LEARNING UNDER ENDOGENOUS SELECTION BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployed classifiers often determine which outcomes become available for future training, creating endogenous selection bias. For example, a pretrial decision system observes whether defendants reoffend or fail to appear in court only for those released, while a lender observes repayment outcomes only for approved applicants. We study this phenomenon in binary classification, where a learner receives a batch of i.i.d. examples in each round and observes labels only for positive predictions. We first show that empirical risk minimization (ERM) on the observed labeled data can fail through self-confirming traps and self-reinforcing mistakes under noisy feedback. We then augment ERM with random exploration and show that, after accounting for initialization, its cumulative excess loss grows sublinearly in both the batch size and the number of rounds, in both realizable and agnostic settings. However, arbitrary exploratory decisions may be infeasible in high-stakes applications. We therefore also develop proper exploration algorithms that implement a single classifier from a prescribed hypothesis class in each round. These algorithms attain comparable guarantees, up to logarithmic factors and an additional cost linear in the batch size, which we show can be unavoidable under fixed initialization.

1 INTRODUCTION

Learning from selectively observed outcomes creates a fundamental tension between decision quality and information acquisition. A lender, for example, observes repayment outcomes only for approved applicants. A college observes academic performance at its institution only for applicants who are admitted and enroll, while a pretrial decision system observes post-release outcomes only for released defendants. In such settings, a conservative decision may avoid an immediate costly mistake, but it also prevents the learner from observing whether a rejected individual would have had a favorable outcome. Obtaining this information requires taking an action that may itself be costly or harmful.

A similar tension has been addressed in the online adversarial setting as Apple Tasting learning (Helmbold et al., 2000; Raman et al., 2024) (a learner aims to learn a classification rule for tasty apples, while minimizing the number of unpalatable apples they bite into. However, they can figure out only the classification of apples they tasted). However, in lending and pretrial decision-making, outcomes are delayed, and institutions may revise decision rules only periodically to maintain consistency across cases. In college admissions, decisions are made over a cohort of applicants. Motivated by these settings, we study a statistical batch model of endogenous selection bias. We investigate a setup of learning taking place in rounds: in each round, the learner receives a batch of n independently drawn instances, makes predictions over the batch, observes labels only for instances predicted positive, and can then update its prediction rule. To reflect realistic scenarios (such as acceptance of students to some program or pretrial bail release), we are especially interested in the regime where the number of rounds is considerably smaller than the batch size.

Before observing any labels, the learner relies on prior knowledge, which we represent by a given initial classifier h_0. We evaluate performance using *initialization-adjusted regret*: we subtract h_0's loss on the initial batch from the learner's cumulative loss and compare the remainder with the loss of the best fixed classifier in $\mathcal{H}$ on subsequent batches. Our main question is:

Can one design algorithms whose expected initialization-adjusted regret is sublinear in both the batch size n and the number of rounds T?

We first show that a basic ERM learner can produce classifiers with higher population loss than its initial classifier, even when the latter is Bayes optimal. Even with noiseless labels, ERM can remain trapped at a suboptimal classifier or, when the hypothesis class is misspecified, cycle indefinitely between suboptimal classifiers. We trace these failures under selective feedback to two underlying mechanisms: self-confirming traps and self-reinforcing mistakes under noisy labels.

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

To overcome these failures, we augment ERM with random exploration: in each batch, the learner predicts positively on a uniformly selected subset of instances to obtain labels. By carefully choosing the amount of exploration, we achieve expected initialization-adjusted regret that is sublinear in both the batch size and the number of rounds.

However, unrestricted random exploration may be unacceptable in high-stakes applications. For example, releasing randomly selected defendants solely to obtain feedback may conflict with institutional constraints. We model these constraints through a prescribed hypothesis class of permissible decision rules, which may encode requirements such as simplicity or interpretability, and require *proper* learning: in each round, the learner must implement a single classifier from this class.

Main contributions. Our main contributions are as follows.

1. **Failure of ERM.** We identify two failure modes arising from the endogenous selection of observed labels. First, ERM can become self-confirming: if the current classifier does not reveal evidence against itself, ERM has no reason to change and may remain trapped at a suboptimal classifier. We formalize this behavior through feedback stability. Even when the optimal hypothesis is uniquely feedback-stable, ERM can incur regret linear in n before convergence in the realizable setting, or cycle indefinitely in the agnostic setting (Proposition 1). Second, noisy labels can make mistakes self-reinforcing: early unlucky observations can cause ERM to reject a truly positive point, preventing further feedback that could correct the mistake. This can happen even when the learner starts at the Bayes optimal classifier (Observation 1). Both failure modes can lead to $\Omega(nT)$ expected initialization-adjusted regret.

2. **Learning through random exploration.** A natural first repair is to augment ERM with a small amount of random exploration. We randomly explore by predicting positively on uniformly selected instances in each batch and applying ERM to the resulting labeled sample. For classes of VC dimension d, this algorithm achieves $\widetilde{O}(\sqrt{dnT})$ expected initialization-adjusted regret in the realizable setting (Theorem 1). We prove an $\Omega(\sqrt{nT})$ lower bound already for a class of VC dimension one, showing that the dependence on nT is optimal up to logarithmic factors. In the agnostic setting, the algorithm achieves $O(d^{1/3}(nT)^{2/3})$ regret.

3. **Proper learning beyond unrestricted exploration.** However, in high-stakes applications such as pretrial decision-making, arbitrary exploration may be infeasible, and decision makers may be restricted to a prescribed class of classifiers. We therefore study proper learning, where the learner implements a single classifier from $\mathcal{H}$ in each round. Under prescribed initialization, we show that $\Omega(n\beta_0)$ regret is unavoidable in general, where β_0 is the probability mass of the disagreement region of $\mathcal{H}$ (Example 1). To achieve sublinear additional regret beyond this unavoidable cost, we develop Positively biased ERM (PBERM) and Capped Exploration ERM (CEERM). Both algorithms achieve $O(n\beta_0) + \widetilde{O}(\sqrt{dnT})$ regret in the realizable setting, while CEERM achieves $O(n\beta_0) + \widetilde{O}(d^{1/3}(nT)^{2/3})$ regret in the agnostic setting (Theorems 2 and 3).

Related work When $n = 1$, our model is a stochastic instance of apple tasting (Helmbold et al., 2000). Recent work characterizes online learnability and minimax rates (Raman et al., 2024) and studies deterministic learners (Chase and Mehalel, 2025). Related selective-labels research examines evaluation under selectively observed outcomes (Lakkaraju et al., 2017) and learning decision policies that account for selective feedback (Kilbertus et al., 2020; Wei, 2021). Pseudo-Label Optimism encourages exploration through optimistic pseudo-labels (Pacchiano et al., 2021), a motivation shared by our positivity-based methods. The theoretical guarantee of Pacchiano et al. (2021), however, assumes that the score-function class contains all constant functions, hence allowing an all-positive classifier. Our proper-learning guarantees also accommodate hypothesis classes that exclude such a classifier, as may be required in some applications.

Other variants incorporate statistical fairness constraints with i.i.d. arrivals (Bechavod et al., 2019) or strategic feature manipulation (Harris et al., 2023). Closely related to our protocol, Bechavod and Roth (2023) study batched one-sided feedback, applying one sampled classifier to the entire batch. Their setting allows adversarial arrivals and includes fairness auditing.

Jiang et al. (2021) also study batched one-sided feedback under a generalized linear response model, allowing stochastic or adversarial feature sequences. They analyze failures of greedy learning and bound a threshold-weighted decision loss relative to an oracle knowing the true model. Our work studies binary classification over general hypothesis classes of finite VC dimension, including arbitrary i.i.d. distributions in the agnostic setting. We compare against the best fixed classifier in the prescribed class, analyze both random and proper exploration, and quantify the effects of batch size, horizon, and prescribed initialization.

2 MODEL

Let $\mathcal{X}$ be the instance space and let $\mathcal{Y} = \{0, 1\}$. A classifier $h : \mathcal{X} \to \mathcal{Y}$ determines which individuals receive the positive action. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{D}_{\mathcal{X}}$ be its marginal over $\mathcal{X}$. At each round t, a fresh batch $S_t = \{(x_{t,1}, y_{t,1}), \ldots, (x_{t,n}, y_{t,n})\} \sim \mathcal{D}^n$ arrives. The learner observes the features $X_t = \{x_{t,1}, \ldots, x_{t,n}\}$, implements a classifier h_t, and then observes labels only for the examples classified as positive:

$$S_t^{\text{obs}}(h_t) = \{(x_{t,i}, y_{t,i}) : h_t(x_{t,i}) = 1\}.$$

The labels of examples with $h_t(x_{t,i}) = 0$ remain unobserved.

We consider the asymmetric classification loss $\ell(h, (x, y)) = c_{\text{fp}} \cdot \mathbb{1}\{h(x) = 1, y = 0\} + c_{\text{fn}} \cdot \mathbb{1}\{h(x) = 0, y = 1\}$, where $c_{\text{fp}}, c_{\text{fn}} \in (0, 1]$. We use $\ell_{\mathcal{D}}(h) := \mathbb{E}[\ell(h, X, Y)]$ to be the expected loss of h. Let $L_t(h) = \sum_{i=1}^{n} \ell(h, (x_{t,i}, y_{t,i}))$ denote the loss of h in round t, and let $\mathcal{L}_T = \sum_{t=1}^{T} L_t(h_t)$ denote the learner's cumulative loss. The learner begins from a fixed initial classifier h_0. We compare the learner to the strategy of using h_0 in the first round and then implementing a fixed $h \in \mathcal{H}$ in the next rounds. Thus, we evaluate performance using the initialization-adjusted regret

$$\mathcal{R}_T := \mathcal{L}_T - L_1(h_0) - \min_{h \in \mathcal{H}} \sum_{t=2}^{T} L_t(h)\,.$$

Our goal is then to achieve initialization-adjusted regret that is sublinear in both n and T in expectation. More precisely, we seek a bound $R(n, T)$ such that $\mathbb{E}[\mathcal{R}_T] \le R(n, T)$ and

$$\lim_{n \to \infty} \frac{R(n, T)}{n} = 0 \quad \text{for every fixed } T, \qquad \lim_{T \to \infty} \frac{R(n, T)}{T} = 0 \quad \text{for every fixed } n.$$

3 FAILURE MODES OF ERM

The most commonly used algorithmic principle is empirical risk minimization: minimize a loss function over the observed labeled data. Under data selection bias, however, the observed labeled data are themselves selected by the previously implemented classifiers. This creates two related failure modes. First, ERM can become self-confirming: if the current classifier does not reveal evidence against itself, ERM has no reason to change. Second, with noisy labels, early unlucky observations can permanently remove truly positive points from the positively-predicted region.

Let $S_{1:t}^{\text{obs}} := \bigcup_{\tau=1}^{t} S_\tau^{\text{obs}}(h_\tau)$ denote the cumulative observed labeled data. For $t \ge 2$, define

$$h_t^{\text{ERM}} \in \arg\min_{h \in \mathcal{H}} \sum_{(x,y) \in S_{1:t-1}^{\text{obs}}} \ell(h, (x, y)). \tag{1}$$

We adopt conservative tie-breaking rules: if h_{t-1} is an empirical risk minimizer, then the learner keeps $h_t^{\text{ERM}} = h_{t-1}$; otherwise, it chooses an arbitrary empirical risk minimizer according to a fixed deterministic tie-breaking rule.

3.1 SELF-CONFIRMING TRAPS

The first failure mode arises because ERM receives feedback only on the region currently labeled positive. If a classifier produces no observable evidence against itself, then the feedback generated under that classifier gives ERM no reason to move away from it.

Definition 1 (Feedback-stable hypothesis)**.** *For $g, h \in \mathcal{H}$, define the risk of g under the feedback induced by h as*

$$\ell_{\mathcal{D},h}(g) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(g, (x, y))\mathbb{1}\{h(x) = 1\}\right].$$

A hypothesis $h \in \mathcal{H}$ is feedback-stable *if*

$$h \in \arg\min_{g \in \mathcal{H}} \ell_{\mathcal{D},h}(g).$$

Feedback stability is a population fixed-point condition: when h is implemented, it minimizes the population risk under the feedback that it generates.

Let $h^\star \in \arg\min_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h)$ denote the optimal hypothesis in $\mathcal{H}$. In the realizable setting where $\ell_{\mathcal{D}}(h^\star) = 0$, since $h^\star$ incurs zero loss on every example, a hypothesis h is feedback-stable if and only if $\Pr_{(x,y)\sim\mathcal{D}}[h(x) = 1, y = 0] = 0$.

Thus, every label revealed while a feedback-stable hypothesis is implemented is positive almost surely. Consequently, any feedback-stable h has zero empirical loss on every feedback sample that it generates, and ERM does not switch away from it under the conservative tie-breaking rule.

In particular, whenever it belongs to $\mathcal{H}$, the all-negative classifier is feedback-stable. In the realizable setting, $h^\star$ is feedback-stable. Moreover, every classifier satisfying $h^{-1}(1) \subseteq h^{\star,-1}(1)$ is feedback-stable. Thus, unless additional restrictions are imposed, ERM can become trapped at a strict under-approximation of the target positive region.

For finite hypothesis classes, if $h^\star$ is the unique feedback-stable hypothesis in $\mathcal{H}$, ERM converges almost surely to $h^\star$ from every initialization. Indeed, if another feedback-stable hypothesis h exists, then ERM initialized at h never leaves it. Conversely, if $h^\star$ is the unique feedback-stable hypothesis, then every $h \neq h^\star$ has a false-positive region of positive probability. Once a false positive of h is observed, h incurs positive cumulative empirical loss, whereas $h^\star$ continues to have zero loss, so h can never be selected again. Since $\mathcal{H}$ is finite, ERM undergoes at most $|\mathcal{H}| - 1$ such informative eliminations before reaching $h^\star$.

However, uniqueness of the feedback-stable hypothesis is a strong condition. But even under this condition, ERM may still converge slowly or get trapped.

Proposition 1 (Self-confirming traps of ERM)**.** *Let $n \geq 1$ and $T \geq 2$. Tie-breaking is conservative: the current hypothesis is retained whenever it is an empirical risk minimizer.*

- *(i)* ERM can get trapped when $h^\star$ is not the unique feedback-stable hypothesis. *There exists a hypothesis class $\mathcal{H}$ with $|\mathcal{H}| = 3$ such that, for every conservative tie-breaking rule, there exist a realizable distribution $\mathcal{D}$ and an initialization h_0 for which ERM retains h_0 forever and satisfies $\mathbb{E}[\mathcal{R}_T] = \Omega(nT)$.*
- *(ii)* ERM can incur regret linear in the batch size despite eventual convergence in the realizable setting. *There exists a hypothesis class $\mathcal{H}$ with $|\mathcal{H}| = 3$ such that, for every conservative tie-breaking rule, there exist a realizable distribution $\mathcal{D}$ and an initialization h_0 for which $h^\star$ is the unique feedback-stable hypothesis in $\mathcal{H}$, and ERM converges almost surely to $h^\star$ after at most two informative updates while satisfying $\mathbb{E}[\mathcal{R}_T] = \Omega(n)$.*
- *(iii)* ERM can cycle in the agnostic setting. *There exists a hypothesis class $\mathcal{H}$ with $|\mathcal{H}| = 3$ such that, for every conservative tie-breaking rule, there exist a distribution $\mathcal{D}$ and an initialization h_0 for which $h^\star$ is both the unique population-risk minimizer and the unique feedback-stable hypothesis in $\mathcal{H}$, yet ERM almost surely switches infinitely often between two suboptimal hypotheses and satisfies $\mathbb{E}[\mathcal{R}_T] = \Omega(nT)$.*

The failure in Proposition 1 is therefore not ordinary overfitting. Rather, it stems from the endogenous selection of the observed data. ERM can either eliminate incorrect hypotheses sequentially or, in the agnostic setting, sustain a cycle through the aggregate feedback generated by several hypotheses, even when none of the hypotheses in the cycle is individually feedback-stable.

3.2 Self-reinforcing mistakes under noisy labels

The second failure mode appears when labels are noisy. Even if the learner is initialized at the optimal classifier, ERM over the observed data can permanently remove truly positive points from the positively predicted region. For this subsection, consider equal costs for false positives and false negatives, $c_{\mathrm{fp}} = c_{\mathrm{fn}} = 1$.

Consider the following example. Let $\mathcal{X} = \{1, \ldots, N\}$, and let $\mathcal{D}_{\mathcal{X}} = \mathrm{Unif}(\mathcal{X})$. Let $f^\star(x) = \mathbb{1}\{x > N/2\}$ be the Bayes optimal classifier. Labels are noisy: for each x,

$$y = \begin{cases} f^\star(x), & \text{with probability } 1 - \lambda(x), \\ 1 - f^\star(x), & \text{with probability } \lambda(x), \end{cases}$$

for some noise function λ with $\lambda(x) < 1/2$ for all x. Let $\mathcal{H} = \{0,1\}^{\mathcal{X}}$ be the class of all binary labeling functions, and initialize the learner at the Bayes optimal classifier $h_0 = f^\star$. At each point x, ERM predicts according to the majority of the labels observed so far at x. We use pointwise conservative tie-breaking: when the observed positive and negative label counts are equal, the learner retains its previous prediction.

