Efficient First-Order Contextual Bandits:
Prediction, Allocation, and Triangular Discrimination

Dylan J. Foster
Microsoft Research, New England
dylanfoster@microsoft.com

Akshay Krishnamurthy
Microsoft Research, NYC
akshaykr@microsoft.com

Abstract

A recurring theme in statistical learning, online learning, and beyond is that faster convergence rates are possible for problems with low noise, often quantified by the performance of the best hypothesis; such results are known as *first-order* or *small-loss* guarantees. While first-order guarantees are relatively well understood in statistical and online learning, adapting to low noise in *contextual bandits* (and more broadly, decision making) presents major algorithmic challenges. In a COLT 2017 open problem, Agarwal et al. [5] asked whether first-order guarantees are even possible for contextual bandits and—if so—whether they can be attained by efficient algorithms. We give a resolution to this question by providing an optimal and efficient reduction from contextual bandits to online regression with the logarithmic (or, cross-entropy) loss. Our algorithm is simple and practical, readily accommodates rich function classes, and requires no distributional assumptions beyond realizability. In a large-scale empirical evaluation, we find that our approach typically outperforms comparable non-first-order methods.

On the technical side, we show that the logarithmic loss and an information-theoretic quantity called the *triangular discrimination* play a fundamental role in obtaining first-order guarantees, and we combine this observation with new refinements to the regression oracle reduction framework of Foster and Rakhlin [29]. The use of triangular discrimination yields novel results even for the classical statistical learning model, and we anticipate that it will find broader use.

1 Introduction

In the contextual bandit problem, a learning agent repeatedly makes decisions based on contextual information, with the goal of learning a decision-making policy that minimizes their total loss over time. This model captures simple reinforcement learning tasks in which the agent must learn to make high-quality decisions in an uncertain environment, but does not need to engage in long-term planning or credit assignment. Owing to the availability of high-quality engineered reward metrics, contextual bandit algorithms are now routinely deployed in production for online personalization systems [4, 61].

Contextual bandits encompass both the general problem of statistical learning with function approximation (specifically, cost-sensitive classification) and the classical multi-armed bandit problem, yet present algorithmic challenges greater than the sum of both parts. In spite of these difficulties, extensive research effort over the past decade has resulted in efficient, general-purpose algorithms, as well as a sharp understanding of the optimal worst-case sample complexity [9, 12, 3, 29, 58].

While the algorithmic and statistical foundations for contextual bandits are beginning to take shape, we still lack an understanding of *adaptive* or *data-dependent* algorithms that can go beyond the worst case and exploit nice properties of real-world instances for better performance. This is in stark contrast to supervised statistical learning, where adaptivity has substantial theory, and where standard

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

algorithms (e.g., empirical risk minimization) are known to automatically adapt to nice data [17]. For contextual bandits, adaptivity poses new challenges that seem to require algorithmic innovation, and a major research frontier is to develop algorithmic principles for adaptivity and an understanding of the fundamental limits.

To highlight the lack of understanding for adaptive and data-dependent algorithms, a COLT 2017 open problem posed by Agarwal, Krishnamurthy, Langford, Luo, and Schapire [5] asks whether there exist contextual bandit algorithms that achieve a certain data-dependent *first-order* regret bound, which scales with the cumulative loss $L^\star$ of the best policy, rather than with the time horizon T. For multi-armed bandits, first-order regret bounds (also known as *small-loss bounds* or *fast rates*) typically scale as $\sqrt{L^\star}$ and imply faster convergence for "easy" problems, interpolating between the optimal $\sqrt{T}$ rate for worst-case instances and constant/logarithmic regret for noise-free instances [7, 31]. Agarwal et al. [5] observed that existing techniques appear to be inadequate to achieve this type of guarantee in contextual bandits. Beyond simply asking whether first-order regret can be achieved, they also asked whether it can be achieved *efficiently*, which is essential for real-world deployment. Subsequently, Allen-Zhu, Bubeck, and Li [6] gave an inefficient algorithm with an optimal first-order regret guarantee, resolving the former question, but the existence of efficient first-order algorithms remained open.

Contributions. We give the first optimal and efficient contextual bandit algorithm with a first-order regret guarantee, providing a resolution to the second open problem raised by Agarwal et al. [5]. Our algorithm, FastCB, builds on a recent line of research that develops efficient contextual bandit algorithms based on the computational primitive of (online/offline) *supervised regression* [43, 32, 29, 58], and is efficient in terms of queries to an *online oracle* for regression with the logarithmic loss. Beyond attaining first-order regret, FastCB inherits all of the benefits of recent algorithms based on regression: it is simple and practical, accommodates flexible function classes, requires no statistical assumptions beyond realizability, and enjoys strong empirical performance.

Technical highlights. By invoking the framework of regression oracles, our algorithm design approach deviates sharply from prior approaches to first-order regret and necessitates the use of techniques that are novel even in the context of statistical learning. At a high-level, the design of FastCB leverages two key techniques:

1. *First-order regret for plug-in classification via logarithmic loss:* We show that algorithms based on regression with least-squares, as used in prior work [29, 58, 68, 34, 23], fail to attain first-order regret, even for the simpler problem of cost-sensitive classification in statistical learning. In spite of this apparent setback, we show that regression with the logarithmic loss *does* lead to first-order regret for statistical learning. This is established through a new analysis based on an information-theoretic quantity called the *triangular discrimination* [66, 44, 62].
2. *Reweighted inverse gap weighting:* Moving from statistical learning to contextual bandits, we transform predictions into distributions over actions using a scale-sensitive refinement to the *inverse-gap weighting scheme* used in the SquareCB algorithm [1, 29]. Our new scheme is tailored to small losses, and we show that its error is controlled by the triangular discrimination.

Summarizing, our approach leverages **prediction** via the logarithmic loss, **allocation** via reweighted inverse gap weighting, and **triangular discrimination** as the bridge from prediction to allocation.

Empirical results. In Section 5, we evaluate FastCB on the large-scale contextual bandit benchmark of Bietti et al. [13] and find that it typically outperforms SquareCB and other non-adaptive baselines [35]. Interestingly, we observe that most of the performance improvement can be attributed to the use of the logarithmic loss, while the reweighted allocation scheme provides modest additional benefit. These findings raise a natural question as to whether simply moving to the logarithmic loss can yield performance improvements in production contextual bandit deployments.

On the regression oracle model. As a disclaimer, we caution that our algorithm is efficient in terms of an oracle for online regression, while Agarwal et al. [5] originally asked for an algorithm that is efficient in terms of a *cost-sensitive classification oracle* capable of solving the policy optimization problem $\mathrm{argmin}_{\pi\in\Pi}\sum_{t=1}^{T}\ell_t(\pi(x_t))$. Hence, while FastCB is the first algorithm with first-order regret that is efficient in *any* oracle model, it does not formally solve the original open problem. Nonetheless, there are strong reasons to prefer a solution based on regression over one based on classification. First, cost-sensitive classification is intractable to implement even for simple function classes for which regression can be solved efficiently [29]. Setting this issue aside, (online)

Algorithm 1 FastCB ("Fast Rates for Contextual Bandits")

```
parameters:
    Learning rate $\gamma > 0$.
    Online regression oracle $\mathbf{Alg}_{\mathsf{KL}}$.
for $t = 1, \dots, T$ do
    Receive context $x_t$.
    // Compute oracle's predictions (Eq. (4)).
    For each action $a \in \mathcal{A}$, compute $\widehat{y}_t(x_t, a) := \mathbf{Alg}^{(t)}_{\mathsf{KL}}(x_t, a\,; \{(x_i, a_i, \ell_i(a_i))\}_{i=1}^{t-1})$.
    Let $b_t \in \operatorname{argmin}_{a \in \mathcal{A}} \widehat{y}_{t,a}$.
    // Reweighted inverse gap weighting.
    For each $a \neq b_t$, define $p_{t,a} = \frac{\widehat{y}_t(x_t,b_t)}{A\widehat{y}_t(x_t,b_t)+\gamma(\widehat{y}_t(x_t,a)-\widehat{y}_t(x_t,b_t))}$. Let $p_{t,b_t} = 1 - \sum_{a \neq b_t} p_{t,a}$.
    Sample $a_t \sim p_t$ and observe loss $\ell_t(a_t)$.
    Update $\mathbf{Alg}_{\mathsf{KL}}$ with example $(x_t, a_t, \ell_t(a_t))$.
end for
```

regression-based algorithms are typically simpler and faster than classification-based algorithms, and multiple empirical evaluations have shown that algorithms based on regression dominate those based on classification [32, 13, 35].

Organization. Section 2 contains our algorithm and main theorem. Section 3 describes the motivation and analysis ideas behind FastCB, beginning from new techniques for statistical learning with regression-based classifiers. Examples for the main theorem are given in Section 4, and experimental results are given in Section 5. Detailed discussion of related work is deferred to Appendix A.

2 Main Result: An Efficient First-Order Algorithm for Contextual Bandits

We begin by formally introducing the contextual bandit model. At each round $t \in [T]$, the learner observes a context $x_t \in \mathcal{X}$, selects an action $a_t \in \mathcal{A}$, then observes a loss $\ell_t(a_t) \in [0, 1]$ for the action they selected. We assume that $A := |\mathcal{A}|$ is finite and that each loss function $\ell_t : \mathcal{A} \to [0, 1]$ is drawn independently from a fixed distribution $\mathbb{P}_{\ell_t}(\cdot \mid x_t)$, where $\mathbb{P}_{\ell_1}, \dots, \mathbb{P}_{\ell_T}$ and $x_1, \dots, x_T$ are selected by a potentially adaptive adversary.

We make a standard *realizability* assumption [24, 2, 32, 29]. Namely, we assume that the learner has access to a class of value functions $\mathcal{F} \subset (\mathcal{X} \times \mathcal{A} \to [0, 1])$ (e.g., neural networks, kernels, or forests) that models the mean of the loss distribution.

Assumption 1 (Realizability)**.** *There exists $f^\star \in \mathcal{F}$ such that for all t, $f^\star(x, a) = \mathbb{E}[\ell_t(a) \mid x_t = x]$.*

The aim of the learner is to minimize their *regret* to the optimal policy $\pi^\star(x) := \operatorname{argmin}_{a \in \mathcal{A}} f^\star(x, a)$:

$$\mathbf{Reg}_{\mathsf{CB}}(T) := \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi^\star(x_t)). \tag{1}$$

For each $f \in \mathcal{F}$, we let $\pi_f(x) := \operatorname{argmin}_{a \in \mathcal{A}} f(x, a)$ be the induced policy. We let $\Pi := \{\pi_f \mid f \in \mathcal{F}\}$ be the induced policy class.

Further notation. We adopt standard big-oh notation, and write $f = \widetilde{\mathcal{O}}(g)$ to denote that $f = \mathcal{O}(g \max\{1, \operatorname{polylog}(g)\})$. We use $\lesssim$ only in informal statements to highlight the most salient elements of an inequality. We use $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

2.1 Algorithm and Main Result

FastCB builds on the SquareCB algorithm of Foster and Rakhlin [29], which provides an efficient, minimax-optimal reduction from contextual bandits to online regression with the square loss. Compared to SquareCB and other subsequent algorithms based on online regression [34, 23], the first twist here is that rather than working with the square loss, we build on the computational primitive of online regression with the *logarithmic loss*. While this point is inconsequential for worst-case guarantees, we show that it is a fundamental distinction for first-order guarantees.

Online regression oracles. In more detail, an online regression oracle, which we denote by $\mathbf{Alg}_{\mathsf{KL}}$ (for "Kullback-Leibler") operates in the following protocol: For each time t, the algorithm receives a context-action pair (x_t, a_t), produces a prediction $\widehat{y}_t \in [0,1]$, then receives a response y_t. The algorithm's prediction error is measured through the binary logarithmic/cross-entropy loss ,

$$\ell_{\log}(\widehat{y}, y) := y \log(1/\widehat{y}) + (1-y)\log(1/(1-\widehat{y})). \tag{2}$$

The algorithm's goal is to ensure that the log loss regret to $\mathcal{F}$ is minimized for all sequences.

Assumption 2. *The algorithm $\mathbf{Alg}_{\mathsf{KL}}$ guarantees that for every (possibly adaptively chosen) sequence $x_{1:T}, a_{1:T}, y_{1:T}$, the log loss regret is bounded by a known function $\mathbf{Reg}_{\mathsf{KL}}(T)$:*

$$\sum_{t=1}^{T} \ell_{\log}(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_{\log}(f(x_t, a_t), y_t) \leq \mathbf{Reg}_{\mathsf{KL}}(T), \tag{3}$$

Online regression with the logarithmic loss (or, *sequential probability assignment*) is a fundamental and well-studied problem in online learning, and there are efficient algorithms available for many function classes of interest [25, 67, 40, 37, 52, 55, 33, 48]; see Section 4 for examples. While log loss regret is a more stringent notion of performance than square loss regret, it nonetheless has a relatively mature theory characterizing optimal rates [57, 51, 20, 14].

The algorithm. FastCB (Algorithm 1) is a reduction that efficiently transforms any online regression oracle satisfying Assumption 2 into a contextual bandit algorithm with an optimal first-order regret bound. At each round t, the algorithm first computes the estimated loss

$$\widehat{y}_t(x_t, a) := \mathbf{Alg}_{\mathsf{KL}}^{(t)}(x_t, a\,; \{(x_i, a_i, \ell_i(a_i))\}_{i=1}^{t-1}) \tag{4}$$

predicted by the regression oracle for each action a (Line 4); see Appendix C.1 for a more detailed formal description of the oracle model. Next, FastCB uses these estimates to assign a probability of being played to each action a via a scale-sensitive refinement to the inverse gap weighting strategy used in SquareCB [1, 29], which we call *reweighted inverse gap weighting* (Line 6). Letting $b_t := \operatorname{argmin}_{a \in \mathcal{A}} \widehat{y}_t(x_t, a)$ be the greedy action according to the predicted losses, we define

$$p_{t,a} := \frac{\widehat{y}_t(x_t, b_t)}{A\widehat{y}_t(x_t, b_t) + \gamma(\widehat{y}_t(x_t, a) - \widehat{y}_t(x_t, b_t))} \quad \forall a \neq b_t, \quad \text{and} \quad p_{t,b_t} := 1 - \sum_{a \neq b_t} p_{t,a}, \tag{5}$$

where $\gamma > 0$ is a learning rate parameter. Given this distribution, FastCB simply samples $a_t \sim p_t$, then updates the oracle with the resulting tuple $(x_t, a_t, \ell_t(a_t))$. Our main theorem shows that this leads to an optimal first-order regret bound.[1]

Theorem 1 (Main theorem)**.** *Suppose Assumptions 1 and 2 hold. Then Algorithm 1 guarantees that for all sequences with $\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\pi^\star(x_t))\right] \leq L^\star$, by choosing $\gamma = \sqrt{AL^\star / 3\mathbf{Reg}_{\mathsf{KL}}(T)} \vee 10A$,*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq 40\sqrt{L^\star \cdot A\mathbf{Reg}_{\mathsf{KL}}(T)} + 600A\mathbf{Reg}_{\mathsf{KL}}(T). \tag{6}$$

The dominant term in this regret bound scales with $\sqrt{L^\star}$ whenever the oracle $\mathbf{Alg}_{\mathsf{KL}}$ attains a fast $\log(T)$-type regret bound. As a simple example, whenever $\mathcal{F}$ is finite, we can instantiate $\mathbf{Alg}_{\mathsf{KL}}$ so that $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \log|\mathcal{F}|$ [67], whereby FastCB enjoys optimal [2] first-order regret:

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq \mathcal{O}\left(\sqrt{L^\star \cdot A\log|\mathcal{F}|} + A\log|\mathcal{F}|\right).$$

Beyond first-order regret, FastCB inherits all of the advantages of online regression-based algorithms:

- *Efficiency and simplicity.* The memory and runtime used by the algorithm—on top of what is required by the regression oracle—scales only as $\mathcal{O}(A)$ per step; implementation is trivial.
- *Flexibility.* Working with regression as a primitive means that the algorithm easily accomodates rich, potentially nonparametric function classes, and we can instantiate Theorem 1 to get provable end-to-end regret guarantees for concrete classes of interest. For example, for linear models in $\mathbb{R}^d$ we can efficiently attain $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \mathcal{O}(d\log(T))$ [25, 40], which yields a first-order regret bound $\mathbf{Reg}_{\mathsf{CB}}(T) \lesssim \sqrt{L^\star \cdot Ad}$; our result is new even for this simple special case. Similar guarantees are available for kernels, generalized linear models, and many other nonparametric classes. On the other hand, even for function classes where provable algorithms are not available, regression is amenable to practical heuristics (e.g., gradient descent). See Section 4 for detailed examples.

