Majority-of-Three is Optimal

Anonymous authors
Paper under double-blind review

Abstract

We give a short proof that the majority vote of three independent consistent classifiers is an optimal learner in the realizable PAC setting. This proves optimality for the simplest voting scheme, while simplifying both the algorithmic structure and the probabilistic analysis of previous voting learners, including the optimal PAC learner of Hanneke (2016) and Larsen's analysis of bagging (Larsen, 2023). In particular, this proves the conjecture of Aden-Ali et al. (2024).

1 Introduction

Determining the optimal sample complexity of PAC learning in the realizable setting has been one of the central problems in statistical machine learning theory since the introduction of VC theory and the PAC model (Vapnik & Chervonenkis, 1971; Valiant, 1984; Blumer et al., 1989). The difficulty is twofold. First, some sample consistent classifiers are known to be suboptimal, so one must identify a suitable algorithmic construction. Second, even for natural candidate algorithms, the required probabilistic analysis has proved delicate.

A major milestone was the optimal PAC learner of Hanneke (2016), building on the construction of Simon (2015). This resolved the question up to universal constants, but the resulting algorithm and its analysis are quite involved. Subsequent work has therefore sought simpler optimal learners or simpler proofs, including bagging (Larsen, 2023) and constructions based on the one-inclusion graph algorithm (Haussler et al., 1994; Aden-Ali et al., 2023a;b).

A particularly elementary candidate in this direction is the majority vote of three independently trained consistent classifiers: split the sample into three equal blocks, train one consistent classifier on each block, and classify by majority vote. This majority-of-three rule was studied by Aden-Ali et al. (2024), who proved the optimal in expectation bound, together with a near optimal PAC bound (up to a doubly logarithmic factor $\log\log(\min\{\frac{n}{d}, \frac{1}{\delta}\})$, where n is the sample size, d is the VC dimension, and δ is the confidence parameter). They conjectured that majority-of-three achieves the optimal PAC bound; here we provide a proof of that claim.

2 Main result

Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ have VC dimension $d \geq 1$, let P be a distribution on $\mathcal{X}$, and let $f^\star \in \mathcal{F}$ be the unknown target. For a training sample $S = (X_1, \ldots, X_m) \sim P^m$, write $(S, f^\star(S))$ for the labeled sample and let $\widehat{f}_S$ be the output of a fixed deterministic consistent selector; thus $\widehat{f}_S \in \mathcal{F}$ and $\widehat{f}_S(X_i) = f^\star(X_i)$ for every sample point X_i. For a classifier f, write

$$\mathrm{err}_P(f) = \mathbb{P}_{X\sim P}\left[f(X) \neq f^\star(X)\right], \qquad \mathrm{Maj}(f_1, f_2, f_3)(x) = \mathbf{1}\{f_1(x) + f_2(x) + f_3(x) > 3/2\}.$$

For simplicity, we assume in what follows that $n = 3m$ for some positive integer m. Given a sample of size n, split it into three independent blocks S_1, S_2, S_3 of equal sizes $m = n/3$, and put $\widehat{f}_i = \widehat{f}_{S_i}$. Importantly, by construction $\widehat{f}_1, \widehat{f}_2, \widehat{f}_3$ are independent. In what follows, we ignore routine measurability issues.

Theorem 2.1. *Assume that $n = 3m$ for some positive integer m. For every $\delta \in (0,1)$, every distribution P, and every $f^\star \in \mathcal{F}$,*

$$\mathbb{P}\left(\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3)\right) \le 10^7 \frac{d + \log(1/\delta)}{n}\right) \ge 1 - \delta.$$

This matches the classic lower bound (Ehrenfeucht et al., 1989) up to a universal constant.

The proof is based on the following VC range space result.