Observation 1 (Noisy-label lockout in the finite-domain model)**.** *Under the finite-domain construction above and pointwise conservative tie-breaking, for every $x > N/2$ with $\mathcal{D}_{\mathcal{X}}(x) > 0$, ERM eventually switches its prediction at x from positive to negative with probability at least $\frac{\lambda(x)}{1-\lambda(x)}(1 - 1/N)^{n-1}$. This lower bound approaches $\lambda(x)/(1-\lambda(x))$ as $n/N \to 0$. Once this happens, the learner never observes labels at x again.*

Suppose $N \gg n$. For a point $x > N/2$, accepting has expected loss $\lambda(x)$, whereas rejecting has expected loss $1 - \lambda(x)$. Every locked-out point therefore contributes expected excess loss $1 - 2\lambda(x) > 0$ on each subsequent appearance.

Since $f^\star$ is Bayes optimal and belongs to $\mathcal{H}$, Observation 1 yields

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\mathcal{R}_T\right]}{T} \gtrsim \frac{n}{N} \sum_{x>N/2} \frac{\lambda(x)\left(1-2\lambda(x)\right)}{1-\lambda(x)}.$$

Thus, whenever $\lambda(x) > 0$ for any $x > N/2$, early unlucky noise can be amplified into permanent false negatives and linear-in-T expected excess loss. When $\lambda(x) = \lambda$ for all $x > N/2$, then we have $\liminf_{T\to\infty} \frac{\mathbb{E}[\mathcal{R}_T]}{T} \gtrsim \frac{n}{2} \cdot \frac{\lambda(1-2\lambda)}{1-\lambda}$. Thus, for fixed $\lambda \in (0, 1/2)$, the expected initialization-adjusted regret is $\Omega(nT)$ as $T \to \infty$.

The two failure modes above motivate explicit exploration. In low-stakes settings, one may randomly label additional points as positive to obtain labels. In high-stakes applications such as pre-trial decision making, however, arbitrary random exploration is not acceptable. This motivates the restricted-exploration setting studied next, where the learner must remain within an interpretable or institutionally feasible class of proper classifiers.

4 Algorithms

The results in the previous section show that ERM can fail under such endogenous data selection bias. The failures come from the fact that ERM is passive: it only learns from the positive region predicted by the current classifier. To obtain positive results, we therefore need alternative algorithms that deliberately collect more labels.

We first study random exploration. In each round, beyond implementing the current classifier, the learner may also select some points that would otherwise be predicted negative and label them positive for the purpose of exploration. This is the most direct way to break the self-confirming traps identified above.

4.1 Random Exploration

Here we allow the learner a budget of k explorations in each round. Let $E_t \subseteq X_t$ denote the set of exploration points in round t, then $|E_t| \le k$ (we also allow $\mathbb{E}\left[|E_t|\right] \le k$). If the learner's current classifier is h_t, then the implemented positive set is $P_t = \{x \in X_t : h_t(x) = 1\} \cup E_t$. The learner observes labels for all points in P_t. The exploration points may create additional false positives, but the additional loss caused by exploration is always at most $|E_t|$ in each round. We now describe our algorithm: on each round $t \in [T]$ select E_t uniformly from X_t. Let $S_0^{\exp} = \emptyset$. For $t \ge 1$:

$$S_t^{\exp} := S_{t-1}^{\exp} \cup \{(x, y) : x \in E_t\}$$

is the set of explored instances. On round $t = 1$ implement the given initial hypothesis, i.e., $h_1 = h_0$. On round $t > 1$ implement ERM over the labeled exploration set $S_{t-1}^{\exp}$. That is,

$$h_t \in \arg\min_{h\in\mathcal{H}} \sum_{(x,y)\in S_{t-1}^{\exp}} \ell(h, x, y).$$

The point of using $S_t^{\exp}$ is that the exploration examples are uniformly sampled from the incoming batches, and hence form an unbiased labeled sample from $\mathcal{D}$. The labels revealed by exploitation may also be useful, but they are selection-biased; the above algorithm ignores them in the ERM step.

Theorem 1 (Random exploration for VC classes)**.** *Let $d = \mathrm{VCdim}(\mathcal{H}) \ge 1$, $T \ge 2$, and $c_{\max} = \max\{c_{\mathrm{fn}}, c_{\mathrm{fp}}\}$, where $c_{\mathrm{fn}}, c_{\mathrm{fp}} > 0$.*

1. *In the realizable setting, random exploration with budget $k = \sqrt{\frac{c_{\max} n d \log(nT) \log T}{c_{\mathrm{fp}} T}}$. guarantees regret*

$$\mathbb{E}\left[\mathcal{R}_T\right] \le O\left(\sqrt{c_{\mathrm{fp}} c_{\max} n T d \log(nT) \log T}\right).$$

2. *For fixed positive costs and sufficiently large nT, there exist a domain with $|\mathcal{X}| = \Theta\left(\sqrt{\frac{c_{\mathrm{fn}} nT}{c_{\mathrm{fp}}}}\right)$, a hypothesis class $\mathcal{H}$ of singleton classifiers, and an initial hypothesis $h_0 \in \mathcal{H}$ such that, for every algorithm, there exists a realizable distribution for which*

$$\mathbb{E}\left[\mathcal{R}_T\right] \ge \Omega\left(\sqrt{c_{\mathrm{fn}} c_{\mathrm{fp}} nT}\right).$$

3. *In the agnostic setting, random exploration with budget $k = \left(\frac{c_{\max}^2 n^2 d}{c_{\mathrm{fp}}^2 T}\right)^{1/3}$ achieves*

$$\mathbb{E}\left[\mathcal{R}_T\right] \le O\left(c_{\mathrm{fp}}^{1/3} c_{\max}^{2/3} (nT)^{2/3} d^{1/3}\right).$$

4.2 Proper Learning with Restricted Exploration

Random exploration gives a direct way to break the failure modes of ERM, but it is not always feasible. In high-stakes applications such as pre-trial decision making, randomly releasing defendants for the purpose of exploration may be unacceptable. Moreover, decision makers often restrict themselves to simple and interpretable rules. This motivates a more constrained form of exploration: the learner must remain proper, in the sense that in every round it implements some classifier $h_t \in \mathcal{H}$. This restriction becomes more significant when $n > 1$ since there might not exist a hypothesis that labels the entire batch positively so we cannot have unrestricted exploration. However, our observations from Proposition 1 suggest that any ERM can incur linear in n regret as it can get trapped in a bad state. The following example further demonstrates that there are cases where any learner that starts with the given h_0 will incur regret that depends linearly on $\beta_0 n$, where $\beta_0 = \mathcal{D}_{\mathcal{X}}(\mathrm{DIS}(\mathcal{H}))$ is the mass of the disagreement region. Here $\mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H}, h(x) \neq h'(x)\}$.

Example 1. *Assume $T \geq 2$. Let $\mathcal{X} = \{0, 1, 2, 3\}$, $\mathcal{H} = \{f_i(x) = \mathbb{1}\{x = i\} : i \in [3]\}$, and suppose $c_{\mathrm{fp}} = c_{\mathrm{fn}} = 1$. For every $\beta_0 \in (0, 1]$, every proper learner, and every initial hypothesis $h_0 \in \mathcal{H}$, there exists a realizable distribution $\mathcal{D}$ with $\mathcal{D}_{\mathcal{X}}(\mathrm{DIS}(\mathcal{H})) = \beta_0$ such that $\mathbb{E}[\mathcal{R}_T] \geq n\beta_0/3$.*

Proof of Example 1. For $j \in [3]$, let $\mathcal{D}_j$ have marginal $\mathcal{D}_{\mathcal{X}}(0) = 1 - \beta_0$ and $\mathcal{D}_{\mathcal{X}}(i) = \beta_0/3$ for $i \in [3]$, with labels $Y = f_j(X)$. By symmetry, assume $h_0 = f_1$. In the initial round, labels are revealed only at $x = 1$, where they are zero under both $\mathcal{D}_2$ and $\mathcal{D}_3$. Couple these environments using the same initial features, second-round features, and learner randomness. Their initialization transcripts coincide, so their second-round predictions coincide, even if the learner uses X_2 when selecting h_2. For every x and every common prediction rule h, $\ell(h, (x, f_2(x))) + \ell(h, (x, f_3(x))) \geq \mathbb{1}\{x \in \{2, 3\}\}$. Summing over the second batch and taking expectations gives

$$\mathbb{E}_2[L_2(h_2)] + \mathbb{E}_3[L_2(h_2)] \geq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\{X_{2,i} \in \{2, 3\}\}\right] = \frac{2n\beta_0}{3}.$$

Thus, in at least one environment, $\mathbb{E}[L_2(h_2)] \geq n\beta_0/3$. The target has zero loss on every realization, and subsequent losses are nonnegative. Since $\mathcal{R}_T$ excludes only the initialization loss, the claimed bound follows. □

This result shows that our original goal of achieving regret sublinear in n is impossible with the given restrictions and a linear term in $\beta_0 n$ is inevitable. Therefore, here we seek a bound of the form $\mathbb{E}[\mathcal{R}_T] \leq \Theta(n\beta_0) + R(n, T)$, where $R(n, T)$ is sublinear in both n and T.

The idea of doing exploration under proper learning restrictions is favoring hypotheses that predict more positives and yet fit the observed data well. Following this intuition, we design two algorithms that balance between exploration and exploitation in two different ways.

Positively biased ERM (PBERM) The first idea is exploration by breaking ties in favor of hypotheses with more positive predictions on sampled instances. For $t \geq 2$, let $\mathcal{U}_t = \bigcup_{j=1}^{t-1} X_j$ be the multiset of instances received before round t, so $|\mathcal{U}_t| = n(t-1)$. Let $V_t := \{h \in \mathcal{H} | \forall (x, y) \in S^{\mathrm{obs}}_{1:t-1}, h(x) = y\}$ be the version space before round t, consisting of the hypotheses consistent with all observed labels, including those revealed during initialization. *Positively biased ERM* (PBERM) selects

$$h_t^{\mathrm{PBERM}} \in \arg\max_{h \in V_t} \sum_{x \in \mathcal{U}_t} h(x).$$

Since the unlabeled instances are i.i.d. from $\mathcal{D}$ we can think of the above maximization as an empirical approximation to maximizing $\mathbb{E}[h(X)] = \Pr_{\mathcal{D}}[h(X) = 1]$. Thus, PBERM favors hypotheses that on average predict more positives.

Define $c_{\max} = \max\{c_{\mathrm{fp}}, c_{\mathrm{fn}}\}$. Let $N = nT$, $d = \mathrm{VCdim}(\mathcal{H})$, $\overline{d} = \min\{d, N\}$, and

$$Q = \min\left\{\log(2|\mathcal{H}|),\ 1 + \overline{d}\log\left(\frac{eN}{\overline{d}}\right)\right\},$$

where the cardinality term is infinite when $\mathcal{H}$ is infinite, and the product involving $\overline{d}$ is zero when $d = 0$. Theorem 2 shows that PBERM achieves the desired regret bound.

Theorem 2 (PBERM in realizable setting)**.** *Suppose $Y = h^*(X)$ almost surely for some $h^* \in \mathcal{H}$, and let $c_{\mathrm{fp}}, c_{\mathrm{fn}} > 0$. Then PBERM achieves*

$$\mathbb{E}[\mathcal{R}_T] \leq C c_{\max}\left(n\beta_0 + \sqrt{nTQ}\right) \tag{2}$$

for a universal constant $C > 0$. For $K = |\mathcal{H}| < \infty$, PBERM also satisfies the horizon-independent bound $\mathbb{E}[\mathcal{R}_T] \leq c_{\max}(21 + 3n\beta_0)(K-1)$.

We can extend PBERM to the agnostic setting by maximizing the positives over all ERMs, which is the same as V_t in the realizable setting. Notice that PBERM is still a specific ERM and thus can fail in the agnostic setting as shown in the construction of Proposition 1 item (iii). In agnostic setting, noisy labels can also trap PBERM. Indeed, assume $n = 1$, $\mathcal{X} = \{x\}$, $\mathcal{H} = \{0, 1\}$, $h_0 = 1$, $c_{\text{fp}} = c_{\text{fn}} = 1$, and $\Pr[Y = 1] = 3/4$. Then PBERM will predict 1 in the beginning, but with probability $1/4$, it will observe label 0 and thus will always predict 0 afterwards as it will be the unique minimizer since no more labels will be revealed. This will lead to $\Omega(T)$ regret. The problem is that PBERM gives up on exploration too early.

Capped Exploration ERM (CEERM) We propose another algorithm that solves this issue by designing a new loss function that incorporates exploration. Intuitively, the new loss function will assign an individual cost for false positives at each instance at the beginning and increase the cost when the point has been well-explored.

Note that our loss function $\ell(h, x, y)$ can be written as

$$\ell(h, x, y) = c_{\text{fn}} \cdot y(1 - h(x)) + c_{\text{fp}} \cdot (1 - y)h(x) = c_{\text{fn}} y - c_{\text{fn}} h(x) + (c_{\text{fn}} + c_{\text{fp}})(1 - y)h(x)\,.$$

Since the first term does not depend on h, minimizing the original loss ℓ is equivalent to minimizing $\widetilde{\ell}(h, x, y) = -c_{\text{fn}} h(x) + (c_{\text{fn}} + c_{\text{fp}})(1 - y)h(x)$. This new loss function assigns loss value 0 when h predicts negatively, and a loss of $-c_{\text{fn}}$ for true positives and c_{fp} for false positives.

Let $C_t(x) := \sum_{\tau=2}^{t-1} h_\tau(x)$ be the count of implemented hypotheses before round t that predict positively on x. Thus, $C_t(x)$ also denotes the number of rounds we would have explored at point x. Given a fixed cap $r > 0$, we reduce the loss of a false positive prediction when $C_t(x) < r$. Our algorithm on round $t > 1$ implements a hypothesis that minimizes an unbiased estimate of the expectation of the following loss function:

$$\ell_r(h, x, y) := -c_{\text{fn}} h(x) + (c_{\text{fn}} + c_{\text{fp}})(1 - y)h(x) \min\{\frac{C_t(x)}{r}, 1\}\,.$$

When $C_t(x) < r$, the factor $\frac{C_t(x)}{r}$ reduces the contribution of false positive correction term to the total loss. Thus, a false positive prediction on an instance x might still have negative loss if x has not been explored enough. This ensures that early false positives do not discourage further exploration too much. By setting r appropriately, minimizing this new designed loss function will not incur much regret. See Algorithm 2 in the appendix for details of the algorithm.

Theorem 3 (CEERM guarantees)**.** *Suppose $Y = h^*(X)$ almost surely for some $h^* \in \mathcal{H}$, i.e. the distribution is realizable. Then CEERM with a suitable choice of r achieves*

$$\mathbb{E}[\mathcal{R}_T] \leq C(c_{\text{fn}} + c_{\text{fp}}) \left(n\beta_0 + \sqrt{\frac{c_{\text{fn}} + c_{\text{fp}}}{c_{\text{fp}}} nTQ} \right) \tag{3}$$

In the agnostic setting, with a different choice of r it achieves the following regret bound:

$$\mathbb{E}[\mathcal{R}_T] \leq C(c_{\text{fn}} + c_{\text{fp}}) \left(n\beta_0 + (nT)^{2/3} Q^{1/3} \right). \tag{4}$$

5 Experiments

In this section we describe our experiments, which demonstrate the possible issues caused by getting apple-tasting feedback. We run our experiments on COMPAS Larson et al. (2016) and MNIST LeCun et al. (1998), which correspond to an agnostic and an almost realizable setting, respectively. Due to space constraints, here we only state important details and results for the COMPAS dataset and defer MNIST dataset, implementation details, additional settings, and additional results to Appendix C.

Algorithms: Here we assume $c_{\text{fp}} = 1$ and consider different values for c_{fn}. Consider a hypothesis class parameterized by θ. Let $p_\theta(x) \in [0, 1]$ denote a model's output on x with prediction rule $\mathbb{1}[p_\theta(x) \geq 1/2]$. We implement ERM over a labeled sample S by optimizing the weighted cross entropy loss $\mathcal{E}(\theta; S) = \frac{1}{|S|} \sum_{(x,y)\in S} \left[-c_{\text{fn}} y \log p_\theta(x) - (1 - y) \log\big(1 - p_\theta(x)\big) \right]$. Simple ERM on round t optimizes $\mathcal{E}(\theta; S^{\text{obs}}_{1:t-1})$. Exploration-only ERM (EO-ERM) implements the algorithm of Section 4.1, i.e. explores k instances E_t on each round and on round t optimizes $\mathcal{E}(\theta; E_{1:t-1})$ ignoring other observed labels. Exploration-augmented ERM does direct exploration, but also uses other observed labels.

Table 1: Final cumulative loss $\mathcal{L}_T$ (mean $\pm$ SEM over all 30 seeds) for COMPAS, $n = 200$, $T = 30$, and i.i.d. sampling. PRERM uses $a_\lambda = 0.01$. Exploration methods use the constant budget with the indicated k. The lowest mean loss among ATF methods in each column is shown in bold.

