Confidence and Budgets under Honest Data Insertions

Pahan Dewasurendra
Johns Hopkins University

Abstract

A monotone adversary observes an i.i.d. labeled sample and appends correctly labeled examples. Recent work determined the worst-budget expected error but left the prescribed-budget and high-confidence laws open. We prove two sharp advances. First, for VC dimension $d \geq 2$, n clean examples, and exact known budget m, the worst-class minimax expected error is at least

$$c \min\left\{1, \frac{d}{n}\left[1 + \log\left(1 + \frac{\min\{m,n\}}{d}\right)\right]\right\}.$$

Thus exactly $m = n$ honest insertions already realize the full $\Theta(1 \wedge (d/n)\log(e+n/d))$ worst-budget penalty. The previous all-learner construction used a much larger budget and gave no interpolation in m. The proof introduces reciprocal completion: after a rare disagreement is selected from the clean sample, the adversary draws the missing side from its conditional clean law. The two target orientations then induce exactly the same final multiset while every inserted label remains correct. Second, we determine the worst-budget PAC sample complexity. At $d = 1$ it is $\Theta(\varepsilon^{-1}\log(1/\delta))$, with no insertion penalty at any confidence. At every $d \geq 2$ it is

$$\Theta\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

The dimension-one upper bound follows from a new nested-disagreement argument that gives the exact tail $\mathbb{P}\{\mathrm{err} > \varepsilon\} \leq (1-\varepsilon)^n$. These results close the high-confidence question and locate a quantitative finite-budget obstruction.

1 Introduction

Training data are often curated after collection. Deduplication, augmentation, and targeted acquisition all use the realized sample to decide which examples remain visible. Even when every visible label is correct, this feedback destroys exchangeability.

Larsen et al. (2026) isolated this issue through *monotone adversarial corruptions*. A clean sample of n i.i.d. examples is labeled by a target concept. After seeing it, an adversary appends exactly m examples whose labels must agree with that same target. The learner receives a uniform shuffle of the $n+m$ examples and is evaluated on the clean distribution. The adversary adds information and never changes a label, yet the additions can make standard optimal PAC learners suboptimal.

For a class of VC dimension d, every hypothesis consistent with the augmented sample is also consistent with the hidden clean sample. Empirical risk minimization therefore retains the classical bound

$$O\left(1 \wedge \frac{d}{n}\log\left(e + \frac{n}{d}\right)\right). \tag{1}$$

Larsen et al. (2026) showed that one-inclusion, bagging, and recursive majority learners can all lose this logarithm under adaptive insertions. Mehrotra (2026) recently proved that the loss is information-theoretic. In the worst case over finite known budgets, the minimax expected error is $\Theta(1/n)$ at $d = 1$ and $\Theta(1 \wedge (d/n)\log(e+n/d))$ at $d \geq 2$.

Two statistical questions remained. The first is the dependence on a prescribed budget m. The construction of Mehrotra (2026) pads every domain point to a common multiplicity, which uses a budget far larger than n and gives no lower curve between bounded and unrestricted budgets. The known upper bounds are

$$R(n,m,d) \lesssim \min\left\{1, \frac{d(m+1)}{n}, \frac{d}{n}\log\left(e + \frac{n}{d}\right)\right\}. \tag{2}$$

The second is the optimal (ε, δ) sample complexity. Expected error does not generally determine a confidence law. This distinction is substantive even for clean one-inclusion learners. Aden-Ali et al. (2023) construct valid one-inclusion rules whose Markov tail is unimprovable.

Our results. We make progress on both questions. For every prescribed budget, we prove the first lower interpolation between the clean and unrestricted regimes:

$$R(n,m,d) \gtrsim \min\left\{1, \frac{d}{n}\left[1+\log\left(1+\frac{\min\{m,n\}}{d}\right)\right]\right\}. \tag{3}$$

At $m = n$, this matches (1). One inserted example per clean example is enough for the full logarithmic penalty. The complete intermediate upper curve remains open. Our lower bound rules out any budget-uniform clean rate and identifies the logarithmic budget scale that an optimal upper bound must confront.

Our second result closes the high-confidence question. For $d \geq 2$, the worst-budget sample complexity is

$$\Theta\left(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}\right). \tag{4}$$

The upper bound is attained by ERM. The lower bound upgrades projective ambiguity from mean loss to constant-probability loss and combines it with the classical confidence obstruction. At $d = 1$, the answer instead remains the clean rate $\Theta(\varepsilon^{-1}\log(1/\delta))$. We prove the stronger tail inequality $(1-\varepsilon)^n$ for the consensus learner of Mehrotra (2026).

