Majority-of-Three is an Optimal PAC Learner

Grigoris Velegkas
Google Research
gvelegkas@google.com

Abstract

Finding the optimal sample complexity of realizable PAC learning was a major open problem in learning theory. This was resolved by the breakthrough work of Hanneke [2016], building on Simon [2015]. Hanneke's algorithm is technically involved and requires training many empirical risk minimizers on carefully constructed subsets of the data. Motivated by this, Aden-Ali et al. [2024] raised the question: what is the *simplest* optimal PAC learner? Since using a single ERM call is provably suboptimal, they proposed the majority-of-three ERMs as a candidate for the simplest optimal PAC learner, and established its optimality in the *in-expectation* regime. Their main open question was whether this algorithm is optimal in the classical *high-probability* regime. In this work, building on their ideas, we answer this question affirmatively.

1 Introduction

In the realizable Probably Approximately Correct (PAC) model [Valiant, 1984, Vapnik and Chervonenkis, 1971], the learner receives labeled samples from an unknown distribution P over an instance space $\mathcal{X}$. The labels are generated by an unknown target function $f^\star$ belonging to a known class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ of VC dimension d. Given a sample $(S, f^\star(S)) = ((X_1, f^\star(X_1)), \dots, (X_n, f^\star(X_n)))$, the learner returns a classifier $\widehat{f}_S$ and is evaluated by

$$\mathrm{err}_P(\widehat{f}_S) = \mathop{\mathbb{P}}_{X\sim P}[\widehat{f}_S(X) \neq f^\star(X)].$$

The most basic learner is empirical risk minimization (ERM): return any $\widehat{f}_S \in \mathcal{F}$ consistent with the labeled sample. Classical uniform convergence [Vapnik and Chervonenkis, 1971, Blumer et al., 1989] gives, for every fixed ERM learner, that with probability $1-\delta$ upon receiving n samples its error is bounded by

$$\mathrm{err}_P(\widehat{f}_S) = O\left(\frac{d}{n}\log\frac{n}{d} + \frac{1}{n}\log\frac{1}{\delta}\right). \tag{1}$$

This guarantee is tight for ERM and, more broadly, proper learning is not sufficient for the optimal PAC rate in full generality [Daniely and Shalev-Shwartz, 2014, Bousquet et al., 2020]. The optimal rate is instead

$$\Theta\left(\frac{d}{n} + \frac{1}{n}\log\frac{1}{\delta}\right), \tag{2}$$

and was first achieved by Hanneke's algorithm [Hanneke, 2016], building on ideas of Simon [2015]. Hanneke's algorithm is technically and computationally involved since it requires training roughly $n^{0.79}$ ERM learners on carefully crafted *overlapping* subsets of the input. This induces high computational overhead both during training time and *inference* time, since one needs to take the majority vote over $n^{0.79}$ classifiers.

For these reasons, subsequent work focused on obtaining *simple* optimal improper learners. Another breakthrough result by Larsen [2023] showed that the celebrated Bagging algorithm proposed by Breiman [1996] is an optimal PAC learner. Subsequently, Høgsgaard [2025] proposed and studied a model that is aiming to capture the *computational* cost of optimal PAC learning by trading off

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

the number of examples being used in every ERM call and the total number of ERM calls. Within that model, they proposed algorithms that achieve optimal sample complexity and approximate the Pareto curve of this trade-off. While these algorithms are optimal, their constructions and analyses are significantly more involved than plain ERM. Importantly, they also require more than a constant number of ERM calls. In a slightly different thread, Aden-Ali et al. [2023] provided a different optimal PAC learner that is not based on ERM, and relies on the One-Inclusion-Graph (OIG). While the OIG is a central object in learning theory, it is generally not a practical primitive, as constructing it is often infeasible.

Aden-Ali et al. [2024] proposed what is arguably the simplest possible improper learner: train three independent copies of the same ERM learner and return their pointwise majority vote. They showed that this majority-of-three learner has expected error $O(d/n)$, which a single ERM cannot guarantee in general. They also proved a near-optimal high-probability bound, with an additional

$$\log\log\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)$$

factor in the d/n term, and conjectured that a different analysis should give the optimal rate (2). In this work, we validate their conjecture.

Main contribution. Let $B_S = \{x : \widehat{f}_S(x) \neq f^\star(x)\}$ denote the error set of ERM trained on S. Our main technical result is an exponential tail bound for the overlap of the error sets of two independent ERMs. In particular, when the size of the training sets S, T is n we have

$$\mathop{\mathbb{P}}_{S,T}\left\{P(B_S \cap B_T) > C\frac{d+t}{n}\right\} \leq e^{-t}, \qquad t \geq 0\,, \tag{3}$$

where C is some numerical constant. Notice that the majority vote errs only at points where at least two of the three ERMs err. Therefore (3), applied to the three pairs and union bounded, yields

$$\mathrm{err}_P\big(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\big) \leq C'\left(\frac{d}{n} + \frac{1}{n}\log\frac{1}{\delta}\right),$$

with probability at least $1-\delta$, where again C' is an absolute numerical constant.

Concurrent and independent work [Rawal and Zhivotovskiy, 2026]. A parallel work by Rawal and Zhivotovskiy [2026] also shows that majority-of-three is an optimal PAC learner. The crux of their technical approach also relies on controlling the moments of (a version of) the overlap of the mass of the error regions of two independent ERMs.[1] The technical implementations differ, however. Rawal and Zhivotovskiy [2026] derive a direct tail bound for conditional error probabilities for arbitrary deterministic consistent selectors and then prove a single (weighted) recursion. In contrast, our proof applies the fixed-set conditional ERM estimate of Aden-Ali et al. [2024] to deterministic level sets of

$$q_E(x) := \mathbb{P}_S\big(E \cap \{x \in B_S\}\big)\,.$$

Their result is stated with explicit constants, whereas our result gives an implicit universal constant.

