

THE SHARP TAIL OF UNIFORM STABILITY

Pahan Dewasurendra
Johns Hopkins University

Nuwan Dewasurendra
MiraCosta College

ABSTRACT

Uniform stability controls how much one training example can change the loss at any test point. A new logarithmic-free upper bound shows that a γ -uniformly stable algorithm with loss in $[0, L]$ has generalization gap at most

$$O\left(\gamma \log(1/\delta) + L\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

with probability $1 - \delta$. Whether an actual bounded-loss learning algorithm can realize the linear dependence on $\log(1/\delta)$ has remained open. The known construction realizes it only for auxiliary weakly dependent random variables whose pointwise range grows with n . The known learning lower bound holds only at constant probability.

We close this gap. For every n , stability level γ , and loss bound L , we construct one deterministic γ -uniformly stable learning problem whose tail satisfies, simultaneously for $1 \leq p \leq cn$,

$$\mathbb{P}\left(R(A_S) - R_S(A_S) \geq c' \min\left\{L, \gamma p + L\sqrt{p/n}\right\}\right) \geq e^{-p}.$$

The construction is ordinary bounded absolute-loss regression with constant labels. Its key is a multiscale collection of rare Rademacher features. A coordinatewise ramp is stable in sup norm, while an odd symmetrized maximum converts a unique extreme feature into a gap of order γp without violating the loss bound. Geometrically spaced ramps put all confidence levels into the same problem. Together with the logarithmic-free upper bound, this determines the optimal high-probability and moment dependence of uniform stability up to universal constants.

1 INTRODUCTION

Algorithmic stability explains generalization through the behavior of the learning rule rather than through uniform convergence over a hypothesis class. A learning algorithm A is γ -uniformly stable when replacing one training example changes its loss at every test point by at most γ (Bousquet & Elisseeff, 2002). This property underlies guarantees for regularized empirical risk minimization, gradient methods, and stochastic optimization (Shalev-Shwartz et al., 2010; Hardt et al., 2016). Its expectation guarantee is simple and sharp: $|\mathbb{E}[R(A_S) - R_S(A_S)]| \leq \gamma$. Obtaining the correct high-probability guarantee took a longer sequence of results.

The classical bounded-differences argument gives a deviation term of order $\gamma\sqrt{n\log(1/\delta)}$ (Bousquet & Elisseeff, 2002). Feldman & Vondrák (2018) replaced the factor $\gamma\sqrt{n}$ by $\sqrt{\gamma L}$, and Feldman & Vondrák (2019) obtained a nearly optimal bound with logarithmic overhead in n . Bousquet et al. (2020) reduced the learning problem to a moment inequality for weakly interacting functions. Their result left one factor $\log n$ in the stability term. Very recently, Nguyen-Cung & Nguyen (2026) removed that factor and proved

$$\|R(A_S) - R_S(A_S)\|_p \leq 33p\gamma + L\sqrt{\frac{2p}{n}}, \quad p \geq 2. \quad (1)$$

The trivial range bound then caps the right side by L . Equation (1) gives the clean tail scale

$$\min\left\{L, \gamma \log(1/\delta) + L\sqrt{\frac{\log(1/\delta)}{n}}\right\}. \quad (2)$$

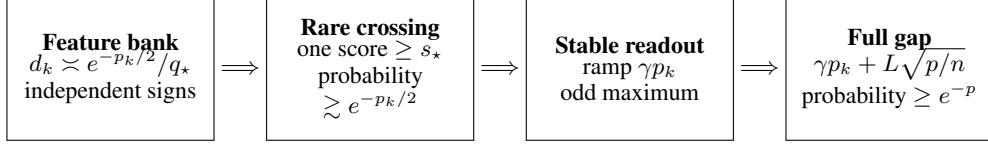


Figure 1: The multiscale mechanism. Reciprocal-probability feature replication makes a rare large-deviation crossing visible. A coordinatewise stable ramp and the odd maximum turn that crossing into empirical bias. Independent memory and sampling terms fill the constant-confidence and square-root parts of the tail. Geometric choices of p_k put every $p \leq cn$ in one learner.

The lower-bound side has not matched this progress. Bousquet et al. (2020) constructed weakly interacting functions with p th moment $\Omega(pn\gamma + L\sqrt{pn})$. Their functions have pointwise magnitude of order $n\gamma$, so the construction does not come from a learning algorithm with a uniform loss bound L . They explicitly asked whether the same behavior is possible for the functions induced by a uniformly stable learner. Liu & Lu (2020) later constructed a bounded-loss learner with gap $\Omega(\gamma + L/\sqrt{n})$ at constant probability, and left the dependence on small failure probability open. The new upper-bound paper also separates its result from exactly this learning lower-bound question (Nguyen-Cung & Nguyen, 2026). Our lower-bound theorem and proof are independent of that preprint. Equation (1) is used only to identify the matching upper scale.

