TIGHT MISTAKE–ABSTENTION TRADEOFFS FOR SEQUENTIAL PREDICTION UNDER ADVERSARIAL INJECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the sequential prediction problem introduced by Goel et al. (2023), where an adversary can inject arbitrarily many corrupted instances in an i.i.d. stream from a clean distribution μ, while the learner is allowed to *abstain* from making a prediction at no cost if the instance was indeed injected. When μ is unknown, Yu & Blanchard (2026b) show a polynomial tradeoff between mistakes and false abstentions for classes with finite VC dimension, but still leaves open a significant gap between upper and lower bounds in their tradeoff landscape. In this paper, we propose the weighted shattering potential, and based on variants of this potential, provide algorithms that essentially characterize the optimal tradeoffs under various settings of interest. When μ is known, our algorithm improves over Goel et al. (2023) and achieves their conjectured $O(d \log T)$ mistake bound. When μ is unknown, for oblivious adversaries, our guarantees are tight both in T and in terms of the VC dimension d, which completely characterizes the abstention-mistake tradeoff in this setting and gives an $O(\sqrt{dT})$ balanced rate. For adaptive adversaries, our results are tight up to $\mathrm{poly}(dD \log T)$ factors for structured classes with finite so-called reduction dimension D. Our guarantees also extend to *censored* feedback, where the label is hidden after an abstention.

1 INTRODUCTION

Consider the sequential prediction problem where a learner sequentially observes each instance and predicts on its label before receiving its true label. We will focus on the realizable setting, where the labels are generated by an unknown target $f^\star$ in a hypothesis class $\mathcal{F}$. The following two assumptions on the sequence of instances are standard in the literature. In the stochastic setting, the instances are drawn i.i.d. from a fixed unknown distribution, whereas in the fully adversarial setting, the instances may be chosen adaptively by an adversary depending on the learner and the past instances. These two settings can have very different complexities. In particular, finite VC dimension is enough for learning in the stochastic setting (Vapnik & Chervonenkis, 1971; 1974; Valiant, 1984), whereas a finite worst-case mistake bound on an adversarial sequence requires finite Littlestone dimension (Littlestone, 1988), which does not even hold for thresholds on the real line.

The clean-label injection model of Goel et al. (2023) lies between these two extremes. In this model, there is an underlying stream of i.i.d. examples from a fixed, unknown clean distribution μ, but an adversary is allowed to inject an arbitrary number of corrupted instances into the stream. The learner observes this combined stream and does not know whether the instances are clean or injected. The learner is allowed to abstain on instances, and its error is measured by the number of mistakes on rounds that the learner makes a prediction, and the number of abstention it makes on clean instances. Thus the learner can safely abstain on an injected instances without any penalty.

When the clean distribution is known, Goel et al. (2023) gave a learner that makes $O(d^2 \log T)$ mistakes and $O(d)$ expected clean abstentions for every class of VC dimension d. On the other hand, when the distribution is unknown and the adversary is oblivious, Yu & Blanchard (2026b) gave a learner that makes $O(T^{3/4})$ mistakes and $O(d^2 (\log T)^{5/3} T^{3/4})$ expected clean abstentions. While these results show that the complexity of learning in this model is governed by the VC dimension, it

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.

is unknown if these are the optimal trade-offs between mistakes and abstentions. This leads to the main question of the paper:

What are the optimal mistake–abstention tradeoffs for sequential prediction under adversarial injections?

Our contributions. We close these gaps in several settings. We write MisErr for the total number of mistakes and AbsErr for the number of abstentions on clean rounds.

Known μ (Theorem 3.1): When the clean distribution is known, for any $\beta > 0$ we give a learner that works against adaptive injections and satisfies :

$$\mathsf{MisErr} \le O(d\beta) \qquad \text{and} \qquad \mathbb{E}[\mathsf{AbsErr}] \le 2T/\beta.$$

The mistake bound holds on every realized interaction. Taking $\beta = T$ gives $O(d \log T)$ mistakes and at most two expected clean abstentions, resolving the question of Goel et al. (2023).

Unknown μ, oblivious adversary (Theorem 4.1): When the clean distribution is unknown and the injections are oblivious, for every $\alpha \in [0, 1)$ we give a learner satisfying :

$$\mathsf{MisErr} = O(T^\alpha) \qquad \text{and} \qquad \mathbb{E}[\mathsf{AbsErr}] = O(dT^{1-\alpha}).$$

For censored feedback, where the label is hidden after an abstention, we also construct a learner essentially achieving the same guarantee up to logarithmic additive factors (Theorem D.1). We complement this with a matching lower bound (Theorem 4.2): for each fixed $\alpha \in [0, 1)$ and $d \ge 1$, there exists a class of VC dimension exactly d and an oblivious clean-label process such that even under full feedback, every learner incurs either

$$\mathbb{E}[\mathsf{MisErr}] = \Omega(T^\alpha) \qquad \text{or} \qquad \mathbb{E}[\mathsf{AbsErr}] = \Omega\big(\min\{dT^{1-\alpha}, T\}\big).$$

Thus the lower bound matches the upper tradeoff throughout the full parameter range.

Unknown μ, adaptive adversary (Theorem 5.1): When the injections are adaptive, we improve the mistake/abstention-error tradeoffs from Yu & Blanchard (2026b) for classes with finite reduction dimension. We use a similar candidate-bank reduction, with (1) the weighted potential supplying a sharper comparator (2) a sample-efficient score compression that largely reduces the number of experts and (3) a sharper expert aggregation based on multiplicative weights. Combining the three leads to the following guarantee: If the function class has $O(d \log T)$-reduction dimension D, for any $\alpha \in [0, 1]$, we give a learner such that

$$\mathsf{MisErr} = O(T^\alpha + d^2 D \log^4 T) \qquad \text{and} \qquad \mathbb{E}[\mathsf{AbsErr}] = O(d^2 D\, T^{1-\alpha} \log^4 T),$$

even under censored feedback. In particular, linear classifiers in $\mathbb{R}^p$ has $D = O(p^3 \log T)$ (Yu & Blanchard, 2026a), hence $\mathsf{MisErr} = O(T^\alpha + p^5 \log^5 T)$ and $\mathbb{E}[\mathsf{AbsErr}] = O(p^5 T^{1-\alpha} \log^5 T)$.

Related work. The closest work is that of Goel et al. (2023), who introduced the model. When the clean distribution is known, their algorithm follows one shattering level at a time and obtains a quadratic dependence on the VC dimension. For unknown distribution, they give algorithms for VC-dimension-one classes and axis-aligned rectangles, and suggest leave-out shattering statistics as a possible route to general classes. Our known-distribution learner combines all shattering levels in one potential. Our unknown-distribution learner uses the same kind of leave-out witnesses, but counts them deterministically instead of treating them as estimates of a population probability.

Yu & Blanchard (2026b) obtained the first unknown-distribution guarantee for every finite-VC class under an oblivious adversary. Their construction builds many candidate labeled banks and aggregates the resulting learners so that at least one candidate is based on clean data. Our adaptive learner builds on their candidate-bank framework but sharpens it using compressed shattering witnesses and direct multiplicative-weights aggregation. Our oblivious learners do not require candidate-bank aggregation.

Edelman & Goel (2026) take a different approach based on robust witnesses: a small labeled subset certifies a prediction and forces progress even when the history is contaminated. They obtain class-specific upper bounds and an $\Omega(\sqrt{T})$ lower bound at VC dimension one. We extend the lower bound to VC dimension d classes. Also, our algorithm follows essentially the template of their potential-based algorithm with the potential instantiated as a weighted shattering.

For a more detailed discussion of related work, we refer the reader to Appendix A.

2 SETUP

We consider the sequential online learning model with clean-label corruptions introduced by Goel et al. (2023). Throughout, we denote by $\mathcal{X}$ the instance space and let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. In the classical sequential prediction problem—without any injections—the VC dimension characterizes the learnability of a function class (Vapnik & Chervonenkis, 1971; 1974; Valiant, 1984).

Definition 2.1 (Shattering and VC dimension)**.** Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class on $\mathcal{X}$. We say that $\mathcal{F}$ *shatters* a set of points $x_1, ..., x_k \in \mathcal{X}$ if for any choice of labels $\epsilon \in \{0,1\}^k$ there exists $f \in \mathcal{F}$ such that $f(x_i) = \epsilon_i$ for all $i \in [k]$. The VC dimension of $\mathcal{F}$ is the size of largest shattered set.

Accordingly, we aim to achieve learning guarantees in the abstention learning setting when the function class $\mathcal{F}$ has finite VC dimension, denoted by d. For adaptive adversaries, we will also consider classes with finite so-called reduction-dimension as defined in Yu & Blanchard (2026a).

Definition 2.2 (ℓ-reduction dimension (Yu & Blanchard, 2026a))**.** For $\ell \geq 1$, the class $\mathcal{F}$ has ℓ-reduction dimension $D_\ell \geq 1$ if, for every $n \geq D_\ell$ and every $x_1, \ldots, x_n \in \mathcal{X}$,

$$\left|\left\{(\mathcal{F} \cap A)|_{\{x_1,\ldots,x_n\}} : A \subseteq \mathcal{X} \times \{0,1\},\ |A| \leq \ell\right\}\right| \leq n^{D_\ell},$$

where $\mathcal{F} \cap A := \{f \in \mathcal{F} : f(x) = y \text{ for every } (x,y) \in A\}$.

Learning protocol. Before the interaction begins, an adversary selects a clean instance distribution μ on $\mathcal{X}$ together with an unknown target $f^\star \in \mathcal{F}$. Then, at each round of interaction, the adversary chooses whether to inject a corrupted instance or draw a clean sample. The learner then sees the resulting instance and may choose to predict a label or abstain. Formally, at each round $t \in [T]$:

1. The adversary decides whether to corrupt the round or not by selecting $c_t \in \{0,1\}$.
2. If $c_t = 0$, the round is *clean* and the new instance is a fresh draw from the clean distribution $x_t \sim \mu$. If $c_t = 1$, the round is *corrupted* and the adversary chooses $x_t \in \mathcal{X}$ arbitrarily.
3. After observing x_t, the learner selects $\widehat{y}_t \in \{0,1,\perp\}$, where $\perp$ denotes an abstention.
4. The true label is $y_t = f^\star(x_t)$, and the feedback rule determines whether the learner observes it.

Importantly, the learner never observes c_t, i.e., whether the round is clean or corrupted. Despite arbitrary corruptions in instances, labels remain realizable and consistent with the fixed target $f^\star$.

The goal of the learner is to make few prediction mistakes while not abstaining on clean rounds, that is, mistakes are counted on every round, whereas abstentions are charged only on clean rounds. Formally, the misclassification and abstention errors are respectively defined as

$$\mathsf{MisErr} = \sum_{t=1}^{T} \mathbf{1}\{\widehat{y}_t \in \{0,1\},\ \widehat{y}_t \neq y_t\} \qquad \text{and} \qquad \mathsf{AbsErr} = \sum_{t=1}^{T} \mathbf{1}\{c_t = 0,\ \widehat{y}_t = \perp\}.$$

Full vs. censored feedback. We consider two different feedback mechanisms. Under *full feedback*, y_t is revealed after every round including those on which the learner abstains. In contrast, under *censored feedback*, y_t is revealed only when a prediction is made: $\widehat{y}_t \in \{0,1\}$.

Oblivious vs. adaptive adversaries. We also distinguish between two adversarial models. An *oblivious* adversary must fix prior to the interaction their corruption rounds $(c_t)_{t=1}^T$ as well as every corrupted instance $(x_t)_{t:c_t=1}$ (potentially using randomness). An *adaptive* adversary may instead be adaptive to the history and depend on all previous learner decisions as well as clean samples on previous clean rounds. We remark that on round t it must nevertheless commit to c_t before observing a possible fresh clean draw $x_t \sim \mu$.

Other notation. Write $[n] := \{1, \ldots, n\}$ and $H_n := \sum_{i=1}^n 1/i$. For a sequence $S = (z_1, \ldots, z_n)$ and $U \subseteq [n]$, let S_{-U} denote the sequence obtained by deleting the coordinates in U, with $S_{-i} := S_{-\{i\}}$; let $S \| z$ denote appending z to S. For labeled sequences, $S[i \leftarrow y]$ replaces the ith label by y, leaving all instances unchanged. For $V \subseteq \mathcal{F}$, $x \in \mathcal{X}$, and $y \in \{0,1\}$, define $V^{x \to y} := \{f \in V : f(x) = y\}$. For $A \subseteq \mathcal{X}$, write $f|_A$ for the restriction of f to A and $V|_A := \{f|_A : f \in V\}$ for the corresponding projection of V. Finally, $V \triangleright A$ means that V shatters A.

Algorithm 1 Known-distribution abstaining learner

1: **Require:** class $\mathcal{F}$, clean distribution μ, and $\beta > 0$.
2: Initialize $V_1 \leftarrow \mathcal{F}$.
3: **For** $t = 1, \ldots, T$:
4: observe x_t and set $P_b \leftarrow \Theta_\beta(V_t^{x_t \to b})$ for $b \in \{0, 1\}$;
5: if $\min\{P_0, P_1\} > \frac{3}{4}\Theta_\beta(V_t)$, output $\widehat{y}_t \leftarrow \perp$;
6: otherwise output $\widehat{y}_t \leftarrow \arg\max_{b \in \{0,1\}} P_b$, breaking ties deterministically;
7: observe y_t; if $\widehat{y}_t \notin \{y_t, \perp\}$, update $V \leftarrow V^{x_t \to y_t}$ ▷ *update version space upon mistakes*

3 Known Clean Distribution

We begin with the setting in which the clean distribution μ is known and the adversary can be adaptive. A natural complexity measure in this setting could be the shattering probabilities:

Definition 3.1 (Shattering probabilities)**.** For $V \subseteq \mathcal{F}$, let $X_1, \ldots, X_k$ be i.i.d. from μ and define

$$\rho_k(V) := \mathbb{P}_{X_1,\ldots,X_k \sim \mu}\big(V \triangleright (X_1, \ldots, X_k)\big).$$

Set $\rho_0(V) := \mathbf{1}\{V \neq \varnothing\}$.

Thus $\rho_k(V)$ measures the ambiguity remaining in V at sample size k. It vanishes for $k > d$. The learner of Goel et al. (2023) follows one such probability at a time. It starts at level d, and a mistake reduces the current level-k shattering probability by a constant factor. Once that probability becomes smaller than a threshold T^{-k}, the learner moves to the next level $k - 1$. This takes $O(\log(1/\alpha_k)) = O(k \log T)$ mistakes at level k; summing over all levels gives $O(d^2 \log T)$ mistakes.

Our learner avoids these restarts. It combines all shattering probabilities using geometric weights, into one potential.

Definition 3.2 (Weighted shattering potential)**.** For $\beta > 0$, define $\Theta_\beta(V) := \sum_{k=0}^{d} \beta^k \rho_k(V)$.

Note, that restricting V can only decrease the potential, and $\Theta_\beta(V) \geq 1$ whenever $V \neq \varnothing$. Since $\rho_k(V) \leq 1$, the potential lies in $[1, (1+\beta)^d]$ while the target remains in the version space.

Our learner uses this one potential Θ_β for all decisions. The following is a short description.

Algorithm description. At the beginning, initialize $V_1 = \mathcal{F}$. At the start of round t, let V_t denote the current version space. Given instance x_t, let $P_b := \Theta_\beta(V_t^{x_t \to b})$ for $b \in \{0, 1\}$. The learner abstains if $\min\{P_0, P_1\} > \frac{3}{4}\Theta_\beta(V_t)$, and otherwise predicts the label b with larger P_b. After observing y_t, it updates $V_{t+1} = V_t^{x_t \to y_t}$ if it predicts and makes a mistake. Otherwise, it keeps $V_{t+1} = V_t$. The learner is formally summarized in Algorithm 1 and enjoys the following guarantee:

Theorem 3.1 (Known distribution)**.** *For $\beta > 0$ and every adaptive adversary, Algorithm 1 satisfies*

$$\mathsf{MisErr} \leq \frac{d \log(1+\beta)}{\log(4/3)} \qquad \textit{and} \qquad \mathbb{E}[\mathsf{AbsErr}] \leq \frac{2T}{\beta}.$$

Proof sketch. The potential is nonincreasing under restriction and lies in $[1, (1+\beta)^d]$ as the target f^* remains in V_t. Whenever the algorithm makes a mistake, it contracts the potential by a factor of at most $3/4$, giving the mistake bound. For abstentions, given V_t, let $\sigma_k(V_t, x)$ be the probability that V_t shatters x together with k fresh draws from μ. A one-point decomposition (Lemma B.1) gives

$$\Theta_\beta(V_t^{x \to 0}) + \Theta_\beta(V_t^{x \to 1}) \leq \Theta_\beta(V_t) + \sum_{k=0}^{d-1} \beta^k \sigma_k(V_t, x).$$

By construction, when the learner abstains, the left-hand side must be at least $\frac{3}{2}\Theta_\beta(V_t)$. Hence, an abstention forces the sum $\sum_{k=0}^{d-1} \beta^k \sigma_k(V_t, x)$ to exceed $\Theta_\beta(V_t)/2$. For a fresh $x \sim \mu$, $\mathbb{E}_x[\sigma_k(V_t, x)] = \rho_{k+1}(V_t)$, so the expectation of this sum is at most $\Theta_\beta(V_t)/\beta$. Markov's inequality then bounds the conditional clean-abstention probability by $2/\beta$; summing over rounds proves the claim. A formal proof appears in Appendix B.