2 Setting and main results

Let $(\mathcal{X}, \mathcal{A})$ be a measurable instance space and let P be a probability distribution on $\mathcal{X}$. Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ have VC dimension d, and assume the realizable setting: the target $f^\star$ belongs to $\mathcal{F}$. A training sample $S = (X_1, \ldots, X_n)$ consists of n independent draws from P, with labels $f^\star(X_i)$. We fix an ERM algorithm: for every labeled sample, it returns a classifier $\widehat{f}_S \in \mathcal{F}$ such that

$$\widehat{f}_S(X_i) = f^\star(X_i), \qquad i = 1, \ldots, n.$$

Throughout this work we assume the algorithm satisfies standard measurability conditions for all displayed events and integrals to be well-defined. For a sample S, define the error set

$$B_S := \{x \in \mathcal{X} : \widehat{f}_S(x) \neq f^\star(x)\}\,. \tag{4}$$

If $d = 0$, all functions in $\mathcal{F}$ agree pointwise with $f^\star$ and the error sets are empty. We therefore focus on the nontrivial case $d \geq 1$. Whenever we use logarithms throughout the paper we assume the natural base.

[1] In fact, their proof is more general and holds for deterministic consistent selectors A.

2.1 Main theorem

We are now ready to state the main technical result of our work.

Theorem 2.1. *There is an absolute numerical constant $C < \infty$ such that the following holds. For every distribution P over $\mathcal{X}$ and for every $n \in \mathbb{N}$, let $S, T \sim P^n$ be independent training samples. Then, for every $t \geq 0$,*

$$\mathop{\mathbb{P}}_{S,T}\left\{ P(B_S \cap B_T) > C\frac{d+t}{n} \right\} \leq e^{-t}\,.$$

As we alluded to earlier, this result is very useful because we can bound the probability that the majority-of-three errs by the probability that at least one of the three pairs errs. This is precisely what Theorem 2.1 controls. Formally, for binary classifiers g_1, g_2, g_3, write

$$\mathrm{Maj}(g_1, g_2, g_3)(x) = \mathbf{1}\{g_1(x) + g_2(x) + g_3(x) \geq 2\}.$$

Taking a union bound over these three events gives the optimal sample complexity of the majority-of-three ERMs as a direct corollary.

Corollary 2.2. *There is a universal constant $C < \infty$ such that, for every distribution P over $\mathcal{X}$, $n \in \mathbb{N}$, and for independent samples $S_1, S_2, S_3 \sim P^n$ and every $\delta \in (0,1)$, it holds that*

$$\mathrm{err}_P\big(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\big) \leq C\left(\frac{d}{n} + \frac{1}{n}\log\frac{1}{\delta}\right)$$

with probability at least $1-\delta$.

For notational convenience, we think of the algorithm as using $3n$ samples. We can easily translate the result to the setting where each ERM is trained on $\lfloor n/3 \rfloor$ samples by adjusting the numerical constant appropriately.

2.2 A technical tool from Aden-Ali et al. [2024]

Our analysis utilizes a technical result of Aden-Ali et al. [2024] that gives a bound on the error of an ERM learner under the conditional distribution P_R, where $R \subseteq \mathcal{X}$ is an arbitrary fixed measurable set. Concretely, for a measurable $R \subseteq \mathcal{X}$ with $P(R) > 0$, let P_R denote P conditioned on R, and write $\mathrm{err}_{P_R}(\widehat{f}_S) = P_R(B_S)$. Let $\mathrm{Log}_+(u) = 1 \vee \log u$, $u > 0$. The next result is a straightforward rewriting of Lemma 11 of Aden-Ali et al. [2024].[2]

Lemma 2.3 ([Aden-Ali et al., 2024])**.** *There is a universal constant $c_0 < \infty$ such that, for every fixed measurable set $R \subseteq \mathcal{X}$ with $P(R) > 0$ and every $\eta \in (0, 1/2]$,*

$$\mathop{\mathbb{P}}_{S}\left\{ \mathrm{err}_{P_R}(\widehat{f}_S) > c_0\frac{d\,\mathrm{Log}_+(eP(R)n/d) + \log(e/\eta)}{P(R)n} \right\} \leq \eta.$$

3 Proof overview

To get some intuition about the proof, it is instructive to start from the quantity we wish to bound. Let

$$Z := P(B_S \cap B_T)\,,$$

where $S, T \sim P^n$ are independent samples and B_S, B_T are the error sets of the corresponding ERMs. The majority vote errs only at points where at least two of the three ERMs err. Hence, the main task is to prove an exponential tail bound for this pairwise overlap Z.

We prove the tail bound through moments. The target estimate we wish to show is

$$\mathbb{E}\,Z^k \leq \left(\frac{C(d+k)}{n}\right)^k, \qquad k \geq 1. \tag{5}$$

We claim that this moment growth has exactly the right shape. Indeed, taking $k \approx \log(1/\delta)$ and applying Markov's inequality gives

$$Z \leq C'\frac{d + \log(1/\delta)}{n}$$

with probability at least $1-\delta$.

[2]This result is Lemma 11 in the conference version of their paper, and Lemma 4.4 in the arXiv version of their paper.