Proposition 2.2 (Moments of two block overlaps)**.** *Let P be a distribution on $\mathcal{X}$ and let $\mathcal{C}$ be a VC class of subsets of $\mathcal{X}$ with $\mathrm{VC}(\mathcal{C}) = d \ge 1$. Let A be any deterministic selector satisfying $A(S) \in \mathcal{C}$ and $A(S) \cap S = \emptyset$ for every sample $S \in \mathcal{X}^m$. If $S, T \sim P^m$ are independent, then for every integer $q \ge 1$,*

$$\|P(A(S) \cap A(T))\|_{L_q} \le 62400 \frac{d+q}{m}.$$

This result is analogous to the overlap estimate of Simon (2015), which is a central tool in subsequent analyses (Hanneke, 2016; Larsen, 2023; Aden-Ali et al., 2024). The difference is that Simon's bound applies to nested samples, whereas here S and T are independent. The nested sample structure is responsible for the additional doubly logarithmic factors in that approach (Aden-Ali et al., 2024).

Proof of Theorem 2.1 assuming Proposition 2.2. Define the collection of error sets

$$\mathcal{C} = \{\{x \in \mathcal{X} : f(x) \ne f^\star(x)\} : f \in \mathcal{F}\}.$$

Observe that $\mathrm{VC}(\mathcal{C}) = \mathrm{VC}(\mathcal{F}) = d$. Thus, if $S \sim P^m$ and $A(S) = \{x : \widehat{f}_S(x) \ne f^\star(x)\}$, then $A(S) \in \mathcal{C}$, and $A(S) \cap S = \emptyset$. Denote $E_i = \{x : \widehat{f}_i(x) \ne f^\star(x)\}$ for $i = 1, 2, 3$ and note that $E_i \in \mathcal{C}$. Just as in Simon (2015), if the majority vote misclassifies x, at least two of the three classifiers are incorrect at x. Therefore, it holds that

$$\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3)\right) \le P(E_1 \cap E_2) + P(E_2 \cap E_3) + P(E_1 \cap E_3).$$

By the triangle inequality and Proposition 2.2 we have

$$\begin{aligned}\left\|\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3)\right)\right\|_{L_q} &\le \|P(E_1 \cap E_2)\|_{L_q} + \|P(E_2 \cap E_3)\|_{L_q} + \|P(E_1 \cap E_3)\|_{L_q} \\ &\le 561600 \frac{d+q}{n}.\end{aligned}$$

Choose $q = \lceil \log(1/\delta) \rceil$. Then $q \le \log(e/\delta)$, and Markov's inequality gives

$$\mathbb{P}\left(\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3)\right) > e \| \mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3)\right) \|_{L_q}\right) \le \exp(-q) \le \delta.$$

Simplifying the constants, this proves Theorem 2.1. □

3 Proof of Proposition 2.2

We start with the moment version of the double counting argument used in the analysis of Aden-Ali et al. (2024). For $S \sim P^m$ and $x_1, \ldots, x_k \in \mathcal{X}$ define

$$p_k(x_{1:k}) = \mathbb{P}_S(x_1, \ldots, x_k \in A(S)). \tag{1}$$

If $S, T \sim P^m$ and $X = (X_1, \ldots, X_q) \sim P^q$ are independent, then, by Fubini and independence, it holds that

$$\mathbb{E}_{S,T} P(A(S) \cap A(T))^q = \mathbb{E}_X \mathbb{P}_S(X_1, \ldots, X_q \in A(S)) \mathbb{P}_T(X_1, \ldots, X_q \in A(T)) = \mathbb{E}_X p_q(X)^2. \tag{2}$$

Thus, by (2), the problem of upper bounding the moment reduces to understanding the behavior of the random variable $p_q(X)$. In what follows, we need the following explicit form of the VC bound (implied by

the realizable PAC bound for consistent classifiers (Blumer et al., 1989, Theorem 2)). For every probability measure Q, every integer $N \geq 0$, and every $\varepsilon \in (0, 1]$,

$$\mathbb{P}_{Y_1,\ldots,Y_N \sim Q}\left(\exists R \in \mathcal{C} : R \cap \{Y_1, \ldots, Y_N\} = \emptyset,\ Q(R) > \varepsilon\right) \leq 2\left(\frac{13}{\varepsilon}\right)^{2d} \exp\left(-\varepsilon N/4\right). \tag{3}$$

To analyze $p_q(X)$, we use a recursive chain rule decomposition. At each step, after fixing a prefix $z = (z_1, \ldots, z_r)$, the relevant quantity is the conditional inclusion probability

$$\alpha_z(x) = \mathbb{P}_S\left(x \in A(S) \mid z_1, \ldots, z_r \in A(S)\right).$$

Thus, we need a uniform tail bound on these conditional probabilities, with explicit dependence on the rarity of the conditioning event $p_z = \mathbb{P}_S\left(z_1, \ldots, z_r \in A(S)\right)$. The next lemma provides exactly this form of conditional control.