Feedback	Method	$c_{\text{fn}} = 0.1$	$c_{\text{fn}} = 0.5$	$c_{\text{fn}} = 1$	$c_{\text{fn}} = 2$
-	Hindsight ERM	326.39 ± 0.86	1376.12 ± 5.01	1932.73 ± 6.44	2424.03 ± 7.40
FF	ERM	371.81 ± 6.79	1420.90 ± 6.16	2004.47 ± 7.07	2514.27 ± 14.04
FF	PRERM	371.87 ± 6.79	1421.25 ± 6.17	2004.47 ± 7.11	2514.43 ± 14.01
ATF	ERM	$\mathbf{370.79 \pm 7.07}$	1517.80 ± 17.79	2509.13 ± 113.77	4113.77 ± 364.54
ATF	PRERM	407.11 ± 2.46	$\mathbf{1452.72 \pm 7.67}$	$\mathbf{2025.70 \pm 8.65}$	$\mathbf{2550.40 \pm 17.44}$
ATF	EA-ERM, $k = 3$	428.36 ± 3.47	1461.00 ± 7.88	2036.57 ± 8.90	2569.53 ± 19.31
ATF	EA-ERM, $k = 21$	634.85 ± 5.03	1573.17 ± 8.28	2096.10 ± 8.07	2565.47 ± 15.61
ATF	EA-ERM, $k = 42$	877.72 ± 6.32	1711.47 ± 7.27	2168.43 ± 5.35	2569.90 ± 12.70
ATF	EO-ERM, $k = 3$	929.89 ± 32.48	1760.27 ± 18.13	2397.13 ± 19.20	3182.40 ± 46.93
ATF	EO-ERM, $k = 21$	679.72 ± 7.54	1640.85 ± 8.97	2202.83 ± 11.79	2639.17 ± 15.23
ATF	EO-ERM, $k = 42$	891.25 ± 7.65	1737.33 ± 7.17	2219.13 ± 10.09	2601.03 ± 12.47

Inspired by our algorithms in Section 4.2, we also consider a *positively regularized ERM* (PRERM). Given labeled and unlabeled datasets S and U, PRERM uses the following objective:

$$\mathcal{J}(\theta; S, U) = \mathcal{E}(\theta; S) + \frac{\lambda}{|U|} \sum_{x \in U} -\log p_\theta(x), \qquad \lambda_t = \frac{a_\lambda}{\sqrt{\max(1, |S^{\text{obs}}_{1:t-1}|)}}, \tag{5}$$

where $\lambda > 0$ is a hyperparameter. Let U_t be the multiset of all unlabeled data available to the learner on round t. PRERM then optimizes $\mathcal{J}(\theta; S^{\text{obs}}_{1:t-1}, U_t)$ with regularization factor λ_t defined in Equation (5). Note that larger values of λ encourage more exploration. Our λ_t schedule decreases as the learner receives more labels, which balances between exploitation and exploration.

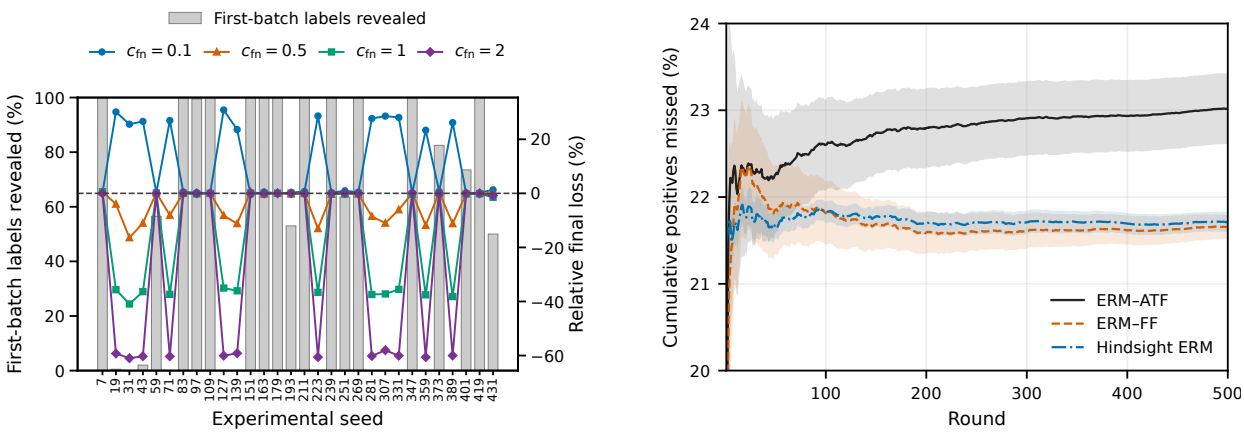
COMPAS: We use the filtered COMPAS dataset, which consists of $6,172$ instances. The positive class corresponds to no recorded recidivism within two years. We use eight features: age, sex, race, juvenile felony count, juvenile misdemeanor count, other juvenile offense count, prior conviction count, and charge degree. We note that the dataset is relatively noisy and out of 3,134 unique feature vectors 612 of them appear with both positive and negative labels.

We consider uniform i.i.d. draws with replacement and use $n = 200$ and $T = 30$. In addition to apple-tasting feedback (ATF), we also consider full feedback (FF), revealing label of all instances, which serves as a baseline. All methods share the same initial weights and stream for each seed. We consider different values $c_{\text{fn}} \in \{0.1, 0.5, 1, 2\}$ and evaluate cumulative loss $\mathcal{L}_T = \sum_{\tau=1}^{T} \big(\text{FP}_\tau + c_{\text{fn}}\text{FN}_\tau\big)$. Here FP_τ and FN_τ are the numbers of false positives and false negatives in round τ. We report mean $\pm$ SEM over 30 seeds. Table 1 reports the results for this setting.

ATF vs FF: Receiving ATF instead of FF increases ERM's average loss when false negatives are more costly. PRERM reduces this gap for $c_{\text{fn}} \in \{0.5, 1, 2\}$, but increases the loss when $c_{\text{fn}} = 0.1$. In that case, the false positives encouraged by its penalty are relatively expensive. Under FF, ERM and PRERM have very similar losses as λ_t vanishes quickly. Explicit exploration also reduces the ATF loss in several cases, although increasing its budget can make the additional false positives outweigh the benefit of receiving more labels. Overall, our results suggest when the cost of a false negative is comparable to or more than cost of false positive, PRERM is able to recover the results of FF setting under ATF.

Self-confirming traps: We observe that the variance of ERM performance under ATF is very high relative to other algorithms. Upon inspecting the results without aggregation, we observe that the performance highly depends on the number of labels the initial hypothesis reveals. Indeed, Figure 1a shows that when this number is not close to zero, the two methods perform similarly. This pattern is consistent with the first failure mode of ERM described in Section 3.1. In 12 of the 30 seeds, the initial hypothesis reveals at most 2% of the first-batch labels. For these seeds, PRERM has lower final loss when $c_{\text{fn}} \in \{0.5, 1, 2\}$, but higher loss when $c_{\text{fn}} = 0.1$. In the remaining 18 seeds, at least 50% of the first-batch labels are revealed, and the final losses of the two methods differ by less than 1.46% relative to ERM for all four costs.

Self-reinforcing mistakes: To account for the bad initialization issue, here we remove the seeds with bad initialization and compare ERM under ATF and FF on the remaining seeds. Keeping $n = 200$, we increase the horizon to $T = 500$. We observe that the false negative rate of ERM under FF rather quickly converges to the baseline. However, under



(a) Self-confirming traps: $n = 200$, $T = 30$.

(b) Self-reinforcing mistakes: $n = 200$, $T = 500$, $c_{\text{fn}} = 1$.

Figure 1: Two failure modes of ERM under ATF on COMPAS with i.i.d. sampling. (a) Each x-axis position corresponds to one of the 30 seeds. Gray bars show the percentage of first-batch labels revealed. The lines show $100(\mathcal{L}_T^{\text{PRERM}} - \mathcal{L}_T^{\text{ERM}})/\mathcal{L}_T^{\text{ERM}}$ under ATF for the four false-negative costs. A negative value indicates that PRERM performs better. PRERM uses $a_\lambda = 0.01$. (b) Cumulative percentage of positive instances missed, $100 \sum_{\tau=1}^{t} \text{FN}_\tau / \sum_{\tau=1}^{t} P_\tau$, where P_τ is the number of actual positive instances in batch τ. Curves show mean $\pm$ SEM over the 18 seeds whose initial hypothesis predicts positively on at least 25% of the first batch instances.

ATF it maintains a considerable gap even at large T as shown in Figure 1b. One potential explanation for this is the self-reinforcing mistakes failure of ERM on noisy data as described in Section 3.2. An observation that supports this argument is that the mentioned gap is notably smaller for the MNIST data, which is relatively noiseless. Specifically, for $c_{\text{fn}} = 1$, we retain 18 COMPAS seeds whose initial hypothesis predicts positively on at least 25% of the first batch instances. Over rounds 401–500, the mean per-round fraction of positives missed is $(23.35 \pm 0.44)\%$ under ATF, $(21.81 \pm 0.13)\%$ under FF, and $(21.79 \pm 0.13)\%$ for the hindsight learner. The paired ATF–FF gap is 1.54 ± 0.38 percentage points, compared with 0.14 ± 0.02 for the 15 retained MNIST seeds under the same setting and filtering rule. These are means $\pm$ SEM across retained seeds, after averaging the per-round rates within each seed over the final 100 rounds.

6 Conclusion

Motivated by applications such as college admissions, we studied a batched variant of apple tasting in which learners observe labels only for instances they predict positively. We showed that the resulting endogenous selection bias can cause ERM to become trapped at suboptimal classifiers. Under noisy labels, early unlucky observations can also produce self-reinforcing mistakes by preventing the learner from receiving corrective feedback.

To address these failures, we developed two approaches to exploration. Random exploration provides unbiased labeled data and achieves regret sublinear in both the batch size and the number of rounds. When decisions must follow a classifier from a predefined hypothesis class, our proper learners, PBERM and CEERM, achieve comparable realizable guarantees with an additional cost linear in the batch size, which can be unavoidable under fixed initialization. CEERM also provides guarantees in the agnostic setting.

Inspired by these algorithms, we introduced PRERM, a practical method that encourages exploration through positivity regularization. Our experiments illustrated the failures of ERM under selective feedback and showed that PRERM can mitigate their impact, although its benefits depend on the relative costs of false positives and false negatives. These results highlight the importance of deliberate exploration when deployed predictions determine future training data.

7 AI USE STATEMENT

In this work, we used generative AI tools for designing an algorithm (CEERM) and proving mathematical claims about it, formulate mathematical claims, for assisting in writing of proofs and their simplification, for cleaning data (removing rows with missing entries from COMPAS), and for implementing our methods. In addition to writing the code for implementation, we used AI for proposing the neural network architecture for MNIST and optimization hyperparameters and budgets. We have not used generative AI tools for developing our theoretical models or conceptual frameworks, for proposing or refining hypotheses, for providing feedback on research methodology or experiments, and for interpreting results. Other tasks including synthetic data generation, translation, qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work including any text, claim, or artifact generated by the aid of generative AI.

8 ETHICS STATEMENT

This work is motivated by selection bias in societal decision-making. Our contributions are primarily theoretical, with offline experiments using publicly available datasets and no real-world interventions. The data attributes and models we use in our experiments are only for demonstration and are not a guidance for what should be used in practice. We do not anticipate direct adverse societal impacts from these analyses and experiments.

REFERENCES

Anthony, M. and Shawe-Taylor, J. (1993). A result of vapnik with applications. *Discrete Applied Mathematics*, 47(3):207–217.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. (2019). Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, volume 32.

Bechavod, Y. and Roth, A. (2023). Individually fair learning with one-sided feedback. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1954–1977.

Chase, Z. and Mehalel, I. (2025). Deterministic apple tasting. In Haghtalab, N. and Moitra, A., editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 888–923. PMLR.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118.

Harris, K., Podimata, C., and Wu, S. (2023). Strategic apple tasting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik–chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232.

Helmbold, D. P., Littlestone, N., and Long, P. M. (2000). Apple tasting. *Information and Computation*, 161(2):85–139.

Jiang, H., Jiang, Q., and Pacchiano, A. (2021). Learning the truth from only one side of the story. In *The 24th International Conference on Artificial Intelligence and Statistics*.

Kilbertus, N., Gomez-Rodriguez, M., Schölkopf, B., Muandet, K., and Valera, I. (2020). Fair decisions despite imperfect predictions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Pacchiano, A., Singh, S., Chou, E., Berg, A. C., and Foerster, J. (2021). Neural pseudo-label optimism for the bank loan problem. In *Advances in Neural Information Processing Systems*, volume 34.

Raman, V., Subedi, U., Raman, A., and Tewari, A. (2024). Apple tasting: Combinatorial dimensions and minimax rates. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4358–4380. PMLR.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Ville, J. (1939). *Étude critique de la notion de collectif*. Gauthier-Villars, Paris.

Wei, D. (2021). Decision-making under selective labels: Optimal finite-domain policies and beyond. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11035–11046.

A PROOFS

A.1 PROOFS FOR SECTION 3

Proof of Proposition 1. Part (i): trapping at a suboptimal feedback-stable hypothesis. Let $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\mathcal{H} = \{g_1, g_2, g_3\}$, where

$$g_i(x_j) = \mathbb{1}\{i = j\}.$$

Let the target be $h^\star = g_1$, initialize at $h_0 = g_2$, and take $\mathcal{D}_\mathcal{X}$ to be uniform over $\{x_1, x_3\}$, with labels determined by $h^\star$. The population risks are

$$\ell_\mathcal{D}(g_1) = 0, \qquad \ell_\mathcal{D}(g_2) = \frac{c_{\text{fn}}}{2}, \qquad \ell_\mathcal{D}(g_3) = \frac{c_{\text{fp}} + c_{\text{fn}}}{2}.$$

Thus, $h^\star$ is the unique population-risk minimizer. Its zero population loss also makes it feedback-stable.

The initializer g_2 rejects both support points, so

$$\ell_{\mathcal{D},g_2}(g) = 0 \qquad \text{for every } g \in \mathcal{H}.$$

Consequently, g_2 is another feedback-stable hypothesis.

Fix any conservative tie-breaking rule. While g_2 is implemented, no labels are revealed, so it remains an empirical risk minimizer. Inductively, conservativeness forces ERM to retain g_2 forever, regardless of how the rule otherwise resolves ties.

The learner incurs a false negative precisely at x_1, whereas the realizable comparator has zero loss. Excluding the initialization round gives

$$\mathbb{E}\left[\mathcal{R}_T\right] = \frac{c_{\text{fn}}}{2}\, n(T-1) = \Theta(nT).$$

Part (ii): regret before convergence in the realizable setting. Use the singleton class from part (i), but take the marginal distribution

$$\Pr[x = x_1] = \Pr[x = x_2] = \frac{1}{4}, \qquad \Pr[x = x_3] = \frac{1}{2},$$

and initialize at $h_0 = g_1$. Fix any conservative tie-breaking rule, which may depend on the observed data and may be randomized. Consider two realizable distributions $\mathcal{D}_2$ and $\mathcal{D}_3$ with this common marginal and respective targets g_2 and g_3.

In the first round, $h_1 = h_0 = g_1$ reveals labels only at x_1, whose label is zero under both targets. Thus, the first-round observations, including the unlabeled examples, have the same distribution under $\mathcal{D}_2$ and $\mathcal{D}_3$. Consequently, the joint distribution of X_2 and the second-round classifier h_2 is also the same under both distributions.

Under this common law, define

$$a := \frac{2}{n}\,\mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}\{x_{2,j} = x_3\}\mathbb{1}\{h_2(x_3) = 1\}\right] \in [0, 1].$$

The expected number of second-round occurrences of x_3 is $n/2$. Thus, the expected numbers of accepted and rejected occurrences of x_3 are $na/2$ and $n(1-a)/2$, respectively.

Under $\mathcal{D}_2$, accepting x_3 incurs false-positive loss. Under $\mathcal{D}_3$, rejecting x_3 incurs false-negative loss. Therefore,

$$\mathbb{E}_{\mathcal{D}_2}[\mathcal{R}_T] \geq \frac{c_{\text{fp}} na}{2}, \qquad \mathbb{E}_{\mathcal{D}_3}[\mathcal{R}_T] \geq \frac{c_{\text{fn}} n(1-a)}{2}.$$

Here the realizable comparator has zero loss, and $\mathcal{R}_T$ includes the second-round loss. Hence

$$\max\left\{\mathbb{E}_{\mathcal{D}_2}[\mathcal{R}_T], \mathbb{E}_{\mathcal{D}_3}[\mathcal{R}_T]\right\} \geq \frac{n}{2}\max\{c_{\text{fp}} a, c_{\text{fn}}(1-a)\} \geq \frac{c_{\text{fp}} c_{\text{fn}}}{2(c_{\text{fp}} + c_{\text{fn}})}\, n.$$

Choose $\mathcal{D} \in \{\mathcal{D}_2, \mathcal{D}_3\}$ attaining the larger expectation, and let $h^\star$ be its target. This fixes the target before sampling and proves the $\Omega(n)$ lower bound.