We show that every term in (2) is necessary. More strongly, one finite-dimensional problem realizes the full curve at all confidence levels e^{-p} with $p \leq cn$. This is not a different problem chosen for each p .

Why the obvious quadratic construction fails. Let $X_1, \dots, X_n$ be Rademacher signs, $T = \sum_i X_i$, and let a learner predict with amplitude γT . A linear loss has generalization gap $\gamma T^2/n$, whose p th scale is γp . The prediction amplitude at that event is $\gamma\sqrt{np}$. Clipping it to the loss range L activates when $p \gtrsim L^2/(\gamma^2 n)$. This is exactly where the term γp begins to dominate $L\sqrt{p/n}$. Thus clipping destroys the part of the tail that needs to be proved.

The new mechanism. We replace one feature by many independent Rademacher features. At a chosen scale, exponentially many coordinates are placed just below a fixed large-deviation threshold. A ramp of width $\Theta(p)$ gives amplitude $\Theta(\gamma p)$ only when one coordinate crosses that threshold. The ramp is coordinatewise γ -stable regardless of the number of coordinates. The loss reads the coordinates through the odd function

$$H_a(x) = \max_j a_j x_j - \max_j (-a_j x_j). \quad (3)$$

It is centered under a symmetric population and is 2-Lipschitz in a under the sup norm. When there is one dominant coordinate, its empirical average is bounded away from zero, so it contributes $\Theta(\gamma p)$ to the gap. Coordinates with smaller ramps can be present without changing the sign of the effect. Coordinates with larger ramps are absent on the clean extreme event.

We use geometrically spaced ramp widths. The event at scale p_k has probability much larger than e^{-p_k} , which leaves enough probability to intersect it with an ordinary sampling-deviation event. A separate stable memorization term handles constant confidence and the range-saturated regime. All components are combined in a scalar predictor under absolute loss.

Contributions.

- We give the first bounded-loss uniformly stable learner with an $\Omega(\gamma \log(1/\delta))$ high-probability generalization gap.
- A single problem realizes $\Omega(\min\{L, \gamma p + L\sqrt{p/n}\})$ simultaneously for all $1 \leq p \leq cn$. Consequently, the moment upper bound in (1) is sharp for actual learning algorithms.
- The construction uses deterministic learning and standard absolute regression loss. We prove uniform stability over all replacement datasets, not only on the high-probability event.

Related extensions. Several recent lines change the assumptions rather than the worst-case lower-bound question. Klochov & Zhivotovskiy (2021) obtain faster excess-risk rates under a Bernstein condition. Zhou et al. (2023) derive PAC-Bayes bounds for randomized uniformly stable algorithms under a sub-exponential condition on their random stability parameter. Yuan & Li (2022) use subbagging to boost confidence for randomized algorithms under the weaker, distribution-dependent notion of L_2 stability. Fan & Lei (2024) introduce pointwise uniform stability and derive function-value and gradient generalization bounds, with applications to SGD. In a complementary 2026 direction, Lei et al. (2026) replace bounded differences by finite- L_p replace-one envelopes and obtain Gaussian-plus-polynomial upper tails for unbounded losses. These results provide sharper conclusions from additional structure or modified algorithms. They do not give a high-probability lower bound for the ordinary deterministic uniform-stability condition (4). Our result concerns exactly that worst-case condition and the bounded-loss minimax gap left open by Bousquet et al. (2020), Liu & Lu (2020), and Nguyen-Cung & Nguyen (2026).

2 SETUP AND RESULT

Let $S = (Z_1, \dots, Z_n) \sim P^n$. An algorithm A maps S to a predictor A_S . For a loss ℓ with values in $[0, L]$, define

$$R(A_S) = \mathbb{E}_{Z \sim P} \ell(A_S, Z), \quad R_S(A_S) = \frac{1}{n} \sum_{i=1}^n \ell(A_S, Z_i), \quad G_S = R(A_S) - R_S(A_S).$$

The algorithm is γ -uniformly stable if, whenever S and S' differ in one coordinate,

$$\sup_z |\ell(A_S, z) - \ell(A_{S'}, z)| \leq \gamma. \quad (4)$$

Our theorem is stated with unspecified universal constants. The appendix tracks concrete constants for the construction.

Theorem 1 (Sharp lower tail). *There are universal constants $c_0, c_1 > 0$ and $n_0 \in \mathbb{N}$ with the following property. For every $n \geq n_0$, $L > 0$, and $0 \leq \gamma \leq L$, there are a finite input space, a distribution P , and one deterministic γ -uniformly stable absolute-loss regression algorithm with loss in $[0, L]$ such that, simultaneously for every real $p \in [1, c_0 n]$,*

$$\mathbb{P}_{S \sim P^n} \left(G_S \geq c_1 \min \left\{ L, \gamma p + L \sqrt{p/n} \right\} \right) \geq e^{-p}. \quad (5)$$

The regression labels are identically zero.

