On the Majority of Three Conjecture

Divit Rawal*

June 1, 2026

Abstract

We prove the conjectured high-probability optimal rate for the Majority-of-Three algorithm of [AAHLZ24].

1 Preliminaries, Notation, Majority-of-Three Algorithm

1.1 Basic Notation

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let P be a probability measure on $(\mathcal{X}, \mathcal{A})$. We write $X \sim P$ for a random point distributed according to P and $S = (X_1, \dots, X_m) \sim P^m$ for an i.i.d. sample of size m. When two samples are denoted by S, T, they are always independent unless explicitly stated otherwise. We identify binary classifiers with their positive regions. Thus a hypothesis class is a family $\mathcal{H} \subseteq \mathcal{A}$ of measurable subsets of $\mathcal{X}$. For $h \in \mathcal{H}$, the prediction is 1 on h and 0 on $\mathcal{X} \setminus h$. The VC dimension of $\mathcal{H}$ is denoted by $d \doteq \mathrm{VC}(\mathcal{H})$. Throughout, we shall assume $d \geq 1$. All constants denoted by $C, c, C_0, c_0, C_1, \dots$ are taken to be positive universal constants whose values may change from line to line unless indicated otherwise. For a measurable set $E \subseteq \mathcal{X}$ with $P(E) > 0$, we write $P_E(B) \doteq P(B \mid E) = \frac{P(B \cap E)}{P(E)}$ for the conditional distribution on E. If $P(E) = 0$, all estimates involving $P(E)$ are understood to be vacuous. For a finite tuple $\mathbf{x} = (x_1, \dots, x_r) \in \mathcal{X}^r$ and a set $h \subseteq \mathcal{X}$, we write $\mathbf{x} \subseteq h$ to mean that $x_j \in h$ for every $j = 1, \dots, r$. Tuples are ordered and repetition is allowed. This convention is convenient since P^r is the law of an ordered i.i.d. test tuple. For a sample $S = (X_1, \dots, x_m)$, we write $h \cap S = \varnothing$ to mean $X_i \notin h$ for every $i \in [m] = 1, \dots, m$. For a nonnegative random variable Z and an integer $q \geq 1$, we write $\|Z\|_q \doteq (\mathbb{E} Z^q)^{1/q}$.

1.2 Realizability

We work in the realizable setting. Let $f^\star \in \mathcal{H}$ be the target hypothesis. For every $h \in \mathcal{H}$, define its error region $g_h \doteq h \Delta f^\star$. Let $\mathcal{H}^\star \doteq \{h \Delta f^\star \mid h \in \mathcal{H}\}$. The following reduction allows us to work directly with error regions.

*University of California, Berkeley

Lemma 1.1. *The class $\mathcal{H}^\star$ satisfies $\varnothing \in \mathcal{H}^\star, \mathrm{VC}(\mathcal{H}) = \mathrm{VC}(\mathcal{H}^\star)$. Moreover, if $h \in \mathcal{H}$ is consistent with a labeled sample S, then $(h\Delta f^\star) \cap S = \varnothing$. Finally, for any three hypotheses h_1, h_2, h_3, $\mathrm{Maj}(h_1, h_2, h_3)\Delta f^\star = Maj(h_1\Delta f^\star, h_2\Delta f^\star, h_3\Delta f^\star)$. Thus, the analysis of the majority vote may be done without loss of generality under the standing assumptions $f^\star = \varnothing, \varnothing \in \mathcal{H}$.*

Proof. Since $f^\star \in \mathcal{H}$, $f^\star\Delta f^\star = \varnothing$, so $\varnothing \in \mathcal{H}^\star$. To check the VC dimension, fix a finite set $Y \subseteq \mathcal{X}$. On Y, the symmetric difference with the fixed set $f^\star \cap Y$ acts as a bijection on $\{0,1\}^Y$. Hence $\mathcal{H}$ and $\mathcal{H}^\star$ realize exactly the same labelings on Y, and therefore shatter exactly the same finite sets. Thus, $\mathrm{VC}(\mathcal{H}) = \mathrm{VC}(\mathcal{H}^\star)$. Now suppose h is consistent with the labeled sample. Then, $h(X_i) = f^\star(X_i)$ for every sample point X_i. Equivalently, $(h\Delta f^\star)(X_i) = 0$ for every i and therefore $(h\Delta f^\star) \cap S = \varnothing$. Finally, for $b_1, b_2, b_3, f \in \{0,1\}$, $\mathrm{Maj}(b_1, b_2, b_3) \oplus f = \mathrm{Maj}(b_1 \oplus f, b_2 \oplus f, b_3 \oplus f)$, which is immediate when $f = 0$ and when $f = 1$ follows from the fact that flipping all three bits flips the majority value. Applying pointwise proves the displayed equality of error regions. □

1.3 Consistent Selectors

After this reduction, we rename $\mathcal{H}^\star$ as $\mathcal{H}$. Thus, henceforth we will assume $f^\star = \varnothing, \varnothing \in \mathcal{H}$. In particular, every consistent hypothesis is simply a set avoiding the sample.

