Active Learning From Positive and Unlabeled Examples

Farnam Mansouri
David Cheriton School of Computer Science
University of Waterloo & Vector Institute
Waterloo, ON, Canada
f5mansou@uwaterloo.ca

Shai Ben-David
David Cheriton School of Computer Science
University of Waterloo & Vector Institute
Waterloo, ON, Canada
shai@uwaterloo.ca

Sandra Zilles
Department of Computer Science
University of Regina & AMI Institute
Regina, SK, Canada
sandra.zilles@uregina.ca

Abstract

Learning from positive and unlabeled data (PU learning) is a weakly supervised variant of binary classification in which the learner receives labels only for (some) positively labeled instances, while all other examples remain unlabeled. Motivated by applications such as advertising and anomaly detection, we study *active* PU learning, where the learner adaptively queries instances from an unlabeled pool, but a label is revealed only when the queried instance is positive and an independent coin flip succeeds; otherwise no label is revealed. We give finite label complexity bounds for active PU learning when the positive-label reveal probability is unknown.

1 Introduction

Learning from positive and unlabeled data (PU learning) is a weakly supervised variant of binary classification in which the training data only consists of positively labeled and unlabeled examples. PU learning appears in many real-world problems, including personal advertisement, land cover classification (Li et al., 2010), prediction of protein similarity (Elkan and Noto, 2008), knowledge base completion (Bekker and Davis, 2020) and disease-gene identification (Yang et al., 2012).

In active learning, a learning algorithm is given access to a large pool of unlabeled examples, and is allowed to request the label of any particular example from that pool. This is in contrast with standard passive learning, where the data are assumed to be collected randomly and labeled independently of the learner's choices. The objective of active learning is to learn an accurate classifier while requesting as few labels as possible, thus in particular reducing the workload of human annotators by carefully selecting the examples from the unlabeled pool that should be labeled.

Just as in standard machine learning, the labeling process can also be difficult for many PU learning applications. For example, in recommender systems, labels are often collected based on whether a recommendation is relevant (a positive outcome), while non-interactions are typically ambiguous and treated as unlabeled. Similarly, in anomaly detection, since anomalies are rare, the learner may only have access to data labeled as normal (which we view as the positive class). In both settings, obtaining labels can be costly, and may potentially benefit from more carefully selecting which examples from the unlabeled pool to query (Perini et al., 2020; Vercruyssen et al., 2018).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We therefore study *active PU learning*, in which the learner may adaptively query instances, but the label request is given to an expert that is only able to detect positive labels for a portion of instances. We make the simplifying assumption that the expert detects each distinct positive instance independently with probability ω. The equivalent *selected-completely-at-random* (SCAR) assumption is often employed in passive PU learning (Liu et al., 2002; Blanchard et al., 2010; Du Plessis et al., 2015; Bekker and Davis, 2018; Mansouri and Ben-David, 2025).

Since the objective of active learning is to identify an accurate classifier with as small a number of queries as possible, a standard measure of evaluation in active learning is the so-called *label complexity*, i.e., the number of label requests that are necessary and sufficient to learn an accurate classifier. In this paper, we give a disagreement-based label complexity analysis for active PU learning with unknown ω. Our bounds depend critically on the *disagreement coefficient* θ, a standard quantity in the theory of active learning (Dasgupta et al., 2007; Hanneke, 2007, 2009, 2011, 2014). Our bounds depend on the positive class prior $\pi_{\mathcal{D}}$ even when the learner is given neither $\pi_{\mathcal{D}}$ nor ω. Ignoring the dependence on ω and $\pi_{\mathcal{D}}$, our bounds incur only a multiplicative θ gap compared to established label complexity bounds for the classical CAL algorithm in standard active learning (Cohn et al., 1994; Hanneke, 2009).

Our analysis assumes that every individual instance has probability zero. This holds, for example, in settings where the instance distribution admits a density, as is common in work on active learning with regular decision boundaries or regression functions (Castro and Nowak, 2007; Wang, 2011; Locatelli et al., 2018; Kpotufe et al., 2022). In our setting the assumption ensures that fresh draws give distinct instances, so the feedback on them is independent conditional on those instances.

Related Work. While active learning has been extensively studied in the fully supervised setting, it has received only limited attention in the context of PU learning. Work on active PU learning and related one-class problems has explored uncertainty sampling methods, including Query-by-Committee, least confidence, margin-based, and entropy-based sampling (Abe et al., 2006; Görnitz et al., 2009; Ghasemi et al., 2011b; Schlachter and Yang, 2018). Prior work has also considered strategies that prioritize queries that are most likely to be positive (He et al., 2006; Ghasemi et al., 2011a). These methods focus on which examples to query. The feedback can differ from ours. For example, Schlachter and Yang (2018) use confirmed examples of both classes to train separate one-class classifiers. Perini et al. (2020) use instance-dependent propensities induced by a given active strategy to estimate the class prior. They show asymptotic convergence as the propensity estimates approach their true values. We study how many queries are needed to learn a classifier when the reveal probability is constant and unknown.

When the reveal probability is known, the feedback can be transformed into symmetric label noise. Existing bounds for noisy active learning then apply, with a linear dependence on θ and a quadratic dependence on $1/\omega$ (Hanneke, 2011; Hanneke and Yang, 2015). Appendix C gives the reduction and compares the bounds.

2 Preliminaries

Active Learning. Let $\mathcal{X}$ be a *domain set*. The *hypothesis class* $\mathcal{H}$ is a set of functions $h : \mathcal{X} \to \{0, 1\}$. By abuse of notation, we also associate h with the set $\{x \in \mathcal{X} \mid h(x) = 1\}$. $\mathcal{D}$ denotes a distribution over $\mathcal{X}$ called the data generating distribution. We use $\ell : \mathcal{X} \to \{0, 1\}$ to denote an underlying labeling rule. In the study of active learning, there is a pool of data $\{x_1, x_2, \ldots\}$ sampled i.i.d. from $\mathcal{D}$. An active learner, at each time t, queries the label of an instance $x \in \mathcal{X}$ from the pool and receives $\ell(x)$. The goal of the learner is to output a function $f : \mathcal{X} \to \{0, 1\}$ that minimizes

$$\mathrm{err}_{\mathcal{D}}(f, \ell) = \mathbb{E}_{x\sim\mathcal{D}}[f(x) \neq \ell(x)]\,.$$

For any multiset $S \subseteq \mathcal{X}$, define

$$\hat{\mathrm{err}}_S(f, \ell) := \frac{1}{|S|} \sum_{x\in S} \mathbb{1}\{f(x) \neq \ell(x)\}\,.$$

Also, define $\hat{\Pr}_S$ as the empirical distribution induced by S, i.e., for any event $A \subseteq \mathcal{X}$, $\hat{\Pr}_S[A]$ is defined as the fraction of elements in the multiset S that also belong to A. Formally,

$$\hat{\Pr}_S[A] := |\{x \in S : x \in A\}|/|S|\,,$$

where both numerator and denominator account for elements in S with their multiplicity in S.