Reciprocal completion. The budget improvement comes from replacing deterministic padding by distributional symmetrization. In a projective-plane block, a row target and a column target disagree at one checker. Once that checker is absent from the clean sample, draw an independent sample from the paired column law conditioned on selecting the same missing checker, and append it to the row sample. Under the column orientation, append the corresponding conditional row sample. Both final multisets have the product law $Q_{\mathrm{row}}Q_{\mathrm{col}}$. The completion is honest because both conditional samples omit the sole disagreement. Diluting the projective blocks with a universally labeled point makes the informative clean count comparable to any prescribed budget.

Related work. The monotone model and its algorithm-specific separations were introduced by Larsen et al. (2026). The worst-budget minimax expectation and the exceptional dimension-one consensus rule are due to Mehrotra (2026). Our work does not alter those expected-rate upper bounds. It localizes the lower bound in the insertion budget, reduces the full-penalty budget to n, determines the confidence law, and gives a tail analysis of the consensus rule.

Optimal clean PAC learning has sample complexity $\Theta((d+\log(1/\delta))/\varepsilon)$ (Hanneke, 2016). Bagging and majority-of-three provide other optimal clean learners (Larsen, 2023; Aden-Ali et al., 2024). High-probability behavior cannot in general be inferred from a leave-one-out expectation (Aden-Ali et al., 2023). Clean-label poisoning also appends correctly labeled examples, but it asks whether a fixed test instance can be made attackable. Its linear budget dependence concerns that stronger instance-targeted objective rather than ordinary population risk on the original distribution (Blum et al., 2021).

In p-tampering, randomly selected examples are replaced by correctly labeled in-support examples, leaving an i.i.d. untouched fraction that preserves realizable PAC learnability (Mahloujifar et al., 2018). Strong budget tampering can instead remove selected rare observations and destroy PAC learnability. Honest insertion differs from both: it retains every clean observation, but the adaptive additions still change the optimal rate. Classical VC generalization and lower bounds are due to Blumer et al. (1989) and Ehrenfeucht et al. (1989).

Table 1 separates the known results from the two contributions here. The prescribed-budget row is a lower bound rather than a complete rate. All other displayed rows are sharp up to universal constants.

2 Model and Main Results

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be nonempty. A target $h^\star \in \mathcal{H}$ and a distribution D generate a clean labeled sequence $S \sim D^n_{h^\star}$. An exact-budget monotone adversary sees S and appends m labeled examples $(x, h^\star(x))$. It may depend on the ordered clean sequence and use private randomness. The learner knows $\mathcal{H}, n, m$, receives a uniform shuffle T of the combined multiset, and outputs an arbitrary binary predictor $\widehat{h}$. Its conditional population error is

$$\mathrm{err}_D(\widehat{h}, h^\star) = \mathbb{P}_{X\sim D}\{\widehat{h}(X) \neq h^\star(X)\}.$$

Randomized adversaries can be replaced by deterministic ones in every lower bound below. A random seed fixes a deterministic map on the finite clean-sample space, and averaging fixes one seed for each learner.

Let $R(n,m,d)$ be the supremum over finite binary classes of VC dimension exactly d of their minimax expected error. The learner may be chosen for the class and budget. We use $a \wedge b = \min\{a,b\}$.

Theorem 1 (Prescribed-budget lower bound)**.** *There is a universal $c > 0$ such that, for all $n \geq 1$, $m \geq 0$, and $d \geq 2$,*

$$R(n,m,d) \geq c\left[1 \wedge \frac{d}{n}\left(1+\log\left(1+\frac{m\wedge n}{d}\right)\right)\right].$$

The hard classes are finite, and the adversary is deterministic after the learner is fixed.

Table 1: Worst-class rates for n clean examples. The notation $R(n, m, d)$ uses an exact known insertion budget, while N takes the worst case over all finite exact budgets.