Under either target, every incorrect singleton g_i reveals only its own negative point, which has positive probability. Thus,

$$\ell_{\mathcal{D},g_i}(g_i) = c_{\text{fp}} \Pr[x = x_i] > \ell_{\mathcal{D},g_i}(h^\star) = 0.$$

Since the target has zero population loss, it is the unique feedback-stable hypothesis.

Part (iii): cycling in the agnostic setting. Let $\mathcal{X} = \{x_0, x_1, x_2\}$ and $\mathcal{H} = \{h^\star, g_1, g_2\}$, with predictions and deterministic labels given by

	$h^\star$	g_1	g_2	y
x_0	1	0	0	1
x_1	1	1	0	0
x_2	1	0	1	0

Set

$$\Pr[x = x_1] = \Pr[x = x_2] = \frac{c_{\mathrm{fn}}}{2(c_{\mathrm{fp}} + 2c_{\mathrm{fn}})}, \qquad \Pr[x = x_0] = \frac{c_{\mathrm{fp}} + c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}}.$$

The population risks satisfy

$$\ell_{\mathcal{D}}(h^\star) = \frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}}$$

and

$$\ell_{\mathcal{D}}(g_1) = \ell_{\mathcal{D}}(g_2) = c_{\mathrm{fn}} \frac{c_{\mathrm{fp}} + c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} + c_{\mathrm{fp}} \frac{c_{\mathrm{fn}}}{2(c_{\mathrm{fp}} + 2c_{\mathrm{fn}})} = \frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} + \frac{c_{\mathrm{fn}}}{2}.$$

Thus, $h^\star$ is the unique population-risk minimizer.

Since $h^\star$ predicts positive everywhere, it reveals the full distribution. Hence $\ell_{\mathcal{D},h^\star}(g) = \ell_{\mathcal{D}}(g)$ for every $g \in \mathcal{H}$, so $h^\star$ is feedback-stable. In contrast,

$$\ell_{\mathcal{D},g_1}(g_2) = 0 < \ell_{\mathcal{D},g_1}(g_1) = \ell_{\mathcal{D},g_1}(h^\star) = \frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{2(c_{\mathrm{fp}} + 2c_{\mathrm{fn}})},$$

and

$$\ell_{\mathcal{D},g_2}(g_1) = 0 < \ell_{\mathcal{D},g_2}(g_2) = \ell_{\mathcal{D},g_2}(h^\star) = \frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{2(c_{\mathrm{fp}} + 2c_{\mathrm{fn}})}.$$

Thus, $h^\star$ is the unique feedback-stable hypothesis.

Initialize at $h_0 = g_1$ and fix any conservative tie-breaking rule. Let $N_i(t)$ count the revealed examples at x_i through round t, for $i = 1, 2$. As long as only g_1 and g_2 have been implemented, all revealed labels are zero, and the unnormalized cumulative empirical losses are

$$\widehat{L}_t(h^\star) = c_{\mathrm{fp}}\big(N_1(t) + N_2(t)\big), \qquad \widehat{L}_t(g_i) = c_{\mathrm{fp}} N_i(t) \quad (i = 1, 2).$$

If both counts are zero, conservativeness retains the incumbent. Otherwise,

$$\min\{N_1(t), N_2(t)\} < N_1(t) + N_2(t),$$

so $h^\star$ is not an empirical risk minimizer. Inductively, ERM never selects $h^\star$. It selects the hypothesis with the smaller count, retaining the incumbent when the counts are equal.

With probability one, both x_1 and x_2 occur infinitely often in the underlying sequence of batches. On this event, finitely many switches would imply that some g_i, $i \in \{1, 2\}$, is eventually implemented forever. Then $N_i(t) \to \infty$, while the other count remains fixed, eventually forcing a switch. This contradiction proves that ERM almost surely switches infinitely often between g_1 and g_2.

Finally, both deployed hypotheses have population risk $\frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} + c_{\mathrm{fn}}/2$. In this construction, the deployed hypothesis is determined by the preceding counts and the incumbent, so it is independent of the fresh batch. Therefore,

$$\mathbb{E}\left[\mathcal{L}_T - L_1(h_0)\right] = \left(\frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} + \frac{c_{\mathrm{fn}}}{2}\right) n(T-1).$$

The hindsight comparator satisfies

$$0 \le \mathbb{E}\left[\min_{h \in \mathcal{H}} \sum_{t=2}^{T} L_t(h)\right] \le \mathbb{E}\left[\sum_{t=2}^{T} L_t(h^\star)\right] = \frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} n(T-1).$$

Consequently,

$$\frac{c_{\mathrm{fn}}}{2} n(T-1) \le \mathbb{E}\left[\mathcal{R}_T\right] \le \left(\frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} + \frac{c_{\mathrm{fn}}}{2}\right) n(T-1).$$

Since $\frac{c_{\mathrm{fp}} c_{\mathrm{fn}}}{c_{\mathrm{fp}} + 2c_{\mathrm{fn}}} \le c_{\mathrm{fn}}$, the upper bound is at most $3c_{\mathrm{fn}} n(T-1)/2$. Thus, $\mathbb{E}\left[\mathcal{R}_T\right] = \Theta(nT)$. □

Proof of Observation 1. Fix $x > N/2$ with $\mathcal{D}(x) > 0$. Consider all labels generated at x, including those not revealed to the learner. Let Z_s be the number of positive labels minus the number of negative labels among the first s such labels, with $Z_0 = 0$. Since x has positive mass, infinitely many such labels are generated almost surely. Each label increases Z_s by 1 with probability $1 - \lambda(x)$ and decreases it by 1 with probability $\lambda(x)$. While ERM predicts positive at x, it observes all labels at x, but updates its prediction only at batch boundaries.

If $\lambda(x) = 0$, all labels at x are positive, so the claim holds. Otherwise, let $r = \lambda(x)/(1 - \lambda(x)) \in (0, 1)$. Since

$$\mathbb{E}\left[r^{Z_{s+1}} \mid Z_s\right] = r^{Z_s}\left((1 - \lambda(x))r + \lambda(x)r^{-1}\right) = r^{Z_s},$$

the process $(r^{Z_s})_{s\geq 0}$ is a martingale.

For an integer $m \geq 1$, let $\tau_m = \inf\{s \geq 0 : Z_s \in \{-1, m\}\}$. This stopping time is almost surely finite, since a run of m upward steps forces an exit and such a run occurs almost surely eventually. The stopped martingale is bounded between r^m and r^{-1}, so optional stopping gives

$$1 = \mathbb{E}\left[r^{Z_{\tau_m}}\right] = r^{-1}\Pr(Z_{\tau_m} = -1) + r^m \Pr(Z_{\tau_m} = m).$$

Thus, $\Pr(Z_{\tau_m} = -1) = (1 - r^m)/(r^{-1} - r^m)$. As $m \to \infty$, the events of hitting -1 before m increase to the event of ever hitting -1. Therefore,

$$\Pr(\exists s \geq 0 : Z_s = -1) = \lim_{m\to\infty} \Pr(Z_{\tau_m} = -1) = r = \frac{\lambda(x)}{1 - \lambda(x)}.$$

To relate this event to ERM's updates, imagine revealing the i.i.d. examples sequentially within each batch. If the walk first reaches -1, its count at every preceding batch boundary is nonnegative. Pointwise conservative tie-breaking therefore ensures that ERM still predicts positive at x during the batch containing this first visit. Conditional on the revealed prefix through this visit, the remaining examples in the batch are independent fresh draws. There are at most $n - 1$ such examples, so with conditional probability at least $(1 - 1/N)^{n-1}$, none equals x. On this event, the count remains -1 at the end of the batch, and ERM switches its prediction at x to negative. Consequently,

$$\begin{aligned}\Pr(\text{ERM eventually switches to negative at } x) &\geq \Pr(\exists s \geq 0 : Z_s = -1)\left(1 - \frac{1}{N}\right)^{n-1} \\ &= \frac{\lambda(x)}{1 - \lambda(x)}\left(1 - \frac{1}{N}\right)^{n-1}.\end{aligned}$$

Once ERM predicts negative at x, no further labels at x are observed. The empirical label counts therefore remain fixed, and the negative prediction persists forever. □

A.2 Proofs from Section 4.1

We implement the prescribed exploration parameter $k > 0$ as follows. If $k \geq 1$, each round explores $\min\{n, \lfloor k \rfloor\}$ uniformly selected distinct batch indices. If $0 < k < 1$, each round independently explores one uniformly selected batch index with probability k. All exploration randomness is independent of the examples. Thus, the expected number of explored positions per round is at most k; for $k < 1$, this is an expected budget rather than a per-round cap. Exploration is allowed during initialization, when $h_1 = h_0$. On subsequent rounds, h_t is an ERM on the previously explored examples, with multiplicities retained; if this sample is empty, we use h_0. All hypothesis choices depend only on the history. Regret is computed from the actual actions, including exploration.

For $0 < k \leq n$, let $m_t = |S_{t-1}^{\text{exp}}|$ be the number of exploration examples available before round t. Conditional on m_t, these examples are i.i.d. from $\mathcal{D}$. If $k \geq 1$, then $m_t = \lfloor k \rfloor (t - 1)$ and $\lfloor k \rfloor \geq k/2$. If $k < 1$, then $m_t \sim \mathrm{Bin}(t - 1, k)$ and

$$\mathbb{E}\left[\frac{1}{m_t + 1}\right] = \frac{1 - (1 - k)^t}{kt} \leq \frac{1}{kt}.$$

Consequently, in both cases, for $t \geq 2$,

$$\mathbb{E}\left[\frac{1}{m_t + 1}\right] \leq \frac{2}{k(t - 1)}, \qquad \mathbb{E}\left[\frac{1}{\sqrt{m_t + 1}}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{m_t + 1}\right]} \leq \sqrt{\frac{2}{k(t - 1)}},$$

where the second estimate uses Jensen's inequality. These estimates will be used in both upper-bound proofs. Throughout, $C > 0$ denotes a universal constant that may change between occurrences.

Proof of Theorem 1, part 1. First suppose that the prescribed budget satisfies $k \le n$. Exploration increases the loss relative to h_t by at most c_{fp} per explored position. Since $h_1 = h_0$, its base loss cancels with the initialization baseline, while any exploratory overrides remain charged. The realizable comparator has zero loss, so

$$\mathbb{E}\left[\mathcal{R}_T\right] \le c_{\mathrm{fp}} kT + \sum_{t=2}^{T} \mathbb{E}\left[L_t(h_t)\right] = c_{\mathrm{fp}} kT + n \sum_{t=2}^{T} \mathbb{E}\left[\ell_{\mathcal{D}}(h_t)\right],$$

where the equality uses the independence of h_t and the fresh batch.

Conditional on m_t, the expected VC bound for consistent ERM (Shalev-Shwartz and Ben-David, 2014) gives

$$\mathbb{E}\left[\ell_{\mathcal{D}}(h_t) \mid m_t\right] \le C c_{\max} \frac{d \log(enT)}{m_t + 1}.$$

This also holds when $m_t = 0$, by $\ell_{\mathcal{D}}(h_t) \le c_{\max}$ and $d \ge 1$. Taking expectations and using the sample-size estimate above,

$$\mathbb{E}\left[\ell_{\mathcal{D}}(h_t)\right] \le C c_{\max} \frac{d \log(enT)}{k(t-1)}.$$

Summing over t and using $T \ge 2$ therefore yields

$$\mathbb{E}\left[\mathcal{R}_T\right] \le c_{\mathrm{fp}} kT + C c_{\max} \frac{nd \log(nT) \log T}{k}.$$

Substituting the prescribed value

$$k = \sqrt{\frac{c_{\max} nd \log(nT) \log T}{c_{\mathrm{fp}} T}}$$

gives

$$\mathbb{E}\left[\mathcal{R}_T\right] \le C \sqrt{c_{\mathrm{fp}} c_{\max} nTd \log(nT) \log T}.$$

If the prescribed value satisfies $k > n$, every position is explored. The actual loss is then at most $c_{\mathrm{fp}} nT$, and both the comparator loss and the subtracted initialization loss are nonnegative. Hence

$$\mathcal{R}_T \le c_{\mathrm{fp}} nT \le c_{\mathrm{fp}} kT = \sqrt{c_{\mathrm{fp}} c_{\max} nTd \log(nT) \log T},$$

which proves the same bound in the remaining case. □

Proof of Theorem 1, part 2. Write $N = nT$ and set

$$M = \left\lceil 16 \sqrt{\frac{c_{\mathrm{fn}} N}{c_{\mathrm{fp}}}} \right\rceil, \qquad \mathcal{X} = \{0, 1, \ldots, M\}, \qquad \mathcal{H} = \{h_j : j = 0, \ldots, M\},$$

where $h_j(x) = \mathbb{1}\{x = j\}$. Let $\mathcal{D}_{\mathcal{X}}$ be uniform on $\{1, \ldots, M\}$, let h_0 be the initial hypothesis, and choose the target h_J with J uniform on $\{1, \ldots, M\}$, independently of the feature stream and the learner's randomness. Every such target defines a realizable distribution.

Fix the entire feature stream and the learner's randomness. Run the learner with every revealed label equal to zero, and let A be the set of distinct instances it predicts positive during this run. Write $q = |A|$ and order these instances by their first positive prediction, breaking ties within a batch arbitrarily. If $J \notin A$, the run is unchanged under target h_J, so every occurrence of J is a false negative. If J is the sth instance in this order, the learner incurs at least $s - 1$ false positives before observing a positive label. This remains true when several of these predictions occur in the same batch.

Let N_j be the number of occurrences of j in the full stream, and write $\mathcal{R}_T(j)$ for the regret under target h_j with the fixed features and learner randomness. The target's loss is zero. Moreover, averaging the subtracted initialization loss over J gives exactly $c_{\mathrm{fn}} n/M$. Consequently,

$$\frac{1}{M} \sum_{j=1}^{M} \mathcal{R}_T(j) \ge \frac{c_{\mathrm{fp}} q(q-1)}{2M} + \frac{c_{\mathrm{fn}}}{M} \sum_{j \notin A} N_j - \frac{c_{\mathrm{fn}} n}{M}.$$

For fixed positive costs and sufficiently large N, $M \ge 16$ and a Chernoff bound gives

$$\Pr\left(\max_{j \in [M]} N_j > \frac{2N}{M} \right) \le M \exp\left(-\frac{N}{3M} \right) \le \frac{1}{4}.$$

On the complementary event, if $q \le M/8$, then $\sum_{j \notin A} N_j \ge 3N/4$. Since $n \le N/2$, the average regret is at least $c_{\mathrm{fn}} N/(4M)$. If $q > M/8$, then $q - 1 \ge q/2$ and

$$\frac{c_{\mathrm{fp}} q(q-1)}{2M} - \frac{c_{\mathrm{fn}} n}{M} \ge \frac{c_{\mathrm{fp}} M}{256} - \frac{c_{\mathrm{fn}} n}{M} \ge \frac{c_{\mathrm{fn}} N}{2M},$$

where the last inequality uses $M^2 \ge 256 c_{\mathrm{fn}} N / c_{\mathrm{fp}}$. On the exceptional event, the average regret is at least $-c_{\mathrm{fn}} n/M$. Taking expectations over features and learner randomness therefore yields

$$\frac{1}{M} \sum_{j=1}^{M} \mathbb{E}\left[\mathcal{R}_T(j)\right] \ge \frac{3}{4} \frac{c_{\mathrm{fn}} N}{4M} - \frac{1}{4} \frac{c_{\mathrm{fn}} n}{M} \ge \frac{c_{\mathrm{fn}} N}{16M} = \Omega\left(\sqrt{c_{\mathrm{fn}} c_{\mathrm{fp}} n T}\right).$$

At least one target attains this lower bound, and $|\mathcal{X}| = M + 1 = \Theta(\sqrt{c_{\mathrm{fn}} n T / c_{\mathrm{fp}}})$, as claimed. □

Proof of Theorem 1, part 3. First suppose that the prescribed budget satisfies $k \le n$, and set

$$L_* := \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h).$$

Conditional on m_t, ERM comparison and the expected VC uniform deviation bound (Shalev-Shwartz and Ben-David, 2014) give

$$\mathbb{E}\left[\ell_{\mathcal{D}}(h_t) - L_* \mid m_t\right] \le C c_{\max} \sqrt{\frac{d}{m_t + 1}}.$$

The empty-sample case follows from boundedness. Thus, by the sample-size estimate above,

$$\mathbb{E}\left[\ell_{\mathcal{D}}(h_t) - L_*\right] \le C c_{\max} \sqrt{\frac{d}{k(t-1)}}.$$

The regret comparator minimizes loss on the realized evaluation stream. Applying the expected uniform deviation bound to the $n(T-1)$ examples in rounds $2, \dots, T$ gives

$$n(T-1) L_* - \mathbb{E}\left[\min_{h \in \mathcal{H}} \sum_{t=2}^{T} L_t(h)\right] \le C c_{\max} \sqrt{n(T-1)d}.$$