Algorithm 2 Weighted-history abstaining learner

Require: class $\mathcal{F}$ and parameter $\beta > 0$.
Initialize $S_0 \leftarrow ()$ and $\lambda_{u,t} \leftarrow \beta(1 + (t-u)\frac{\beta}{d})$ for all $1 \le u \le t \le T$.
For $t = 1, \dots, T$:
  observe x_t and set $P_b \leftarrow \widehat{\Theta}_\lambda(S_{t-1} \| (x_t, b))$ for $b \in \{0, 1\}$;
  if $\min\{P_0, P_1\} > \frac{3}{4}\widehat{\Theta}_\lambda(S_{t-1})$, output $\widehat{y}_t \leftarrow \perp$;
  otherwise output $\widehat{y}_t \leftarrow \arg\max_{b \in \{0,1\}} P_b$, breaking ties deterministically;
  observe y_t and update $S_t \leftarrow S_{t-1} \| (x_t, y_t)$.

4 Unknown Distribution with Oblivious Corruptions

For a labeled history $S = (x_u, y_u)_{u=1}^t$, the natural analogue of a fresh shattered sample is a set of coordinates whose labels remain free after all other observed labels have been enforced. Accordingly, we say a set $U \subseteq [n]$ is *shattered in* S if, for every $b \in \{0,1\}^U$, there is an $f \in \mathcal{F}$ such that

$$f(x_j) = y_j \quad (j \notin U), \qquad f(x_j) = b_j \quad (j \in U).$$

Let $\mathrm{Sh}(S)$ be the family of such subsets. Membership $U \in \mathrm{Sh}(S)$ means that the observed labels remain enforced at every coordinate outside U, while the labels on U may vary arbitrarily.

This definition suggests a direct empirical replacement for ρ_k using S by counting the number of shattered k-sets in S and applying normalization. However, such a normalized empirical analogue of ρ_k is unsuitable because by injecting duplicates, the adversary can arbitrarily alter its denominator without changing the numerator — in general, normalized leave-out frequencies need not reflect the shattering behavior of clean draws under μ. We therefore retain only the combinatorial witnesses but discard the normalization and the probability interpretation.

Definition 4.1 (Weighted history potential)**.** For $\beta > 0$, define $\widehat{\Theta}_\beta(S) := \sum_{U \in \mathrm{Sh}(S)} \beta^{|U|}$.

If S is realizable, then $\varnothing \in \mathrm{Sh}(S)$ and every $U \in \mathrm{Sh}(S)$ has $|U| \le d$. The simple potential $\widehat{\Theta}_\beta(S)$ assigns equal weight to all witnesses of the same size. We next define its age-weighted variant:

Definition 4.2 (Age-weighted history potential)**.** For $\beta > 0$ and a history S of length t, define $\widehat{\Theta}_\lambda(S) := \sum_{U \in \mathrm{Sh}(S)} \prod_{u \in U} \lambda_{u,t}$, where $\lambda_{u,t} := \beta(1 + (t-u)\frac{\beta}{d})$.

The complete algorithm based on age-weighted history potential is detailed in Algorithm 2. As a remark, if we replace $\widehat{\Theta}_\lambda(S)$ by the simpler potential $\widehat{\Theta}_\beta(S)$ in Algorithm 2, we lose a $\log T$ factor in the abstention error (see Theorem C.1). In comparison, the age-dependent potential places more mass on older datapoints and since the learner aims to minimize the potential, this incentivizes the learner to resolve older ambiguities first. For intuition, resolving older datapoints ensures that if the learner made abstentions, it was because the adversary significantly modified the distributions of observed instances—in turn, this prevents the abstention error from growing logarithmically due to old observations. Altogether, Algorithm 2 enjoys the following guarantee.

Theorem 4.1 (Unknown distribution, oblivious corruption)**.** *There is a universal constant $C > 0$ such that for every $\beta \in (0, 1/4]$, Algorithm 2 satisfies, against any oblivious adversary,*

$$\mathsf{MisErr} \le \frac{2\beta}{\log(4/3) + 2\beta} T, \qquad \mathbb{E}[\mathsf{AbsErr}] \le C\frac{d}{\beta} + Cd \log\left(1 + \frac{\beta T}{d}\right).$$

Hence, the mistake rate is proportional to β, whereas the abstention error is of order d/β. For $T = \Omega(d)$, taking $\beta = \sqrt{d/T}$ results in a balanced $O(\sqrt{dT})$ rate for both MisErr and $\mathbb{E}[\mathsf{AbsErr}]$. Next, we provide some intuitions on how Algorithm 2 achieves its performance guarantee. For simplicity, we focus on Algorithm 2 with $\widehat{\Theta}_\lambda(S)$ replaced by its simplified version $\widehat{\Theta}_\beta(S)$.

First, notice that the history potential increases by at most a factor $1 + \beta$ when a new labeled example is appended. Indeed, a shattered set in the extended history either excludes the new coordinate, or includes it and becomes a shattered set of the old history when that coordinate is removed.

Lemma 4.1 (One-step growth)**.** *For any labeled history S and example z, $\widehat{\Theta}_\beta(S\|z) \le (1+\beta)\widehat{\Theta}_\beta(S)$.*

A formal combinatorial proof of the growth lemma appears in Appendix C.1. Combining this growth bound with the algorithm's $3/4$-contraction on mistakes immediately gives the proposition below.

Proposition 4.1 (History-potential mistake bound)**.** *For every realizable interaction, Algorithm 2 with non-age-dependent potential satisfies* $\mathsf{MisErr} \le \frac{\log(1+\beta)}{\log(4(1+\beta)/3)}\, T$.

Next, we bound $\mathbb{E}[\mathsf{AbsErr}]$ for Algorithm 2. The argument has three steps.

Step 1: replay one coordinate. Fix a realizable history $S = (z_1, \dots, z_n)$, delete coordinate i, and imagine presenting x_i after all the other examples. We recall the notation $S[i \leftarrow y]$ for the sequence obtained by replacing the i-th label with y. If the replayed point receives candidate label y, permutation invariance shows that the resulting branch potential is $P_i(y) := \widehat{\Theta}_\beta(S[i \leftarrow y])$. Thus $P_i(0)$ and $P_i(1)$ are exactly the two values that the learner would compare in this replay. We call i *bad in* S, or $i \in \mathrm{Bad}(S)$, if the learner would abstain assuming index i is displayed last:

$$\min\{P_i(0), P_i(1)\} > \tfrac{3}{4}\widehat{\Theta}_\beta(S_{-i}).$$

Step 2: only a few coordinates are bad. For each coordinate, let $W_i(S) := \sum_{U \in \mathrm{Sh}(S), i \in U} \beta^{|U|-1}$, which quantifies how much of the empirical potential can be charged to coordinate i. Each shattered set U contributes once to $W_i(S)$ for every $i \in U$. Consequently,

$$\sum_{i=1}^{n} W_i(S) = \sum_{U \in \mathrm{Sh}(S)} |U|\beta^{|U|-1} \le \frac{d}{\beta}\widehat{\Theta}_\beta(S), \tag{1}$$

where the last inequality uses $|U| \le d$. The next lemma relates the two branch potentials to $W_i(S)$.

Lemma 4.2 (Replay-branch decomposition)**.** *For every realizable history S and every $i \in [n]$,*

$$P_i(0) + P_i(1) \le \widehat{\Theta}_\beta(S_{-i}) + (1 + 2\beta)W_i(S).$$

The proposition then sums those local weights to show that there are few bad positions.

Proposition 4.2 (Few bad indices)**.** *For every realizable history S, $|\mathrm{Bad}(S)| \le 2d(1+2\beta)(1+\beta)/\beta$.*

Indeed, if $i \in \mathrm{Bad}(S)$, the replay-branch decomposition and the one-step growth lemma would force $W_i(S)$ to carry at least a fixed fraction of $\widehat{\Theta}_\beta(S)$. The total-weight bound in (1) then gives Proposition 4.2. This deterministic counting argument is proved in Appendix C.1.

Step 3: use exchangeability to identify the current clean coordinate. Condition on the oblivious adversary's seed, which fixes the injected examples and their locations. To bound the probability of abstaining on the i-th clean round, we may further condition on the first i clean sample pool, but not their assignment to clean rounds. Since clean samples are exchangeable, the probability of abstaining on the i-th clean round becomes the proportion of bad coordinates within the i clean coordinate pool. This is at most $|\mathrm{Bad}(S)|/i$. Combining this observation with Proposition 4.2 gives the next proposition. The formal replay and conditioning argument appears in Appendix C.1.

Proposition 4.3 (Clean-round exchangeability)**.** *For the ith clean round s_i of any fixed oblivious corruption schedule,* $\mathbb{P}(\widehat{y}_{s_i} = \perp) \le \frac{2d(1+2\beta)(1+\beta)}{\beta i} = O(\frac{d}{\beta i})$.

Finally, summing this abstention probability over all clean rounds i gives the desired abstention error $\mathbb{E}[\mathsf{AbsErr}] = O(\frac{d}{\beta}\log T)$. This, together with the deterministic $O(\beta T)$ mistake bound in Proposition 4.1, provides the guarantee in Theorem 4.1 up to a logarithm factor.

Importantly, we show that the dependence on d and T in Theorem 4.1 cannot be improved.

Theorem 4.2 (Oblivious lower bound)**.** *Fix $d \ge 1$, $T \ge 32$ and an integer abstention budget $1 \le A \le T/32$. There is a class of VC dimension d such that every full-feedback learner faces some oblivious adversary, for which either $\mathbb{E}[\mathsf{AbsErr}] > A$ or $\mathbb{E}[\mathsf{MisErr}] \ge \frac{1}{1024}\min\left\{\frac{dT}{A}, T\right\}$.*

Compared with the lower bound tradeoffs in Edelman & Goel (2026) or Yu & Blanchard (2026b), Theorem 4.2 tightens the dependence in d, which matches our upper bound in Theorem 4.1. This characterizes the mistake-abstention tradeoff landscape for the oblivious full-feedback setting.

4.1 Handling Censored Feedback for Oblivious Adversaries

We can also extend the strategy described above to the censored model, where the label is revealed only when the learner makes a prediction. A tempting direct approach is to store every label that happens to be revealed and run the age-weighted history learner on that labeled subsequence. However, this fails because censored labels are retained precisely on rounds when the learner chooses to predict, making the retained clean instances no longer exchangeable.

We instead retain every observed instance without its label, forming an unlabeled history $U = (x_1, \dots, x_t)$. Revealed labels are used only after mistakes, when they restrict the current version space which will be denoted by V_t. Thus censorship does not select which instances enter the history; it only selects a sequence of truthful version-space restrictions. Similar to the previous history potentials—Definitions 4.1 and 4.2—we then use as potential the partition function of subsets of U that are shattered by the current version space V:

Definition 4.3 (Unlabeled-history potential)**.** For $V \subseteq \mathcal{F}$ and unlabeled history $U = (x_1, \dots, x_t)$, define the age-weighted unlabeled-history potential via $\widehat{\Theta}^u_t(V; U) := \sum_{I \subseteq [t]} \mathbf{1}\{V \triangleright U_I\} \prod_{i \in I} \lambda_{i,t}$. For the empty history, set $\widehat{\Theta}^u_0(V; ()) = \mathbf{1}\{V \neq \emptyset\}$.

The learner uses the potential to make predictions or abstain as in the previous algorithms. However, the learner only updates the version space after each mistake, and deliberately ignores labels revealed after correct predictions. Also, for uniform concentration purposes, we periodically restart the version space $V \leftarrow \mathcal{F}$ every $W_\beta := \lceil d \log(T+1)/\beta \rceil$ iterations, forming epochs of length W_β. This will in turn ensure that the potential number of version spaces arising throughout the procedure can be controlled. The algorithm is summarized in Algorithm 5 and has the same guarantee as in Theorem 4.1 for oblivious full-feedback, up to logarithmic additive factors (see Theorem D.1 in Appendix D).

5 Learning Against Adaptive Adversaries

In this section, we handle the more challenging case of adaptive adversaries. Recall that the weighted-history learner in Section 4 fixes the injected sequence and replays the clean examples in a random order. Against an adaptive adversary, later injected examples may depend on earlier clean draws, so replaying those draws also changes the injected history. The adaptive proof must therefore isolate the i.i.d. clean sequence rather than replay it through the contaminated history.

We build on the adaptive full-feedback reduction introduced by Yu & Blanchard (2026b). Their reduction has three parts. First, it analyzes an expert whose stored examples are all clean; these examples form its *clean bank*. Second, it uses reduction dimension, which bounds the number of constrained subclasses that can appear on a fixed sample, to obtain one guarantee for every version space the expert may reach. Third, since the learner does not know which rounds are clean, it aggregates over candidate banks and compares itself with the actual clean-bank expert. We improve these three parts using a normalized-mean expert, a leave-one-out bound together with a uniform concentration argument to compress the expert, and direct multiplicative-weights aggregation.

For comparison, their theorem for adaptive adversary gives, for $\alpha \in [0, 1/3]$,

$$\mathsf{MisErr} \lesssim T^{3\alpha}, \qquad \mathbb{E}[\mathsf{AbsErr}] \lesssim d^2 (D \log D + \log T)^{2/3} \log(T)\, T^{1-\alpha},$$

where D is the $O(d^2 \log T)$-reduction dimension of $\mathcal{F}$. We obtain the following result.

Theorem 5.1 (Adaptive adversaries via direct aggregation)**.** *For every $T \geq 1$, $\alpha \in [0, 1]$, the following holds. Let $\ell_T := \lceil d \log(T+1)/\log(16/15) \rceil + 1$ and suppose that $\mathcal{F}$ has ℓ_T-reduction dimension $D_T < \infty$. Then there is a learner that satisfies against any adaptive adversary with censored feedback, without knowledge of μ:*

$$\mathsf{MisErr} \lesssim T^{\alpha} + d^2 D_T \log^4(2T), \qquad \mathbb{E}[\mathsf{AbsErr}] \lesssim d^2 D_T T^{1-\alpha} \log^4(2T).$$

This matches the tradeoff from the oblivious case up to factors in $d, \log T$, and D_T. Hence, the dependency of the tradeoff in the horizon T is tight up to logarithmic factors. We prove this result in Appendix E and give the main ingredients below for the full-feedback case.

A compressed shattering-witness expert. We start by constructing a clean-bank expert. Partition the clean bank into m blocks $S_1, \dots, S_m$ of size $N \geq d$. For a block $S = (z_1, \dots, z_N) \in$

Algorithm 3 Reference expert

Require: Class $\mathcal{F}$, start time t_0, witness-index lists $(\mathcal{Z}_r)_{0\le r\le M_T}$ in $[t_0-1]$
Abstain $\widehat{y}_t \leftarrow \perp$ for all $t < t_0$ then initialize $V \leftarrow \mathcal{F}$, $r \leftarrow 0$
For each round $t \ge t_0$:
  observe x_t and compute $\widehat{a}_b(x_t; V)$ for $b \in \{0,1\}$ as in (2) using witness list $\mathcal{Z}_r$
  if $\min_{b\in\{0,1\}} \widehat{a}_b(x_t) > \frac{3}{4}$, abstain $\widehat{y}_t \leftarrow \perp$
  otherwise, output $\widehat{y}_t \leftarrow \arg\max_{b\in\{0,1\}} \widehat{a}_b(x_t)$, breaking ties deterministically
  observe y_t, and if $\widehat{y}_t = 1 - y_t$, update $V \leftarrow V^{x_t \to y_t}$ and $r \leftarrow r+1$

$\mathcal{X}^N$ and any class $V \subseteq \mathcal{F}$, we define the symmetric shattered-subset statistic $\widehat{\rho}_k^S(V) := \binom{N}{k}^{-1} \sum_{I\subseteq[N],|I|=k} \mathbf{1}\{V \triangleright (x_i)_{i\in I}\}$ and set $\widehat{\rho}_0^S(V) := \mathbf{1}\{V \neq \varnothing\}$. As in the weighted-potential learner from the known-distribution case, we use $\beta = T$ and combine the statistics at different levels into the single block potential $\Theta^S(V) := \sum_{k=0}^{d} T^k \widehat{\rho}_k^S(V)$.