The problem may depend on (n, L, γ) , as is standard for a minimax lower bound. Crucially, it does not depend on p or δ . The simultaneous statement immediately gives moment sharpness.

Corollary 2 (Moment sharpness). *For the problem in Theorem 1 and every $p \in [2, c_0 n]$,*

$$\|G_S\|_p \geq c_2 \min \left\{ L, \gamma p + L \sqrt{p/n} \right\} \quad (6)$$

for a universal $c_2 > 0$.

Indeed, (5) implies $\|G_S\|_p \geq e^{-1}$ times its displayed threshold. Combining Corollary 2 with (1) and the range bound yields the minimax law

$$\sup_{(P, A, \ell)} \|G_S\|_p \asymp \min \left\{ L, \gamma p + L \sqrt{p/n} \right\}, \quad 2 \leq p \leq c_0 n, \quad (7)$$

where the supremum is over γ -stable algorithms with loss in $[0, L]$. The upper and lower constants in (7) are universal.

3 THE MULTISCALE LEARNER

We now define the single problem used for every p . A training input is

$$Z = (X, J, \Sigma, U).$$

All components are independent. The vector X consists of independent Rademacher coordinates, J is uniform on $[D]$, and Σ, U are Rademacher signs. The dimension of X is divided into groups indexed by a geometric set of scales.

Let

$$p_k = 16 \cdot 4^k, \quad r_k = p_k/16 = 4^k, \quad p_k \leq \min\{n/64, L/\gamma\}, \quad (8)$$

where $L/0 = +\infty$. Let B_n be a sum of n independent Rademacher signs. Choose the first attainable score $s_\star \geq n/4$ and set

$$q_\star = \mathbb{P}(B_n \geq s_\star), \quad d_k = \left\lfloor \frac{8e^{-p_k/2}}{q_\star} \right\rfloor, \quad t_k = s_\star - 2r_k. \quad (9)$$

Group k contains d_k independent coordinates. For a sample, write $S_{kj} = \sum_{i=1}^n X_{i,kj}$ and define the ramp amplitude

$$a_{kj}(S) = \min \left\{ \gamma r_k, \frac{\gamma}{2} (S_{kj} - t_k)_+ \right\}. \quad (10)$$

If the set of scales is empty, X has one dummy coordinate with amplitude zero. The total feature dimension is finite and typically exponential in n because $q_\star = e^{-\Theta(n)}$. Section 7 explains why this multiplicity is used.

The memory component uses $D = 8n^2$ cells. Let

$$b_j(S) = c_{\text{mem}} \operatorname{sgn} \left(\sum_{i: J_i=j} \Sigma_i \right), \quad c_{\text{mem}} = \min\{\gamma/4, L/8\}, \quad (11)$$

with $\operatorname{sgn}(0) = 0$. Collect all ramp amplitudes in a vector $a(S)$ and use the symmetrized maximum H_a from (3). The predictor is

$$h_S(X, J, \Sigma, U) = \frac{L}{2} - \frac{1}{4} H_{a(S)}(X) - b_J(S) \Sigma - \frac{L}{4} U. \quad (12)$$

The regression label is zero and the loss is absolute error. We will show that $h_S \in [0, L]$, so the loss equals h_S itself.

The four pieces in (12) have distinct roles. The center $L/2$ enforces a valid loss range. The symmetrized maximum produces the γp tail. The cell memory produces a constant-order stability gap and handles scales below 16 or above L/γ . The last sign produces the ordinary $L\sqrt{p/n}$ sampling tail.

4 STABILITY AND CENTERING

Two elementary Lipschitz properties make the construction work.

Lemma 3 (Symmetrized maximum). *For nonnegative vectors a, a' and every sign vector x ,*

$$H_a(-x) = -H_a(x), \quad |H_a(x)| \leq 2\|a\|_\infty, \quad |H_a(x) - H_{a'}(x)| \leq 2\|a - a'\|_\infty. \quad (13)$$

The proof applies the ℓ_∞ Lipschitz property of a maximum twice. Replacing one training point changes every score S_{kj} by at most two. Hence (10) gives $\|a(S) - a(S')\|_\infty \leq \gamma$. The $H/4$ term changes by at most $\gamma/2$. At a fixed memory cell, a replacement can change b_j from $-c_{\text{mem}}$ to c_{mem} , so the memory term changes by at most $2c_{\text{mem}} \leq \gamma/2$. The U term does not depend on the sample. This proves (4).

The largest ramp cap is at most $L/16$. Lemma 3 gives $|H_a| \leq L/8$. Together with $c_{\text{mem}} \leq L/8$, the sample-dependent offset from $L/2$ in (12) is at most

$$L/32 + L/8 + L/4 = 13L/32.$$

Thus $h_S \in [3L/32, 29L/32] \subset [0, L]$.