This comparison does not require the population infimum to be attained. As in part 1, the base initialization loss cancels and the expected exploration cost is at most $c_{\mathrm{fp}} k T$. Consequently,

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}_T\right] &\le c_{\mathrm{fp}} k T + C n c_{\max} \sum_{t=2}^{T} \sqrt{\frac{d}{k(t-1)}} + C c_{\max} \sqrt{nTd} \\ &\le c_{\mathrm{fp}} k T + C c_{\max} n \sqrt{\frac{dT}{k}}, \end{aligned}$$

where $\sum_{s=1}^{T-1} s^{-1/2} \le 2\sqrt{T}$ and $k \le n$ absorbs the comparator term. Substituting the prescribed value

$$k = \left(\frac{c_{\max}^2 n^2 d}{c_{\mathrm{fp}}^2 T}\right)^{1/3}$$

gives

$$\mathbb{E}\left[\mathcal{R}_T\right] \le C c_{\mathrm{fp}}^{1/3} c_{\max}^{2/3} (nT)^{2/3} d^{1/3}.$$

If the prescribed value satisfies $k > n$, every position is explored, and the same direct argument as in part 1 gives

$$\mathcal{R}_T \le c_{\mathrm{fp}} n T \le c_{\mathrm{fp}} k T = c_{\mathrm{fp}}^{1/3} c_{\max}^{2/3} (nT)^{2/3} d^{1/3}.$$

This completes the proof. □

A.3 Proofs from Section 4.2

Here for simplicity we assume that the game starts on round $t = 0$ where all learners implement the initial policy h_0 on X_0 and then proceed to predict on rounds $t \in [T]$. Although in our model the learners are allowed to select h_t based on X_t in addition to the history, our proposed learners will only use the history and ignore X_t when selecting h_t. We make use of this fact in our analysis. Then we have the following:

$$\mathcal{R}_T = \sum_{t=1}^{T}\sum_{i=1}^{n} \ell(h_t, X_{t,i}, Y_{t,i}) - \min_{h\in\mathcal{H}} \sum_{t=1}^{T}\sum_{i=1}^{n} \ell(h, X_{t,i}, Y_{t,i}) = \sum_{t} L_t(h_t) - \min_{h\in\mathcal{H}} \sum_{t} L_t(h)$$

Since h_t is independent of X_t, $\mathbb{E}\left[L_t(h_t)\right] = n\mathbb{E}\left[\ell_{\mathcal{D}}(h_t)\right]$. Let $N = nT$, $d = \mathrm{VCdim}(\mathcal{H})$ and $\overline{d} = \min\{d, N\}$. Define,

$$Q = Q(\mathcal{H}) := \min\{\log(2|\mathcal{H}|), 1 + \overline{d}\log(\frac{eN}{\overline{d}})\}$$

Then by standard uniform convergence arguments (Shalev-Shwartz and Ben-David, 2014, Chapters 6 and 26), we also have $\min_{h\in\mathcal{H}} N\ell_{\mathcal{D}}(h) - \mathbb{E}\left[\min_{h\in\mathcal{H}} \sum_t L_t(h)\right] = \mathcal{O}(c_{\max}\sqrt{NQ})$ where $c_{\max} := \max\{c_{\mathrm{fp}}, c_{\mathrm{fn}}\}$. Therefore, we focus on the following objective:

$$\mathcal{R}_T^{\mathrm{off}} := n\sum_{t=1}^{T} \mathbb{E}\left[\ell_{\mathcal{D}}(h_t) - \ell_{\mathcal{D}}(h^*)\right],$$

where $h^* \in \arg\min_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h)$. The preceding identities imply

$$0 \le \mathbb{E}\left[\mathcal{R}_T\right] - \mathcal{R}_T^{\mathrm{off}} \le Cc_{\max}\sqrt{NQ}.$$

Thus, an upper bound on $\mathcal{R}_T^{\mathrm{off}}$ yields an upper bound on $\mathbb{E}\left[\mathcal{R}_T\right]$ with this additional term. Under realizability, where $Y = h^*(X)$ almost surely, both comparator losses are zero almost surely, and hence $\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{R}_T^{\mathrm{off}}$ exactly.

We will also need the following definitions:

$$B = \mathrm{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \text{there exist } h, g \in \mathcal{H} \text{ with } h(x) \neq g(x)\}, \qquad \beta_0 = \mathcal{D}_X(B)$$

We use $\mathbb{1}_B(x) := \mathbb{1}\{x \in B\}$ to denote the indicator of the disagreement region.

We next state our key lemma. All datasets are treated as a multiset of instances and the set operations are defined accordingly (i.e., they keep the multiplicity of each instance). Fix any sequence $h_1, \ldots, h_T$ selected before its fresh batches and an integer $1 \le r \le T$. For $1 \le t \le T + 1$, let

$$\mathcal{U}_t = \bigcup_{j=0}^{t-1} X_j, \qquad C_t(x) = \sum_{j=1}^{t-1} h_j(x). \tag{6}$$

Thus $|\mathcal{U}_t| = nt$, $\mathcal{U}_1 = X_0$, and $C_1(x) = 0$. The cap r here is used for analysis and the results hold for any learning algorithm. For any fixed $h \in \mathcal{H}$, define

$$Z_{j,i}(h) = \mathbb{1}_B(X_{j,i})h(X_{j,i})h_j(X_{j,i})\mathbb{1}\{C_j(X_{j,i}) < r\}(1 - Y_{j,i}),$$
$$K_t(h) = \sum_{j=1}^{t-1}\sum_{i=1}^{n} Z_{j,i}(h).$$

Let $\mathcal{F}_j$ denote the history after selecting h_j and before drawing its batch, and set

$$\Lambda_t(h) = \sum_{j=1}^{t-1}\sum_{i=1}^{n} \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{F}_j\right], \qquad M_t(h) = K_t(h) - \Lambda_t(h).$$

The mask B restricts the counts to the region on which candidates can disagree.

For any x and any j, if $h_j(x) = 1$, then $C_{j+1}(x) = C_j(x) + 1$ and hence

$$\sum_{j=1}^{t-1} h_j(x)\mathbb{1}\{C_j(x) < r\} = \min\{C_t(x), r\}, \tag{7}$$

and therefore

$$\Lambda_t(h) = n\mathbb{E}\left[\mathbb{1}_B(X)h(X)(1-Y)\min\{C_t(X), r\}\right]. \tag{8}$$

Here and throughout, the expectations are over a fresh draw of $(X, Y) \sim \mathcal{D}$ independent of the history. Finally, let $[z]_+ = \max\{0, z\}$ denote the positive part of z.

Lemma 1 (Concentration of the capped correction)**.** *Let $1 \le r \le T$. For any sequence of hypotheses selected before their fresh batches, the preceding quantities satisfy*

$$\mathbb{E}\left[\sup_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} |M_t(h)|\right] \le C\left(\sqrt{nr\beta_0 Q} + Q\right), \tag{9}$$

$$\mathbb{E}\left[\sup_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} [(1-\gamma)\Lambda_t(h) - K_t(h)]_+\right] \le \frac{CQ}{\gamma}, \qquad 0 < \gamma \le \frac{1}{2}. \tag{10}$$

Moreover, writing

$$p(h) = \mathbb{E}\left[\mathbb{1}_B(X)h(X)\right], \qquad \widehat{p}_t(h) = \frac{1}{nt}\sum_{x\in\mathcal{U}_t} \mathbb{1}_B(x)h(x),$$

for every $t \in [T]$,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |\widehat{p}_t(h) - p(h)|\right] \le C\sqrt{\frac{\beta_0 Q}{nt}}. \tag{11}$$

The proof of Lemma 1 is given in Subsubsection A.3.3.

A.3.1 PBERM

Here we describe the positively biased ERM (PBERM) algorithm and prove its regret in the realizable setting. PBERM is a special case of ERM algorithm that breaks ties in favor of hypotheses that produce more positive predictions.

Let $\mathcal{U}_t$ be the multiset of all instances observed before round t as in (6). Let $\mathcal{S}_t := S^{\text{obs}}_{0:t-1}$ be the set of all observed labeled instances before round t. Define the version space

$$V_t = \{h \in \mathcal{H} : h(x) = y \text{ for every } (x, y) \in \mathcal{S}_t\}.$$

On round t, PBERM selects

$$h_t \in \arg\max_{h\in V_t} \sum_{x\in\mathcal{U}_t} h(x). \tag{12}$$

Thus, the learner selects a consistent hypothesis with the largest number of positive predictions on the accumulated instances. After implementing h_t on X_t, update

$$\mathcal{U}_{t+1} = \mathcal{U}_t \cup X_t, \qquad \mathcal{S}_{t+1} = \mathcal{S}_t \cup \{(X_{t,i}, Y_{t,i}) : h_t(X_{t,i}) = 1\}.$$

The algorithm is summarized in Algorithm 1. Notice that it does not require knowledge of the horizon T.

Algorithm 1 PBERM

1: Implement h_0 on X_0 and initialize $\mathcal{U}_1$ and $\mathcal{S}_1$.
2: **for** $t = 1, \ldots, T$ **do**
3: Choose h_t according to (12).
4: Implement h_t on the fresh batch X_t and observe labels where $h_t(X_{t,i}) = 1$.
5: Append all current feature vectors to $\mathcal{U}_t$, obtaining $\mathcal{U}_{t+1}$.
6: Append all revealed labeled examples to $\mathcal{S}_t$, obtaining $\mathcal{S}_{t+1}$.
7: **end for**

Theorem 4 (Realizable regret of PBERM). *Suppose $Y = h^*(X)$ almost surely for some $h^* \in \mathcal{H}$, and let $c_{\mathrm{fp}}, c_{\mathrm{fn}} > 0$. Algorithm 1 satisfies*

$$\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{R}_T^{\mathrm{off}} \le C c_{\max}\left(n\beta_0 + \sqrt{nTQ}\right), \tag{13}$$

where $C > 0$ is universal. If $K = |\mathcal{H}| < \infty$, it also satisfies

$$\mathcal{R}_T^{\mathrm{off}} \le c_{\max}(21 + 3n\beta_0)(K-1). \tag{14}$$

Proof. Under realizability, $h^* \in V_t$ for every t, and $\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{R}_T^{\mathrm{off}}$ as observed in the notation above. Write

$$F(h) = \mathbb{E}\left[h(X)(1-h^*(X))\right], \qquad G(h) = \mathbb{E}\left[(1-h(X))h^*(X)\right].$$

Then $\ell_{\mathcal{D}}(h) = c_{\mathrm{fp}}F(h) + c_{\mathrm{fn}}G(h) \le c_{\max}\{F(h)+G(h)\}$. Using the notation of Lemma 1, set

$$\Delta_t = 2\sup_{h\in\mathcal{H}} |\widehat{p}_t(h) - p(h)|.$$

By (11),

$$\mathbb{E}\left[\Delta_t\right] \le C\sqrt{\frac{\beta_0 Q}{nt}}. \tag{15}$$

Since every $h \in \mathcal{H}$ agrees with h^* outside B, $F(h) - G(h) = p(h) - p(h^*)$. Moreover, feasibility of h^* and (12) imply $\widehat{p}_t(h_t) \ge \widehat{p}_t(h^*)$. Consequently, for $h_t \in V_t \subseteq \mathcal{H}$,

$$G(h_t) \le F(h_t) + \Delta_t. \tag{16}$$

We first prove (13). If $d = 0$, every hypothesis agrees with h^* and the regret is zero. If $d > N$, then $Q \ge N\log 2$, so the trivial bound $\mathcal{R}_T^{\mathrm{off}} \le c_{\max}N\beta_0$ proves the claim. Thus, it remains to consider $1 \le d \le N$. Fix an integer $1 \le r \le T$, used only in the analysis, and define

$$A_t(h) = \frac{1}{r}\mathbb{E}\left[\mathbb{1}_B(X)h(X)(1-h^*(X))\min\{C_t(X), r\}\right],$$
$$u_t(h) = F(h) - A_t(h) = \mathbb{E}\left[\mathbb{1}_B(X)h(X)(1-h^*(X))\left(1 - \frac{C_t(X)}{r}\right)_+\right].$$

These expectations integrate a fresh feature vector while holding the realized history fixed. Under realizability, (8) gives $\Lambda_t(h) = nrA_t(h)$. Also, consistency gives $K_t(h_t) = 0$, since h_t rejects every previously observed negative, including those counted by the analytical cap. Set

$$J = \sup_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} \left[\tfrac{7}{8}\Lambda_t(h) - K_t(h)\right]_+.$$

By (10) with $\gamma = 1/8$, $\mathbb{E}\left[J\right] \le CQ$. Hence, simultaneously for all $t \in [T]$,

$$\frac{7}{8}nrA_t(h_t) \le J, \qquad F(h_t) \le u_t(h_t) + \frac{8J}{7nr}.$$

For each x, $C_{j+1}(x) = C_j(x) + h_j(x)$. Therefore,

$$\sum_{t=1}^{T} h_t(x)\left(1 - \frac{C_t(x)}{r}\right)_+ \le \sum_{k=0}^{r-1}(1-k/r) = \frac{r+1}{2}.$$

Multiplying by $\mathbb{1}_B(x)(1-h^*(x))$ and integrating gives $\sum_{t=1}^{T} u_t(h_t) \le \beta_0(r+1)/2$ pathwise. Combining the preceding estimates with (16) and (15), and using $\sum_{t=1}^{T} t^{-1/2} \le 2\sqrt{T}$, yields

$$\mathcal{R}_T^{\mathrm{off}} \le C c_{\max}\left(n\beta_0 r + \frac{TQ}{r} + \sqrt{n\beta_0 TQ}\right). \tag{17}$$

Choose $r = \min\{T, \max\{1, \lceil\sqrt{TQ/n}\rceil\}\}$. If $Q \le nT$ and $TQ \le n$, then $r = 1$ and $TQ \le \sqrt{nTQ}$. If $Q \le nT$ and $TQ > n$, then $\sqrt{TQ/n} \le r \le 1 + \sqrt{TQ/n}$. In both cases, substitution in (17) and $\beta_0 \le 1$ give (13). If $Q > nT$, the trivial bound $\mathcal{R}_T^{\mathrm{off}} \le c_{\max}nT\beta_0$ gives the same conclusion.

We next prove (14). Let Z_t count false positives in batch t. Conditional on $\mathcal{F}_t$, $Z_t \sim \mathrm{Bin}(n, F(h_t))$. Every round with $Z_t > 0$ permanently eliminates h_t from the version space, whereas h^* is never eliminated. Thus, pathwise,

$$\sum_{t=1}^{T} \mathbb{1}\{Z_t > 0\} \le K - 1.$$

For $p \in [0,1]$, $(1-p)^n \le (1+np)^{-1}$, and hence $np \le (1+np)\{1-(1-p)^n\}$. Since $F(h_t) \le \beta_0$,

$$\mathbb{E}\left[Z_t \mid \mathcal{F}_t\right] = nF(h_t) \le (1+n\beta_0)\Pr(Z_t > 0 \mid \mathcal{F}_t).$$

Taking expectations and summing gives

$$n\sum_{t=1}^{T} \mathbb{E}\left[F(h_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} Z_t\right] \le (1+n\beta_0)(K-1). \tag{18}$$

Partition the fixed hypotheses according to whether $G(h) \le 2F(h)$. For this first group, $\ell_{\mathcal{D}}(h) \le 3c_{\max}F(h)$, so (18) bounds its contribution to $\mathcal{R}_T^{\mathrm{off}}$ by $3c_{\max}(1+n\beta_0)(K-1)$.

For a fixed h in the second group, $G(h) > 2F(h)$. If $h_t = h$, then (12) and feasibility of h^* imply

$$\sum_{x \in \mathcal{U}_t} \{h(x) - h^*(x)\} \ge 0.$$

For a fresh feature vector, let $W = h(X) - h^*(X)$ and $\lambda = \log(4/3)$. Then

$$\mathbb{E}\left[e^{\lambda W}\right] = 1 + F(h)/3 - G(h)/4 \le 1 - G(h)/12 \le e^{-G(h)/12}.$$

The nt features in $\mathcal{U}_t$ are i.i.d., so Markov's inequality gives

$$\Pr(h_t = h) \le \Pr\left(\sum_{x \in \mathcal{U}_t} \{h(x) - h^*(x)\} \ge 0\right) \le e^{-ntG(h)/12}.$$

This is an unconditional estimate for a fixed candidate. Since $\ell_{\mathcal{D}}(h) \le 3c_{\max}G(h)/2$, its contribution to $\mathcal{R}_T^{\mathrm{off}}$ is at most

$$n\ell_{\mathcal{D}}(h)\sum_{t=1}^{T}\Pr(h_t = h) \le \frac{3c_{\max}nG(h)/2}{e^{nG(h)/12} - 1} \le 18c_{\max},$$

where the last inequality uses $e^z - 1 \ge z$ for $z \ge 0$. There are at most $K-1$ hypotheses in this group, since it does not contain h^*. Adding the two contributions gives

$$\mathcal{R}_T^{\mathrm{off}} \le c_{\max}\{3(1+n\beta_0) + 18\}(K-1),$$

which proves (14). □

For finite classes, PBERM therefore satisfies the minimum of (13) and (14). The parameter r introduced in the proof is used only for analysis and is not an input to PBERM.

