)).$$

While the above definition permits arbitrary sequences of convex loss functions, we only require an oracle that can handle linear losses $\ell_i(y) = \lambda y$ or label loss $\ell_i(\cdot) = \ell(\cdot, y)$ for some label y; weights in our analysis satisfy $\|w\|_\infty \lesssim \sqrt{\log(T)}$ w.h.p.

Related work. Due to its robustness and fundamental nature, online learning has been extensively studied over the past decades with a relatively complete understanding of its theoretical foundations (Freund & Schapire, 1997; Ben-David et al., 2009; Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz, 2012; Rakhlin et al., 2011; Block et al., 2021) and computational properties (Hazan & Koren, 2016; Kalai & Vempala, 2005; Abernethy et al., 2014; Rakhlin et al., 2012). On the other hand, because of the strong statistical and computational lower bounds, more recently a number of works have explored online learning under various relaxations and assumptions to achieve tractable algorithms and improved performance guarantees (Rakhlin et al., 2011). Of particular relevance are prior works (Haghtalab et al., 2020; Block et al., 2022; Haghtalab et al., 2022; Block & Polyanskiy, 2023; Block et al., 2024a), which focus on the statistical and computational advantages of smoothed online learning under the assumption that the base measure is *known*, as well as applications to other domains like game theory (Daskalakis et al., 2024). Other work sought to extend these results, either by taking advantage of additional structure for specific base measures (Block et al., 2023b;a; Block & Simchowitz, 2022), or considering more general loss functions (Bhatt et al., 2023).

In all of the above work relevant to smoothed online learning, some knowledge of the base measure was required, which limits the applicability of the results. A separate line of work has sought to develop methods that do not rely on knowing the base measure, thereby broadening the scope and applicability of smoothed online learning techniques (Wu et al., 2023; 2024), although the former work focuses on oracle-inefficient algorithms under substantially stronger assumptions whereas the latter provides oracle-efficient solutions assuming the features are independent. In Block et al. (2024b), the authors demonstrate that it is possible to achieve strong performance guarantees even without knowledge of the base measure using just empirical risk minimization, but they require the labels to be *well-specified* in the sense that they can be perfectly predicted by the features. On the other end of the spectrum, Blanchard (2025) consider the agnostic setting where no such well-specification

assumption is made, but do not consider oracle-efficient algorithms. Unlike these prior works, ours is the first to propose an oracle-efficient algorithm (similar in spirit to one in Block et al. (2022)) that is capable of achieving sublinear regret under an arbitrary unknown base measure; moreover, our algorithm's regret scales optimally in both time and smoothness, losing only a $\sqrt{d}$ factor compared to the optimal rate.

3 Main Results

The main results of our paper are two oracle-efficient regret bounds, for the binary and the real-valued loss cases. We begin with the binary case. Recall that Gaussian FTPL, in each round t, computes:

$$\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \left[\ell(f(X_s), Y_s) + G_s^{(t)} f(X_s) \right], \tag{6}$$

and predicts $\hat{Y}_t = \hat{f}_t(X_t)$. Note that the above can be implemented using one call to the weighted ERM oracle. For this algorithm, we obtain the following regret guarantee.

Theorem 3.1. *Consider a binary class $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ of VC dimension $d \geq 1$, and let $\ell(\hat{y}, y) = -\hat{y}y$ be the binary loss. For every horizon $T \geq 1$ and every covariate process that is σ-smooth with respect to a fixed μ, the learner in (6) satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T \lesssim d \log^2(2T/\sigma) \sqrt{\frac{T}{\sigma}}. \tag{7}$$

A complete proof of the above statement is given in Section B. Note that the regret bound above is obtained in a completely parameter-free manner. In particular, the learner in (6) requires no knowledge of the base measure μ, the smoothness parameter σ, or the horizon T.

Extending the above guarantee to convex Lipschitz losses requires an additional averaging step, which, intuitively, helps reduce the variance of the predictor. Formally, in round t, our learner resamples the noise $G^{(t)}$ independently $n_t := t$ times, and computes (conditionally) independent solutions $\hat{f}_t^{(i)}$, $i \in [n_t]$ to the optimization problem in Eq. (6). The prediction is then aggregated in the following way:

$$\hat{Y}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{f}_t^{(i)}(X_t). \tag{8}$$

The above can be recognized as a Monte-Carlo estimate of the expectation $\mathbb{E}_{G^{(t)}}\, \hat{f}_t(X_t)$. By Jensen's inequality $\ell(\mathbb{E}_{G^{(t)}}\, \hat{f}_t(X_t), Y_t) \leq \mathbb{E}_{G^{(t)}}\, \ell(\hat{f}_t(X_t), Y_t)$, so, up to the Monte Carlo error, the averaged prediction incurs no more loss than a single Gaussian FTPL prediction in conditional expectation.

Theorem 3.2. *Consider a class $\mathcal{F}\colon \mathcal{X} \to [-1, 1]$ and let $\ell(\hat{y}, y)$ be convex and 1-Lipschitz in its first argument, and suppose the covariate process is σ-smooth with respect to a fixed μ. Then, the learner in (8) satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T \lesssim \inf_{0 < \varepsilon \leq 1/16} \left\{ \left(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}) + \sqrt{T}\varepsilon\right) \sqrt{\frac{T}{\sigma}} \log^{7/2}\left(2T/\sigma\varepsilon\right) \right\}, \tag{9}$$

where c is a universal constant.

A complete proof of the above statement is given in Section C. Theorem 3.2 gives explicit rates for two important complexity profiles. For parametric classes satisfying $\mathrm{fat}_\varepsilon(\mathcal{F}) \lesssim d \log(1/\varepsilon)$ at sufficiently small scales,

$$\mathbb{E}\,\mathrm{Reg}_T \leq \widetilde{O}\left(d\sqrt{\frac{T}{\sigma}} \right).$$

As in binary classification, the dependence on d is suboptimal by a factor of $\sqrt{d}$, up to logarithmic factors (Blanchard, 2025). For nonparametric classes satisfying $\mathrm{fat}_\varepsilon(\mathcal{F}) \lesssim \varepsilon^{-\alpha}$, $\alpha > 0$, choosing $\varepsilon \asymp_\alpha T^{-1/(2\alpha+2)}$ gives

$$\mathbb{E}\,\mathrm{Reg}_T \leq \widetilde{O}_\alpha\left(\frac{T^{(2\alpha+1)/(2\alpha+2)}}{\sqrt{\sigma}} \right).$$

For fixed σ, this is sublinear for every $\alpha > 0$. It matches the leading horizon and smoothness exponents in Blanchard (2025, Theorem 4). That regression guarantee places the comparator infimum outside the expectation; here it remains inside the expectation.

4 Analysis Techniques

We give an overview of the proof of Theorem 3.1 for the binary case, since it is simpler and yet contains most important ideas needed for the general case. The complete proofs for the binary case are deferred to Section B, and the proofs for the real-valued case are contained in Appendix C. The proof has three ingredients: a sharper surprise lemma (Section 4.1), a reduction to exponential weights (Section 4.2), and a regret bound for Gaussian-perturbed exponential weights (Section 4.3).

4.1 Surprise Lemma

For smoothed online learning algorithms that aim to extrapolate behavior over the past into the future, surprise lemmas have proven useful. Intuitively, for an arbitrary sequence of (predictable) test functions $\{h_t\}$, surprise lemmas quantify how often the value $h_t(X_t)$ can be significantly larger than the historical values $h_t(X_1), \dots, h_t(X_{t-1})$. Formally, we prove the following version of the surprise lemma, which controls expected absolute values of test functions as a function of their historical second moments.

Lemma 4.1. *Let $\mathcal{F}$ be a class of VC dimension d and let $\{h_t\}_{t\in[T]}$ be a predictable sequence of functions with $h_t \in \mathrm{conv}(\pm\mathcal{F})$. Then*

$$\mathbb{E}\sum_{t=1}^{T}|h_t(X_t)| \lesssim \sqrt{\frac{\log(2T)}{\sigma}\,\mathbb{E}\sum_{t=1}^{T}\sum_{s<t}h_t(X_s)^2} + \sqrt{\frac{dT}{\sigma}}\log^2\left(\frac{2T}{\sigma}\right). \tag{10}$$

The above is proven in two steps. We first use a result from Xie et al. (2022, Section 3.2, equations (3)–(5)), which establishes an analogous estimate where the second moments are evaluated on a *tangent* sequence $\bar{X}_{1:T}$ as opposed to realized covariates $X_{1:T}$. A (decoupled) tangent sequence $\bar{X}_{1:T}$ is defined by drawing independent samples $\bar{X}_t \mid \mathcal{H}_{t-1} \sim p_t$. Compared to the analogous statement in Block et al. (2024b, Lemma 12), which involves first moments of the test functions, keeping second moments allows for a sharper transfer from past to current covariates. The next step is to extend the norm comparison inequality of (Block et al., 2024b, Theorem 10) to convex hulls of VC classes, for which we utilize Maurey's empirical method (Pisier, 1981). It is important to allow $h_t \in \mathrm{conv}(\pm\mathcal{F})$ as we wish to apply Lemma 4.1 to functions of predictions of exponential weights, which live in the convex hull of the class. The full proof is in Appendix B.1.3.

4.2 Reduction to exponential weights

In this section, we show that predictions of our Gaussian FTPL algorithm are close to predictions of an auxiliary exponential weights algorithm that is run over the cover $\mathcal{G}$ of $\mathcal{F}$. Prior work establishes existence of covers of controlled size that approximate $\mathcal{F}$ well along a draw of a smoothed covariate trajectory, using the coupling lemma (Haghtalab et al., 2024).

Proposition 4.2. *Let $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ be a binary class with VC dimension d. For every probability measure μ, horizon $T \geq 1$, and $\sigma \in (0, 1]$, there exists a cover $\mathcal{G} \subseteq \mathcal{F}$ such that $\log|\mathcal{G}| \lesssim d\log(2T/\sigma)$ and, for every covariate process that is σ-smooth with respect to μ,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}}\sum_{s\leq T}\mathbb{1}\{f(X_s)\neq g(X_s)\} > Cd\log(2T/\sigma)\right) \leq \frac{1}{T},$$

where $C > 0$ is a universal constant.

Now, the auxiliary exponential weights learner is defined as in Eq. (3) using the cover $\mathcal{G}$ from Proposition 4.2. We show that predictions of such an algorithm are close on average to the predictions of our Gaussian FTPL learner Eq. (6).

Lemma 4.3 (Reduction to exponential weights)**.** *Consider a binary function class $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ of VC dimension d. For $\hat{f}_t$ and $\tilde{f}_t$ as defined in Eqs. (3) and (6), respectively, we have the following bound:*

$$\mathbb{E} \sum_{t=1}^{T} |\, \mathbb{E}_{G^{(t)}}\, \hat{f}_t(X_t) - \mathbb{E}_{G^{(t)}}\, \tilde{f}_t(X_t)| \lesssim d \log^2(2T/\sigma) \sqrt{\frac{T}{\sigma}}. \tag{11}$$

Since binary $\ell(\hat{y}, y) = -\hat{y}y$ is Lipschitz, the above immediately implies

$$\mathbb{E}\, \mathrm{Reg}_T - \mathbb{E}\, \mathrm{Reg}_T^{\mathrm{EW}} \lesssim d \log^2(2T/\sigma) \sqrt{\frac{T}{\sigma}}.$$

We prove Eq. (11) in two steps. First, we show that $\hat{f}_t$ and $\tilde{f}_t$ agree on *past covariates*. Then, using predictability of $\hat{f}_t$ and $\tilde{f}_t$, we leverage the surprise lemma (Lemma 4.1) to transfer that guarantee into closeness on current covariates. In particular, we obtain the following estimate.

Lemma 4.4. *For every $t \le T$,*

$$\mathbb{E} \sum_{s<t} \left(\mathbb{E}_{G^{(t)}}\, \hat{f}_t(X_s) - \mathbb{E}_{G^{(t)}}\, \tilde{f}_t(X_s) \right)^2 \lesssim d^2 \log^2(2T/\sigma). \tag{12}$$

It shows that the Gaussian-averaged predictions $\mathbb{E}_{G^{(t)}}\, \hat{f}_t$ and $\mathbb{E}_{G^{(t)}}\, \tilde{f}_t$ closely agree on the past data points in expected squared norm.

Proof sketch of Lemma 4.4. The case $t = 1$ is immediate. For $t \ge 2$, fix the history $(X_{1:t-1}, Y_{1:t-1})$, write $G = G^{(t)}$, and introduce two potentials:

$$\Phi_t(G) = \max_{f \in \mathcal{F}} \sum_{s<t} (Y_s - G_s) f(X_s),$$

$$\widetilde{\Phi}_t(G) = \ln \sum_{f \in \mathcal{G}} \exp\left(\sum_{s<t} (Y_s - G_s) f(X_s) \right) = \max_{w \in \Delta(\mathcal{G})} \left[\mathbb{E}_{f \sim w} \sum_{s<t} (Y_s - G_s) f(X_s) + H(w) \right],$$

where $H(w) = \sum_{f \in \mathcal{G}} w(f) \ln(1/w(f))$ denotes entropy. Then, almost surely,[2] we have $\hat{f}_t(X_s) = -\partial_{G_s} \Phi_t(G)$ and $\tilde{f}_t(X_s) = -\partial_{G_s} \widetilde{\Phi}_t(G)$ for every $s < t$. These identities follow from Danskin's theorem for the finite maximum (Bertsekas, 2016, Appendix B), and using the fact that w_t is the maximizer in the variational formula for $\tilde{\Phi}_t$; in the context of online learning, this idea also appears in Abernethy et al. (2014). Thus,

$$\begin{aligned}
\sum_{s<t} (\mathbb{E}_G\, \hat{f}_t(X_s) - \mathbb{E}_G\, \tilde{f}_t(X_s))^2 &= \left\| \mathbb{E}_G\, \nabla_G (\Phi_t(G) - \widetilde{\Phi}_t(G)) \right\|_2^2 \\
&\overset{(a)}{=} \left\| \mathbb{E}_G[G(\Phi_t(G) - \widetilde{\Phi}_t(G))] \right\|_2^2 \\
&= \left\| \mathbb{E}_G \left[G \left(\Phi_t(G) - \widetilde{\Phi}_t(G) - \mathbb{E}_G[\Phi_t(G) - \widetilde{\Phi}_t(G)] \right) \right] \right\|_2^2 \\
&= \sup_{\|u\|_2 = 1} \mathrm{Cov}_G(u^\top G, \Phi_t - \widetilde{\Phi}_t)^2 \\
&\overset{(b)}{\le} \sup_{\|u\|_2 = 1} \mathbb{E}_G[(u^\top G)^2]\, \mathrm{Var}_G(\Phi_t - \widetilde{\Phi}_t) \\
&= \mathrm{Var}_G(\Phi_t - \widetilde{\Phi}_t),
\end{aligned}$$

where (a) uses Stein's lemma (Lemma A.3) and (b) uses the Cauchy–Schwarz inequality. It remains to control $\Phi_t - \widetilde{\Phi}_t$. The two potentials differ in that (1) Φ_t maximizes over $\mathcal{F}$ rather than $\mathcal{G}$, which Proposition 4.2 controls, and (2) $\widetilde{\Phi}_t$ adds an entropy term, which costs at most $\ln |\mathcal{G}|$. Combining these, we obtain:

$$\mathrm{Var}_G(\Phi_t - \widetilde{\Phi}_t) \le \mathbb{E}_G\, |\widetilde{\Phi}_t - \Phi_t|^2 \lesssim d^2 \log^2(2T/\sigma) + \ln^2 |\mathcal{G}| \lesssim d^2 \log^2(2T/\sigma),$$

as desired. □

[2] In particular, it suffices that any two maximizers f and f' of either potential give the same predictions on $X_{1:t-1}$, which is true almost surely by absolute continuity of the Gaussians.