Quantity	$d = 1$	$d \geq 2$
Clean expected risk	$\Theta(1/n)$	$\Theta(1 \wedge d/n)$
Worst-budget expected risk (Mehrotra, 2026)	$\Theta(1/n)$	$\Theta(1 \wedge (d/n)\log(e + n/d))$
Prescribed-budget lower bound, this work	$\Omega(1/n)$	$\Omega\big(1 \wedge \frac{d}{n}[1 + \log(1 + (m \wedge n)/d)]\big)$
Worst-budget PAC complexity, this work	$\Theta(\log(1/\delta)/\varepsilon)$	$\Theta((d\log(1/\varepsilon) + \log(1/\delta))/\varepsilon)$

The clean lower bound supplies the first term. The new logarithm is nontrivial when m exceeds d. Combining Theorem 1 with the ERM upper bound yields an exact linear-budget law.

Corollary 2 (One-for-one insertions suffice)**.** *For every $d \geq 2$,*

$$R(n, n, d) \asymp 1 \wedge \frac{d}{n}\log\Big(e + \frac{n}{d}\Big).$$

The same conclusion holds for every $m \geq n$.

For confidence, define $N(d, \varepsilon, \delta)$ as the smallest n for which every finite class of VC dimension at most d has, for every exact budget m, a learner satisfying

$$\sup_{h^\star, D, A} \mathbb{P}\{\mathrm{err}_D(\widehat{h}, h^\star) > \varepsilon\} \leq \delta.$$

The supremum includes adaptive monotone adversaries. Standard measurability conditions extend the upper bounds to infinite classes.

Theorem 3 (High-confidence phase diagram)**.** *There are universal constants $c, C > 0$ such that, for $0 < \varepsilon, \delta \leq 1/8$,*

$$c\frac{\log(1/\delta)}{\varepsilon} \leq N(1, \varepsilon, \delta) \leq C\frac{\log(1/\delta)}{\varepsilon}.$$

For every $d \geq 2$,

$$N(d, \varepsilon, \delta) \geq c\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon},$$
$$N(d, \varepsilon, \delta) \leq C\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}.$$

The $d\log(1/\varepsilon)$ lower term is witnessed with exact budget $m = n$.

Dimension zero has sample complexity zero. The discontinuity between dimensions one and two occurs at every fixed confidence and persists through the entire confidence range.

3 Reciprocal Completion

We isolate the symmetry used in Theorem 1. A flag f has two orientations, called row and column. Conditional on t informative observations, let E_f^{r} be the event that a symmetric selection rule chooses f under the row orientation. Define E_f^{c} analogously.

Lemma 4 (Reciprocal completion)**.** *Suppose for every flag and count,*

$$\mathbb{P}(E_f^{\mathrm{r}} \mid t) = \mathbb{P}(E_f^{\mathrm{c}} \mid t) = \pi_{f,t}.$$

Let $Q_{f,t}^{\mathrm{r}}$ and $Q_{f,t}^{\mathrm{c}}$ be the conditional clean-sample laws given these events. Assume the two targets agree on the union of the supports of both conditional laws. Then an honest adversary using t insertions can give the two orientations identical final-multiset laws on the selection event.

Proof. Under the row orientation, retain $A \sim Q_{f,t}^{\mathrm{r}}$ and append an independent $B \sim Q_{f,t}^{\mathrm{c}}$. Under the column orientation, retain $B \sim Q_{f,t}^{\mathrm{c}}$ and append an independent $A \sim Q_{f,t}^{\mathrm{r}}$. For every f, A, B, both joint masses equal

$$\pi_{f,t}Q_{f,t}^{\mathrm{r}}(A)Q_{f,t}^{\mathrm{c}}(B).$$

The final object is the same unordered union $A \uplus B$. Every appended label is correct because the paired targets agree on the conditional supports. □

The equality is pointwise, not a total-variation approximation. One can equivalently draw (A, B) first and then draw a fair orientation bit. The learner's observation is independent of that bit.

4 A Budgeted Projective Adversary

We instantiate Lemma 4 using a finite projective plane. A plane of prime order q has $K = q^2 + q + 1$ points and lines, with $k = q + 1$ points per line and lines per point (Hirschfeld, 1998). For every flag (r, M) introduce a checker c_{rM} and a selector s_{rM}. The class has row concepts u_p and column concepts v_L with

$$u_p(c_{rM}) = \mathbf{1}\{p \in M,\ p \neq r\},$$
$$v_L(c_{rM}) = \mathbf{1}\{L = M\}.$$

On selectors, set $u_p(s_{rM}) = \mathbf{1}\{p = r\}$ and $v_L(s_{rM}) = \mathbf{1}\{L = M\}$. The selectors have zero clean probability but ensure the exact dimension.