For any considered class V, the expert estimates the weighted potential of version spaces $V^{x\to b} = \{f \in V : f(x) = b\}$ through a normalized combination of block estimates as defined below, which resembles the idea of median of block estimates from Yu & Blanchard (2026a). We define the *score*

$$a_b(x; V) := \frac{1}{m} \sum_{j=1}^{m} \frac{\Theta^{S_j}(V^{x\to b})}{\Theta^{S_j}(V)}, \qquad b \in \{0,1\}.$$

The normalization is performed separately within each block: Each averaging term is simply the ratio of potentials between the prior class V and updated version space $V^{x\to b}$ of the corresponding block. If we directly use the scores defined above, the learner therefore has to consider $T^{O(Nm)}$ experts, corresponding to all possible choices of m blocks. However, the learner also aims to use as few experts as possible (see aggregation strategy below). To this end, we compress the description of scores using a uniform-concentration argument as follows. Choose a block j uniformly, then a subset $I \subseteq [N]$, $|I| \le d$, with probability

$$\frac{T^{|I|}\mathbf{1}\{V \triangleright S_{j,I}\}}{\binom{N}{|I|}\Theta^{S_j}(V)}.$$

Let ν_V denote the resulting distribution of the witness tuple $Z = S_{j,I}$. Crucially, since a tuple shattered by $V^{x\to b}$ is also shattered by V, we have

$$a_b(x; V) = \mathbb{P}_{Z\sim\nu_V}[V^{x\to b} \triangleright Z].$$

We then estimate score $a_b(x)$ by an empirical mean of samples $Z \sim \nu_V$. The reduction dimension is used to achieve uniform concentration of these empirical means. For $s := \lceil C_s D_T \log(2dD_T)\rceil$ it turns out that there exist witnesses $Z_1, \ldots, Z_s$ for some universal constant $C_s > 0$, satisfying

$$\sup_{x,b} |\widehat{a}_b(x; V) - a_b(x; V)| \le \frac{1}{64}, \quad \text{where} \quad \widehat{a}_b(x; V) := \frac{1}{s}\sum_{u=1}^{s} \mathbf{1}[V^{x\to b} \triangleright Z_u], \tag{2}$$

whenever V is defined by at most $\ell_T - 1$ constraints (see Lemma E.2 for a proof). Importantly, each witness uses at most d bank instances; hence the total number of bank instances required to compute the approximate score $\widehat{a}_b(x)$ is at most ds, independent of N and m.

We then construct the compressed shattering witness expert using approximate scores. Similarly to previous algorithms, the expert abstains if both approximate scores exceed $\frac{3}{4}$, using valid lists of witnesses satisfying (2); otherwise it predicts a label of maximum approximate score. It restricts the version space V to the true-label branch $V^{x_t\to y_t}$ only when it makes a mistake. Whenever the version space V is updated, a new witness list $\mathcal{Z} = (Z_1, \ldots, Z_s)$ is used—this is required because the witness distribution ν_V is dependent on the version space V. To run this procedure, the compressed expert needs Nm clean rounds. We may consider that it abstains until these clean rounds have passed: starting from the following round t_0, it uses a sequence of witness lists $(\mathcal{Z}_r)_{0\le r<\ell_T}$ constructed from the clean bank. We summarize this expert in Algorithm 3.

Algorithm 4 Direct aggregation of candidate-bank experts

Require: expert family $\mathcal{C}$, margin $\kappa \in (0, 1/2)$, and update rate $\gamma \in (0, 1]$.
Initialize $w_{c,1} \leftarrow 1/|\mathcal{C}|$ for every $c \in \mathcal{C}$.
For $t = 1, \dots, T$:
  observe x_t, obtain each expert's recommendation $\widehat{y}_{c,t} \in \{0, 1, \perp\}$, and set

$$\Phi_t \leftarrow \sum_{c \in \mathcal{C}} w_{c,t}, \qquad P_b \leftarrow \sum_{c \in \mathcal{C}} w_{c,t} \mathbf{1}\{\widehat{y}_{c,t} \in \{b, \perp\}\}, \quad b \in \{0, 1\}$$

  If $\min\{P_0, P_1\} > (1 - \kappa)\Phi_t$ then
    output $\widehat{y}_t \leftarrow \perp$ and update $w_{c,t+1} \leftarrow w_{c,t}(1 + \gamma)^{\mathbf{1}[\widehat{y}_{c,t} \neq \perp]}$ for each $c \in \mathcal{C}$
  Else
    output $\widehat{y}_t \leftarrow \arg\max_{b \in \{0,1\}} P_b$, breaking ties deterministically
    observe y_t and update $w_{c,t+1} \leftarrow w_{c,t} 2^{-\mathbf{1}[\widehat{y}_{c,t} = 1 - y_t]}$ for each $c \in \mathcal{C}$

With a similar mistake-contraction argument as before, we show that the expert makes at most $M_T = \ell_T - 1 \asymp d \log T$ mistakes. Meanwhile, an abstention forces more than $m/6$ blocks to have, roughly speaking, both branch potentials greater than $5/8$ of their current block potential (Lemma E.3). Using a replay argument, we then show that for a fixed clean bank of size Nm, the abstention error rate is essentially $O(d/N)$ for each new clean round (Lemma E.5). Altogether, the abstention error bound for this expert is essentially $Nm + O(dT/N)$ (Proposition E.1). We can further remove this Nm term using a recursive phase strategy—in each phase, the strategy uses clean rounds from previous phases as the clean bank. Details are deferred to Algorithm 6 in Appendix E).

Direct Aggregation over candidate experts. So far, we have obtained a good reference expert with guaranteed mistake and abstention error bounds. Since the correct start time t_0 and witness-index lists $(\mathcal{Z}_r)_{0 \le r \le M_T}$ are unknown, we consider all experts arising from any possible choice for these quantities. Let $\mathcal{C}$ be the resulting family: their number is $|\mathcal{C}| \le (T + 1)^{ds(M_T+1)+1}$, which, thanks to the score compression, greatly improves over the $T^{O(Nm)}$ expert count in Yu & Blanchard (2026b).

It remains to describe how the learner can achieve comparable performance to the reference expert. Yu and Blanchard handle this final step with their layered BOOSTING procedure. If one expert has at most M mistakes, their guarantee adds $O(\varepsilon T)$ abstentions at the cost of $O((M + \log T)(\log L)^2/\varepsilon)$ mistakes, where L is the number of candidate-bank experts. Using a single weight potential, we give a sharper comparison: it adds the same $O(\varepsilon T)$ abstentions with only $O((M + \log L)/\varepsilon)$ mistakes. This is our replacement for the final step of their reduction. A formal guarantee appears in Appendix E.

Our proposed aggregate is summarized in Algorithm 4. It assigns a weight to each of the L experts. On each round, the score P_b is the total weight of experts that do not oppose label b: those that predict b or abstain. The learner abstains precisely when less than a κ fraction of the total weight predicts each label, then rewards the experts that did predict by multiplying their weight by $1 + \gamma$. Otherwise, the learner predicts the branch with larger weight, and halves the weights of experts making a mistake. Consequently, a learner mistake contracts a fixed fraction of the total weight. Conversely, whenever the learner abstains while the reference expert predicts, the update increases that expert's relative weight. Comparing total weight with weight of the reference expert gives the following guarantee.

Proposition 5.1 (Direct aggregation)**.** *Consider L experts. Let $M \ge 0$ and $R := 1 + \log L + (\log 2)M$. Fix $\kappa \in (0, 1/2)$, suppose $R \le 2\kappa T$, and that there is an expert which makes at most M mistakes. Run Algorithm 4 with $\gamma = R/(2\kappa T)$. Then, we have*

$$\mathsf{MisErr} \le \frac{4R}{\kappa}.$$

Moreover, simultaneously for every expert with at most M mistakes and every $U \subseteq [T]$, the learner abstains on at most $8\kappa T$ more rounds of U than that expert.

The weight comparison argument proving this proposition appears in Appendix E.3.

AI USE STATEMENT

The authors used GPT 5.6 Pro and GPT 6 Astra during the later stages of the project to improve parts of the adaptive full-feedback and oblivious censored-feedback proofs. The main results and initial proofs were developed by the authors prior to this use. The authors have reviewed and verified all AI-assisted arguments and writing and take full responsibility for the final contents of the paper.

REFERENCES

Gautam Chandrasekaran, Adam R. Klivans, Vasilis Kontonis, Konstantinos Stavropoulos, and Arsen Vasilyan. Efficient discrepancy testing for learning with distribution shift. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f8238b414fce7d91fedb896253545300-Abstract-Conference.html.

Gautam Chandrasekaran, Adam R. Klivans, Lin Lin Lee, and Konstantinos Stavropoulos. Learning neural networks with distribution shift: Efficiently certifiable guarantees. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94d64b1d0647329bd3de5d7cfbbdfdb1-Abstract-Conference.html.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1059–1067. PMLR, 2018. URL https://proceedings.mlr.press/v80/cortes18a.html.

Ezra Edelman and Surbhi Goel. Reliable abstention under adversarial injections: Tight lower bounds and new upper bounds, 2026. URL https://arxiv.org/abs/2602.20111v1. Version 1, 23 February 2026.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Surbhi Goel, Steve Hanneke, Shay Moran, and Abhishek Shetty. Adversarial resilience in sequential prediction via abstention. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8027–8047, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1967f962c7c2083618236d80eeb9d1ac-Abstract-Conference.html.

Shafi Goldwasser, Adam Tauman Kalai, Yael Tauman Kalai, and Omar Montasser. Identifying unpredictable test examples with worst-case guarantees. In *2020 IEEE Information Theory and Applications Workshop (ITA)*, pp. 1–14. IEEE, 2020. doi: 10.1109/ITA50056.2020.9244996.

Adam Tauman Kalai and Varun Kanade. Towards optimally abstaining from prediction with OOD test examples. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12774–12785, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6a26c75d6a576c94654bfc4dda548c72-Abstract.html.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Learning intersections of halfspaces with distribution shift: Improved algorithms and SQ lower bounds. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2944–2978. PMLR, 2024a. URL https://proceedings.mlr.press/v247/klivans24b.html.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2887–2943. PMLR, 2024b.

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

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3030–3048. PMLR, 2020. URL https://proceedings.mlr.press/v125/neu20a.html.

Shyamal Patel, Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Equivalence of coarse and fine-grained models for learning with distribution shift. In *The Thirty Ninth Annual Conference on Learning Theory*, pp. 4022–4049. PMLR, 2026.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

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

Jialin Yu and Moïse Blanchard. Distribution-free sequential prediction with abstentions. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6976–7011. PMLR, 2026a. URL https://proceedings.mlr.press/v336/yu26a.html.

Jialin Yu and Moïse Blanchard. Distribution-free sequential prediction with abstentions, 2026b. URL https://arxiv.org/abs/2602.17918v2. Version 2, 17 March 2026.

Chicheng Zhang and Kamalika Chaudhuri. The extended Littlestone's dimension for learning with mistakes and abstentions. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1584–1616. PMLR, 2016. URL https://proceedings.mlr.press/v49/zhang16a.html.

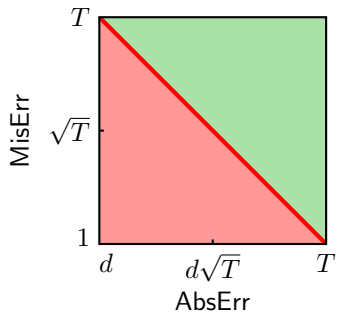

Figure 1: This is a schematic summary of the tradeoffs between AbsErr and MisErr that we achieve in this paper. The plot is exact for VC-dimension-d classes under oblivious injections (with full or censored feedback) when $d < \sqrt{T}$. Axes are log-scale in T. The red region is ruled out by Theorem 4.2. The green region is achieved by Theorems 4.1 and D.1 under full and censored feedback, respectively, giving a complete characterization. For adaptive injections, the lower bound still applies, and we achieve a similar upper bound in Theorem E.1, tight up to $\mathrm{poly}(dD \log T)$ factors, where D is the $O(d \log T)$-reduction dimension of the class.

A Additional Related Work

Transductive abstention under distribution shift. The offline analogue of the online adversarial model was introduced by Goldwasser et al. (2020). In their PQ learning model, the learner is given an i.i.d. labeled training sample and an unlabeled test set, some of which may be inserted/modified by an adversary. The goal is to output a predictor and selector such that the error on the test and false abstentions on the training are simultaneously minimized. Kalai & Kanade (2021) also assume transductive access and study the optimal tradeoff when every abstention has a fixed cost. Klivans et al. (2024b) introduced the decision version of this problem called Testable Distribution Shift (TDS) learning where the learner may abstain on the entire test set with the motivation that the decision problem would allow for computationally efficient algorithms, with subsequent advances for intersections of halfspaces, discrepancy testing, and neural-network regression (Klivans et al., 2024a; Chandrasekaran et al., 2024; 2025). Surprisingly, recent work by Patel et al. (2026) showed that these two models of learning are equivalent in the distribution-free setting.

Other abstention objectives. Classical selective classification studies the error–coverage tradeoff under a single test distribution (Chow, 1970; El-Yaniv & Wiener, 2010). Online formulations either charge every abstention or combine mistakes and abstentions in one regret objective (Zhang & Chaudhuri, 2016; Cortes et al., 2018; Neu & Zhivotovskiy, 2020). The injection model is different in two ways: prediction errors are counted on both clean and injected rounds, and only clean abstentions are charged. This hidden distinction between the two sources is what allows the statistical clean subsequence to help without revealing which inputs came from it.

B Algorithm for the Known Clean Distribution and its Proof

We give the formal proof of Theorem 3.1. Throughout, let $\beta > 0$, $V \subseteq \mathcal{F}$, and $d = \mathrm{VC}(\mathcal{F})$.

The proof has two parts. Every mistake follows the smaller branch and contracts the potential, which proves the realized-interaction mistake bound. An abstention requires both branches to be large; the one-point decomposition shows that this is rare when the current point is a clean draw.

B.1 Mistakes contract the potential

Proposition B.1 (Mistake bound for every realized interaction)**.** *On every realized interaction, Algorithm 1 satisfies*

$$\mathsf{MisErr} \le \frac{d \log(1+\beta)}{\log(4/3)}.$$

Proof. Set $\Phi_t := \Theta_\beta(V_{t+1})$ for $t = 0, \dots, T$. Realizability ensures that the target $f^\star$ remains in the version space, so $\Phi_t \geq 1$, while $\Phi_0 \leq (1+\beta)^d$. Monotonicity gives $\Phi_t \leq \Phi_{t-1}$ on every round. If a mistake occurs, the update follows the branch not selected by the learner, which has the smaller potential. Since the learner predicted rather than abstained,

$$\Phi_t = \min\{P_0, P_1\} \leq \tfrac{3}{4}\Phi_{t-1}.$$

Writing $M = \mathsf{MisErr}$ and iterating yields

$$1 \leq \Phi_T \leq \left(\frac{3}{4}\right)^M (1+\beta)^d.$$

Taking logarithms proves the claim. □

B.2 A CLEAN POINT RARELY TRIGGERS ABSTENTION

Potential contraction alone cannot bound abstentions, because an abstention need not change the version space. Instead, we show that both branches can be large only when the current point extends unusually many shattered samples. We first quantify this extension event. For $k \geq 1$, define

$$\sigma_k(V, x) := \mathbb{P}_{X_1,\dots,X_k \sim \mu}\big(V \triangleright (x, X_1, \dots, X_k)\big).$$

Set $\sigma_0(V, x) := \mathbf{1}\{V^{x\to 0} \neq \varnothing,\ V^{x\to 1} \neq \varnothing\}$. The tuple $(x, X_1, \dots, X_d)$ cannot be shattered: either it contains a repeated point, or it contains $d+1$ distinct points, which is impossible for a class of VC dimension d. Hence $\sigma_d(V, x) = 0$. Moreover, for $0 \leq k \leq d-1$,

$$\mathbb{E}_{x\sim\mu}[\sigma_k(V, x)] = \rho_{k+1}(V). \tag{3}$$

Thus, if the anchor x is drawn independently from μ, the mean probability of an anchored k-extension is exactly $\rho_{k+1}(V)$.

The next lemma is the weighted form of the fixed-level one-point inequality of Goel et al. (2023).

Lemma B.1 (One-point decomposition)**.** *For every $V \subseteq \mathcal{F}$ and $x \in \mathcal{X}$,*

$$\Theta_\beta(V^{x\to 0}) + \Theta_\beta(V^{x\to 1}) \leq \Theta_\beta(V) + \sum_{k=0}^{d-1} \beta^k \sigma_k(V, x).$$

Proof. For $A = (X_1, \dots, X_k)$, pointwise

$$\mathbf{1}\{V^{x\to 0} \triangleright A\} + \mathbf{1}\{V^{x\to 1} \triangleright A\} \leq \mathbf{1}\{V \triangleright A\} + \mathbf{1}\{V \triangleright (x, A)\}.$$

This is the fixed-level indicator inequality of Goel et al. (2023). Taking expectations gives

$$\rho_k(V^{x\to 0}) + \rho_k(V^{x\to 1}) \leq \rho_k(V) + \sigma_k(V, x).$$

Multiplying by β^k, summing over $k \leq d$, and using $\sigma_d = 0$ proves the claim. □

If the learner abstains, the two branch potentials sum to more than $3\Theta_\beta(V)/2$, so the decomposition forces the anchored extension sum to exceed $\Theta_\beta(V)/2$. For a clean x, the expected anchored extension sum is at most $\Theta_\beta(V)/\beta$. Markov's inequality therefore gives the following conditional bound.