We begin by writing the k-th moment of Z. For $x = (x_1, \ldots, x_k) \in \mathcal{X}^k$, define the k-point "forced-error" probability

$$p_k(x_1, \ldots, x_k) := \mathop{\mathbb{P}}_{S}\{x_1, \ldots, x_k \in B_S\}.$$

Thus, $p_k(x)$ is the probability that one ERM errs on all of the prescribed points $x_1, \ldots, x_k$. Expanding Z^k using the product measure P^k and then using the independence of S and T gives

$$\mathbb{E}\, Z^k = \int_{\mathcal{X}^k} p_k(x_1, \ldots, x_k)^2 \, dP^k(x_1, \ldots, x_k)\,. \tag{6}$$

The squared term that appears above is the benefit of using two independent ERMs instead of one. The bulk of the proof controls the right-hand side of (6) by analyzing the effect of adding one point at a time and controlling how quickly the value of this expression drops. Let $E = E(S)$ be any event depending only on the sample S used to train the ERM, and define

$$a_E := \mathop{\mathbb{P}}_{S}(E), \qquad q_E(x) := \mathop{\mathbb{P}}_{S}\big(E \cap \{x \in B_S\}\big).$$

The event E can be thought of as the constraint already imposed on the ERM sample, and $q_E(x)$ is the probability that this constraint is satisfied and that the ERM also errs at the new point x. Equivalently, if $a_E > 0$, then

$$\frac{q_E(x)}{a_E} = \mathop{\mathbb{P}}_{S}\{x \in B_S \mid E\}.$$

Thus q_E is the conditional error profile given E, multiplied by the probability a_E of the event E itself. The reason for allowing arbitrary events E is the moment recursion. Fix a tuple $x = (x_1, \ldots, x_k)$ and let

$$E_x := \{S : x_1, \ldots, x_k \in B_S\}.$$

This is the event that the ERM errs on the k prescribed points $x_1, \ldots, x_k$. Then

$$\mathop{\mathbb{P}}_{S}(E_x) = p_k(x_1, \ldots, x_k),$$

and, for a new point y,

$$q_{E_x}(y) = \mathop{\mathbb{P}}_{S}\big(E_x \cap \{y \in B_S\}\big) = p_{k+1}(x_1, \ldots, x_k, y).$$

So this is precisely the $(k+1)$-point forced-error probability, viewed as a function of the new point y.

The main advantage of considering q_{E_x} is that its level sets are fixed. Although the event E_x may depend on the sample, the function $q_E : \mathcal{X} \to [0,1]$ is deterministic once E_x, the distribution, the target, the class, and the ERM selector have been fixed. This lets us apply Lemma 2.3 to the level sets of q_{E_x}. This is where our proof departs from the high-probability analysis of Aden-Ali et al. [2024]. Their proof controls pairwise intersections through conditional two-ERM estimates on domains obtained after conditioning on a realized error set. Here, we apply Lemma 2.3 directly to deterministic level sets of $q_{E_x}(y)$. We elaborate more on the differences in the two approaches in Section 3.1.

We now describe the level sets we consider. Consider an event E and let

$$b_j := 2^{-(j+1)}, \qquad R_{E,j} := \{x \in \mathcal{X} : a_E b_j < q_E(x) \le a_E 2^{-j}\}, \qquad j = 0, 1, 2, \ldots.$$

The sets $R_{E,j}$ are deterministic. Let $\mu_j := P(R_{E,j})$ and, when $\mu_j > 0$, define

$$Y_j(S) := \mathrm{err}_{P_{R_{E,j}}}(\widehat{f}_S) = P_{R_{E,j}}(B_S).$$

This is the conditional error of the ERM inside the fixed set $R_{E,j}$.

The first step is a lower bound on the average of Y_j under the event E. By Fubini's theorem and the definition of R_j,

$$\mathop{\mathbb{E}}_{S}[\mathbf{1}_E Y_j(S)] = \frac{1}{\mu_j} \int_{R_j} q_E(x)\, dP(x) \ge a_E b_j.$$

Indeed, every $x \in R_j$ satisfies $q_E(x) > a_E b_j$. Since $0 \le Y_j \le 1$, the last bound implies: $\mathbb{P}_S\{Y_j(S) \ge b_j/2\} \ge \frac{a_E b_j}{2}$. Now we apply Lemma 2.3 to the fixed set $R_{E,j}$ with confidence parameter η_j of order $a_E b_j$. The lemma gives an upper bound on $\mathbb{P}_S\{Y_j(S) > \tau_j\}$, where

$$\tau_j = c_0 \frac{d \operatorname{Log}_+(e\mu_j n/d) + \log(e/\eta_j)}{\mu_j n}.$$

If $\tau_j < b_j/2$, then the event $\{Y_j \ge b_j/2\}$ would be contained in the exceptional event of the previous bound, contradicting the lower bound above. Therefore $\tau_j \ge b_j/2$. Solving the resulting inequality yields

$$P(R_{E,j}) \le \frac{C}{nb_j}\big(d(j+1) + \ell(a_E) + j + 1\big), \qquad \ell(a_E) := 1 + \log(1/a_E). \tag{7}$$

This is the basic level-set estimate. It turns a lower bound on an appropriate average error over $R_{E,j}$ into an upper bound on the mass of $R_{E,j}$.

The logarithmic term $\ell(a_E)$ comes from the confidence parameter. Since η_j is of order $a_E b_j$, Lemma 2.3 contributes

$$\log(e/\eta_j) \lesssim 1 + \log\frac{1}{a_E} + \log\frac{1}{b_j} \lesssim \ell(a_E) + j + 1.$$

Thus, rare events incur a logarithmic cost. The rest of the proof is a way of carrying this cost through the moment recursion without losing the additive form $d + \log(1/\delta)$.