Lemma 3.1 (Tail bound for conditional inclusion probabilities)**.** *Let $S \sim P^m$ and let $A(S) \in \mathcal{C}$ satisfy $A(S) \cap S = \emptyset$, where $\mathrm{VC}(\mathcal{C}) = d \geq 1$. Fix $z = (z_1, \ldots, z_r)$ and define $\mathcal{E}_z = \{z_1, \ldots, z_r \in A(S)\}$, and $p_z = \mathbb{P}_S(\mathcal{E}_z)$. If $p_z > 0$, set*

$$\alpha_z(x) = \mathbb{P}_S(x \in A(S) \mid \mathcal{E}_z) = \frac{\mathbb{P}_S(x, z_1, \ldots, z_r \in A(S))}{\mathbb{P}_S(z_1, \ldots, z_r \in A(S))}.$$

Then, for every $a \in (0, 1]$, the following tail bound for the conditional inclusion probability holds for $X \sim P$,

$$\mathbb{P}\left(\alpha_z(X) > a\right) \leq \frac{160}{am}\left(d \log \frac{e}{a} + \log \frac{e}{p_z}\right).$$

Proof. Let $G = \{x : \alpha_z(x) > a\}$, and $\beta = P(G) = \mathbb{P}\left(\alpha_z(X) > a\right)$. There is nothing to prove if $\beta = 0$, so assume $\beta > 0$, and let P_G denote P conditioned on G. By the definition of G, conditioning on $\mathcal{E}_z$, the random set $A(S)$ has average P_G-mass larger than a. Indeed, by Fubini,

$$\begin{aligned}\mathbb{E}\left[P_G(A(S)) \mid \mathcal{E}_z\right] &= \mathbb{E}\left[\frac{1}{\beta}\int_G \mathbf{1}\{x \in A(S)\}\,\mathrm{d}P(x) \,\middle|\, \mathcal{E}_z\right] \\ &= \frac{1}{\beta}\int_G \mathbb{P}_S(x \in A(S) \mid \mathcal{E}_z)\,\mathrm{d}P(x) = \frac{1}{\beta}\int_G \alpha_z(x)\,\mathrm{d}P(x) > a.\end{aligned}$$

Since $0 \leq P_G(A(S)) \leq 1$, the inequality

$$a < \mathbb{E}[P_G(A(S)) \mid \mathcal{E}_z] \leq \frac{a}{2} + \mathbb{P}\left(P_G(A(S)) > a/2 \mid \mathcal{E}_z\right),$$

multiplied by $p_z = \mathbb{P}(\mathcal{E}_z)$, implies

$$\mathbb{P}\left(P_G(A(S)) > a/2\right) \geq \frac{ap_z}{2}. \tag{4}$$

We now bound the same probability using the VC bound (3). By the definition of $A(S)$ and (4), it holds that

$$\{P_G(A(S)) > a/2\} \subseteq \{\exists R \in \mathcal{C} :\ R \cap S = \emptyset,\ P_G(R) > a/2\}. \tag{5}$$

Let $N = \sum_{i=1}^m \mathbf{1}\{X_i \in G\}$, where $S = (X_1, \ldots, X_m)$. Conditional on N, the points of $S \cap G$ are N independent samples from P_G. Applying (3) to the restricted range space $\{R \cap G : R \in \mathcal{C}\}$, which still has VC dimension at most d, with $Q = P_G$ and $\varepsilon = a/2$, gives

$$\mathbb{P}\left(\exists R \in \mathcal{C} : R \cap S = \emptyset,\ P_G(R) > a/2 \mid N\right) \leq 2\left(\frac{26}{a}\right)^{2d} \exp(-aN/8).$$