Conditional on the training sample, a fresh X is symmetric. Lemma 3 therefore gives $\mathbb{E} H_{a(S)}(X) = 0$. A fresh Σ and a fresh U are centered as well, so

$$R(A_S) = L/2 \quad (14)$$

for every sample. Consequently,

$$G_S = \frac{1}{4n} \sum_{i=1}^n H_{a(S)}(X_i) + \frac{1}{n} \sum_{i=1}^n b_{J_i}(S) \Sigma_i + \frac{L}{4n} \sum_{i=1}^n U_i. \quad (15)$$

This exact identity is the bridge from rare features to generalization.

5 RARE EXTREMES CREATE THE LINEAR TAIL

For each scale k , let E_k be the event that exactly one coordinate in group k has score at least s_* , and every other coordinate in group k or a larger group $\ell > k$ has score strictly below its threshold t_ℓ . Smaller groups are unrestricted.

Lemma 4 (Clean extreme). *For every scale in (8),*

$$\mathbb{P}(E_k) \geq 6e^{-p_k/2}. \quad (16)$$

On E_k ,

$$\frac{1}{n} \sum_{i=1}^n H_{a(S)}(X_i) \geq \frac{3}{32} \gamma r_k = \frac{3}{512} \gamma p_k. \quad (17)$$

We give the main argument because it contains the new idea. Hoeffding's inequality yields $q_* \leq e^{-n/32}$. Since $p_k \leq n/64$, the floor in (9) is negligible at the required constant scale and

$$d_k q_* \geq 7e^{-p_k/2}. \quad (18)$$

Moving the threshold down by r_k binomial steps multiplies its upper tail by at most $(r_k + 1)2^{r_k} \leq 4^{r_k} < e^{p_k/8}$. Therefore, the expected number of coordinates in group k at or above the forbidden threshold t_k , and similarly in every larger group, is at most

$$8e^{-p_k/2+p_k/8} = 8e^{-3p_k/8}. \quad (19)$$

The scales grow by four, so the sum of these expectations from group k upward is below 0.021. Independence across coordinates and a union bound with (18) prove (16).

On E_k , the exceptional coordinate reaches its cap $a = \gamma r_k$. Every coordinate in a smaller group has amplitude at most $a/4$, while every other coordinate in the same or a larger group has amplitude zero. Let x_i^* be the exceptional coordinate. If $x_i^* = 1$, then $H_a(X_i) \geq a - a/4$. If $x_i^* = -1$, then $H_a(X_i) \geq -a$. Since $\sum_i x_i^* \geq s_* \geq n/4$, summing these inequalities gives

$$\frac{1}{n} \sum_i H_a(X_i) \geq a \left(\frac{1}{4} - \frac{1+1/4}{8} \right) = \frac{3a}{32}.$$

This proves (17). The smaller scales may fire, but the geometric cap makes them too weak to reverse the contribution of the exceptional feature.

The role of the factor $e^{-p_k/2}$ is deliberate. It is more probability than the theorem needs. We spend the surplus by intersecting E_k with the independent sampling event below.

6 PROOF OF THE SHARP TAIL

We use a standard one-sided anti-concentration fact. A self-contained block proof from the Paley–Zygmund inequality is included in Appendix A.6.

Lemma 5 (Rademacher lower tail). *For all sufficiently large n , every Rademacher sum T_n and $1 \leq p \leq n$ satisfy*

$$\mathbb{P}\left(T_n \geq \frac{1}{64} \sqrt{np}\right) \geq \frac{1}{8} e^{-p/8}. \quad (20)$$

For $1 \leq p < 16$, the probability on the left is at least 0.45.

The constants are elementary. Paley–Zygmund and symmetry give $\mathbb{P}(V_m \geq \sqrt{m/2}) \geq 1/24$ for a Rademacher block sum. For $p \geq 32 \log 24$, split the signs into $q = \lfloor p/(16 \log 24) \rfloor$ blocks and require this event in every block, with a nonnegative leftover. The resulting threshold is at least $\sqrt{np}/64$ and the probability is at least $(1/2)24^{-q} \geq (1/8)e^{-p/8}$. One block handles $16 \leq p < 32 \log 24$. For $p < 16$, the maximum binomial atom bound $n^{-1/2}$ gives the stated 0.45. Appendix A.6 records the arithmetic.

Let F be the event that the memory indices $J_1, \dots, J_n$ are distinct. Since $D = 8n^2$,

$$\mathbb{P}(F) \geq 1 - \frac{n(n-1)}{2D} \geq \frac{15}{16}. \quad (21)$$

On F , each observed cell contains one sign. Hence $b_{J_i}(S) = c_{\text{mem}} \Sigma_i$ and the middle term in (15) is exactly c_{mem} .