Proposition B.2 (Conditional abstention bound)**.** *Conditional on the pre-round history and on round t being clean,*

$$\mathbb{P}(\widehat{y}_t = \perp \mid \mathcal{H}_{t^-}, c_t = 0) \leq \frac{2}{\beta}.$$

Proof. Condition on the realized pre-draw history and write $V = V_t$. On a clean round, $x_t \sim \mu$. If the learner abstains, both restrictions have potential greater than $3\Theta_\beta(V)/4$. Hence, Lemma B.1 implies

$$\sum_{k=0}^{d-1} \beta^k \sigma_k(V, x_t) > \tfrac{1}{2}\Theta_\beta(V).$$

By Equation (3),

$$\mathbb{E}_{x_t\sim\mu}\left[\sum_{k=0}^{d-1}\beta^k\sigma_k(V,x_t)\right]=\sum_{k=0}^{d-1}\beta^k\rho_{k+1}(V)=\frac{1}{\beta}\sum_{j=1}^{d}\beta^j\rho_j(V)\le\frac{1}{\beta}\Theta_\beta(V).$$

The result follows from Markov's inequality. □

Proof of Theorem 3.1. The mistake claim is Proposition B.1. On a corrupted round the charged abstention indicator is zero. On a clean round, Proposition B.2 bounds its conditional expectation by $2/\beta$. Summing over $t\in[T]$ proves the expectation bound. □

C Unknown-Distribution Proofs

We prove the upper and lower bounds for an unknown clean distribution under oblivious adversaries and full feedback. This section is organized as follows. In Appendix C.1 we prove the misclassification and abstention guarantee for the algorithm using the simpler weighted history potential (see Definition 4.1). We give a deterministic mistake guarantee, then a replace-one argument bounds the number of bad positions in any realized labeled sequence, where a position is bad if the same rule would abstain when that position is replayed last. Exchangeability turns this count into a clean-round abstention probability. Next, in Appendix C.2 we adapt the proof structure to the age-weighted history potential. The mistake guarantee extends immediately, while the exchangeability argument requires additional care to bound certain fractions of partition functions. Finally, in Appendix C.3 we prove the lower bound. To do so, we lift the one-dimensional construction of Edelman & Goel (2026) to VC dimension d by a direct sum and derive the resulting mistake–abstention tradeoff.

C.1 Upper bound for the Weighted History Potential

Proof of Lemma 4.1. Let S have length n, so the appended point has index $n+1$. Every $J\in\mathrm{Sh}(S\|z)$ falls into one of two cases. If $n+1\notin J$, then $J\in\mathrm{Sh}(S)$: deleting the new consistency constraint can only make shattering easier. If $n+1\in J$, set $I:=J\setminus\{n+1\}$. Then $I\in\mathrm{Sh}(S)$, since a labeling of I can be realized by invoking the shattering of J with the new coordinate fixed, say, to zero. Therefore

$$\begin{aligned}\widehat{\Theta}_\beta(S\|z)&=\sum_{J\in\mathrm{Sh}(S\|z)}\beta^{|J|}\\&\le\sum_{J\in\mathrm{Sh}(S)}\beta^{|J|}+\sum_{I\in\mathrm{Sh}(S)}\beta^{|I|+1}=(1+\beta)\widehat{\Theta}_\beta(S).\end{aligned}$$

□

Proof of Proposition 4.1. Set $\Phi_t:=\widehat{\Theta}_\beta(S_t)$ for $t=0,\dots,T$. Every prefix is realizable by $f^\star$, so $\varnothing\in\mathrm{Sh}(S_t)$ and $\Phi_t\ge1$. Moreover, $\Phi_0=\widehat{\Theta}_\beta(())=1$. By Lemma 4.1, every round satisfies

$$\Phi_t\le(1+\beta)\Phi_{t-1}.$$

On a mistake round, the learner predicted a branch of maximum potential and the update follows the other branch, which has minimum potential. Since the learner predicted rather than abstained,

$$\Phi_t=\min\{P_0,P_1\}\le\tfrac{3}{4}\Phi_{t-1}.$$

Writing $M=\mathsf{MisErr}$ and iterating gives

$$1\le\Phi_T\le\left(\frac{3}{4}\right)^M(1+\beta)^{T-M}.$$

Taking logarithms yields

$$M\le\frac{T\log(1+\beta)}{\log(4(1+\beta)/3)}.$$

□

Proof of Lemma 4.2. Fix $i \in [n]$. For each $I \subseteq [n] \setminus \{i\}$ and $y \in \{0,1\}$, define

$$a_y(I) := \mathbf{1}\{I \in \mathrm{Sh}(S[i \leftarrow y])\}, \qquad b(I) := \mathbf{1}\{I \in \mathrm{Sh}(S_{-i})\}, \qquad q(I) := \mathbf{1}\{I \cup \{i\} \in \mathrm{Sh}(S)\}.$$

For every such I,

$$a_0(I) + a_1(I) \le b(I) + q(I). \tag{4}$$

Indeed, if $a_y(I) = 1$ for either y, deleting the constraint at i can only help, so $b(I) = 1$. If both $a_0(I) = a_1(I) = 1$, then any labeling of $I \cup \{i\}$ can be realized by choosing a witness from $S[i \leftarrow 0]$ or $S[i \leftarrow 1]$ according to the desired label at i; hence $q(I) = 1$.

Decompose $P_i(y)$ according to whether the shattered set contains i. If a set contains i, its membership in $\mathrm{Sh}(S[i \leftarrow y])$ is independent of the stored label at i, and hence is the same as its membership in $\mathrm{Sh}(S)$. Therefore

$$P_i(y) = \sum_{I \subseteq [n] \setminus \{i\}} \beta^{|I|} a_y(I) + \beta W_i(S).$$

Summing over y and using (4) yields

$$\begin{aligned} P_i(0) + P_i(1) &\le \sum_{I \subseteq [n] \setminus \{i\}} \beta^{|I|} \big(b(I) + q(I)\big) + 2\beta W_i(S) \\ &= \widehat{\Theta}_\beta(S_{-i}) + W_i(S) + 2\beta W_i(S), \end{aligned}$$

where $\sum_I \beta^{|I|} q(I) = W_i(S)$. This is the claimed inequality. □

Proof of Proposition 4.2. Fix $i \in \mathrm{Bad}(S)$. Badness and Lemma 4.2 imply

$$\frac{3}{2}\widehat{\Theta}_\beta(S_{-i}) < P_i(0) + P_i(1) \le \widehat{\Theta}_\beta(S_{-i}) + (1 + 2\beta) W_i(S),$$

and consequently

$$W_i(S) > \frac{\widehat{\Theta}_\beta(S_{-i})}{2(1 + 2\beta)}. \tag{5}$$

The sequence S is a permutation of $S_{-i} \| z_i$. Permutation invariance and Lemma 4.1 therefore give

$$\widehat{\Theta}_\beta(S) \le (1 + \beta)\widehat{\Theta}_\beta(S_{-i}).$$

Combining with (5), every bad index satisfies

$$W_i(S) > \frac{\widehat{\Theta}_\beta(S)}{2(1 + 2\beta)(1 + \beta)}.$$

Summing this strict inequality over the bad indices gives

$$\frac{|\,\mathrm{Bad}(S)|}{2(1 + 2\beta)(1 + \beta)} \widehat{\Theta}_\beta(S) < \sum_{i \in \mathrm{Bad}(S)} W_i(S) \le \sum_{i=1}^{n} W_i(S). \tag{6}$$

On the other hand, every shattered set has size at most d, so

$$\begin{aligned} \sum_{i=1}^{n} W_i(S) &= \sum_{U \in \mathrm{Sh}(S)} |U| \beta^{|U|-1} \\ &\le \frac{d}{\beta} \sum_{U \in \mathrm{Sh}(S)} \beta^{|U|} = \frac{d}{\beta} \widehat{\Theta}_\beta(S). \end{aligned}$$

Since S is realizable, $\widehat{\Theta}_\beta(S) \ge 1$. Substituting the preceding upper bound on $\sum_{i=1}^{n} W_i(S)$ into (6) and canceling the positive potential gives

$$|\,\mathrm{Bad}(S)| \le \frac{2d(1 + 2\beta)(1 + \beta)}{\beta}.$$

□

Proof of Proposition 4.3. Fix the adversary's private seed. The clean rounds $1 \le s_1 < \cdots < s_N \le T$ and all corrupted examples are now fixed. Let

$$Z_j := (x_{s_j}, f^\star(x_{s_j})), \qquad j \in [N].$$

Because the adversary is oblivious, $Z_1, \dots, Z_N$ are i.i.d. labeled draws. For $i \in [N]$, let C_i be the labeled sequence of corrupted rounds preceding s_i and form the reordered sequence

$$R_i := C_i \| Z_1 \| \cdots \| Z_i.$$

The actual history immediately before round s_i is a permutation of R_i with its final entry removed. Since the weighted history potential is permutation invariant, the learner abstains at s_i exactly when the final index of R_i is bad:

$$\mathbb{P}(\widehat{y}_{s_i} = \perp) = \mathbb{P}\big(|R_i| \in \mathrm{Bad}(R_i)\big). \tag{7}$$

The prefix C_i is fixed, while the i.i.d. clean examples fill the last i positions. Permuting these positions leaves the law of R_i unchanged. It also permutes the membership of the clean positions in $\mathrm{Bad}(R_i)$, because deletion, relabeling, and $\widehat{\Theta}_\beta$ are equivariant under reordering. All clean positions are therefore equally likely to be bad, and

$$\begin{aligned}
\mathbb{P}\big(|R_i| \in \mathrm{Bad}(R_i)\big) &= \frac{1}{i}\mathbb{E}\big[|\,\mathrm{Bad}(R_i) \cap \{\text{clean positions}\}|\big] \\
&\le \frac{1}{i}\mathbb{E}[|\,\mathrm{Bad}(R_i)|] \\
&\le \frac{2d(1+2\beta)(1+\beta)}{\beta i},
\end{aligned}$$

where the last step is Proposition 4.2. Combine with (7). □

We are now ready to prove a misclassification and abstention error bound for the algorithm using the weighted history potential $\widehat{\Theta}_\beta$.

Theorem C.1 (Unknown distribution, oblivious corruption)**.** *For every $\beta > 0$, Algorithm 2 with non-age-dependent potential $\widehat{\Theta}_\beta$ satisfies, on every realizable labeled interaction,*

$$\mathsf{MisErr} \le \frac{\log(1+\beta)}{\log(4(1+\beta)/3)}\, T. \tag{8}$$

Against an oblivious adversary,

$$\mathbb{E}[\mathsf{AbsErr}] \le \frac{2d(1+2\beta)(1+\beta)}{\beta}\, H_T. \tag{9}$$

The expectation is over the clean draws and any randomization of the adversary and learner.

Proof. The mistake claim is Proposition 4.1. For abstentions, condition on the oblivious adversary's private seed and retain the notation from the preceding proof. If $N = 0$, then $\mathsf{AbsErr} = 0$ and there is nothing to prove. Otherwise, Proposition 4.3 gives

$$\begin{aligned}
\mathbb{E}[\mathsf{AbsErr} \mid \text{seed}] &\le \frac{2d(1+2\beta)(1+\beta)}{\beta} \sum_{i=1}^{N} \frac{1}{i} \\
&= \frac{2d(1+2\beta)(1+\beta)}{\beta} H_N \\
&\le \frac{2d(1+2\beta)(1+\beta)}{\beta} H_T.
\end{aligned}$$

Taking expectation over the adversary's seed proves the abstention error claim. □

C.2 Upper bound for the Age-Weighted History Potential

Bounding the misclassification error for this potential follows the same template as for the simpler potential $\widehat{\Theta}_\beta$. The following is the direct analog of Lemma 4.1.

Lemma C.1 (One-step growth)**.** *For every labeled history S and any new labeled example z,*

$$\widehat{\Theta}_\lambda(S \parallel z) \le e^{2\beta}\widehat{\Theta}_\lambda(S).$$

Proof. Since z is added last, its age-dependent weight within the new dataset $S \parallel z$ is β. Additionally, note that for any $u \le t$ one has

$$\frac{\lambda_{u,t+1}}{\lambda_{u,t}} \le 1 + \frac{\beta}{d}.$$

We then compute

$$\begin{aligned}
\widehat{\Theta}_\lambda(S \parallel z) &\le \sum_{U \in \mathrm{Sh}(S)} \prod_{u \in U} \lambda_{u,t+1} + \sum_{U \subseteq [n] : U \cup \{z\} \in \mathrm{Sh}(S \parallel z)} \beta \prod_{u \in U} \lambda_{u,t+1} \\
&\le (1+\beta) \sum_{U \in \mathrm{Sh}(S)} \prod_{u \in U} \lambda_{u,t+1} \\
&\le (1+\beta) \sum_{U \in \mathrm{Sh}(S)} \left(1 + \frac{\beta}{d}\right)^{|U|} \prod_{u \in U} \lambda_{u,t} \le (1+\beta) \left(1 + \frac{\beta}{d}\right)^{d} \widehat{\Theta}_\lambda(S).
\end{aligned}$$

Using $1 + x \le e^x$ ends the proof. □

From there, the same arguments give the following misclassification bound. Specifically, if M mistakes have been made in total and S_T denotes the dataset after all T rounds, we have

$$1 \le \widehat{\Theta}_\lambda(S_T) \le e^{2\beta(T-M)} \left(\frac{3}{4}\right)^M.$$

This corresponds to the following misclassification bound:

Proposition C.1. *For every realizable interaction, Algorithm 2 with the age-dependent potential satisfies*

$$\mathsf{MisErr} \le \frac{2\beta}{\log(4/3) + 2\beta} T.$$

We next turn to bounding the abstention error with the new potential. The end goal of this part is to prove the following abstention probability on the i-th clean sample:

Proposition C.2. *For the ith clean round s_i of any fixed oblivious corruption schedule,*

$$\mathbb{P}(\widehat{y}_{s_i} = \perp) \lesssim \left(\frac{d}{\beta i}\right)^2 + \frac{d}{i}.$$

We defer the proof to the next section Section C.2.1. Compared to Proposition 4.3 which gave a bound $O(\frac{d}{\beta i})$ for the simpler potential $\widehat{\Theta}_\beta$, Proposition C.2 shows a tighter tail decay for the age-weighted potential $\widehat{\Theta}_\beta$ ($\sim 1/i^2$ instead of $\sim 1/i$). This in turn removes a $\log T$ factor for the abstention error once these probabilities are summed.

Proposition C.3. *Fix $\beta \in (0, 1/4]$. Against any oblivious adversary, Algorithm 2 with the age-dependent potential satisfies*

$$\mathbb{E}[\mathsf{AbsErr}] \lesssim \frac{d}{\beta} + d \log\left(1 + \frac{\beta T}{d}\right).$$

Proof. The expected abstention error is exactly the sum of probabilities of abstention for each clean round, which are bounded in Proposition C.2. Hence, letting N be the number of clean rounds,

$$\mathbb{E}[\mathsf{AbsErr}] \lesssim \sum_{i=1}^{N} \min\left(\left(\frac{d}{\beta i}\right)^2 + \frac{d}{i}, 1\right) \lesssim \frac{d}{\beta} + d \log\left(1 + \frac{\beta N}{d}\right).$$

Using $N \le T$ ends the proof. □

Together, Propositions C.1 and C.3 prove Theorem 4.1.

C.2.1 Bounding abstention probabilities

For convenience, we introduce a few definitions. Given a dataset $S \subseteq \mathcal{X} \times \{0,1\}$, and weights $\lambda : S \to \mathbb{R}_+$, for any family $T \subseteq 2^S$ we define the corresponding partition function

$$Z_\lambda(T) := \sum_{U \in T} \prod_{u \in U} \lambda(u).$$

In particular, if we denote by λ_t the age-dependent weights, then the age-dependent potential is obtained by taking the partition function over all shattered sets in S: $\widehat{\Theta}_\lambda(S) = Z_{\lambda_t}(\mathrm{Sh}(S))$.

We will consider two different families. *Old* shattered sets correspond to those that are shattered without using the last datapoint:

$$\mathcal{O} := \{U \in \mathrm{Sh}(S) : t \notin U\} \subseteq \mathrm{Sh}(S_{-t}).$$

In particular, $\mathcal{O} \subseteq \mathrm{Sh}(S_{-t})$. Conversely, *new* shattered sets are those that were created when adding the new datapoint:

$$\mathcal{N} := \{U \subseteq [t-1] : U \cup \{t\} \in \mathrm{Sh}(S)).$$

We start by observing that the decomposition from the previous potential $\widehat{\Theta}_\beta$ still applies here up to minor modifications. By slight abuse of notation, we denote by $P_t(0)$ and $P_t(1)$ the two next-round potentials if the label is 0 or 1. In particular, the learner abstains if

$$\min_{y \in \{0,1\}} P_t(y) > \frac{3}{4} \widehat{\Theta}_{\lambda,t-1}(S_{-t}).$$

Similar arguments as in Lemma 4.2 up to book-keeping weights show the following identity when a new point triggers an abstention.