We now explain the bookkeeping. For $u \in (0, 1]$, let

$$\ell(u) := 1 + \log(1/u).$$

For integers $r \ge 0$, define

$$W_r(u) := \big(d + 4r + \ell(u)\big)^r.$$

These weights are chosen so that the level-set estimates (7) can be summed over all j in a stable way. They satisfy the following one-step inequality: for every event $E = E(S)$ and every integer $r \ge 0$,

$$\int q_E(x)^2 W_r(q_E(x))\, dP(x) \le \frac{C}{n} a_E^2 W_{r+1}(a_E), \qquad a_E = \mathop{\mathbb{P}}_S(E). \tag{8}$$

Let us indicate why (8) follows from the level-set bound. On R_j,

$$q_E(x) \le a_E 2^{-j}, \qquad \ell(q_E(x)) \le \ell(a_E) + j + 1.$$

Therefore, the contribution of R_j to the left-hand side of (8) is at most

$$a_E^2 2^{-2j} P(R_j)\big(d + 4r + \ell(a_E) + j + 1\big)^r.$$

Substituting (7) and using $b_j = 2^{-(j+1)}$ gives:

$$a_E^2 2^{-2j} P(R_j) \lesssim \frac{a_E^2}{n} 2^{-j}\big(d(j+1) + \ell(a_E) + j + 1\big).$$

Hence the entire integral is bounded by

$$\frac{C a_E^2}{n} \sum_{j \ge 0} 2^{-j}\big(d(j+1) + \ell(a_E) + j + 1\big)\big(d + 4r + \ell(a_E) + j + 1\big)^r.$$

The choice of W_r is made so that this sum is at most

$$C\big(d + 4(r+1) + \ell(a_E)\big)^{r+1} = C W_{r+1}(a_E).$$

We can now turn (8) into the desired moment recursion. For integers $k, r \ge 0$, define

$$F_{k,r} := \int_{\mathcal{X}^k} p_k(x)^2 W_r(p_k(x))\, dP^k(x).$$

The desired moment is $F_{k,0}$, because $W_0 \equiv 1$ and (6) gives

$$\mathbb{E}\, Z^k = F_{k,0}.$$

Fix $x = (x_1, \dots, x_k)$ and take $E_x = \{S : x_1, \dots, x_k \in B_S\}$. Then

$$\mathop{\mathbb{P}}_S(E_x) = p_k(x), \qquad q_{E_x}(y) = p_{k+1}(x, y).$$

Applying (8) with $E = E_x$ gives

$$\int p_{k+1}(x,y)^2 W_r(p_{k+1}(x,y))\, dP(y) \le \frac{C}{n} p_k(x)^2 W_{r+1}(p_k(x)).$$

Integrating this inequality over $x \sim P^k$ yields $F_{k+1,r} \le \frac{C}{n} F_{k,r+1}$. Iterating,

$$F_{k,0} \le \frac{C}{n} F_{k-1,1} \le \left(\frac{C}{n}\right)^2 F_{k-2,2} \le \cdots \le \left(\frac{C}{n}\right)^k F_{0,k}.$$

For $k = 0$ the tuple is empty, so the event that the ERM errs on all points of the empty tuple is always true; hence $p_0 \equiv 1$. Therefore $F_{0,k} = W_k(1) = \big(d + 4k + \ell(1)\big)^k \le \big(C(d+k)\big)^k$. Combining this with $F_{k,0} = \mathbb{E}\, Z^k$ gives $\mathbb{E}\, Z^k \le \left(\frac{C(d+k)}{n}\right)^k$.

3.1 Comparison with the Analysis of Aden-Ali et al. [2024]

We now explain the point of departure from the previous high-probability analysis. Both arguments begin with the same observation: the majority vote can err only on pairwise overlaps of ERM error sets. Thus, in both papers, the central task is to obtain a high-probability bound on $Z = P(B_S \cap B_T)$, where $S, T \sim P^n$ are independent samples. Both arguments also use the same single-ERM fixed-set inequality (Lemma 2.3), that was proposed by Aden-Ali et al. [2024]: for a fixed measurable set $R \subseteq \mathcal{X}$, one can control $\mathrm{err}_{P_R}(\widehat{f}_S)$ in terms of $P(R)$, the VC-dimension d, and the desired failure probability.

The main difference lies in the sets to which this inequality is applied. In Aden-Ali et al. [2024], the pairwise-overlap analysis proceeds by partitioning the space according to the pointwise one-ERM error probability $p(x) = \mathbb{P}_S\{x \in B_S\}$. On these levels, their proof establishes conditional two-ERM estimates bound under (conditional) distributions P_R. This route requires controlling several local error scales, and the combination over scales produces the additional doubly logarithmic factor in their high-probability theorem.

Our proof follows a different path. For an arbitrary event $E = E(S)$, we consider $q_E(x) = \mathbb{P}_S\big(E \cap \{x \in B_S\}\big)$. Although E may depend on the sample, the function q_E is deterministic after the event (and other relevant quantities like the distribution, target, hypothesis class, and ERM algorithm) have been fixed. Hence, the level sets $R_j = \{x : a_E 2^{-j-1} < q_E(x) \le a_E 2^{-j}\}, a_E = \mathbb{P}_S(E)$, are fixed measurable sets. Thus, Lemma 2.3 can be applied directly to each R_j. Then, we can instantiate the event E appropriately to do a recursive analysis on the probability of error on k points. Leveraging this analysis we can show

$$\mathbb{E}\, Z^k \le \left(\frac{C(d+k)}{n}\right)^k.$$

Finally, Markov's inequality gives $Z \lesssim \frac{d+\log(1/\delta)}{n}$, with probability at least $1-\delta$.

4 Formal Proof Details

In this section we give the formal details of our proof. Throughout this section all logarithms are natural. For $u \in (0,1]$, we let $\ell(u) := 1 + \log(1/u)$. For integers $r \ge 0$ we also let

$$W_r(u) := \big(d + 4r + \ell(u)\big)^r, \qquad u \in (0,1]. \tag{9}$$

We interpret $u^2 W_r(u)$ as zero at $u = 0$. The following two technical results will be useful for our derivations. Their proofs follow by direct algebraic manipulations and are deferred to Appendix A.