We therefore fix a deterministic measurable selector $A : \mathcal{X}^m \to \mathcal{H}$ such that $A(S) \cap S = \varnothing$ for every sample $S \in \mathcal{X}^m$. We shall not assume any additional properties on the selector.

Majority-of-Three Rule

Let $n \geq 1$ and define $m \doteq \lfloor n/3 \rfloor$. The majority-of-three learner partitions the sample into three disjoint blocks of size m. If n is not divisible by 3, the remaining $n - 3m$ observations are discarded. Thus, we obtain independent blocks $S^{(1)}, S^{(2)}, S^{(3)} \sim P^m$. Applying the selector A to each block gives three hypotheses $h_a \doteq A(S^{(a)}), a \in [3]$. The majority-of-three prediction is the pointwise majority vote $\mathbb{1}\{\mathbb{1}_{h_1}(x) + \mathbb{1}_{h_2}(x) + \mathbb{1}_{h_3}(x) \geq 2\}$. Since we work under the convention $f^\star = \varnothing$, each h_a is itself an error region. Therefore, the majority vote errs at a point x iff at least two of the three sets h_1, h_2, h_3 contain x. The majority error region is therefore

$$\mathcal{E}_{\mathrm{Maj}} = (h_1 \cap h_2) \cup (h_2 \cap h_3) \cup (h_3 \cap h_1)$$
$$\implies P(\mathcal{E}_{\mathrm{Maj}}) \leq P(h_1 \cap h_2) + P(h_2 \cap h_3) + P(h_3 \cap h_1)$$

Thus the analysis of the majority vote reduces to controlling the overlap of two independent ERM outputs.

1.4 Overlap Moments

Theorem 1.2. *For independent samples $S, T \sim P^m$, define $Y \doteq P(A(S) \cap A(T))$. For each integer $q \geq 1$, define $R_q \doteq \mathbb{E}_{S,T}[P(A(S) \cap A(T))^q] = \mathbb{E}Y^q$.*

Then,

$$R_q \le \left(\frac{C(d+q)}{m}\right)^q. \tag{1}$$

Once eq. (1) is established, the high-probability majority-vote bound follows from Minkowski's inequality and Markov's inequality with $q \asymp \log(1/\delta)$.

Proposition 1.3. *Assume that eq.* (1) *holds. Then there exists a universal constant* C *such that for every* $\delta \in (0,1)$*,*

$$P(\mathcal{E}_{\mathrm{Maj}}) \le \frac{C(d+\log(1/\delta))}{n},$$

with probability $1-\delta$*.*

Proof. Define $Y_{ab} \doteq P(h_a \cap h_b)$ for $1 \le a < b \le 3$. From the union bound for the majority error region, $P(\mathcal{E}_{\mathrm{Maj}}) \le Y_{12} + Y_{23} + Y_{31}$. Let $q \ge 1$. Taking L^q norms and applying Minkowski's inequality,

$$\|P(\mathcal{E}_{\mathrm{Maj}})\|_q \le \|Y_{12}\|_q + \|Y_{23}\|_q + \|Y_{31}\|_q.$$

Each pair $(S^{(a)}, S^{(b)})$ consists of two independent P^m-samples. Hence

$$\|Y_{ab}\|_q^q = \mathbb{E}\left[P\left(A(S^{(a)}) \cap A(S^{(b)})\right)^q\right] = R_q.$$

From eq. (1), $\|Y_{ab}\|^q = R_q^{1/q} \le \frac{C_\star(d+q)}{m}$. Therefore, $\|P(\mathcal{E}_{\mathrm{Maj}})\|_q \le \frac{3C_\star(d+q)}{m}$. Let $\lambda = \frac{3eC_\star(d+q)}{m}$. Markov's inequality gives

$$\Pr[P(\mathcal{E}_{\mathrm{Maj}}) > \lambda] = \Pr[P(\mathcal{E}_{\mathrm{Maj}})^q > \lambda^q] \le \frac{\mathbb{E}[P(\mathcal{E}_{\mathrm{Maj}})^q]}{\lambda^q} = \left(\frac{\|P(\mathcal{E}_{\mathrm{Maj}})\|_q}{\lambda}\right)^q.$$