Definition 2.1. *The pseudo-metrics $\rho_{\mathcal{D}}, \rho_S : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \Re^{\geq 0}$, are defined by, for all $f, g \in 2^{\mathcal{X}}$:*

$$\rho_{\mathcal{D}}(f,g) := \Pr_{x\sim\mathcal{D}}[f(x) \neq g(x)] \;\; \textit{and} \;\; \rho_S(f,g) := \hat{\Pr}_S[f(x) \neq g(x)] \,.$$

Definition 2.2 (Hanneke (2007))**.** *The region of disagreement of any set $V \subseteq \mathcal{H}$ is defined as*

$$\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h_1, h_2 \in V \textit{ s.t. } h_1(x) \neq h_2(x)\} \,.$$

Definition 2.3 (Hanneke (2007))**.** *The* disagreement rate*, and the* empirical disagreement rate*, respectively, of any set $V \subseteq \mathcal{H}$ are defined as*

$$\Delta_{\mathcal{D}}(V) := \Pr_{x\in\mathcal{D}}[x \in \mathrm{DIS}(V)] \;\; \textit{and} \;\; \hat{\Delta}_S(V) := \hat{\Pr}_S[\mathrm{DIS}(V)] \,.$$

Definition 2.4. *Let $f : \mathcal{X} \to \{0,1\}$, $r > 0$, and $S \subseteq \mathcal{X}$ a multiset. The r-*ball around *f with respect to $(\mathcal{D}, \mathcal{H})$ and the* empirical *r-*ball around *f with respect to $(S, \mathcal{H})$, resp., are given by*

$$\mathcal{B}_{\mathcal{D}}(f,r) = \{h \in \mathcal{H} : \rho_{\mathcal{D}}(f,h) \leq r\} \textit{ and } \hat{\mathcal{B}}_S(f,r) = \{h \in \mathcal{H} : \rho_S(f,h) \leq r\} \,.$$

Definition 2.5 (Hanneke (2007))**.** *The disagreement coefficient is defined as*

$$\theta = \sup_{h\in\mathcal{H}, r>0} \frac{\Delta_{\mathcal{D}}(\mathcal{B}_{\mathcal{D}}(h,r))}{r} \,.$$

Passive PU Learning. In passive PU learning, the learner has access only to positive examples and unlabeled data. We state an existing result for this setting, which we will use later:

Theorem 2.6 (Liu et al. (2002))**.** *Let d be the VC dimension of $\mathcal{H}$. Then there is a constant $M_1 > 1$ such that for any $\varepsilon, \delta > 0$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ that is realized by $\mathcal{H}$, if*

$$k \;\geq\; M_1 \frac{d\ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon},$$

then the following holds. Let S^U be an unlabeled sample of size k drawn i.i.d. from $\mathcal{D}$, and let S^P be a positive sample of size k drawn i.i.d. from $\mathcal{D}(\cdot \mid y = 1)$. Define

$$h^{PU} \;:=\; \arg\min_{h\in\mathcal{H}(S^P)} \left|h \cap S^U\right| \;\textit{ s.t. }\; \mathcal{H}(S^P) \;:=\; \{h \in \mathcal{H} :\, \forall x \in S^P,\; h(x) = 1\} \,.$$

Then with probability at least $1 - \delta$, we have $\mathrm{err}_{\mathcal{D}}(h^{PU}, \ell) \leq \varepsilon$.

3 Active PU Learning Setting

Fix $\omega \in (0,1]$. For every instance x with $\ell(x) = 0$, set $s(x) = 0$. For each distinct instance x with $\ell(x) = 1$, draw $s(x)$ independently from a Bernoulli distribution with mean ω. When the learner queries x from the unlabeled pool, it receives feedback 1 if $s(x) = 1$ and $\star$ otherwise. The feedback is fixed for each instance, so querying it again returns the same response. This models an expert who may consistently miss a positive instance. Our algorithms do not require knowledge of ω.

Denote $\pi_{\mathcal{D}} := \mathcal{D}\big(\{x \in \mathcal{X} : \ell(x) = 1\}\big)$, and, for any multiset $S \subseteq \mathcal{X}$, let $\hat{\pi}_S := \frac{|\{x\in S : \ell(x)=1\}|}{|S|}$. Moreover, define the positive disagreement region of $\mathcal{H}$ as $\mathrm{DIS}^P(\mathcal{H}) := \{x \in \mathrm{DIS}(\mathcal{H}) : \ell(x) = 1\}$. Throughout this paper, we focus on the realizable setting, so we assume $\ell \in \mathcal{H}$. The VC dimension of $\mathcal{H}$ is denoted by $\mathrm{VCD}(\mathcal{H}) = d$. We take $\pi_{\mathcal{D}} > 0$ in the unknown-prior analysis, since estimating ω requires some positive instances. The algorithms that use θ can take any supplied finite upper bound on the disagreement coefficient that is at least one. Their query bounds then use that supplied value.

Our proof techniques rely on the following continuity assumption, which rules out repeated draws of the same instance almost surely.

Assumption 3.1. *$\mathcal{D}$ has zero mass on singletons, i.e., $\mathcal{D}(\{x\}) = 0$ for every $x \in \mathcal{X}$.*

Remark 3.2. *Let $\mathcal{D}$ be a distribution that satisfies Assumption 3.1. Let S be an i.i.d. sample of $\mathcal{D}(\cdot \mid x \in A)$ for $A \subseteq \mathcal{X}$. Then conditional on its size, the revealed subset $S' = \{x \in S : s(x) = 1\}$ is an i.i.d. sample from $\mathcal{D}(\cdot \mid x \in A, \ell(x) = 1)$ whenever it is nonempty.*

Algorithm 1 Active PU learning with known $\pi_{\mathcal{D}}$

Input: $\mathcal{H}$, $\pi_{\mathcal{D}}$, θ, accuracy ε, confidence δ.
If $d = 0$, **return** any $h \in \mathcal{H}$.
$\gamma \leftarrow \varepsilon/(64\theta)$.
Draw S of size $\lceil C\big(d\ln(2/\gamma) + \ln(8/\delta)\big)/\gamma^2 \rceil$ from $\mathcal{D}$.
$V_0 \leftarrow \{h \in \mathcal{H} : \hat{\Pr}_S[h] \le \pi_{\mathcal{D}} + \gamma\}$.
If $V_0 = \varnothing$, **return** any $h \in \mathcal{H}$.
for $t = 0, 1, \ldots$ **do**
  Draw a fresh sample S'_t of size $\lceil C\ln(16(t+1)^2/\delta)/\varepsilon \rceil$ from $\mathcal{D}$.
  if $\hat{\Delta}_{S'_t}(V_t) \le \varepsilon/2$ **then**
    return any $\hat{h} \in V_t$.
  end if
  Rejection sample x from $\mathcal{D}$ conditioned on $\mathrm{DIS}(V_t)$ and query it.
  If the feedback is 1, set $V_{t+1} \leftarrow \{h \in V_t : h(x) = 1\}$.
  If the feedback is $\star$, set $V_{t+1} \leftarrow V_t$.
end for

4 Analysis with Known $\pi_{\mathcal{D}}$

First, we study active PU learning for the case when $\pi_{\mathcal{D}}$ is known. We propose Algorithm 1, which is simply the CAL algorithm of Cohn et al. (1994) applied to the restricted hypothesis class consisting of all $h \in \mathcal{H}$ whose predicted positive mass over an unlabeled sample is approximately at most $\pi_{\mathcal{D}}$.

In Theorem 4.2 we analyze the label complexity of this setup. The proof is inspired by Hanneke's (2009) analysis of the CAL algorithm for active learning. The key ideas behind the analysis are as follows: (i) The initial pruning of the hypothesis class ensures that, for every h,

$$\Pr[\ell(x) = 0, h(x) = 1] - \Pr[\ell(x) = 1, h(x) = 0] = \Pr[h(x) = 1] - \Pr[\ell(x) = 1] \lessapprox 0.$$

The false positive mass is at most the false negative mass plus the pruning tolerance. Thus, using positively labeled data, we can bound the false negative rate with standard PAC bounds. Consequently, also the total error can be bounded. (ii) Using this with an argument similar to Hanneke's (2009), we can show that after every $\tilde{O}(d\theta)$ instances with labels revealed, the disagreement rate $\Delta_{\mathcal{D}}(V_t)$ can be cut in half with high probability. (iii) Using Lemma 4.1 (which is proven in the appendix), we guarantee that the proportion of positively labeled points in $\mathrm{DIS}(V_t)$ has a lower bound of $\Omega(1/\theta)$ whenever another query is needed. Thus the response rate is at least $\Omega(\omega/\theta)$.