Lemma C.2. *Let $\beta \in (0, 1/4]$. For every history $S = (x_u, y_u)_{u<t}$ and new point (x_t, y_t), if x_t triggered an abstention, then denoting by λ_t the age-dependent weights,*

$$Z_{\lambda_t}(\mathcal{N}) \geq \frac{1}{10} Z_{\lambda_t}(\mathcal{O}). \tag{10}$$

Proof. The same arguments as in Lemma 4.2 up to book-keeping weights gives the following similar potential decomposition:

$$P_i(0) + P_i(1) \leq Z_{\lambda_t}(\mathrm{Sh}(S_{-t})) + (1 + 2\beta) Z_{\lambda_t}(\mathcal{N}). \tag{11}$$

Next, recall that for any $u < t$ one has $\lambda_{u,t} \leq (1 + \beta/d)\lambda_{u,t-1}$. As a result, when a learner abstains one has

$$P_i(0) + P_i(1) \geq \frac{3}{2} \widehat{\Theta}_{\lambda_{t-1}}(S_{-t}) \geq \frac{3}{2(1+\beta/d)^d} Z_{\lambda_t}(\mathrm{Sh}(S_{-t})) \geq \frac{3}{2e^\beta} Z_{\lambda_t}(\mathrm{Sh}(S_{-t}))$$

Together with Equation (11) this implies

$$Z_{\lambda_t}(\mathcal{N}) \geq \frac{\frac{3}{2}e^{-\beta} - 1}{1 + 2\beta} Z_{\lambda_t}(\mathrm{Sh}(S_{-t})) \geq \frac{1}{10} Z_{\lambda_t}(\mathrm{Sh}(S_{-t})).$$

In the last inequality we used $\beta \leq 1/4$. Noting that $\mathcal{O} \subseteq \mathrm{Sh}(S_{-t})$ ends the proof. □

Finally, we use an exchangeability argument to bound the probability of this event happening. This is the main step of the proof. Denote the number of clean rounds by N. In the following, we fix $i \in [N]$ and let τ_i denote the time of the i-th clean round. The following result bounds the probability of the event $\{Z_{\lambda_t}(\mathcal{N}) \geq \frac{1}{10} Z_{\lambda_t}(\mathcal{O})\}$ when randomly rotating the order of clean samples arrival.

Fix corrupted samples on all rounds and also fix the instances within clean rounds: $\{x_{\tau_j} : j \in [i]\}$ but not their assignment to clean rounds $C_i := \{\tau_j : j \leq i\}$. We have the following abstention probability bound where the probability is taken only over this random assignment. This immediately implies Proposition C.2.

Proposition C.4. *Let $i \in [N]$. Fix all corrupted instances, as well as clean round instances before τ_i (but not their assignment to clean rounds). Let $\pi \in \mathrm{Sym}(C_i)$ be a uniform reordering of the clean rounds. We denote by $S_i^\pi = (x_{\pi(u)}, y_{\pi(u)})_{u \le t}$ the corresponding reordered dataset. Similarly, we denote by $\mathcal{N}_i^\pi$ and $\mathcal{O}_i^\pi$ the corresponding old and new shattered set families, and λ^π the corresponding age-dependent weights. Then, for $\beta \in (0, 1/4]$ and any universal constant $c \in (0, 1]$,*

$$\mathbb{P}_\pi \left[Z_{\lambda^\pi}(\mathcal{N}_i^\pi) \ge c Z_{\lambda^\pi}(\mathcal{O}_i^\pi) \right] \lesssim \frac{d}{i} + \left(\frac{d}{\beta i} \right)^2 .$$

Proof. For convenience, we use the notation $t = \tau_i$, and denote by $S_0 := \{(x_{\tau_j}, y_{\tau_j}) : j \in [i]\}$ the clean instances that have been fixed (without their assignment to clean slots), and $S = \{(x_u, y_u), u \le t\}$ the complete dataset. For any clean datapoint $v \in S_0$, we introduce

$$\mathcal{O}(v) := \{U \in \mathrm{Sh}(S) : v \notin U\}$$
$$\mathcal{N}(v) := \{U \subseteq S \setminus \{v\} : U \cup \{v\} \in \mathrm{Sh}(S)\}.$$

In particular, for a fixed assignment π, denoting by $e = e(\pi)$ the corresponding last datapoint observed at time t, we have $\mathcal{O}_i^\pi = \mathcal{O}(e)$ and $\mathcal{N}_i^\pi = \mathcal{N}(e)$. Last, we introduce distributions μ_π and ν_π corresponding to the distribution on $\mathrm{Sh}(S)$ and $\mathcal{N}(e)$ corresponding to the partition function:

$$\mu_\pi(U) := \frac{\prod_{u \in U} \lambda^\pi(u)}{Z_{\lambda^\pi}(\mathrm{Sh}(S))}, \quad U \in \mathrm{Sh}(S). \quad \text{and} \quad \nu_\pi(U) := \frac{\prod_{u \in U} \lambda^\pi(u)}{Z_{\lambda^\pi}(\mathcal{N}(e))}, \quad U \in \mathcal{N}(e).$$

For any clean datapoint $v \in S_0$, we denote by $p_v(\pi)$ and $q_v(\pi)$ the probability that v belongs to a random shattered set according to μ_π and ν_π respectively:

$$p_v(\pi) := \mathbb{P}_{U \sim \mu_\pi}[v_r \in U] \quad \text{and} \quad q_v(\pi) := \mathbb{P}_{U \sim \nu_\pi}[v_r \in U].$$

In particular, since shattered sets contain at most d points, we have

$$\sum_{v \in S_0} p_v(\pi) \le \mathbb{E}_{U \sim \mu_\pi}[|U|] \le d \quad \text{and} \quad \sum_{v \in S_0} q_v(\pi) \le d. \tag{12}$$

We suppose in the following that the assignment π satisfies the considered property

$$Z_{\lambda^\pi}(\mathcal{N}_i^\pi) \ge c Z_{\lambda^\pi}(\mathcal{O}_i^\pi). \tag{13}$$

Fix $r \in [i]$ and denote by v_r the datapoint assigned to the r-th clean round in π. We consider the assignment $\pi^{(r)}$ obtained by swapping the clean observation in the r-th clean round with the last i-th clean round (corresponding to time $t = \tau_i$). Under $\pi^{(r)}$, the age-dependent weight of v_r is reduced to β while the weight of e becomes

$$a_r := \beta \left(1 + (t - \tau_r) \frac{\beta}{d} \right) \ge \beta.$$

All other age-dependent weights are kept unchanged. In particular, all sets in $\mathcal{N}(e)$ which didn't include v_r have the same weight under π and $\pi^{(r)}$. Therefore,

$$\begin{aligned} Z_{\lambda^{\pi^{(r)}}}(\mathcal{N}(e)) \ge \mathbb{P}_{U \sim \nu_\pi}[v_r \notin U] \cdot Z_{\lambda^\pi}(\mathcal{N}(e)) &= (1 - q_{v_r}(\pi)) Z_{\lambda^\pi}(\mathcal{N}(e)) \\ &\overset{(i)}{\ge} c(1 - q_{v_r}(\pi)) Z_{\lambda^\pi}(\mathcal{O}(e)) \\ &\overset{(ii)}{\ge} c(1 - q_{v_r}(\pi)) Z_{\lambda^{\pi^{(r)}}}(\mathcal{O}(e)). \end{aligned}$$

In (i) we used the assumption Equation (13) and in (ii) we used the fact that old shattered sets can only have reduced weight when the datapoint v_r is brought closer to the end time t.

We now turn our attention to the probability $p_e(\pi^{(r)})$ that e belongs to the distribution on $Sh(S)$ induced by $\pi^{(r)}$. Since e has weight a_r under $\pi^{(r)}$ we have

$$p_e(\pi^{(r)}) = \frac{a_r Z_{\lambda^{\pi^{(r)}}}(\mathcal{N}(e))}{Z_{\lambda^{\pi^{(r)}}}(\mathcal{O}(e)) + a_r Z_{\lambda^{\pi^{(r)}}}(\mathcal{N}(e))}$$

$$\overset{(i)}{\geq} \frac{ca_r(1-q_{v_r}(\pi))}{1+ca_r(1-q_{v_r}(\pi))}$$
$$\overset{(ii)}{\geq} \frac{1}{2}\min\left(ca_r(1-q_{v_r}(\pi)),1\right),$$

where in (i) we used the previous inequality and in (ii) we used $\frac{x}{1+x} \geq \frac{1}{2}\min(x,1)$ for $x \geq 0$. Summing these inequalities over r then gives

$$\begin{aligned}
2\sum_{r\in[i]} p_e(\pi^{(r)}) &\geq \sum_{r\in[i]} \min(ca_r,1)(1-q_{v_r}(\pi)) \\
&\geq c\sum_{r\in[i]}\min(a_r,1) - \sum_{r\in[i]} q_{v_r}(\pi) \\
&\overset{(i)}{\geq} c\sum_{r\in[i]}\min(a_r,1) - d.
\end{aligned}$$

In (i) we used Equation (12). As an important reminder, this bound holds under the event from Equation (13). On the other hand,

$$\sum_{r\in[i]}\min(a_r,1) \geq \sum_{r\in[i]}\min\left(\beta\left(1+(i-r)\frac{\beta}{d}\right),1\right) \gtrsim i\cdot\min\left(\beta+\frac{\beta^2(i-1)}{d},1\right).$$

Importantly, the mapping $(\pi, r) \to (\pi^{(r)}, e)$ is injective: given a target (σ, v), π is recovered by swapping v with the t-th datapoint of σ, and r is the index of v within σ. As a result, the last inequality implies that if σ is a uniform assignment,

$$\mathbb{E}_\sigma\left[\sum_{r\in[i]} p_e(\sigma^{(r)})\right] \leq \mathbb{E}_\sigma\left[\sum_{v\in S_0} p_v(\pi)\right] \leq d,$$

where in the last inequality we used Equation (12). Taking the last three inequalities together implies that for some universal constant $c' > 0$,

$$\mathbb{P}_\sigma[Z_{\lambda^\sigma}(\mathcal{N}_i^\sigma) \geq cZ_{\lambda^\sigma}(\mathcal{O}_i^\sigma)] \leq \frac{2d}{c'i\cdot\min\left(\beta+\frac{\beta^2(i-1)}{d},1\right)-d}.$$

For $i \lesssim d/\beta$ we can simply bound the probability by one. For $i \gtrsim d/\beta$ the bound above yields

$$\mathbb{P}_\sigma[Z_{\lambda^\sigma}(\mathcal{N}_i^\sigma) \geq cZ_{\lambda^\sigma}(\mathcal{O}_i^\sigma)] \lesssim \frac{d}{i\min\left(\frac{\beta^2 i}{d},1\right)} \lesssim \frac{d}{i}+\frac{d^2}{\beta^2 i^2}.$$

This ends the proof. □

C.3 The direct-sum lower bound

Proof of Theorem 4.2. We first prove the budget tradeoff by averaging over a family of problem instances constructed below. Step 4 then selects one fixed instance that satisfies one of the theorem's two alternatives. The $\Omega(\sqrt{dT})$ total-cost bound follows by a suitable choice of the abstention budget A. Throughout the proof, expectations include the random choice of instance, the fresh clean draws, and any internal randomness of the learner.

Step 1: the class. Set

$$m := \min\{d, A\}.$$

Because d and A are positive integers, $1 \leq m \leq d$. Fix integers $b \geq 2$ and $L \geq 1$. We use the rooted b-ary tree-path class appearing in the VC-1 lower bounds of Edelman & Goel (2026) and Yu & Blanchard (2026b), truncated at depth L: a hypothesis chooses a root-to-leaf path and labels exactly its nodes by one. The hidden-clean-block sampling below is the Edelman–Goel construction;

Yu–Blanchard use the same type of VC-1 tree class with a different recursive adversary. Either source establishes that one copy has VC dimension one. For completeness, no two nodes are shattered: incomparable nodes cannot both be positive, while for comparable nodes the labeling that makes the ancestor negative and the descendant positive is impossible; any non-root node is itself shattered.

Take m disjoint active copies and let a hypothesis choose its path in each copy independently. Their direct sum has VC dimension exactly m. To obtain dimension exactly d, take the product with the class of all labelings of $d - m$ additional isolated points. These additional points never appear in the process, so this padding changes neither the transcript nor either loss. The resulting class has VC dimension $m + (d - m) = d$.

For the tradeoff below, the displayed finite class may depend on T and A through b and L. Equivalently, one can fix the active tree class once and for all using m disjoint countably infinite rooted trees and realize every finite construction inside them.

Step 2: the schedule and the hidden clean block. In active copy $r \in [m]$, let $v_{r,j}$ be the hidden path node at depth j, with $v_{r,0}$ the root. At each depth j, choose $v_{r,j}$ independently and uniformly among the b children of $v_{r,j-1}$. For $j \in [L]$, let P_j be uniform over all children of the current depth-$(j-1)$ path nodes across the m active copies:

$$P_j := \mathrm{Unif}\big\{u : u \text{ is a child of } v_{r,j-1} \text{ for some } r \in [m]\big\}.$$

Choose a hidden clean block $J \sim \mathrm{Unif}([L])$ and set the clean distribution to $\mu := P_J$. Split the first

$$T_0 := Lmb$$

rounds into L blocks of length mb. In block j, if $j = J$, all rounds are clean i.i.d. draws from P_j; otherwise every round is corrupted and its instance is drawn from P_j before the interaction begins. All labels are the truthful labels induced by the same hidden paths.

Whenever $T_0 < T$, pad the remaining $T - T_0$ rounds with corrupted repetitions of the root of the first active tree, which is labeled one by every target in the active class. This padding is oblivious and clean-label. Abstentions on the padded rounds are free, and any additional mistakes only increase the lower-bound cost, so it is enough to analyze the first T_0 rounds. This also specifies a full horizon-T instance rather than silently discarding unused rounds.

The paths, the index J, and all corrupted examples are sampled before the interaction begins. Once these choices are fixed, they determine the target, the clean distribution $\mu = P_J$, and the entire corruption schedule and injected sequence. Thus each realization defines a valid instance with an oblivious adversary; clean examples are then drawn independently from μ on the designated clean rounds.

Under the distributional construction, the labeled transcript has the same law for every value of J: block j always consists of mb draws from P_j with the labels of the same hidden paths. The learner never observes the corruption indicators. Let $N_\perp(j)$ be its number of abstentions in block j and let

$$N_\perp := \sum_{j=1}^{L} N_\perp(j)$$

be its total number of abstentions in the first T_0 rounds. Since only block J is clean and J is uniform,

$$\mathbb{E}[\mathsf{AbsErr}] = \frac{1}{L}\sum_{j=1}^{L} \mathbb{E}[N_\perp(j) \mid J = j] = \frac{1}{L}\sum_{j=1}^{L} \mathbb{E}[N_\perp(j)] = \frac{1}{L}\mathbb{E}[N_\perp]. \tag{14}$$

Step 3: uncertain rounds. In block j, each draw first selects an active copy r and then a child of the current path node $v_{r,j-1}$. Call the round *uncertain* if it is the first occurrence of the selected child, the positive child of that parent has not appeared earlier in the block, and at least two children remain possible before this occurrence. We now extend the one-copy coupon calculation of Edelman & Goel (2026) to the uniform mixture of the m active copies.

Fix one block and first expose the copy index of each of its mb draws. For active copy $r \in [m]$, let N_r be its number of draws, and list their child indices in within-copy order as $C_{r,1}, \ldots, C_{r,N_r} \in [b]$. Then

$$N_r \sim \mathrm{Bin}(mb, 1/m), \qquad \mathbb{E}[N_r] = b, \qquad \mathrm{Var}(N_r) \le b.$$

Let $\theta_r \in [b]$ be the positive child in copy r, and put $q := \lfloor b/2 \rfloor$. Chebyshev's inequality gives, for $b \geq 8$,

$$\mathbb{P}(N_r \geq q) \geq 1 - \mathbb{P}\big(|N_r - b| \geq b/2\big) \geq 1 - \frac{4}{b} \geq \frac{1}{2}.$$

On the event $\{N_r \geq q\}$, the first q within-copy child indices are independent uniform draws from the same b children. Let $U_r^{(q)}$ be the number of their first occurrences that come no later than the first occurrence of the positive child. For a child c, let τ_c be its first occurrence among these draws, with $\tau_c = \infty$ if it does not occur. By symmetry, for every $c \neq \theta_r$,

$$\mathbb{P}(\tau_c \leq q,\ \tau_c < \tau_{\theta_r} \mid N_r \geq q) = \frac{1}{2}\left(1 - \left(1 - \frac{2}{b}\right)^q\right).$$

Indeed, this probability is

$$\sum_{t=1}^{q} \left(1 - \frac{2}{b}\right)^{t-1} \frac{1}{b}.$$

The positive child itself contributes when $\tau_{\theta_r} \leq q$. Because the copy count is independent of the within-copy child indices,

$$\mathbb{E}[U_r^{(q)} \mid N_r \geq q] = 1 - \left(1 - \frac{1}{b}\right)^q + \frac{b-1}{2}\left(1 - \left(1 - \frac{2}{b}\right)^q\right) \geq \frac{b}{4}.$$

For $q = \lfloor b/2 \rfloor$ and $b \geq 8$, we have $2q/b \geq 8/9$, and hence

$$1 - \left(1 - \frac{2}{b}\right)^q \geq 1 - e^{-2q/b} \geq 1 - e^{-8/9} > \frac{4}{7}.$$

Thus the second term alone is greater than $2(b-1)/7 \geq b/4$. Before any of these first occurrences, at most $q - 1 < b - 1$ children have been ruled out, so at least two possible positive children remain. Thus all $U_r^{(q)}$ counted rounds are uncertain.