Fix $p \in [16, c_0 n]$, where $c_0 \leq 1/64$. If a scale is available, choose the largest available $p_k \leq p$. In particular, this is the terminal scale when p exceeds the largest constructed scale. The geometric spacing gives

$$p_k \geq \frac{1}{4} \min\{p, L/\gamma\}. \quad (22)$$

On the independent intersection of E_k , F , and the event in Lemma 5 for $\sum_i U_i$, equations (15) and (17) give

$$G_S \geq c \min\{L, \gamma p\} + cL\sqrt{p/n}. \quad (23)$$

Its probability is at least

$$6e^{-p_k/2} \cdot \frac{15}{16} \cdot \frac{1}{8} e^{-p/8} \geq e^{-p}.$$

The inequality uses $p_k \leq p$ and $p \geq 16$.

If no scale is available, then $L/\gamma < 16$. In that case $c_{\text{mem}} \geq L/64$, which already gives a constant fraction of $\min\{L, \gamma p\}$. For $p < 16$, the same memory term gives a constant fraction of $\min\{L, \gamma p\}$. We additionally require that no ramp is active. Equation (19) shows that this event has probability above 0.97. A constant-scale version of Lemma 5, together with (21), leaves probability above e^{-p} and gives the required $L\sqrt{p/n}$ term. This proves Theorem 1.

7 DISCUSSION

The auxiliary lower bound is now realizable. The construction of Bousquet et al. (2020) produces the quadratic Rademacher chaos $\gamma(T_n^2 - n)$, which has the right moments but arises from functions whose pointwise range is $\Theta(n\gamma)$. The ramps in our construction move the large deviation into the feature index. At any fixed test point, the loss reads a maximum and remains bounded. At the training sample, a rare coordinate carries a persistent empirical bias. This separates pointwise range from aggregate tail size.

All confidence levels require multiple scales. A single ramp can match one prescribed p , which would prove a pointwise minimax lower bound. The geometric family is stronger. Smaller ramps have at most one quarter of the target amplitude, while larger ramps are inactive with overwhelming probability. This scale separation is why the same learner matches the full tail rather than a confidence level fixed in advance.

Implication for stability-based analyses. A common analysis of an optimizer first proves a uniform stability constant and then invokes a distribution-free generalization theorem that sees only (γ, L) . Our lower bound shows that this black-box second step must retain the $\gamma \log(1/\delta)$ term. A sharper guarantee for a particular optimizer must use more information, such as curvature, data-dependent stability, algorithmic randomness, or representation constraints.

The loss is standard, the learner is adversarial. Our predictor is scalar and is evaluated by absolute loss against the constant label zero. The construction is nonetheless a worst-case learner with exponentially many Rademacher features. This multiplicity is deliberate. A coordinate has upper-tail probability $q_\star = e^{-\Theta(n)}$, and using on the order of $1/q_\star$ independent coordinates

makes one crossing occur at a controlled probability. Within an independent-threshold bank of this form, the union bound shows that polynomial multiplicity cannot suffice. Probability e^{-p} requires $d \geq e^{-p}/q_* \geq e^{n/64}$ when $p \leq n/64$. We do not claim that exponential dimension is necessary outside this mechanism. It is the price paid by this construction to encode all confidence levels in one problem. This is appropriate for determining what uniform stability alone can guarantee because that condition places no restriction on dimension. Additional structure, such as convex optimization, a fixed low-dimensional representation, or distributional stability, may yield smaller tails. The theorem shows that such improvements cannot follow from uniform stability and bounded loss by themselves.

Confidence range. The interval $p \leq cn$ is the natural nondegenerate range for bounded i.i.d. sampling. It is also the range in which the weak-dependence lower bound of Bousquet et al. (2020) has its two-level form. Beyond it, the loss-range cap dominates and constants depend on how the extreme endpoint is parameterized. Theorem 1 covers every confidence level used in the usual exponential generalization regime.

REPRODUCIBILITY STATEMENT

The proof is finite and nonasymptotic. The supplement gives every omitted constant calculation. It also provides three short audit programs: `extreme_ramp_check.py` evaluates the exact binomial probabilities, `multiscale_check.py` checks every scale and case split, and `stability_bruteforce.py` exhausts the replacement inequalities on small instances. These computations check the proof and are not used as assumptions.

REFERENCES

- Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 610–626, 2020.
- Jun Fan and Yunwen Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.
- Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9770–9780, 2018.
- Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1270–1279, 2019.
- Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, 2016.
- Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Qianqian Lei, Soham Bonnerjee, Yuefeng Han, and Wei Biao Wu. Stability beyond bounded differences: Sharp generalization bounds under finite l_p moments. In *Proceedings of the 43rd International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, 2026. arXiv:2606.06855.
- Qinghua Liu and Zhou Lu. A tight lower bound for uniformly stable algorithms. *arXiv preprint arXiv:2012.13326*, 2020.
- Thanh Nguyen-Cung and Binh T. Nguyen. Logarithmic-free moment and generalization bounds for uniformly stable algorithms. *arXiv preprint arXiv:2608.09870*, 2026.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

Xiao-Tong Yuan and Ping Li. Boosting the confidence of generalization for l_2 -stable randomized learning algorithms. *arXiv preprint arXiv:2206.03834*, 2022.