Lemma 4.1. *Let $\varepsilon > 0$. Suppose $V \subseteq \mathcal{B}_{\mathcal{D}}(\mathbf{0}, \pi_{\mathcal{D}} + \frac{\varepsilon}{4})$, $\ell \in V$ and $\sup_{h \in V} \mathrm{err}_{\mathcal{D}}(h, \ell) > \varepsilon$. Then*

$$\Pr[\ell(x) = 1 \mid x \in \mathrm{DIS}(V)] \ge \frac{1}{4\theta}\,.$$

Theorem 4.2. *For a sufficiently large absolute constant C, for any $\varepsilon, \delta \in (0,1)$ and under Assumption 3.1, with probability at least $1 - \delta$ Algorithm 1 outputs $\hat{h}$ with $\mathrm{err}_{\mathcal{D}}(\hat{h}, \ell) \le \varepsilon$ and makes at most*

$$O\left(\frac{\theta^2 \ln(4/\varepsilon)}{\omega}\left[d\ln(e\theta) + \ln\left(\frac{\ln(4/\varepsilon)}{\delta}\right)\right]\right)$$

label requests.

Proof. By uniform convergence for VC classes (Haussler, 1992; Vapnik, 2006), the initial sample guarantees $|\hat{\Pr}_S[h] - \Pr_{\mathcal{D}}[h]| \le \gamma$ for every $h \in \mathcal{H}$ with probability at least $1 - \delta/4$, for a sufficiently large absolute constant C. On this event $\ell \in V_0$, and for every $h \in V_0$,

$$\Pr_{\mathcal{D}}[h(x) = 1, \ell(x) = 0] - \Pr_{\mathcal{D}}[h(x) = 0, \ell(x) = 1] = \Pr_{\mathcal{D}}[h] - \pi_{\mathcal{D}} \le 2\gamma.$$

Every update preserves ℓ.

The stopping test uses a fresh sample conditional on the history up to V_t. Chernoff bounds give two implications simultaneously for all tests, except on an event of probability at most $\delta/4$. The learner

stops if $\Delta_{\mathcal{D}}(V_t) \le \varepsilon/4$, and it continues if $\Delta_{\mathcal{D}}(V_t) > \varepsilon$. Each test violates these implications with conditional probability at most $\delta/(8(t+1)^2)$ for sufficiently large C.

Set $N = \lceil \log_2(4/\varepsilon) \rceil$ and

$$\lambda = \left\lceil C_1 \frac{\theta^2}{\omega} \big(d \ln(e\theta) + \ln(4N/\delta) \big) \right\rceil .$$

Consider a block of λ queries starting from a version space V at which the learner continues. Write $\Delta = \Delta_{\mathcal{D}}(V)$. On the stopping-test event, $\Delta > \varepsilon/4$. By the definition of θ, $\sup_{h \in V} \mathrm{err}_{\mathcal{D}}(h, \ell) \ge \Delta/\theta > \varepsilon/(8\theta)$. Since $2\gamma = \varepsilon/(32\theta)$, Lemma 4.1 with accuracy parameter $\varepsilon/(8\theta)$ implies $\Pr_{\mathcal{D}}(\ell(x) = 1 \mid x \in \mathrm{DIS}(V)) \ge 1/(4\theta)$.

Conditional on the history at the start of the block, consider the first λ subsequent points in the fixed region $\mathrm{DIS}(V)$ from the fresh pool stream used for rejection sampling. If the learner completes the block, it scans at least these points. A positive label revealed at a point it queries eliminates every inconsistent hypothesis. If it passes over a point because the current version space agrees there, all surviving hypotheses already agree with ℓ there. Thus all hypotheses surviving the block are consistent with the revealed positives among the first λ points in the fixed region.

Each of these points has reveal probability at least $\omega/(4\theta)$. Conditional on their number, the revealed positives are i.i.d. from $\mathcal{D}(\cdot \mid \mathrm{DIS}^P(V))$ by Remark 3.2. For sufficiently large C_1, a Chernoff bound and the realizable VC bound (Blumer et al., 1989) show that, with conditional failure probability at most $\delta/(2N)$, every surviving hypothesis has false negative error at most $1/(8\theta)$ under $\mathcal{D}(\cdot \mid \mathrm{DIS}^P(V))$. On that event a surviving h has false negative mass at most $\Delta/(8\theta)$ and hence

$$\mathrm{err}_{\mathcal{D}}(h, \ell) \le \frac{\Delta}{4\theta} + 2\gamma \le \frac{\Delta}{2\theta}.$$

The definition of θ then bounds the new disagreement mass by $\Delta/2$.

A union bound over the first N blocks gives halving in each completed block with total failure probability at most $\delta/2$. If the learner completes all N blocks, the disagreement mass is at most $\varepsilon/4$ and the stopping test ends the algorithm. At any successful stopping test the disagreement mass is at most ε, which bounds the error of every remaining hypothesis. □

5 Analysis with Unknown $\pi_{\mathcal{D}}$

In this section, we study active PU learning in the more general setting where $\pi_{\mathcal{D}}$ is unknown. For this case, we propose Algorithm 3, with the formal analysis presented in Theorem 5.5. When $\pi_{\mathcal{D}}$ is unknown, the learner is unable to restrict the version space to hypotheses whose predicted positive rate over a pool of unlabeled data is close to $\pi_{\mathcal{D}}$. Controlling the response rate of queries as done in Lemma 4.1 thus becomes impossible. The key idea in our analysis is to use a binary-search-style procedure to control the positive mass and restrict the version space.

First, Algorithm 2 estimates ω to within a constant factor. It uses revealed positives and an unlabeled sample to find a region where the remaining hypotheses all predict positive, then measures feedback on fresh instances from that region. We then draw an unlabeled sample S_1 to measure the positive mass predicted by each hypothesis. At iteration i, we keep only hypotheses whose predicted positive mass on S_1 is at most u_i, an upper bound on $\hat{\pi}_{S_1}$. We initialize $u_0 = 1$. A fresh unlabeled sample gives us a lower bound b_i on the mass of the region where all hypotheses in V_i predict positive. Since $\ell \in V_i$, this also gives a lower bound on $\pi_{\mathcal{D}}$.

If the feedback rate within $\mathrm{DIS}(V_i)$ is small, Lemma 5.4 and the feedback-rate test show that we can safely replace u_i by $(u_i + b_i)/2$. The gap $u_i - b_i$ can be halved only $O(\log(1/\varepsilon))$ times before the error is at most ε. On rounds where the test takes the other branch, the feedback rate is at least $\Omega(\omega/\theta)$ with high probability. A batch of $\widetilde{O}(d\theta^2/\omega)$ queries then supplies the $\widetilde{O}(d\theta)$ positive responses needed to halve $\Delta_{\mathcal{D}}(V_i)$. We also use a fresh unlabeled sample to check whether the disagreement mass is already small enough to stop.