An application of the surprise lemma (Lemma 4.1) to Lemma 4.4 finishes the proof of Lemma 4.3. The full proof is in Appendix B.3.

4.3 Regret of Gaussian-perturbed exponential weights

It remains to bound the regret of the exponential weights algorithm. In particular, we show the following upper bound.

Lemma 4.5. *The regret of the exponential weights algorithm in (3) satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{EW}} \lesssim \sqrt{\frac{dT}{\sigma}}\log^2(2T/\sigma). \tag{13}$$

The proof has three steps. First, we adapt the variance-based analysis of exponential weights (Cesa-Bianchi et al., 2007, Lemmas 3 and 4), using the Bernstein bound of de Rooij et al. (2014, Lemma 4) to handle the Gaussian-perturbed losses. This results in the following claim.

Proposition 4.6. *Let $\mathcal{G} \subseteq \mathcal{F}$ be a nonempty finite class, and let $\ell(\hat{y}, y) = -y\hat{y}$ be the binary loss. For every data sequence, the predictor $\tilde{f}_t$ defined in (3) satisfies the following inequality,*

$$\sum_{t=1}^{T} \mathbb{E}_{G^{(t)}}\,\ell(\tilde{f}_t(X_t), Y_t) - \min_{f\in\mathcal{G}} \sum_{t=1}^{T} \ell(f(X_t), Y_t) \lesssim \log|\mathcal{G}| + \sum_{t=1}^{T} \mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot\,|\,G^{(t)})}(f(X_t)). \tag{14}$$

By Proposition 4.2, the comparator losses over the entire class $\mathcal{F}$ and over the cover $\mathcal{G}$ are close, so it remains to bound the cumulative variance on the right-hand side of (14). Second, we show that the Gaussian perturbations alone keep these variances small on the covariates from the past, even though the learning rate is constant.

Lemma 4.7. *Let $\mathcal{G}$ be a nonempty finite class of functions $\mathcal{X} \to [-1, 1]$ and let ℓ be any loss. For every $t \le T$ and every fixed history $(X_s, Y_s)_{s<t}$, the weights in (3) satisfy*

$$\sum_{s<t} \left(\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot\,|\,G^{(t)})}(f(X_s))\right)^2 \le 2\log|\mathcal{G}|. \tag{15}$$

The proof, given in Appendix B.4, combines Stein's lemma (Lemma A.3) with a Gaussian maximal inequality. Third, we transfer this bound to the current covariates. For binary predictions, the variance can be written as

$$\frac{1}{2}\,\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot\,|\,G^{(t)})}(f(x)) = \mathbb{E}_{G^{(t)}} \sum_{f,g\in\mathcal{G}} w_t(f \mid G^{(t)}) w_t(g \mid G^{(t)}) \mathbb{1}\{f(x) \ne g(x)\}. \tag{16}$$

This is a predictable convex mixture over the disagreement class $\{\mathbb{1}\{f \ne g\} : f, g \in \mathcal{F}\}$, whose VC dimension is $O(d)$ (see Appendix B.4). Applying Lemma 4.1 to this mixture and summing (15) over $t \le T$ gives

$$\sum_{t=1}^{T} \mathbb{E}\,\mathrm{Var}_{f\sim w_t(\cdot\,|\,G^{(t)})}(f(X_t)) \lesssim \sqrt{\frac{T\log|\mathcal{G}|\log(2T)}{\sigma}} + \sqrt{\frac{(d+1)T}{\sigma}}\log^2(2T/\sigma).$$

Substituting into (14) and adding the expected cover approximation cost $O(d\log(2T/\sigma))$ gives Lemma 4.5.

Conclusion

Our results show oracle-efficient learning from smooth data is possible even when the data are agnostic and the base measure unknown. Moreover, the algorithm that we propose is close to the best that could be hoped for: entirely parameter free, requiring no knowledge of the base measure, the smoothness parameter, or the horizon, and conceptually simple. Modulo the remaining $\sqrt{d}$ gap between that which we have shown is achievable by our algorithm and the statistical optimum, our work essentially closes the question of oracle-efficient learning from smooth data.

Statement on the use of AI

In this work, we used generative AI tools to help develop theoretical arguments, formulate and refine mathematical claims and hypotheses, propose proof ingredients, write and revise proofs, and interpret theoretical results. Over several months, the authors made initial progress by establishing a $\widetilde{O}(T^{3/4}d^{1/4}/\sqrt{\sigma})$ regret bound for the algorithm of Wu et al. (2024). The proof analyzed the stability of Rademacher potentials with perturbations on subsampled past covariates, using an information-theoretic decoupling argument to control the effect of dependencies among covariates. The authors supplied the initial proof idea, and the argument was developed and refined through iterative work with GPT 5.6 Sol. GPT 6 Astra subsequently proposed the modification to Gaussian FTPL used in the present manuscript. The corresponding proof was completed through further iterations with the model. We additionally used GPT 6 Astra, GPT 5.6 Sol, and Claude Opus 5.5 to identify and compare related literature, draft and edit manuscript text, improve exposition and organization, and edit LaTeX. All mathematical claims in the present manuscript were independently verified by the authors. The authors take responsibility for the final manuscript, including all text, claims, and proofs produced with AI assistance. We did not use generative AI tools for translation. Synthetic dataset generation, the design of empirical experiments, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this theoretical work.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 263–274, 2008.

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 807–823. PMLR, 2014. URL https://proceedings.mlr.press/v35/abernethy14.html.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

Ingo Althöfer. On sparse approximations to randomized strategies and convex combinations. *Linear Algebra and its Applications*, 199:339–355, 1994. doi: 10.1016/0024-3795(94)90357-3.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3 edition, 2016. URL https://web.mit.edu/dimitrib/www/NLP_Appendixes_Refs.pdf.

Alankrita Bhatt, Nika Haghtalab, and Abhishek Shetty. Smoothed analysis of sequential probability assignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79808–79831, 2023.

Moïse Blanchard. Agnostic smoothed online learning without knowledge of the base measure. *arXiv preprint arXiv:2410.05124v3*, 2025.

Moïse Blanchard, Abhishek Shetty, and Alexander Rakhlin. Characterizing online and private learnability under distributional constraints via generalized smoothness. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 729–759. PMLR, 2026. URL https://proceedings.mlr.press/v336/blanchard26a.html.

Adam Block and Yury Polyanskiy. The sample complexity of approximate rejection sampling with applications to smoothed online learning. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 228–273. PMLR, 2023.

Adam Block and Max Simchowitz. Efficient and near-optimal smoothed online learning for generalized linear functions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7477–7489, 2022.

Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 587–590. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/block21a.html.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Proceedings of the Thirty-Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1716–1786. PMLR, 2022.

Adam Block, Max Simchowitz, and Alexander Rakhlin. Oracle-efficient smoothed online learning for piecewise continuous decision making. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1618–1665. PMLR, 2023a.

Adam Block, Max Simchowitz, and Russ Tedrake. Smoothed online learning for prediction in piecewise affine systems. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41663–41674, 2023b.

Adam Block, Mark Bun, Rathin Desai, Abhishek Shetty, and Zhiwei Steven Wu. Oracle-efficient differentially private learning with public data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113191–113233, 2024a.

Adam Block, Alexander Rakhlin, and Abhishek Shetty. On the performance of empirical risk minimization with smoothed data. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 596–629. PMLR, 2024b.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2–3):321–352, 2007. doi: 10.1007/s10994-006-5001-7. URL https://stoltz.perso.math.cnrs.fr/Publications/CBMS-SOE-ML.pdf.

Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 224–232. ACM, 2018. doi: 10.1145/3240323.3240370.

Thomas M. Cover. Behavior of sequential predictors of binary sequences. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1965)*, pp. 263–272, Prague, 1967. Publishing House of the Czechoslovak Academy of Sciences.

Constantinos Daskalakis, Noah Golowich, Nika Haghtalab, and Abhishek Shetty. Smooth Nash equilibria: Algorithms and complexity. In *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*, volume 287 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 37:1–37:22, 2024. doi: 10.4230/LIPIcs.ITCS.2024.37.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014. URL https://jmlr.org/papers/volume15/rooij14a/rooij14a.pdf.

Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, volume 30 of *Proceedings of Machine Learning Research*, pp. 460–473, 2013.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9203–9215, 2020.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for smoothed adversaries. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4072–4084, 2022.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *Journal of the ACM*, 71(3):1–34, 2024. doi: 10.1145/3656638.

James Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (eds.), *Contributions to the Theory of Games, Volume III*, number 39 in Annals of Mathematics Studies, pp. 97–139. Princeton University Press, 1957.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. doi: 10.1016/0097-3165(95)90052-7.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 128–141, 2016. doi: 10.1145/2897518.2897536.

Marcus Hutter and Jan Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390. ACM, 2019. doi: 10.1145/3306618.3314288.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005. doi: 10.1016/j.jcss.2004.10.016.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. doi: 10.1007/BF00116827.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. doi: 10.1006/inco.1994.1009.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Shahar Mendelson. On the size of convex hulls of small sets. *Journal of Machine Learning Research*, 2:1–18, 2001. URL https://www.jmlr.org/papers/v2/mendelson01a.html.

Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire d'analyse fonctionnelle*, pp. 1–12, 1981. URL https://numdam.org/item/SAF_1980-1981____A5_0/. Exposé no. V, séminaire 1980–1981.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1764–1772, 2011.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2141–2149, 2012.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972. doi: 10.1016/0097-3165(72)90019-2.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. doi: 10.1561/2200000018. URL https://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf.

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2–3):115–142, 2007. doi: 10.1007/s10994-007-5014-x.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pp. 371–386. Morgan Kaufmann, 1990.

Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Online learning in dynamically changing environments. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 325–358. PMLR, 2023.

Changlong Wu, Jin Sima, and Wojciech Szpankowski. Oracle-efficient hybrid online learning with unknown distribution. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4992–5018. PMLR, 2024.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022. URL https://arxiv.org/abs/2210.04157.

A Additional preliminaries

We collect preliminaries for smoothed online learning, variance-based regret bounds for exponential weights, Gaussian integration by parts, and Maurey's empirical method.

A.1 Covers & smoothed online learning

It is well-established in the smoothed online learning literature that covers w.r.t. $L^1(\mu)$ norm provide approximation guarantees along the draw of a smoothed trajectory Haghtalab et al. (2024). In particular, the following proposition follows using the coupling construction of Haghtalab et al. (2024, Theorem 2.1 and Appendix B.2).

Proposition 4.2. *Let $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ be a binary class with VC dimension d. For every probability measure μ, horizon $T \geq 1$, and $\sigma \in (0,1]$, there exists a cover $\mathcal{G} \subseteq \mathcal{F}$ such that $\log|\mathcal{G}| \lesssim d\log(2T/\sigma)$ and, for every covariate process that is σ-smooth with respect to μ,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}}\sum_{s\leq T}\mathbb{1}\{f(X_s)\neq g(X_s)\} > Cd\log(2T/\sigma)\right) \leq \frac{1}{T},$$

where $C>0$ is a universal constant.

Proof of Proposition 4.2. Set $\delta = 1/(2T)$ and $n = \lceil \sigma^{-1}\log(T/\delta)\rceil$. The recursive coupling construction of Haghtalab et al. (2024, Theorem 2.1 and Appendix B.2) implies existence of a coupling between the smooth covariate sequence $X_{1:T}$ and an array $Z_{1:T,1:n}$, drawn i.i.d. from μ, so that

$$\mathbb{P}\left[X_t \in \{Z_{t,1},\dots,Z_{t,n}\} \text{ for every } t\leq T\right] \geq 1 - T(1-\sigma)^n \geq 1-\delta. \tag{17}$$

On the displayed event, for any pair of functions f, g, we can upper bound the discrepancy between f and g on $X_{1:T}$ as:

$$\sum_{t=1}^{T}\mathbb{1}\{f(X_t)\neq g(X_t)\} \leq \sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{1}\{f(Z_{t,i})\neq g(Z_{t,i})\}.$$

The VC covering bound of Haussler (1995, Theorem 1), in its probability-measure formulation (Boucheron et al., 2013, Lemma 13.6) implies the existence of a cover $\mathcal{G}\subseteq\mathcal{F}$ such that

$$\log|\mathcal{G}| \lesssim d\log(2T/\sigma), \qquad \forall f\in\mathcal{F}\ \exists g\in\mathcal{G}:\quad \mathbb{E}_{Z\sim\mu}|f(Z)-g(Z)| \leq \frac{\sigma}{T}. \tag{18}$$

Now, consider the disagreement class

$$\mathcal{H} := \mathcal{F}\Delta\mathcal{F} = \{x\mapsto \mathbb{1}\{f(x)\neq g(x)\} : f,g\in\mathcal{F}\}.$$

The VC dimension of such a class is in $O(d)$ Blumer et al. (1989). Results for norm comparison inequalities for VC classes (Boucheron et al., 2013, Exercise 12.4) thus imply that, with probability at least $1-\delta$ over the i.i.d. array $Z_{1:T,1:n}$,

$$\sup_{h\in\mathcal{H}}\left[\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}h(Z_{t,i}) - 2\,\mathbb{E}_\mu h\right] \lesssim \frac{d\log(2Tn)+\log(1/\delta)}{Tn}.$$