Using the previous estimate $\Pr[P(\mathcal{E}_{\mathrm{Maj}}) > \lambda] \le e^{-q}$. Choose $q = \lceil \log(1/\delta) \rceil$. Then $e^{-q} \le \delta$. Since $d \ge 1$, $d+q \le d + \log(1/\delta) + 1 \le 2(d+\log(1/\delta))$. Hence, with probability at least $1-\delta$,

$$P(\mathcal{E}_{\mathrm{Maj}}) \le \frac{6eC_\star(d+\log(1/\delta))}{m}.$$

Finally, if $n \ge 6$, $m = \lfloor n/3 \rfloor \ge n/4$. Therefore, $P(\mathcal{E}_{\mathrm{Maj}}) \le \frac{24eC_\star(d+\log(1/\delta))}{n}$. For $n \le 5$, the conclusion is absorbed into the universal constant and we have finished the proof. □

2 Main Results

Having reduced the problem to proving eq. (1), the remainder of this section is dedicated to proving eq. (1).

For an ordered test tuple $X = (x_1, \ldots, x_q) \in \mathcal{X}^q$, define $p_X \doteq \Pr_S[X \subseteq A(S)]$, where $S \sim P^m$. Thus p_X is the probability that a single ERM output contains every point of the tuple X. Note that the overlap moment R_q is exactly the second moment of the random quantity p_X.

Lemma 2.1. *For every integer $q \geq 1$, $R_q = \mathbb{E}_{X\sim P^q}[p_X^2]$.*

Proof. Let $W \doteq A(S) \cap A(T)$. For fixed S and T, $P(W)^q = P^q(W^q)$. Equivalently, $P(W)^q = \Pr_{X\sim P^q}[X \subseteq W]$. Taking expectation over S and T and applying Tonelli's theorem,

$$R_q = \mathbb{E}_{X\sim P^q} \Pr_{S,T}[X \subseteq A(S) \cap A(T)].$$

Since S and T are independent, $\Pr_{S,T}[X \subseteq A(S)\cap A(T)] = \Pr_S[X \subseteq A(S)] \Pr_T[X \subseteq A(T)]$. Both factors equal p_X. Therefore $\Pr_{S,T}[X \subseteq A(S) \cap A(T)] = p_X^2$. Substituting gives $R_q = \mathbb{E}_{X\sim P^q}[p_X^2]$. □

For our test tuple X and for $1 \leq \ell \leq q$, define the prefix $X_{<\ell} \doteq (x_1, \ldots, x_{\ell-1})$, with the convention that $X_{<1} = \varnothing$. For every prefix $X_{<\ell}$ with $\Pr_S[X_{<\ell} \subseteq A(S)] > 0$, define the conditional inclusion probability $q_{X_{<\ell}}(x_\ell) \doteq \Pr_S[x_\ell \in A(S) \mid X_{<\ell} \subseteq A(S)]$. When $\ell = 1$, this becomes $q_\varnothing(x_1) = \Pr_S[x_1 \in A(S)]$. The probability p_X admits the following chain rule factorization.

Lemma 2.2. *For every tuple X with $p_X > 0$, $p_X = \prod_{\ell=1}^q q_{X_{<\ell}}(x_\ell)$.*

Proof. By definition, $p_X = \Pr_S[x_1, \ldots, x_q \in A(S)]$. Applying the chain rule for conditional probabilities,

$$p_X = \Pr_S[x_1 \in A(S)] \prod_{\ell=2}^{q} \Pr_S[x_\ell \in A(S) \mid x_1, \ldots, x_{\ell-1} \in A(S)].$$

The factors are exactly $q_{X_{<\ell}}(x_\ell)$, which proves the claim. □

The proof of theorem 1.2 will proceed by analyzing the conditional probabilities $q_{X_{<\ell}}(x_\ell)$ on dyadic scales. The primary difficulty is that conditioning on the event $X_{<\ell} \subseteq A(S)$ can substantially change the distribution of the ERM output. We shall spend the next couple lemmas developing the tools needed to control the conditioned process.

Let $E \subseteq \mathcal{X}$ be measurable with $P(E) = \beta > 0$. Recall that $P_E(B) = P(B \mid E) = \frac{P(B\cap E)}{\beta}$ denotes the conditional distribution on E. The following lemma is the only place we shall use VC theory.