Since $N \sim \mathrm{Bin}(m, \beta)$, we obtain

$$\mathbb{E}\exp(-aN/8) = (1 - \beta + \beta\exp(-a/8))^m = (1 - \beta(1 - \exp(-a/8)))^m \leq \exp(-a\beta m/16),$$

where we used $1 - \exp(-t) \geq t/2$ for $0 \leq t \leq 1/8$. Therefore,

$$\mathbb{P}\left(\exists R \in \mathcal{C}: \ R \cap S = \emptyset, \ P_G(R) > a/2\right) \leq 2\left(\frac{26}{a}\right)^{2d} \exp\left(-a\beta m/16\right). \tag{6}$$

Combining the lower bound (4) with the upper bound (6), using the inclusion (5), we get

$$\frac{ap_z}{2} \leq 2\left(\frac{26}{a}\right)^{2d} \exp\left(-a\beta m/16\right).$$

Taking logarithms and simplifying yields

$$\frac{a\beta m}{16} \leq \log\frac{4}{ap_z} + 2d\log\frac{26}{a} \leq 10d\log\frac{e}{a} + \log\frac{e}{p_z} \leq 10\left(d\log\frac{e}{a} + \log\frac{e}{p_z}\right).$$

Since $\beta = \mathbb{P}\left(\alpha_z(X) > a\right)$, the claim now follows. □

We now prove Proposition 2.2.

Proof of Proposition 2.2. Let $X_1, \ldots, X_q \sim P$ be independent test points, independent of the training sample. Define for $0 \leq k \leq q$, following (1), $p_k(x_{1:k}) = \mathbb{P}_S(x_1, \ldots, x_k \in A(S))$, with $p_0 = 1$. We set

$$M_k = \mathbb{E}_{X_{1:k}}\left[p_k(X_{1:k})^2\left(64(d+q) + \log\frac{e}{p_k(X_{1:k})}\right)^{q-k}\right],$$

with the convention that the integrand is zero when $p_k(X_{1:k}) = 0$. It can be seen as a weighted version of the functional (2). An immediate observation is that when $k = q$ by (2) we have

$$\mathbb{E}_{S,T}P(A(S) \cap A(T))^q = M_q.$$

At the same time, the following recursion holds: for $k = 1, \ldots, q$,

$$M_k \leq \frac{960}{m}M_{k-1}, \tag{7}$$

which implies, iterating,

$$\mathbb{E}_{S,T}P(A(S) \cap A(T))^q = M_q \leq \left(\frac{960}{m}\right)^q M_0 = \left(\frac{960}{m}\right)^q (64(d+q)+1)^q \leq \left(62400\,\frac{d+q}{m}\right)^q, \tag{8}$$

where we used $64(d+q)+1 \leq 65(d+q)$. Note that (8) is exactly the claim of the proposition.

It remains only to prove the recursion (7). Consider first the deterministic sequence $x_1, \ldots, x_k \in \mathcal{X}$ and denote

$$\alpha(x) = \alpha_{(x_1,\ldots,x_{k-1})}(x) = \mathbb{P}_S(x \in A(S) \mid x_1, \ldots, x_{k-1} \in A(S)).$$

By the definition of the conditional probability we have

$$\begin{aligned} &p_k(x_{1:k})^2\left(64(d+q) + \log\frac{e}{p_k(x_{1:k})}\right)^{q-k} \\ &\quad = p_{k-1}(x_{1:k-1})^2\alpha^2(x_k)\left(64(d+q) + \log\frac{e}{p_{k-1}(x_{1:k-1})} + \log\frac{1}{\alpha(x_k)}\right)^{q-k}. \end{aligned} \tag{9}$$

Denote for brevity $p_{k-1} = p_{k-1}(x_{1:k-1})$. For $a \in (0,1]$, consider the real-valued function $\phi : (0,1] \to \mathbb{R}$ given by $\phi(a) = p_{k-1}^2 a^2\left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{a}\right)^{q-k}$. A direct differentiation shows that

$$0 \leq \phi'(a) \leq 2p_{k-1}^2 a\left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{a}\right)^{q-k},$$

and thus, since $\lim_{a\to 0_+} \phi(a) = 0$, for any $y \in (0,1]$,

$$\phi(y) \le \int_0^y 2p_{k-1}^2 a \left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{a}\right)^{q-k} \mathrm{d}a. \tag{10}$$