For each $f\in\mathcal{F}$, choose a cover element $g_f\in\mathcal{G}$ satisfying Eq. (18). On the preceding event,

$$\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{1}\{f(Z_{t,i})\neq g_f(Z_{t,i})\} \lesssim \sigma n + d\log(2Tn)+\log(1/\delta) \lesssim d\log(2T/\sigma),$$

where the last inequality uses $\sigma n \leq \log(2T^2)+1$, $\log(2Tn)\lesssim\log(2T/\sigma)$, and $d\geq 1$. Thus,

$$\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{1}\{f(Z_{t,i})\neq g(Z_{t,i})\} \lesssim \sigma n + d\log(2Tn)+\log(1/\delta) \lesssim d\log(2T/\sigma).$$

Intersecting this high probability event with one in Eq. (17), we obtain with probability $1 - 2\delta = 1 - 1/T$:

$$\sup_{f \in \mathcal{F}} \inf_{g \in \mathcal{G}} \sum_{t=1}^{T} \mathbb{1}\{f(X_t) \neq g(X_t)\} \lesssim \sigma n + d \log(2Tn) + \log(1/\delta) \lesssim d \log(2T/\sigma).$$

This concludes the proof. □

For real-valued classes, the above can be extended to bounds that depend on the *fat-shattering dimension* of the class

Proposition A.1. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. There are universal constants $c, C > 0$ such that the following hold.*

(i) For every $m \geq 1$, points $z_{1:m} \in \mathcal{X}^m$, and $0 < \varepsilon \leq 1$, there exist a subset $\mathcal{G} \subset \mathcal{F}$ such that

$$\log |\mathcal{G}| \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2(Cm/\varepsilon), \tag{19}$$

and, for every $f \in \mathcal{F}$, there exists $g \in \mathcal{G}$ such that

$$\sup_{i \in [m]} |f(z_i) - g(z_i)| \leq \varepsilon.$$

(ii) For every μ, horizon $T \geq 1$, $\sigma \in (0,1]$, and $0 < \varepsilon \leq 1/16$ with $1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}) \leq T$, there exist a finite subset $\mathcal{G} \subseteq \mathcal{F}$ and a representative $g_f \in \mathcal{G}$ for each $f \in \mathcal{F}$, with $g_f = f$ for $f \in \mathcal{G}$, such that

$$\log |\mathcal{G}| \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon} \tag{20}$$

and, for every covariate process that is σ-smooth with respect to μ, with probability at least $1 - (8T)^{-4}$, we have

$$\sup_{f \in \mathcal{F}} \sum_{s=1}^{T} \mathbb{1}\{|f(X_s) - g_f(X_s)| > 2\varepsilon\} \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon}. \tag{21}$$

Moreover, for every $f \in \mathcal{F}$,

$$\mu\left(\{x \in \mathcal{X} : |f(x) - g_f(x)| > 2\varepsilon\}\right) \lesssim \frac{\sigma}{T}(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon}. \tag{22}$$

Proof. Part (i) is the packing estimate of Alon, Ben-David, Cesa-Bianchi, and Haussler, in the form stated by Mendelson (2001, Theorem 1.3). To prove part (ii), we use a double-sampling trick. First, let

$$\delta = (8T)^{-4}, \quad n = \left\lceil \sigma^{-1} \log(2T/\delta) \right\rceil, \quad N = Tn, \quad k = \left\lceil C(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon} \right\rceil.$$

Let $Z_{1:N}$ and $Z'_{1:N}$ be independent i.i.d. samples from μ. We claim that

$$\mathbb{P}\left[\sup_{\substack{f,g \in \mathcal{F}, \\ \max_{j \leq N} |f(Z'_j) - g(Z'_j)| \leq \varepsilon}} \sum_{j=1}^{N} \mathbb{1}\{|f(Z_j) - g(Z_j)| > 2\varepsilon\} \geq k \right] \leq \frac{\delta}{2}. \tag{23}$$

Condition on the unordered multiset $\{Z_j, Z'_j\}, j \in [N]$. Choose an $\varepsilon/4$-cover $\mathcal{G}' \subseteq \mathcal{F}$ on the $2N$ points $\{Z_j, Z'_j\}, j \leq N$ using part (i), so that $\mathcal{G}'$ depends only on the multiset of $2N$ points, but not on their ordering. By part (i), its cardinality satisfies

$$\log |\mathcal{G}'| \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2(CN/\varepsilon) \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon}.$$

If f, g satisfy the event in (23), choose functions $h, h' \in \mathcal{G}'$ from the cover such that $|h - f| \leq \varepsilon/4$ and $|h' - g| \leq \varepsilon/4$ on all $2N$ points. Then, on the event Eq. (23), we have

$$\max_{j \leq N} |h(Z'_j) - h'(Z'_j)| \leq \frac{3\varepsilon}{2}, \qquad \sum_{j=1}^{N} \mathbb{1}\left\{ |h(Z_j) - h'(Z_j)| > \frac{3\varepsilon}{2} \right\} \geq k.$$

Fix $h, h' \in \mathcal{G}'$. Now, recall that so far we have conditioned on the multiset of $2N$ observations, but not on their ordering. Call a sample $Z \in \{Z_j, Z'_j\}_{j \le N}$ good if $|h(Z) - h'(Z)| \le 3\varepsilon/2$, and call it bad otherwise. Then, on the above event, all bad samples are from $Z_{1:N}$, and there are at least k of them. By exchangeability, the probability that all $\ge k$ bad samples are assigned to $Z_{1:N}$ is at most 2^{-k}. Thus, conditional on the multiset of $2N$ points, we have for any fixed pair h, h':

$$\mathbb{P}\left[\max_{j\le N}|h(Z'_j) - h'(Z'_j)| \le \frac{3\varepsilon}{2}, \quad \sum_{j=1}^{N} \mathbb{1}\left\{|h(Z_j) - h'(Z_j)| > \frac{3\varepsilon}{2}\right\} \ge k \,\middle|\, \{Z_j, Z'_j\}_{j\le N}\right] \le 2^{-k}.$$

A union bound over $h, h' \in \mathcal{G}'$ therefore bounds the conditional probability in (23) by

$$|\mathcal{G}'|^2 2^{-k} = \exp\left(2\log|\mathcal{G}'| - k\log 2\right) \le \frac{\delta}{2},$$

where the last inequality follows from the cardinality bound and $\log(2/\delta) \lesssim \log(2T)$, for C sufficiently large. The above (23).

Now, to construct the desired cover, note that Eq. (23) implies existence of at least one *fixed* tuple $z'_{1:N}$ such that, when $Z_{1:N} \sim \mu$, we have

$$\mathbb{P}\left[\sup_{\substack{f,g\in\mathcal{F},\\ \max_{j\le N}|f(z'_j)-g(z'_j)|\le\varepsilon}} \sum_{j=1}^{N} \mathbb{1}\{|f(Z_j) - g(Z_j)| > 2\varepsilon\} \ge k\right] \le \frac{\delta}{2}.$$

Construct an ε-cover $\mathcal{G}$ on $z'_{1:N}$ using part (i), and choose, for each $f \in \mathcal{F}$, a representative $g_f \in \mathcal{G}$ such that

$$\max_{j\le N}|f(z'_j) - g_f(z'_j)| \le \varepsilon,$$

taking $g_f = f$ for $f \in \mathcal{G}$. Part (i) gives (20). By (23), with probability at least $1 - \delta/2$ over a fresh i.i.d. sample $Z_{1:N}$,

$$\sup_{f\in\mathcal{F}} \sum_{j=1}^{N} \mathbb{1}\{|f(Z_j) - g_f(Z_j)| > 2\varepsilon\} < k.$$

The conclusion follows from (17): with probability at least $1 - \delta/2$, we can couple $X_{1:T}$ to $Z_{1:N}$ such that $X_{1:T} \subset Z_{1:N}$. Applying union bound gives:

$$\sup_{f\in\mathcal{F}} \sum_{s=1}^{T} \mathbb{1}\{|f(X_s) - g_f(X_s)| > 2\varepsilon\} \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon},$$

with probability $1 - \delta$, as desired.

Finally, we deduce the bound in (22). Fix $f \in \mathcal{F}$ and let

$$p = \mu\left(\{x \in \mathcal{X} : |f(x) - g_f(x)| > 2\varepsilon\}\right).$$

Since f and its representative g_f are fixed before drawing $Z_{1:N}$, the number of indices $j \le N$ with $|f(Z_j) - g_f(Z_j)| > 2\varepsilon$ has the binomial distribution with N trials and success probability p. The bound above for this sample gives

$$\mathbb{P}\left[\sum_{j=1}^{N} \mathbb{1}\{|f(Z_j) - g_f(Z_j)| > 2\varepsilon\} < k\right] \ge 1 - \frac{\delta}{2} > \frac{1}{2}.$$

Suppose, for a contradiction, that $p \ge 2k/N$. Then the mean Np of this count is at least $2k$, so a count below k is less than half its mean. The multiplicative Chernoff bound therefore gives

$$\mathbb{P}\left[\sum_{j=1}^{N} \mathbb{1}\{|f(Z_j) - g_f(Z_j)| > 2\varepsilon\} < k\right] \le e^{-Np/8} \le e^{-k/4} \le \frac{1}{2},$$

where the last inequality holds for C sufficiently large. This contradicts the preceding lower bound on the same probability. Thus,

$$p < \frac{2k}{N} \le \frac{2k\sigma}{T} \lesssim \frac{\sigma}{T}(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon},$$

where we used $N = Tn \ge T/\sigma$ and the definition of k. This proves (22). □

A.2 Variance-based regret bounds

Variance-based bounds are a standard tool for analyzing exponential weights (Cesa-Bianchi et al., 2007). The following is a one-round Bernstein bound of de Rooij et al. (2014, Lemma 4), which will be useful in our analysis of exponential weights.

Proposition A.2. *Let $\mathcal{G}$ be a nonempty finite class, let $w \in \Delta(\mathcal{G})$ be a set of weights, and let $\ell\colon \mathcal{G} \to \mathbb{R}$ be a loss function. Suppose,*

$$\max_{f\in\mathcal{G}} \ell(f) - \min_{f\in\mathcal{G}} \ell(f) \le s,$$

that is, s upper bounds the range of the loss. Then, for every $\eta > 0$,

$$\mathbb{E}_{f\sim w}\, \ell(f) + \frac{1}{\eta} \ln \sum_{f\in\mathcal{G}} w(f) e^{-\eta\ell(f)} \le \frac{e^{\eta s} - \eta s - 1}{\eta s^2} \operatorname{Var}_{f\sim w}(\ell(f)),$$

where the coefficient is interpreted as $\eta/2$ when $s = 0$.

A.3 Stein's lemma

The following identity is due to Stein (1981) and will be handy throughout the proof.

Lemma A.3 (Stein's lemma)**.** *Let $G = (G_1, \dots, G_n)$ have independent standard Gaussian coordinates, and let $g : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz. Then, for every $s \le n$,*

$$\mathbb{E}_G\, \partial_{G_s} g(G) = \mathbb{E}_G[G_s g(G)].$$

A.4 Convex hull approximations

We use the following consequence of Maurey's empirical method (Pisier, 1981), in the ℓ_∞ form due to Althöfer (1994).

Proposition A.4 (Maurey's empirical method)**.** *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. For any $h \in \operatorname{conv}(\pm\mathcal{F})$, integers $m, K \ge 1$, and points $Z_1, \dots, Z_m \in \mathcal{X}$, there exist functions $f_1, \dots, f_K \in \pm\mathcal{F}$ such that*

$$\max_{s\in[m]} \left| h(Z_s) - \frac{1}{K} \sum_{i=1}^{K} f_i(Z_s) \right| \le \sqrt{\frac{2\log(2m)}{K}}.$$

B Proofs for the binary case

The binary proof has three ingredients: a surprise lemma using second moments, a reduction to exponential weights, and a regret bound for the exponential-weights reference.

B.1 Surprise lemma

We first bound the cumulative expected values of predictable test functions by their historical second moments on a tangent sequence. A norm comparison then replaces the tangent moments by their realized counterparts.

B.1.1 Surprise lemma with tangent sequence

The following lemma follows from the change-of-measure argument of Xie et al. (2022, Section 3.2, equations (3)–(5)). We restate the result in our notation, and provide a proof for completeness.

Lemma B.1. *For predictable test functions $h_t\colon \mathcal{X} \to [0,1]$, we have*

$$\mathbb{E} \sum_{t=1}^{T} h_t(X_t) \lesssim \sqrt{\frac{\log(2T)}{\sigma}\, \mathbb{E} \sum_{t=1}^{T} \sum_{s<t} h_t(\bar{X}_s)^2} + \frac{\log(2T)}{\sigma}. \tag{24}$$

Proof. We begin by letting

$$a_t(x) = \sigma \frac{\mathrm{d}p_t}{\mathrm{d}\mu}(x) \in [0, 1].$$

Similar to Xie et al. (2022), for each $x \in \mathcal{X}$ let

$$\tau(x) := \min \left\{ t \colon \sum_{s \le t} a_s(x) \ge 1 \right\}$$

be the index of the first round in which the probability mass on x exceeds 1. Intuitively, we can partition the surprise contribution $\mathbb{E}\, h_t(X_t)$ in each round x's that are still in the "burn-in" period ($\tau(x) < t$) and those that have left the burn-in phrase. Observe that the burn-in contribution can be upper bounded as follows:

$$\frac{1}{\sigma} \int \sum_{t=1}^{\tau(x)} a_t(x) h_t(x) \mathrm{d}\mu(x) \le \frac{1}{\sigma} + \frac{1}{\sigma} \int \sum_{t=1}^{\tau(x)-1} a_t(x) h_t(x) \mathrm{d}\mu(x) \le \frac{2}{\sigma}, \tag{25}$$

where we used $|h_t| \le 1$. It remains to bound the contribution of those covariates that have left the burn-in period. From Xie et al. (2022, Lemma 4), we have with probability 1:

$$\sum_{t=1}^{T} \frac{a_t(x)}{1 + \sum_{s<t} a_s(x)} \lesssim \log(2T).$$

On the event $\{\sum_{s<t} a_s(x) \ge 1\}$, we also have

$$\frac{a_t(x)^2}{\sum_{s<t} a_s(x)} \le \frac{2a_t(x)}{1 + \sum_{s<t} a_s(x)}.$$

Thus, using Cauchy–Schwarz, we can upper bound the contribution of covariates that have left the burn-in period is bounded by

$$\begin{aligned}
&\frac{1}{\sigma}\, \mathbb{E} \int \sum_{t=\tau(x)+1}^{T} a_t(x) h_t(x) \mathrm{d}\mu(x) \\
&\quad \le \left(\frac{1}{\sigma}\, \mathbb{E} \int \sum_{t=\tau(x)+1}^{T} \frac{a_t(x)^2}{\sum_{s<t} a_s(x)} \mathrm{d}\mu(x) \right)^{1/2} \left(\frac{1}{\sigma}\, \mathbb{E} \int \sum_{t=1}^{T} \left(\sum_{s<t} a_s(x) \right) h_t(x)^2 \mathrm{d}\mu(x) \right)^{1/2} \\
&\quad \lesssim \sqrt{\frac{\log(2T)}{\sigma} \frac{1}{\sigma}\, \mathbb{E} \int \sum_{t=1}^{T} \left(\sum_{s<t} a_s(x) \right) h_t(x)^2 \mathrm{d}\mu(x)} \\
&\quad = \sqrt{\frac{\log(2T)}{\sigma}\, \mathbb{E} \sum_{t=1}^{T} \sum_{s<t} \mathbb{E}_{p_s} h_t^2}.
\end{aligned}$$

Combing the above with Eq. (25), we obtain:

$$\mathbb{E} \sum_{t=1}^{T} h_t(X_t) \lesssim \sqrt{\frac{\log(2T)}{\sigma}\, \mathbb{E} \sum_{t=1}^{T} \sum_{s<t} h_t(\bar{X}_s)^2} + \frac{\log(2T)}{\sigma},$$

as desired. □

For the same predictable functions $0 \le h_t \le 1$, Block et al. (2024b, Lemma 12) gives, up to universal constants,

$$\mathbb{E} \sum_{t=1}^{T} h_t(X_t) \lesssim \frac{\log(2T/\sigma)}{\sigma} + \log(2T/\sigma) \sqrt{\frac{T}{\sigma}\, \mathbb{E} \sum_{t=1}^{T} \frac{1}{t} \sum_{s<t} \mathbb{E}_{p_s} h_t}.$$

Since $h_t^2 \le h_t$ and $t \le T$,

$$\mathbb{E} \sum_{t=1}^{T} \sum_{s<t} \mathbb{E}_{p_s} h_t^2 \le T\, \mathbb{E} \sum_{t=1}^{T} \frac{1}{t} \sum_{s<t} \mathbb{E}_{p_s} h_t.$$

Thus (24) also implies this first-moment bound in Block et al. (2024b).