[1] While we assume that an upper bound on the optimal loss is known for simplicity, one can extend to the unknown case by running the algorithm in epochs, setting γ in terms of the algorithm's estimated loss $L_t = \sum_{\tau=1}^{t} \ell_\tau(a_\tau)$, and applying the doubling trick. Theorem 1 also readily extends to high probability.

3 Overview of Analysis

We now outline the algorithmic principles and analysis ideas behind FastCB. First, in Section 3.1, we take a step back and consider the sub-problem of cost-sensitive classification in statistical learning. We establish that approaches based on least-squares fail to attain first-order regret (Theorem 2) for cost-sensitive classification, then show how to fix this problem using log loss regression (Theorem 3); this analysis serves as an introduction to the triangular discrimination. With this result in hand, we move to the contextual bandit setting and transform predictions into distributions over actions using the reweighted inverse-gap weighting scheme in (5), which exploits small losses. Our main result here shows that this scheme satisfies a first-order variant of the *per-round minimax inequality* of Foster and Rakhlin [29], which links the instantaneous contextual bandit regret to the triangular discrimination for the regression oracle on a per-round basis (Theorem 4). Full proofs are deferred to Appendices B and C.

3.1 Warmup: First-Order Regret Bounds for Plug-In Classifiers

For the simpler problem of cost-sensitive classification in statistical learning, the literature on *plug-in classification* shows that whenever realizability conditions such as Assumption 1 hold, we can obtain optimal worst-case regret by taking the greedy policy/classifier induced by a least-squares estimator. We first show that this approach fails to attain first-order regret.

The statistical learning setting we consider is as follows. We receive a dataset D_n consisting of n context-loss pairs $(x_t, \ell_t) \sim \mathcal{D}$ i.i.d., where the entire loss function $\ell_t : \mathcal{A} \to [0,1]$ is observed. Analogously to Assumption 1, we assume access to a function class $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ such that $\mathbb{E}_{\mathcal{D}}[\ell(a) \mid x] = f^\star(x,a)$ for some $f^\star \in \mathcal{F}$, and take $\Pi := \{\pi_f \mid f \in \mathcal{F}\}$ as the induced class of policies. Our goal is to learn a policy $\widehat{\pi} : \mathcal{X} \to \mathcal{A}$ such that the regret (or, excess risk)

$$L(\widehat{\pi}) - L^\star \tag{7}$$

is small, where $L(\pi) := \mathbb{E}_{\mathcal{D}}[\ell(\pi(x))]$ and $L^\star := L(\pi^\star)$, with $\pi^\star := \pi_{f^\star}$. Formally, this an easier problem than contextual bandits, since any algorithm with a regret bound for contextual bandits yields a bound on the cost-sensitive classification regret (7) via online-to-batch conversion.

A classical result in statistical learning [65, 54, 59] shows that if we compute the policy/classifier $\widehat{\pi} := \operatorname{argmin}_{\pi \in \Pi} \sum_{t=1}^{n} \ell_t(\pi(x_t))$ that minimizes the empirical risk, we obtain a first-order regret bound of the form[2]

$$\mathbb{E}[L(\widehat{\pi})] - L^\star \lesssim \sqrt{\frac{L^\star \cdot \log|\mathcal{F}|}{n}} + \frac{\log|\mathcal{F}|}{n}. \tag{8}$$

This is an optimal first-order guarantee, but computing $\widehat{\pi}$ is typically computationally intractable, even for relatively simple policy classes. As an alternative, the approach of plug-in classification aims to use the realizability assumption to develop algorithms based on the more tractable primitive of regression. Here, another classical result (e.g., Audibert and Tsybakov [8][3]), shows that if we perform least-squares via

$$\widehat{f}_{\mathsf{LS}} := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} (f(x_t, a) - \ell_t(a))^2,$$

and take $\widehat{\pi}_{\mathsf{LS}} := \pi_{\widehat{f}_{\mathsf{LS}}}$ as our classifier, then under the realizability assumption we are guaranteed

$$\mathbb{E}[L(\widehat{\pi}_{\mathsf{LS}})] - L^\star \lesssim \sqrt{\frac{A \log|\mathcal{F}|}{n}}. \tag{9}$$

While this result is rate-optimal, it is not first-order, and first-order regret bounds for plug-in classification are conspicuously absent from the literature. We show that this is fundamental.

Theorem 2 (Failure of least-squares for plug-in classification)**.** *Let* $\mathcal{A} = \{1,2\}$ *and* $\mathcal{X} = \{1,2\}$. *For every* $n > 10^8$, *there exists a function class* $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ *with* $|\mathcal{F}| = 2$, *and a realizable distribution* $\mathcal{D}$ *such that* $L^\star \le \frac{2^7}{n} < 1$, *yet* $L(\widehat{\pi}_{\mathsf{LS}}) - L^\star \ge 2^{-5}\sqrt{\frac{1}{n}}$ *with probability at least* $1/10$.

[2] Following the convention in contextual bandit literature, we focus on finite classes with $|\mathcal{F}| < \infty$ in this discussion, but one can extend our observations to general classes, e.g., using the machinery of Zhang [71].

[3] This result is well-known in the binary setting. We are not aware of a reference for the multiclass/cost-sensitive version here, though it is implicit in many recent works on contextual bandits.

Since the instance in this theorem has $\sqrt{\frac{L^\star \cdot A \log|\mathcal{F}|}{n}} \lesssim \frac{1}{n}$, we conclude that plug-in classification with least-squares fails to attain the first-order regret bound in (8) with constant probability; a lower bound in expectation follows immediately.

3.1.1 Fast Rates for Plug-In Classifiers: Triangular Discrimination and Logarithmic Loss

It would appear we are at an impasse, as Theorem 2 shows that square loss regression oracles of the type used in Foster and Rakhlin [29] are unlikely to attain first-order regret bounds on their own. However, the plug-in classification approach is not completely doomed. All we need to do to fix this issue is change the loss function and instead perform regression with the *logarithmic loss*.

To understand why plug-in least-squares fails and how it can be improved, it will be helpful to review the key steps in the analysis leading to the rate (9).

`Step 1`. First, using a generic regret decomposition based on realizability, for any f we have

$$L(\pi_f) - L^\star \le 2 \max_{\pi \in \{\pi_f, \pi^\star\}} \mathbb{E}_{\mathcal{D}} \big| f(x, \pi(x)) - f^\star(x, \pi(x)) \big|. \tag{10}$$

`Step 2`. Next, by Cauchy-Schwarz, for any policy π we have

$$\mathbb{E}_{\mathcal{D}} \big| f(x, \pi(x)) - f^\star(x, \pi(x)) \big| \le \left(\mathbb{E}_{\mathcal{D}} \big| f(x, \pi(x)) - f^\star(x, \pi(x)) \big|^2 \right)^{1/2}, \tag{11}$$

which we may further upper bound by $\left(\sum_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{D}} \big| f(x,a) - f^\star(x,a) \big|^2 \right)^{1/2}$.

`Step 3`. Finally, under realizability, a standard concentration argument based on Bernstein's inequality implies that the least-squares estimator satisfies

$$\mathbb{E}\left[\sum_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{D}} \big| \widehat{f}_{\mathsf{LS}}(x,a) - f^\star(x,a) \big|^2 \right] \lesssim \frac{A \log|\mathcal{F}|}{n}. \tag{12}$$

Combining this bound with `Step 2`, we conclude that $\mathbb{E}[L(\widehat{\pi}_{\mathsf{LS}})] - L^\star \le \sqrt{A \log|\mathcal{F}|/n}$.

The issue here is that even in the presence of low noise, the squared error in (12) shrinks no faster than $\frac{1}{n}$. This holds even if $L^\star \propto \frac{1}{n}$, as in the lower bound construction for Theorem 2. Consequently, once we apply Cauchy-Schwarz in `Step 2`, we lose all hope of attaining a first-order bound.

Our starting point toward improving this result is a refined application of Cauchy-Schwarz, by which we can replace the right hand side of (11) with

$$\left(\mathbb{E}_{\mathcal{D}}[f(x,\pi(x)) + f^\star(x,\pi(x))] \cdot \mathbb{E}_{\mathcal{D}}\left[\frac{\left(f(x,\pi(x)) - f^\star(x,\pi(x))\right)^2}{f(x,\pi(x)) + f^\star(x,\pi(x))} \right] \right)^{1/2}. \tag{13}$$

The ratio term above is closely related to the *triangular discrimination*, an information-theoretic divergence measure which we define for $p, q \in \mathbb{R}_+^A$ as[4]

$$D_\Delta(p \,\|\, q) := \sum_a \frac{(p_a - q_a)^2}{p_a + q_a}. \tag{14}$$

The triangular discrimination—also known as the symmetric χ^2-divergence and Vincze-Le Cam distance—is a fundamental, often-overlooked quantity in information theory [66, 44, 62]. Since readers may be unfamiliar, we record some basic facts.

Proposition 1 (Topsøe [62]). *The triangular discrimination D_Δ, over the domain Δ_A, i) is the f-divergence given by $f(t) = \frac{(t-1)^2}{t+1}$, ii) is the square of a distance metric, and iii) is equivalent (up to a multiplicative constant) to both Hellinger distance and Jensen-Shannon divergence.*

[4] The triangular discrimination is traditionally defined over the simplex Δ_A, but for our application it is useful to work with the entire positive orthant.

The triangular discrimination turns out to be "just right" for our purposes, in that it is both i) large enough to facilitate the scale-sensitive application of Cauchy-Schwarz in (13), and ii) small enough (compared to the more standard χ^2-divergence) to facilitate minimizing from samples.

Returning to (13), we can upper bound with the triangular discrimination and leverage a certain *self-bounding* property that it satisfies to arrive at the following improvement on **Step 1**/**Step 2**.

Lemma 1 (Regret decomposition for triangular discrimination)**.** *For any* $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$*,*

$$L(\pi_f) - L^\star \leq 8(L^\star \cdot \mathbb{E}_{\mathcal{D}}[D_\triangle(f^\star(x,\cdot) \,\|\, f(x,\cdot))])^{1/2} + 17\,\mathbb{E}_{\mathcal{D}}[D_\triangle(f^\star(x,\cdot) \,\|\, f(x,\cdot))]. \tag{15}$$

Lemma 1 shows that low triangular discrimination (i.e. $\mathbb{E}_{\mathcal{D}}[D_\triangle(f^\star(x,\cdot) \,\|\, f(x,\cdot))] \propto 1/n$) suffices for an optimal first-order regret bound. What remains is to find an estimator $\widehat{f}$ that minimizes this quantity given only samples. Our key observation here is that the triangular discrimination satisfies a refined variant of Pinsker's inequality (originally due to Topsøe [62]), which allows us to bound it by the Kullback-Leibler divergence:

$$D_\triangle(f^\star(x,\cdot) \,\|\, f(x,\cdot)) = \sum_a \frac{(f(x,a) - f^\star(x,a))^2}{f(x,a) + f^\star(x,a)} \leq 2\sum_a d_{\mathrm{KL}}(f^\star(x,a) \,\|\, f(x,a)), \tag{16}$$

where $d_{\mathrm{KL}}(p \,\|\, q) := p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the binary KL-divergence. Note that the triangular discrimination is critical here, as the *opposite* inequality holds for χ^2-divergence. This bound suggests that we should minimize the logarithmic loss, since—under the realizability assumption—this loss is closely related to the KL-divergence. In particular, we show (Theorem 6 in Appendix B), that by taking the estimator

$$\widehat{f}_{\mathsf{KL}} := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \ell_{\log}(f(x_t, a), \ell_t(a)),$$

we are guaranteed that with high probability, $\mathbb{E}_{\mathcal{D}}\big[D_\triangle(f^\star(x,\cdot) \,\|\, \widehat{f}_{\mathsf{KL}}(x,\cdot))\big] \lesssim \frac{A\log|\mathcal{F}|}{n}$. Putting everything together, we arrive at a first-order regret bound for the plug-in classifier $\widehat{\pi}_{\mathsf{KL}} := \pi_{\widehat{f}_{\mathsf{KL}}}$.[5]

Theorem 3 (First-order regret bound for plug-in classification)**.** *Let* $\delta \in (0, 1)$*. Suppose that Assumption 3 holds. Then with probability at least* $1 - \delta$*, we have*

$$L(\widehat{\pi}_{\mathsf{KL}}) - L^\star \leq 16\sqrt{\frac{L^\star \cdot A\,(\log|\mathcal{F}| + \log(A/\delta))}{n}} + 68\frac{A\,(\log|\mathcal{F}| + \log(A/\delta))}{n}.$$

Interestingly, applications of the triangular discrimination similar to Lemma 1 have recently been discovered across a number of branches of mathematics, including theoretical computer science (communication complexity lower bounds), probability, and group theory (e.g., construction of group homomorphisms) [70, 28, 11, 53]. Additionally, Bubeck and Sellke [18] use a related *non-negative* χ^2*-divergence* to provide first-order Bayesian regret bounds for Thompson sampling for the multi-armed bandit.

3.2 Moving to Contextual Bandits: Inverse Gap Weighting meets Triangular Discrimination

FastCB builds on the development for plug-in classifiers in Section 3.1 but with two key differences. First, since we need to make decisions on the fly for arbitrary sequences of contexts, the algorithm estimates losses using an *online* regression oracle for the logarithmic loss, as described in Assumption 2. Second, and more importantly, since the algorithm receives partial feedback, the strategy for selecting actions is critical. Here our main technical result shows that the reweighted inverse gap weighting strategy (5) satisfies a certain *per-round* inequality that links the instantaneous contextual bandit error to the triangular discrimination between the oracle's prediction $\widehat{y}_t$ and the true loss function $f^\star$.

Theorem 4 (First-order per-round inequality)**.** *Let* $y \in [0,1]^A$ *be given and* $b \in \operatorname{argmin}_a y_a$*. Define* $p_a = \frac{y_b}{Ay_b + \gamma(y_a - y_b)}$ *for* $a \neq b$*, and* $p_b = 1 - \sum_{a \neq b} p_a$*. If* $\gamma \geq 2A$*, then for all* $f \in [0,1]^A$ *and* $a^\star \in \operatorname{argmin}_a f_a$*, we have*

$$\underbrace{\sum_a p_a(f_a - f_{a^\star})}_{\text{CB regret}} \leq \underbrace{\frac{5A}{\gamma}\sum_a p_a f_a}_{\text{bias from } \textit{exploring}} + \underbrace{7\gamma\sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}}_{\text{error from } \textit{exploiting}}. \tag{17}$$

[5] The dependence on A in this result can be improved under additional assumptions on the loss distribution. As an example, in Appendix B we remove the leading A factor for the special case of multiclass classification.

The inequality (17) may be thought of as an algorithmic analogue of the refined Cauchy-Schwarz lemma (15), with the learning rate γ modulating the tradeoff between exploration and exploitation. Applying the inequality for each step t (with $p = p_t$, $y = \widehat{y}_t(x_t, \cdot)$, and $f = f^\star(x_t, \cdot)$), and using the Pinsker-type inequality (16), we are guaranteed that

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq \frac{5A}{\gamma}\,\mathbb{E}[L_T] + 14\gamma \cdot \mathbf{Reg}_{\mathsf{KL}}(T), \tag{18}$$

where $L_T := \sum_{t=1}^{T} \ell_t(a_t)$. By a standard argument, this implies the main result in Theorem 1.

Compared to the per-round inequality used to analyze the original version of SquareCB in Foster and Rakhlin [29], the main improvement given by Theorem 4 is that, by reweighting—which leads to less exploration when the optimal loss is small—we are able to replace a constant exploration bias of order $\frac{A}{\gamma}$ incurred by SquareCB with the scale-sensitive bias term $\frac{A}{\gamma} \cdot \sum_a p_a f_a$ in (17), leading to a first-order bound. The price for this improvement is that we must now minimize the triangular discrimination rather than the squared error used by SquareCB, but this is taken care of by the log loss oracle.