Lemma 4.1. *For every $C_0 < \infty$ there exists a constant $C_{\mathrm{fp}} = C_{\mathrm{fp}}(C_0) < \infty$ such that the following holds. Let $d \ge 1$, let $j \ge 0$, let $b = 2^{-(j+1)}$, and let $L \ge 1$ and $y > 0$. If*

$$yb \le C_0\Big(d\,\mathrm{Log}_+(ey/d) + L + j + 1\Big),$$

then

$$y \le \frac{C_{\mathrm{fp}}}{b}\Big(d(j+1) + L + j + 1\Big).$$

Lemma 4.2. *There exists a numerical constant $C_{\mathrm{sum}} < \infty$ such that, for every integer $r \ge 0$, every $D \ge 0$, and every $L \ge 1$,*

$$\sum_{j=0}^{\infty} 2^{-j}\big(D(j+1) + L + j + 1\big)\big(D + 4r + L + j + 1\big)^r \le C_{\mathrm{sum}}\big(D + 4(r+1) + L\big)^{r+1}.$$

The next lemma is a crucial technical tool for our result. We highlight that the constant C_{act} that appears in the lemma is an absolute numerical constant and does not depend on n, d, r, j, a, E, the distribution, or the concept class.

Lemma 4.3. *There exists an absolute numerical constant $C_{\mathrm{act}} < \infty$ so that the following holds. Let E be any event depending only on the training sample $S \sim P^n$. Define*

$$a := \mathop{\mathbb{P}}_S(E), \qquad q(x) := \mathop{\mathbb{P}}_S\big(E \cap \{x \in B_S\}\big), \qquad x \in \mathcal{X},$$

and

$$\int q(x)^2 W_r(q(x))\, dP(x) \le \frac{C_{\mathrm{act}}}{n} a^2 W_{r+1}(a), \qquad r \ge 0.$$

Proof. If $a = 0$, then $q = 0$ P-almost surely and the claim is trivially correct. Assume $a > 0$. By definition, $q(x) = \mathbb{P}_S\big(E \cap \{x \in B_S\}\big) \le \mathbb{P}_S(E)$, hence $q(x) \in [0, a]$. For $j = 0, 1, 2, \ldots$, let

$$b_j := 2^{-(j+1)}, \qquad R_j := \{x \in \mathcal{X} : ab_j < q(x) \le a2^{-j}\}, \qquad \mu_j := P(R_j).$$

We first prove a bound on the mass of R_j. We claim that there is a numerical constant $C_{\mathrm{lev}} < \infty$ such that, for every $j \ge 0$,

$$\mu_j \le \frac{C_{\mathrm{lev}}}{nb_j}\big(d(j+1) + \ell(a) + j + 1\big). \tag{10}$$

If $\mu_j = 0$, the claim holds. Assume $\mu_j > 0$ and define $Y_j(S) := \mathrm{err}_{P_{R_j}}(\widehat{f}_S) = P_{R_j}(B_S)$. By Fubini's theorem,

$$\mathop{\mathbb{E}}_S[\mathbf{1}_E Y_j(S)] = \frac{1}{\mu_j}\int_{R_j} \mathbb{P}(E \cap \{x \in B_S\})\, dP(x) = \frac{1}{\mu_j}\int_{R_j} q(x)\, dP(x) \ge ab_j, \tag{11}$$

where the last inequality uses the definition of R_j.

Since $0 \le Y_j(S) \le 1$, $\mathbb{E}_S[\mathbf{1}_E Y_j(S)] \le \mathbb{P}\{E \cap (Y_j(S) \ge b_j/2)\} + \frac{b_j}{2}\,\mathbb{P}(E)$. Combining this with (11) and using $\mathbb{P}(E) = a$ gives

$$\mathbb{P}\{E \cap (Y_j \ge b_j/2)\} \ge \frac{ab_j}{2}. \tag{12}$$

In particular,

$$\mathbb{P}\{Y_j \ge b_j/2\} \ge \frac{ab_j}{2}. \tag{13}$$

We apply Lemma 2.3 to the fixed set R_j with $\eta_j := \frac{ab_j}{4}$. Since $a \le 1$ and $b_j \le 1/2$, we have $\eta_j \le 1/8$. Lemma 2.3 gives

$$\mathbb{P}\left\{Y_j > c_0 \frac{d\,\mathrm{Log}_+(e\mu_j n/d) + \log(e/\eta_j)}{\mu_j n}\right\} \le \eta_j.$$

Let

$$\tau_j := c_0 \frac{d\,\mathrm{Log}_+(e\mu_j n/d) + \log(e/\eta_j)}{\mu_j n}.$$

If $\tau_j < b_j/2$, then $\{Y_j \ge b_j/2\} \subseteq \{Y_j > \tau_j\}$, and hence (13) would imply $\mathbb{P}\{Y_j > \tau_j\} \ge \frac{ab_j}{2} = 2\eta_j$, contradicting the previous bound. Therefore $\tau_j \ge b_j/2$, i.e.

$$\frac{b_j}{2} \le c_0 \frac{d\,\mathrm{Log}_+(e\mu_j n/d) + \log(e/\eta_j)}{\mu_j n}.$$