B.1.2 Comparison inequality for signed convex hulls

We next compare tangent and realized second moments uniformly for elements of signed convex hulls.

Lemma B.2. *Under the same assumptions as in Lemma 4.1, we have*

$$\mathbb{E} \sup_{h \in \mathrm{conv}(\pm\mathcal{F})} \left\{ \sum_{t \le T} \mathbb{E}_{\bar{X}_t} h^2(\bar{X}_t) - 12 \sum_{t \le T} h(X_t)^2 \right\} \lesssim d \log^3(2T/\sigma),$$

Proof of Lemma B.2. In the proof, we first reduce the claim to comparison for binary function class using layer-cake identity for expectations and Maurey's empirical method Proposition A.4. Then, the proof is concluded by invoking Block et al. (2024b, Theorem 10). To start, using the layer-cake identity, we have

$$\sum_{s=1}^{T} \mathbb{E}_{\bar{X}_s} h^2(\bar{X}_s) \le 2 + \int_{1/\sqrt{T}}^{1} 2\lambda \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \ge \lambda] \mathrm{d}\lambda.$$

Also, note that we have:

$$\sum_{s=1}^{T} h(X_s)^2 \ge \int_{1/\sqrt{T}}^{1} \frac{\lambda}{2} \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \ge \lambda/2\} \mathrm{d}\lambda.$$

Thus,

$$\begin{aligned} &\sum_{s=1}^{T} \mathbb{E}_{\bar{X}_s} h^2(\bar{X}_s) - 12 \sum_{s=1}^{T} h(X_s)^2 \\ &\qquad \le 2 + \int_{1/\sqrt{T}}^{1} 2\lambda \left\{ \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \ge \lambda] - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \ge \lambda/2\} \right\} \mathrm{d}\lambda. \end{aligned} \tag{26}$$

For $1/\sqrt{T} \le \lambda \le 1$, take $K = \lceil 32\lambda^{-2} \log(4T) \rceil$, and define

$$\mathcal{F}_\lambda = \left\{ x \mapsto \mathbb{1}\left\{ \left| \frac{1}{K} \sum_{i=1}^{K} f_i(x) \right| \ge \frac{3\lambda}{4} \right\} : f_1, \dots, f_K \in \pm\mathcal{F} \right\},$$

to be the class obtained by averaging K functions from $\mathcal{F}$. Applying Proposition A.4 to the $2T$ points $X_{1:T}, \bar{X}_{1:T}$, for every $h \in \mathrm{conv}(\pm\mathcal{F})$ there exists $g \in \mathcal{F}_\lambda$ such that

$$\mathbb{1}\{|h(x)| \ge \lambda\} \le g(x) \le \mathbb{1}\{|h(x)| \ge \lambda/2\} \tag{27}$$

for every $x \in \{X_s, \bar{X}_s : s \le T\}$. Set $m = \lceil 2T \log(T)/\sigma \rceil$ and let $Z_{1:m} \sim \mu^m$. Using Block et al. (2024b, Theorem 10) and a bound on Will's functional using growth function of the class, we have:

$$\begin{aligned} &\mathbb{E} \sup_{h \in \mathrm{conv}(\pm\mathcal{F})} \left\{ \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \ge \lambda] - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \ge \lambda/2\} \right\} \\ &\quad \le \mathbb{E} \sup_{h \in \mathrm{conv}(\pm\mathcal{F})} \left\{ \sum_{s=1}^{T} \mathbb{1}\{|h(\bar{X}_s)| \ge \lambda\} - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \ge \lambda/2\} \right\} \\ &\quad \le \mathbb{E} \sup_{g \in \mathcal{F}_\lambda} \sum_{s=1}^{T} \left(g(\bar{X}_s) - 3g(X_s) \right) \\ &\quad \lesssim 1 + \log \sup_{Z_{1:m}} \left| \mathcal{F}_\lambda|_{Z_{1:m}} \right|. \end{aligned} \tag{28}$$

We now bound growth function of $\mathcal{F}_\lambda$. Note that, on any m points $Z_1, \dots, Z_m$, by Sauer's lemma (Sauer, 1972), we have

$$\left| \pm\mathcal{F}|_{Z_{1:m}} \right| \le 2(2m)^d.$$

Thus, via a simple counting argument, we have:

$$\left|\mathcal{F}_\lambda|_{Z_{1:m}}\right| \le [2(2m)^d]^K.$$

Thus, we have

$$1 + \log \sup_{Z_{1:m}} \left|\mathcal{F}_\lambda|_{Z_{1:m}}\right| \lesssim dK \log(2m) \lesssim d\lambda^{-2} \log^2(2T/\sigma).$$

Plugging this into Eqs. (26) and (28) gives

$$\mathbb{E} \sup_{h \in \mathrm{conv}(\pm\mathcal{F})} \left\{ \sum_{s=1}^T \mathbb{E}_{\bar{X}_s} h^2(\bar{X}_s) - 12 \sum_{s=1}^T h(X_s)^2 \right\}$$
$$\lesssim 1 + \int_{1/\sqrt{T}}^1 2\lambda \left(d\lambda^{-2} \log^2(2T/\sigma) \right) \mathrm{d}\lambda \lesssim d \log^3(2T/\sigma),$$

as desired. □

B.1.3 Proof of the surprise lemma

It remains to combine the change-of-measure bound in Lemma B.1 with the uniform norm comparison in Lemma B.2.

Lemma 4.1. *Let $\mathcal{F}$ be a class of VC dimension d and let $\{h_t\}_{t\in[T]}$ be a predictable sequence of functions with $h_t \in \mathrm{conv}(\pm\mathcal{F})$. Then*

$$\mathbb{E} \sum_{t=1}^T |h_t(X_t)| \lesssim \sqrt{\frac{\log(2T)}{\sigma} \mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2} + \sqrt{\frac{dT}{\sigma}} \log^2\left(\frac{2T}{\sigma}\right). \tag{10}$$

Proof of Lemma 4.1. If $\sigma T < 1$, the right-hand side of (10) is at least T, so the claim follows from $|h_t| \le 1$. Otherwise, apply Lemma B.1 to $|h_t|$ and the uniform bound in Lemma B.2 to h_t on each prefix $X_{1:t-1}$. This gives

$$\mathbb{E} \sum_{t=1}^T |h_t(X_t)| \lesssim \sqrt{\frac{\log(2T)}{\sigma} \left(\mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2 + dT \log^3(2T/\sigma) \right)} + \frac{\log(2T)}{\sigma}$$
$$\lesssim \sqrt{\frac{\log(2T)}{\sigma} \mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2} + \sqrt{\frac{dT}{\sigma}} \log^2(2T/\sigma),$$

where the last step uses $\sigma T \ge 1$, $d \ge 1$, and $\log(2T) \le \log(2T/\sigma)$. □

The proofs of Lemmas B.2 and 4.1 use the VC dimension only to count the restrictions of $\pm\mathcal{F}$ to finitely many points. The same argument therefore applies to finite classes of real-valued functions, which we use in Appendix C.

Corollary B.3. *Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ be a nonempty finite class, and let $\{h_t\}_{t\in[T]}$ be a predictable sequence of functions with $h_t \in \mathrm{conv}(\pm\mathcal{F})$. Then,*

$$\mathbb{E} \sum_{t=1}^T |h_t(X_t)| \lesssim \sqrt{\frac{\log(2T)}{\sigma} \mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2} + \sqrt{\frac{\log(2|\mathcal{F}|)T}{\sigma}} \log^2\left(\frac{2T}{\sigma}\right).$$

B.2 Proof of the prediction comparison on the past

We begin by proving that predictions of FTPL Eq. (2) and exponential weights Eq. (3) in round t are close on past covariates X_s with $s < t$.

Lemma 4.4. *For every $t \le T$,*

$$\mathbb{E} \sum_{s<t} \left(\mathbb{E}_{G^{(t)}} \hat{f}_t(X_s) - \mathbb{E}_{G^{(t)}} \tilde{f}_t(X_s) \right)^2 \lesssim d^2 \log^2(2T/\sigma). \tag{12}$$

Proof of Lemma 4.4. Fix the history $(X_{1:t-1}, Y_{1:t-1})$ and let $G := G^{(t)}$ to lighten the notational load. Use the same potentials as in the proof sketch:

$$\Phi_t(G) = \max_{f\in\mathcal{F}} \sum_{s<t} (Y_s - G_s) f(X_s),$$

$$\widetilde{\Phi}_t(G) = \ln \sum_{f\in\mathcal{G}} \exp\left(\sum_{s<t} (Y_s - G_s) f(X_s)\right)$$

$$= \max_{w\in\Delta(\mathcal{G})} \left[\mathbb{E}_{f\sim w} \sum_{s<t} (Y_s - G_s) f(X_s) + H(w)\right].$$

$w_t(\cdot \mid G)$ from (3) maximize the latter potential in its variational form. Note that, by absolute continuity of the Gaussians, the set of all functions that maximize the former or the latter potential share predictions on the past covariates X_s. Differentiating the two potentials therefore gives, for every $s < t$ and almost every G, by Danskin's theorem for the finite maximum (Bertsekas, 2016, Appendix B):

$$\partial_{G_s} \Phi_t(G) = -\hat{f}_t(X_s), \qquad \partial_{G_s} \widetilde{\Phi}_t(G) = -\tilde{f}_t(X_s).$$

Both potentials are Lipschitz functions of G, so Stein's lemma (Lemma A.3) yields

$$\sum_{s<t} \left(\mathbb{E}_G\, \hat{f}_t(X_s) - \mathbb{E}_G\, \tilde{f}_t(X_s)\right)^2$$

$$= \left\|\mathbb{E}_G\, \nabla_G(\Phi_t(G) - \widetilde{\Phi}_t(G))\right\|_2^2 = \left\|\mathbb{E}_G[G(\Phi_t(G) - \widetilde{\Phi}_t(G))]\right\|_2^2.$$

Since $\mathbb{E}_G\, G = 0$ and $\mathrm{Var}_G(u^\top G) = 1$ for every unit vector $u \in \mathbb{R}^{t-1}$, Cauchy–Schwarz gives

$$\sum_{s<t} \left(\mathbb{E}_G\, \hat{f}_t(X_s) - \mathbb{E}_G\, \tilde{f}_t(X_s)\right)^2 = \sup_{\|u\|_2=1} \left|\mathrm{Cov}_G(u^\top G, \Phi_t - \widetilde{\Phi}_t)\right|^2$$

$$\leq \mathrm{Var}_G(\Phi_t - \widetilde{\Phi}_t) \leq \mathbb{E}_G(\Phi_t - \widetilde{\Phi}_t)^2. \tag{29}$$

Now, let

$$\Phi_t^\dagger(G) = \max_{f\in\mathcal{G}} \sum_{s<t} (Y_s - G_s) f(X_s),$$

by the unregularized potential over the cover $\mathcal{G}$. Since entropy is upper bounded by $\ln|\mathcal{G}|$, we have:

$$0 \leq \widetilde{\Phi}_t(G) - \Phi_t^\dagger(G) \leq \ln|\mathcal{G}| \lesssim d\log(2T/\sigma). \tag{30}$$

By Proposition 4.2, with probability at least $1 - 1/T$ over the history,

$$\sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \sum_{s<t} \mathbb{1}\{f(X_s) \neq g(X_s)\} \leq Cd\log(2T/\sigma).$$

On the above event, using $Y_t \in \{\pm 1\}$, we have:

$$0 \leq \Phi_t(G) - \Phi_t^\dagger(G) \leq 2Cd\log(2T/\sigma)$$

$$+ \max_{\substack{f\in\mathcal{F},\ g\in\mathcal{G}:\\ \sum_{s<t}\mathbb{1}\{f(X_s)\neq g(X_s)\}\leq Cd\log(2T/\sigma)}} \left|\sum_{s<t} G_s(f(X_s) - g(X_s))\right|. \tag{31}$$

For each pair in the maximum, its Gaussian sum has variance

$$\mathrm{Var}_G\left(\sum_{s<t} G_s(f(X_s) - g(X_s))\right) = 4\sum_{s<t} \mathbb{1}\{f(X_s) \neq g(X_s)\} \leq 4Cd\log(2T/\sigma).$$

By Sauer–Shelah lemma, the maximum in Eq. (31) is taken over at most $(2t)^{2(d+1)}$ points. Applying Gaussian tail inequality with a union bound then gives:

$$\mathbb{E}_G\left[\max_{\substack{f\in\mathcal{F},\ g\in\mathcal{G}:\\ \sum_{s<t}\mathbb{1}\{f(X_s)\neq g(X_s)\}\leq Cd\log(2T/\sigma)}} \left|\sum_{s<t} G_s(f(X_s) - g(X_s))\right|^2\right]$$

$$\lesssim d\log(2T/\sigma)(d+1)\log(2t) \lesssim d^2\log^2(2T/\sigma).$$

It follows that $\mathbb{E}_G(\Phi_t - \Phi_t^\dagger)^2 \lesssim d^2\log^2(2T/\sigma)$. Combining this with Eq. (30), we have:

$$\begin{aligned}\mathbb{E}_G(\Phi_t - \widetilde{\Phi}_t)^2 &\le 2\,\mathbb{E}_G(\Phi_t - \Phi_t^\dagger)^2 + 2\,\mathbb{E}_G(\widetilde{\Phi}_t - \Phi_t^\dagger)^2 \\ &\lesssim d^2\log^2(2T/\sigma).\end{aligned}$$

Plugging this into Eq. (29) concludes the proof. □

B.3 Proof of the reduction to exponential weights

We now apply the surprise lemma Lemma 4.1 to the difference of predictions.