4 Examples

In this section we take advantage of the extensive literature on regression with the logarithmic loss [25, 67, 40, 37, 52, 55, 33, 48] and instantiate Theorem 1 to give provable and efficient first-order regret bounds for a number of function classes of interest. To the best of our knowledge, our results are new for each of these special cases.

Example 1 (Finite function classes)**.** *If $\mathcal{F}$ is a finite class, Vovk's aggregating algorithm [67] guarantees that*[6] $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \log|\mathcal{F}|$. *With this choice, FastCB satisfies* $\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq \mathcal{O}\big(\sqrt{L^\star \cdot A \log|\mathcal{F}|} + A\log|\mathcal{F}|\big)$.

Example 2 (Low-dimensional linear functions)**.** *Suppose that $\mathcal{F}$ takes the form* $\mathcal{F} = \{(x,a) \mapsto \langle w, \phi(x,a)\rangle \mid w \in \Delta_d\}$, *where $\phi(x,a) \in \mathbb{R}_+^d$ is a fixed feature map with $\|\phi(x,a)\|_\infty \leq 1$. Then the continuous exponential weights algorithm ensures that* $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \mathcal{O}(d\log(T/d))$, *and can be implemented in* $\mathrm{poly}(d,T)$ *time per step using log-concave sampling [25, 40].*[7] *With this choice, FastCB satisfies* $\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq \mathcal{O}\big(\sqrt{L^\star \cdot Ad\log(T/d)} + Ad\log(T/d)\big)$.

Beyond attaining first-order regret, the bound in this example is minimax optimal when the number of actions is constant [46]. A natural direction for future work is to extend the result to large action spaces. Another more practical choice for the oracle in this setting is the algorithm of Luo et al. [48], which has slightly worse regret $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \widetilde{\mathcal{O}}(d^2)$, but runs in time $\mathcal{O}(Td^{2.5})$ per step.

While first-order regret bounds for contextual bandits have primarily been investigated for finite classes prior to this work, an advantage of working within the regression oracle framework is that we can easily lift our first-order guarantees to rich, nonparametric function classes.

Example 3 (High/infinite-dimensional linear functions)**.** *Suppose that $\mathcal{F}$ takes the form* $\mathcal{F} = \big\{(x,a) \mapsto \frac{1}{2}(1 + \langle w, \phi(x,a)\rangle) \mid \|w\|_2 \leq 1\big\}$, *where $\|\phi(x,a)\|_2 \leq 1$ is a fixed feature map. For this setting, Rakhlin and Sridharan [55, Section 6.1] show that the FTRL algorithm with log-barrier regularization has*[8] $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \mathcal{O}(\sqrt{T\log(T)})$. *This algorithm can be implemented in time $\mathcal{O}(d)$ per step, and satisfies the dimension-independent rate* $\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \leq \mathcal{O}\big((AL^\star)^{1/2}T^{1/4} + A\sqrt{T}\big)$.

Let us interpret the bound in Example 3. First, we recall that the minimax optimal rate for this setting is $A^{1/2}T^{3/4}$, which the bound above always achieves in the worst case [1, 29]; this "worse-than-$\sqrt{T}$" rate is the price we pay for working with an expressive function class. On the other hand, if $L^\star$ is constant, the bound in Example 3 improves to $\mathcal{O}(A\sqrt{T})$, which beats the worst-case rate. While one might hope that a tighter rate of the form, e.g., $(L^\star)^{3/4}$, might be possible, by adapting a lower bound in Srebro et al. [59, Section 4], one can show that this result cannot be improved.

Lastly, we highlight that the logarithmic loss is well-suited to generalized linear models.

[6] See Proposition 6 for a proof that the loss $\ell_{\log}(\widehat{y}, y)$ is mixable over the domain $[0,1]$.

[7] Our setup directly reduces to universal portfolio selection as follows: When y_t is binary, we reduce by using features $\phi(x_t, a_t)$ when $y_t = 1$, and using features $\mathbf{1}_d - \phi(x_t, a_t)$ when $y_t = 0$. The case where $y_t \in [0,1]$ can be reduced to this setting by sampling from $\mathrm{Ber}(y_t)$.

[8] This is technically only proven for the case where $y \in \{0,1\}$, but the proof easily extends to $y \in [0,1]$.

Example 4 (Generalized linear models)**.** *Let* $\mathcal{F} = \left\{(x,a) \mapsto \sigma(\langle w, \phi(x,a)\rangle) \mid w \in \mathbb{R}^d, \|w\|_2 \leq 1\right\}$*, where* $\sigma(t) = 1/(1+e^{-t})$ *is the logistic link function and* $\phi(x,a)$ *is a fixed feature map. In this case, the map* $w \mapsto \ell_{\log}(\sigma(\langle w, \phi(x,a)\rangle), y)$ *is equivalent to the standard logistic loss function applied to* $\langle w, \phi(x,a)\rangle$*, and we can use the algorithm from Foster et al. [33] to obtain* $\mathbf{Reg}_{\mathsf{KL}}(T) \leq \mathcal{O}(d\log(T/d))$ *and* $\mathbf{Reg}_{\mathsf{CB}}(T) \leq \widetilde{\mathcal{O}}(\sqrt{L^\star \cdot Ad} + Ad)$*.*

Beyond the algorithmic examples above, for general function classes Bilodeau et al. [14] provide a tight characterization for the minimax optimal rates for online regression with the logarithmic loss in terms of *sequential covering numbers* [55] for the class $\mathcal{F}$. We can use this result in tandem with Theorem 1 to give new regret bounds for general classes.

5 Experiments

We compared the performance of FastCB to that of the de-facto alternative, SquareCB [29] in the large-scale contextual bandit evaluation suite ("bake-off") of Bietti et al. [13]. We found that FastCB typically enjoys improved performance, particularly on datasets where the optimal loss $L^\star$ is small. As a secondary observation, we found that using generalized linear models with the logarithmic loss rather than a linear model with the square loss (as in prior work [13, 35]) leads to substantial improvements, even without changing the SquareCB allocation rule. We summarize results here; further details are given in Appendix E.

Datasets. The *contextual bandit bake-off* is a collection of over 500 multiclass, multilabel, and cost-sensitive classification datasets available on the `openml.org` platform [64]. The collection was introduced in Bietti et al. [13] for the purpose of benchmarking oracle-based contextual bandit algorithms. Following Bietti et al. [13], we use the multiclass classification datasets from the collection (each context x has a "correct" label y associated with it) to simulate bandit feedback by assigning loss 0 if the learner predicts the correct label and 1 otherwise.

Algorithms and oracle. We use the standard implementation of SquareCB in the Vowpal Wabbit (VW) online learning library,[9] as used by Foster et al. [35]. We also implement FastCB in VW.

For both algorithms, we instantiate the oracle as performing online logistic regression with a fixed dataset-dependent feature map. This choice is convenient because i) it naturally produces predictions in $[0,1]$, as required by FastCB, and ii), it formally meets our oracle requirements, since it is equivalent to online log loss regression with a generalized linear model. It can also be viewed as an admissible online square loss oracle, as required by SquareCB (see Appendix E for further discussion). We additionally instantiate SquareCB with a linear model and the square loss, which was shown to be the strongest non-adaptive method in prior evaluations [35]. We do not compare with high-performing adaptive algorithms like RegCB and AdaCB [13, 35] as these algorithmic modifications are somewhat complementary, and we expect they can be incorporated into FastCB. All oracles are trained with the default VW learning rule, which performs online gradient descent with adaptive updates [27, 41, 56].

For both FastCB and SquareCB, we apply inverse gap weighting (the reweighted and original version, respectively) with a time-varying learning rate schedule in which we set $\gamma = \gamma_t$ in Line 6 of Algorithm 1 at round t, and likewise for SquareCB. Following Foster et al. [35], we set $\gamma_t = \gamma_0 t^\rho$, where $\gamma_0 \in \{10, 50, 100, 400, 700, 10^3\}$ and $\rho \in \{.25, .5\}$ are hyperparameters.

Evaluation. We evaluate the performance of each algorithm using *progressive validation* (PV) loss, defined as $L_{\mathsf{PV}}(T) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(a_t)$ [15]. Following Bietti et al. [13], we define a given algorithm as beating another algorithm *significantly* on a given dataset using an approximate Z-test. For each pair (a,b) of algorithms, Figure 1 (top row) displays the number of datasets where a beats b significantly, minus the number of datasets where b beats a significantly. Figure 1 (bottom row) shows the progressive validation loss for the best-performing hyperparameter configuration for each algorithm as a function of the number of examples. We consider 10 replicates for each dataset, where each replicate has the example order randomly permuted, and plot the average progressive validation loss across the replicates. Error bands in each plot correspond to significance $p < 0.05$ under the Z-test (cf. (41)). See Appendix E for details.

Results. We find (Figure 1, top row) that FastCB with the logistic loss oracle (FastCB.L) has a positive win-loss difference against SquareCB with both logistic and square loss oracles

[9] https://vowpalwabbit.org

↓ vs →	S.S	S.L	F.L
SquareCB.S	-	-55	-66
SquareCB.L	55	-	-11
FastCB.L	**66**	**11**	-

↓ vs →	S.S	S.L	F.L
SquareCB.S	-	-54	-64
SquareCB.L	54	-	-3
FastCB.L	**64**	**3**	-

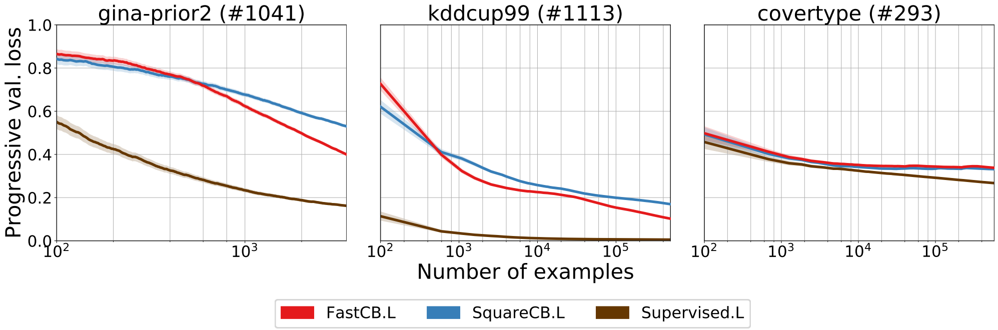

Figure 1: *Top:* Head-to-head win-loss differences. Each entry indicates the statistically significant win-loss difference between the row algorithm and the column algorithm. *Top-Left:* All hyperparameters are optimized on each dataset. *Top-Right:* Best fixed hyperparameter configuration across all datasets; only the oracle's learning rate is optimized per-dataset. *Bottom:* Progressive validation results for representative datasets depicting significant wins for FastCB.L (left, center) and a loss (right).

(SquareCB.L/SquareCB.S), indicating the strongest overall performance. This holds both when hyperparameters are optimized on a per-dataset basis and for the best global hyperparameter configuration.

Perhaps surprisingly, our results suggest that the largest gains come from switching from the square loss oracle to the logistic loss oracle (SquareCB.S vs. SquareCB.L), while the gains from switching from the original inverse gap weighting strategy to our reweighted version (SquareCB.L vs. FastCB.L) are more marginal. Inspecting the results in more detail, we find that when we compare FastCB.L and SquareCB.L with hyperparameters optimized on a per-dataset basis, FastCB.L wins on 14/17 of the datasets in which either algorithm wins significantly, and that all but two of these 14 datasets have $L^\star \leq 0.2$. This suggests that the reweighted inverse gap weighting strategy is indeed helpful when $L^\star$ is small. Figure 1 (bottom row) displays progressive validation performance for FastCB.L and SquareCB.L for three representative datasets which illustrate this phenomenon.

The fact that FastCB.L does not strictly improve over SquareCB.L on every dataset, in spite of being very similar, might be attributed to the fact that the constants in the per-round inequality (17) are worse than those in the corresponding inequality for SquareCB.L, suggesting worse performance when $L^\star$ is not small. Thus, a fruitful future direction might be to find a strategy with optimal constants for (17).

6 Discussion

We have given the first efficient algorithm with optimal first-order regret for contextual bandits, resolving a variant of the open problem posed by Agarwal et al. [5]. Let us briefly mention some extensions. First, we believe that our techniques can also be used to obtain first-order guarantees for stochastic contextual bandits with an *offline* log loss oracle (à la Simchi-Levi and Xu [58])—albeit with a more technical analysis. As another extension, in Appendix D we show how to use our method to efficiently obtain a first-order regret bound when working with rewards rather than losses. Such a guarantee is useful when no policy accumulates much reward, as is common in personalization applications. Several other extensions appear to be straightforward, including accommodating infinite action spaces [34].

We close with some directions for future work. Directly relevant to our theoretical results is to continue the investigation into adaptivity in contextual bandits and reinforcement learning. More broadly, while triangular discrimination has been used in various mathematics disciplines, we are not aware of many applications in algorithm design. Are there other uses for the triangular discrimination in machine learning? We look forward to pursuing these directions.

Acknowledgments and Disclosure of Funding

We thank Sivaraman Balakrishnan, John Langford, Zakaria Mhammedi, and Sasha Rakhlin for helpful discussions. We also thank Sasha Rakhlin for providing Google Cloud credits used for experiments.

References

[1] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, 1999.

[2] A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. E. Schapire. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[3] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

[4] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, S. Sen, and A. Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.

[5] A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, and R. E. Schapire. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, 2017.

[6] Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. *International Conference on Machine Learning*, 2018.

[7] C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, 2006.

[8] J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 2007.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

[10] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 2002.

[11] I. Benjamini, H. Duminil-Copin, G. Kozma, and A. Yadin. Disorder, entropy and harmonic functions. *Annals of Probability*, 2015.

[12] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[13] A. Bietti, A. Agarwal, and J. Langford. A contextual bandit bake-off. *arXiv:1802.04064*, 2018.

[14] B. Bilodeau, D. J. Foster, and D. Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, 2020.

[15] A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for K-fold and progressive cross-validation. In *Conference on Computational Learning Theory*, 1999.

[16] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[17] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, 2003.

[18] S. Bubeck and M. Sellke. First-order bayesian regret analysis of thompson sampling. In *International Conference on Algorithmic Learning Theory*, 2020.

[19] R. J. Carroll. Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, 1982.

[20] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Conference on Computational Learning Theory*, 1999.

[21] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[22] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 2007.

[23] S. Chen, F. Koehler, A. Moitra, and M. Yau. Online and distribution-free robustness: Regression and contextual bandits with Huber contamination. *arXiv:2010.04157*, 2020.

[24] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[25] T. M. Cover. Universal portfolios. *Mathematical Finance*, 1991.

[26] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013.

[27] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[28] A. Erschler and A. Karlsson. Homomorphisms to $\mathbb{R}$ constructed from random walks. *Annales de l'Institut Fourier*, 2010.

[29] D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

[30] D. J. Foster, A. Rakhlin, and K. Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems*, 2015.

[31] D. J. Foster, Z. Li, T. Lykouris, K. Sridharan, and É. Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, 2016.

[32] D. J. Foster, A. Agarwal, M. Dudík, H. Luo, and R. E. Schapire. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018.

[33] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018.

[34] D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 2020.

[35] D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, 2021.

[36] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

[37] E. Hazan and S. Kale. An online portfolio selection algorithm with regret logarithmic in price variation. *Mathematical Finance*, 2015.

[38] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007.

[39] S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 2020.

[40] A. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 2002.

[41] N. Karampatziakis and J. Langford. Online importance weight aware updates. In *Conference on Uncertainty in Artificial Intelligence*, 2011.

[42] W. M. Koolen and T. van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, 2015.

[43] A. Krishnamurthy, A. Agarwal, T.-K. Huang, H. Daumé III, and J. Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, 2017.

[44] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer, 1986.

[45] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on the World Wide Web*, 2010.

[46] Y. Li, Y. Wang, and Y. Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, 2019.

[47] H. Luo and R. E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Conference on Learning Theory*, 2015.

[48] H. Luo, C.-Y. Wei, and K. Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.

[49] T. Lykouris, K. Sridharan, and É. Tardos. Small-loss bounds for online learning with partial information. *Conference on Learning Theory*, 2018.

[50] G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, 2015.