Sijia Zhou, Yunwen Lei, and Ata Kabán. Toward better PAC-Bayes bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29602–29614, 2023.

A DETAILED PROOFS

This appendix proves the claims in the main text with one concrete choice of the scale constants. No optimization of numerical constants is intended.

A.1 NOTATION AND ELEMENTARY BOUNDS

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher signs and $B_n = \sum_i \varepsilon_i$. Define

$$k_\star = \left\lceil \frac{5n}{8} \right\rceil, \quad s_\star = 2k_\star - n, \quad q_\star = \mathbb{P}(B_n \geq s_\star).$$

Then

$$\frac{1}{4} \leq \frac{s_\star}{n} \leq \frac{1}{4} + \frac{2}{n}. \quad (24)$$

Hoeffding’s inequality gives

$$q_\star \leq \exp\left(-\frac{s_\star^2}{2n}\right) \leq e^{-n/32}. \quad (25)$$

For $r \leq n/1024$, put

$$q(r) = \mathbb{P}(B_n \geq s_\star - 2r).$$

The event on the right contains the upper tail at $k_\star$ and at most r additional binomial levels. Moving down one level multiplies the point mass by

$$\frac{k}{n - k + 1} \leq \frac{5n/8 + 1}{3n/8 - r} < 2$$

for all sufficiently large n . Since $q_\star$ is at least the mass at $k_\star$, we obtain

$$\frac{q(r)}{q_\star} \leq (r + 1)2^r \leq 4^r. \quad (26)$$

The last inequality holds for every integer $r \geq 1$.

A.2 THE SYMMETRIZED MAXIMUM

Proof of Lemma 3. Write $M_a(x) = \max_j a_j x_j$. Then

$$H_a(x) = M_a(x) - M_a(-x).$$

This immediately gives $H_a(-x) = -H_a(x)$. Since each $a_j x_j$ lies in $[-\|a\|_\infty, \|a\|_\infty]$, the range bound follows. Finally,

$$|M_a(x) - M_{a'}(x)| \leq \max_j |(a_j - a'_j)x_j| \leq \|a - a'\|_\infty.$$

Applying this inequality at x and $-x$ and using the triangle inequality proves the final claim. $\square$

A.3 THE LEARNER IS STABLE AND BOUNDED

We verify the claims around (12). Suppose S and S' differ at observation i . For every ramp coordinate,

$$|S_{kj} - S'_{kj}| = |X_{i,kj} - X'_{i,kj}| \leq 2.$$

The scalar map in (10) is $\gamma/2$ -Lipschitz. Therefore

$$\|a(S) - a(S')\|_\infty \leq \gamma. \quad (27)$$

Lemma 3 implies that the $H/4$ term changes by at most $\gamma/2$ at every test point.

Only the old and new memory cells can change. At either cell, b_j takes values in $\{-c_{\text{mem}}, 0, c_{\text{mem}}\}$, so

$$|b_j(S) - b_j(S')| \leq 2c_{\text{mem}} \leq \gamma/2. \quad (28)$$

For a fixed test point, only its own cell is read. Adding (27) and (28) proves γ -uniform stability.

If the scale set is nonempty, its largest element satisfies $p_K \leq L/\gamma$. Thus

$$\|a(S)\|_\infty \leq \gamma r_K = \frac{\gamma p_K}{16} \leq \frac{L}{16}. \quad (29)$$

The same bound is trivial if there are no ramps. Lemma 3 gives $|H_a(X)|/4 \leq L/32$. Since $|b_J \Sigma| \leq L/8$ and $|LU/4| = L/4$,

$$3L/32 \leq h_S(Z) \leq 29L/32.$$

This proves the range claim. With regression label zero, absolute loss is exactly $h_S(Z)$.

For a fresh input, $X \stackrel{d}{=} -X$, and Lemma 3 gives $\mathbb{E}[H_a(X) \mid S] = 0$. The fresh signs Σ and U are independent of S and centered. This proves (14) and (15).

A.4 PROBABILITY OF A CLEAN EXTREME

We prove the probability part of Lemma 4. For a scale $p_k = 16r_k \leq n/64$, equation (25) gives

$$q_\star \leq e^{-2p_k}. \quad (30)$$

By the definition of d_k ,

$$d_k q_\star \geq 8e^{-p_k/2} - q_\star \geq 7e^{-p_k/2}, \quad d_k q_\star \leq 8e^{-p_k/2}. \quad (31)$$

In particular, $d_k \geq 2$.