Lemma 4.3 (Reduction to exponential weights)**.** *Consider a binary function class $\mathcal{F}\colon \mathcal{X} \to \{\pm 1\}$ of VC dimension d. For $\hat{f}_t$ and $\tilde{f}_t$ as defined in Eqs. (3) and (6), respectively, we have the following bound:*

$$\mathbb{E}\sum_{t=1}^{T} |\,\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t) - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)| \lesssim d\log^2(2T/\sigma)\sqrt{\frac{T}{\sigma}}. \tag{11}$$

Proof of Lemma 4.3. Apply Lemma 4.1 to the sequence

$$h_t = \frac{1}{2}\left(\mathbb{E}_{G^{(t)}}\,\hat{f}_t - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t\right), \qquad t \in [T].$$

Note that both $\mathbb{E}_{G^{(t)}}\,\hat{f}_t$ and $\mathbb{E}_{G^{(t)}}\,\tilde{f}_t$ are elements of $\mathrm{conv}(\pm\mathcal{F})$ and are predictable. Thus, using Lemma 4.1 gives

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\left|\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t) - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)\right| \\ &\lesssim \left(\frac{\log(2T)}{\sigma}\,\mathbb{E}\sum_{t=1}^{T}\sum_{s<t}\left(\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_s) - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_s)\right)^2\right)^{1/2} \\ &\qquad + \sqrt{\frac{dT}{\sigma}}\log^2(2T/\sigma).\end{aligned}$$

By Lemma 4.4, the expected double sum is in $O(d^2T\log^2(2T/\sigma))$. Substituting this bound and using $\log(2T) \le \log(2T/\sigma)$ gives

$$\mathbb{E}\sum_{t=1}^{T}\left|\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t) - \mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)\right| \lesssim d\log^2(2T/\sigma)\sqrt{\frac{T}{\sigma}}.$$

This proves (11). □

B.4 Regret of Gaussian-perturbed exponential weights

We first prove Proposition 4.6, which reduces the regret bound to controlling cumulative prediction variance. We then prove Lemma 4.7, which bounds the variances on the historical covariates, and combine it with the surprise lemma to prove Lemma 4.5.

Proposition 4.6. *Let $\mathcal{G} \subseteq \mathcal{F}$ be a nonempty finite class, and let $\ell(\hat{y}, y) = -y\hat{y}$ be the binary loss. For every data sequence, the predictor $\tilde{f}_t$ defined in (3) satisfies the following inequality,*

$$\sum_{t=1}^{T}\mathbb{E}_{G^{(t)}}\,\ell(\tilde{f}_t(X_t), Y_t) - \min_{f\in\mathcal{G}}\sum_{t=1}^{T}\ell(f(X_t), Y_t) \lesssim \log|\mathcal{G}| + \sum_{t=1}^{T}\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_t)). \tag{14}$$

Proposition 4.6 is the special case of the following proposition in which $\mathcal{G} \subseteq \mathcal{F}$ consists of binary functions and $\ell(\hat{y}, y) = -y\hat{y}$, which is linear, and hence convex and 1-Lipschitz, in its first argument. We use the general form in Appendix C.4.

Proposition B.4. *Let $\mathcal{G}$ be a nonempty finite class of functions $\mathcal{X} \to [-1,1]$, and let $\ell(\hat{y}, y)$ be convex and 1-Lipschitz in its first argument. For every data sequence, the predictor $\tilde{f}_t$ defined in (3) satisfies the following inequality,*

$$\sum_{t=1}^{T} \mathbb{E}_{G^{(t)}}\, \ell(\tilde{f}_t(X_t), Y_t) - \min_{f\in\mathcal{G}} \sum_{t=1}^{T} \ell(f(X_t), Y_t) \lesssim \log|\mathcal{G}| + \sum_{t=1}^{T} \mathbb{E}_{G^{(t)}}\, \mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_t)).$$

Proof of Proposition B.4. Recall the log-partition potential from the proof of Lemma 4.4, written for a general loss for each t:

$$\widetilde{\Phi}_t(G^{(t)}) == \ln \sum_{f\in\mathcal{G}} \exp\left(-\sum_{s<t}\left[\ell(f(X_s), Y_s) + G_s^{(t)} f(X_s)\right]\right),$$

For a fixed round t, we have the following identity:

$$\widetilde{\Phi}_{t+1}(G^{(t+1)}) - \widetilde{\Phi}_t(G^{(t+1)}_{1:t-1}) = \ln \sum_{f\in\mathcal{G}} w_t(f \mid G^{(t+1)}_{1:t-1}) \exp\left(-\ell(f(X_t), Y_t) - G_t^{(t+1)} f(X_t)\right).$$

Apply Proposition A.2 to these weights and the losses $\ell(f(X_t), Y_t) + G_t^{(t+1)} f(X_t)$ at learning rate 1. Their range is at most $s_t := 2(1 + |G_t^{(t+1)}|)$, and

$$\mathrm{Var}_{f\sim w_t(\cdot|G^{(t+1)}_{1:t-1})}(\ell(f(X_t), Y_t) + G_t^{(t+1)} f(X_t)) \le (1 + |G_t^{(t+1)}|)^2\, \mathrm{Var}_{f\sim w_t(\cdot|G^{(t+1)}_{1:t-1})}(f(X_t)),$$

since $y \mapsto \ell(y, Y_t) + G_t^{(t+1)} u$ is $(1 + |G_t^{(t+1)}|)$-Lipschitz on $[-1,1]$. Thus, cancelling s_t^2 from numerator and the denominator in Proposition A.2, we have

$$\begin{aligned}
&\sum_{f\in\mathcal{G}} w_t(f \mid G^{(t+1)}_{1:t-1})[\ell(f(X_t), Y_t) + G_t^{(t+1)} f(X_t)] \\
&\quad\le \widetilde{\Phi}_t(G^{(t+1)}_{1:t-1}) - \widetilde{\Phi}_{t+1}(G^{(t+1)}) \\
&\qquad + \frac{e^{2(1+|G_t^{(t+1)}|)} - 2(1 + |G_t^{(t+1)}|) - 1}{4}\, \mathrm{Var}_{f\sim w_t(\cdot|G^{(t+1)}_{1:t-1})}(f(X_t)).
\end{aligned}$$

This inequality also holds when the loss range is zero. Note that the last coordinate $G_t^{(t+1)}$ is centered and independent of the first $t-1$ coordinates. In particular,

$$\mathbb{E}_{G_t^{(t+1)}} \frac{e^{2(1+|G_t^{(t+1)}|)} - 2(1 + |G_t^{(t+1)}|) - 1}{4} \le \frac{e^2}{4}\, \mathbb{E}_{G_t^{(t+1)}}\left(e^{2G_t^{(t+1)}} + e^{-2G_t^{(t+1)}}\right) = \frac{e^4}{2}.$$

Now, $G^{(t+1)}_{1:t-1}$ has the same law as $G^{(t)}$ conditionally on the history up to round t. Thus, using convexity of the loss,

$$\begin{aligned}
\mathbb{E}_{G^{(t)}}\, \ell(\tilde{f}_t(X_t), Y_t) &\le \mathbb{E}_{G^{(t)}}\, \widetilde{\Phi}_t(G^{(t)}) - \mathbb{E}_{G^{(t+1)}}\, \widetilde{\Phi}_{t+1}(G^{(t+1)}) \\
&\quad + \frac{e^4}{2}\, \mathbb{E}_{G^{(t)}}\, \mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_t)).
\end{aligned} \tag{32}$$

Note that we have $\widetilde{\Phi}_1 = \ln|\mathcal{G}|$, and

$$\begin{aligned}
\mathbb{E}_{G^{(T+1)}}\, \widetilde{\Phi}_{T+1}(G^{(T+1)}) &\ge \max_{f\in\mathcal{G}} \mathbb{E}_{G^{(T+1)}}\left[-\sum_{s=1}^{T}\left(\ell(f(X_s), Y_s) + G_s^{(T+1)} f(X_s)\right)\right] \\
&= -\min_{f\in\mathcal{G}} \sum_{s=1}^{T} \ell(f(X_s), Y_s),
\end{aligned}$$

by Jensen's inequality. Combining this with Eq. (32) and telescoping gives the desired result. □

Lemma 4.7. *Let $\mathcal{G}$ be a nonempty finite class of functions $\mathcal{X} \to [-1,1]$ and let ℓ be any loss. For every $t \le T$ and every fixed history $(X_s, Y_s)_{s<t}$, the weights in (3) satisfy*

$$\sum_{s<t} \left(\mathbb{E}_{G^{(t)}}\, \mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_s))\right)^2 \le 2\log|\mathcal{G}|. \tag{15}$$

Proof of Lemma 4.7. Write $G = G^{(t)}$ throughout this proof for notational brevity. Differentiating each weight w_t from Eq. (3) w.r.t. a Gaussian perturbations G_s, we obtain:

$$\begin{aligned}\partial_{G_s} w_t(f \mid G) &= \partial_{G_s} \frac{\exp\left(-\sum_{r<t}[\ell(f(X_r), Y_r) + G_r f(X_r)]\right)}{\sum_{g\in\mathcal{G}} \exp\left(-\sum_{r<t}[\ell(g(X_r), Y_r) + G_r g(X_r)]\right)} \\ &= -w_t(f \mid G) f(X_s) + w_t(f \mid G) \sum_{g\in\mathcal{G}} w_t(g \mid G) g(X_s) \\ &= -w_t(f \mid G)\left(f(X_s) - \tilde{f}_t(X_s)\right).\end{aligned}$$

Multiplying by $f(X_s)$ and summing over $f \in \mathcal{G}$ gives

$$\begin{aligned}\partial_{G_r} \tilde{f}_t(X_s) &= -\left[\sum_{f\in\mathcal{G}} w_t(f \mid G) f(X_s) f(X_r) - \tilde{f}_t(X_s)\tilde{f}_t(X_r)\right] \\ &= -\operatorname{Cov}_{f\sim w_t(\cdot\mid G)}\left(f(X_s), f(X_r)\right)\end{aligned}$$

and in particular, taking $r = s$,

$$\partial_{G_s} \tilde{f}_t(X_s) = -\operatorname{Var}_{f\sim w_t(\cdot\mid G)}(f(X_s)).$$

Applying Lemma A.3 therefore gives:

$$\mathbb{E}_G \operatorname{Var}_{f\sim w_t(\cdot\mid G)}(f(X_s)) = -\mathbb{E}_G\left[G_s \tilde{f}_t(X_s)\right].$$

To bound all these correlations together, fix $\alpha \in \mathbb{R}^{t-1}$ and consider rescaling of summands by α:

$$\begin{aligned}&\sum_{s<t} \alpha_s \mathbb{E}_G \operatorname{Var}_{f\sim w_t(\cdot\mid G)}(f(X_s)) \\ &= \mathbb{E}_G \sum_{f\in\mathcal{G}} w_t(f \mid G) \sum_{s<t} \alpha_s(-G_s) f(X_s) \\ &\le \mathbb{E}_G \max_{f\in\mathcal{G}} \sum_{s<t} \alpha_s(-G_s) f(X_s).\end{aligned}$$

For each $f \in \mathcal{G}$, the sum inside the maximum is a centered Gaussian with variance $\sum_{s<t} \alpha_s^2 f(X_s)^2 \le \|\alpha\|_2^2$. Thus, for every $\lambda > 0$, Jensen's inequality and the Gaussian moment-generating function give

$$\begin{aligned}\mathbb{E}_G \max_{f\in\mathcal{G}} \sum_{s<t} \alpha_s(-G_s) f(X_s) &\le \frac{1}{\lambda} \ln \sum_{f\in\mathcal{G}} \mathbb{E}_G\, e^{\lambda \sum_{s<t} \alpha_s(-G_s) f(X_s)} \\ &\le \frac{\log|\mathcal{G}|}{\lambda} + \frac{\lambda\|\alpha\|_2^2}{2}.\end{aligned}$$

Taking the infimum over $\lambda > 0$ bounds the Gaussian maximum by $\sqrt{2\log|\mathcal{G}|}\|\alpha\|_2$. Thus,

$$\begin{aligned}\sqrt{\sum_{s<t}\left(\mathbb{E}_G \operatorname{Var}_{f\sim w_t(\cdot\mid G)}(f(X_s))\right)^2} &= \sup_{\|\alpha\|_2\le 1} \sum_{s<t} \alpha_s \mathbb{E}_G \operatorname{Var}_{f\sim w_t(\cdot\mid G)}(f(X_s)) \\ &\le \sqrt{2\log|\mathcal{G}|}.\end{aligned}$$

After squaring, this implies the desired bound. □

Lemma 4.5. *The regret of the exponential weights algorithm in (3) satisfies*

$$\mathbb{E}\operatorname{Reg}_T^{\mathrm{EW}} \lesssim \sqrt{\frac{dT}{\sigma}} \log^2(2T/\sigma). \tag{13}$$

Proof of Lemma 4.5. Apply Proposition 4.6 to the cover from Proposition 4.2. The difference in the comparator terms over $\mathcal{G}$ and $\mathcal{F}$ is upper bounded as:

$$0 \le \min_{g\in\mathcal{G}} \sum_{t=1}^{T} \ell(g(X_t), Y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$
$$\le 2 \sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \sum_{t=1}^{T} \mathbb{1}\{f(X_t) \neq g(X_t)\}.$$

By Proposition 4.2, the expectation of this difference is at most $Cd\log(2T/\sigma) + 2$. Thus, applying this estimate to (14), we have:

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{EW}} \lesssim \log|\mathcal{G}| + \mathbb{E}\sum_{t=1}^{T} \mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_t)) + Cd\log(2T/\sigma) + 2. \tag{33}$$

The above reduces the proof to bounding the cumulative prediction variance. We know from Lemma 4.7 that variance of $f \sim w_t$ on the past covariates X_s with $s < t$ can be suitably upper bounded. To upper bound the above, we will show that variance terms can be expressed as predictable test functions of a VC class, and apply the surprise lemma Lemma 4.1. Indeed, for every t and $x \in \mathcal{X}$, we have

$$\frac{1}{2}\,\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(x)) = \mathbb{E}_{G^{(t)}} \sum_{f,g\in\mathcal{G}} w_t(f \mid G^{(t)}) w_t(g \mid G^{(t)}) \mathbb{1}\{f(x) \neq g(x)\}.$$

This is a predictable test-function that belongs to the convex hull of the disagreement class:

$$\{x \mapsto \mathbb{1}\{f(x) \neq g(x)\} : f, g \in \mathcal{F}\}.$$

From Blumer et al. (1989), its VC dimension is in $O(d)$. Thus, using Lemma 4.7 and applying Lemma 4.1 to the sequence of variances gives

$$\sum_{t=1}^{T} \mathbb{E}\,\mathrm{Var}_{f\sim w_t(\cdot|G^{(t)})}(f(X_t))$$
$$\lesssim \sqrt{\frac{T\log|\mathcal{G}|\log(2T)}{\sigma}} + \sqrt{\frac{(d+1)T}{\sigma}}\log^2(2T/\sigma).$$

Plugging this into Eq. (33) yields:

$$\mathbb{E}\,\mathrm{Reg}_T^{\mathrm{EW}} \lesssim \log|\mathcal{G}| + \sqrt{\frac{T\log|\mathcal{G}|\log(2T)}{\sigma}} + \sqrt{\frac{(d+1)T}{\sigma}}\log^2(2T/\sigma) + d\log(2T/\sigma).$$

To obtain (13), we use Proposition 4.2 and note that $\mathrm{Reg}_T^{\mathrm{EW}} \le 2T$ with probability 1, and thus we can absorb the last term by $\sqrt{T \cdot d\log(2T/\sigma)}$. This concludes the proof. □

Theorem 3.1 follows by combining Lemma 4.3 and Lemma 4.5.

C Proofs for the real-valued loss case

The real-valued proof has the same three ingredients as the binary proof: a surprise lemma, a reduction to an exponential-weights reference algorithm that runs over a finite cover, and a regret bound for exponential weights.

First, we change slightly the way the reference exponential algorithm is defined. Let $\mathcal{G}$ be the cover from Proposition A.1 and, for each f, let g_f be the element of the cover that corresponds to f. One can consider a partition of $\mathcal{F}$ into cells where each cell contains all functions f that share a corresponding element g_f. Within each cell, we consider clipping of functions towards their corresponding cover center; in particular, for a fixed valued $\varepsilon < 1/16$, and for $f \in \mathcal{F}$, define

$$\bar{f}(x) = g_f(x) + \max\{-2\varepsilon, \min\{2\varepsilon, f(x) - g_f(x)\}\}.$$

Then, we modify exponential weights as follows. Each element $g \in \mathcal{G}$ is averaged w.r.t. using the best loss of a clipped function within the cell of g, formally, we let

$$\Phi_{t,g}(G) = \sup_{f\in\mathcal{F}:g_f=g} \left\{ -\sum_{s<t} \left[\ell(\bar{f}(X_s), Y_s) + G_s \bar{f}(X_s)\right] \right\}, \qquad \widetilde{\Phi}_t(G) = \ln \sum_{g\in\mathcal{G}} e^{\Phi_{t,g}(G)},$$

where $\ln$ denotes the natural logarithm, and define

$$w_t(g \mid G) = e^{\Phi_{t,g}(G) - \widetilde{\Phi}_t(G)}, \qquad \tilde{f}_t = \sum_{g\in\mathcal{G}} w_t(g \mid G)\, g. \tag{34}$$

Intuitively, averaging w.r.t. to clipped functions as opposed to the centers g accomplishes two goals. First, by selecting the minimizer within each cell, we improve the comparison between exponential weights and FTPL. Second, by introducing clipping, we ensure that the predictions of this modified exponential weights algorithm are not too far from those that use the cover centers $\mathcal{G}$ to compute averaging weights.

Formally, we establish the following guarantee for the clipping we use.

Lemma C.1. *Let $\mathcal{G}$ and the representatives g_f be as in part (ii) of Proposition A.1, and define $\bar{f}$ for $f \in \mathcal{F}$ as above.*

(i) For every $f \in \mathcal{F}$ and $x \in \mathcal{X}$, we have $\bar{f}(x) \in [-1, 1]$, $|\bar{f}(x) - g_f(x)| \le 2\varepsilon$, and

$$|\bar{f}(x) - f(x)| \le 2\mathbb{1}\{|f(x) - g_f(x)| > 2\varepsilon\}.$$

(ii) With probability at least $1 - (8T)^{-4}$,

$$\sup_{f\in\mathcal{F}} \sum_{s=1}^{T} |f(X_s) - \bar{f}(X_s)| \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon}.$$

Proof. For part (i), $\bar{f} - g_f$ is $f - g_f$ clipped to $[-2\varepsilon, 2\varepsilon]$. Hence $\bar{f}(x)$ lies between $g_f(x)$ and $f(x)$, so it belongs to $[-1, 1]$ and satisfies $|\bar{f}(x) - f(x)| \le |f(x) - g_f(x)| \le 2$. It equals $f(x)$ whenever $|f(x) - g_f(x)| \le 2\varepsilon$. For part (ii), part (i) gives $|f(X_s) - \bar{f}(X_s)| \le 2\mathbb{1}\{|f(X_s) - g_f(X_s)| > 2\varepsilon\}$. Summing over $s \le T$ and applying (21) proves the claim. □

C.1 Finite-scale surprise bounds

We begin by the surprise bound for real-valued classes from its binary counterpart.

Lemma C.2. *Let $0 < \varepsilon \le 1/16$, and let $\{h_t\}_{t\in[T]}$ be a predictable sequence of functions, each of which is a convex combination or a probability mixture of elements of $\pm\mathcal{F}$. Then*

$$\mathbb{E}\sum_{t=1}^{T} |h_t(X_t)| \lesssim T\varepsilon\sqrt{\frac{\log(2T)}{\sigma}} + \sqrt{\frac{\log(2T)}{\sigma}\,\mathbb{E}\sum_{t=1}^{T}\sum_{s<t} h_t(X_s)^2} + \sqrt{\frac{(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))\,T}{\sigma}} \log^3 \frac{CT}{\sigma\varepsilon}. \tag{35}$$

Compared with Lemma 4.1, the VC dimension is replaced by the fat-shattering dimension at scale $c\varepsilon$, at the price of the additive term $T\varepsilon\sqrt{\log(2T)/\sigma}$. This term pays for replacing each function by its representative, which is within 2ε of it except at a few covariates.

Proof of Lemma C.2. Throughout this proof, suprema over h range over convex combinations and probability mixtures of elements of $\pm\mathcal{F}$. In the proof, we first establish the following analogue of Lemma B.2 for second moments truncated at 2ε:

$$\mathbb{E}\sup_h \left\{ \sum_{s=1}^{T} \mathbb{E}_{\bar{X}_s}(|h(\bar{X}_s)| - 2\varepsilon)_+^2 - 12\sum_{s=1}^{T} h(X_s)^2 \right\}$$

$$\lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^4 \frac{CT}{\sigma\varepsilon}. \tag{36}$$

As in the proof of Lemma B.2, we reduce this claim to comparison for binary function class using layer-cake identity for expectations and Maurey's empirical method (Proposition A.4), and conclude it by invoking Block et al. (2024b, Theorem 10). Then, the proof is concluded by combining (36) with Lemma B.1, as in the proof of Lemma 4.1.

To start, using the layer-cake identity, we have

$$\sum_{s=1}^{T} \mathbb{E}_{\bar{X}_s}(|h(\bar{X}_s)| - 2\varepsilon)_+^2 = \int_{2\varepsilon}^{1} 2(\lambda - 2\varepsilon) \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \geq \lambda] \mathrm{d}\lambda.$$

Also, note that we have:

$$\sum_{s=1}^{T} h(X_s)^2 \geq \int_{2\varepsilon}^{1} \frac{\lambda - 2\varepsilon}{2} \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \geq \lambda/2\} \mathrm{d}\lambda.$$

Thus,

$$\sum_{s=1}^{T} \mathbb{E}_{\bar{X}_s}(|h(\bar{X}_s)| - 2\varepsilon)_+^2 - 12 \sum_{s=1}^{T} h(X_s)^2$$
$$\leq \int_{2\varepsilon}^{1} 2(\lambda - 2\varepsilon) \left\{ \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \geq \lambda] - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \geq \lambda/2\} \right\} \mathrm{d}\lambda. \tag{37}$$

Set $m = \lceil CT \log(2T)/\sigma \rceil$, let $Z_{1:m} \sim \mu$ and consider a coupling Block et al. (2024b, Lemma 25) such that

$$X_{1:T}, \bar{X}_{1:T} \subset Z_{1:m}$$

with probability at least $1 - 1/T$. For $2\varepsilon \leq \lambda \leq 1$, let $\mathcal{G}_\lambda \subseteq \mathcal{F}$ be a $\lambda/32$-cover of $\mathcal{F}$ on $Z_{1:m}$ given by part (i) of Proposition A.1, so that every $f \in \mathcal{F}$ has a representative $g \in \mathcal{G}_\lambda$ with $\max_{s \leq m} |f(Z_s) - g(Z_s)| \leq \lambda/32$. Take $K = \lceil C\lambda^{-2} \log(2m) \rceil$, and define

$$\mathcal{F}_\lambda = \left\{ x \mapsto \mathbb{1}\left\{ \left| \frac{1}{K} \sum_{i=1}^{K} f_i(x) \right| \geq \frac{3\lambda}{4} \right\} : f_1, \ldots, f_K \in \pm\mathcal{G}_\lambda \right\},$$

to be the class obtained by averaging K functions from $\pm\mathcal{G}_\lambda$. Applying Proposition A.4 to the m points $Z_{1:m}$, for C large enough and every $h \in \mathrm{conv}(\pm\mathcal{F})$ there exist $f_1, \ldots, f_K \in \pm\mathcal{F}$ such that

$$\sup_{i \in [m]} \left| h(Z_i) - \frac{1}{K} \sum_{j=1}^{K} f_j(Z_i) \right| \leq \lambda/16.$$

Since $|f_i - g_i| \leq \lambda/32$ on $Z_{1:m}$, the average $\bar{g} = \frac{1}{K} \sum_{i=1}^{K} g_i$ satisfies $|h - \bar{g}| \leq 3\lambda/32$ on $Z_{1:m}$. Consequently, $g = \mathbb{1}\{|u| \geq 3\lambda/4\} \in \mathcal{F}_\lambda$ satisfies

$$\mathbb{1}\{|h(x)| \geq \lambda\} \leq g(x) \leq \mathbb{1}\{|h(x)| \geq \lambda/2\} \tag{38}$$

for every $x \in \{Z_1, \ldots, Z_m\}$. Applying Block et al. (2024b, Theorem 10) to $\mathcal{F}_\lambda$ and upper bounding Will's functional by a growth function of the class, we get

$$\mathbb{E} \sup_h \left\{ \sum_{s=1}^{T} \mathbb{P}[|h(\bar{X}_s)| \geq \lambda] - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \geq \lambda/2\} \right\}$$
$$\leq \mathbb{E} \sup_h \left\{ \sum_{s=1}^{T} \mathbb{1}\{|h(\bar{X}_s)| \geq \lambda\} - 3 \sum_{s=1}^{T} \mathbb{1}\{|h(X_s)| \geq \lambda/2\} \right\}$$
$$\leq 1 + \mathbb{E} \sup_{g \in \mathcal{F}_\lambda} \sum_{s=1}^{T} \big(g(\bar{X}_s) - 3g(X_s) \big)$$

$$\lesssim 1 + \sup_{Z_{1:m}} \log |\mathcal{F}_\lambda|. \tag{39}$$

The third line uses the sandwich bound Eq. (38) on the event that $X_{1:T}, \bar{X}_{1:T} \subset Z_{1:m}$ and upper bounds by worst case bound T on the low probability $1/T$ complement.

We now bound the cardinality of $\mathcal{F}_\lambda$. Note that, on any m points $Z_1, \dots, Z_m$, by (19), we have

$$\log |\mathcal{G}_\lambda| \lesssim (1 + \mathrm{fat}_{c\lambda}(\mathcal{F})) \log^2(Cm/\lambda).$$

Thus, via a simple counting argument, we have:

$$|\mathcal{F}_\lambda| \le (2|\mathcal{G}_\lambda|)^K.$$

Thus, we have

$$1 + \sup_{Z_{1:m}} \log |\mathcal{F}_\lambda| \lesssim K(1 + \mathrm{fat}_{c\lambda}(\mathcal{F})) \log^2(Cm/\lambda) \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))\lambda^{-2} \log^3 \frac{CT}{\sigma\lambda},$$

where the last step uses $\log(2m) \lesssim \log(CT/\sigma)$ and $\mathrm{fat}_{c\lambda}(\mathcal{F}) \le \mathrm{fat}_{c\varepsilon}(\mathcal{F})$ for $\lambda \ge \varepsilon$. Plugging this into Eqs. (37) and (39) gives

$$\begin{aligned}
&\mathbb{E} \sup_h \left\{ \sum_{s=1}^T \mathbb{E}_{\bar{X}_s} (|h(\bar{X}_s)| - 2\varepsilon)_+^2 - 12 \sum_{s=1}^T h(X_s)^2 \right\} \\
&\lesssim 1 + \int_{2\varepsilon}^1 2(\lambda - 2\varepsilon)(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))\lambda^{-2} \log^3 \frac{CT}{\sigma\lambda} \mathrm{d}\lambda \\
&\lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^4 \frac{CT}{\sigma\varepsilon},
\end{aligned}$$

where the last step uses $2(\lambda - 2\varepsilon)\lambda^{-2} \le 2/\lambda$ and $\int_{2\varepsilon}^1 \lambda^{-1} \mathrm{d}\lambda \le \log(1/\varepsilon)$. This proves (36).