[51] M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, 1999.

[52] L. Orseau, T. Lattimore, and S. Legg. Soft-bayes: Prod for mixtures of experts with log-loss. In *International Conference on Algorithmic Learning Theory*, 2017.

[53] N. Ozawa. A functional analysis proof of Gromov's polynomial growth theorem. *Annales Scientifiques de l'École Normale Supérieure*, 2018.

[54] D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 2002.

[55] A. Rakhlin and K. Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv:1501.07340*, 2015.

[56] S. Ross, P. Mineiro, and J. Langford. Normalized online learning. In *Uncertainty in Artificial Intelligence*, 2013.

[57] Y. M. Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 1987.

[58] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *arXiv:2003.12699*, 2020.

[59] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, 2010.

[60] A. Takeshi. *Advanced econometrics*. Harvard university press, 1985.

[61] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 2017.

[62] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 2000.

[63] S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[64] J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 2014.

[65] V. N. Vapnik and A. A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of Complexity*, 1971.

[66] I. Vincze. On the concept and measure of information contained in an observation. In *Contributions to Probability*. Elsevier, 1981.

[67] V. Vovk. A game of prediction with expert advice. In *Conference on Computational Learning Theory*, 1995.

[68] Y. Xu and A. Zeevi. Upper counterfactual confidence bounds: A new optimism principle for contextual bandits. *arXiv:2007.07876*, 2020.

[69] Y. Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 1999.

[70] A. Yehudayoff. Pointer chasing via triangular discrimination. *Combinatorics, Probability and Computing*, 2020.

[71] T. Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 2006.

[72] Z. Zhang, J. Yang, X. Ji, and S. S. Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture MDP. *Neural Information Processing Systems (NeurIPS)*, 2021.

[73] D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

A Further Related Work

First-order regret bounds have a long history throughout statistical learning [65, 54, 59], online learning [36, 10, 21, 22, 47, 42, 30], and bandits [7, 31, 5, 49, 6]. Below we highlight some of the most relevant lines of work.

Statistical learning and plug-in classification. Beginning with the work of Vapnik and Chervonenkis [65] for VC classes, classical work in statistical learning [54, 59] provides first-order regret (or, excess risk) bounds for *empirical risk minimization* which, in our setting, corresponds to the (typically intractable) policy optimization problem $\operatorname{argmin}_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(\pi(x_t))$. These results are also sometimes referred to as relative deviation bounds.

In the realizable setting (i.e., under Assumption 1), the process of fitting a model $\widehat{f}$ for the losses using regression and then performing classification with the induced classifier $\pi_{\widehat{f}}$ is often referred to as *plug-in classification* [69, 8, 26]. While these works establish worst-case optimal guarantees for plug-in classifiers, first-order regret bounds are—to the best of our knowledge—unexplored, and our observations regarding the suboptimality of least-squares and optimality of log loss regression are new.

Bandits. First-order regret bounds for multi-armed bandits appear in Allenberg et al. [7] (see also Foster et al. [31], Bubeck and Sellke [18]), and have been extended to the semi-bandit framework [50, 49] and linear bandits [39]. For contextual bandits, Agarwal et al. [5] show that many common algorithms fall short of achieving first-order regret, and we are not aware of any optimal first-order algorithms outside the solution of Allen-Zhu et al. [6], even if one disregards efficiency or considers additional assumptions such as realizability.

On the technical side, Bubeck and Sellke [18] provide first-order regret bounds for Thompson sampling for the multi-armed bandit in the Bayesian setting. Their approach takes advantage of a certain *nonnegative* χ^2*-divergence* which is closely related to the triangular discrimination we work with. Curiously, their analysis uses this divergence to measure distance between (posterior) distributions over actions, whereas we use the triangular discrimination to measure distance between regression functions. It would be interesting to understand whether there are deeper (e.g., primal-dual) connections between these approaches.

Fast rates under margin/gap conditions. Another line of work on plug-in classifiers aims for faster rates under various margin assumptions, and—similar to our work—observes that least-squares can be suboptimal in certain settings [8]. Fast rates based on margin conditions are distinct from first-order bounds (neither type of bound implies the other in general), but it would be interesting to understand their relationship more closely. Recent work [35] extends these developments to contextual bandits and provides logarithmic regret bounds based on similar gap/margin conditions. As in statistical learning, these types of guarantees are incomparable to first-order regret bounds.

Heteroscedastic regression. Our observations regarding suboptimality of least-squares for plug-in classification are also closely related to regression with heteroscedastic noise (Carroll [19]; Takeshi [60, Chapter 6]). Consider a regression setting where we receive variables $\{(x_i, y_i)\}_{i=1}^{n}$ i.i.d., with $y_i = f^\star(x_i) + \varepsilon_i$ for some $f^\star \in \mathcal{F}$, where $\mathbb{E}[\varepsilon_i \mid x_i] = 0$, and our goal is to produce an estimator such that the L_1-error $\mathbb{E}|\widehat{f}(x) - f^\star(x)|$ is small. In the heteroscedastic model, the noise variance $\sigma_x^2 := \mathbb{E}[\varepsilon_i^2 \mid x_i = x]$ may vary as a function of x. Using the same construction as Theorem 2, one can show that standard least-squares incurs error scaling with the worst-case variance $\sup_x \sigma_x^2$, while, if the variances were known, weighted least-squares with weights $w_x := 1/\sigma_x^2$ would yield error scaling with the more favorable average variance $\mathbb{E}\big[\sigma_x^2\big]$. Key to our results is that for responses in $[0, B]$, we have $\mathbb{E}\big[\sigma_x^2\big] \leq B \cdot \mathbb{E}[f^\star(x)]$ and, as we show, the logarithmic loss achieves error scaling with the latter quantity *without knowledge of the variances*. We mention in passing that regression with heteroscedastic noise has found recent use in the context of reinforcement learning with linear function approximation [73, 72].

B Proofs for Plug-In Classification Results (Section 3.1)

B.1 Proof of Theorem 2

Theorem 2 (Failure of least-squares for plug-in classification)**.** *Let $\mathcal{A} = \{1, 2\}$ and $\mathcal{X} = \{1, 2\}$. For every $n > 10^8$, there exists a function class $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \to [0, 1])$ with $|\mathcal{F}| = 2$, and a realizable distribution $\mathcal{D}$ such that $L^\star \leq \frac{2^7}{n} < 1$, yet $L(\widehat{\pi}_{\mathsf{LS}}) - L^\star \geq 2^{-5}\sqrt{\frac{1}{n}}$ with probability at least $1/10$.*

Proof. Let $\widehat{L}_{\mathsf{LS}}(f) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a\in\mathcal{A}}(f(x_t, a) - \ell_t(a))^2$ be the empirical square loss, so that $\widehat{f}_{\mathsf{LS}} = \mathrm{argmin}_{f\in\mathcal{F}}\,\widehat{L}_{\mathsf{LS}}(f)$. We adopt the shorthand $\varepsilon_n = 1/n$ throughout the proof.

Construction. We define $\mathcal{X} = \left\{x^{(1)}, x^{(2)}\right\}$ and $\mathcal{A} = \left\{a^{(1)}, a^{(2)}\right\}$, so that there are only two possible contexts and actions.

The data-generating process for our construction has three parameters, μ_n, ν_n, and p_n. We choose $\mathbb{P}_{\mathcal{D}}(x = x^{(1)}) = 1 - p_n$, and define $f^\star$ and the conditional loss distribution as follows:

- $f^\star(x^{(1)}, a^{(1)}) = \mu_n$ and $f^\star(x^{(1)}, a^{(2)}) = \nu_n$, where $\mu_n < \nu_n$. We choose $\ell(a^{(1)}) \sim \mathrm{Ber}(\mu_n) \mid x^{(1)}$ and $\ell(a^{(2)}) = \nu_n$ a.s. $\mid x^{(1)}$.
- $f^\star(x^{(2)}, a^{(1)}) = f^\star(x^{(2)}, a^{(2)}) = \frac{1}{2}$. We choose $\ell(a^{(1)}) \sim \mathrm{Ber}(\frac{1}{2}) \mid x^{(2)}$ and $\ell(a^{(2)}) = \frac{1}{2}$ a.s. $\mid a^{(2)}$.

We take $\mathcal{F} = \{f^\star, \tilde{f}\}$, where $\tilde{f}$ will be fully specified in the sequel, but is chosen to satisfy $\tilde{f}(x, a^{(2)}) = f^\star(x, a^{(2)})$ for all x. This, combined with the fact that $\ell(a^{(2)})$ is deterministic conditioned on x, means that our analysis will only concern the realized outcomes for $\ell(a^{(1)})$.

The high level idea for our construction is to set $p_n, \mu_n \propto \varepsilon_n = 1/n$, which ensures that $L^\star \leq (1 - p_n)\mu_n + p_n \lesssim \frac{1}{n}$, then show that if we choose $\tilde{f}(\cdot, a^{(1)}) \approx (\sqrt{\varepsilon_n}, 0)$, we have $\widehat{f}_{\mathsf{LS}} = \tilde{f}$ with constant probability. We then choose $\nu_n \approx \sqrt{\varepsilon_n}/2$, which implies that $\pi_{\tilde{f}}(x^{(1)}) = a^{(2)} \neq \pi^\star(x^{(1)})$, and consequently

$$L(\widehat{\pi}_{\mathsf{LS}}) = L(\pi_{\tilde{f}}) \gtrsim (1 - p_n) \cdot f^\star(x^{(1)}, \pi_{\tilde{f}}(x^{(1)})) = (1 - p_n) \cdot \nu_n \gtrsim \sqrt{\varepsilon_n}.$$

We make this approach formal below.

Bad event. Let n_1 and n_2 be the number of examples for which $x = x^{(1)}$ and $x = x^{(2)}$. Let $n_1(0)$ and $n_1(1)$ be the number of examples for which $x = x^{(1)}$ and $\ell(a^{(1)}) = 0$ or $\ell(a^{(1)}) = 1$, respectively, and let $n_2(0)$ and $n_2(1)$ be defined likewise. We restrict to $n \geq 4$ going forward so that $\varepsilon_n \leq 1/4$.

Let $\widehat{\mu}_1 = \frac{1}{n_1}\sum_{i:x_i = x^{(1)}}\ell(a^{(1)})$ (whenever $n_1 > 0$), and let $\widehat{\mu}_2$ be defined likewise.

We prove the following proposition, which states that a certain event that is unfavorable for the least-squares estimator occurs with constant probability.

Proposition 2. *Let $n \geq 256$. Then if we set $p_n = \varepsilon_n$ and $\mu_n = 2^7\varepsilon_n$, the following event holds with probability at least $1/10$.*

1. $n_2 = n_2(0) = 1$*, and in particular* $\widehat{\mu}_2 = 0$.
2. $n_1 \geq \frac{3}{8}n$.
3. $\widehat{\mu}_1 \leq \frac{3}{2}\mu_n$.

Going forward, we adopt the parameter setting in Proposition 2 and condition on the event in the proposition, which we denote by $\mathscr{E}$. Note that this parameter setting ensures that

$$L^\star = (1 - \varepsilon_n)f^\star(x^{(1)}, a^{(1)}) + \varepsilon_n f^\star(x^{(2)}, a^{(1)}) = (1 - \varepsilon_n)\mu_n + \frac{\varepsilon_n}{2} \leq 2^8\varepsilon_n,$$

as long as $\mu_n < \nu_n$.

Lower bound under the bad event. Next, we observe that for both $f \in \mathcal{F}$, since $f(x, a^{(2)})$ perfectly predicts $\ell(a^{(2)})$ for all x, we have

$$\widehat{L}_{\mathsf{LS}}(f) \equiv \frac{n_1}{n}(f(x^{(1)}, a^{(1)}) - \widehat{\mu}_1)^2 + \frac{n_2}{n}(f(x^{(2)}, a^{(1)}) - \widehat{\mu}_2)^2,$$

up to additive noise that depends only on the realization of the dataset, not on the function f under consideration. Since our argument only depends on the relative value of $\widehat{L}_{\mathsf{LS}}$, we identify $\widehat{L}_{\mathsf{LS}}$ with this representation going forward. We first observe that conditioned by Proposition 2 (Item 1), we have $\widehat{\mu}_2 = 0$, so that

$$\widehat{L}_{\mathsf{LS}}(f^\star) \geq \frac{n_2}{n}(f^\star(x^{(2)}, a^{(1)}) - \widehat{\mu}_2)^2 = \varepsilon_n \cdot (f^\star(x^{(2)}, a^{(1)}))^2 = \frac{\varepsilon_n}{4}.$$

Here we use that $n_2 = 1$ under the bad event and that $\varepsilon_n = 1/n$. On the other hand, if we set $\tilde{f}(x^{(2)}, a^{(1)}) = 0$, we have

$$\begin{aligned}\widehat{L}_{\mathsf{LS}}(\tilde{f}) = \frac{n_1}{n}(\tilde{f}(x^{(1)}, a^{(1)}) - \widehat{\mu}_1)^2 &\leq 2(\tilde{f}(x^{(1)}, a^{(1)}))^2 + 2\widehat{\mu}_1^2 \\ &\leq 2(\tilde{f}(x^{(1)}, a^{(1)}))^2 + 2^3\mu_n^2 \\ &\leq 2(\tilde{f}(x^{(1)}, a^{(1)}))^2 + 2^{17}\varepsilon_n^2,\end{aligned}$$

where we have used Proposition 2 (Item 3). Note that as long as $\varepsilon_n < 2^{-20}$, we have $2^{17}\varepsilon_n^2 < \varepsilon_n/8$. If this is satisfied, then by choosing $\tilde{f}(x^{(1)}, a^{(1)}) = \sqrt{\varepsilon_n/16}$, we have

$$\widehat{L}_{\mathsf{LS}}(\tilde{f}) < \frac{\varepsilon_n}{4} \leq \widehat{L}_{\mathsf{LS}}(f^\star),$$

and we conclude that $\widehat{f}_{\mathsf{LS}} = \tilde{f} \neq f^\star$ whenever $\mathscr{E}$ occurs.

To conclude, we set $\nu_n = \sqrt{\varepsilon_n}/8$. Since $\tilde{f}(x^{(1)}, a^{(2)}) = \nu_n$, we have $\tilde{f}(x^{(1)}, a^{(2)}) < \tilde{f}(x^{(1)}, a^{(1)})$, so that $\pi_{\tilde{f}}(x^{(1)}) = a^{(2)} \neq \pi^\star(x^{(1)})$; this choice satisfies $\mu_n < \nu_n$ as required as long as $\varepsilon_n < 2^{-20}$. Finally, we observe that

$$L(\pi_{\tilde{f}}) - L^\star = (1 - \varepsilon_n)(\nu_n - \mu_n) \geq \frac{1}{2}(\sqrt{\varepsilon_n}/8 - 2^7\varepsilon_n) > 2^{-5}\sqrt{\varepsilon_n},$$

as long as $\varepsilon_n < 2^{-22}$. □

Proof of Proposition 2. Let $\mathscr{E}_1$, $\mathscr{E}_2$, and $\mathscr{E}_3$ denote the respective events in Proposition 2. We lower bound their probabilities one by one.

Event $\mathscr{E}_1$. We calculate

$$\mathbb{P}(n_2 = 1) = \sum_{i=1}^{n} \varepsilon_n \cdot (1 - \varepsilon_n)^{n-1} = \frac{1}{1 - \varepsilon_n}(1 - \varepsilon_n)^{1/\varepsilon_n} \geq e^{-1},$$

where we have used that $(1 - 1/x)^x \geq e^{-1}(1 - 1/x)$ for $x \geq 1$. Hence, since $\ell(a^{(1)}) \sim \mathrm{Ber}(\frac{1}{2})$ given $x^{(2)}$, $\mathscr{E}_1$ happens with probability at least $1 - \delta_1$ for $\delta_1 := 1 - e^{-1}/2$.

Event $\mathscr{E}_2$. We recall a standard multiplicative variant of the Chernoff bound.

Lemma 2 (Chernoff bound (e.g., Boucheron et al. [16]))**.** *Let $Y_i \sim \mathrm{Ber}(\mu)$ i.i.d.. Then for any $x \in [0, 1/2]$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq (1 + x)\mu n\right) \vee \mathbb{P}\left(\sum_{i=1}^{n} Y_i \leq (1 - x)\mu n\right) \leq e^{-\frac{1}{4}x^2\mu n}.$$

As long as $p_n = 1/n \leq 1/4$, Lemma 2 implies that $n_1 \geq \frac{3n}{8}$ with probability at least $1 - e^{-\frac{3n}{64}} =: 1 - \delta_2$, so that event $\mathscr{E}_2$ holds.