The clean event forbids every nonexceptional coordinate whose score is at least $t_k = s_\star - 2r_k$. Equations (26) and (31) show

$$d_k q(r_k) \leq 8e^{-p_k/2} 4^{r_k} \leq 8e^{-3p_k/8}, \quad (32)$$

because $4^{r_k} = e^{(\log 4)p_k/16} < e^{p_k/8}$. The same inequality holds at every larger scale. Since $p_{k+j} = 4^j p_k$,

$$\sum_{\ell \geq k} d_\ell q(r_\ell) \leq 8 \sum_{j \geq 0} \exp\left(-\frac{3}{8} 4^j p_k\right) < 0.021 \quad (33)$$

for $p_k \geq 16$.

Choose which coordinate in group k is the exceptional one. Independence of all Rademacher coordinates gives

$$\mathbb{P}(E_k) = d_k q_\star (1 - q(r_k))^{d_k - 1} \prod_{\ell > k} (1 - q(r_\ell))^{d_\ell} \quad (34)$$

$$\geq d_k q_\star \left(1 - \sum_{\ell \geq k} d_\ell q(r_\ell)\right) \quad (35)$$

$$\geq 7e^{-p_k/2} (1 - 0.021) \geq 6e^{-p_k/2}. \quad (36)$$

The second line is the union bound applied to all forbidden coordinates. This proves (16).

A.5 GAP ON THE CLEAN EXTREME EVENT

On E_k , let (k, j_*) denote the exceptional coordinate and put $a = \gamma r_k$. Its score is at least $s_* = t_k + 2r_k$, so its ramp reaches the cap a . Every other coordinate in group k or a larger group has amplitude zero. Every smaller scale has cap at most $a/4$.

Write $x_i^* = X_{i, kj_*}$ and let n_+ and n_- count its positive and negative signs. There is at least one zero-amplitude coordinate because $d_k \geq 2$. If $x_i^* = 1$, the first maximum in H is a and the second is at most $a/4$, so

$$H_{a(S)}(X_i) \geq 3a/4. \quad (37)$$

If $x_i^* = -1$, the second maximum is a and the first is at least zero, so

$$H_{a(S)}(X_i) \geq -a. \quad (38)$$

Let $\rho = (n_+ - n_-)/n$. Equations (37) and (38) imply

$$\frac{1}{n} \sum_i H_{a(S)}(X_i) \geq \frac{3a}{4} \frac{1+\rho}{2} - a \frac{1-\rho}{2} = a \left(\frac{7}{8}\rho - \frac{1}{8} \right). \quad (39)$$

On E_k , $\rho \geq s_*/n \geq 1/4$. The right side of (39) is increasing in ρ , so it is at least $3a/32$. Substituting $a = \gamma r_k = \gamma p_k/16$ proves (17).

A.6 RADEMACHER ANTI-CONCENTRATION

We give an elementary block proof of Lemma 5. If V_m is a sum of m independent Rademacher signs, then

$$\mathbb{E}V_m^2 = m, \quad \mathbb{E}V_m^4 = 3m^2 - 2m \leq 3m^2.$$

Paley-Zygmund applied to V_m^2 therefore gives

$$\mathbb{P}(|V_m| \geq \sqrt{m/2}) \geq \frac{1}{12}. \quad (40)$$

By symmetry, either one-sided event has probability at least $1/24$.

Partition the first qm signs of T_n into q blocks of size $m = \lfloor n/q \rfloor$, leaving at most $q-1$ signs. If every block sum is at least $\sqrt{m/2}$ and the leftover sum is nonnegative, then, whenever $q \leq n/2$,

$$T_n \geq q\sqrt{m/2} \geq \frac{1}{2}\sqrt{nq}. \quad (41)$$

The event has probability at least

$$\frac{1}{2} 24^{-q}. \quad (42)$$

We now track constants. First suppose $1 \leq p < 16$. The largest atom of T_n is at most $n^{-1/2}$ by the central binomial coefficient bound. Positive attainable values are spaced by two, so there are at most $\sqrt{n}/32 + 1$ of them below $\sqrt{n}/16$. Symmetry gives

$$\mathbb{P}(T_n \geq \sqrt{n}/16) \geq \frac{1}{2} - \frac{1}{2\sqrt{n}} - \left(\frac{1}{32} + \frac{1}{\sqrt{n}} \right) = \frac{15}{32} - \frac{3}{2\sqrt{n}} \geq 0.45 \quad (43)$$

for all sufficiently large n . Since $p < 16$, the threshold is at least $\sqrt{np}/64$.

For $16 \leq p < 32 \log 24$, use one block. Equations (40) and symmetry give

$$\mathbb{P}(T_n \geq \sqrt{n/2}) \geq 1/24 \geq (1/8)e^{-p/8},$$

and $\sqrt{n/2} \geq \sqrt{np}/64$. Finally, suppose $32 \log 24 \leq p \leq n$ and take

$$q = \left\lfloor \frac{p}{16 \log 24} \right\rfloor.$$

Then $p/(32 \log 24) \leq q \leq p/(16 \log 24)$ and $q \leq n/2$. Equations (41) and (42) yield

$$T_n \geq \frac{1}{2\sqrt{32 \log 24}} \sqrt{np} \geq \frac{1}{64} \sqrt{np}$$

with probability at least

$$\frac{1}{2} 24^{-q} \geq \frac{1}{2} e^{-p/16} \geq \frac{1}{8} e^{-p/8}.$$

The three ranges prove Lemma 5 with no asymptotic normal approximation.