It remains to combine (36) with Lemma B.1. Apply Lemma B.1 to $(|h_t| - 2\varepsilon)_+$ and the uniform bound (36) to h_t on each prefix $X_{1:t-1}$. Since $|h_t| \le (|h_t| - 2\varepsilon)_+ + 2\varepsilon$, this gives

$$\begin{aligned}
\mathbb{E} \sum_{t=1}^T |h_t(X_t)| &\le 2T\varepsilon + \mathbb{E} \sum_{t=1}^T (|h_t(X_t)| - 2\varepsilon)_+ \\
&\lesssim T\varepsilon + \frac{\log(2T)}{\sigma} \\
&\quad + \left[\frac{\log(2T)}{\sigma} \left(\mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2 + (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))T \log^4 \frac{CT}{\sigma\varepsilon} \right) \right]^{1/2} \\
&\lesssim T\varepsilon + \sqrt{\frac{\log(2T)}{\sigma} \mathbb{E} \sum_{t=1}^T \sum_{s<t} h_t(X_s)^2} \\
&\quad + \sqrt{\frac{(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))T}{\sigma}} \log^{5/2} \frac{CT}{\sigma\varepsilon} + \frac{\log(2T)}{\sigma},
\end{aligned}$$

where the last step uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\log(2T) \le \log(CT/(\sigma\varepsilon))$. This proves (35). □

C.2 Proof of the prediction comparison on the past

We begin by proving that predictions of FTPL Eq. (6) and the reference from Eq. (34) in round t are close on past covariates X_s with $s < t$.

Lemma C.3. *For every $t \le T$,*

$$\mathbb{E} \sum_{s<t} \left(\mathbb{E}_{G^{(t)}} \hat{f}_t(X_s) - \mathbb{E}_{G^{(t)}} \tilde{f}_t(X_s) \right)^2 \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))^2 \log^5 \frac{CT}{\sigma\varepsilon} + T\varepsilon^2 + 1. \tag{40}$$

Proof of Lemma C.3. Fix the history $(X_{1:t-1}, Y_{1:t-1})$ and let $G := G^{(t)}$ to lighten the notational load. The case $t = 1$ is immediate, so suppose that $t \geq 2$. Use the same potentials as in the proof of Lemma 4.4, with the clipped reference:

$$\begin{aligned}\Phi_t(G) &= \sup_{f \in \mathcal{F}} \left\{ -\sum_{s<t} [\ell(f(X_s), Y_s) + G_s f(X_s)] \right\}, \\ \widetilde{\Phi}_t(G) &= \ln \sum_{g \in \mathcal{G}} e^{\Phi_{t,g}(G)} \\ &= \max_{w \in \Delta(\mathcal{G})} \left[\mathbb{E}_{g \sim w} \, \Phi_{t,g}(G) + H(w) \right].\end{aligned}$$

$w_t(\cdot \mid G)$ from Eq. (34) maximize the latter potential in its variational form. Note that, by absolute continuity of the Gaussians, the set of all functions that maximize the former potential share predictions on the past covariates X_s. Differentiating the two potentials therefore gives, for every $s < t$ and almost every G, by Danskin's theorem (Bertsekas, 2016, Appendix B) and the chain rule:

$$\partial_{G_s} \Phi_t(G) = -\hat{f}_t(X_s), \qquad \partial_{G_s} \widetilde{\Phi}_t(G) = \sum_{g \in \mathcal{G}} w_t(g \mid G) \partial_{G_s} \Phi_{t,g}(G).$$

Each $\Phi_{t,g}$ is a supremum of affine functions whose slopes in G_s are within 2ε of $-g(X_s)$. Hence, almost everywhere,

$$|\partial_{G_s} \Phi_{t,g}(G) + g(X_s)| \leq 2\varepsilon, \qquad \left| \partial_{G_s} \widetilde{\Phi}_t(G) + \tilde{f}_t(X_s) \right| \leq 2\varepsilon. \tag{41}$$

Both potentials are Lipschitz functions of G. For such potentials, the argument leading to (29) in the proof of Lemma 4.4, which combines Stein's lemma (Lemma A.3) with Cauchy–Schwarz, shows that

$$\left\| \mathbb{E}_G \, \nabla_G (\Phi_t(G) - \widetilde{\Phi}_t(G)) \right\|_2^2 \leq \mathbb{E}_G (\Phi_t - \widetilde{\Phi}_t)^2.$$

Together with (41) and $(a+b)^2 \leq 2a^2 + 2b^2$ in each coordinate, this yields

$$\sum_{s<t} \left(\mathbb{E}_G \, \hat{f}_t(X_s) - \mathbb{E}_G \, \tilde{f}_t(X_s) \right)^2 \leq 2 \, \mathbb{E}_G (\Phi_t - \widetilde{\Phi}_t)^2 + 8(t-1)\varepsilon^2. \tag{42}$$

It remains to bound the difference of the potentials. Since the entropy of a distribution on $\mathcal{G}$ lies in $[0, \ln |\mathcal{G}|]$, the variational form of $\widetilde{\Phi}_t$ gives

$$\max_{g \in \mathcal{G}} \Phi_{t,g}(G) \leq \widetilde{\Phi}_t(G) \leq \max_{g \in \mathcal{G}} \Phi_{t,g}(G) + \ln |\mathcal{G}|.$$

The cells $\{f \in \mathcal{F} : g_{\underline{f}} = g\}$ partition $\mathcal{F}$, so $\max_{g \in \mathcal{G}} \Phi_{t,g}(G)$ is the supremum defining $\Phi_t(G)$ with each f replaced by $\underline{f}$. For each $s < t$, the map $y \mapsto \ell(y, Y_s) + G_s u$ is $(1 + |G_s|)$-Lipschitz on $[-1, 1]$. Hence, on the event of Lemma C.1(ii), which has probability at least $1 - (8T)^{-4}$ over the history, using (20) to bound $\ln |\mathcal{G}|$,

$$\begin{aligned}|\Phi_t(G) - \widetilde{\Phi}_t(G)| &\leq \ln |\mathcal{G}| + \left(1 + \max_{s<t} |G_s| \right) \sup_{f \in \mathcal{F}} \sum_{s<t} |f(X_s) - \bar{f}(X_s)| \\ &\lesssim \left(1 + \max_{s<t} |G_s| \right) (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F})) \log^2 \frac{CT}{\sigma\varepsilon}.\end{aligned}$$

Since $\mathbb{E}_G (1 + \max_{s<t} |G_s|)^2 \lesssim \log(2T)$, it follows that

$$\mathbb{E}_G (\Phi_t - \widetilde{\Phi}_t)^2 \lesssim (1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}))^2 \log^5 \frac{CT}{\sigma\varepsilon}.$$

Plugging this into Eq. (42) gives the claimed bound on the above event. The complement of this high probability event contributes at most $4T(8T)^{-4} \leq 1$. This concludes the proof. □

C.3 Proof of the reduction to exponential weights

We now apply the surprise lemma Lemma C.2 to the difference of predictions.

Lemma C.4. *For the reference defined in Eq. (34), we have the following bound:*

$$\mathbb{E}\sum_{t=1}^{T}\left|\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t)-\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)\right| \lesssim \left[(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\sqrt{\frac{T}{\sigma}}+\frac{T\varepsilon}{\sqrt{\sigma}}\right]\log^3\frac{CT}{\sigma\varepsilon}+\frac{\log(2T)}{\sigma}. \quad (43)$$

Proof of Lemma C.4. Apply Lemma C.2 to the sequence

$$h_t=\frac{1}{2}\left(\mathbb{E}_{G^{(t)}}\,\hat{f}_t-\mathbb{E}_{G^{(t)}}\,\tilde{f}_t\right),\qquad t\in[T].$$

Note that both $\mathbb{E}_{G^{(t)}}\,\hat{f}_t$ and $\mathbb{E}_{G^{(t)}}\,\tilde{f}_t$ are probability mixtures of elements of $\mathcal{F}$ and are predictable. Thus, h_t is a probability mixture of elements of $\pm\mathcal{F}$, and using Lemma C.2 gives

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\left|\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t)-\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)\right| \\
&\lesssim T\varepsilon\sqrt{\frac{\log(2T)}{\sigma}}+\left(\frac{\log(2T)}{\sigma}\,\mathbb{E}\sum_{t=1}^{T}\sum_{s<t}\left(\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_s)-\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_s)\right)^2\right)^{1/2} \\
&\quad+\sqrt{\frac{(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))T}{\sigma}}\log^3\frac{CT}{\sigma\varepsilon}.
\end{aligned}$$

By Lemma C.3, the expected double sum is in $O(T(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))^2\log^5(CT/(\sigma\varepsilon))+T^2\varepsilon^2+T)$. Substituting this bound and using $\log(2T)\le\log(CT/(\sigma\varepsilon))$ gives

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\left|\mathbb{E}_{G^{(t)}}\,\hat{f}_t(X_t)-\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t)\right| \\
&\lesssim\left[(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\sqrt{\frac{T}{\sigma}}+\frac{T\varepsilon}{\sqrt{\sigma}}\right]\log^3\frac{CT}{\sigma\varepsilon}+\frac{\log(2T)}{\sigma}.
\end{aligned}$$

This proves (43). □

C.4 Regret of Gaussian-perturbed exponential weights

We first prove Proposition C.5, which reduces the regret bound to controlling cumulative prediction variance. We then prove Lemma C.6, which bounds the variances on the historical covariates, and combine it with the surprise lemma to prove Lemma C.7.

Proposition C.5. *For every data sequence, the predictor $\tilde{f}_t$ defined in Eq. (34) satisfies the following inequality,*

$$\begin{aligned}
&\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t),Y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t),Y_t) \\
&\lesssim\log|\mathcal{G}|+T\varepsilon+\sum_{t=1}^{T}\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{g\sim w_t(\cdot|G^{(t)})}(g(X_t)).
\end{aligned} \quad (44)$$

Proof of Proposition C.5. We follow the proof of Proposition B.4, with the log-partition potential from Eq. (34)

$$\widetilde{\Phi}_t(G^{(t)})=\ln\sum_{g\in\mathcal{G}}e^{\Phi_{t,g}(G^{(t)})},\qquad 1\le t\le T+1.$$

For a fixed round t, we have the following identity:

$$\widetilde{\Phi}_{t+1}(G^{(t+1)})-\widetilde{\Phi}_t(G^{(t+1)}_{1:t-1})$$

$$= \ln \sum_{g \in \mathcal{G}} w_t(g \mid G_{1:t-1}^{(t+1)}) \exp\left(\Phi_{t+1,g}(G^{(t+1)}) - \Phi_{t,g}(G_{1:t-1}^{(t+1)})\right).$$

By Lemma C.1(i), the score increments satisfy

$$\left|\Phi_{t,g}(G_{1:t-1}^{(t+1)}) - \Phi_{t+1,g}(G^{(t+1)}) - \ell(g(X_t), Y_t) - G_t^{(t+1)} g(X_t)\right| \le 2\varepsilon(1 + |G_t^{(t+1)}|).$$

Consequently,

$$\begin{aligned}
&\widetilde{\Phi}_{t+1}(G^{(t+1)}) - \widetilde{\Phi}_t(G_{1:t-1}^{(t+1)}) \\
&\le \ln \sum_{g \in \mathcal{G}} w_t(g \mid G_{1:t-1}^{(t+1)}) e^{-\ell(g(X_t), Y_t) - G_t^{(t+1)} g(X_t)} + 2\varepsilon(1 + |G_t^{(t+1)}|).
\end{aligned}$$

Apply Proposition A.2 to these weights and the losses $\ell(g(X_t), Y_t) + G_t^{(t+1)} g(X_t)$ at learning rate 1. Their range is at most $s_t := 2(1 + |G_t^{(t+1)}|)$, and

$$\begin{aligned}
&\mathrm{Var}_{g \sim w_t(\cdot \mid G_{1:t-1}^{(t+1)})}(\ell(g(X_t), Y_t) + G_t^{(t+1)} g(X_t)) \\
&\le (1 + |G_t^{(t+1)}|)^2 \, \mathrm{Var}_{g \sim w_t(\cdot \mid G_{1:t-1}^{(t+1)})}(g(X_t)),
\end{aligned}$$

since $u \mapsto \ell(u, Y_t) + G_t^{(t+1)} u$ is $(1 + |G_t^{(t+1)}|)$-Lipschitz. Thus, cancelling s_t^2 from numerator and the denominator in Proposition A.2, we have

$$\begin{aligned}
&\sum_{g \in \mathcal{G}} w_t(g \mid G_{1:t-1}^{(t+1)})[\ell(g(X_t), Y_t) + G_t^{(t+1)} g(X_t)] \\
&\quad \le \widetilde{\Phi}_t(G_{1:t-1}^{(t+1)}) - \widetilde{\Phi}_{t+1}(G^{(t+1)}) + 2\varepsilon(1 + |G_t^{(t+1)}|) \\
&\qquad + \frac{e^{2(1+|G_t^{(t+1)}|)} - 2(1 + |G_t^{(t+1)}|) - 1}{4} \mathrm{Var}_{g \sim w_t(\cdot \mid G_{1:t-1}^{(t+1)})}(g(X_t)).
\end{aligned}$$

Note that the last coordinate $G_t^{(t+1)}$ is centered and independent of the first $t-1$ coordinates. In particular,

$$\begin{aligned}
&\mathbb{E}_{G_t^{(t+1)}} \frac{e^{2(1+|G_t^{(t+1)}|)} - 2(1 + |G_t^{(t+1)}|) - 1}{4} \\
&\quad \le \frac{e^2}{4} \mathbb{E}_{G_t^{(t+1)}} \left(e^{2G_t^{(t+1)}} + e^{-2G_t^{(t+1)}}\right) = \frac{e^4}{2}.
\end{aligned}$$

Now, $G_{1:t-1}^{(t+1)}$ has the same law as $G^{(t)}$ conditionally on the history up to round t. Thus, using convexity of the loss and $\mathbb{E}\,|G_t^{(t+1)}| \le 1$,

$$\begin{aligned}
\ell(\mathbb{E}_{G^{(t)}} \tilde{f}_t(X_t), Y_t) &\le \mathbb{E}_{G^{(t)}} \widetilde{\Phi}_t(G^{(t)}) - \mathbb{E}_{G^{(t+1)}} \widetilde{\Phi}_{t+1}(G^{(t+1)}) \\
&\quad + 4\varepsilon + \frac{e^4}{2} \mathbb{E}_{G^{(t)}} \mathrm{Var}_{g \sim w_t(\cdot \mid G^{(t)})}(g(X_t)).
\end{aligned} \tag{45}$$