Event $\mathscr{E}_3$. We observe that conditioned on the realization of $x_1, \ldots, x_n$, Lemma 2 implies that

$$\widehat{\mu}_1 \leq \frac{3}{2}\mu_n$$

with probability at least $1 - e^{-\frac{1}{16}\mu_n n_1}$. Conditioned on $\mathscr{E}_2$, this probability is at least $1 - e^{-\frac{3}{128}\mu_n n}$. Since $\mu_n = 128/n$, which is admissible whenever $n \geq 256$, we conclude that $\mathscr{E}_3$ holds with probability at least $1 - e^{-3} =: 1 - \delta_3$ given $\mathscr{E}_2$.

Wrapping up. Taking a union bound, we have that $\mathscr{E} = \bigcup_{i=1}^{3} \mathscr{E}_i$ occurs with probability at least $1 - \sum_{i=1}^{3} \delta_i \geq e^{-1}/2 - e^{-12} - e^{-3} \geq 1/10$.

□

B.2 Proof of Theorem 3

B.2.1 Overview of Results

Recall that we work in the plug-in classification setting of Section 3.1, where $\mathcal{X}$ is the feature/context space, $\mathcal{A}$ is the label/action space, and $\mathcal{D}$ is the joint distribution over context-loss pairs (x, ℓ). We take a class of regression functions $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{A} \to [0,1])$ as a given and make the following realizability assumption.

Assumption 3. *Define $f^\star(x,a) = \mathbb{E}_{\mathcal{D}}[\ell(a) \mid x]$. We assume $f^\star \in \mathcal{F}$.*

Under realizability, the optimal classifier is $\pi^\star(x) := \operatorname{argmin}_{a \in \mathcal{A}} f^\star(x,a)$, and we have $L(\pi) = \mathbb{E}[f^\star(x, \pi(x))]$. Motivated by realizability, the plug-in approach to classification finds and estimator $\widehat{f} \in \mathcal{F}$ and returns the induced classifier $\widehat{\pi}(x) := \operatorname{argmin}_{a \in \mathcal{A}} \widehat{f}(x,a)$. In this section, we estimate the losses using the following log loss regression problem.

$$\widehat{f}_{\mathsf{KL}} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \ell_i(a) \log(1/f(x_i,a)) + (1 - \ell_i(a)) \log(1/(1 - f(x_i,a))). \tag{19}$$

For the resulting classifier $\widehat{\pi}_{\mathsf{KL}} := \pi_{\widehat{f}_{\mathsf{KL}}}$, we prove the following theorem.

Theorem 3 (First-order regret bound for plug-in classification)**.** *Let $\delta \in (0,1)$. Suppose that Assumption 3 holds. Then with probability at least $1 - \delta$, we have*

$$L(\widehat{\pi}_{\mathsf{KL}}) - L^\star \leq 16\sqrt{\frac{L^\star \cdot A\left(\log|\mathcal{F}| + \log(A/\delta)\right)}{n}} + 68\frac{A\left(\log|\mathcal{F}| + \log(A/\delta)\right)}{n}.$$

Multiclass classification. We also provide a refinement of Theorem 3 for the important special case of multiclass classification. Here, rather than observing a cost function $\ell \in [0,1]^A$ we simply observe a label $y \in \mathcal{A}$ and the goal is to predict the correct label. Formally, the distribution $\mathcal{D}$ is supported on $\mathcal{X} \times \mathcal{A}$ and we measure the error of a classifier as $\operatorname{err}(\pi) := \mathbb{P}_{\mathcal{D}}[\pi(x) \neq y]$. This can be seen as a special case of cost-sensitive classification by defining loss function $\ell(a) = \mathbb{1}\{a \neq y\}$, and the realizability assumption is as before, so that $f^\star(x,a) = \mathbb{P}_{\mathcal{D}}[a \neq y \mid x]$.

In this setting, rather than reducing to Bernoulli MLE, it is more natural to reduce to multinomial MLE. Since our function class is designed to predict the probability that a given action is *wrong* (that is, $\mathbb{P}_{\mathcal{D}}[y = a \mid x] = 1 - f^\star(x,a)$), the multinomial MLE problem is

$$\widehat{f}_{\mathsf{KL}} \leftarrow \operatorname*{argmax}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \log(1 - f(x_i, y_i)).$$

The resulting policy is $\widehat{\pi}_{\mathsf{KL}} := \operatorname{argmin}_a \widehat{f}_{\mathsf{KL}}(x,a)$, for which we establish the following guarantee.

Theorem 5. *Let $\delta \in (0,1)$ and consider the multiclass classification setting under Assumption 3. Then with probability at least $1 - \delta$,*

$$\operatorname{err}(\widehat{\pi}_{\mathsf{KL}}) - \operatorname{err}(\pi^\star) \leq 8\sqrt{\frac{\operatorname{err}(\pi^\star) \cdot 2\log(|\mathcal{F}|/\delta)}{n}} + 34\frac{\log(|\mathcal{F}|/\delta)}{n}.$$

Compared to Theorem 3, we see that by working in the simpler multiclass classification setting, we can remove the dependence on A from the theorem.

B.2.2 Preliminaries

For discrete distributions $p, q \in \Delta_A$, the Hellinger distance is defined as

$$D^2_{\mathrm{H}}(p \,\|\, q) = \frac{1}{2}\sum_a (\sqrt{p_a} - \sqrt{q_a})^2.$$

For scalars $p, q \in [0,1]$ we overload notation and interpret $D^2_{\mathrm{H}}(p \,\|\, q) \equiv D^2_{\mathrm{H}}((p, 1-p) \,\|\, (q, 1-q))$ as the Hellinger divergence between the implied Bernoulli distributions. We similarly overload $D_{\triangle}(p \,\|\, q) \equiv D_{\triangle}((p, 1-p) \,\|(q, 1-q))$ as the Bernoulli triangular discrimination when given scalar arguments.

The following useful results relate Hellinger distance to the triangular discrimination for Bernoulli distributions and to a related quantity for multinomial distributions.

Proposition 3. *For all $p, q \in [0,1]$, we have*

$$D^2_{\mathrm{H}}(p \,\|\, q) \geq \frac{1}{4} D_{\triangle}(p \,\|\, q) \geq \frac{1}{4}\frac{(p-q)^2}{(p+q)}.$$

Proposition 4. *Let $p, q \in \Delta(\mathcal{A})$ be probability mass functions. Then*

$$\max_{a \in \mathcal{A}} \frac{(p_a - q_a)^2}{(1-p_a) + (1-q_a)} \leq 4 D^2_{\mathrm{H}}(p \,\|\, q).$$

B.2.3 Proof of Theorem 3 and Theorem 5

We focus on proving Theorem 3 and provide a sketch for Theorem 5, which is quite similar. For the former, the core of the argument is a generalization guarantee for $\widehat{f}_{\mathsf{KL}}$.

Theorem 6. *Under the conditions of Theorem 3, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{\mathcal{D}}\left[\sum_{a \in \mathcal{A}} \frac{(\widehat{f}_{\mathsf{KL}}(x,a) - f^\star(x,a))^2}{\widehat{f}_{\mathsf{KL}}(x,a) + f^\star(x,a)}\right] \leq \frac{4A\,(\log|\mathcal{F}| + \log(A/\delta))}{n}. \tag{20}$$

Theorem 6 builds on classical convergence results for maximum-likelihood estimators in well-specified settings, which provide bounds of the form

$$\mathbb{E}_{\mathcal{D}}\Big[D^2_{\mathrm{H}}(\widehat{f}_{\mathsf{KL}}(x,a) \,\|\, f^\star(x,a))\Big] \leq \mathcal{O}\bigg(\frac{\log(|\mathcal{F}|/\delta)}{n}\bigg)$$

for any fixed action [cf. 63, 71]. Theorem 6 follows quickly from this classical analysis by applying Proposition 3, which shows that the Hellinger divergence between Bernoulli distributions upper bounds the triangular discrimination that appears on the left-hand side of (20).

Theorem 3 immediately follows by combining Theorem 6 with the refined Cauchy-Schwarz lemma (Lemma 1) which we restate and prove here.

Lemma 1 (Regret decomposition for triangular discrimination)**.** *For any $f : \mathcal{X} \times \mathcal{A} \to [0,1]$,*

$$L(\pi_f) - L^\star \leq 8(L^\star \cdot \mathbb{E}_{\mathcal{D}}[D_{\triangle}(f^\star(x,\cdot) \,\|\, f(x,\cdot))])^{1/2} + 17\,\mathbb{E}_{\mathcal{D}}[D_{\triangle}(f^\star(x,\cdot) \,\|\, f(x,\cdot))]. \tag{15}$$

Proof of Lemma 1. Let $f \in \mathcal{F}$ be fixed. We first state a simple technical lemma.

Lemma 3. *For any function $f \in \mathcal{F}$ and policy $\pi : \mathcal{X} \to \mathcal{A}$,*

$$\mathbb{E}_{\mathcal{D}}[f^\star(x, \pi(x)) + f(x, \pi(x))] \leq \mathbb{E}_{\mathcal{D}}[D_{\triangle}(f^\star(x,\cdot) \,\|\, f(x,\cdot))] + 4L(\pi).$$

Going forward, define $\gamma(x,a) := f^\star(x,a) - f(x,a)$ and $s(x,a) := f^\star(x,a) + f(x,a)$, and $\Delta := \mathbb{E}_{\mathcal{D}}[D_{\triangle}(f^\star(x,\cdot) \,\|\, f(x,\cdot))]$. Let us adopt the shorthand $\mathbb{E} \equiv \mathbb{E}_{\mathcal{D}}$. We proceed to bound the cost-sensitive regret:

$$\begin{aligned}
L(\pi_f) - L(\pi^\star) &\leq \mathbb{E}[f^\star(x,\pi_f(x)) - f(x,\pi_f(x)) + f(x,\pi^\star(x)) - f^\star(x,\pi^\star(x))]\\
&\leq \mathbb{E}\left[\sqrt{\frac{\max\{s(x,\pi_f(x)), s(x,\pi^\star(x))\}}{\max\{s(x,\pi_f(x)), s(x,\pi^\star(x))\}}}\cdot(|\gamma(x,\pi_f(x))| + |\gamma(x,\pi^\star(x))|)\right]\\
&\leq \sqrt{\mathbb{E}[\max\{s(x,\pi_f(x)), s(x,\pi^\star(x))\}]}\cdot\left(\sum_{\pi\in\{\pi_f,\pi^\star\}}\sqrt{\mathbb{E}\left[\frac{|\gamma(x,\pi(x))|^2}{\max\{s(x,\pi_f(x)), s(x,\pi^\star(x))\}}\right]}\right)\\
&\leq \sqrt{\mathbb{E}[s(x,\pi_f(x)) + s(x,\pi^\star(x)))]}\cdot\left(\sqrt{\mathbb{E}\left[\frac{\gamma(x,\pi_f(x))^2}{s(x,\pi_f(x))}\right]} + \sqrt{\mathbb{E}\left[\frac{\gamma(x,\pi^\star(x))^2}{s(x,\pi^\star(x))}\right]}\right)\\
&\leq \sqrt{\mathbb{E}[(s(x,\pi_f(x)) + s(x,\pi^\star(x)))]}\cdot 2\sqrt{\mathbb{E}\left[\sum_a \frac{\gamma(x,a)^2}{s(x,a)}\right]}\\
&= \sqrt{\mathbb{E}[(s(x,\pi_f(x)) + s(x,\pi^\star(x)))]}\cdot 2\sqrt{\Delta}.
\end{aligned}$$

Here, the first inequality uses that $f(x,\pi_f(x)) \leq f(x,\pi^\star(x))$ by the definition of π_f. The second inequality introduces the s and γ quantities, while the third follows from Cauchy-Schwarz. In the fourth we use that $s(x,\pi_f(x)) \leq \max\{s(x,\pi_f(x)), s(x,\pi^\star(x))\}$ and analogously for $\pi^\star$. Finally we sum over all actions to eliminate the dependence on the policies to introduce the triangular discrimination Δ. Applying Lemma 3, we additionally observe that

$$\mathbb{E}\left[s(x,\pi_f(x)) + s(x,\pi^\star(x))\right] \leq 2\Delta + 4\left(L(\pi_f) + L(\pi^\star)\right).$$

After applying standard simplifications, this yields

$$\begin{aligned}
L(\pi_f) - L(\pi^\star) &\leq 2\sqrt{\Delta}\cdot\sqrt{2\Delta + 4(L(\pi_f) + L(\pi^\star))} \leq 2\sqrt{2}\Delta + 4\sqrt{L(\pi^\star)\Delta} + 4\sqrt{L(\pi_f)\Delta} \qquad (21)\\
&\leq 6\sqrt{2}\Delta + (L(\pi_f) + L(\pi^\star))/2.
\end{aligned}$$

Re-arranging, we deduce that $L(\pi_f) \leq 12\sqrt{2}\Delta + 3L(\pi^\star)$, and plugging this back into the first inequality in (21) gives

$$\begin{aligned}
L(\pi_f) - L(\pi^\star) &\leq 2\sqrt{\Delta}\cdot\sqrt{2\Delta + 4(L(\pi_f) + L(\pi^\star))} \leq 2\sqrt{\Delta}\cdot\sqrt{(2 + 48\sqrt{2})\Delta + 16L(\pi^\star)}\\
&\leq 8\sqrt{L(\pi^\star)\Delta} + 17\Delta.
\end{aligned}$$

□

Proof sketch for Theorem 5. The majority of the calculations in this proof are very similar to those of Theorem 3, so we highlight the two main differences. First, rather than use the triangular discrimination-type bound in Theorem 6, we use a Hellinger bound on the maximum likelihood estimate of the multinomial parameters. Specifically, using essentially the same argument as in Theorem 6, we can prove that with probability at least $1-\delta$,

$$\mathbb{E}_{\mathcal{D}}\left[D^2_{\mathrm{H}}(\widehat{p}(\cdot \mid x) \,\|\, p^\star(\cdot \mid x))\right] \leq \frac{2\log|\mathcal{F}|/\delta}{n},$$

where $p^\star(\cdot \mid x) := \mathbb{P}_{\mathcal{D}}[y = \cdot \mid x] = 1 - f^\star(x,\cdot)$ and $\widehat{p}(\cdot \mid x) := 1 - \widehat{f}_{\mathsf{KL}}(x,\cdot)$.