Keeping all variables fixed, we now treat x_k as a random variable $X \sim P$. Integrating the right-hand side of the previous inequality with respect to this variable and using (10) and Fubini, we have

$$\begin{aligned}
&p_{k-1}^2 \mathbb{E}_X \alpha^2(X) \left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{\alpha(X)}\right)^{q-k} \\
&\le p_{k-1}^2 \int_0^1 2a \left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{a}\right)^{q-k} \mathbb{P}(a < \alpha(X))\,\mathrm{d}a \\
&\le \frac{320}{m} p_{k-1}^2 \int_0^1 \left(d\log\frac{e}{a} + \log\frac{e}{p_{k-1}}\right)\left(64(d+q) + \log\frac{e}{p_{k-1}} + \log\frac{1}{a}\right)^{q-k} \mathrm{d}a \qquad (11) \\
&= \frac{320}{m} p_{k-1}^2 \int_0^\infty \left(d(1+u) + \log\frac{e}{p_{k-1}}\right)\left(64(d+q) + \log\frac{e}{p_{k-1}} + u\right)^{q-k} \exp(-u)\,\mathrm{d}u, \qquad (12)
\end{aligned}$$

where in (11) we used Lemma 3.1 and in (12) we used the change of variables $u = \log(1/a)$. Denote $r = 64(d+q) + \log\frac{e}{p_{k-1}}$. Observe that $d(1+u) + \log\frac{e}{p_{k-1}} \le r(1+u)$. Moreover, since $r \ge 64(q-k)$, we have

$$(r+u)^{q-k} = r^{q-k}\left(1 + \frac{u}{r}\right)^{q-k} \le r^{q-k}\exp\left(\frac{(q-k)u}{r}\right) \le r^{q-k}\exp(u/64).$$

Thus, (12) can be bounded by

$$\frac{320}{m} p_{k-1}^2 \int_0^\infty (r(1+u))\, r^{q-k} \exp(-63u/64)\,\mathrm{d}u \le \frac{960}{m} p_{k-1}^2 r^{q-k+1},$$

where we used $\int_0^\infty (1+u)\exp(-63u/64) \le 3$. Plugging this into (9) and integrating with respect to the remaining variables $X_1, \ldots, X_{k-1}$ (using joint independence of $X_1, \ldots, X_k$), we get

$$\begin{aligned}
&\mathbb{E}_{X_{1:k}}\left[p_k(X_{1:k})^2 \left(64(d+q) + \log\frac{e}{p_k(X_{1:k})}\right)^{q-k}\right] \\
&\le \frac{960}{m}\mathbb{E}_{X_{1:k-1}}\left[p_{k-1}(X_{1:k-1})^2 \left(64(d+q) + \log\frac{e}{p_{k-1}(X_{1:k-1})}\right)^{q-k+1}\right],
\end{aligned}$$

which is the desired recursion (7). The claim follows. □

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. The one-inclusion graph algorithm is not always optimal. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 72–88. PMLR, 12–15 Jul 2023a. URL https://proceedings.mlr.press/v195/aden-ali23a.html.

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)*, pp. 1203–1223. IEEE Computer Society, 2023b. doi: 10.1109/FOCS57990.2023.00071. URL https://doi.org/10.1109/FOCS57990.2023.00071.

Ishaq Aden-Ali, Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 22–45. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/aden-ali24a.html.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. URL https://www.jmlr.org/papers/v17/15-389.html.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994. doi: 10.1006/inco.1994.1097.

Kasper Green Larsen. Bagging is an optimal PAC learner. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 450–468. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/larsen23a.html.

Hans U. Simon. An almost optimal PAC algorithm. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1552–1563, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Simon15a.html.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972.

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.