A.3.2 Capped Exploration ERM

Here we describe and analyze the capped exploration ERM (CEERM) algorithm. We first apply a regret-preserving transformation of the loss so that the loss of the selected action can always be computed from the available feedback. This gives a reduction from apple tasting to contextual bandit feedback, following the same principle as Bechavod et al. (2019, Proposition 2.3). Let $a = c_{\mathrm{fn}}$ and $b = c_{\mathrm{fn}} + c_{\mathrm{fp}}$. We have:

$$\begin{aligned}
\ell(h,x,y) &= c_{\mathrm{fn}}(1-h(x))y + c_{\mathrm{fp}}h(x)(1-y) \\
&= c_{\mathrm{fn}}y - c_{\mathrm{fn}}h(x)y + c_{\mathrm{fp}}h(x)(1-y) \\
&= c_{\mathrm{fn}}y - c_{\mathrm{fn}}h(x) + c_{\mathrm{fn}}h(x)(1-y) + c_{\mathrm{fp}}h(x)(1-y) \\
&= ay - ah(x) + bh(x)(1-y).
\end{aligned}$$

Since ay is independent of the prediction, removing it preserves both regret benchmarks. Define

$$\widetilde{\ell}(h,x,y) = -ah(x) + bh(x)(1-y).$$

Then the expected distributional regret computed with $\widetilde{\ell}$ is exactly $\mathcal{R}_T^{\mathrm{off}}$.

We now describe our algorithm. Fix an integer $r \geq 1$, and use $\mathcal{U}_t$ and C_t from (6). Let $\mathcal{N}_t$ be the multiset of retained observed negative examples (i.e., false positives), initialized with $\mathcal{N}_1 = \emptyset$. At the beginning of round t, select

$$h_t \in \arg\min_{h \in \mathcal{H}} \left\{ -\frac{a}{nt} \sum_{x \in \mathcal{U}_t} h(x) + \frac{b}{nr} \sum_{(x,0) \in \mathcal{N}_t} h(x) \right\}. \tag{19}$$

The learner implements h_t on the fresh batch X_t and observes labels only where $h_t(X_{t,i}) = 1$. It then updates

$$\mathcal{U}_{t+1} = \mathcal{U}_t \cup X_t$$

and

$$\mathcal{N}_{t+1} = \mathcal{N}_t \cup \{(X_{t,i}, 0) : h_t(X_{t,i}) = 1,\ Y_{t,i} = 0,\ C_t(X_{t,i}) < r\}.$$

Labels revealed during initialization are not used in the correction term. CEERM is summarized in Algorithm 2.

Algorithm 2 Capped Exploration ERM

1: Implement h_0 on X_0.
2: Initialize $\mathcal{U}_1 = X_0$, $\mathcal{N}_1 = \emptyset$, and $C_1 = 0$.
3: **for** $t = 1, \ldots, T$ **do**
4: Choose h_t according to (19).
5: Implement h_t on the fresh batch X_t and observe labels where $h_t(X_{t,i}) = 1$.
6: Add each observed negative with $C_t(X_{t,i}) < r$ to $\mathcal{N}_t$, obtaining $\mathcal{N}_{t+1}$.
7: Set $\mathcal{U}_{t+1} = \mathcal{U}_t \cup X_t$.
8: Set $C_{t+1}(x) = C_t(x) + h_t(x)$ for every x.
9: **end for**

Theorem 5 (Realizable regret of CEERM)**.** *Suppose $Y = h^*(X)$ almost surely for some $h^* \in \mathcal{H}$, and let $c_{\mathrm{fp}}, c_{\mathrm{fn}} > 0$. For every integer $1 \leq r \leq T$, Algorithm 2 satisfies*

$$\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{R}_T^{\mathrm{off}} \leq C\left(bn\beta_0 r + \frac{b^2}{c_{\mathrm{fp}}} \frac{TQ}{r} + a\sqrt{n\beta_0 TQ} \right), \tag{20}$$

where $C > 0$ is universal. Choosing

$$r = \min\left\{ T, \max\left\{ 1, \left\lceil \sqrt{\frac{bTQ}{c_{\mathrm{fp}} n}} \right\rceil \right\} \right\}$$

gives

$$\mathcal{R}_T^{\mathrm{off}} \leq Cb\left(n\beta_0 + \sqrt{\frac{b}{c_{\mathrm{fp}}} nTQ} \right). \tag{21}$$

Proof. Under realizability, $\mathbb{E}\left[\mathcal{R}_T\right] = \mathcal{R}_T^{\mathrm{off}}$ by the regret comparison above. If $d = 0$, the regret is zero. If $d > N$, then $Q \geq N \log 2$, and both bounds follow from $\mathcal{R}_T^{\mathrm{off}} \leq c_{\max} N\beta_0$, using $r \leq T$ and $b^2/c_{\mathrm{fp}} \geq b \geq c_{\max}$. Thus, assume $1 \leq d \leq N$.

Use F, G, A_t, u_t, and Δ_t defined in the proof of Theorem 4, with the policies now selected by CEERM and r its cap. In particular, $\Lambda_t(h) = nrA_t(h)$ and $\mathbb{E}\left[\Delta_t\right] \leq C\sqrt{\beta_0 Q/(nt)}$. Let

$$\gamma = \frac{c_{\mathrm{fp}}}{2b}, \qquad J = \sup_{\substack{1 \leq t \leq T+1 \\ h \in \mathcal{H}}} \left[(1-\gamma)\Lambda_t(h) - K_t(h) \right]_+.$$

Since $0 < \gamma < 1/2$, (10) gives $\mathbb{E}\left[J\right] \leq CbQ/c_{\mathrm{fp}}$.

The contributions outside B in (19) are independent of the candidate h, so removing them leaves the minimizers unchanged. Also, realizability gives $K_t(h^*) = 0$ almost surely. Comparing the objectives at h_t and h^* therefore yields

$$\begin{aligned} \frac{b}{nr} K_t(h_t) &\leq a\{\widehat{p}_t(h_t) - \widehat{p}_t(h^*)\} \\ &\leq a\{F(h_t) - G(h_t) + \Delta_t\}. \end{aligned}$$

By the definition of J, $K_t(h_t) \geq (1-\gamma)nrA_t(h_t) - J$. Substituting this lower bound and $A_t(h_t) = F(h_t) - u_t(h_t)$ into the preceding inequality gives

$$\{(1-\gamma)b - a\}F(h_t) + aG(h_t) \leq (1-\gamma)b\, u_t(h_t) + \frac{bJ}{nr} + a\Delta_t.$$

Our choice of γ ensures $(1-\gamma)b - a = c_{\mathrm{fp}}/2$. The left side is therefore at least $\ell_{\mathcal{D}}(h_t)/2$, so

$$\ell_{\mathcal{D}}(h_t) \leq 2b\, u_t(h_t) + \frac{2bJ}{nr} + 2a\Delta_t. \tag{22}$$

The pointwise counting argument in the proof of Theorem 4 applies to any sequence of policies:

$$\sum_{t=1}^{T} h_t(x)\left(1 - \frac{C_t(x)}{r}\right)_+ \leq \frac{r+1}{2}.$$

Multiplying by $\mathbb{1}_B(x)(1-h^*(x))$ and integrating gives $\sum_{t=1}^{T} u_t(h_t) \leq \beta_0(r+1)/2$ pathwise. Taking expectations in (22), multiplying by n, and summing over t gives

$$\mathcal{R}_T^{\mathrm{off}} \leq bn\beta_0(r+1) + \frac{2bT}{r}\mathbb{E}\left[J\right] + 2an\sum_{t=1}^{T}\mathbb{E}\left[\Delta_t\right].$$

Using the bounds on $\mathbb{E}\left[J\right]$ and $\mathbb{E}\left[\Delta_t\right]$, $r+1 \leq 2r$, and $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$ proves (20).

For the optimized bound, write $\rho = \sqrt{bTQ/(c_{\mathrm{fp}}n)}$. If $\rho \leq T$, the stated choice satisfies $\rho \leq r \leq 1+\rho$, and

$$bn\beta_0 r \leq bn\beta_0 + bn\rho, \qquad \frac{b^2TQ}{c_{\mathrm{fp}}r} \leq bn\rho, \qquad a\sqrt{n\beta_0 TQ} \leq bn\rho.$$

Substitution gives (21). If $\rho > T$, the trivial bound $\mathcal{R}_T^{\mathrm{off}} \leq c_{\max}nT\beta_0 \leq bn\rho$ gives the same conclusion. □

For fixed positive costs, (21) gives $O(n\beta_0 + \sqrt{nTQ})$ regret.

Theorem 6 (Agnostic regret of capped exploration ERM)**.** *Let $c_{\mathrm{fp}}, c_{\mathrm{fn}} > 0$. For every integer $1 \leq r \leq T$, Algorithm 2 satisfies*

$$\mathcal{R}_T^{\mathrm{off}} \leq \frac{b}{2}n\beta_0(r+1) + Cb\left(T\sqrt{\frac{n\beta_0 Q}{r}} + \frac{TQ}{r}\right) + Ca\sqrt{n\beta_0 TQ}, \tag{23}$$

where $C > 0$ is universal. Choosing

$$r = \min\left\{T, \max\left\{1, \left\lceil (T^2Q/n)^{1/3}\right\rceil\right\}\right\}$$

gives

$$\mathcal{R}_T^{\mathrm{off}} \leq Cb\left(n\beta_0 + (nT)^{2/3}Q^{1/3}\right). \tag{24}$$

The same optimized bound holds for $\mathbb{E}\left[\mathcal{R}_T\right]$.

Proof. Since all hypotheses agree outside B, pathwise,

$$\mathcal{R}_T \leq c_{\max}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{1}_B(X_{t,i}).$$

Consequently, $\mathcal{R}_T^{\mathrm{off}} \leq \mathbb{E}\left[\mathcal{R}_T\right] \leq c_{\max}N\beta_0$. If $d = 0$, both regrets are zero. If $d > N$, then $Q \geq N\log 2$, and the trivial bound proves the claims, using $TQ/r \geq Q$ and $b \geq c_{\max}$. Thus, assume $1 \leq d \leq N$.

Define

$$\mu(h) = \mathbb{E}\left[\mathbb{1}_B(X)\widetilde{\ell}(h, X, Y)\right].$$

Agreement outside B and the loss transformation imply

$$\mu(h) - \mu(h^*) = \ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(h^*).$$

Define the score

$$q_t(h) = -a\,p(h) + \frac{b}{nr}\Lambda_t(h).$$

By (8), its deficit relative to $\mu(h)$ is

$$d_t(h) := \mu(h) - q_t(h) = b\mathbb{E}\left[\mathbb{1}_B(X)h(X)(1-Y)\left(1 - \frac{C_t(X)}{r}\right)_+\right] \geq 0.$$

These expectations integrate a fresh draw while holding the realized history fixed. The same pointwise counting argument as in the realizable proof gives, pathwise,

$$\sum_{t=1}^{T} d_t(h_t) \leq \frac{b}{2}\beta_0(r+1). \tag{25}$$

Removing contributions outside B from (19) gives

$$\widehat{q}_t(h) = -a\,\widehat{p}_t(h) + \frac{b}{nr}K_t(h).$$

These removed contributions are independent of h, so h_t minimizes $\widehat{q}_t$. Moreover,

$$\widehat{q}_t(h) - q_t(h) = -a\{\widehat{p}_t(h) - p(h)\} + \frac{b}{nr}M_t(h).$$

Since $q_t(h^*) \leq \mu(h^*)$, we obtain

$$\begin{aligned}
\ell_{\mathcal{D}}(h_t) - \ell_{\mathcal{D}}(h^*) &= \mu(h_t) - \mu(h^*) \\
&\leq d_t(h_t) + q_t(h_t) - q_t(h^*) \\
&\leq d_t(h_t) + 2\sup_{h\in\mathcal{H}} |\widehat{q}_t(h) - q_t(h)| \\
&\leq d_t(h_t) + 2a\sup_{h\in\mathcal{H}} |\widehat{p}_t(h) - p(h)| + \frac{2b}{nr}\sup_{h\in\mathcal{H}} |M_t(h)|.
\end{aligned}$$

Taking expectations, multiplying by n, and summing over t, Lemma 1 and (25) give

$$\mathcal{R}_T^{\mathrm{off}} \leq \frac{b}{2}n\beta_0(r+1) + Cb\left(T\sqrt{\frac{n\beta_0 Q}{r}} + \frac{TQ}{r}\right) + Ca\sqrt{n\beta_0 TQ},$$

where we used $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$. This proves (23).

It remains to choose the cap. Suppose first that $Q \leq nT$, and write $\rho = (T^2Q/n)^{1/3}$. Then $\rho \leq T$, and the stated choice satisfies $\rho \leq r \leq 1 + \rho$. Using $\beta_0 \leq 1$ and $Q \leq nT$, we have

$$\begin{aligned}
\frac{1}{2}n\beta_0(r+1) &\leq n\beta_0 + (nT)^{2/3}Q^{1/3}, \\
T\sqrt{\frac{n\beta_0 Q}{r}} &\leq (nT)^{2/3}Q^{1/3}, \\
\frac{TQ}{r} &\leq (nT)^{2/3}Q^{1/3}, \\
\sqrt{n\beta_0 TQ} &\leq (nT)^{2/3}Q^{1/3}.
\end{aligned}$$

Since $a \leq b$, substitution proves (24). If $Q > nT$, the trivial bound

$$\mathcal{R}_T^{\mathrm{off}} \leq c_{\max}nT\beta_0 \leq b(nT)^{2/3}Q^{1/3}$$

gives the same conclusion.

Finally, when $Q \leq N$, the regret comparison at the start of this subsection gives

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \mathcal{R}_T^{\mathrm{off}} + Cc_{\max}\sqrt{NQ},$$

and $\sqrt{NQ} \leq N^{2/3}Q^{1/3}$ absorbs the additional term into (24). When $Q > N$, the trivial bound $\mathbb{E}\left[\mathcal{R}_T\right] \leq c_{\max}N\beta_0$ proves the same optimized rate directly. □

The realizable and agnostic rates use different choices of r. The agnostic theorem also applies to realizable distributions, while (21) uses the realizable choice of the cap.

A.3.3 Proof of the Key Lemma

Lemma 1 (Concentration of the capped correction)**.** *Let $1 \le r \le T$. For any sequence of hypotheses selected before their fresh batches, the preceding quantities satisfy*

$$\mathbb{E}\left[\sup_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} |M_t(h)|\right] \le C\left(\sqrt{nr\beta_0 Q} + Q\right), \tag{9}$$

$$\mathbb{E}\left[\sup_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} [(1-\gamma)\Lambda_t(h) - K_t(h)]_+\right] \le \frac{CQ}{\gamma}, \qquad 0 < \gamma \le \frac{1}{2}. \tag{10}$$

Moreover, writing

$$p(h) = \mathbb{E}\left[\mathbb{1}_B(X)h(X)\right], \qquad \widehat{p}_t(h) = \frac{1}{nt}\sum_{x\in\mathcal{U}_t} \mathbb{1}_B(x)h(x),$$

for every $t \in [T]$,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |\widehat{p}_t(h) - p(h)|\right] \le C\sqrt{\frac{\beta_0 Q}{nt}}. \tag{11}$$

Proof of Lemma 1. Let us first assume $\mathcal{H}$ is finite and that $Q = \log(2|\mathcal{H}|)$. Fix $h \in \mathcal{H}$ and flatten the indices by $k = (j-1)n + i$, where $j \in [T]$ and $i \in [n]$. For the analysis, reveal the examples in this order. Let $(\mathcal{G}_k)_{k=0}^N$ be the resulting filtration, including initialization and each hypothesis choice before its batch is revealed. Thus, $\mathcal{G}_{k-1}$ contains the previous batches, the current policy h_j, and the first $i-1$ examples of batch j. Since each batch is drawn independently from $\mathcal{D}^n$ and h_j is selected before that batch,

$$\mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{G}_{k-1}\right] = \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{F}_j\right].$$

Define

$$\xi_k = Z_{j,i}(h) - \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{F}_j\right], \qquad S_k = \sum_{\ell=1}^{k} \xi_\ell, \qquad S_0 = 0.$$

Then $\mathbb{E}\left[\xi_k \mid \mathcal{G}_{k-1}\right] = 0$, so (S_k) is a martingale with respect to $(\mathcal{G}_k)$. Moreover, $Z_{j,i}(h) \in \{0,1\}$ implies $|\xi_k| \le 1$ almost surely. At batch boundaries,

$$S_{n(t-1)} = K_t(h) - \Lambda_t(h) = M_t(h).$$

The conditional variance of a binary random variable is at most its conditional mean. Therefore, the predictable quadratic variation from Lemma 4 satisfies

$$\begin{aligned} V_N &= \sum_{k=1}^{N} \mathbb{E}\left[\xi_k^2 \mid \mathcal{G}_{k-1}\right] \\ &\le \sum_{j=1}^{T}\sum_{i=1}^{n} \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{F}_j\right] \\ &\le n\sum_{j=1}^{T} \mathbb{E}\left[\mathbb{1}_B(X)h_j(X)\mathbb{1}\{C_j(X) < r\}\right] \\ &= n\mathbb{E}\left[\mathbb{1}_B(X)\min\{C_{T+1}(X), r\}\right] \\ &\le nr\beta_0. \end{aligned}$$

The equality uses (7). The expectations involving X integrate a fresh feature vector while holding the realized policies fixed, so this bound holds for every realized history.