Multiplying by $\mu_j n$ gives $\mu_j n b_j \le 2c_0\Big(d\,\mathrm{Log}_+(e\mu_j n/d) + \log(e/\eta_j)\Big)$. Since $\eta_j = ab_j/4$ and $b_j = 2^{-(j+1)}$,

$$\log(e/\eta_j) = \log\frac{4e}{ab_j} = \log(4e) + \log(1/a) + (j+1)\log 2.$$

Thus, there is a numerical constant $C_{\log} < \infty$ such that, for all $a \in (0, 1]$ and all $j \ge 0$, $\log(e/\eta_j) \le C_{\log}\big(\ell(a) + j + 1\big)$. Consequently, $\mu_j n b_j \le C_0^\star\Big(d\,\mathrm{Log}_+(e\mu_j n/d) + \ell(a) + j + 1\Big)$, where $C_0^\star := 2c_0 \max\{1, C_{\log}\}$. Next, we apply Lemma 4.1 with $y = \mu_j n, b = b_j, L = \ell(a)$, and with the constant $C_0^\star$. Let $C_{\mathrm{fp}}^\star := C_{\mathrm{fp}}(C_0^\star)$ be the constant obtained from Lemma 4.1. Then, $\mu_j n \le \frac{C_{\mathrm{fp}}^\star}{b_j}\big(d(j+1) + \ell(a) + j + 1\big)$. Thus, (10) holds with $C_{\mathrm{lev}} := C_{\mathrm{fp}}^\star$.

On R_j, we have $q(x) \le a2^{-j}$ and $q(x) > ab_j = a2^{-(j+1)}$. Thus,

$$\ell(q(x)) = 1 + \log(1/q(x)) \le \ell(a) + (j+1)\log 2 \le \ell(a) + j + 1.$$

Therefore, on R_j, $W_r(q(x)) \le \big(d + 4r + \ell(a) + j + 1\big)^r$. Using (10) across all R_j we get,

$$\begin{aligned}
\int q^2 W_r(q)\, dP &= \sum_{j\ge 0}\int_{R_j} q(x)^2 W_r(q(x))\, dP(x) \\
&\le \sum_{j\ge 0} a^2 2^{-2j}\mu_j\big(d + 4r + \ell(a) + j + 1\big)^r \\
&\le \frac{C_{\mathrm{lev}} a^2}{n}\sum_{j\ge 0}\frac{2^{-2j}}{b_j}\big(d(j+1) + \ell(a) + j + 1\big)\big(d + 4r + \ell(a) + j + 1\big)^r.
\end{aligned}$$

Since $b_j = 2^{-(j+1)}$, $\frac{2^{-2j}}{b_j} = 2^{-2j}2^{j+1} = 2 \cdot 2^{-j}$. Hence,

$$\int q^2 W_r(q)\, dP \le \frac{2C_{\mathrm{lev}}a^2}{n} \sum_{j\ge 0} 2^{-j}\big(d(j+1) + \ell(a) + j + 1\big)\big(d + 4r + \ell(a) + j + 1\big)^r.$$

Now we can apply Lemma 4.2 with $D = d$ and $L = \ell(a)$. The sum is at most $C_{\mathrm{sum}}\big(d + 4(r+1) + \ell(a)\big)^{r+1} = C_{\mathrm{sum}} W_{r+1}(a)$. Therefore

$$\int q^2 W_r(q)\, dP \le \frac{2C_{\mathrm{lev}}C_{\mathrm{sum}}}{n} a^2 W_{r+1}(a).$$

The lemma follows with $C_{\mathrm{act}} := 2C_{\mathrm{lev}}C_{\mathrm{sum}}$. □

We now iterate the previous inequality. For $k \ge 0$, define the "forced-error" probability

$$p_k(x_1, \dots, x_k) := \mathop{\mathbb{P}}_S\{x_1, \dots, x_k \in B_S\}.$$

For $k = 0$, the tuple is empty and $p_0 \equiv 1$. For integers $k, r \ge 0$, set

$$F_{k,r} := \int_{\mathcal{X}^k} p_k(x_1, \dots, x_k)^2 W_r\big(p_k(x_1, \dots, x_k)\big)\, dP^k(x_1, \dots, x_k), \tag{14}$$

again interpreting $0^2 W_r(0)$ as zero.

Lemma 4.4. *For all integers $k, r \ge 0$, $F_{k+1,r} \le \frac{C_{\mathrm{act}}}{n} F_{k,r+1}$.*

Proof. Fix $x = (x_1, \dots, x_k) \in \mathcal{X}^k$ and define $E_x := \{S : x_1, \dots, x_k \in B_S\}$. Then $\mathbb{P}_S(E_x) = p_k(x)$, and, for $y \in \mathcal{X}$, $\mathbb{P}_S(E_x \cap \{y \in B_S\}) = p_{k+1}(x, y)$. Applying Lemma 4.3 with $E = E_x$ gives

$$\int_{\mathcal{X}} p_{k+1}(x, y)^2 W_r(p_{k+1}(x, y))\, dP(y) \le \frac{C_{\mathrm{act}}}{n} p_k(x)^2 W_{r+1}(p_k(x)).$$

Integrating over $x \sim P^k$ proves the claim. □

We are now ready to prove the key technical result.

Lemma 4.5. *There is a numerical constant $C_{\mathrm{mom}} < \infty$ such that the following holds. Let $S, T \sim P^n$ be independent training samples and let $Z := P(B_S \cap B_T)$. Then, for every integer $k \ge 1$, $\mathbb{E}\, Z^k \le \big(C_{\mathrm{mom}} \frac{d+k}{n}\big)^k$.*

Proof. First expand the kth power of the overlap:

$$Z^k = \left(\int \mathbf{1}_{B_S \cap B_T}(x)\, dP(x)\right)^k = \int_{\mathcal{X}^k} \prod_{i=1}^k \mathbf{1}_{B_S}(x_i)\mathbf{1}_{B_T}(x_i)\, dP^k(x_1, \dots, x_k).$$

Taking expectation over the independent samples S and T and using Fubini gives

$$\mathop{\mathbb{E}}_{S,T} Z^k = \int_{\mathcal{X}^k} \mathop{\mathbb{P}}_S\{x_1, \dots, x_k \in B_S\}^2\, dP^k(x_1, \dots, x_k) = F_{k,0},$$

because $W_0 \equiv 1$. Iterating Lemma 4.4 k times gives $F_{k,0} \le \big(\frac{C_{\mathrm{act}}}{n}\big)^k F_{0,k}$. Since $p_0 \equiv 1$, $F_{0,k} = W_k(1) = \big(d + 4k + \ell(1)\big)^k$. As $\ell(1) = 1$ and $d \ge 1$, $d + 4k + \ell(1) = d + 4k + 1 \le 5(d + k)$. Therefore $F_{0,k} \le \big(5(d+k)\big)^k$. Combining the last bounds, $\mathbb{E}\, Z^k = F_{k,0} \le \left(\frac{5C_{\mathrm{act}}(d+k)}{n}\right)^k$. The lemma follows with $C_{\mathrm{mom}} := 5C_{\mathrm{act}}$. □

We are now ready to prove our main result.