Algorithm 2 $\mathrm{EstRate}(\mathcal{H}, \delta)$

Input: Hypothesis class $\mathcal{H}$, confidence parameter δ
$d_+ \leftarrow \max\{1, d\}$
for $j = 0, 1, \ldots$ **do**
  $k_j \leftarrow 2^j d_+$, $\delta_j \leftarrow \delta/[16(j+1)^2]$
  Query fresh i.i.d. instances from $\mathcal{D}$ until k_j labels are revealed
  Let R_j be the revealed positives and T_j the number of requests
  Sample a fresh unlabeled set S_j of size k_j from $\mathcal{D}$
  $V_j \leftarrow \{h \in \mathcal{H} : \forall x \in R_j,\ h(x) = 1\}$
  $h_j^* \leftarrow \mathrm{argmin}_{h \in V_j} |h \cap S_j|$
  $\gamma_j \leftarrow \frac{(M_1+M_2)(d_+ \ln(2k_j/d_+)+\ln(1/\delta_j))}{k_j}$
  $V_j \leftarrow V_j \cap \hat{\mathcal{B}}_{S_j}(h_j^*, 3\gamma_j)$
  Sample a fresh unlabeled set W_j of size T_j from $\mathcal{D}$
  $P_j \leftarrow \{x \in W_j : \forall h \in V_j,\ h(x) = 1\}$
  Query each instance in P_j and let r_j be the number of revealed labels
  if $r_j \geq \lceil 16 \ln(1/\delta_j) \rceil$ **then**
    Return: $r_j/|P_j|$
  end if
end for

Algorithm 3 Active PU learning with unknown $\pi_{\mathcal{D}}$

Input: $\mathcal{H}$, upper bound $\theta \geq 1$ on the disagreement coefficient, ε, δ.
If $d = 0$, **return** any $h \in \mathcal{H}$.
$\hat{\omega} \leftarrow \mathrm{EstRate}(\mathcal{H}, \delta/8)$
$u_0 \leftarrow 1$, $V_0 \leftarrow \mathcal{H}$, $b_{-1} \leftarrow 0$, $N \leftarrow 2\lceil \log_2(6/\varepsilon) \rceil$, $\tau \leftarrow \varepsilon/192$
$\lambda_1 \leftarrow \lceil 768\theta \ln(8N/\delta)/\hat{\omega} \rceil$
$\lambda_2 \leftarrow \lceil M_3\theta^2\big(d\ln(2\theta) + \ln(8N/\delta)\big)/\hat{\omega} \rceil$
Draw an unlabeled sample S_1 of size $\lceil C\big(d\ln(2/\tau) + \ln(8/\delta)\big)/\tau^2 \rceil$ from $\mathcal{D}$.
for $i = 0, \ldots, N$ **do**
  If $V_i = \varnothing$, **return** any $h \in \mathcal{H}$.
  Draw a fresh unlabeled sample T_i of size $\lceil \ln(32(N+1)/\delta)/(2\tau^2) \rceil$ from $\mathcal{D}$.
  $b_i \leftarrow \max\{b_{i-1}, \hat{\Pr}_{T_i}[\forall h \in V_i : h(x) = 1] - \tau\}$
  If $u_i - b_i \leq \varepsilon/6$, $\hat{\Delta}_{T_i}(V_i) \leq \varepsilon/3$, or $i = N$, **return** any $\hat{h} \in V_i$.
  Rejection sample λ_1 instances in $\mathrm{DIS}(V_i)$ from $\mathcal{D}$ and query them. Let R_1 contain those with feedback 1.
  if $|R_1|/\lambda_1 < \hat{\omega}/(128\theta)$ **then**
    $u_{i+1} \leftarrow (u_i + b_i)/2$
    $V_{i+1} \leftarrow \{h \in V_i : \forall x \in R_1,\ h(x) = 1\} \cap \hat{\mathcal{B}}_{S_1}(\mathbf{0}, u_{i+1})$
  else
    $u_{i+1} \leftarrow u_i$
    Rejection sample λ_2 instances in $\mathrm{DIS}(V_i)$ from $\mathcal{D}$ and query them. Let R_2 contain those with feedback 1.
    If $|R_2| < 2d$, **return** any $h \in V_i$.
    Draw $|R_2|$ fresh unlabeled instances S_2 from $\mathcal{D}$ conditioned on $\mathrm{DIS}(V_i)$.
    $W \leftarrow \{h \in V_i : \forall x \in R_2,\ h(x) = 1\}$
    If $W = \varnothing$, **return** any $h \in V_i$.
    $h_{i+1} \leftarrow \mathrm{argmin}_{h \in W} |h \cap S_2|$
    $\gamma_{i+1} \leftarrow (2M_1 + M_2)\big(d\ln(2|R_2|/d) + \ln(8N/\delta)\big)/|R_2|$
    $V_{i+1} \leftarrow W \cap \hat{\mathcal{B}}_{S_2}(h_{i+1}, \gamma_{i+1})$
  end if
end for

Theorem 5.1 (Haussler (1992))**.** *Suppose that* $\mathrm{VCD}(\mathcal{H}) = d$ *and let* $f : \mathcal{X} \to \{0, 1\}$*. Then, for* $0 < \alpha < 1$ *and* $0 < \nu \leq 8$*, a sample* S *of size* n *drawn i.i.d. from* $\mathcal{D}$ *satisfies*

$$\Pr\big[\exists h \in \mathcal{H} :\ |\rho_{\mathcal{D}}(h, f) - \rho_S(h, f)| > \alpha\big(\rho_{\mathcal{D}}(h, f) + \rho_S(h, f) + \nu\big)\big] \leq 8\left(\frac{16e}{\alpha\nu}\ln\frac{16e}{\alpha\nu}\right)^d e^{-\alpha^2\nu n/8}.$$

For $0 < \varepsilon < 1$, apply the theorem with $\alpha = 1/3$, $\nu = 2\varepsilon$, and $f = \mathbf{0}$ to $\mathcal{H}\Delta\mathcal{H} = \{h \oplus h' : h, h' \in \mathcal{H}\}$. For $d \geq 1$, this class has VC dimension $O(d)$ because its growth function is bounded by the square of that of $\mathcal{H}$, and Sauer's lemma bounds the latter by $(en/d)^d$ for $n \geq d$. If $d = 0$, all members of $\mathcal{H}$ agree pointwise and the following inequalities are immediate. This gives the following corollary.

Corollary 5.2. *For some constant $M_2 > 1$, if S is drawn i.i.d. from $\mathcal{D}$ with $|S| \geq \max\{1, d\}$, then with probability at least $1 - \delta$ we have (i) for all $h, h' \in \mathcal{H}$*

$$\rho_{\mathcal{D}}(h, h') \leq 2\rho_S(h, h') + \varepsilon \text{ and } \rho_S(h, h') \leq 2\rho_{\mathcal{D}}(h, h') + \varepsilon\,;$$

(ii) for all $r > 0$ and all $f \in \mathcal{H}$

$$\mathcal{B}_{\mathcal{D}}(f, r) \subseteq \hat{\mathcal{B}}_S\left(f, 2r + \varepsilon\right) \text{ and } \hat{\mathcal{B}}_S(f, r) \subseteq \mathcal{B}_{\mathcal{D}}(f, 2r + \varepsilon)\,, \tag{1}$$

where $\varepsilon = M_2 \frac{d_+ \ln(2|S|/d_+) + \ln(1/\delta)}{|S|}$ and $d_+ = \max\{1, d\}$.

Subsequently, M_1 and M_2 refer to the constants given in Theorem 2.6 and Corollary 5.2, resp.

Theorem 5.3. *Suppose Assumption 3.1 holds, $\pi_{\mathcal{D}} > 0$, and $\theta < \infty$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, Algorithm 2 returns an estimate in $[\omega/2, 2\omega]$ after making*

$$O\left(\frac{\bar{\theta}\big(d_+ \ln(e\bar{\theta}/\pi_{\mathcal{D}}) + \ln(8/\delta)\big)}{\pi_{\mathcal{D}}^2 \omega}\right)$$

label requests, where $\bar{\theta} = \max\{1, \theta\}$ and $d_+ = \max\{1, d\}$.