Lemma 2.3. *There exist universal constants $C_0, c_0 > 0$ such that the following holds. Let $E \subseteq \mathcal{X}$ be measurable with $P(E) = \beta > 0$. Then, for every $u \in (0, 1]$,*

$$\Pr_S[P_E(A(S) \cap E) \geq u] \leq \exp(C_0 d \log(e/u) - c_0 m\beta u).$$

Proof. Write $\beta = P(E)$, let $N \doteq \sum_{i=1}^m \mathbb{1}_E(X_i) = |S \cap E|$, and set $\mathcal{H}_E \doteq \{h \cap E \mid h \in \mathcal{H}\}$. Restricting every set to E cannot increase the number of distinct labelings induced on a finite subset of E, so $\mathrm{VC}(\mathcal{H}_E) \leq d$.

Suppose $P_E(A(S) \cap E) \geq u$ and put $g \doteq A(S) \cap E \in \mathcal{H}_E$. Then $P_E(g) = P_E(A(S) \cap E) \geq u$. Since $A(S) \cap S = \varnothing$ we have $g \cap S = \varnothing$, and as $g \subseteq E$ any

sample point lying in g would belong to E; hence $g \cap (S \cap E) = g \cap S = \varnothing$. Therefore

$$\{P_E(A(S) \cap E) \geq u\} \subseteq B \doteq \{\exists\, g \in \mathcal{H}_E \mid P_E(g) \geq u,\ g \cap (S \cap E) = \varnothing\},$$

and the event B depends on S only through the points falling in E, i.e. through $S \cap E$.

Condition on N. Given N, the points of $S \cap E$ are N i.i.d. draws from P_E, independent of N and of the points outside E. Hence $\Pr[B \mid N]$ equals the probability that these N points fail to form a u-net for $\mathcal{H}_E$ under P_E, i.e. that some $g \in \mathcal{H}_E$ with $P_E(g) \geq u$ avoids all of them. Put $\delta \doteq \max\{2, d\}$, so that $\mathrm{VC}(\mathcal{H}_E) \leq d \leq \delta$ and $\delta \geq 2$.

By the ε-net theorem [Wel12, Lemma 15.10], for a range space of VC dimension $\delta \geq 2$ and a sample of $M \geq 8/\varepsilon$ points drawn i.i.d. from a probability measure μ, the probability that the sample misses some range of μ-measure at least ε is at most $2\,\Phi_\delta(2M)\,2^{-\varepsilon M/2}$, where $\Phi_\delta(k) = \sum_{i=0}^{\delta} \binom{k}{i}$. (The reference states this for the empirical measure on a finite point set; its proof uses only that the points are i.i.d., so it holds verbatim for an arbitrary μ.) Apply it with $\mu = P_E$, ranges $\mathcal{H}_E$, $M = N$, $\varepsilon = u$. When $N \geq 8/u$ and $2N \geq \delta$,

$$\Pr[B \mid N] \leq\ 2\,\Phi_\delta(2N)\,2^{-uN/2} \leq\ 2\,(2eN/\delta)^{\delta}\,2^{-uN/2},$$

the last step by Sauer's lemma [Wel12, Lemma 15.6] in the form $\Phi_\delta(k) \leq (ek/\delta)^\delta$ for $k \geq \delta$. Taking natural logarithms (the base affects only the constant) and writing $\log(2eN/\delta) = \log(2e/u) + \log(uN/\delta)$, the inequality $\log t \leq \alpha t - \log\alpha - 1$ (all $t, \alpha > 0$) with $t = uN/\delta$ and $\alpha = (\log 2)/4$ gives $\delta \log(uN/\delta) \leq \frac{\log 2}{4} Nu + C'\delta$ for a universal C'. Combining with $\log(2e/u) \leq (1 + \log 2)\log(e/u)$ and $\log(e/u) \geq 1$, the terms free of Nu are at most $C_0\, d \log(e/u)$ for a universal C_0 (using $\delta \leq 2d$), while the Nu terms combine to $-\big(\frac{\log 2}{2} - \frac{\log 2}{4}\big)Nu = -\frac{\log 2}{4} Nu$; hence $\Pr[B \mid N] \leq \exp(C_0 d \log(e/u) - c_0 Nu)$ with $c_0 = (\log 2)/4 \leq 1$. In the remaining range $N < 8/u$ or $2N < \delta$, we have $Nu < \max\{8, \delta/2\} \leq \max\{8, d\}$, so the exponent is at least $C_0 d - c_0 \max\{8, d\} \geq 0$ once $C_0 \geq 8c_0$ (using $\log(e/u) \geq 1$); there $\exp(C_0 d \log(e/u) - c_0 Nu) \geq 1 \geq \Pr[B \mid N]$ holds trivially. Therefore, for every value of N,

$$\Pr[B \mid N]\ \leq\ \exp\big(C_0\, d \log(e/u) - c_0\, Nu\big), \qquad c_0 \leq 1.$$