Proof of Theorem 2.1. Let $t \ge 0$ and set $k := \max\{1, \lceil t \rceil\}$. By Lemma 4.5 and Markov's inequality, $\mathbb{P}\left\{Z > eC_{\mathrm{mom}} \frac{d+k}{n}\right\} \le e^{-k}$. Since $d \ge 1$ and $k \le t + 1$,

$$d + k \le d + t + 1 \le 2(d + t),$$

where the last inequality holds for all $t \ge 0$ because $d \ge 1$. Also $e^{-k} \le e^{-t}$. Therefore $\mathbb{P}\left\{Z > 2eC_{\mathrm{mom}} \frac{d+t}{n}\right\} \le e^{-t}$. Thus Theorem 2.1 holds with, e.g., $C_{\mathrm{tail}} := 2eC_{\mathrm{mom}}$. □

Proof of Corollary 2.2. Let $B_i := B_{S_i}, i = 1, 2, 3$, and write $M := \mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})$. For every $x \in \mathcal{X}$, if $M(x) \neq f^\star(x)$, then at least two of the three ERMs err at x. Therefore

$$\mathbf{1}\{M(x) \neq f^\star(x)\} \leq \mathbf{1}_{B_1 \cap B_2}(x) + \mathbf{1}_{B_1 \cap B_3}(x) + \mathbf{1}_{B_2 \cap B_3}(x).$$

Integrating with respect to P gives $\mathrm{err}_P(M) \leq P(B_1 \cap B_2) + P(B_1 \cap B_3) + P(B_2 \cap B_3)$.

We then apply Theorem 2.1 to each of the three independent pairs with $t := \log(3/\delta)$. For each pair, the failure probability is at most $e^{-t} = \frac{\delta}{3}$. By the union bound, with probability at least $1 - \delta$, all three pairwise bounds hold simultaneously. On this event, $\mathrm{err}_P(M) \leq 3C_{\mathrm{tail}} \frac{d + \log(3/\delta)}{n}$. Since $d \geq 1$ and $\delta \in (0, 1)$, $d + \log(3/\delta) = d + \log 3 + \log(1/\delta) \leq (1 + \log 3)\big(d + \log(1/\delta)\big)$. Hence

$$\mathrm{err}_P(M) \leq C_{\mathrm{maj}} \left(\frac{d}{n} + \frac{1}{n} \log \frac{1}{\delta} \right),$$

where $C_{\mathrm{maj}} := 3C_{\mathrm{tail}}(1 + \log 3)$. Renaming C_{maj} as the universal constant C in the corollary statement proves the result. □

5 Conclusion

We have resolved the question of Aden-Ali et al. [2024] by showing that majority-of-three ERMs is an optimal PAC learner. Our proof directly bounds the moments of the overlap of the error regions of two independent ERMs. Subsequent work by Mathiasen et al. [2026] establishes the optimal high-probability agnostic PAC rate, up to universal constants, at every fixed best-in-class error, closing the gap left by earlier bounds [Hanneke et al., 2024, Asilis et al., 2025]. Their algorithm uses an agnostic one-inclusion graph. An interesting direction is to obtain these guarantees with simpler learners and to determine whether our moment-based analysis can help.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal pac bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1203–1223. IEEE, 2023.

Ishaq Aden-Ali, Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In *The Thirty Seventh Annual Conference on Learning Theory*, pages 22–45. PMLR, 2024.

Julian Asilis, Mikael Møller Høgsgaard, and Grigoris Velegkas. On agnostic pac learning in the small error regime. *arXiv preprint arXiv:2502.09496*, 2025.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

Steve Hanneke, Kasper Green Larsen, and Nikita Zhivotovskiy. Revisiting agnostic pac learning. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1968–1982. IEEE, 2024.

Mikael Møller Høgsgaard. Efficient optimal pac learning. *arXiv preprint arXiv:2502.03620*, 2025.

Kasper Green Larsen. Bagging is an optimal pac learner. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 450–468. PMLR, 2023.

Markus Engelund Mathiasen, Jian Qian, and Nikita Zhivotovskiy. An optimal agnostic PAC algorithm. *arXiv preprint arXiv:2608.06363*, 2026.

Divit Rawal and Nikita Zhivotovskiy. Majority-of-three is optimal. *arXiv preprint arXiv:2606.13614*, 2026.

Hans U Simon. An almost optimal pac algorithm. In *Conference on Learning Theory*, pages 1552–1563. PMLR, 2015.

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

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

A Deferred Proofs

Proof of Lemma 4.1. Let $y > 0$ satisfy the given inequality. We will rewrite the inequality in terms of a normalized variable to decouple the dependencies. Let $x = \frac{yb}{d}$, which implies $y = \frac{xd}{b}$.