The second change concerns the way we bound the quantity

$$(\widehat{f}_{\mathsf{KL}}(x,\pi(x)) - f^\star(x,\pi(x)))^2/(\widehat{f}_{\mathsf{KL}}(x,\pi(x)) + f^\star(x,\pi(x))),$$

which is done throughout the proof of Lemma 1. Rather than naively introduce a sum over all actions as was done previously, we instead apply Proposition 4, which relates the multinomial Hellinger divergence to the triangular discrimination-type quantity above. As a result, for any policy π we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\frac{\widehat{f}_{\mathsf{KL}}(x,\pi(x)) - f^\star(x,\pi(x)))^2}{\widehat{f}_{\mathsf{KL}}(x,\pi(x)) + f^\star(x,\pi(x))}\right] &= \mathbb{E}_{\mathcal{D}}\left[\frac{(p^\star(\pi(x) \mid x) - \widehat{p}(\pi(x) \mid x))^2}{(1 - p^\star(\pi(x) \mid x)) + (1 - \widehat{p}(\pi(x) \mid x))}\right]\\
&\leq 2\mathbb{E}_{\mathcal{D}}\left[D^2_{\mathrm{H}}(\widehat{p}(\cdot \mid x) \,\|\, p^\star(\cdot \mid x))\right].
\end{aligned}$$

All other calculations are unaffected. □

B.2.4 Proofs for Supporting Results

Proof of Proposition 3. Observe that we can write

$$D^2_{\mathrm{H}}(p \,\|\, q) = \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + \frac{1}{2}(\sqrt{1-p} - \sqrt{1-q})^2.$$

For each of these terms, we create a difference of squares as follows

$$(\sqrt{x} - \sqrt{y})^2 = \frac{(x-y)^2}{(\sqrt{x}+\sqrt{y})^2} \geq \frac{(x-y)^2}{2(x+y)},$$

where the last inequality uses the fact that $2\sqrt{xy} \leq x + y$. Applying this argument to both terms yields the result. □

Proof of Proposition 4. This is an immediate consequence of the data processing inequality for Hellinger divergence and Proposition 3. Indeed, by data processing, we have

$$D^2_{\mathrm{H}}(p \,\|\, q) \geq D^2_{\mathrm{H}}((p_a, 1-p_a) \,\|\, (q_a, 1-q_a)),$$

since the latter is the distribution of the random variable $Y := \mathbb{1}\{X = a\}$ when $X \sim p$ (resp. q). Now that we have passed to the Bernoulli Hellinger divergence, we simply apply Proposition 3 and drop one of the two terms. □

Proof of Theorem 6. The initial steps of this proof parallel the classical analysis of maximum likelihood estimators [see, e.g., 71]. We start by establishing a symmetrization inequality. Let $D := \{(x_i, \ell_i)\}_{i=1}^n$ and $D' := \{(x'_i, \ell'_i)\}_{i=1}^n$ denote two i.i.d. datasets of n examples, let $C(f, D)$ be any function of a regression function f and dataset D, and let $\widehat{f}$ be any estimator that takes the dataset D and outputs a function in $\mathcal{F}$. We first show that

$$\mathbb{E}_D\left[\exp\Big(C(\widehat{f}(D), D) - \log \mathbb{E}_{D'}\Big[\exp(C(\widehat{f}(D), D'))\Big]\Big) - \log|\mathcal{F}|)\right] \leq 1. \tag{22}$$

This is a symmetrization inequality because it relates the "training error" $C(\widehat{f}(D), D)$ to the error $C(\widehat{f}(D), D')$ measured on the "ghost sample" D'. The unusual form of the expression involving the ghost sample is to accommodate the fact that C may be unbounded.

To prove (22), let μ denote the uniform distribution over $\mathcal{F}$, and observe that for any distribution $\widehat{\mu} \in \Delta(\mathcal{F})$ and any function $g : \mathcal{F} \to \mathbb{R}$, we have

$$\sum_{f \in \mathcal{F}} \widehat{\mu}(f) g(f) \leq \max_{f \in \mathcal{F}} g(f) \leq \log \sum_{f \in \mathcal{F}} \exp(g(f)) = \log\left(\mathbb{E}_{f \sim \mu} \exp(g(f))\right) + \log|\mathcal{F}|.$$

Now for any D we take $\widehat{\mu}(f) := \mathbb{1}\{f = \widehat{f}(D)\}$ and $g(f) := C(f, D) - \log \mathbb{E}_{D'} \exp(C(f, D'))$ to obtain

$$C(\widehat{f}(D), D) - \log \mathbb{E}_{D'} \exp(C(\widehat{f}(D), D')) \leq \log\left(\mathbb{E}_{f \sim \mu} \frac{\exp(C(f, D))}{\mathbb{E}_{D'} \exp(C(f, D'))}\right) + \log|\mathcal{F}|.$$

We will exponentiate this inequality and take expectation over the initial dataset D. When we do this, the first term on the right-hand side simplifies to

$$\mathbb{E}_D \exp\left(\log\left(\mathbb{E}_{f \sim \mu}\left[\frac{\exp(C(f, D))}{\mathbb{E}_{D'} \exp(C(f, D'))}\right]\right)\right) = \mathbb{E}_{f \sim \mu}\left[\frac{\mathbb{E}_D \exp(C(f, D))}{\mathbb{E}_{D'} \exp(C(f, D'))}\right] = 1.$$

Re-arranging, we obtain (22). With the exponential moment bound in (22), a standard application of the Chernoff method yields that for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ we have

$$-\log \mathbb{E}_{D'} \exp(C(\widehat{f}(D), D')) \leq -C(\widehat{f}(D), D) + \log|\mathcal{F}| + \log(1/\delta).$$

This high-probability bound holds for any fixed functional C. To apply it, for each $a \in \mathcal{A}$, we define

$$C_a(f, D) := -\frac{1}{2} \sum_{i=1}^n \ell_i(a) \log(f^\star(x_i, a)/f(x_i, a)) + (1 - \ell_i(a)) \log((1 - f^\star(x_i, a))/(1 - f(x_i, a))),$$

where $\ell_i(a)$ is defined as in (19). We apply the bound for each C_a, then take a union bound over all $a \in \mathcal{A}$ and sum up the resulting inequalities, which gives that with probability at least $1-\delta$,

$$\sum_{a\in\mathcal{A}} -\log \mathbb{E}_{D'} \exp(C_a(\widehat{f}(D), D')) \le \sum_{a\in\mathcal{A}} -C_a(\widehat{f}(D), D) + A\left(\log|\mathcal{F}| + \log(A/\delta)\right).$$

We will apply this inequality with $\widehat{f}_{\mathsf{KL}}$, which is the maximum likelihood estimate. Then, since $f^\star \in \mathcal{F}$ and $\widehat{f}_{\mathsf{KL}}$ minimizes the log loss, we have that $\sum_a -C_a(\widehat{f}_{\mathsf{KL}}(D), D) \le 0$. On the other hand, for each action $a \in \mathcal{A}$, the corresponding term on the left-hand side can be simplified to

$$-\log \mathbb{E}_{D'} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\ell'_i(a)\log\frac{f^\star(x'_i,a)}{\widehat{f}_{\mathsf{KL}}(x'_i,a)} + (1-\ell'_i(a))\log\frac{1-f^\star(x'_i,a)}{1-\widehat{f}_{\mathsf{KL}}(x'_i,a)}\right)\right)$$

Now, let $y'_i(a) \sim \mathrm{Ber}(\ell'_i(a))$. Then by Jensen's inequality, we have

$$\begin{aligned}
&\ge\; -n\log \mathbb{E}_{x',\ell'}\, \mathbb{E}_{y'|\ell'} \exp\left(-\frac{1}{2}\left(y'(a)\log\frac{f^\star(x',a)}{\widehat{f}_{\mathsf{KL}}(x',a)} + (1-y'(a))\log\frac{1-f^\star(x',a)}{1-\widehat{f}_{\mathsf{KL}}(x',a)}\right)\right)\\
&= -n\log \mathbb{E}_{x',\ell'}\, \mathbb{E}_{y'|\ell'}\left[\left(\frac{f^\star(x',a)}{\widehat{f}_{\mathsf{KL}}(x',a)}\right)^{-y'(a)/2}\left(\frac{1-f^\star(x',a)}{1-\widehat{f}_{\mathsf{KL}}(x',a)}\right)^{-(1-y'(a))/2}\right]\\
&= -n\log \mathbb{E}_{x'}\left[\sqrt{f^\star(x',a)\widehat{f}_{\mathsf{KL}}(x',a)} + \sqrt{(1-f^\star(x',a))(1-\widehat{f}_{\mathsf{KL}}(x',a))}\right].
\end{aligned}$$

Here the last line holds because the model is well-specified; in particular $\mathbb{P}[y'(a) = 1 \mid x'] = f^\star(x',a)$. Continuing, observe that for any random variables u, v taking values in $[0,1]$ we have

$$-\log \mathbb{E}\left[\sqrt{uv} + \sqrt{(1-u)(1-v)}\right] = -\log\left(1 - \mathbb{E}\left[1 - \sqrt{uv} - \sqrt{(1-u)(1-v)}\right]\right) \ge \frac{1}{2}\mathbb{E}\left[D^2_{\mathrm{H}}(u \,\|\, v)\right], \tag{23}$$

where the last step uses that $x \le -\log(1-x)$ for $x \in [0,1]$ along with the definition of the Hellinger divergence. Together, these inequalities establish that

$$\frac{1}{2}\sum_{a\in\mathcal{A}} \mathbb{E}_x\left[D^2_{\mathrm{H}}(f^\star(x,a) \,\|\, \widehat{f}_{\mathsf{KL}}(x,a))\right] \le \frac{A\left(\log|\mathcal{F}| + \log(A/\delta)\right)}{n}.$$

To conclude, we simply apply Proposition 3, which yields the result. □

Proof of Lemma 3. Let $f \in \mathcal{F}$ be fixed and define $\gamma(x,a) := f^\star(x,a) - f(x,a)$ and $s(x,a) := f^\star(x,a) + f(x,a)$. By the triangle inequality, the AM-GM inequality, and an application of Theorem 6, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[s(x,\pi(x))] &\le \mathbb{E}_{\mathcal{D}}|\gamma(x,\pi(x))| + 2L(\pi^\star)\\
&\le \mathbb{E}_{\mathcal{D}}\left[\sqrt{s(x,\pi(x))}\frac{|\gamma(x,\pi(x))|}{\sqrt{s(x,\pi(x))}}\right] + 2L(\pi^\star)\\
&\le \frac{1}{2}\mathbb{E}_{\mathcal{D}}[s(x,\pi(x))] + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\frac{\gamma(x,\pi(x))^2}{s(x,\pi(x))}\right] + 2L(\pi^\star)\\
&\le \frac{1}{2}\mathbb{E}_{\mathcal{D}}[s(x,\pi(x))] + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\sum_a \frac{\gamma(x,a)^2}{s(x,a)}\right] + 2L(\pi^\star)\\
&\le \frac{1}{2}\mathbb{E}_{\mathcal{D}}[s(x,\pi(x))] + \frac{1}{2}\mathbb{E}_{\mathcal{D}}[D_\Delta(f^\star(x,\cdot) \,\|\, f(x,\cdot))] + 2L(\pi^\star).
\end{aligned}$$

Re-arranging yields the result. □

C Proofs for Contextual Bandit Results (Section 2)

C.1 Online Regression Oracles

In this section we briefly formalize the notion of an online regression oracle sketched in the introduction and Assumption 2. The treatment here follows Foster and Rakhlin [29].

We consider the following model for the oracle $\mathbf{Alg}_{\mathsf{KL}}$.

For $t = 1, \ldots, T$:

- Nature selects context-action pair $(x_t, a_t) \in \mathcal{X} \times \mathcal{A}$.
- Algorithm produces prediction $\widehat{y}_t \in [0, 1]$.
- Nature selects outcome $y_t \in [0, 1]$.

We model the oracle as a sequence of mappings $\mathbf{Alg}_{\mathsf{KL}}^{(t)} : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^{t-1} \to [0, 1]$, so that $\widehat{y}_t = \mathbf{Alg}_{\mathsf{KL}}^{(t)}\big(x_t, a_t\,; \{(x_i, , a_i, y_i)\}_{i=1}^{t-1}\big)$ above. Any algorithm of this type induces a mapping

$$\widehat{y}_t(x, a) := \mathbf{Alg}_{\mathsf{KL}}^{(t)}\big(x, a\,; \{(x_i, , a_i, y_i)\}_{i=1}^{t-1}\big), \tag{24}$$

which may be understood as the prediction the algorithm would make at time t if we froze its internal state and selected $(x_t, a_t) = (x, a)$.

C.2 Proof of Theorem 1

Theorem 1 (Main theorem)**.** *Suppose Assumptions 1 and 2 hold. Then Algorithm 1 guarantees that for all sequences with* $\mathbb{E}\big[\sum_{t=1}^{T} \ell_t(\pi^\star(x_t))\big] \le L^\star$*, by choosing* $\gamma = \sqrt{AL^\star/3\mathbf{Reg}_{\mathsf{KL}}(T)} \vee 10A$*,*

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \le 40\sqrt{L^\star \cdot A\mathbf{Reg}_{\mathsf{KL}}(T)} + 600A\mathbf{Reg}_{\mathsf{KL}}(T). \tag{6}$$

Proof. Define $L_T = \sum_{t=1}^{T} \ell_t(a_t)$ and $L_T^\star = \sum_{t=1}^{T} \ell_t(\pi^\star(x_t))$. All of the effort in this proof will be to show that for any choice $\gamma \ge 10A$, Algorithm 1 has

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \le \frac{10A}{\gamma}\,\mathbb{E}[L_T^\star] + 28\gamma \cdot \mathbf{Reg}_{\mathsf{KL}}(T). \tag{25}$$

The bound in (6) immediately follows from this guarantee by using choice of γ in the theorem statement.

Define a filtration

$$\mathfrak{F}_{t-1} = \sigma((x_1, a_1, \ell_1(a_1)), \ldots, (x_{t-1}, a_{t-1}, \ell_{t-1}(a_{t-1})), x_t) \tag{26}$$

and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathfrak{F}_t]$. Next, define the following conditional-expected versions of the contextual bandit regret and log loss regret, respectively

$$\overline{\mathbf{Reg}}_{\mathsf{CB}}(T) = \sum_{t=1}^{T} \mathbb{E}_{t-1}[\ell_t(a_t) - \ell_t(\pi^\star(x_t))] = \sum_{t=1}^{T}\sum_{a} p_{t,a}(f^\star(x_t, a) - f^\star(x_t, \pi^\star(x_t)))$$

and

$$\overline{\mathbf{Reg}}_{\mathsf{KL}}(T) = \sum_{t=1}^{T} \mathbb{E}_{t-1}[\ell_{\log}(\widehat{y}_t(x_t, a_t), \ell_t(a_t)) - \ell_{\log}(f^\star(x_t, a_t), \ell_t(a_t))].$$

Our starting point is to observe that $\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] = \mathbb{E}\big[\overline{\mathbf{Reg}}_{\mathsf{CB}}(T)\big]$ and $\mathbb{E}\big[\overline{\mathbf{Reg}}_{\mathsf{KL}}(T)\big] \le \mathbf{Reg}_{\mathsf{KL}}(T)$, where the latter holds since $\mathbf{Reg}_{\mathsf{KL}}(T)$ is a deterministic upper bound on the log loss regret of the oracle. So it suffices to relate the conditional-expected versions of these quantities.

The main step of the proof is to upper bound $\overline{\mathbf{Reg}}_{\mathsf{CB}}(T)$, using the first-order per-round inequality Theorem 4 (proven in Appendix C.3), which we restate here for completeness.

Theorem 4 (First-order per-round inequality)**.** *Let $y \in [0,1]^A$ be given and $b \in \operatorname{argmin}_a y_a$. Define $p_a = \frac{y_b}{Ay_b + \gamma(y_a - y_b)}$ for $a \neq b$, and $p_b = 1 - \sum_{a \neq b} p_a$. If $\gamma \geq 2A$, then for all $f \in [0,1]^A$ and $a^\star \in \operatorname{argmin}_a f_a$, we have*

$$\underbrace{\sum_a p_a(f_a - f_{a^\star})}_{\text{CB regret}} \leq \underbrace{\frac{5A}{\gamma}\sum_a p_a f_a}_{\text{bias from \textit{exploring}}} + \underbrace{7\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}}_{\text{error from \textit{exploiting}}}. \tag{17}$$

Applying Theorem 4 for each round t, we are guaranteed that

$$\begin{aligned}\overline{\mathbf{Reg}}_{\mathsf{CB}}(T) &\leq \frac{5A}{\gamma}\sum_{t=1}^{T}\sum_a p_{t,a} f^\star(x_t,a) + 7\gamma \sum_{t=1}^{T}\sum_a p_{t,a}\frac{(\widehat{y}_t(x_t,a) - f^\star(x_t,a))^2}{\widehat{y}_t(x_t,a) + f^\star(x_t,a)} \\ &= \frac{5A}{\gamma}\overline{L}_T + 7\gamma \cdot \overline{\mathbf{Err}}_{\Delta}(T),\end{aligned}$$

where $\overline{L}_T := \sum_{t=1}^{T}\sum_a p_{t,a} f^\star(x_t,a)$ and

$$\overline{\mathbf{Err}}_{\Delta}(T) := \sum_{t=1}^{T}\sum_a p_{t,a}\frac{(\widehat{y}_t(x_t,a) - f^\star(x_t,a))^2}{\widehat{y}_t(x_t,a) + f^\star(x_t,a)}.$$

Next, we relate the triangular discrimination-type error $\overline{\mathbf{Err}}_{\Delta}(T)$ to the log loss regret using the following proposition (proven in the sequel).