Lemma 5.4. *At an iteration of Algorithm 3, let $B_i = \Pr_{x\sim\mathcal{D}}[\forall h \in V_i : h(x) = 1]$. Suppose $\ell \in V_i$, $u_i - b_i > \varepsilon/6$, $b_i \leq B_i \leq b_i + 2\tau$, and $|\hat{\Pr}_{S_1}[h] - \Pr_{x\sim\mathcal{D}}[h(x) = 1]| \leq \tau$ for all $h \in \mathcal{H}$. If $\hat{\pi}_{S_1} > (u_i + b_i)/2$, then*

$$\Pr_{x\sim\mathcal{D}}[\ell(x) = 1 \mid x \in \mathrm{DIS}(V_i)] \geq \frac{1}{32\theta}.$$

The proofs of Theorem 5.3 and Lemma 5.4 are deferred to the appendix.

Theorem 5.5. *There exist constants $C, M_3 > 1$ such that for $\varepsilon, \delta \in (0, 1)$, under Assumption 3.1 and with $\pi_{\mathcal{D}} > 0$, Algorithm 3 with probability at least $1 - \delta$ outputs a hypothesis $\hat{h}$ with $\mathrm{err}_{\mathcal{D}}(\hat{h}, \ell) \leq \varepsilon$ and makes at most*

$$O\Bigg(\frac{\ln(6/\varepsilon)\,\theta^2\big(d\ln(2\theta) + \ln\ln(6/\varepsilon) + \ln(1/\delta)\big)}{\omega} + \frac{\theta\big(d\ln(2\theta/\pi_{\mathcal{D}}) + \ln(1/\delta)\big)}{\pi_{\mathcal{D}}^2\,\omega}\Bigg)$$

label requests. A class with $d = 0$ requires no queries since all hypotheses agree everywhere.

Proof. By Theorem 5.3, with probability at least $1 - \delta/8$, $\hat{\omega} \in [\omega/2, 2\omega]$ and the estimation uses at most the second term in the bound. Standard VC uniform convergence gives

$$\left|\hat{\Pr}_{S_1}[h] - \Pr_{x\sim\mathcal{D}}[h(x) = 1]\right| \leq \tau \qquad \text{for all } h \in \mathcal{H}$$

with probability at least $1 - \delta/8$, for a sufficiently large C. At every iteration the fresh validation sample T_i is independent of V_i given the past. Hoeffding's inequality and a union bound over at most $N + 1$ validation samples give, with probability at least $1 - \delta/8$,

$$\left|\hat{\Delta}_{T_i}(V_i) - \Delta_{\mathcal{D}}(V_i)\right| \leq \tau, \qquad \left|\hat{\Pr}_{T_i}[\forall h \in V_i : h(x) = 1] - B_i\right| \leq \tau$$

at every validation step. Since $V_{i+1} \subseteq V_i$, the true masses B_i are nondecreasing. Thus the update of b_i guarantees $b_i \leq B_i \leq b_i + 2\tau$ on this event. The class V_i is always restricted to hypotheses predicting positive mass at most u_i on S_1.

For any iteration that proceeds to query, let $\mathcal{D}_i$ be $\mathcal{D}$ conditioned on $\mathrm{DIS}(V_i)$, and let $p_i = \Pr_{x\sim\mathcal{D}_i}[\ell(x) = 1]$. The validation rule ensures that $\Delta_{\mathcal{D}}(V_i) > \varepsilon/3 - \tau > \varepsilon/6$ on the validation event. We establish three claims about these iterations. Each per-round probability is conditional on the history through the relevant draw, whenever the preceding conditions hold. The union bounds apply to failures along such histories.

Claim 1. With probability at least $1 - \delta/8$ over the query batches, provided $\hat{\omega} \in [\omega/2, 2\omega]$, at every executed query iteration (i) if $p_i \geq 1/(32\theta)$ the algorithm does not enter the `if` branch in Line 14, and (ii) if $p_i \leq 1/(512\theta)$ the algorithm enters that branch.

Proof. Conditional on the history before this batch, $|R_1|$ is binomial with parameters λ_1 and $q_i = \omega p_i$. Write $a = \hat{\omega}/(128\theta)$. In (i), $q_i \geq 2a$, so the lower-tail Chernoff bound gives

$$\Pr[|R_1|/\lambda_1 < a] \leq \exp(-\lambda_1 q_i/8) \leq \exp(-\lambda_1 a/4) \leq \delta/(8N).$$

In (ii), $q_i \leq a/2$. Stochastic domination by a binomial random variable with success probability $a/2$, followed by the upper-tail Chernoff bound, gives

$$\Pr[|R_1|/\lambda_1 \geq a] \leq \exp(-\lambda_1 a/6) \leq \delta/(8N).$$

A union bound over at most N query iterations gives the claim. □

Claim 2. On the events above, the new cap in Line 16 preserves ℓ whenever $\ell \in V_i$.

Proof. Every point in R_1 has true label 1. The new cap can exclude ℓ only if $\hat{\pi}_{S_1} > (u_i + b_i)/2$. In that case Lemma 5.4 gives $p_i \geq 1/(32\theta)$, and Claim 1(i) ensures that the algorithm takes the `else` branch. □

Claim 3. On the events above, with probability at least $1 - 3\delta/8$ over the fresh samples in the `else` branch, every such update for which $\ell \in V_i$ preserves ℓ and halves $\Delta_{\mathcal{D}}(V_i)$.

Proof. When the algorithm takes the `else` branch, Claim 1(ii) implies $p_i > 1/(512\theta)$. Conditional on the history, $|R_2|$ is binomial with mean at least $\lambda_2\omega/(512\theta)$. Set $m_0 = \lambda_2\omega/(1024\theta)$. By the lower-tail Chernoff bound, $\Pr[|R_2| < m_0] \leq e^{-m_0/4} \leq \delta/(8N)$ for sufficiently large M_3. We choose M_3 so that $m_0 \geq 2d$.

Conditional on $m = |R_2| \geq m_0$, R_2 consists of m i.i.d. positives under $\mathcal{D}_i$ by Remark 3.2, and S_2 consists of m independent draws from $\mathcal{D}_i$. The classifier h_{i+1} minimizes the predicted positive count on S_2 over the hypotheses in V_i consistent with R_2. Let $g_m = (d\ln(2m/d) + \ln(8N/\delta))/m$. Theorem 2.6 applied to the restricted class V_i gives $\mathrm{err}_{\mathcal{D}_i}(h_{i+1}, \ell) \leq M_1 g_m$ except with probability $\delta/(8N)$. Corollary 5.2 gives both relative error bounds for S_2 with additive term $M_2 g_m$, except with probability $\delta/(8N)$. Therefore,

$$\rho_{S_2}(h_{i+1}, \ell) \leq (2M_1 + M_2)g_m = \gamma_{i+1},$$

so ℓ remains in V_{i+1}. For every $h \in V_{i+1}$, the reverse error bound gives

$$\mathrm{err}_{\mathcal{D}_i}(h, \ell) \leq \rho_{\mathcal{D}_i}(h, h_{i+1}) + \mathrm{err}_{\mathcal{D}_i}(h_{i+1}, \ell) \leq (5M_1 + 3M_2)g_m \leq \frac{1}{2\theta}.$$

The last inequality follows from $m \geq m_0$, the definition of λ_2, and a sufficiently large M_3. Each such h agrees with ℓ outside $\mathrm{DIS}(V_i)$, so

$$V_{i+1} \subseteq \mathcal{B}_{\mathcal{D}}\left(\ell, \frac{\Delta_{\mathcal{D}}(V_i)}{2\theta}\right), \qquad \Delta_{\mathcal{D}}(V_{i+1}) \leq \frac{\Delta_{\mathcal{D}}(V_i)}{2}.$$