Note that we have $\widetilde{\Phi}_1 = \ln |\mathcal{G}|$, and

$$\begin{aligned}
\mathbb{E}_{G^{(T+1)}} \widetilde{\Phi}_{T+1}(G^{(T+1)}) &\ge \sup_{f \in \mathcal{F}} \mathbb{E}_{G^{(T+1)}} \left[-\sum_{s=1}^{T} \left(\ell(\bar{f}(X_s), Y_s) + G_s^{(T+1)} \bar{f}(X_s)\right)\right] \\
&= -\inf_{f \in \mathcal{F}} \sum_{s=1}^{T} \ell(\bar{f}(X_s), Y_s),
\end{aligned}$$

by Jensen's inequality. Combining this with Eq. (45) and telescoping gives the desired result. □

Lemma C.6. *For every $t \le T$ and every fixed history $(X_s, Y_s)_{s<t}$, the weights from Eq. (34) satisfy*

$$\sum_{s<t} \left(\mathbb{E}_{G^{(t)}} \mathrm{Var}_{g \sim w_t(\cdot \mid G^{(t)})}(g(X_s))\right)^2 \lesssim \log(e|\mathcal{G}|) + T\varepsilon^2. \tag{46}$$

Proof of Lemma C.6. Write $G = G^{(t)}$ throughout this proof for notational brevity. Differentiating each weight w_t from Eq. (34) w.r.t. a Gaussian perturbation G_s, we obtain:

$$\begin{aligned}\partial_{G_s} w_t(g \mid G) &= \partial_{G_s} \frac{e^{\Phi_{t,g}(G)}}{\sum_{g' \in \mathcal{G}} e^{\Phi_{t,g'}(G)}} \\ &= w_t(g \mid G)\partial_{G_s}\Phi_{t,g}(G) - w_t(g \mid G)\sum_{g' \in \mathcal{G}} w_t(g' \mid G)\partial_{G_s}\Phi_{t,g'}(G) \\ &= w_t(g \mid G)\left(\partial_{G_s}\Phi_{t,g}(G) - \sum_{g' \in \mathcal{G}} w_t(g' \mid G)\partial_{G_s}\Phi_{t,g'}(G)\right).\end{aligned}$$

Multiplying by $g(X_s)$ and summing over $g \in \mathcal{G}$ gives

$$\begin{aligned}\partial_{G_r}\tilde{f}_t(X_s) &= \sum_{g \in \mathcal{G}} w_t(g \mid G)g(X_s)\partial_{G_r}\Phi_{t,g}(G) - \tilde{f}_t(X_s)\sum_{g \in \mathcal{G}} w_t(g \mid G)\partial_{G_r}\Phi_{t,g}(G) \\ &= -\operatorname{Cov}_{g \sim w_t(\cdot \mid G)}(g(X_s), g(X_r)) \\ &\quad + \operatorname{Cov}_{g \sim w_t(\cdot \mid G)}(g(X_s), \partial_{G_r}\Phi_{t,g}(G) + g(X_r)),\end{aligned}$$

and in particular, taking $r = s$,

$$\left|\partial_{G_s}\tilde{f}_t(X_s) + \operatorname{Var}_{g \sim w_t(\cdot \mid G)}(g(X_s))\right| \le 2\varepsilon.$$

Here the covariance error is bounded by (41) and Cauchy–Schwarz. All derivatives hold almost everywhere, and $\tilde{f}_t(X_s)$ is Lipschitz in G. Applying Lemma A.3 therefore gives:

$$\mathbb{E}_G \operatorname{Var}_{g \sim w_t(\cdot \mid G)}(g(X_s)) \le -\mathbb{E}_G\left[G_s\tilde{f}_t(X_s)\right] + 2\varepsilon.$$

To bound all these correlations together, fix a nonnegative $\alpha \in \mathbb{R}^{t-1}$ and consider rescaling of summands by α:

$$\begin{aligned}&\sum_{s<t} \alpha_s \,\mathbb{E}_G \operatorname{Var}_{g \sim w_t(\cdot \mid G)}(g(X_s)) \\ &\le \mathbb{E}_G \sum_{g \in \mathcal{G}} w_t(g \mid G)\sum_{s<t} \alpha_s(-G_s)g(X_s) + 2\varepsilon\sum_{s<t}\alpha_s \\ &\le \mathbb{E}_G \max_{g \in \mathcal{G}} \sum_{s<t} \alpha_s(-G_s)g(X_s) + 2\varepsilon\sqrt{t-1}\|\alpha\|_2.\end{aligned}$$

For each $g \in \mathcal{G}$, the sum inside the maximum is a centered Gaussian with variance $\sum_{s<t}\alpha_s^2 g(X_s)^2 \le \|\alpha\|_2^2$. Thus, for every $\lambda > 0$, Jensen's inequality and the Gaussian moment-generating function give

$$\begin{aligned}\mathbb{E}_G \max_{g \in \mathcal{G}} \sum_{s<t} \alpha_s(-G_s)g(X_s) &\le \frac{1}{\lambda}\ln \sum_{g \in \mathcal{G}} \mathbb{E}_G\, e^{\lambda \sum_{s<t}\alpha_s(-G_s)g(X_s)} \\ &\le \frac{\log|\mathcal{G}|}{\lambda} + \frac{\lambda\|\alpha\|_2^2}{2}.\end{aligned}$$

Taking the infimum over $\lambda > 0$ bounds the Gaussian maximum by $\sqrt{2\log|\mathcal{G}|}\|\alpha\|_2$. Thus,

$$\begin{aligned}&\sqrt{\sum_{s<t}\left(\mathbb{E}_G \operatorname{Var}_{g \sim w_t(\cdot \mid G)}(g(X_s))\right)^2} \\ &= \sup_{\substack{\alpha \in [0,\infty)^{t-1} \\ \|\alpha\|_2 \le 1}} \sum_{s<t} \alpha_s \,\mathbb{E}_G \operatorname{Var}_{g \sim w_t(\cdot \mid G)}(g(X_s)) \\ &\le \sqrt{2\log|\mathcal{G}|} + 2\varepsilon\sqrt{t-1}.\end{aligned}$$

After squaring, this implies the desired bound. □

Finally, combining the preceding bounds, we obtain the following result.

Lemma C.7. *The regret of the modified exponential weights algorithm from Eq. (34) satisfies*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t),Y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t),Y_t)\right]\\
&\lesssim(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon}+T\varepsilon\\
&\quad+\sqrt{\frac{T}{\sigma}}\left[\sqrt{\log(e|\mathcal{G}|)}\log^{5/2}\frac{2T}{\sigma}+\sqrt{T\log(2T)}\varepsilon\right]+\frac{\log(2T)}{\sigma}.
\end{aligned}\tag{47}$$

Proof of Lemma C.7. Apply Proposition C.5 to the cover from Proposition A.1. The difference in the comparator terms over the clipped functions and $\mathcal{F}$ is upper bounded as:

$$\begin{aligned}
&\left|\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(\bar{f}(X_t),Y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t),Y_t)\right|\\
&\qquad\le\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}|\bar{f}(X_t)-f(X_t)|,
\end{aligned}$$

since $\ell(\cdot,Y_t)$ is 1-Lipschitz. The right-hand side is at most $2T$, so, by Lemma C.1(ii), its expectation is at most $C(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2(CT/(\sigma\varepsilon))+2T(8T)^{-4}$. Thus, applying this estimate to (44), we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t),Y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t),Y_t)\right]\\
&\lesssim\log|\mathcal{G}|+T\varepsilon+\mathbb{E}\sum_{t=1}^{T}\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{g\sim w_t(\cdot|G^{(t)})}(g(X_t))\\
&\qquad+(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon}.
\end{aligned}\tag{48}$$

The above reduces the proof to bounding the cumulative prediction variance. We know from Lemma C.6 that the variance of $g(X_s)$ under $g\sim w_t$ on the past covariates X_s, $s<t$, can be suitably upper bounded. To upper bound the above, we will show that variance terms can be expressed as predictable test functions in the convex hull of a finite class, and apply Corollary B.3. Indeed, for every t and $x\in\mathcal{X}$, we have

$$\frac{1}{2}\,\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{g\sim w_t(\cdot|G^{(t)})}(g(x))=\mathbb{E}_{G^{(t)}}\sum_{g,g'\in\mathcal{G}}w_t(g\mid G^{(t)})w_t(g'\mid G^{(t)})\frac{(g(x)-g'(x))^2}{4}.$$

This is a predictable test-function that belongs to the convex hull of the squared-difference class:

$$\left\{x\mapsto(g(x)-g'(x))^2/4:g,g'\in\mathcal{G}\right\}.$$

It has at most $|\mathcal{G}|^2$ elements. If $|\mathcal{G}|=1$, all variances vanish; otherwise, using Lemma C.6 and applying Corollary B.3 to the sequence of variances, with the finite class from above in place of $\mathcal{F}$, gives

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\mathbb{E}_{G^{(t)}}\,\mathrm{Var}_{g\sim w_t(\cdot|G^{(t)})}(g(X_t))\\
&\lesssim\sqrt{\frac{T\log(2T)}{\sigma}\left(\log(e|\mathcal{G}|)+T\varepsilon^2\right)}+\sqrt{\frac{T\log|\mathcal{G}|}{\sigma}}\log^{5/2}\frac{2T}{\sigma}+\frac{\log(2T)}{\sigma}.
\end{aligned}\tag{49}$$

Plugging this into Eq. (48) yields:

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\,\tilde{f}_t(X_t),Y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t),Y_t)\right]$$

$$\lesssim \log|\mathcal{G}| + T\varepsilon + \sqrt{\frac{T\log(2T)}{\sigma}\left(\log(e|\mathcal{G}|) + T\varepsilon^2\right)}$$
$$+ \sqrt{\frac{T\log|\mathcal{G}|}{\sigma}}\log^{5/2}\frac{2T}{\sigma} + \frac{\log(2T)}{\sigma} + (1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon}.$$

To obtain (47), we use Proposition A.1 to bound $\log|\mathcal{G}|$ and use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. This concludes the proof. □

C.5 Proof of the real-valued theorem

The preceding bounds control the Gaussian mean of FTPL. It remains to account for the finite average in (8) and combine the prediction and reference-regret bounds.

Theorem 3.2. *Consider a class $\mathcal{F}\colon \mathcal{X} \to [-1,1]$ and let $\ell(\hat{y}, y)$ be convex and 1-Lipschitz in its first argument, and suppose the covariate process is σ-smooth with respect to a fixed μ. Then, the learner in (8) satisfies*

$$\mathbb{E}\,\mathrm{Reg}_T \lesssim \inf_{0<\varepsilon\le 1/16}\left\{\left(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F})+\sqrt{T}\varepsilon\right)\sqrt{\frac{T}{\sigma}}\log^{7/2}\left(2T/\sigma\varepsilon\right)\right\}, \tag{9}$$

where c is a universal constant.

Proof of Theorem 3.2. Fix $0 < \varepsilon \le 1/16$ with $\mathrm{fat}_{c\varepsilon}(\mathcal{F}) < \infty$. If $1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}) > T$ or $T < 1/\sigma$, then $2T$ is at most a constant times the expression inside the infimum in (9), so the bound at this ε follows from $\mathrm{Reg}_T \le 2T$. Otherwise, $1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}) \le T$ and $T \ge 1/\sigma$, and we construct the fixed cover and the modified exponential weights algorithm from Eq. (34).

Conditional on $\mathcal{H}_{t-1}, X_t, Y_t$, the average of the t independent oracle predictions has variance at most $1/t$ about $\mathbb{E}_{G^{(t)}}\hat{f}_t(X_t)$. Cauchy–Schwarz and Lipschitzness give

$$\mathbb{E}\left[\ell(\hat{Y}_t, Y_t) - \ell(\mathbb{E}_{G^{(t)}}\hat{f}_t(X_t), Y_t) \mid \mathcal{H}_{t-1}, X_t, Y_t\right] \le t^{-1/2}.$$

Summing over t and using $\sum_{t\le T} t^{-1/2} \le 2\sqrt{T}$, we obtain

$$\mathbb{E}\,\mathrm{Reg}_T \le 2\sqrt{T} + \mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\hat{f}_t(X_t), Y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t), Y_t)\right]. \tag{50}$$

We decompose the regret of the Gaussian mean on the right-hand side as

$$\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\hat{f}_t(X_t), Y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t), Y_t)$$
$$= \sum_{t=1}^{T}\left[\ell(\mathbb{E}_{G^{(t)}}\hat{f}_t(X_t), Y_t) - \ell(\mathbb{E}_{G^{(t)}}\tilde{f}_t(X_t), Y_t)\right]$$
$$+ \left[\sum_{t=1}^{T}\ell(\mathbb{E}_{G^{(t)}}\tilde{f}_t(X_t), Y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(X_t), Y_t)\right].$$

Since the loss is 1-Lipschitz, the first sum is bounded by Lemma C.4 in expectation. Lemma C.7 bounds the expectation of the second bracket. Plugging these bounds into (50), and using (20) to bound $\sqrt{\log(e|\mathcal{G}|)} \lesssim \sqrt{1+\mathrm{fat}_{c\varepsilon}(\mathcal{F})}\log(CT/(\sigma\varepsilon))$ and $\log(2T/\sigma) \le \log(CT/(\sigma\varepsilon))$, yields

$$\mathbb{E}\,\mathrm{Reg}_T \lesssim \sqrt{T} + \left[(1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\sqrt{\frac{T}{\sigma}} + \frac{T\varepsilon}{\sqrt{\sigma}}\right]\log^3\frac{CT}{\sigma\varepsilon} + \frac{\log(2T)}{\sigma}$$
$$+ (1+\mathrm{fat}_{c\varepsilon}(\mathcal{F}))\log^2\frac{CT}{\sigma\varepsilon} + T\varepsilon$$
$$+ \sqrt{\frac{T}{\sigma}}\left[\sqrt{1+\mathrm{fat}_{c\varepsilon}(\mathcal{F})}\log^{7/2}\frac{CT}{\sigma\varepsilon} + \sqrt{T\log(2T)}\varepsilon\right].$$

To obtain (9), note that each term on the right-hand side is at most a constant times

$$\left(1 + \mathrm{fat}_{c\varepsilon}(\mathcal{F}) + \sqrt{T}\varepsilon\right) \sqrt{\frac{T}{\sigma}} \log^{7/2} \frac{CT}{\sigma\varepsilon}.$$

For the term $\log(2T)/\sigma$, this uses $T \geq 1/\sigma$, which gives $1/\sigma \leq \sqrt{T/\sigma}$; for the remaining terms, it uses $\sigma \leq 1$ and $\log(CT/(\sigma\varepsilon)) \geq 1$. Moreover, $T/(\sigma\varepsilon) \geq 16$ gives $\log(CT/(\sigma\varepsilon)) \lesssim \log(2T/(\sigma\varepsilon))$. The learner does not depend on ε, so taking the infimum over ε proves (9). □