A AI disclosure and proof simplification

This appendix describes how LLM assistance was used, after an earlier proof had been obtained, to simplify the argument. All mathematical claims in the final manuscript were checked and written by the authors.

An informal "one-shot" AI test. On June 8, 2026, we informally tested public versions of OpenAI GPT-5.5 Pro and Anthropic Claude Opus 4.8 on the problem statement alone, without the hints described below. In these tests, neither model produced a complete proof of Theorem 2.1. An example output of GPT-5.5 Pro is available at https://chatgpt.com/share/6a25eaef-7bd4-83ea-a6bb-65e4f0975f69.

AI assistance disclosure. The initial proof of Theorem 2.1 was obtained by the first author before the LLM-assisted simplification stage. During this stage, GPT-5.5 Pro was only used to check intermediate calculations and to search for local simplifications. The resulting proof was then verified by the authors, but it was substantially longer than the one presented here (a sketch of the earlier proof is given in Appendix B). A subsequent stage then used LLMs to simplify and reorganize this argument: starting from the original proof, we explored alternative arguments, one of which led to the shorter recursive form given above.

Proof simplification. We include the following LLM-assisted proof simplification process. Based on the earlier proof, our initial goal was first to compress that proof into a short outline and to use that outline to explore alternative versions of the argument.

First, we prepared a prompt containing the problem statement together with a short list of hints capturing the main ideas. We then repeatedly removed hints and retested the prompt, testing whether the remaining

outline still contained enough information to recover the main steps of the proof in some runs of GPT-5.5 Pro.

The resulting prompt was then used as a starting point for exploring alternative arguments. We ran the prompt multiple times and inspected the proposed solutions. One of these iterations led to the form of the recursive argument presented in this note. The final proof was then simplified, checked, and written by the authors.

Distilled prompt. The final distilled prompt used in the proof simplification process is reproduced below. In tests performed on June 8, 2026, this prompt sometimes produced versions of the main recursive construction. The key hints were: the reduction to bounding moments of two block overlaps; the reduction to the identity (2); and the instruction to study the conditional probabilities $\mathbb{P}\left(x_k \in A(S) \mid x_1, \ldots, x_{k-1} \in A(S)\right)$.

Distilled prompt, GPT-5.5 Pro, June 2026

```
Let \(\mathcal X\) be an arbitrary domain, and let
\[
\mathcal H \subseteq \{0,1\}^{\mathcal X}
\]
be a binary concept class of VC dimension \(d\ge 1\). Let \(P\) be an arbitrary distribution on \(\mathcal X\), and assume the realizable setting: there exists \(h^\star
        \in\mathcal H\) such that the labels are generated by \(h^\star\).

Let \(A\) be an arbitrary deterministic consistent selector: whenever \(S\) is a realizable labeled sample, \(A(S)\in\mathcal H\) agrees with all labels in \(S\).

Given \(n\) iid labeled examples, split the sample into three independent blocks \(S_1,S_2,S_3\) of sizes as equal as possible. Let
\[
\widehat h_i=A(S_i),\qquad i=1,2,3,
\]
and define the majority-of-three learner
\[
\widehat h(x)
=
\operatorname{maj}\{\widehat h_1(x),\widehat h_2(x),\widehat h_3(x)\}.
\]

Resolve the following problem completely: prove that there exists a universal constant \(C>0\) such that, for every \(\delta\in(0,1)\),
\[
\Pr\left(
P\{x:\widehat h(x)\neq h^\star(x)\}
\le
C\frac{d+\log(e/\delta)}{n}
\right)
\ge 1-\delta .
\]

The result must hold uniformly over all distributions \(P\), all VC classes \(\mathcal H\) of dimension \(d\), all targets \(h^\star\in\mathcal H\), and all
        deterministic consistent selectors \(A\). No special ERM rule, compression scheme, margin assumption, disagreement coefficient, distributional regularity, or
        additional structure may be assumed.

A complete solution must prove the optimal high-probability realizable VC bound
\[
O\left(\frac{d+\log(1/\delta)}{n}\right).
\]

Bounds with extra logarithmic factors, expectation-only guarantees, special cases, or algorithm-specific results do not count unless they are upgraded to the full
        theorem above.

You may ignore only routine measurability issues.

A minimal hint toward the intended proof:

First reduce to the case \(h^\star\equiv 0\), so every learned hypothesis is an error set avoiding its training block. Then observe that the majority vote errs only
        where at least two of the three independently learned error sets overlap.

The main step should be a two-block moment bound. For two independent blocks \(S,T\) of size \(m\), prove that for every integer \(q\ge 1\),
\[
\left\|P(A(S)\cap A(T))\right\|_{L_q}
\lesssim
\frac{d+q}{m}.
\]

To prove this, introduce an independent test tuple \(X=(x_1,\dots,x_q)\sim P^q\) and rewrite the overlap moment in terms of
\[
\mathbb E_X \Pr_S(X\subseteq A(S))^2.
\]

For prefixes \(x_1,\dots,x_{k-1}\), study the conditional inclusion probabilities
\[
\Pr\left(x_k\in A(S)\mid x_1,\dots,x_{k-1}\in A(S)\right).
\]
```