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

$$\Pr[B] \leq \exp(C_0 d \log(e/u))\mathbb{E}e^{-c_0 uN} = \exp(C_0 d \log(e/u))(1 - \beta + \beta e^{-c_0 u})^m.$$

Then

$$(1 - \beta + \beta e^{-c_0 u})^m \leq \exp\big(-m\beta(1 - e^{-c_0 u})\big).$$

For $u \in (0, 1]$, $1 - e^{-c_0 u} \geq c' u$ for a universal $c' > 0$. Hence,

$$\Pr[B] \leq \exp(C_0 d \log(e/u) - c_0 m\beta u)$$

as claimed. □

The above lemma states that if a measurable set E has P-mass β, then the probability that a consistent ERM output captures a u-fraction of E decays like $\exp(-cm\beta u)$, up to the VC complexity factor $\exp(Cd\log(e/u))$. The estimate is "relative" because the relevant mass parameter is not $P(A(S))$, but the conditional mass $P_E(A(S)\cap E)$. This will be useful later when E is chosen to be a dyadic level set of a conditioned inclusion probability.

We now convert the avoidance estimate of lemma 2.3 into a capacity bound for conditioned ERM processes. Let $\mathbf{a} = (a_1, \dots, a_r) \in \mathcal{X}^r$ be a finite anchor tuple. Define $p_{\mathbf{a}} \doteq \Pr_S[\mathbf{a} \in A(S)]$. Whenever $p_{\mathbf{a}} > 0$, we write $Q_{\mathbf{a}} \doteq P^m \mid \{\mathbf{a} \subseteq A(S)\}$ for the law of the sample conditioned on the event that the ERM contains the anchor tuple. For $x \in \mathcal{X}$, define the conditional inclusion probability $q_{\mathbf{a}}(x) \doteq \Pr_S[x \in A(S) \mid \mathbf{a} \subseteq A(S)]$. The quantity $q_{\mathbf{a}}(X)$ measures how likely the conditioned ERM process is to contain the point x. To study $q_{\mathbf{a}}$, we decompose it into dyadic levels. For each integer $k \geq 0$, define

$$F_k(\mathbf{a}) \doteq \left\{x \mid 2^{-k-1} < q_{\mathbf{a}}(x) \leq 2^{-k}\right\}.$$

The following lemma is an important structural estimate.

Lemma 2.4. *There exists a universal constant C_1 such that for every anchor tuple $\mathbf{a}$ with $p_{\mathbf{a}} > 0$ and every integer $k \geq 0$,*

$$P(F_k(\mathbf{a})) \leq \frac{C_1(d(k+1) + \log(e/p_{\mathbf{a}}))2^k}{m}.$$

Proof. If $P(F_k(\mathbf{a})) = 0$, the argument is trivial, so assume $P(F_k(\mathbf{a})) > 0$. Fix $k \geq 0$, and write $F \doteq F_k(\mathbf{a}), t \doteq 2^{-k}, \beta \doteq P(F)$. By definition of F, $\frac{t}{2} \leq q_{\mathbf{a}}(x) \leq t$ for every $x \in F$. Consider the random variable $Z \doteq P_F(A(S)\cap F)$. Under the conditioned law $Q_{\mathbf{a}}$, $\mathbb{E}_{Q_{\mathbf{a}}} Z = \frac{1}{\beta}\int_F q_{\mathbf{a}}(x)dP(x)$. Since $q_{\mathbf{a}}(X) \in (t/2, t]$ on F, $\frac{t}{2} \leq \mathbb{E}_{Q_{\mathbf{a}}} \leq t$. Because $0 \leq Z \leq 1$, it follows that $Q_{\mathbf{a}}\left[Z \geq \frac{t}{4}\right] \geq ct$ for a universal constant $c > 0$. Indeed, if $Q_{\mathbf{a}}[Z \geq t/4] \leq t/4$, then $\mathbb{E}_{Q_{\mathbf{a}}} Z \leq t/4 + t/4 = t/2$, contradicting the previous lower bound.