Proposition 5. *If $y \in [0,1]$ is a random variable with $\mathbb{E}[y] = \mu$, then for any $\widehat{y} \in [0,1]$,*

$$\mathbb{E}[\ell_{\log}(\widehat{y}, y) - \ell_{\log}(\mu, y)] = d_{\mathrm{KL}}(\mu \,\|\, \widehat{y}) \geq \frac{1}{2} \cdot \frac{(\widehat{y} - \mu)^2}{\widehat{y} + \mu}. \tag{27}$$

In particular, since a_t and ℓ_t are conditionally independent given $\mathfrak{F}_{t-1}$, this implies that

$$\overline{\mathbf{Err}}_{\Delta}(T) \leq 2\sum_{t=1}^{T}\sum_a p_{t,a} d_{\mathrm{KL}}(f^\star(x_t,a) \,\|\, \widehat{y}_t(x,a_t)) = 2\overline{\mathbf{Reg}}_{\mathsf{KL}}(T),$$

so that

$$\overline{\mathbf{Reg}}_{\mathsf{CB}}(T) \leq \frac{5A}{\gamma}\overline{L}_T + 14\gamma \cdot \overline{\mathbf{Reg}}_{\mathsf{KL}}(T).$$

To conclude, let $\overline{L}_T^\star = \sum_{t=1}^{T} f^\star(x_t, \pi^\star(x_t))$. Then this inequality can be written as

$$\overline{L}_T - \overline{L}_T^\star \leq \frac{5A}{\gamma}\overline{L}_T + 14\gamma \cdot \overline{\mathbf{Reg}}_{\mathsf{KL}}(T).$$

Since $1/(1-\varepsilon) \leq 1 + 2\varepsilon$ for all $\varepsilon \leq 1/2$, this implies that whenever $\gamma \geq 10A$,

$$\overline{L}_T - \overline{L}_T^\star \leq \frac{10A}{\gamma}\overline{L}_T^\star + 28\gamma \cdot \overline{\mathbf{Reg}}_{\mathsf{KL}}(T).$$

Noting that $\mathbb{E}[\overline{L}_T^\star] = \mathbb{E}[L_T^\star]$ and $\mathbb{E}[\overline{L}_T] = \mathbb{E}[L_T]$, this establishes (25).

□

Proof of Proposition 5. For the equality in (27), we have

$$\mathbb{E}[\ell_{\log}(\widehat{y}, y) - \ell_{\log}(\mu, y)] = \mathbb{E}[y\log(\mu/\widehat{y}) + (1-y)\log((1-\mu)/(1-\widehat{y}))] = d_{\mathrm{KL}}(\mu \,\|\, \widehat{y}).$$

To prove the inequality, let $f_{\widehat{y}}(\mu) = d_{\mathrm{KL}}(\mu \,\|\, \widehat{y})$. By Taylor's theorem, we have

$$f_{\widehat{y}}(\mu) = f_{\widehat{y}}(\widehat{y}) + f_{\widehat{y}}'(\widehat{y})(\mu - \widehat{y}) + \frac{1}{2}f_{\widehat{y}}''(\bar{y})(\mu - \widehat{y})^2,$$

for some $\bar{y} \in \operatorname{conv}(\{\widehat{y}, \mu\})$. Observe that

$$f_{\widehat{y}}'(z) = \log(z/\widehat{y}) - \log((1-z)/(1-\widehat{y})),$$

so that we have $f_{\widehat{y}}(\widehat{y}) = f_{\widehat{y}}'(\widehat{y}) = 0$. Further

$$f_{\widehat{y}}''(\bar{y}) = \frac{1}{\bar{y}} + \frac{1}{1-\bar{y}} \geq \frac{1}{\max\{\widehat{y}, \mu\}} \geq \frac{1}{\widehat{y} + \mu},$$

which establishes the result. □

C.3 Proof of Theorem 4

Theorem 4 (First-order per-round inequality)**.** *Let $y \in [0,1]^A$ be given and $b \in \operatorname{argmin}_a y_a$. Define $p_a = \frac{y_b}{Ay_b + \gamma(y_a - y_b)}$ for $a \neq b$, and $p_b = 1 - \sum_{a \neq b} p_a$. If $\gamma \geq 2A$, then for all $f \in [0,1]^A$ and $a^\star \in \operatorname{argmin}_a f_a$, we have*

$$\underbrace{\sum_a p_a(f_a - f_{a^\star})}_{\text{CB regret}} \leq \underbrace{\frac{5A}{\gamma}\sum_a p_a f_a}_{\text{bias from } \textit{exploring}} + \underbrace{7\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}}_{\text{error from } \textit{exploiting}}. \tag{17}$$

Proof. To begin, we observe that by the AM-GM inequality,

$$\begin{aligned}\sum_a p_a(f_a - f_{a^\star}) &= \sum_{a \neq a^\star} p_a(y_a - f_{a^\star}) + \sum_{a \neq a^\star} p_a(f_a - y_a) \\ &\leq \sum_{a \neq a^\star} p_a(y_a - f_{a^\star}) + \frac{1}{4\gamma}\sum_{a \neq a^\star} p_a(f_a + y_a) + \gamma \sum_{a \neq a^\star} p_a \frac{(y_a - f_a)^2}{y_a + f_a}. \end{aligned} \tag{28}$$

We focus on bounding the first term in (28), then return to the other terms at the end of the proof. We have

$$\begin{aligned}\sum_{a \neq a^\star} p_a(y_a - f_{a^\star}) &= \sum_{a \neq a^\star} p_a(y_a - y_b) + (1 - p_{a^\star})(y_b - f_{a^\star}) \\ &= \sum_{a \notin \{a^\star, b\}} p_a(y_a - y_b) + (1 - p_{a^\star})(y_b - f_{a^\star}). \end{aligned} \tag{29}$$

Recall that for $a \neq b$ we set $p_a = \frac{y_b}{Ay_b + \gamma(y_a - y_b)}$ and for p_b we set $p_b = 1 - \sum_{a \neq b} p_a$. With this setting, the first term in (29) is bounded as

$$\sum_{a \notin \{a^\star, b\}} p_a(y_a - y_b) \leq \sum_{a \notin \{a^\star, b\}} \frac{y_b(y_a - y_b)}{Ay_b + \gamma(y_a - y_b)} \leq A\frac{y_b}{\gamma}. \tag{30}$$

It remains to bound the term

$$(1 - p_{a^\star})(y_b - f_{a^\star}).$$

If $f_{a^\star} \geq y_b$ this is trivially negative, so we assume going forward that $f_{a^\star} \leq y_b$, and upper bound as

$$(1 - p_{a^\star})(y_b - f_{a^\star}) \leq y_b - f_{a^\star}.$$

We now appeal to the following lemma.

Lemma 4. *The distribution p in Theorem 4 ensures that*

$$y_b - f_{a^\star} \leq \frac{A}{4\gamma} y_b + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}}. \tag{31}$$

Combining (28), (30), and (31), we arrive at the bound.

$$\sum_a p_a(f_a - f_{a^\star}) \leq \frac{1}{4\gamma}\sum_a p_a(f_a + y_a) + 2\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{2A}{\gamma} y_b. \tag{32}$$

To conclude, we relate the non-triangular terms above to $\sum_a p_a f_a$, which corresponds to the learner's expected loss. For the first term, we use the following basic result.

Lemma 5. *For any distribution $p \in \Delta_A$,*

$$\sum_a p_a y_a \leq 3\sum_a p_a f_a + \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

Applying this gives

$$\sum_a p_a(f_a - f_{a^\star}) \leq \frac{1}{\gamma}\sum_a p_a f_a + 3\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{2A}{\gamma} y_b,$$

where we have used that $\gamma \geq 1$ to simplify. Our final step is to relate the last term above to $f_{a^\star}$. To do this, we observe that if $\gamma \geq 2A$, then Lemma 4 implies (after rearranging), that

$$y_b \leq 2f_{a^\star} + 4\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}},$$

so that

$$\frac{2A}{\gamma} y_b \leq \frac{4A}{\gamma} f_{a^\star} + 8Ap_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}} \leq \frac{4A}{\gamma} f_{a^\star} + 4\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}}.$$

With this, we have

$$\sum_a p_a(f_a - f_{a^\star}) \leq \frac{1}{\gamma} \sum_a p_a f_a + 7\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{4A}{\gamma} f_{a^\star},$$

Finally, since $a^\star \in \operatorname{argmin}_a f_a$, we have $f_{a^\star} \leq \sum_a p_a f_a$, so we can simplify to

$$\sum_a p_a(f_a - f_{a^\star}) \leq \frac{5A}{\gamma} \sum_a p_a f_a + 7\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

□

C.3.1 Proofs for Supporting Lemmas

Proof of Lemma 4. Assume that $y_b \geq f_{a^\star}$, or else we are done. We consider two cases.

Case 1: $a^\star = b$. In this case, by the AM-GM inequality

$$y_b - f_{a^\star} = y_{a^\star} - f_{a^\star} \leq \frac{y_{a^\star} + f_{a^\star}}{8\gamma p_{a^\star}} + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}}.$$

Since $a^\star = b$, we have

$$p_{a^\star} = p_b = 1 - \sum_{a \neq b} \frac{y_b}{Ay_b + \gamma(y_a - y_b)} \geq 1/A,$$

so we can further upper bound by

$$\frac{A}{8\gamma}(y_{a^\star} + f_{a^\star}) + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}} \leq \frac{A}{4\gamma} y_{a^\star} + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}} = \frac{A}{4\gamma} y_b + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}},$$

where we have used that $f_{a^\star} \leq y_b = y_{a^\star}$, where the latter holds since, for this case, we are assuming $a^\star = b$.

Case 2: $a^\star \neq b$. Observe that in this case, we have

$$y_{a^\star} \geq y_b, \quad \text{and} \quad f_b \geq f_{a^\star}. \tag{33}$$

Since $a^\star \neq b$, using the definition of $p_{a^\star}$, we have

$$\begin{aligned} y_b - f_{a^\star} &= p_{a^\star} \frac{Ay_b + \gamma(y_{a^\star} - y_b)}{y_b}(y_b - f_{a^\star}) \\ &= Ap_{a^\star}(y_b - f_{a^\star}) + \gamma \cdot p_{a^\star} \frac{(y_{a^\star} - y_b)(y_b - f_{a^\star})}{y_b}, \end{aligned}$$

which we can rewrite as

$$y_b - f_{a^\star} = \underbrace{Ap_{a^\star}(y_b - f_{a^\star}) - \gamma \cdot p_{a^\star} \frac{(y_b - f_{a^\star})^2}{y_b}}_{\mathbf{A}} + \underbrace{\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})(y_b - f_{a^\star})}{y_b}}_{\mathbf{B}}.$$

For the term **A** above, we observe that by the AM-GM inequality,

$$Ap_{a^\star}(y_b - f_{a^\star}) \leq \frac{A^2}{4\gamma} p_{a^\star} y_b + \gamma p_{a^\star} \frac{(y_b - f_{a^\star})^2}{y_b}, \tag{34}$$

so that

$$\mathbf{A} \leq \frac{A^2}{4\gamma} p_{a^\star} y_b \leq \frac{A}{4\gamma} y_b,$$

where we have used that $p_{a^\star} \leq 1/A$ when $a^\star \neq b$.

Next, to bound $\mathbf{B}$, we observe that $y_{a^\star} \geq y_b \geq f_{a^\star} \geq 0$. Since the function $a \mapsto \frac{(a-b)}{a}$ is increasing for $a, b \geq 0$, we have that $\frac{(y_b - f_{a^\star})}{y_b} \leq \frac{(y_{a^\star} - f_{a^\star})}{y_{a^\star}}$ and consequently

$$\frac{(y_{a^\star} - f_{a^\star})(y_b - f_{a^\star})}{y_b} \leq \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star}} \leq 2\frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}},$$

where the second inequality uses that $y_{a^\star} \geq f_{a^\star}$.

Altogether, we have that when $a^\star \neq b$,

$$y_b - f_{a^\star} = \mathbf{A} + \mathbf{B} \leq \frac{A}{4\gamma} y_b + 2\gamma \cdot p_{a^\star} \frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}}. \tag{35}$$

The result now follows by combining the two cases. □

Proof of Lemma 5. First, we write

$$\sum_a p_a y_a = \sum_a p_a f_a + \sum_a p_a (y_a - f_a).$$

By the AM-GM inequality, we have

$$\sum_a p_a (y_a - f_a) \leq \frac{1}{2} \sum_a p_a (y_a + f_a) + \frac{1}{2} \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a},$$

so that

$$\sum_a p_a y_a \leq \frac{1}{2} \sum_a p_a y_a + \frac{3}{2} \sum_a p_a f_a + \frac{1}{2} \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a},$$

and after rearranging,

$$\sum_a p_a y_a \leq 3 \sum_a p_a f_a + \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

□

C.4 Auxiliary Results

Proposition 6. *When $\widehat{y}, y \in [0,1]$, the logarithmic loss $\widehat{y} \mapsto \ell_{\log}(\widehat{y}, y)$ is 1-exp-concave and 1-mixable.*

Proof of Proposition 6. Let $f_y(\widehat{y}) = \ell_{\log}(\widehat{y}, y)$. From Hazan et al. [38], the loss is α-exp-concave if and only if $f''_y(\widehat{y}) \geq \alpha(f'_y(\widehat{y}))^2$ for all $\widehat{y}, y \in [0,1]$. We observe that $f'_y(\widehat{y}) = -\frac{y}{\widehat{y}} + \frac{1-y}{1-\widehat{y}}$ and $f''_y(\widehat{y}) = \frac{y}{\widehat{y}^2} + \frac{1-y}{(1-\widehat{y})^2}$. Since $y \in [0,1]$, Jensen's inequality implies that

$$(f'_y(\widehat{y}))^2 \leq y\left(\frac{-1}{\widehat{y}}\right)^2 + (1-y)\left(\frac{1}{1-\widehat{y}}\right)^2 = f''_y(\widehat{y}),$$

so we may take $\alpha = 1$.

Mixability is an immediate consequence of exp-concavity [21]. □

D Extensions

D.1 Small Rewards

In this section we sketch an extension of FastCB to the setting where the learner observes rewards $r_t(a) \in [0,1]$ rather than losses $\ell_t(a)$, and aims to achieve high reward rather than low loss. As before, we assume access to a function class $\mathcal{F}$ such that the Bayes predictor $f^\star(x,a) := \mathbb{E}[r(a) \mid x] \in \mathcal{F}$. Formally, we define regret for this setting as

$$\mathbf{Reg}_{\mathsf{CB}}(T) = \sum_{t=1}^{T} r_t(\pi^\star(x_t)) - \sum_{t=1}^{T} r_t(a_t),$$

where $\pi^\star(x) := \operatorname{argmax}_{a\in\mathcal{A}} f^\star(x,a)$ is the optimal policy.

Our aim here is to provide regret bounds that adapt whenever the reward of the optimal policy is small. This type of guarantee is natural if we believe a-priori that rewards are typically very small, which is common in personalization and recommendation applications, where clicks are often used as reward signal, yet click-through rates are typically well below 1%. In such settings, it is favorable to have regret scaling with the reward $R^\star$ of the optimal policy. Note that this is *not* equivalent to an $L^\star$ bound after the translation $r_t(a) = 1 - \ell_t(a)$, since having low reward corresponds to having high loss.

FastCB can be adapted to the small-reward setting achieve

$$\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}(T)] \le \mathcal{O}\Big(\sqrt{R^\star \cdot A\mathbf{Reg}_{\mathsf{KL}}(T)} + A\mathbf{Reg}_{\mathsf{KL}}(T)\Big)$$

whenever $\mathbb{E}\Big[\sum_{t=1}^{T} r_t(\pi^\star(x_t))\Big] \le R^\star$. The algorithm remains essentially as described in Algorithm 1, with the only difference being that we change the reweighted inverse gap weighting strategy used in Line 6. The new strategy and corresponding per-round inequality are described in the following theorem.