Since $\mathbb{P}(N_r \geq q) \geq 1/2$, the expected number of uncertain rounds contributed by active copy r is at least

$$\mathbb{P}(N_r \geq q)\,\mathbb{E}[U_r^{(q)} \mid N_r \geq q] \geq \frac{b}{8}.$$

Summing over the m active copies shows that every length-mb block contains at least $mb/8$ uncertain rounds in expectation, and hence the L blocks contain at least $T_0/8$ in expectation. This step uses only linearity; the copywise counts need not be independent. Conditional on the learner's information after it observes the current child on an uncertain round, symmetry leaves the positive child uniform among the s children not yet ruled out, where $2 \leq s \leq b$. Predicting zero errs with probability $1/s \geq 1/b$, while predicting one errs with probability $1 - 1/s \geq 1/2 \geq 1/b$. Thus every non-abstaining prediction on an uncertain round has conditional mistake probability at least $1/b$.

Abstention can avoid the mistake on an uncertain round, so subtract the abstentions on those rounds and upper bound their number by $N_\perp$. We obtain

$$\mathbb{E}[\mathsf{MisErr}] \geq \frac{1}{b}\left(\frac{T_0}{8} - \mathbb{E}[N_\perp]\right). \tag{15}$$

Step 4: the budget tradeoff. Now set

$$b := \left\lceil \frac{16A}{m} \right\rceil, \qquad L := \left\lfloor \frac{T}{mb} \right\rfloor.$$

Because $m \leq A$ and $A \leq T/32$,

$$b \geq 16, \qquad b \leq \frac{32A}{m}, \qquad mb \leq T, \qquad L \geq \frac{T}{2mb}.$$

In particular, $T_0 = Lmb \in [T/2, T]$. As in Step 2, pad the remaining $T - T_0$ rounds with corrupted repetitions of the first active tree's root.

Algorithm 5 Censored learner with age-dependent weights and periodic restarts

Require: Class $\mathcal{F}$, parameter $\beta \in (0, 1]$
$U_0 \leftarrow ()$ and $W_\beta \leftarrow \lceil d \log(T+1)/\beta \rceil$
For $t = 1, \ldots, T$:
  If $t \equiv 1 \pmod{W_\beta}$, $V \leftarrow \mathcal{F}$ ▷ *Restart version space at beginning of epoch*
  Observe x_t and set $P_b \leftarrow \widehat{\Theta}^u_t(V^{x_t \to b}; U_{t-1})$ for $b \in \{0, 1\}$
  If $\min\{P_0, P_1\} > \frac{3}{4}\widehat{\Theta}^u_t(V; U_{t-1})$, output $\widehat{y}_t \leftarrow \perp$
  Else
    output $\widehat{y}_t \leftarrow \arg\max_{b \in \{0,1\}} P_b$, breaking ties deterministically, and observe y_t
    if $\widehat{y}_t \neq y_t$, update $V \leftarrow V^{x_t \to y_t}$ ▷ *update only after a mistake*
  update $U_t \leftarrow U_{t-1} \,\|\, x_t$

If some fixed member of the hard family has $\mathbb{E}[\mathsf{AbsErr}] > A$, the first alternative of the theorem holds. Otherwise every member has expected clean abstentions at most A, so the uniform mixture also satisfies $\mathbb{E}[\mathsf{AbsErr}] \leq A$. Equation (14) then gives $\mathbb{E}[N_\perp] \leq AL$. Substitute this into (15) and use $T_0 = Lmb$:

$$\begin{aligned}\mathbb{E}[\mathsf{MisErr}] &\geq L\left(\frac{m}{8} - \frac{A}{b}\right) \geq \frac{Lm}{16} \\ &\geq \frac{T}{32b} \geq \frac{mT}{1024A} \\ &= \frac{1}{1024}\min\left\{\frac{dT}{A}, T\right\}.\end{aligned}$$

Therefore some fixed member of the family forces the stated expected mistake lower bound.

□

D Unlabeled History Potential under Censored Feedback

In this section, we first summarize our algorithm for censored feedback in Algorithm 5, then prove its performance guarantee:

Theorem D.1 (Censored feedback with oblivious adversaries)**.** *There is a universal constant $C > 0$ such that, for every $\beta \in (0, 1]$, Algorithm 5 satisfies, on every realizable interaction,*

$$\mathrm{MisErr} \leq \frac{4\beta T}{\log(4/3)}. \tag{16}$$

Also, against every oblivious adversary,

$$\mathbb{E}[\mathrm{AbsErr}] \leq C\left(\frac{d}{\beta} + d\log^3(2T)\right).$$

The expectation is over the clean draws and any randomization of the oblivious adversary and learner.

Note that up to the additive abstention error $O(d\log^3(2T))$, this censored algorithm achieves the same guarantee as the full feedback algorithm Algorithm 2 up to constant factors.

For uniform concentration purposes, we periodically restart the version space $V \leftarrow \mathcal{F}$ every

$$W_\beta := \left\lceil \frac{d\log(T+1)}{\beta} \right\rceil, \tag{17}$$

iterations, forming epochs of length W_β. This will in turn ensures that the potential number of version spaces arising throughout the procedure can be controlled.

The proof has the same two deterministic ingredients as the unknown-distribution proof. First, with Sauer's lemma, at the start of each restart period, the potential is at most $(T+1)^{2d}$. Hence each period contributes at most $O(d \log T)$ mistakes. Summing across all periods gives $O(d \log T \cdot T/W_\beta) = O(\beta T)$ mistakes.

Second, the replay-branch decomposition is essentially the same as in Lemma 4.2. The main new issue is that the version space is selected by the previous mistakes and may therefore depend on the clean ordering. We therefore prove a uniform version of the replay argument Proposition 4.3 for each potential version space V. Once the version space is fixed, we can use a similar exchangeability argument as in the non-censored case: conditionally on the oblivious adversary's seed, and the unordered pool of clean samples, the samples observed on each clean round are exchangeable and can be permuted. The replay argument from Proposition 4.3 essentially applies to each such fixed version space.

The version spaces arising throughout the procedure depend on the clean ordering. We therefore use a uniform concentration argument to bound the abstention probability simultaneously over every potential version space V. By restarting periodically, we drastically reduce the number of all potential version spaces. Union bounding over the remaining version spaces then gives the additional $d \log^3(2T)$ term.

We now prove the misclassification and abstention bounds of Theorem D.1 separately in the following two sections.

D.1 Mistake accounting and the number of active constraints

Lemma D.1 (Misclassification bound)**.** *On every realizable interaction, Algorithm 5 satisfies (16). Moreover, it makes at most*

$$K_T := \left\lceil \frac{4d \log(T+1) + 2\beta}{\log(4/3)} \right\rceil$$

mistakes in each block. Every current version space is consequently specified by at most K_T truthful constraints from the current block.

Proof. We first note that since every shattered set has size at most d and weights are bounded by $\lambda_{u,t} \le T+1$ for any $t \le T$, we have for any version spaces $V, V' \subseteq \mathcal{F}$

$$1 \le \widehat{\Theta}^u_t(V; U_t) \le \sum_{k=0}^{d} \binom{T}{k} (T+1)^k \le (T+1)^{2d} \le (T+1)^{2d} \widehat{\Theta}^u_{t-1}(V'; U_{t-1}),$$

provided that V and V' are non-empty (realizable).

We now denote by V_t the version space at the end of round t, and by m_t the indicator whether a mistake was made at that round. If there was no restart of the version space at time t, one has

$$\begin{aligned}
\widehat{\Theta}^u_t(V_t; U_t) \le (1+\beta)\widehat{\Theta}^u_t(V_t; U_{t-1}) &\le (1+\beta)\left(\frac{3}{4}\right)^{m_t} \widehat{\Theta}^u_t(V_{t-1}; U_{t-1}) \\
&\le (1+\beta)\left(1+\frac{\beta}{d}\right)^{d}\left(\frac{3}{4}\right)^{m_t} \widehat{\Theta}^u_{t-1}(V_{t-1}; U_{t-1}) \\
&\le e^{2\beta}\left(\frac{3}{4}\right)^{m_t} \widehat{\Theta}^u_{t-1}(V_{t-1}; U_{t-1}).
\end{aligned}$$

In the first an third inequality we used the same arguments as in the non-censored case: appending a new point x_t with weight β increases the potential by at most a factor $1+\beta$ and $\lambda_{u,t} \le (1+\beta/d)\lambda_{u,t-1}$ for all $u \le t-1$.

In particular, let r denote the number of restarts and M the total number of mistakes. Since the potential is always between $[1, (T+1)^{2d}]$, every restart inflates the potential by at most $(T+1)^{2d}$, hence

$$1 \le \widehat{\Theta}^u_T(V_T; U_T) \le (T+1)^{2dr} e^{2\beta T}\left(\frac{3}{4}\right)^{M} \le e^{4\beta T}\left(\frac{3}{4}\right)^{M}.$$

In the last inequality we used the fact that $r \le T/W_\beta \le \beta T/(d\log(T+1))$. Rearranging gives the desired misclassification bound.

Next, within each block that has length at most W_β, there is at most 1 restart (at the beginning). Hence the same arguments as above show that the corresponding number of mistakes m satisfies

$$1 \le (T+1)^{2d} e^{2\beta W_\beta} \left(\frac{3}{4}\right)^m \le e^2 (T+1)^{4d} \left(\frac{3}{4}\right)^m.$$

Rearranging proves the desired bound on the number m of mistakes per block. □

D.2 Uniform concentration with capped weights

We fix the optimal function $f^\star$. For any subset $I \subseteq [T]$ we use the notation $w_t(I) := \prod_{u\in I} \lambda_{u,t}$ for the weight of I. For any set $C \subseteq [T]$ we also define the version space corresponding to mistakes on rounds in C via

$$V_C := \{f \in \mathcal{F} : \forall s \in C : f(x_s) = f^\star(x_s)\}.$$

In particular, each version space V_C is non-empty since it contains $f^\star$. For any non-empty version space $V \subseteq \mathcal{F}$ we then define

$$a_t(U, V, x) := \mathbf{1}\left\{\min_{b\in\{0,1\}} \widehat{\Theta}_t^u(V^{x\to b}; U) > \frac{3}{4}\widehat{\Theta}_t^u(V; U)\right\},$$

which is the indicator of whether round t led to an abstention of the algorithm, provided the unlabeled history is U, the current version space is V, and the new test point is x.

We start by relating the probability of an abstention to comparisons between weighted sums and potentials, similar to the replay-branch decomposition Lemma 4.2.

Lemma D.2. *For any $t \in [T]$, non-empty version space $V \subseteq \mathcal{F}$ and unlabeled history U,*

$$\{a_t(U, V, x) = 1\} \subseteq \left\{2 \sum_{I\subseteq[t-1]} w_t(I)\mathbf{1}[V \triangleright (x_I, x)] \ge \widehat{\Theta}_t^u(V; U)\right\}.$$

Proof. As in the non-censored case, for every I and x, the one-point shattering inequality is

$$\mathbf{1}\{V^{x\to 0} \triangleright x_I\} + \mathbf{1}\{V^{x\to 1} \triangleright x_I\} \le \mathbf{1}\{V \triangleright x_I\} + \mathbf{1}\{V \triangleright (x_I, x)\}.$$

Multiplying by the weight $w_t(I)$ and summing over the subsets I shows that if an abstention is triggerred at round t ($a_t(U, V, x) = 1$):

$$\frac{1}{2}\widehat{\Theta}_t^u(V; U) \le \widehat{\Theta}_t^u(V^{x\to 0}; U) + \widehat{\Theta}_t^u(V^{x\to 1}; U) - \widehat{\Theta}_t^u(V; U) \le \sum_{I\subseteq[t-1]} w_t(I)\mathbf{1}\{V \triangleright (x_I, x)\}.$$

This ends the proof. □

We are now ready to bound the probability of a bad abstention on a clean round t. We denote by $J := \{j < t : c_j = 0\}$ the set of previous clean rounds and by $i = |J|$ their number. We also fix an oblivious corruption schedule and all injected contexts. For convenience, we introduce the quantity

$$h_j := \beta\left(1 + \frac{\beta j}{d}\right), j \ge 1,$$

which lower-bounds the weight of the j-th previous clean sample before t.

Lemma D.3. *Fix $\delta \in (0, 1)$ and a positive integer $k \le i - d$. We introduce*

$$H := \sum_{j=1}^{i-d-k} h_j \quad \text{and} \quad h := 6h_i((k+d)\log(T+1) + \log(1/\delta)).$$

Then, with probability at least $1-\delta$, simultaneously over all subsets $C \subseteq [t-1]$ with $|C| \le k$, one has

$$\mu\{x \in \mathcal{X} : a_t(U_{t-1}, V_C, x) = 1\} \le \frac{4(d+h)}{H},$$

where the probability is taken only over the clean samples on clean rounds before t. In particular, with $\delta = 1/T^2$, if $i \geq 4(d+k)$ we have for some universal constant $c_0 > 0$, for any such subset $C \subseteq [t-1]$,

$$\mu\left\{x \in \mathcal{X} : a_t(U_{t-1}, V_C, x) = 1\right\}] \leq c_0 \left(\frac{d}{\beta i}\right)^2 + c_0 \frac{(k+d)\log(T+1)}{i}$$

Proof. For any subset $I \subseteq [t-1]$ we introduce

$$Z_{t,C,I} := \sum_{j \in J \setminus (C \cup I)} \lambda_{j,t} \mathbf{1}[V_C \triangleright (x_I, x_j)] \quad \text{and} \quad q_{t,C,I} := \mathbb{P}_{x \sim \mu}[V_C \triangleright (x_I, x)].$$

We first lower-bound $Z_{t,C,I}$ as a function of $q_{t,C,I}$. For convenience, we introduce

$$\bar{\lambda}_{j,t} := \min(\lambda_{j,t}, h_i),\ j \in J \setminus (C \cup I).$$

Conditionally on $x_{C \cup I}$, the variables $\mathbf{1}[V_C \triangleright (x_I, x_j)]$ for $j \in J \setminus (C \cup I)$ are i.i.d. Bernoulli variables of parameter $q_{t,C,I}$. Hence, Bernstein's inequality implies that with probability at least $1-\delta$ conditionally on $x_{C \cup I}$,

$$\begin{aligned} Z_{t,C,I} \geq \sum_{j \in J \setminus (C \cup I)} \bar{\lambda}_{j,t} \mathbf{1}[V_C \triangleright (x_I, x_j)] &\geq \frac{1}{2} q_{t,C,I} \sum_{j \in J \setminus (C \cup I)} \bar{\lambda}_{j,t} - 6h_i \log(1/\delta) \\ &\geq \frac{1}{2} q_{t,C,I} \sum_{j=1}^{|J \setminus (C \cup I)|} h_j - 6h_i \log(1/\delta). \end{aligned}$$

In the first inequality we used the definition of $\lambda_{j,t}$. Hence, taking the union bound over all subsets $I, C \subseteq [t-1]$ with $|I| \leq d$ and $|C| \leq k$ we obtain simultaneously over all subsets $C \subseteq [t-1]$ of size at most k, with probability at least $1-\delta$:

$$\sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) Z_{t,C,I} \geq \frac{H}{2} \sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) q_{t,C,I} - h \widehat{\Theta}^u_t(V_C; U_{t-1}). \tag{18}$$

Let $\mathcal{E}_t$ be the event on U_{t-1} on which Equation (18) holds simultaneously for all $C \subseteq [t-1]$ with $|C| \leq k$. Then $\mathbb{P}(U_{t-1} \in \mathcal{E}_t) \geq 1-\delta$. On the other hand, we note that

$$\begin{aligned} \sum_{\substack{I \subseteq [t-1] \\ V \triangleright x_I}} w_t(I) Z_{t,C,I} &\leq \sum_{\substack{I \subseteq [t-1] \\ V \triangleright x_I}} w_t(I) \sum_{j \in [t-1] \setminus I} \lambda_{j,t}\, \mathbf{1}[V_C \triangleright (x_I, x_j)] \\ &= \sum_{\substack{J \subseteq [t-1] \\ V \triangleright x_J}} |J| w_t(J) \leq d \widehat{\Theta}^u_t(V_C; U_{t-1}), \end{aligned}$$

where in the last inequality we used the fact that each shattered set in V_C has at most d elements.

Fix any $U_{t-1} \in \mathcal{E}_t$ and any $C \subseteq [t-1]$ with $|C| \leq k$. Together with Equation (18), Lemma D.2 implies that

$$\begin{aligned} \mathbb{E}_{x \sim \mu}[a_t(U_{t-1}, V_C, x)] &\leq \mathbb{P}_{x \sim \mu}\left[\frac{4(d+h)}{H} \sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) \mathbf{1}[V_C \triangleright (x_I, x)] \geq \sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) q_{t,C,I}\right] \\ &\leq \frac{4(d+h)}{H} \mathbb{E}_{x \sim \mu}\left[\frac{\sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) \mathbf{1}[V_C \triangleright (x_I, x)]}{\sum_{\substack{I \subseteq [t-1] \\ V_C \triangleright x_I}} w_t(I) q_{t,C,I}}\right] \\ &= \frac{4(d+h)}{H}. \end{aligned}$$

The equality follows from the definition of $q_{t,C,I}$. With at least $1-\delta$ probabiltiy over U_{t-1}, this holds simultaneously over all eligible C. Simplifying the expression proves the second claim. □

Lemma D.3 bound the probability of a bad abstention uniformly over the potential version spaces. During the run of the algorithm, these version spaces are defined by at most K_T constraints as per Lemma D.1. Putting these together then gives the desired abstention bound.