On the other hand,

$$Q_{\mathbf{a}}\left[Z \geq \frac{t}{4}\right] = \frac{\Pr_S[\mathbf{a} \subseteq A(S), Z > t/4]}{p_{\mathbf{a}}} \leq \frac{\Pr_S[Z \geq t/4]}{p_{\mathbf{a}}}.$$

Applying lemma 2.3 with $E = F, u = t/4$, gives

$$\Pr_S[Z \geq t/4] \leq \exp(C_0 d\log(e/t) - c_0 m\beta t).$$

Therefore, $ct \leq \exp(C_0 d\log(e/t) - c_0 m\beta t)/p_{\mathbf{a}}$. Taking logarithms and absorbing constants yields

$$m\beta t \leq C(d\log(e/t) + \log(e/p_{\mathbf{a}})).$$

Since $t = 2^{-k}$, we have $\log(e/t) \asymp k+1$. Hence

$$\beta = P(F_k(\mathbf{a})) \leq \frac{C_1(d(k+1) + \log(e/p_{\mathbf{a}}))2^k}{m}$$

which proves the claim. ∎

The important feature of lemma 2.4 is that the VC contribution remains exactly $d(k+1)$; i.e. the anchor tuple does not increase the VC dimension. Instead, the cost of conditioning appears only through the "rarity" term $\log(e/p_{\mathbf{a}})$. The remainder of the proof of theorem 1.2 is devoted to showing that the rarity term can be absorbed by a suitable multiscale decomposition of the inclusion probabilities p_X.

Recall from lemma 2.2 that $p_X = \prod_{\ell=1}^{q} q_{X_{<\ell}}(x_\ell)$. The quantities $q_{X_{<\ell}}(x_\ell)$ will be grouped according to dyadic magnitude. For each ℓ, let $k_\ell \geq 0$ be the unique integer such that $2^{-k_\ell-1} < q_{X_{<\ell}}(x_\ell) \leq 2^{-k_\ell}$. We shall refer to the vector $\mathbf{k} = (k_1, \ldots, k_q)$ as the dyadic profile of the tuple X. Define

$$\Gamma_{\mathbf{k}} \doteq \left\{ X \in \mathcal{X}^q \mid 2^{-k_\ell-1} < q_{X_{<\ell}}(x_\ell) \leq 2^{-k_\ell} \text{ for every } \ell \right\}$$

The profile partitions all tuples with positive inclusion probability. For a profile $\mathbf{k}$, write $i \doteq \sum_{\ell=1}^{q} k_\ell$. Since $q_{X_{<\ell}}(x_\ell) \leq 2^{-k_\ell}$, we obtain $p_X = \prod_{l=1}^{q} q_{X_{<\ell}}(x_\ell) \leq 2^{-i}$ for every $X \in \Gamma_{\mathbf{k}}$. Substituting this into lemma 2.1 yields the basic profile decomposition.

Lemma 2.5. *For every integer $q \geq 1$, $R_q \leq \sum_{\mathbf{k}} 2^{-2i} P^q(\Gamma_{\mathbf{k}})$, where $i = \sum_{\ell=1}^{q} k_\ell$.*

Proof. By lemma 2.1, $R_q = \int_{\mathcal{X}^q} p_X^2 dP^q(X)$. Decompose $\mathcal{X}^q$ into dyadic profile classes: $R_q = \sum_{\mathbf{k}} \int_{\Gamma_{\mathbf{k}}} p_X^2 dP^q(X)$. On $\Gamma_{\mathbf{k}}$, $p_X^2 \leq 2^{-2i}$. Therefore,

$$\int_{\Gamma_{\mathbf{k}}} p_X^2 dP^q(X) \leq 2^{-2i} P^q(\Gamma_{\mathbf{k}}).$$

Summing over profiles gives the result. □

The problem is now reduced to estimating the profile probabilities $P^q(\Gamma_{\mathbf{k}})$. Currently, the rarity term from lemma 2.4 appears uncontrolled. The important observation is that on a fixed profile, the rarity of each conditioning event is already determined by the previous levels.

For a profile $\mathbf{k}$, define the partial sums $K_\ell \doteq \sum_{j<\ell} k_j, 1 \leq \ell \leq q$. The next lemma shows that every profile automatically controls the rarity of its own conditioning events.

Lemma 2.6. *Let $X \in \Gamma_{\mathbf{k}}$. Then for every ℓ, $p_{X_{<\ell}} > 2^{-K_\ell-(\ell-1)}$. Consequently,*

$$\log\left(\frac{e}{p_{X_{<\ell}}}\right) \leq C(K_\ell + \ell)$$

for a universal constant C.