If $\beta_0 = 0$, the process is zero almost surely. Otherwise, applying Lemma 4 with $m = N$, $L = 1$, and $v = nr\beta_0$ gives, for every $u > 0$,

$$\Pr\left(\max_{1\le t\le T+1} |M_t(h)| \ge \sqrt{2nr\beta_0 u} + \frac{2u}{3}\right) \le 2e^{-u}.$$

A union bound over h, followed by integration of the tail, proves (9).

For the one-sided estimate, fix $h \in \mathcal{H}$ and $\lambda > 0$. Using the same flattened indices and filtration, write

$$p_k = \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{G}_{k-1}\right] = \mathbb{E}\left[Z_{j,i}(h) \mid \mathcal{F}_j\right], \qquad k = (j-1)n + i.$$

Since $Z_{j,i}(h)$ is binary,

$$\mathbb{E}\left[e^{-\lambda Z_{j,i}(h)} \mid \mathcal{G}_{k-1}\right] = 1 - p_k(1 - e^{-\lambda}) \le \exp\{-p_k(1 - e^{-\lambda})\}.$$

Define

$$\mathcal{E}_0 = 1, \qquad \mathcal{E}_k = \mathcal{E}_{k-1} \exp\{(1 - e^{-\lambda})p_k - \lambda Z_{j,i}(h)\}.$$

The preceding inequality gives $\mathbb{E}\left[\mathcal{E}_k \mid \mathcal{G}_{k-1}\right] \le \mathcal{E}_{k-1}$. Thus $(\mathcal{E}_k)_{k=0}^N$ is a nonnegative supermartingale with initial value one. At batch boundaries,

$$\mathcal{E}_{n(t-1)} = \exp\{(1 - e^{-\lambda})\Lambda_t(h) - \lambda K_t(h)\}.$$

Take $\lambda = 2\gamma$. Since

$$\frac{1 - e^{-\lambda}}{\lambda} \ge 1 - \frac{\lambda}{2} = 1 - \gamma$$

and $\Lambda_t(h) \ge 0$, we have

$$\mathcal{E}_{n(t-1)} \ge \exp\left\{2\gamma\big((1-\gamma)\Lambda_t(h) - K_t(h)\big)\right\}.$$

Consequently, Lemma 5 implies, for every $u > 0$,

$$\begin{aligned}\Pr\left(\max_{1\le t\le T+1}\{(1-\gamma)\Lambda_t(h) - K_t(h)\} > u\right) &\le \Pr\left(\max_{0\le k\le N} \mathcal{E}_k > e^{2\gamma u}\right) \\ &\le e^{-2\gamma u}.\end{aligned}$$

Then (10) follows from a union bound and integration of the tail bound of $\max_{\substack{1\le t\le T+1\\ h\in\mathcal{H}}} \left[(1-\gamma)\Lambda_t(h) - K_t(h)\right]_+$.

Now suppose $\mathcal{H}$ has VC dimension $1 \le d \le N$. By Lemma 3, take a fixed $L_1(\mathcal{D}_X)$ cover $V \subseteq \mathcal{H}$ of radius $\varepsilon = d/N$, with $\log(2|V|) \le CQ$. For each $h \in \mathcal{H}$, choose $v_h \in V$ with $\mathcal{D}_X(h \,\Delta\, v_h) \le \varepsilon$. The cover and these choices are independent of the observed data. Also note that the cover is used only in the proof and need not be known by the learner.

Let $(\overline{X}_k)_{k=1}^N$ denote the flattened batches of data, excluding the initial round. For all t, h, pathwise,

$$\begin{aligned}|K_t(h) - K_t(v_h)| &\le \sum_{k=1}^{N} \mathbb{1}\{h(\overline{X}_k) \ne v_h(\overline{X}_k)\}, \\ |\Lambda_t(h) - \Lambda_t(v_h)| &\le nr\mathcal{D}_X(h \,\Delta\, v_h) \le N\varepsilon = d.\end{aligned}$$

The second inequality follows from (8) and $r \le T$. By (29), the supremum over h of the first right-hand side has expectation at most CQ. Also, $d \le CQ$: the VC complexity term is at least d, and, when $\mathcal{H}$ is finite, shattering d points implies $|\mathcal{H}| \ge 2^d$.

Consequently, replacing h by v_h changes $M_t(h)$ and $(1-\gamma)\Lambda_t(h) - K_t(h)$ by quantities whose absolute values have expected supremum at most CQ, uniformly over t, h. The same holds after taking positive parts, since $|[z]_+ - [z']_+| \le |z - z'|$. Apply the preceding finite-class argument with the supremum restricted to V, keeping the original policies and mask B. Using $\log(2|V|) \le CQ$ and $\gamma \le 1/2$, and adding the approximation errors, proves (9) and (10) when $1 \le d \le N$.

Finally, since $\mathcal{U}_t$ consists of nt i.i.d. feature vectors, (11) follows from symmetrization, Massart's lemma, and the cardinality or Sauer bound on the number of traces (Shalev-Shwartz and Ben-David, 2014, Chapters 6 and 26). The factor $\sqrt{\beta_0}$ follows by bounding the squared norm of each masked trace by $\sum_{x\in\mathcal{U}_t} \mathbb{1}_B(x)$ and applying Jensen's inequality, since this count has expectation $nt\beta_0$.

□

B Background Technical Lemmas

We collect the standard inequalities used in the preceding proofs and state them without proof.

For a nonempty class $\mathcal{A}$ of measurable subsets of $\mathcal{X}$, define its growth function by

$$\Pi_{\mathcal{A}}(m) = \sup_{x_1,\dots,x_m \in \mathcal{X}} |\{(\mathbb{1}_A(x_1),\dots,\mathbb{1}_A(x_m)) : A \in \mathcal{A}\}|.$$

For a binary hypothesis class, its growth function is defined by identifying each hypothesis with its positive set.

Lemma 2 (Relative deviation bounds (Anthony and Shawe-Taylor, 1993)). *Let $m \geq 1$, let $W_1,\dots,W_m$ be an i.i.d. sample from a probability distribution P on $\mathcal{X}$, and write*

$$\widehat{P}_m(A) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_A(W_i), \qquad q_m = \log\{2\Pi_{\mathcal{A}}(2m)\}.$$

For every $u > 0$, with probability at least $1 - e^{-u}$, simultaneously for all $A \in \mathcal{A}$,

$$|\widehat{P}_m(A) - P(A)| \leq C\left(\sqrt{\frac{P(A)(q_m+u)}{m}} + \frac{q_m+u}{m}\right), \tag{26}$$

$$|\widehat{P}_m(A) - P(A)| \leq C\left(\sqrt{\frac{\widehat{P}_m(A)(q_m+u)}{m}} + \frac{q_m+u}{m}\right). \tag{27}$$

On the same event,

$$\widehat{P}_m(A) \leq 2P(A) + \frac{C(q_m+u)}{m}, \qquad P(A) \leq 2\widehat{P}_m(A) + \frac{C(q_m+u)}{m}. \tag{28}$$

We will apply this result to the disagreement class

$$\mathcal{H}\,\Delta\,\mathcal{H} = \{\{x : h(x) \neq v(x)\} : h, v \in \mathcal{H}\}.$$

Its growth function satisfies $\Pi_{\mathcal{H}\Delta\mathcal{H}}(m) \leq \Pi_{\mathcal{H}}(m)^2$. If $1 \leq d \leq N$, Sauer's lemma (or cardinality for finite classes) gives $\log\{2\Pi_{\mathcal{H}\Delta\mathcal{H}}(2N)\} \leq CQ$. For an i.i.d. sample $\overline{X}_1,\dots,\overline{X}_N$ from $\mathcal{D}_X$, applying (26) to sets of probability at most d/N bounds their empirical counts by $C(Q+u)$, simultaneously, with probability at least $1-e^{-u}$. Integrating gives

$$\mathbb{E}\left[\sup_{\substack{h,v\in\mathcal{H}\\ \mathcal{D}_X(h\Delta v)\leq d/N}} \sum_{k=1}^{N} \mathbb{1}\{h(\overline{X}_k) \neq v(\overline{X}_k)\}\right] \leq CQ. \tag{29}$$

Here $h\,\Delta\,v = \{x : h(x) \neq v(x)\}$.

Lemma 3 (Haussler's packing and covering bound (Haussler, 1995)). *Suppose $1 \leq d = \mathrm{VCdim}(\mathcal{H}) < \infty$. For $h, v \in \mathcal{H}$, write $h\Delta v = \{x \in \mathcal{X} : h(x) \neq v(x)\}$. For a probability distribution P on $\mathcal{X}$, define the $L_1(P)$ pseudometric*

$$\rho_P(h,v) = P(h\,\Delta\,v).$$

For every $0 < \varepsilon \leq 1$, every subset $V \subseteq \mathcal{H}$ whose distinct elements have distance greater than ε satisfies

$$|V| \leq (C/\varepsilon)^d. \tag{30}$$

Consequently, there is a finite $V \subseteq \mathcal{H}$ such that

$$|V| \leq (C/\varepsilon)^d, \qquad \sup_{h\in\mathcal{H}} \min_{v\in V} \rho_P(h,v) \leq \varepsilon. \tag{31}$$

In particular, taking $P = \mathcal{D}_X$ and $\varepsilon = d/N$ gives $\log(2|V|) \leq CQ$ when $1 \leq d \leq N$.

Lemma 4 (Freedman's maximal inequality (Freedman, 1975)). *Let $(S_k)_{k\geq 0}$ be a real-valued martingale with respect to a filtration $(\mathcal{F}_k)_{k\geq 0}$, with $S_0 = 0$. Define*

$$\xi_k = S_k - S_{k-1}, \qquad V_k = \sum_{j=1}^{k} \mathbb{E}\left[\xi_j^2 \mid \mathcal{F}_{j-1}\right], \qquad V_0 = 0.$$

Suppose $|\xi_k| \le L$ almost surely for every $k \ge 1$, where $L > 0$. Then, for every $x, v > 0$,

$$\Pr\left(\exists k \ge 0 : \ |S_k| \ge x \text{ and } V_k \le v\right) \le 2\exp\left\{-\frac{x^2}{2(v + Lx/3)}\right\}. \tag{32}$$

In particular, if $V_m \le v$ almost surely for an integer $m \ge 1$ and a deterministic $v > 0$, then, for every $u > 0$,

$$\Pr\left(\max_{0\le k\le m} |S_k| \ge \sqrt{2vu} + \frac{2Lu}{3}\right) \le 2e^{-u}. \tag{33}$$

Lemma 5 (Ville's inequality (Ville, 1939))**.** *Let $(Z_k)_{k\ge0}$ be a nonnegative supermartingale with respect to a filtration $(\mathcal{F}_k)_{k\ge0}$. Then, for every $u > 0$,*

$$\Pr\left(\sup_{k\ge0} Z_k \ge u\right) \le \frac{\mathbb{E}\left[Z_0\right]}{u}. \tag{34}$$

In particular, if $\mathbb{E}\left[Z_0\right] \le 1$, then, for every $\delta \in (0, 1)$,

$$\Pr\left(\sup_{k\ge0} Z_k \ge \frac{1}{\delta}\right) \le \delta.$$

C Additional experimental details and results

In this section we provide implementation details and additional results for the experiments. We first describe the data representation and optimization for COMPAS and MNIST, then present the MNIST results and the additional figures for both datasets. All experiments use uniform i.i.d. draws with replacement.

C.1 Implementation details

COMPAS: We use the two-year recidivism data with the following filters: screening is within 30 days of arrest, the recidivism indicator is available and is not -1, the charge degree is available and is not `O`, and the risk-score category is available and is not `N/A`. This leaves $6,172$ instances. We set $y = 1$ for no recorded recidivism within two years and $y = 0$ otherwise.

We use age, sex, race, juvenile felony count, juvenile misdemeanor count, other juvenile offense count, prior conviction count, and charge degree. Numeric features are kept in their original units. In particular, the learner does not use a population mean or standard deviation to normalize them. For each categorical feature, we sort its categories and omit the first category as a reference. This gives 12 feature coordinates. The risk-score category is used only for filtering and is not an input feature. We use a linear model with an intercept, with $p_\theta(x)$ given by the sigmoid of its output. The 13 initial coefficients, including the intercept, are independent normal draws with mean zero and standard deviation 0.25.

We optimize the weighted cross-entropy objective using damped Newton steps with backtracking, with at most $1,000$ iterations per update. The gradient tolerance is 10^{-8} and the step tolerance is 10^{-12}. We use the finite model returned by this optimization even when the convergence criterion is not met within the iteration budget. However, in most cases the convergence criterion was indeed met. The damping used in the Newton step is a numerical stabilization and does not add a term to the objective.

MNIST: We use digits 5 and 8 from the training split as the negative and positive classes, respectively. There are $11,272$ images, consisting of $5,421$ fives and $5,851$ eights. Pixel values are divided by 255. We use a convolutional neural network with two convolutional layers: the first has 8 output channels and a 5×5 kernel, and the second has 16 output channels and a 3×3 kernel. Both convolutions preserve spatial size through padding and are followed by ReLU and 2×2 max pooling. The resulting features are passed through a 32-unit ReLU layer and one output logit. We use the default PyTorch layer initialization.

Each online update uses 50 AdamW steps, with learning rate 10^{-3}, weight decay 10^{-4}, and minibatch size 128. At each step, we draw a minibatch without replacement from the available labeled occurrences, or use all of them if fewer than 128 are available. PRERM independently draws a minibatch in the same way from the feature history for its positivity term. These minibatches are drawn again at each step. Each update starts from the previous network weights with a new optimizer state. We use a fixed number of steps rather than a convergence test.

Round order and available data: All methods and costs share the same stream and initial weights for each seed. Round 1 uses this initial hypothesis without training. On round t, the learner first predicts on the n instances in the batch. It then receives the labels of its positive predictions under ATF, or all labels under FF. These labels are added to $S^{\text{obs}}_{1:t}$, and the model for the next round is trained using the data from the completed rounds.

In the PRERM objective, U_t contains all $n(t-1)$ feature occurrences. Repeated occurrences are retained with their multiplicities in both the labeled data and feature history. The weighted cross entropy is normalized by the number of labeled occurrences, rather than the sum of their class weights. We use

$$\lambda_t = \frac{a_\lambda}{\sqrt{\max(1, |S^{\text{obs}}_{1:t-1}|)}}, \qquad a_\lambda = 0.01, \qquad t \geq 2,$$

for both datasets and all costs. If the labeled set is empty, ERM keeps its initial hypothesis, while PRERM omits the labeled term and optimizes only the positivity term.

Exploration: On each round, EO-ERM and EA-ERM choose k batch positions uniformly without replacement and predict positively on them. They use their current ERM hypothesis on the remaining positions. The exploration budget is constant over rounds and is given by

$$k = \min\left\{n, \left\lceil a_e \sqrt{n/T} \right\rceil\right\}, \qquad a_e \in \{1, 8, 16\}.$$

For the COMPAS comparison with $n = 200, T = 30$, this gives $k \in \{3, 21, 42\}$. For MNIST with $n = 200, T = 55$, it gives $k \in \{2, 16, 31\}$. EO-ERM trains only on the deliberately explored occurrences, while EA-ERM trains on all revealed labels. The exploration positions are shared across the two methods and all costs for a given seed. Larger budgets use nested subsets of the same random permutation. We report all three budgets, and evaluation includes the errors made by exploratory predictions.

Hindsight ERM: For each seed and cost, we train a separate model on all Tn occurrences in the sampled stream, starting from the same initial weights as the online methods. We then apply this fixed model in every round. COMPAS uses the same objective and optimizer as in online training. For MNIST, we use 50 full shuffled epochs with the same AdamW parameters and minibatch size. This baseline optimizes the weighted cross-entropy surrogate; it is not guaranteed to be a minimizer of the binary prediction loss.

Evaluation: We evaluate predictions using all stream labels, including those hidden from a learner under ATF. The cumulative loss is $\mathcal{L}_t = \sum_{\tau=1}^{t}(\text{FP}_\tau + c_{\text{fn}}\text{FN}_\tau)$, with $c_{\text{fn}} \in \{0.1, 0.5, 1, 2\}$ in the main comparisons. All cumulative quantities include round 1. Unless otherwise stated, we report means $\pm$ SEM across 30 independent seeds, paired across algorithms and feedback models. In aggregate figures, shaded regions show SEM across seeds.