The three failure probabilities sum to $3\delta/(8N)$ at each update. A union bound proves the claim. □

On the simultaneous event, $\ell \in V_i$ at every validation step. Every query iteration either halves $\Delta_{\mathcal{D}}(V_i)$ or decreases $u_i - b_i$ by at least a factor of two. Both quantities are nonincreasing on other rounds and start at most 1. After $\lceil \log_2(6/\varepsilon) \rceil$ updates of either kind, the corresponding stopping criterion holds, since $\varepsilon/6 + \tau < \varepsilon/3$. Thus one of the two accuracy tests holds by iteration N. If the algorithm stops because of the disagreement test, $\mathrm{err}_{\mathcal{D}}(\hat{h}, \ell) \leq \Delta_{\mathcal{D}}(V_i) \leq \varepsilon/3 + \tau$. If the width test holds, both $\hat{h}$ and ℓ label the unanimous-positive region as positive and each has true positive mass at most $u_i + \tau$. Hence

$$\mathrm{err}_{\mathcal{D}}(\hat{h}, \ell) \leq 2(u_i + \tau - B_i) \leq 2(u_i - b_i + \tau) \leq \varepsilon.$$

After estimating ω, there are at most $N(\lambda_1 + \lambda_2)$ requests. Using $\hat{\omega} \geq \omega/2$ gives the first term of the bound. The failure probabilities above sum to at most δ. This completes the proof. □

6 Conclusion and Open Problems

Theorem 5.5 gives a finite label complexity bound when both ω and $\pi_{\mathcal{D}}$ are unknown. Its term that depends on ε differs from the classical CAL bound of Hanneke (2009) by a factor of θ/ω. In our analysis, each round needs on the order of θ revealed positives, up to logarithmic and VC dimension factors, to halve the disagreement mass. The fraction of positive instances in the disagreement region can itself be of order $1/\theta$, giving a second factor of θ in the query bound.

When ω is known, the reduction in Appendix C gives a linear dependence on θ at the cost of a quadratic dependence on $1/\omega$. An open question is whether one can obtain both linear dependencies without the separate cost of estimating an unknown reveal rate. The interval example in Appendix D gives an $\Omega(1/(\pi_{\mathcal{D}}\omega))$ lower bound, even when both parameters are known. The necessity of the squared dependence on $1/\pi_{\mathcal{D}}$ in the upper bound for unknown ω remains open.

A second direction is to study discrete data distributions. With atoms in the data distribution, an instance of substantial mass may remain unlabelable even after repeated queries. Appendix D gives a simple example and discusses when the continuous analysis can still be used. A third direction is to analyze active PU learning in the agnostic setting.

Acknowledgments and Disclosure of Funding

Shai Ben-David's research is supported by an NSERC Discovery grant and through a CIFAR AI Chair. Farnam Mansouri is supported by a NSERC Canada Graduate Scholarship-Doctoral (CGS-D), David Cheriton Scholarship and a Vector Institute Research Grant. Sandra Zilles is supported through a Canada CIFAR AI Chair at the Alberta Machine Intelligence Institute (Amii), through an NSERC Canada Research Chair, through the New Frontiers in Research Fund under grant no. NFRFE-2023-00109 and through the NSERC Discovery Grants program under application no. RGPIN-2026-05856.

We thank the anonymous reviewer who pointed out the connection to active learning under label noise.

References

Abe, N., Zadrozny, B., and Langford, J. (2006). Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 504–509.

Bekker, J. and Davis, J. (2018). Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760.

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.

Castro, R. M. and Nowak, R. D. (2007). Minimax bounds for active learning. In *Proceedings of the 20th Conference on Learning Theory (COLT)*.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

Dasgupta, S., Hsu, D. J., and Monteleoni, C. (2007). A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Du Plessis, M., Niu, G., and Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pages 1386–1394. PMLR.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220.

Ghasemi, A., Manzuri, M. T., Rabiee, H. R., Rohban, M. H., and Haghiri, S. (2011a). Active one-class learning by kernel density estimation. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE.

Ghasemi, A., Rabiee, H. R., Fadaee, M., Manzuri, M. T., and Rohban, M. H. (2011b). Active learning from positive and unlabeled data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 244–250.

Görnitz, N., Kloft, M., and Brefeld, U. (2009). Active and semi-supervised data domain description. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–422. Springer.

Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 353–360.

Hanneke, S. (2009). *Theoretical foundations of active learning*. Carnegie Mellon University.

Hanneke, S. (2011). Rates of convergence in active learning. *The Annals of Statistics*, pages 333–361.

Hanneke, S. (2014). Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3).

Hanneke, S. and Yang, L. (2015). Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(109):3487–3602.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150.

He, J., Li, M., Zhang, H.-J., Tong, H., and Zhang, C. (2006). Generalized manifold-ranking-based image retrieval. *IEEE Transactions on Image Processing*, 15(10):3170–3177.

Kääriäinen, M. (2006). Active learning in the non-realizable case. In *Algorithmic Learning Theory*, pages 63–77. Springer.

Kpotufe, S., Yuan, G., and Zhao, Y. (2022). Nuances in margin conditions determine gains in active learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8112–8126.

Li, W., Guo, Q., and Elkan, C. (2010). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE transactions on Geoscience and Remote Sensing*, 49(2):717–725.

Liu, B., Lee, W. S., Yu, P. S., and Li, X. (2002). Partially supervised classification of text documents. In *ICML*, volume 2, pages 387–394. Sydney, NSW.

Locatelli, A., Carpentier, A., and Kpotufe, S. (2018). An adaptive strategy for active learning with smooth decision boundary. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 547–571.

Mansouri, F. and Ben-David, S. (2025). Learning from positive and unlabeled examples -finite size sample bounds. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*.

Motwani, R. and Raghavan, P. (1996). Randomized algorithms. *ACM Computing Surveys (CSUR)*, 28(1):33–37.

Perini, L., Vercruyssen, V., and Davis, J. (2020). Class prior estimation in active positive and unlabeled learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2915–2921.

Schlachter, P. and Yang, B. (2018). Active learning for one-class classification using two one-class classifiers. In *26th European Signal Processing Conference (EUSIPCO)*, pages 1197–1201. IEEE.

Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media.

Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., and Davis, J. (2018). Semi-supervised anomaly detection with an application to water analytics. In *2018 IEEE International Conference on Data Mining (ICDM)*, volume 2018, pages 527–536. IEEE.

Wang, L. (2011). Smoothness, disagreement coefficient, and the label complexity of active learning. *Journal of Machine Learning Research*, 12:2269–2292.

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., and Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647.

A Multiplicative Chernoff Bounds

The following well-known lemma is used in some of our proofs.