Proof. By the chain rule,

$$p_{X_{<\ell}} = \prod_{j<\ell} q_{X_{<j}}(x_j).$$

Since $X \in \Gamma_{\mathbf{k}}$, $q_{X_{<j}}(x_j) > 2^{-k_j-1}$. Therefore,

$$p_{X_{<\ell}} > \prod_{j<\ell} 2^{-k_j-1} = 2^{-K_\ell-(\ell-1)}.$$

Taking logarithms gives

$$\log\left(\frac{e}{p_{X_{<\ell}}}\right) \le 1 + (K_\ell + \ell - 1)\log 2 \le C(K_\ell + \ell).$$

□

We now estimate the profile probabilities $P^q(\Gamma_{\mathbf{k}})$. The main point is that lemma 2.6 converts the anchor rarity term in lemma 2.4 into a deterministic function of the profile. Fix a profile $\mathbf{k} = (k_1, \ldots, k_q)$, and recall the partial sums $K_\ell \doteq \sum_{j<\ell} k_j$. For a prefix $X_{<\ell} = (x_1, \ldots, x_{\ell-1})$, define $B_\ell(X_{<\ell}) \doteq \left\{x \in \mathcal{X} \mid 2^{-k_\ell-1} < q_{X_{<\ell}}(x) \le 2^{-k_\ell}\right\}$. Thus, $X \in \Gamma_{\mathbf{k}}$ iff $x_\ell \in B_\ell(X_{<\ell})$ for every ℓ. The following lemma provides an important estimate for our main proof.

Lemma 2.7. *There exists a universal constant C_2 such that for every profile $\mathbf{k}$, every level $1 \le \ell \le q$, and every prefix $X_{<\ell}$ consistent with the profile up to level $\ell - 1$,*

$$P(B_\ell(X_{<\ell})) \le \frac{C_2(d(k_\ell+1) + K_\ell + \ell)2^{k_\ell}}{m}.$$

Proof. Fix a profile-consistent prefix $X_{<\ell}$. Applying lemma 2.4 with anchor tuple $\mathbf{a} = X_{<\ell}$, gives

$$P(B_\ell(X_{<\ell})) \le \frac{C_1(d(k_\ell+1) + \log\big(e/p_{X_{<\ell}}\big))2^{k_\ell}}{m}.$$

Since the prefix is profile-consistent, lemma 2.6 yields $\log\big(e/p_{x_{<\ell}}\big) \le C(K_\ell + \ell)$. Substituting this estimate gives

$$P(B_\ell(X_{<\ell})) \le \frac{C_2(d(k_\ell+1) + K_\ell + \ell)2^{k_\ell}}{m}$$

as desired. □

The importance of lemma 2.7 is that the right-hand side depends only on the profile and not on the particular prefix. Consequently, every profile-consistent prefix has the same one-step upper bound, permitting an iterative application of Fubini's theorem.

Proposition 2.8. *For every profile $\mathbf{k} = (k_1, \ldots, k_q)$,*

$$P^q(\Gamma_{\mathbf{k}}) \le \prod_{\ell=1}^{q} \frac{C_2(d(k_\ell+1) + K_\ell + \ell)2^{k_\ell}}{m}.$$

Proof. For $0 \le \ell \le q$, let $\Gamma_{\mathbf{k}}^{(\ell)}$ denote the set of prefixes $(x_1, \ldots, x_\ell)$ satisfying the profile conditions up to level ℓ. Thus, $\Gamma_{\mathbf{k}}^{(0)} = \{\varnothing\}$, $\Gamma_{\mathbf{k}}^{(q)} = \Gamma_{\mathbf{k}}$. For $\ell \ge 1$,

$$P^\ell(\Gamma_{\mathbf{k}}^{(\ell)}) = \int_{\Gamma_{\mathbf{k}}^{(\ell-1)}} P(B_\ell(X_{<\ell}))dP^{\ell-1}(X_{<\ell}).$$

By lemma 2.7, $P(B_\ell(X_{<\ell})) \le M_\ell$ where $M_\ell \doteq \frac{C_2(d(k_\ell+1)+K_\ell+\ell)2^{k_\ell}}{m}$. Since M_ℓ is independent of the particular prefix, $P^\ell(\Gamma_{\mathbf{k}}^{(\ell)}) \le M_\ell P^{\ell-1}(\Gamma_{\mathbf{k}}^{(\ell-1)})$. Iterating from $\ell = 1$ to q yields

$$P^q(\Gamma_{\mathbf{k}}) = P^q(\Gamma_{\mathbf{k}}^{(q)}) \le \prod_{\ell=1}^{q} M_\ell.$$