C.2 MNIST RESULTS

MNIST: Here we investigate whether similar issues arise in an almost realizable setting with a neural-network hypothesis class. We use digits 5 and 8 as the negative and positive classes, respectively, and draw the stream uniformly with replacement from the $11,272$ training images. We use $n = 200$ and $T = 55$, with the architecture and optimization described above.

We consider the same four false-negative costs as in the COMPAS comparison. PRERM uses $a_\lambda = 0.01$, and EO-ERM and EA-ERM use $k \in \{2, 16, 31\}$. All methods share the stream and initial weights for each seed. Table 2 reports the final cumulative loss over all 30 seeds, including the first round.

ATF vs FF: Receiving ATF instead of FF considerably increases ERM's average loss. PRERM reduces this loss for all four costs, with a larger improvement when false negatives are more costly. Under FF, ERM and PRERM perform similarly. EA-ERM with $k = 2$ has the lowest mean ATF loss when $c_{\text{fn}} \in \{0.1, 0.5\}$, while PRERM has the lowest mean when $c_{\text{fn}} \in \{1, 2\}$. Increasing the exploration budget introduces more false positives and increases the loss of EA-ERM in this setting.

Self-confirming traps: As in COMPAS, ERM's performance under ATF varies considerably across seeds and depends on how many labels the initial hypothesis reveals. In 15 of the 30 seeds, at most 1% of the first-batch labels are revealed. In the remaining 15 seeds, at least 95.5% are revealed. Figure 2 shows that PRERM's largest improvements occur among the initial hypotheses that reveal few labels. When $c_{\text{fn}} \in \{1, 2\}$, PRERM has lower final loss in all 15 of these seeds. However, it does not improve the loss for every seed or every cost.

C.3 SELF-REINFORCING MISTAKES

Here we compare ERM under ATF and FF after removing seeds with bad initialization. We use $n = 200, T = 500$, and $c_{\text{fn}} = 1$ for both datasets. To separate this issue from an initial hypothesis that reveals few labels, we retain a seed if its initial hypothesis predicts positively on at least 25% of the first batch instances. This is only for diagnosis and the learner does not use any extra information. This leaves 18 COMPAS seeds and 15 MNIST seeds.

Table 2: Final cumulative loss $\mathcal{L}_T$ (mean $\pm$ SEM over all 30 seeds) for MNIST, $n = 200$, $T = 55$, and i.i.d. sampling. PRERM uses $a_\lambda = 0.01$. Exploration methods use the constant budget with the indicated k. The lowest mean loss among ATF methods in each column is shown in bold.

Feedback	Method	$c_{\text{fn}} = 0.1$	$c_{\text{fn}} = 0.5$	$c_{\text{fn}} = 1$	$c_{\text{fn}} = 2$
-	Hindsight ERM	0.30 ± 0.15	0.07 ± 0.07	0.00 ± 0.00	2.07 ± 2.07
FF	ERM	96.58 ± 7.69	136.72 ± 5.41	184.47 ± 3.05	315.50 ± 15.23
FF	PRERM	94.84 ± 7.54	135.17 ± 5.37	182.30 ± 3.09	315.77 ± 14.92
ATF	ERM	364.08 ± 38.34	1521.92 ± 246.87	2957.27 ± 509.76	5849.73 ± 1031.53
ATF	PRERM	334.36 ± 42.06	452.88 ± 125.07	$\mathbf{320.00 \pm 20.78}$	$\mathbf{436.70 \pm 28.81}$
ATF	EA-ERM, $k = 2$	$\mathbf{201.92 \pm 3.80}$	$\mathbf{275.48 \pm 8.19}$	338.97 ± 16.84	478.67 ± 31.89
ATF	EA-ERM, $k = 16$	532.67 ± 6.22	587.07 ± 4.44	645.43 ± 7.82	764.83 ± 18.09
ATF	EA-ERM, $k = 31$	919.02 ± 7.24	964.02 ± 4.64	1007.70 ± 5.63	1123.70 ± 15.42
ATF	EO-ERM, $k = 2$	680.69 ± 33.05	1084.32 ± 75.94	1358.70 ± 38.39	2010.17 ± 54.66
ATF	EO-ERM, $k = 16$	593.01 ± 8.54	695.37 ± 7.75	798.17 ± 9.84	1003.80 ± 19.91
ATF	EO-ERM, $k = 31$	947.39 ± 7.64	1017.13 ± 5.43	1090.00 ± 5.18	1247.53 ± 15.84

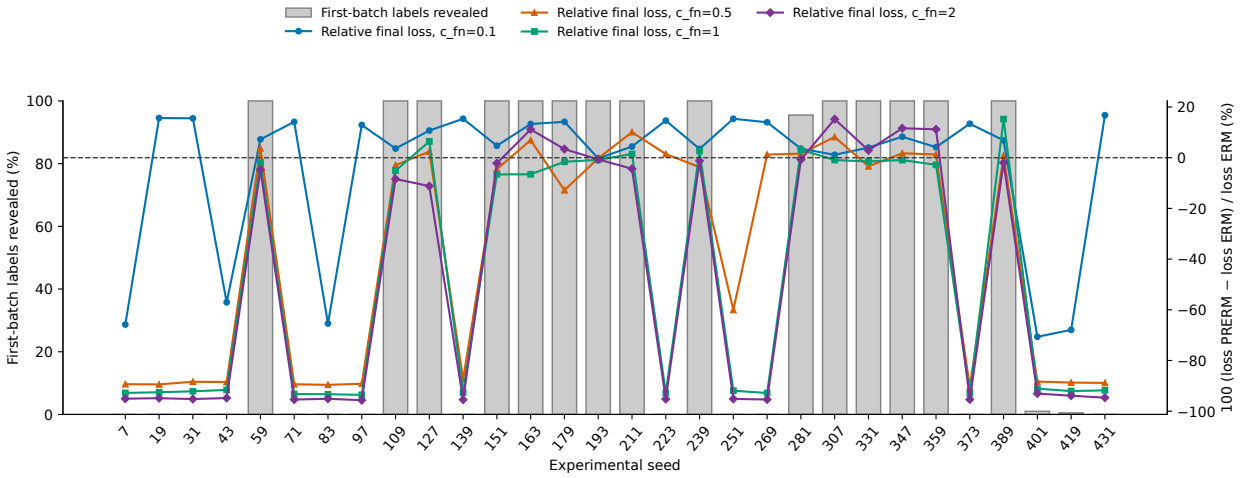


Figure 2: The effect of the number of labels revealed in the first batch on MNIST, with $n = 200$, $T = 55$, and i.i.d. sampling. Each x-axis position corresponds to one of the 30 seeds. Gray bars show the percentage of first-batch labels revealed. The lines show $100(\mathcal{L}_T^{\text{PRERM}} - \mathcal{L}_T^{\text{ERM}})/\mathcal{L}_T^{\text{ERM}}$ under ATF for the four false-negative costs. A negative value indicates that PRERM performs better. PRERM uses $a_\lambda = 0.01$.

Let P_t be the number of instances with true label $y = 1$ in batch t, including those missed by the learner. We consider the percentage of positives missed in each batch and its cumulative counterpart:

$$100\frac{\text{FN}_t}{P_t}, \qquad 100\frac{\sum_{\tau=1}^{t}\text{FN}_\tau}{\sum_{\tau=1}^{t}P_\tau}. \tag{35}$$

We compute each ratio within a seed before averaging across retained seeds. In particular, the denominator here is the number of actual positive instances, rather than the total number of instances or the number correctly predicted positive.

Table 3 reports the mean per-round percentage of positives missed over the final 100 rounds. We first average rounds 401–500 within each seed, then compute the mean and SEM across seeds. The ATF–FF difference is also computed within each seed before aggregation.

For COMPAS, ERM under FF approaches the hindsight baseline, while ERM under ATF continues to miss a larger fraction of positive instances. This is consistent with the self-reinforcing mistakes issue described in the main text: after a learner predicts negatively on a region, ATF does not reveal its mistakes there. For MNIST, the ATF–FF gap is much smaller, although it is not zero. This supports the observation that the remaining feedback gap is more noticeable for the noisier COMPAS data in these experiments. Figures 3 and 4 show both the per-round and cumulative rates. We also observe that, over later rounds, the cumulative fraction of positive instances missed by ERM under ATF increases

Table 3: Percentage of positives missed in rounds 401–500, with $n = 200$, $T = 500$, $c_{\mathrm{fn}} = 1$, and i.i.d. sampling. Values are mean $\pm$ SEM across the retained seeds. The paired ATF–FF gap is in percentage points.

Dataset	Seeds	ERM, ATF	ERM, FF	Hindsight ERM	ATF–FF gap
COMPAS	18	23.350 ± 0.439	21.811 ± 0.130	21.787 ± 0.134	1.538 ± 0.381
MNIST	15	0.142 ± 0.018	0.001 ± 0.001	0.000 ± 0.000	0.142 ± 0.018

slightly for COMPAS, whereas it decreases for MNIST. Together with the persistent ATF–FF gap for COMPAS, this is consistent with our hypothesis.

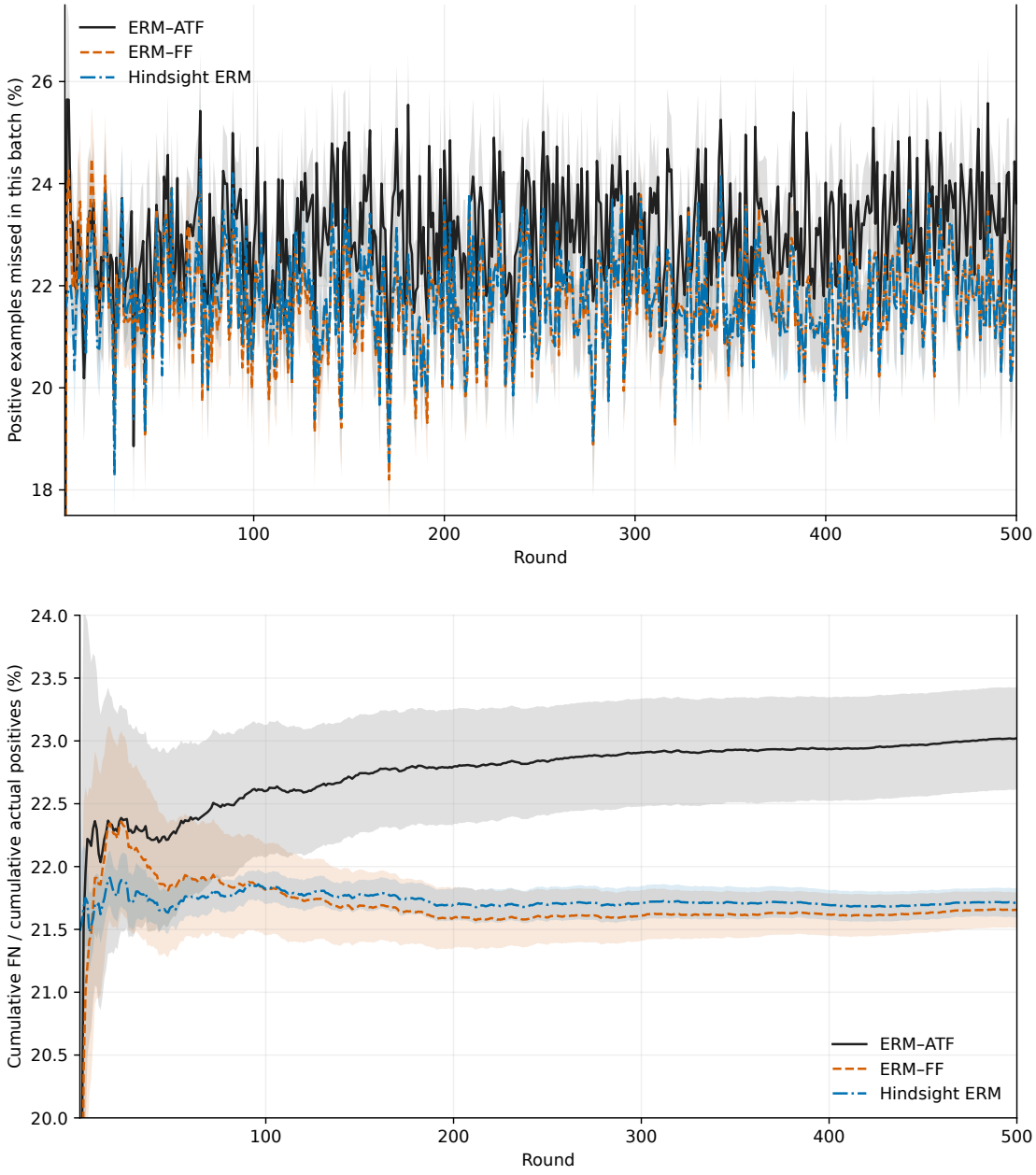


Figure 3: Self-reinforcing mistakes on COMPAS, with $n = 200$, $T = 500$, $c_{\mathrm{fn}} = 1$, and i.i.d. sampling. The top panel shows the percentage of positives missed in each batch; the bottom panel shows the cumulative percentage of positives missed. Curves show mean $\pm$ SEM over the 18 seeds retained by the initial-coverage criterion. Both panels include ERM under ATF and FF and the fixed hindsight learner.

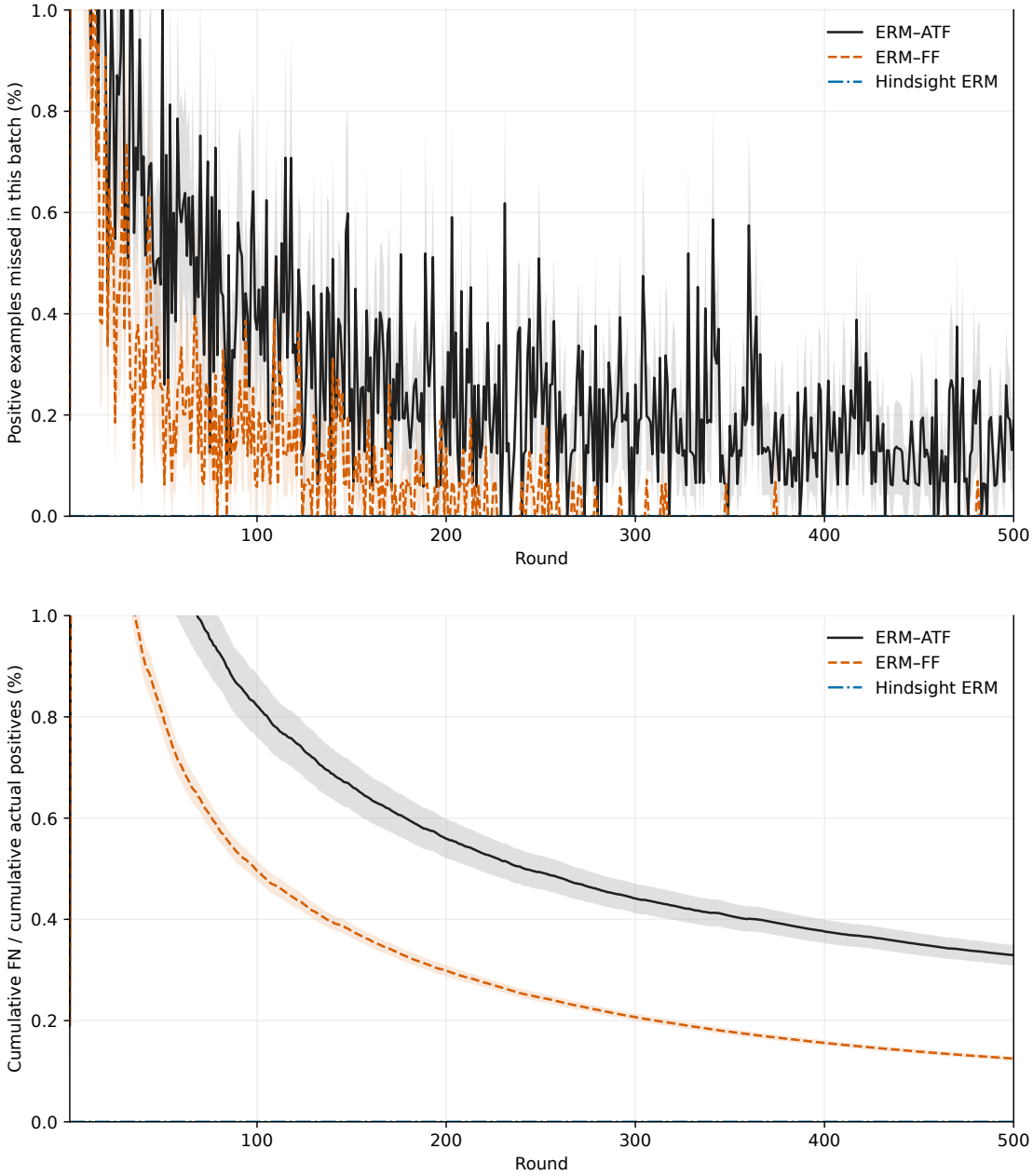


Figure 4: Self-reinforcing mistakes on MNIST, with the same setting as Figure 3. Curves show mean ± SEM over the 15 retained seeds. The top panel shows the per-round percentage of positives missed, and the bottom panel shows its cumulative counterpart. The scales differ from the COMPAS panels.