Lemma A.1 (Multiplicative Chernoff Bounds Motwani and Raghavan (1996))**.** *Let $X_1, \ldots, X_m$ be independent random variables drawn according to some distribution $\mathcal{D}$ with mean p and support included in $[0,1]$. Then, for any $\gamma \in \left[0, \frac{1}{p} - 1\right]$, the following inequalities hold for $\widehat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$:*

$$\mathbb{P}[\widehat{p} \geq (1+\gamma)p] \leq e^{-\frac{mp\gamma^2}{3}}, \quad \mathbb{P}[\widehat{p} \leq (1-\gamma)p] \leq e^{-\frac{mp\gamma^2}{2}}.$$

B Proofs Omitted from the Paper

B.1 Proof of Lemma 4.1

Lemma 4.1. *Let $\varepsilon > 0$. Suppose $V \subseteq \mathcal{B}_{\mathcal{D}}(\mathbf{0}, \pi_{\mathcal{D}} + \frac{\varepsilon}{4})$, $\ell \in V$ and $\sup_{h \in V} \mathrm{err}_{\mathcal{D}}(h, \ell) > \varepsilon$. Then*

$$\Pr[\ell(x) = 1 \mid x \in \mathrm{DIS}(V)] \geq \frac{1}{4\theta}.$$

Proof. Denote $b := \Pr_{x \sim \mathcal{D}}[\forall h \in V : h(x) = 1]$. By definition we have

$$\Pr[\ell(x) = 1,\, x \in \mathrm{DIS}(V)] \;=\; \pi_{\mathcal{D}} - b\,. \tag{2}$$

Denote $f = \bigcap_{h \in V} h$. Since $V \subseteq \mathcal{B}_{\mathcal{D}}(\mathbf{0}, \pi_{\mathcal{D}} + \frac{\varepsilon}{4})$, we also have $V \subseteq \mathcal{B}_{\mathcal{D}}(f,\, \pi_{\mathcal{D}} - b + \frac{\varepsilon}{4})$. Since $\ell \in V$, we obtain $V \subseteq \mathcal{B}_{\mathcal{D}}\left(\ell,\, 2\left(\pi_{\mathcal{D}} - b + \frac{\varepsilon}{4}\right)\right)$. Thus,

$$2\left(\pi_{\mathcal{D}} - b + \frac{\varepsilon}{4}\right) \;\geq\; \sup_{h \in V} \mathrm{err}_{\mathcal{D}}(h, \ell) \;>\; \varepsilon\,,$$

which implies $\pi_{\mathcal{D}} - b > \frac{\varepsilon}{4}$. This yields $V \subseteq \mathcal{B}(\ell,\, 4(\pi_{\mathcal{D}} - b))$. Therefore,

$$\Delta_{\mathcal{D}}(V) \;\leq\; \Delta_{\mathcal{D}}\left(\mathcal{B}_{\mathcal{D}}(\ell, 4(\pi_{\mathcal{D}} - b))\right) \;\leq\; 4\theta(\pi_{\mathcal{D}} - b)\,. \tag{3}$$

Combining this with (2) completes the proof. □

B.2 Proof of Theorem 5.3

Theorem 5.3. *Suppose Assumption 3.1 holds, $\pi_{\mathcal{D}} > 0$, and $\theta < \infty$. For any $\delta \in (0,1)$, with probability at least $1 - \delta$, Algorithm 2 returns an estimate in $[\omega/2, 2\omega]$ after making*

$$O\left(\frac{\bar{\theta}\big(d_+ \ln(e\bar{\theta}/\pi_{\mathcal{D}}) + \ln(8/\delta)\big)}{\pi_{\mathcal{D}}^2 \omega}\right)$$

label requests, where $\bar{\theta} = \max\{1, \theta\}$ and $d_+ = \max\{1, d\}$.

Proof. Write $\pi = \pi_{\mathcal{D}}$, $p = \pi\omega$ and $L_j = \lceil 16 \ln(1/\delta_j) \rceil$. For the analysis, we can generate the samples for all stages, including stages after the algorithm returns. By Assumption 3.1, the instances drawn in these stages are almost surely distinct.

In each stage, the feedback from the initial uniform queries is a sequence of i.i.d. Bernoulli trials with success probability p. The first k_j revealed positives form an i.i.d. sample from $\mathcal{D}(\cdot \mid \ell(x) = 1)$, independently of the fresh unlabeled sample S_j. By Theorem 2.6 and Corollary 5.2, with probability at least $1 - 2\delta_j$,

$$\rho_{\mathcal{D}}(h_j^*, \ell) \le \gamma_j \quad \text{and} \quad \mathcal{B}_{\mathcal{D}}(h_j^*, \gamma_j) \subseteq \hat{\mathcal{B}}_{S_j}(h_j^*, 3\gamma_j) \subseteq \mathcal{B}_{\mathcal{D}}(h_j^*, 7\gamma_j).$$

Indeed, the additive term of the corollary is at most γ_j, and when $\gamma_j < 1$, $\ln(1/\gamma_j) \le \ln(2k_j/d_+)$ ensures that k_j meets the sample requirement of Theorem 2.6. If $\gamma_j \ge 1$, the error bound is automatic. On this event $\ell \in V_j$ after pruning, and the region $E_j = \{x : \forall h \in V_j,\ h(x) = 1\}$ contains only positives.

Conditional on the samples used to construct E_j and the fresh validation instances W_j, if E_j contains only positives then r_j is binomial with mean $\mu = \omega|P_j|$. An underestimate at a returning stage requires $r_j \le \mu/2$ and $r_j \ge L_j$, hence $\mu \ge 2L_j$. Its probability is at most $\exp(-L_j/4)$ by the lower-tail Chernoff bound. An overestimate requires $r_j \ge \max\{L_j, 2\mu\}$. Since $\mathbb{E}[2^{r_j}] \le e^{\mu}$, this probability is at most

$$\exp\big(\mu - \ln(2)\max\{L_j, 2\mu\}\big) \le \exp\big(-L_j(\ln 2 - 1/2)\big).$$

These two probabilities sum to at most δ_j. A union bound over all stages gives a probability of at most $3\sum_{j\ge 0}\delta_j$ of ever returning an inaccurate estimate.

For the query bound, set $a = \bar{\theta}/\pi \ge 1$ and $B = d_+ \ln(ea) + \ln(8/\delta)$. Let j_* be the first stage with $k_{j_*} \ge CaB$ for a sufficiently large absolute constant C. Then $k_{j_*} < 2CaB$, and $\ln(j_* + 1) \le \ln(k_{j_*}/d_+)$. These inequalities give $\gamma_{j_*} \le 1/(14a)$ and $k_{j_*} \ge 8L_{j_*}$ for sufficiently large C. To see the first inequality, use

$$\ln(2k_{j_*}/d_+) \le \ln(4Ca) + \ln(B/d_+).$$

After multiplication by d_+, both terms on the right are $O((1 + \ln C)B)$. The same bound controls $\ln(1/\delta_{j_*}) = \ln(16/\delta) + 2\ln(j_* + 1)$.

On the event that both the error bound and the ball inclusions hold in stage j_*, the definition of the disagreement coefficient gives

$$\begin{aligned}\Delta_{\mathcal{D}}(V_{j_*}) &\le \Delta_{\mathcal{D}}(\hat{\mathcal{B}}_{S_{j_*}}(h_{j_*}^*, 3\gamma_{j_*})) \\ &\le \Delta_{\mathcal{D}}(\mathcal{B}_{\mathcal{D}}(h_{j_*}^*, 7\gamma_{j_*})) \le 7\theta\gamma_{j_*} \le \pi/2.\end{aligned}$$

Since $\ell \in V_{j_*}$, every positive instance outside $\mathrm{DIS}(V_{j_*})$ belongs to E_{j_*}. Hence $\mathcal{D}(E_{j_*}) \ge \pi/2$.

Write $k = k_{j_*}$. The event $T_{j_*} < k/(2p)$ requires at least k successes in at most $\lfloor k/(2p) \rfloor$ Bernoulli trials of mean p. Its probability is at most $\exp(-k/8)$ by the same exponential-moment bound used above. Given $T_{j_*} \ge k/(2p)$ and $\mathcal{D}(E_{j_*}) \ge \pi/2$, the count r_{j_*} in the fresh validation sample has conditional mean $T_{j_*}\omega\mathcal{D}(E_{j_*}) \ge k/4$. The lower-tail Chernoff bound gives a conditional probability at most $\exp(-k/32)$ of falling short of L_{j_*}, since $k \ge 8L_{j_*}$. In particular, the algorithm returns by stage j_* except on these events or the failure of that stage's error bound or ball inclusions.