Substituting y into the hypothesis yields:

$$xd \le C_0 \left(d \operatorname{Log}_+ \left(\frac{ex}{b} \right) + L + j + 1 \right).$$

Divide both sides by d:

$$x \le C_0 \operatorname{Log}_+ \left(\frac{ex}{b} \right) + \frac{C_0}{d}(L + j + 1).$$

Since $b = 2^{-(j+1)}$, we have $\log(1/b) = (j+1)\log 2$. Using the basic property of logarithms that $\operatorname{Log}_+(uv) \le \operatorname{Log}_+(u) + \operatorname{Log}_+(v)$ (or simply expanding the logarithm), we can bound the Log_+ term:

$$\operatorname{Log}_+ \left(\frac{ex}{b} \right) \le \operatorname{Log}_+(ex) + (j+1)\log 2.$$

Substituting this back into our inequality gives:

$$x \le C_0 \operatorname{Log}_+(ex) + C_0(\log 2)(j+1) + \frac{C_0}{d}(L + j + 1).$$

Now, let $K = C_0(\log 2)(j+1) + \frac{C_0}{d}(L + j + 1)$. The inequality simplifies to:

$$x - C_0 \operatorname{Log}_+(ex) \le K.$$

It is a standard property of the logarithm that for any constant $C_0 > 0$, the function $f(x) = x - C_0 \operatorname{Log}_+(ex)$ diverges to $+\infty$ linearly as $x \to \infty$. Therefore, there exists a universal constant $C_1 > 0$ such that $x \le C_1 \max(1, K)$.

Since $j \ge 0$, $L \ge 1$, and $d \ge 1$, we know K is bounded away from zero by at least $C_0 \log 2$. Thus, we can find another constant $C_2 = C_2(C_0) < \infty$ such that $x \le C_2 K$. Substituting K into this bound yields:

$$x \le C_2 \left(C_0(\log 2)(j+1) + \frac{C_0}{d}(L + j + 1) \right).$$

We can pull the constants into a single constant $C_3(C_0)$:

$$x \le C_3 \left((j+1) + \frac{L + j + 1}{d} \right).$$

Finally, to solve for y, we multiply both sides by $\frac{d}{b}$ (since $y = \frac{xd}{b}$):

$$y \le \frac{C_3}{b} \left(d(j+1) + L + j + 1 \right).$$

Setting $C_{\mathrm{fp}} = C_3$ yields the desired result. □

Proof of Lemma 4.2. Let $X = D + 4r + L$. Since $D \ge 0$ and $L \ge 1$, we know that $X \ge 1$ and $X \ge 4r$.

First, we bound the linear term in the summand. Observe that:

$$D(j+1) + L + j + 1 \le (D + L + 1)(j+1).$$

Because $D + L \ge 1$, it follows that $D + L + 1 \le 2(D + L)$. Since $2(D + L) \le 2(D + 4r + L) = 2X$, we have:

$$D(j+1) + L + j + 1 \le 2X(j+1).$$

Next, we bound the polynomial term, $(D + 4r + L + j + 1)^r = (X + j + 1)^r$. By factoring out X and applying the standard inequality $1 + x \le e^x$ (valid for all $x \ge 0$), we get:

$$(X + j + 1)^r = X^r \left(1 + \frac{j+1}{X} \right)^r \le X^r \exp\left(\frac{r(j+1)}{X} \right).$$

Because $X \ge 4r$, we know that $\frac{r}{X} \le \frac{1}{4}$. Substituting this upper bound into the exponent yields:

$$(X + j + 1)^r \le X^r \exp\left(\frac{j+1}{4} \right).$$

Now, we substitute both of these bounds back into the original summation:

$$\sum_{j=0}^{\infty} 2^{-j}\big(D(j+1)+L+j+1\big)\big(X+j+1\big)^r \le \sum_{j=0}^{\infty} 2^{-j}\big(2X(j+1)\big)\left(X^r \exp\left(\frac{j+1}{4}\right)\right)$$
$$= 2X^{r+1}\sum_{j=0}^{\infty}(j+1)2^{-j}e^{(j+1)/4}.$$

Let $q = \frac{e^{1/4}}{2}$. Because $e < 16$, we are guaranteed that $e^{1/4} < 2$, which means $0 < q < 1$. Therefore, the infinite sum converges to a finite numerical constant:

$$\sum_{j=0}^{\infty}(j+1)2^{-j}e^{(j+1)/4} = e^{1/4}\sum_{j=0}^{\infty}(j+1)q^j = \frac{e^{1/4}}{(1-q)^2} := M < \infty.$$

Thus, the entire series is strictly bounded by:

$$2MX^{r+1} = 2M(D+4r+L)^{r+1}.$$

Because $D+4r+L < D+4(r+1)+L$, we can slightly loosen the base to match the desired right-hand side:

$$\sum_{j=0}^{\infty} 2^{-j}\big(D(j+1)+L+j+1\big)\big(D+4r+L+j+1\big)^r \le 2M\big(D+4(r+1)+L\big)^{r+1}.$$

Setting the numerical constant $C_{\text{sum}} = 2M$ completes the proof. Importantly, M depends only on e and 2, making C_{sum} purely numerical and independent of r, D, and L. □

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: The theoretical results introduced in the main papaer match the abstract.
   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: We discuss limitations in the conclusion. In particular, our paper studies a theoretical model and doesn't reflect all practical aspects of data collection.
   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: Our paper provided full proofs for each result and precisely states all necessary assumptions.
   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: We do not include any 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: We do not provide any experiments.
   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: We do not inlcude 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: We do not provide 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: We do not provide any 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: We have read and followed the code of ethics.
   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: This is a theoretical paper and we do not foresee any immediate societal impacts.
    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 theoretical paper and does not pose such risks.
    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: This is a theoretical paper that does not use such 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 produce such 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.
    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 these aspects.
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: [Yes]
Justification: We used LLMs to assist with writing, editing, formatting, and technical calculations.
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.