Theorem 7. *Let $y \in [0,1]^A$ be given and $b := \operatorname{argmax}_a y_a$. Define $p_a = \frac{y_b}{Ay_b + \gamma(y_b - y_a)}$ for $a \neq b$ and $p_b = 1 - \sum_{a\neq b} p_a$. If $\gamma \ge 4A$, then for all $f \in [0,1]^A$ and $a^\star \in \operatorname{argmax}_a f_a$, we have*

$$\sum_a p_a(f_{a^\star} - f_a) \le \frac{9A}{\gamma}\sum_a p_a f_a + 10\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

Observe that the left hand side is the per-round regret of the learner when f is the *reward* (rather than loss) model, which contrasts with the left-hand side in Theorem 4. On the other hand, the right-hand side only differs from that of Theorem 4 in the constants. As such, it naturally yields an $R^\star$ bound when applied with $y = \widehat{y}_t(x_t,\cdot)$ as in Algorithm 1.

It should be noted that achieving $R^\star$-based first-order bounds for contextual bandits appears to be considerably easier than achieving $L^\star$-based bounds. Indeed, the standard analysis of the Exp4 algorithm already yields a $\mathcal{O}(\sqrt{R^\star \cdot A\log|\Pi|})$ regret bound, under the benign assumption that the policy class contains the policy that selects actions uniformly at random on every context [9, Theorem 7.1]. On the other hand, Exp4 cannot achieve an $L^\star$-based bound without modifications [6].

Proof of Theorem 7. The proof parallels that of Theorem 4. We start by adding and subtracting y_a and applying the AM-GM inequality

$$\begin{aligned}\sum_a p_a(f_{a^\star} - f_a) &= \sum_{a\neq a^\star} p_a(f_{a^\star} - y_a) + \sum_{a\neq a^\star} p_a(y_a - f_a)\\ &\le \sum_{a\neq a^\star} p_a(f_{a^\star} - y_a) + \frac{1}{4\gamma}\sum_{a\neq a^\star} p_a(y_a + f_a) + \gamma\sum_{a\neq a^\star} p_a\frac{(y_a - f_a)^2}{y_a + f_a}.\end{aligned}$$

For the first term above, let us consider two cases.

Case 1. First, if $y_b \ge f_{a^\star}$ then

$$\sum_{a\neq a^\star} p_a(f_{a^\star} - y_a) \le \sum_{a\notin\{a^\star,b\}} p_a(f_{a^\star} - y_a) \le \sum_{a\notin\{a^\star,b\}} p_a(f_{a^\star} - y_a)\mathbb{1}\{f_{a^\star} \ge y_a\}.$$

Here we have simply dropped negative terms. Now, using the definition of p_a for $a \neq b$, we have

$$p_a(f_{a^\star} - y_a)\mathbb{1}\{f_{a^\star} \geq y_a\} = \frac{y_b(f_{a^\star} - y_a)}{Ay_b + \gamma(y_b - y_a)}\mathbb{1}\{f_{a^\star} \geq y_a\} \leq \frac{y_b(f_{a^\star} - y_a)}{\gamma(y_b - y_a)}\mathbb{1}\{f_{a^\star} \geq y_a\}.$$

Observe that $y_b/(y_b - y_a) \leq f_{a^\star}/(f_{a^\star} - y_a)$, since $y_b \geq f_{a^\star} \geq y_a \geq 0$. This yields

$$\mathbb{1}\{f_{a^\star} \geq y_a\}\frac{y_b(f_{a^\star} - y_a)}{\gamma(y_b - y_a)} \leq \mathbb{1}\{f_{a^\star} \geq y_a\}\frac{f_{a^\star}(f_{a^\star} - y_a)}{\gamma(f_{a^\star} - y_a)} \leq \frac{f_{a^\star}}{\gamma}.$$

And so, if $y_b \geq f_{a^\star}$ we have the bound

$$\sum_a p_a(f_{a^\star} - f_a) \leq \frac{Af_{a^\star}}{\gamma} + \frac{1}{4\gamma}\sum_{a \neq a^\star} p_a(f_a + y_a) + \gamma \sum_{a \neq a^\star} p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

Case 2. If $y_b \leq f_{a^\star}$ then for the first term, we write

$$\sum_{a \neq a^\star} p_a(f_{a^\star} - y_a) = \sum_{a \notin \{a^\star, b\}} p_a(y_b - y_a) + (1 - p_{a^\star})(f_{a^\star} - y_b) \leq \sum_{a \notin \{a^\star, b\}} p_a(y_b - y_a) + (f_{a^\star} - y_b). \tag{36}$$

For the first term in (36), using the definition of p_a, we have

$$\sum_{a \notin \{a^\star, b\}} p_a(y_b - y_a) = \sum_{a \notin \{a^\star, b\}} \frac{y_b(y_b - y_a)}{Ay_b + \gamma(y_b - y_a)} \leq \sum_{a \notin \{a^\star, b\}} \frac{y_b}{\gamma} \leq \frac{Ay_b}{\gamma} \leq \frac{Af_{a^\star}}{\gamma}. \tag{37}$$

For the second term, we first note that $p_b = 1 - \sum_{a \neq b} p_a \geq 1 - \sum_{a \neq b} \frac{y_b}{Ay_b} \geq \frac{1}{A}$, then consider two subcases.

Case 2a ($y_b \leq f_{a^\star}$ and $a^\star = b$). Here we simply use the AM-GM inequality to show that

$$\begin{aligned} f_{a^\star} - y_b = f_{a^\star} - y_{a^\star} &\leq \frac{f_{a^\star} + y_{a^\star}}{8\gamma p_{a^\star}} + 2\gamma p_{a^\star}\frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}} \\ &\leq \frac{A}{4\gamma}f_{a^\star} + 2\gamma p_{a^\star}\frac{(y_{a^\star} - f_{a^\star})^2}{y_{a^\star} + f_{a^\star}}. \end{aligned}$$

Here the first inequality is AM-GM, while the second uses that $y_{a^\star} = y_b \leq f_{a^\star}$ (by the conditions for this case), along with the fact that $p_{a^\star} = p_b \geq 1/A$.

Case 2b ($y_b \leq f_{a^\star}$ and $a^\star \neq b$). In this case, we have

$$y_b \geq y_{a^\star}, \quad \text{and} \quad f_{a^\star} \geq f_b.$$

Using the definition for $p_{a^\star}$, we have

$$\begin{aligned} f_{a^\star} - y_b = p_{a^\star}\frac{Ay_b + \gamma(y_b - y_{a^\star})}{y_b}(f_{a^\star} - y_b) &= p_{a^\star}A(f_{a^\star} - y_b) + p_{a^\star}\gamma\frac{(y_b - y_{a^\star})(f_{a^\star} - y_b)}{y_b} \\ &\leq p_{a^\star}A(f_{a^\star} - y_b) + p_{a^\star}\gamma\frac{(f_{a^\star} - y_{a^\star})(f_{a^\star} - y_b)}{f_{a^\star}} \\ &= p_{a^\star}A(f_{a^\star} - y_b) + p_{a^\star}\gamma\frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star}} + p_{a^\star}\gamma\frac{(f_{a^\star} - y_{a^\star})(y_{a^\star} - y_b)}{f_{a^\star}} \\ &\leq p_{a^\star}A(f_{a^\star} - y_b) + p_{a^\star}\gamma\frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star}} \\ &\leq p_{a^\star}A(f_{a^\star} - y_{a^\star}) + p_{a^\star}\gamma\frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star}}. \end{aligned}$$

Here, in the first inequality we use that $a \mapsto (a - b)/a$ is increasing in a, for $a, b \geq 0$ along with the fact that $f_{a^\star} \geq y_b \geq y_{a^\star}$. The second and third inequalities both use that $y_{a^\star} \leq y_b$.

Now by the AM-GM inequality, we have

$$\begin{aligned}p_{a^\star} A(f_{a^\star} - y_{a^\star}) &\leq \frac{p_{a^\star} A^2}{4\gamma} f_{a^\star} + p_{a^\star}\gamma \frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star}} \\ &\leq \frac{A}{4\gamma} f_{a^\star} + p_{a^\star}\gamma \frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star}},\end{aligned}$$

where the second inequality uses the fact that $p_{a^\star} \leq 1/A$ since $a_\star \neq b$. Finally, we use that $y_{a^\star} \leq f_{a^\star}$ to conclude that in this case,

$$f_{a^\star} - y_b \leq \frac{A}{4\gamma} f_{a^\star} + 4\gamma p_{a^\star} \frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star} + y_{a^\star}}. \tag{38}$$

This bound applies to both Case 2a and 2b.

Wrapping up. Returning to Case 2 and combining (36), (37), and (38), we have

$$\sum_{a \neq a^\star} p_a (f_{a^\star} - y_a) \leq \frac{2Af_{a^\star}}{\gamma} + 4\gamma p_{a^\star} \frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star} + y_{a^\star}}.$$

Combining this with our initial calculation, we have

$$\begin{aligned}\sum_a p_a(f_{a^\star} - f_a) &\leq \frac{2Af_{a^\star}}{\gamma} + 4\gamma p_{a^\star} \frac{(f_{a^\star} - y_{a^\star})^2}{f_{a^\star} + y_{a^\star}} + \frac{1}{4\gamma}\sum_{a \neq a^\star} p_a(y_a + f_a) + \gamma \sum_{a \neq a^\star} p_a \frac{(y_a - f_a)^2}{y_a + f_a} \\ &\leq 4\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{1}{4\gamma}\sum_a p_a(y_a + f_a) + \frac{2A}{\gamma} f_{a^\star}.\end{aligned}$$

Next, we can apply Lemma 5 as-is, which yields

$$\sum_a p_a(f_{a^\star} - f_a) \leq \frac{1}{\gamma}\sum_a p_a f_a + 5\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{2A}{\gamma} f_{a^\star}.$$

This inequality, after using assumption the that $\gamma \geq 4A$ and rearranging, implies

$$f_{a^\star} \leq 2(1 + 1/\gamma)\sum_a p_a f_a + 10\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} \leq 4\sum_a p_a f_a + 10\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a}.$$

Plugging this bound in for the final expression gives

$$\begin{aligned}\sum_a p_a(f_{a^\star} - f_a) &\leq \frac{1}{\gamma}\sum_a p_a f_a + 5\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} + \frac{8A}{\gamma}\sum_a p_a f_A + 20A \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a} \\ &\leq \frac{9A}{\gamma}\sum_a p_a f_a + 10\gamma \sum_a p_a \frac{(y_a - f_a)^2}{y_a + f_a},\end{aligned}$$

as desired. □

E Details for Experiments

E.1 Assets and Computing Resources

Assets. The code for the contextual bandit evaluation setup of Bietti et al. [13], which we used as a starting point, is publicly available at `https://github.com/albietz/cb_bakeoff`. Likewise, the source code for Vowpal Wabbit, upon which our implementation is built, is publicly available at `https://github.com/vowpalwabbit/vowpal_wabbit/`. The source code used to run the experiments is included in the supplementary material.

All datasets used in the experiments are publicly available via the OpenML collection (`https://www.openml.org`). Readers can refer to the information page for each respective dataset (e.g., `https://www.openml.org/d/1041` for dataset `1041`) for copyright information.

Computing resources. Experiments were run on a single `n1-highcpu-32` instance on Google Compute Engine. The total compute time required to run the experiments was under 12 hours.

E.2 Additional Details

Datasets. We restrict to a subset of the bake-off suite consisting of 516 multiclass classification datasets in the same fashion as Foster et al. [35].

Algorithms and oracle. For SquareCB.L and FastCB.L we take $\mathcal{F}$ to be a class of generalized linear models:

$$\mathcal{F} = \left\{(x,a) \mapsto \sigma(\langle w, \phi(x,a)\rangle) \mid w \in \mathbb{R}^d\right\}, \tag{39}$$

where $\sigma(t) = 1/(1+e^{-t})$ is the logistic link function and $\phi(x,a)$ is a fixed (dataset-dependent) feature map. This choice is convenient because i) it naturally produces predictions in $[0,1]$, as required by FastCB, and ii), we have that $\ell_{\log}(\sigma(\langle w, \phi(x,a)\rangle), y) = \ell_{\text{logistic}}(\langle w, \phi(x,a)\rangle, y)$, so that online regression with the logarithmic loss is equivalent to online logistic regression (cf. Example 4).

Even though SquareCB is designed for the square loss rather than the log loss, one can show that under the realizability assumption (Assumption 1), any log loss oracle is admissible for SquareCB. Indeed, for any log loss oracle satisfying Assumption 2, realizability and Pinsker's inequality imply that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\widehat{y}_t(x_t,a_t) - f^\star(x_t,a_t))^2\right] \le 2\,\mathbb{E}\left[\sum_{t=1}^{T} d_{\mathrm{KL}}(f^\star(x_t,a_t) \,\|\, \widehat{y}_t(x_t,a_t))\right] \le 2\mathbf{Reg}_{\mathsf{KL}}(T), \tag{40}$$

which means that the oracle is a valid square loss oracle for SquareCB in the sense of Assumption 2b in Foster and Rakhlin [29].

The oracle is trained with the default VW learning rule, which performs online gradient descent with adaptive updates [27, 41, 56]. We treat the algorithm's step size parameter as a tunable hyperparameter.

For SquareCB.S, we configure SquareCB exactly as described in Foster et al. [35]. We take $\mathcal{F}$ to be the class of linear models

$$\mathcal{F} = \left\{(x,a) \mapsto \langle w, \phi(x,a)\rangle \mid w \in \mathbb{R}^d\right\},$$

and the oracle applies the default VW learning rule to the square loss. We use the same hyperparameter range as for SquareCB.L and FastCB.L, both for the SquareCB learning rate and for the VW learning rule's step size.

Tables in Figure 1. For both tables, each cell (a,b) plots the number of datasets in which algorithm a significantly beats b, minus the number of datasets in which b significantly beats a. Following Bietti et al. [13], we define a significant win using a heuristic based on an approximate Z-test. If p_a and p_b are the final PV loss values for algorithms a and b, respectively, we say that a significantly beats b if

$$1 - \Phi\left(\frac{p_a - p_b}{\sqrt{\frac{p_a(1-p_a)}{n} + \frac{p_b(1-p_b)}{n}}}\right) < 0.05, \tag{41}$$

where n is the number of examples and Φ is the Gauss error function.

In the left table, we choose the configuration (hyperparameters for SquareCB/FastCB and learning rate for the VW learner) with lowest final PV loss for each algorithm on a per-dataset basis. In the right table, for each algorithm we choose the hyperparameter configuration with best performance on a held-out collection of 200 datasets using the method described in Bietti et al. [13]. We keep this configuration fixed and tune only the learning rate for the VW learner on each dataset.

Plots in Figure 1. Each plot shows the progressive validation loss $L_{\mathsf{PV}}(t)$ as a function of the number of examples t, for the best-performing (in terms of final PV loss) hyperparameter configuration for each algorithm. We consider 10 replicates for each dataset, where each replicate has the example order randomly permuted, and plot the average progressive validation loss across the replicates. Error bands in each plot correspond to significance $p < 0.05$ under the Z-test in (41), setting $n = t \cdot (\#\text{replicates})$ at each time t.

The algorithm Supervised.L included in each of the plots is an oracle benchmark that runs online logistic regression using the true label for each example (which the bandit algorithms do not have access to). The only hyperparameter for this algorithm is the learning rate for the VW learning rule.

E.3 Additional Figures

Figure 2 shows the results for the experiment in Figure 1 (Top-Left) with two additional adaptive algorithms, AdaCB and RegCB, included. These algorithms were found to have the strongest overall performance on the bake-off suite in Foster et al. [35] using the same online square loss oracle as SquareCB.S, and are considered state-of-the-art [13, 35]. We see in that switching SquareCB from regression with the square loss to the logistic loss (SquareCB.L) is already enough to outperform AdaCB and RegCB, and that the performance of FastCB.L is even stronger. It would be interesting to understand how the performance of AdaCB and RegCB improves if we switch to the generalized linear model (39) in the same fashion as SquareCB.L/FastCB.L, but it is unclear how to efficiently compute the confidence sets required by these algorithms in this case. We leave this for future work.

$\downarrow$ vs $\rightarrow$	R.S	A.S	S.S	S.L	F.L
RegCB.S	-	6	46	-6	-12
AdaCB.S	-6	-	42	-8	-18
SquareCB.S	-46	-42	-	-55	-66
SquareCB.L	6	8	55	-	-11
FastCB.L	**12**	**18**	**66**	**11**	-

Figure 2: Head-to-head win-loss differences. Each entry indicates the statistically significant win-loss difference between the row algorithm and the column algorithm. Hyperparameters are per-dataset.