Finally, the total number of requests in the initial uniform batches through stage j_* is the waiting time for $K = \sum_{j=0}^{j_*} k_j < 2k$ successes in i.i.d. trials of success probability p. With probability at least $1 - \exp(-K/4)$ this takes at most $\lceil 2K/p \rceil$ requests, by the lower-tail Chernoff bound applied to that many trials. Each validation batch makes at most T_j additional requests. Combining these bounds and $\sum_{j\ge 0}\delta_j \le \delta/8$ gives the simultaneous guarantee with probability at least $1 - \delta$. The total query bound is $O(k/p) = O(\bar{\theta}B/(\pi^2\omega))$. □

B.3 Proof of Lemma 5.4

Proof. Write $w_i = u_i - b_i$. Since $\ell \in V_i$, its positive mass in $\mathrm{DIS}(V_i)$ is $\pi_{\mathcal{D}} - B_i$. The assumed midpoint inequality and the two estimation bounds give

$$\pi_{\mathcal{D}} - B_i \ge \hat{\pi}_{S_1} - \tau - (b_i + 2\tau) > w_i/2 - 3\tau \ge w_i/4,$$

where the last inequality uses $w_i > \varepsilon/6$ and $\tau = \varepsilon/192$.

Every $h \in V_i$ predicts positive mass at most u_i on S_1, so its true positive mass is at most $u_i + \tau$. It predicts positive throughout the region of mass B_i where all members of V_i agree on label 1. Comparing h and ℓ to the indicator of this region gives $\rho_{\mathcal{D}}(h, \ell) \le 2(u_i + \tau - B_i)$. By the definition of θ,

$$\Delta_{\mathcal{D}}(V_i) \le 2\theta(u_i + \tau - B_i) \le 2\theta(w_i + \tau) \le 3\theta w_i.$$

Dividing the bound on $\pi_{\mathcal{D}} - B_i$ by the bound on $\Delta_{\mathcal{D}}(V_i)$ gives a positive fraction of at least $1/(12\theta)$, which implies the claim. □

C Connection to Noisy Active Learning When ω Is Known

Write $Y = \ell(X)$, and encode feedback $\star$ as $\widetilde{Y} = 0$ and feedback 1 as $\widetilde{Y} = 1$. The resulting label has one-sided class-conditional noise,

$$\Pr(\widetilde{Y} = 0 \mid Y = 1) = 1 - \omega, \qquad \Pr(\widetilde{Y} = 1 \mid Y = 0) = 0.$$

These forward corruption probabilities depend only on ω. In contrast, the fraction of positives among instances with $\widetilde{Y} = 0$ is $(1 - \omega)\pi_{\mathcal{D}}/(1 - \omega\pi_{\mathcal{D}})$ whenever the denominator is positive.

When ω is known, set $\eta = (1 - \omega)/(2 - \omega)$. Keep each revealed label as 1, and replace $\star$ by 1 with probability η and by 0 otherwise. Use an independent draw for each distinct instance and cache the synthetic label if that instance is queried again. For a negative instance the synthetic label flips with probability η. For a positive instance the flip probability is $(1 - \omega)(1 - \eta) = \eta$. The synthetic labels therefore have symmetric classification noise of rate $\eta < 1/2$. If $L_\eta(h)$ denotes the error against the synthetic labels and $g = 1 - 2\eta = \omega/(2 - \omega)$, then

$$L_\eta(h) - L_\eta(\ell) = g \ \mathrm{err}_{\mathcal{D}}(h, \ell).$$

To achieve clean error ε, we apply noisy active learning with excess-error tolerance $g\varepsilon$.

The disagreement-based bound of Hanneke (2011, 2014), also discussed in Section 7.1 of Hanneke and Yang (2015), gives $\widetilde{O}(d\theta/\omega^2)$ label requests. The bounded-noise upper bound in Theorem 4 of Hanneke and Yang (2015) gives

$$\widetilde{O}\left(\frac{1}{\omega^2} \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}\right),$$

where $\mathfrak{s}$ is the star number of $\mathcal{H}$ (Hanneke and Yang, 2015) and $\widetilde{O}$ suppresses logarithmic factors. The d/ε term uses clean error. It comes from the term $(1 - 2\eta)d/\tau$ in the noisy excess-error bound at tolerance $\tau = g\varepsilon$. The pool model of Hanneke and Yang (2015) permits using one synthetic label per sampled instance. By contrast, procedures that need fresh labels on repeated queries of the same instance (Kääriäinen, 2006) require feedback our model does not supply.

These results give a way to use existing noisy active learning methods when ω is known, without knowledge of $\pi_{\mathcal{D}}$. The disagreement-based bound improves the dependence on θ from quadratic to linear, while the bounds in Sections 4 and 5 have linear dependence on $1/\omega$ and allow ω to be unknown. The comparison leaves a gap between the two bounds, including the separate cost of estimating an unknown reveal rate.

D Lower Bounds and Feedback Assumptions

A lower bound under continuous sampling. Partition $[0, 1]$ into $m \ge 2$ intervals of equal length, take $\mathcal{D}$ uniform, and let each hypothesis label just one of the intervals positive. Then $\pi_{\mathcal{D}} = 1/m$, and $\theta = m/2$ because distinct hypotheses disagree on two intervals and the full class disagrees on all of $[0, 1]$. Choose the target interval uniformly at random. Conditional on the pool and the learner's random choices, consider the sequence of at most q queries it would make if every answer were $\star$. Up to the first reveal, its actual queries follow that sequence. Each query in the sequence belongs to the target interval with probability $1/m$ and then reveals a label with probability ω. By a union bound, the probability of any reveal is at most $q\omega/m$. The output on the all-$\star$ transcript can have error below $1/m$ for at most one target, since distinct targets disagree on a set of mass $2/m$. Thus the average success probability for $\varepsilon < 1/m$ is at most $q\omega/m + 1/m$. Achieving success probability at least $3/4$ for every target requires $q = \Omega(m/\omega) = \Omega(1/(\pi_{\mathcal{D}}\omega))$. This example also shows that the positive fraction in a disagreement region can be of order $1/\theta$.

Continuity and repeated queries. In our model, querying the same instance multiple times does not produce new independent responses. The feedback associated with each instance remains fixed. This models an expert whose ability to recognize a particular instance does not change when that instance is presented again.

For discrete distributions, this can prevent learning with high probability. Suppose two hypotheses $h_0, h_1 \in \mathcal{H}$ differ only on an instance x, with $\mathcal{D}(\{x\}) = 1/2$ and $\omega = 1/2$. When x is positive, its feedback is $\star$ with probability $1/2$ and remains $\star$ whenever it is queried. This feedback is also consistent with the hypothesis that labels x negative, while queries elsewhere give the same information under both hypotheses.

One possible modification is to allow the learner to abstain from predicting on instances for which it received feedback $\star$. Analyzing discrete distributions would still be more complicated. When an atom occurs repeatedly but its feedback stays fixed, the revealed positive instances obtained through random queries need no longer form an i.i.d. sample from the conditional positive distribution. This prevents us from directly applying the passive PU learning results used in our analysis.

A further difficulty arises when an instance x is labeled positive by every hypothesis in $\mathcal{H}$, with $\mathcal{D}(\{x\}) = 1/2 - 2\varepsilon$, while the remaining positive instances have total mass 2ε. Obtaining even one revealed positive instance other than x through random querying requires $\Omega(1/(\omega\varepsilon))$ queries. Thus collecting informative positives by random querying can preclude an exponential improvement in $1/\varepsilon$ without further assumptions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction provide informal summaries of the formal results in the main body of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In our conclusions section, the limitations of our work are discussed as potential lines of future research.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are made explicit in the premises of the formal results. Proofs are provided either in the main body or in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed. The work described in this paper is purely theoretical and not tied to any particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a purely theoretical paper that provides no data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.