A.7 SMALL SCALES, SATURATION, AND SIMULTANEITY

We fill in the cases abbreviated in the main proof. First, the memory indices are distinct with probability at least $15/16$ by (21). On this event,

$$\frac{1}{n} \sum_i b_{J_i}(S) \Sigma_i = c_{\text{mem}}. \quad (44)$$

Suppose $1 \leq p < 16$. If $\gamma \leq L/2$, then $c_{\text{mem}} = \gamma/4$ and

$$c_{\text{mem}} \geq \frac{1}{64} \min\{L, \gamma p\}.$$

If $\gamma > L/2$, then $c_{\text{mem}} = L/8$ and the same inequality is immediate. The probability that any ramp is active is at most the sum in (33), hence below 0.021. By the compact-interval part of Lemma 5, the sampling event can be chosen to have probability at least 0.45 while still giving a universal multiple of $L\sqrt{p/n}$. The three events are independent, and $0.979 \cdot (15/16) \cdot 0.45 > e^{-1}$. Their intersection therefore exceeds e^{-p} . This proves (5) for $p < 16$.

Now suppose $p \geq 16$ but $L/\gamma < 16$. Then $\gamma > L/16$ and

$$c_{\text{mem}} = \min\{\gamma/4, L/8\} \geq L/64. \quad (45)$$

Thus memory alone supplies a constant fraction of the saturated stability term. Intersecting (44) with Lemma 5 proves the result.

It remains to justify (22). Let $P = \min\{n/64, L/\gamma\}$. If $P \geq 16$, the largest geometric scale below P is greater than $P/4$. For a query $p \leq c_0 n \leq n/64$, the largest available scale not exceeding p , or the terminal scale if $p > P$, therefore obeys

$$p_k \geq \frac{1}{4} \min\{p, L/\gamma\}.$$

The construction of all groups uses only (n, L, γ) . It is completed before p is chosen. Thus the proof supplies the probability statement for all p using the same distribution and algorithm. This establishes the simultaneity claimed in Theorem 1.

Finally, on the intersection used in the proof,

$$G_S \geq \frac{1}{4} \frac{3\gamma p_k}{512} + c_{\text{mem}} + \frac{cL}{4} \sqrt{p/n} \quad (46)$$

$$\geq c' \left(\min\{L, \gamma p\} + L\sqrt{p/n} \right) \quad (47)$$

$$\geq c' \min\left\{L, \gamma p + L\sqrt{p/n}\right\}. \quad (48)$$

This also makes clear that every term is one-sided and no cancellation is used.

A.8 MOMENT CONSEQUENCE

Let

$$u_p = c_1 \min\left\{L, \gamma p + L\sqrt{p/n}\right\}.$$

Theorem 1 gives $\mathbb{P}(G_S \geq u_p) \geq e^{-p}$. Hence

$$\|G_S\|_p^p \geq u_p^p e^{-p}, \quad \|G_S\|_p \geq e^{-1} u_p.$$

This proves Corollary 2. The upper direction of (7) is (1) combined with $|G_S| \leq L$. The lower direction is the corollary.

B EXACT COMPUTATIONAL AUDITS

The construction contains only binomial probabilities and maximum inequalities. We nevertheless performed three audits.

Exact clean-event probabilities. For each $n \in \{4096, 8192, 16384, 32768, 65536\}$ and every geometric scale below $n/64$, we evaluated q_* , each lowered tail $q(r_k)$, the floor in d_k , and the exact product in (36) in log space. Table 1 shows representative results. The final column divides the rigorous clean-event lower bound by $e^{-p_k/2}$. The analytic proof uses the weaker constant 6.

Table 1: Exact audit of the multiscale event.

n	p_k	$-\log q_*$	$\log(q(r_k)/q_*)/p_k$	$\mathbb{P}(E_k)e^{p_k/2}$	lower bound
4096	16	132.81	0.0320		7.964
4096	64	132.81	0.0319		8.000
16384	256	521.60	0.0318		8.000
65536	1024	2074.71	0.0318		8.000

Exhaustive stability audit. For $n = d = 3$, we enumerated every Rademacher dataset, every one-point replacement, and every test input. We separately enumerated every memory dataset with three cells. With $\gamma = 0.2$, the largest ramp loss change was 0.1, the largest memory loss change was 0.1, and their sum was exactly 0.2. The maximum coordinate-amplitude change was exactly γ . This checks the sharp cases of (27) and (28).

The audit scripts use exact enumeration for stability and double-precision log binomial masses for the probability calculation. The displayed margins are several orders of magnitude larger than floating-point error.