Proof of Theorem D.1. Lemma D.1 proves the misclassifiction bound. For abstentions, we condition on $f^\star$, injected contexts, and clean rounds, which we denote $s_1 < \ldots < s_N$. We apply Lemma D.3 with $k = K_T$ to each i-th clean round for $i \geq 4(d/\beta + K_T)$, which gives

$$\begin{aligned}
\mathbb{E}[\mathrm{AbsErr}] &= \mathbb{E}\left[\sum_{i=1}^{N} a_{s_i}(U_{s_i-1}, V_{s_i}, x_{s_i})\right] \\
&\leq 4\left(\frac{d}{\beta} + K_T\right) + \sum_{i \geq 4(d/\beta + K_T)} \left(c_0\left(\frac{d}{\beta i}\right)^2 + c_0\frac{(K_T + d)\log(T+1)}{i} + \frac{1}{T^2}\right) \\
&\leq c_1\left(\frac{d}{\beta} + d\log^3(T+1))\right),
\end{aligned}$$

for some universal $c_1 > 0$. This ends the proof. □

E Proofs for Learning Against Adaptive Adversaries

This appendix supplies the detailed construction and proofs for the compressed-witness learner in Section 5. We prove witness compression and the mistake bound in Appendix E.1, and the single-phase abstention guarantee in Appendix E.2. We then prove our guarantee for direct aggregation in Appendix E.3. Finally, Appendix E.4 develops the multi-phase reference expert, its implementation under censored feedback, and the parameter choices that yield the main tradeoff.

We use the notation from the main text:

$$M_T := \left\lceil \frac{d\log(T+1)}{\log(16/15)} \right\rceil, \qquad \ell_T = M_T + 1.$$

Here D_T is the ℓ_T-reduction dimension, and $\Theta^S = \widehat{\Theta}_T^S$ denotes the block potential. For every integer $\ell \geq 0$, define

$$\begin{aligned}
\mathcal{R}_\ell &:= \{\mathcal{F} \cap C : C \subseteq \mathcal{X} \times \{0,1\},\ |C| \leq \ell\}, \\
\mathcal{F} \cap C &:= \{f \in \mathcal{F} : f(x) = y \text{ for every } (x,y) \in C\}.
\end{aligned}$$

Thus $\mathcal{R}_\ell$ is the family of version spaces defined by at most ℓ labeled constraints.

E.1 Preliminaries on the witness compression

We first check that the reduction dimension can be used to bound the VC dimension of witness classes.

Lemma E.1 (VC dimension of the witness class)**.** *Fix $\ell \geq 0$. On the disjoint union $\mathcal{X}^{\leq d} := \bigcup_{k=0}^{d} \mathcal{X}^k$, the function class*

$$\mathcal{G} := \{Z \in \mathcal{X}^{\leq d} \mapsto \mathbf{1}[W \triangleright Z] : W \in \mathcal{R}_\ell\}$$

has VC dimension at most $C_0 D\log(2dD)$ for a universal constant C_0, where D is the ℓ-reduction dimension of $\mathcal{F}$.

Proof. It suffices to consider $r \geq D$. Fix r tuples in $\mathcal{X}^{\leq d}$. All r shattering indicators are determined by the trace of W on the union of their coordinates, which has $n \leq rd$ points. As $rd \geq D$, the reduction-dimension assumption bounds the number of possible indicator vectors by $(rd)^D$. If the r tuples are shattered by $\mathcal{G}$, then $2^r \leq (rd)^D$. In particular, we must have $r \leq C_0 D\log(2dD)$ for some universal C_0. □

We are now ready to check that short witnesses are well-defined.

Lemma E.2 (Existence of a short witness list)**.** *There is a universal constant $C_s > 0$ such that the following holds. Let D_T be the ℓ_T-reduction dimension of $\mathcal{F}$. Fix any bank $S_1, \ldots, S_m$ and any nonempty $V \in \mathcal{R}_{\ell_T - 1}$. Let $s = \lceil C_s D_T \log(2dD_T) \rceil$. There is a list of s witness tuples $Z_1, \ldots, Z_s$, each using at most d bank coordinates among $S_1, \ldots, S_m$, such that*

$$\sup_{x \in \mathcal{X}, b \in \{0,1\}} \left| \frac{1}{s} \sum_{u=1}^{s} \mathbf{1}[V^{x \to b} \triangleright Z_u] - a_b(x; V) \right| \leq \frac{1}{64}.$$

Proof. Fix (x, b). Since any shattered set in $V^{x \to b}$ must also be shattered in V, one has

$$\mathbb{P}_{Z \sim \nu_V}[V^{x \to b} \triangleright Z] = \frac{1}{m} \sum_{j=1}^{m} \sum_{I \subseteq [N]} \frac{T^{|I|} \mathbf{1}[V \triangleright S_{j,I}]}{\binom{N}{|I|} \Theta^{S_j}(V)} \mathbf{1}[V^{x \to b} \triangleright S_{j,I}] = a_b(x; V).$$

Apply uniform convergence to the VC class in Lemma E.1, at accuracy $1/64$ and success probability $1/2$, under the fixed distribution ν_V (e.g. see (Shalev-Shwartz & Ben-David, 2014, Theorem 6.8)). This precisely ensures the existence of the desired witnesses $Z_1, \ldots, Z_s$ such that

$$\sup_{W \in \mathcal{R}_{\ell_T}} \left| \frac{1}{s} \sum_{u=1}^{s} \mathbf{1}[W \triangleright Z_u] - \mathbb{E}_{Z \sim \nu_V}[W \triangleright Z]. \right| \leq \frac{1}{64}.$$

It then suffices to check that for any $(x, b) \in \mathcal{X} \times \{0, 1\}$ we have $V^{x \to b} \in \mathcal{R}_{\ell_T}$. □

For a nonempty class $W \subseteq \mathcal{F}$ and a bank $S \subseteq \mathcal{X}$, we define

$$g_W(S, x) = \mathbf{1}\left\{ \Theta^S(W^{x \to 0}) + \Theta^S(W^{x \to 1}) > \frac{5}{4} \Theta^S(W) \right\}.$$

The reference process and the matching reference expert. Given Lemma E.2, we define the *reference process* using Algorithm 3. On a sequence of instances with adaptive injections, the process first abstains until the first Nm clean rounds have passed. Then, it follows the reference expert's decision rules with sequentially selected witness lists: starting from $V = \mathcal{F}$, it uses the Nm clean rounds as $S_1, ..., S_m$ and uses Lemma E.2 to construct the first witness-index list $\mathcal{Z}_1 = \{Z_{11}, ..., Z_{1s}\}$. Then, it runs the reference expert's decision rules with $\mathcal{Z}_1$ until the version space V is updated. For this updated V, it again constructs a witness-index list using $S_1, ..., S_m$, and the construction repeats until it ends up with some $V \notin R_{\ell_T - 1}$ or $t = T$.

Assume the reference process makes at most $M_T = \ell_T - 1$ mistakes, we then have $V \in R_{\ell_T - 1}$ throughout the run, hence the process continues until $t = T$. Moreover, though the reference process is not an implementable learner, we are guaranteed that in hindsight, there is always a reference expert that behaves exactly the same — as long as the reference process uses at most M_T witness-index lists, there is always a reference expert whose start time t_0 and witness-index lists happen to match exactly with the reference process, which we will denote as *the matching reference expert*.

Next, we check that using the threshold $3/4$ on the approximate score $\widehat{a}_b(x; V)$ to trigger an abstention has adequate properties in terms of the estimated weighted potentials. In particular, we check if the reference process really makes at most M_T mistakes.

Lemma E.3. *Fix V and a new test point x. For the reference process:*

1. *On a mistaken prediction with true label y,*

$$\left(\prod_{j \in [m]} \Theta^{S_j}(V^{x \to y}) \right)^{1/m} < \frac{15}{16} \left(\prod_{j \in [m]} \Theta^{S_j}(V) \right)^{1/m}.$$

2. *On an abstention, $|\{j \in [m] : g_V(S_j, x) = 1\}| > m/6$.*

In particular, the reference process makes at most M_T mistakes.

Proof. On a mistaken prediction, we have by construction $\widehat{a}_y(x;V) \le 3/4$. Hence, using the arithmetico-geometric inequality:

$$\left(\prod_{j=1}^{m} \frac{\Theta^{S_j}(V^{x\to y})}{\Theta^{S_j}(V)}\right)^{1/m} \le \frac{1}{m}\sum_{j=1}^{m} \frac{\Theta^{S_j}(V^{x\to y})}{\Theta^{S_j}(V)} = a_y(x;V) \le \widehat{a}_y(x;V) + \frac{1}{64} \le \frac{3}{4} + \frac{1}{64} < \frac{15}{16}.$$

Next, for any nonempty class W and block S,

$$1 \le \Theta^{S}(W) \le \sum_{k=0}^{d} T^k \le (1+T)^d.$$

Let $P(V) = (\prod_j \Theta^{S_j}(V))^{1/m}$. After r mistakes with valid witness lists, $1 \le P(V) \le (1+T)^d(15/16)^r$. This precisely implies the desired bound $r \le M_T$ on the number of mistakes.

As a result, the version space V_t of the reference process at the beginning of any iteration $t \ge t_0$ has at most $M_T = \ell_T - 1$ constraints: $V \in \mathcal{R}_{\ell_T - 1}$. Then, Lemma E.2 implies that $|a_b(x;V) - \widehat{a}_b(x;V)| \le 1/64$ for both labels $b \in \{0,1\}$. In turn, on an abstention we must have for both labels $a_b(x;V) > 3/4 - 1/64 > 11/16$. Therefore,

$$\begin{aligned}\frac{11}{8} < a_0(x;V) + a_1(x;V) &= \frac{1}{m}\sum_{j\in[m]} \frac{\Theta^{S_j}(V^{x\to 0})}{\Theta^{S_j}(V)} + \frac{\Theta^{S_j}(V^{x\to 1})}{\Theta^{S_j}(V)} \\ &\le \frac{1}{m}\sum_{j\in[m]} \left(2g_V(S_j,x) + \frac{5}{4}(1 - g_V(S_j,x))\right).\end{aligned}$$

Rearranging gives the desired bound on the number of blocks $j \in [m]$ for which $g_V(S_j,x) = 1$. This ends the proof. □

E.2 A Leave-One-Out Bound for Clean Abstentions

We start by proving a bound on the expectation of the indicator $g_W(S,x)$ when the elements of the bank S and the test-point x are exchangeable, using a leave-one-out argument.

Lemma E.4 (One-block leave-one-out bound)**.** *Fix any nonempty class $W \subseteq \mathcal{F}$ and a tuple $R = (z_1, \dots, z_{N+1}) \in \mathcal{X}^{N+1}$. Let i be a random uniform index within $[N+1]$. Then,*

$$\mathbb{E}_i[g_W(R_{-i}, z_i)] \le \frac{4}{T} + \frac{d}{N+1}.$$

Proof. Define for each $i \in [N+1]$ the quantity

$$E_i = \sum_{I \subseteq [N+1]\setminus\{i\}} \frac{T^{|I|}}{\binom{N}{|I|}} \mathbf{1}[W \triangleright (z_i, R_I)].$$

As in the previous sections, if either branch $W^{z_i\to b}$ shatters a tuple, then W shatters it; if both branches shatter it, then W shatters its extension by z_i:

$$\Theta^{R_{-i}}(W^{z_i\to 0}) + \Theta^{R_{-i}}(W^{z_i\to 1}) \le \Theta^{R_{-i}}(W) + E_i.$$

In particular, if $g_W(R_{-i}, z_i) = 1$, the previous inequality implies $E_i > \Theta^{R_{-i}}(W)/4$.

Next, we compare $\Theta^R(W)$ to $\Theta^{R_{-i}}(W)$, the potential where one sample was removed. Splitting terms according to whether they contain i gives

$$\begin{aligned}\Theta^R(W) &\le \Theta^{R_{-i}}(W) + \sum_{\substack{I\subseteq[N+1]\\ i\in I}} \frac{T^{|I|}}{\binom{N+1}{|I|}} \mathbf{1}[W \triangleright R_I] \\ &\overset{(i)}{\le} \Theta^{R_{-i}}(W) + \frac{d}{N+1}\sum_{\substack{I\subseteq[N+1]\\ i\in I}} \frac{T^{|I|}}{\binom{N}{|I|-1}} \mathbf{1}[W \triangleright R_I] = \Theta^{R_{-i}}(W) + \frac{dT}{N+1} E_i.\end{aligned}$$

In (i) we used the fact that shattered sets in $W \subseteq \mathcal{F}$ have size at most d. Therefore, if $g_W(R_{-i}, z_i) = 1$, we have

$$E_i > \frac{\Theta^R(W)}{4 + Td/(N+1)}.$$

On the other hand, we can bound the sum of the quantities E_i by counting each nonempty shattered subset once for each distinguished coordinate:

$$\begin{aligned}\sum_{i=1}^{N+1} E_i &= \sum_{\varnothing \neq I \subseteq [N+1]} |I| \frac{T^{|I|-1}}{\binom{N}{|I|-1}} \mathbf{1}[W \triangleright R_I] \\ &= \frac{N+1}{T} \sum_{\varnothing \neq I \subseteq [N+1]} \frac{T^{|I|}}{\binom{N+1}{|I|}} \mathbf{1}[W \triangleright R_I] \\ &= \frac{N+1}{T}\big(\Theta^R(W) - 1\big) \leq \frac{N+1}{T}\Theta^R(W).\end{aligned}$$

Together with the previous lower bound on E_i when $g_W(R_{-i}, z_i) = 1$, this shows that

$$|\{i \in [N+1] : g_W(R_{-i}, z_i) = 1\}| \leq \frac{N+1}{T}\left(4 + \frac{Td}{N+1}\right) = \frac{4(N+1)}{T} + d.$$

Dividing by $N+1$ ends the proof. □

Essentially, Lemma E.4 shows that each block of size N leads to a clean abstention error rate of order $O(1/N)$ for a fixed class W. Using the reduction dimension, we lift this abstention error rate for all classes $W \in \mathcal{R}_{\ell_T}$ constructed using at most ℓ_T constraints. This is sufficient since all version spaces arising during the run of the matching reference expert have at most ℓ_T constraints. Using a few blocks m precisely ensures a sufficiently small failure probability so that we can apply uniform concentration over all such classes in $\mathcal{R}_{\ell_T}$.

Lemma E.5 (Uniform block certificate)**.** *There is a universal constant $C_m > 0$ such that the following holds. Let $\delta \in (0,1)$, $N \geq 1$, $\ell \geq 0$, and suppose that $m \geq C_m(D\log(2ND) + \log\frac{2}{\delta})$ where D is the ℓ-reduction dimension of $\mathcal{F}$. Sample independently m blocks $S_1, \ldots, S_m \sim \mu^N$ of N samples each. Next, define*

$$Q_\ell(S_1, \ldots, S_m) := \sup_{\emptyset \neq W \in \mathcal{R}_\ell} \mu\left\{x : \frac{1}{m}\sum_{j=1}^{m} g_W(S_j, x) > \frac{1}{6}\right\}.$$

Then,

$$\mathbb{E}_{S_1,\ldots,S_m}[Q(S_1, \ldots, S_m)] \leq \frac{32}{T} + \frac{8d}{N+1} + \delta.$$

Proof. On r pairs (S, x), the indicator vector $(g_W(S, x))$ for classes $W \in \mathcal{R}_\ell$ is determined by the trace of such W on at most $r(N+1)$ coordinates. The counting argument from Lemma E.1 therefore bounds the VC dimension of this class by $C_1 D\log(2ND)$ for some universal $C_1 > 0$. Then, uniform convergence at the constant gap $1/24$ shows that for an appropriate choice of the constant C_m, for each fixed $x_0 \in \mathcal{X}$, with probability at least $1 - \delta$ over $S_1, \ldots, S_m$,

$$\sup_{W \in \mathcal{R}_\ell} \left| \frac{1}{m}\sum_{j \in [m]} g_W(S_j, x_0) - \mathbb{E}_{S \sim \mu^N}[g_W(S, x_0)] \right| < \frac{1}{24}.$$

Additionally taking expectation over $x \sim \mu$ gives,

$$\mathbb{E}_{S_1,\ldots,S_m}[\mu(\mathcal{B})] \leq \delta,$$

where

$$\mathcal{B} := \left\{ x : \begin{array}{c} \exists W \in \mathcal{R}_\ell, \quad \mathbb{E}_{S \sim \mu^N}[g_W(S, x)] \leq \frac{1}{8} \text{ and} \\ \frac{1}{m}\sum_{j \in [m]} g_W(S_j, x) > \frac{1}{6} \end{array} \right\}.$$

Next, since i.i.d. samples from μ are exchangeable, Lemma E.4 shows that for any non-empty class W,

$$\mathbb{E}_{S\sim\mu^N, x\sim\mu}[g_W(S,x)] \le \frac{4}{T} + \frac{d}{N+1}.$$

In particular, Markov's inequality gives for every non-empty class $W \in \mathcal{R}_\ell$,

$$\begin{aligned}\mu\left\{\frac{1}{m}\sum_{j\in[m]} g_W(S_j,x) > \frac{1}{6}\right\} &\le \mu\left(\mathcal{B}\cup\left\{x : \mathbb{E}_{S\sim\mu^N}[g_W(S,x)] > \frac{1}{8}\right\}\right)\\ &\le \frac{32}{T} + \frac{8d}{N+1} + \mu(\mathcal{B}).\end{aligned}$$

Taking the supremum over all non-empty $W \in \mathcal{R}_\ell$ then the expectation over $S_1,\dots,S_m$ ends the proof, since $\mathbb{E}[\mu(\mathcal{B})] \le \delta$. □

We are now ready to bound the abstention error of the matching reference expert.