Substituting the definition of M_ℓ gives the claim. □

Corollary 2.9. *For every profile* $\mathbf{k}$*,*

$$2^{-2i}P^q(\Gamma_{\mathbf{k}}) \le \left(\frac{C_2}{m}\right)^q 2^{-i} \prod_{\ell=1}^{q}(d(k_\ell+1)+K_\ell+\ell),$$

where $i = \sum_{\ell=1}^{q} k_\ell$*.*

Proof. By proposition 2.8,

$$P^q(\Gamma_{\mathbf{k}}) \le \left(\frac{C_2}{m}\right)^q \prod_{\ell=1}^{q}(d(k_\ell+1)+K_\ell+\ell)2^{k_\ell}.$$

Since $\prod_{\ell=1}^{q} 2^{k_\ell} = 2^i$, multiplying by 2^{-2i} yields

$$2^{-2i}P^q(\Gamma_{\mathbf{k}}) \le \left(\frac{C_2}{m}\right)^q 2^{-i} \prod_{\ell=1}^{q}(d(k_\ell+1)+K_\ell+\ell).$$

□

Combining corollary 2.9 with lemma 2.5 gives

$$R_q \le \left(\frac{C_2}{m}\right)^q \sum_{\mathbf{k}} 2^{-i} \prod_{\ell=1}^{q}(d(k_\ell+1)+K_\ell+\ell).$$

All that remains is to bound the profile sum by $(C(d+q))^q$.

Proposition 2.10. *For every integer* $q \ge 1$*,* $\widetilde{B}_q(d) \le (16(d+q))^q$*, where*

$$\widetilde{B}_q(d) \doteq \sum_{\mathbf{k}\in\mathbb{N}_0^q} 2^{-i} \prod_{\ell=1}^{q}\big(d(k_\ell+1)+K_\ell+\ell\big).$$

Proof. Substitute $a_\ell \doteq k_\ell + 1 \geq 1$ and set $A \doteq \sum_{\ell=1}^{q} a_\ell$, so that $i = A - q$. Define $M_\ell \doteq \sum_{j<\ell} a_j$. Since $K_\ell = \sum_{j<\ell}(a_j - 1) = M_\ell - (\ell - 1)$, each factor satisfies

$$d(k_\ell + 1) + K_\ell + \ell = da_\ell + M_\ell + 1 \leq da_\ell + A,$$

using $M_\ell + 1 \leq A$. Fixing A and applying AM-GM,

$$\prod_{\ell=1}^{q}(da_\ell + A) \leq \left(\frac{1}{q}\sum_{\ell=1}^{q}(da_\ell + A)\right)^q = \left(\frac{(d+q)A}{q}\right)^q.$$

The number of q-tuples $(a_1, \ldots, a_q)$ of positive integers summing to A is $\binom{A-1}{q-1}$. Using $2^{-i} = 2^{q-A}$ and grouping by A,

$$\widetilde{B}_q(d) \leq \sum_{A=q}^{\infty} 2^{q-A}\binom{A-1}{q-1}\left(\frac{(d+q)A}{q}\right)^q.$$

Applying $\binom{A-1}{q-1} \leq A^{q-1}/(q-1)!$ and extending the sum to $A \geq 1$,

$$\widetilde{B}_q(d) \leq \frac{2^q(d+q)^q}{q^q(q-1)!}\sum_{A=1}^{\infty} A^{2q-1}2^{-A}.$$

The bound $\sum_{A=1}^{\infty} A^r 2^{-A} \leq 2^r r!$ with $r = 2q-1$ gives $\sum_{A=1}^{\infty} A^{2q-1}2^{-A} \leq 2^{2q-1}(2q-1)!$, and therefore

$$\widetilde{B}_q(d) \leq \frac{2^{3q-1}(d+q)^q(2q-1)!}{q^q(q-1)!}.$$

Since $(2q-1)!/(q-1)! = q(q+1)\cdots(2q-1) \leq (2q)^q$, we have $(2q-1)!/(q^q(q-1)!) \leq 2^q$. Hence $\widetilde{B}_q(d) \leq 2^{4q-1}(d+q)^q \leq (16(d+q))^q$. □

Substituting proposition 2.10 into the bound of lemma 2.5 together with corollary 2.9 gives $R_q \leq (16C_2(d+q)/m)^q$. Absorbing $16C_2$ into the universal constant C completes the proof of theorem 1.2.

References

[AAHLZ24] Ishaq Aden-Ali, Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner?, 2024.

[Wel12] Emo Welzl. Computational geometry: Chapter 15 – epsilon nets. Lecture notes, ETH Zürich, 2012. Lemma 15.10.