B Sketch of the earlier proof of Proposition 2.2

For completeness we sketch the original argument, which reaches Proposition 2.2 through an explicit multi-scale decomposition rather than the single recursion (7); a complete write-up is available at [may be provided

upon request, redacted for anonymization]. Write $R_q = \mathbb{E}_{S,T}\, P(A(S) \cap A(T))^q$ and recall from (2) the identity $R_q = \mathbb{E}_X\, p_q(X)^2$, together with the chain-rule factorization

$$p_q(X) = \prod_{\ell=1}^{q} \alpha_{X_{<\ell}}(X_\ell), \qquad \alpha_{X_{<\ell}}(x) = \mathbb{P}_S(x \in A(S) \mid X_{<\ell} \subseteq A(S))\,,$$

where $X_{<\ell} = (X_1, \ldots, X_{\ell-1})$. The earlier proof controls R_q by tracking these conditional inclusion probabilities across dyadic scales.

For an anchor tuple a with $p_a = \mathbb{P}_S(a \subseteq A(S)) > 0$ and an integer $k \geq 0$, consider the level set $F_k(a) = \{x : 2^{-k-1} < \alpha_a(x) \leq 2^{-k}\}$. The VC dimension enters exactly once, through the relative estimate (3), which gives

$$P(F_k(a)) \leq \frac{C\,(d(k+1) + \log(e/p_a))\,2^k}{m}.$$

The key feature is that conditioning on the anchor never inflates the VC term, which stays $d(k+1)$; the entire cost of conditioning is carried by the rarity term $\log(e/p_a)$. Each test tuple X is then assigned a dyadic profile $k = (k_1, \ldots, k_q)$, with k_ℓ the scale of $\alpha_{X_{<\ell}}(X_\ell)$. Writing Γ_k for the tuples of profile k and $i = \sum_\ell k_\ell$, the factorization gives $p_q(X) \leq 2^{-i}$ on Γ_k, so by (2), $R_q \leq \sum_k 2^{-2i}\, P^q(\Gamma_k)$.

The rarity terms are absorbed by the following observation: on a fixed profile the rarity of each conditioning event is already determined by the earlier levels, since $p_{X_{<\ell}} \geq 2^{-K_\ell - (\ell-1)}$ with $K_\ell = \sum_{j<\ell} k_j$, whence $\log(e/p_{X_{<\ell}}) \leq C(K_\ell + \ell)$. Substituting this into the level-set bound makes the one step estimate depend only on the profile and not on the particular prefix, so an iterated application of Fubini yields $P^q(\Gamma_k) \leq \prod_{\ell=1}^{q} C\,(d(k_\ell+1) + K_\ell + \ell)\,2^{k_\ell}/m$. The whole estimate then collapses to the elementary profile sum

$$R_q \leq \left(\frac{C}{m}\right)^q \sum_{k \in \mathbb{N}_0^q} 2^{-i} \prod_{\ell=1}^{q} (d(k_\ell+1) + K_\ell + \ell) \leq \left(\frac{C(d+q)}{m}\right)^q.$$

This recovers Proposition 2.2 up to the universal constant.