Proposition E.1. *There is a universal constant $C_m > 0$ such that the following holds. Let $N \ge 1$ and $\delta \in (0,1)$, and suppose that $\mathcal{F}$ has ℓ_T-reduction dimension D_T. If*

$$m \ge C_m\big(D_T \log(2TD_T) + \log(2/\delta)\big),$$

then the matching reference expert makes at most M_T mistakes on every realized interaction and has expected clean-abstention error at most

$$\frac{8dT}{N+1} + Nm + 32 + \delta T.$$

Proof of Proposition E.1. We already bounded the misclassification error within Lemma E.3. We now focus on the abstention error.

To simplify the analysis, we view all clean samples as coming from an infinite sequence of i.i.d. draws $Z_1, Z_2, \dots \overset{iid}{\sim} \mu$: the sample observed on the i-th clean round is Z_i.

Using the first mN elements of the sequence $(Z_i)_{i\ge1}$ we construct the m blocks $S_1,\dots,S_m$. Importantly, we define these even if fewer than mN clean rounds occur during the interaction. Now focus on any clean round $t \ge t_0$ (recall that t_0 is the time immediately after the mN-th clean round that occurred). For convenience, we denote by V_t the version space at the beginning of round t and by $\mathcal{H}_t$ the history up to time t included. Using Lemma E.3, we can bound the probability of a bad abstention on round t by

$$\begin{aligned}\mathbb{P}[\widehat{y}_t = \perp \mid \mathcal{H}_{t-1}, c_t = 0] &\le \mathbb{P}\left[\frac{1}{m}\sum_{j\in[m]} g_{V_t}(S_j,x_t) > \frac{1}{6} \mid \mathcal{H}_{t-1}, c_t = 0\right]\\ &= \mu\left\{x : \frac{1}{m}\sum_{j\in[m]} g_{V_t}(S_j,x) > \frac{1}{6}\right\} \le Q_{\ell_T}(S_1,\dots,S_m). \qquad (19)\end{aligned}$$

In the last inequality we used the fact that $V \in \mathcal{R}_{\ell_T}$. Before time t_0, the matching reference expert always abstains, incurring abstention error at most mN. We next use the notation $c(T) = |\{t \ge t_0 : c_t = 0\}|$ for the number of clean samples after the start time t_0. Starting from t_0, the expected abstention error is then

$$\begin{aligned}\mathbb{E}\left[\sum_{t\ge t_0}^{T} \mathbf{1}[\widehat{y}_t = \perp]\mathbf{1}[c_t = 0]\right] &\le \mathbb{E}[c(T) Q_{\ell_T}(S_1,\dots,S_m)] \qquad (20)\\ &\le T\mathbb{E}[Q_{\ell_T}(S_1,\dots,S_m)]\\ &\le \frac{8dT}{N+1} + 32 + \delta T,\end{aligned}$$

where in the last inequality we used the assumption on m and Lemma E.5.

Adding the two contributions to the expected abstention error proves the desired bound. □

E.3 Direct Aggregation over Candidate Experts

Write $L := |\mathcal{C}|$, $R := 1 + \log L + (\log 2)M$, and $\gamma := R/(2\kappa T)$. Assume $R \le 2\kappa T$, so that $\gamma \le 1$.

Proof of Proposition 5.1. The comparison uses the total expert weight in two directions. Learner mistakes give an upper bound by contracting the total weight, while the weight of one expert gives a lower bound and records rounds on which the learner abstains but that expert predicts.

Suppose first that the learner predicts and is wrong. The true label has minimum aggregate mass P_{y_t}, possibly tied, and this mass is at most $(1-\kappa)\Phi_t$. The complementary weight, consisting exactly of experts that predicted the learner's wrong label, is at least $\kappa\Phi_t$. Halving those weights gives

$$\Phi_{t+1} \le \left(1 - \frac{\kappa}{2}\right)\Phi_t \le e^{-\kappa/2}\Phi_t.$$

If the learner abstains, the weight predicting either label is less than $\kappa\Phi_t$. The total weight of experts that predict is therefore less than $2\kappa\Phi_t$, which gives

$$\Phi_{t+1} \le (1 + 2\gamma\kappa)\Phi_t \le e^{2\gamma\kappa}\Phi_t.$$

On a correct learner prediction, the total weight cannot increase.

For an expert c, let $D_c := |\{t : \widehat{y}_t = \perp,\ \widehat{y}_{c,t} \ne \perp\}|$, and let $\mathsf{MisErr}^{(c)}$ be the total number of its mistakes. Since $\Phi_1 = 1$, iterating the one-round bounds and using at most T learner abstentions gives

$$\Phi_{T+1} \le \exp\left(2\gamma\kappa T - \frac{\kappa}{2}\mathsf{MisErr}\right).$$

On the other hand, the two updates give the following lower bound on expert c's final weight. The expert is rewarded whenever the learner abstains and it predicts; it is penalized on at most $\mathsf{MisErr}^{(c)}$ rounds, because labels are used only when the learner predicts:

$$w_{c,T+1} \ge \frac{1}{L}(1+\gamma)^{D_c} 2^{-\mathsf{MisErr}^{(c)}}.$$

Combining $w_{c,T+1} \le \Phi_{T+1}$ with these two bounds gives

$$\frac{\kappa}{2}\mathsf{MisErr} + \log(1+\gamma)D_c \le \log L + (\log 2)\mathsf{MisErr}^{(c)} + 2\gamma\kappa T. \tag{21}$$

If $\mathsf{MisErr}^{(c)} \le M$, the right side is at most $2R$. Dropping the second term on the left gives $\mathsf{MisErr} \le 4R/\kappa$. Dropping the first term instead and using $\log(1+\gamma) \ge \gamma/2$ gives $D_c \le 4R/\gamma = 8\kappa T$. Finally, on every round in U where the learner abstains, either expert c also abstains or the round contributes to D_c. Thus the learner abstains on at most D_c more rounds of U than expert c, which proves the simultaneous comparison. □

E.4 Completing the Adaptive Theorem Proofs

We first give the full construction of the multi-phase reference expert outlined in the main text. Then, we implement the comparison in the censored-feedback model and derive the tradeoff in Theorem 5.1.

E.4.1 Multi-phase reference expert

The reference expert in Algorithm 3 abstains until it observes a large enough clean bank of size Nm. This results in an Nm abstention error term as shown in Proposition E.1. To further remove this term, instead of simply abstaining on this phase, we can use a similar reference expert which uses a smaller clean bank. We repeat this idea over at most $J := \lceil \log_2 T \rceil$ phases, using the compression to successively increase the clean bank size throughout the interaction. We denote the number of phases by $h \le J$. We refer to this as the *multi-phase reference expert* as defined in Algorithm 6.

Algorithm 6 Multi-phase reference expert

1: **Require:** Class $\mathcal{F}$, phase start-times $\tau_0 < \cdots < \tau_{h-1} < \tau_h := T+1$, witness-index lists $(\mathcal{Z}_{j,r})_{j<h,\, 0\le r\le M_T}$
2: Abstain $\widehat{y}_t \leftarrow \perp$ for all $t < \tau_0$
3: **For** each phase $j = 0, \dots, h-1$:
4: **For** rounds $t \in \{\tau_j, \dots, \tau_{j+1} - 1\}$:
5: Run reference expert Algorithm 3 for $\mathcal{F}$, start time τ_j, and witness-index list $(\mathcal{Z}_{j,r})_{0\le r\le M_T}$

The multi-phase reference process and its matching multi-phase reference expert. Similar to the reference process one can define the *multi-phase reference process*. The process first abstains until it has observed md clean instances. For $j \ge 0$, let $n_j = md2^j$ and $N_j = d2^j$. Let τ_j be the round immediately after the n_j-th clean round. Phase j starts on round τ_j immediately after the n_j-th clean round. On each phase, it runs the reference process from Appendix E.1 using the $n_j = mN_j$ clean bank already available, until the next phase starts. In particular, the version space is restarted at the beginning of each phase.

Since each phase of this multi-phase process runs the reference process, in hindsight, there exists a *matching multi-phase reference expert* that behaves the same as the multi-phase process, under the right choice of phase start times $\tau_0, ..., \tau_h$ as well as witness-index lists — those choices basically come from the reference processes on all h phases.

To be more specific, on each phase j the matching multi-phase reference expert coincides with the matching reference expert starting at time τ_j, as they share the witness-index lists $(\mathcal{Z}_{j,r})_{0\le r\le M_T}$ on that phase. Hence, the misclassification and abstention error bounds from Proposition E.1 still apply on each phase separately: on phase j, the matching multi-phase expert makes at most M_T mistakes, and an abstention error roughly $O(d(n_{j+1} - n_j)/N_j) = O(dm)$. Summing over all phases gives the following guarantee for the matching multi-phase reference expert:

Proposition E.2 (Clean reference process)**.** *Let $\delta \in (0,1)$ and suppose that $\mathcal{F}$ has ℓ_T-reduction dimension D_T. If $m \ge C_m(D_T \log(2TD_T) + \log(2/\delta))$, then the multi-phase reference process has at most JM_T mistakes on every realized interaction and expected abstention error at most*

$$(8dm + 32 + \delta T)J + md. \tag{22}$$

Proof of Proposition E.2. On each phase j, the matching multi-phase reference expert coincides with the matching reference expert which starts at time τ_j and uses the parameters m and N_j. Hence, Proposition E.1 already shows that it makes at most M_T mistakes on that phase. Since there are at most J phases, this results in at most JM_T misclassification error.

For the abstention error, the same arguments as in the proof of Proposition E.1 (see Equations (19) and (20)) show that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbf{1}[\widehat{y}_t = \perp]\mathbf{1}[c_t = 0]\right] &\le \mathbb{E}\left[\sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbf{1}[c_t = 0] Q_{\ell_T}(S_1^{(j)}, \dots, S_m^{(j)})\right] \\
&\le \min(n_{j+1} - n_j, T)\mathbb{E}[Q_{\ell_T}(S_1^{(j)}, \dots, S_m^{(j)})] \\
&\le \frac{8d(n_{j+1} - n_j)}{N_j + 1} + 32 + \delta T \le 8dm + 32 + \delta T,
\end{aligned}$$

where $S_1^{(j)}, \dots, S_m^{(j)}$ denote the N_j-size clean-bank blocks used for phase j. Summing over at most J phases and recalling that the multi-phase expert abstains for the first md clean rounds, gives the desired expected abstention error bound. □

Let $\mathcal{C}'$ be the family of all admissible multi-phase candidates, including the always-abstaining candidate $h = 0$. A direct counting gives

$$|\mathcal{C}'| \le (T+1)^{J + Jds(M_T+1)}. \tag{23}$$

In particular, after any realized interaction and in hindsight, one member of this family agrees with the multi-phase reference process defined above.

E.4.2 Extension to censored feedback

As discussed in Appendix E.3, the weight updates of Algorithm 4 use the label only when the aggregate predicts. Its candidate experts, however, may need to update after mistakes on rounds when the aggregate abstains and receives no label. Following the encoding argument of Yu & Blanchard (2026b), we include these updates in the description of each candidate.

Up to increasing the number of experts, we may consider that the multi-phase reference experts are given their phase start-times $\{\tau_j\}_{j<h}$ and witness-index lists $(\mathcal{Z}_{j,r})_{j<h,\,0\le r\le M_T}$ together with a set of at most JM_T rounds $S \subseteq [T]$ and corresponding labels $z_S \in \{0,1\}^S$. The expert then runs the previous multi-phase reference expert Algorithm 6 but only updates the version space V at times $t \in S$ using the prescribed label z_t. Formally, if $t \in S$ it updates $V \leftarrow V^{x_t \to z_t}$. In particular, the new experts do not rely on observations at all and hence can be implemented in the censored model. For convenience, denote by $\widetilde{\mathcal{C}}$ the collection of all these experts. Due to the choice of rounds S and labels z_S we have

$$|\widetilde{\mathcal{C}}| \le |\mathcal{C}'|(2T+1)^{JM_T} \le (T+1)^{J+Jds(M_T+1)}(2T+1)^{JM_T}.$$

Since the matching multi-phase reference expert from Proposition E.2 makes at most JM_T mistakes, one of the previous censored experts emulates it, enjoying the same mistake and abstention guarantee. We can then apply directly the guarantee from the direct aggregation algorithm Proposition 5.1 with the only difference that $L = |\widetilde{\mathcal{C}}|$. This increases the term R by an extra $JM_T \log(2T)$ additive term as compared to the full-feedback case. With the parameters above, $JM_T \log(2T) = O(\log|\mathcal{C}'|)$. Hence, this modification only affects the results up to a constant factor.

E.4.3 Deriving the adaptive tradeoff

We are now ready to prove the main result—which is essentially putting together the multi-phase reference expert guarantee Proposition E.2 and the direct aggregation guarantee Proposition 5.1.

Theorem E.1 (Adaptive guarantee with an abstention-rate parameter)**.** *There is a universal constant $C > 0$ such that the following holds for every $T \ge 1$ and $\epsilon \in (0,1]$. Let $\mathcal{F}$ have VC dimension $d \ge 1$, and set $\ell_T := \lceil d\log(T+1)/\log(16/15)\rceil + 1$. Suppose that $\mathcal{F}$ has ℓ_T-reduction dimension $D_T < \infty$. Then there is a learner, without knowledge of μ, that satisfies against any adaptive adversary with censored feedback*

$$\mathsf{MisErr} \le C\frac{d^2 D_T \log^4(2T)}{\epsilon}, \qquad \mathbb{E}[\mathsf{AbsErr}] \le \epsilon T + CdD_T \log^2(2T).$$

The mistake bound holds on every realized interaction.

Proof of Theorem E.1. The cases $T = 1$, $d > T$, or $D_T > T$ can be handled directly by either always abstaining (0 misclassification) or always predicting (0 abstention). We assume $T \ge 2$ and $1 \le d, D_T \le T$ from now on. We choose $\delta = 1/T$ and $m = \lceil C_m\big[D_T \log(2TD_T) + \log(2T)\big]\rceil$. For the direct aggregation procedure, we use the parameter $\kappa = \epsilon/8$ and the family of experts $\mathcal{C}$ obtained for any choice of start times $\tau_0 < \ldots < \tau_{h-1} \le T$ for $h \le J = \lceil \log_2 T\rceil$ phases, and corresponding witness-index lists $(\mathcal{Z}_{j,r})_{j<h,0\le r\le M_T}$ which each contain $s = \lceil C_s D_T \log(2dD_T)\rceil$ tuples of at most d indices in $[T]$. In particular, $L := |\widetilde{\mathcal{C}}| \le (T+1)^{J+J(M_T+1)ds}(2T+1)^{JM_T}$. We also use the parameter $\gamma = R/(2\kappa T)$, where $R := 1 + \log L + (\log 2)JM_T = O(d^2 D_T \log^4 T)$. Note that if $R \ge 2\kappa T = \varepsilon T/4$, the desired misclassification bound can trivially be achieved by always predicting arbitrarily. We then suppose that $R \le 2\kappa T$ from now.

By Proposition E.2, we can bound the misclassification and abstention errors $M_{\mathrm{ref}}, A_{\mathrm{ref}}$ of the multi-phase reference expert by

$$M_{\mathrm{ref}} \le JM_T \le Cd\log^2 T \quad \text{and} \quad \mathbb{E}A_{\mathrm{ref}} \le CdD_T \log^2 T.$$

for some universal constant $C > 0$. Then, by Proposition 5.1, the direct aggregation guarantee gives

$$\mathrm{MisErr} \le 32R/\varepsilon \quad \text{and} \quad \mathbb{E}\,\mathrm{AbsErr} \le \mathbb{E}[A_{\mathrm{ref}} + 8\kappa T] \le \varepsilon T + CdD_T \log^2 T.$$

This ends the proof. □

Proof of Theorem 5.1. Write $K := d^2 D_T \log^4(2T)$ and $B := dD_T \log^2(2T)$, and fix $\alpha \in [0, 1]$. If $K \geq T^\alpha$, always abstaining gives $\mathsf{MisErr} = 0$ and $\mathbb{E}[\mathsf{AbsErr}] \leq T \leq KT^{1-\alpha}$. Otherwise apply Theorem E.1 with $\epsilon := K/T^\alpha \in (0, 1)$. Then

$$\mathsf{MisErr} \leq CT^\alpha, \qquad \mathbb{E}[\mathsf{AbsErr}] \leq KT^{1-\alpha} + CB \lesssim KT^{1-\alpha},$$

because $B \lesssim K$ and $T^{1-\alpha} \geq 1$. Both cases imply the asserted bounds $\mathsf{MisErr} \lesssim T^\alpha + d^2 D_T \log^4(2T)$ and $\mathbb{E}[\mathsf{AbsErr}] \lesssim d^2 D_T T^{1-\alpha} \log^4(2T)$. □