Lemma 5 (Projective block dimension Mehrotra, 2026)**.** *The checker-selector class has VC dimension two.*

Proof. Fix p and distinct lines L, M through p. The pair s_{pL}, s_{pM} has traces $11, 10, 01, 00$ under u_p, v_L, v_M, u_r for any $r \neq p$. For the upper bound, associate each domain point x with its line M_x and let $S_x = \{p : u_p(x) = 1\}$. Each S_x is either a singleton or M_x with one point removed. If three associated lines are distinct, the column concepts supply the zero and singleton traces. The row traces 110 and 111 would then put two points in $S_{x_1} \cap S_{x_2}$, contradicting the unique intersection of two lines. If exactly two lines agree, shattering requires both directed differences of the corresponding two S_x sets to have size at least two, which is impossible. If all lines agree, any selector is positive under only one row concept, and three distinct checkers have row traces only in $\{000, 011, 101, 110, 111\}$. No triple is shattered. □

Under row target u_p, the clean block distribution is uniform on $\{c_{pM} : M \ni p\}$, where every label is zero. Under column target v_L, it is uniform on $\{c_{rL} : r \in L\}$, where every label is one.

For a flag $f = (p, L)$, the paired concepts disagree only at c_{pL} on the union of these supports. Independently rank the k support checkers and select the first missing checker. Conditional on t draws, symmetry gives

$$\mathbb{P}\{f \text{ selected} \mid u_p, t\} = \mathbb{P}\{f \text{ selected} \mid v_L, t\} = \frac{\rho_t}{k}, \quad (5)$$

where ρ_t is the coupon-missing probability. Both conditional laws omit c_{pL}, so reciprocal completion applies.

Take $r = \lfloor d/2 \rfloor$ disjoint blocks and product the concepts. Zero-mass free coordinates give exact dimension d when it is odd. Add one common point on which every concept is zero. Give all projective blocks total probability θ, split uniformly over blocks, and give the common point mass $1 - \theta$.

Let T be the number of informative clean observations and T_i the count in block i. If $T \leq m$, reciprocally complete each block that has a missing checker using T_i insertions. Do not complete blocks with no missing checker. Fill every unused insertion slot with the common point. This uses exactly m honest insertions. A fallback handles $T > m$.

Each selected checker has test mass $\theta/(rk)$. Conditional on all latent variables except a successfully completed orientation, the final sample is independent of that fair orientation bit. Hence every successful block contributes $\theta/(2rk)$ to expected error.

Lemma 6 (Many missing checkers)**.** *For one block, let Z be the number of unseen support checkers. Then*

$$\mathbb{P}(Z > 0) \geq \frac{\mu}{1+\mu}, \qquad \mu = k\left(1 - \frac{\theta}{rk}\right)^n \geq k \exp\left(-\frac{2n\theta}{rk}\right).$$

Proof. The first two moments satisfy $\mathbb{E}Z = \mu$ and $\mathbb{E}Z^2 \leq \mu + \mu^2$, since the probability that two fixed cells are both unseen is at most the product of their individual missing probabilities. Paley–Zygmund gives the first inequality. The last inequality uses $\log(1 - x) \geq -2x$ for $x \leq 1/2$. □

We now prove Theorem 1. Let J count successfully completed blocks on the branch $T \leq m$. Averaging over the independent fair orientation priors and then taking a worst target gives

$$\mathbb{E}\,\mathrm{err}_D(\widehat{h}, h^\star) \geq \frac{\theta}{2rk}\mathbb{E}J. \quad (6)$$

For each block, success is the event that some checker is missing, except that all blocks are discarded when $T > m$. Therefore

$$\mathbb{E}J \geq r\left(\frac{\mu}{1+\mu} - \mathbb{P}(T > m)\right). \quad (7)$$

First suppose $m < n$. Choose $\theta = m/(8n)$, so $T \sim \mathrm{Bin}(n, \theta)$ has mean $m/8$. A multiplicative Chernoff bound gives $\mathbb{P}(T > m) \leq e^{-c_0 m}$ for a universal $c_0 > 1$. Set $s = m/r$ and $L = \log(e + s)$. For s above a universal constant, Bertrand's postulate supplies a prime-plane support size satisfying

$$\frac{8s}{L} < k \leq \frac{20s}{L}. \quad (8)$$

Lemma 6 and $n\theta = m/8$ now give

$$\mu \geq k \exp\left(-\frac{s}{4k}\right) \geq \frac{8s}{L}(e+s)^{-1/32}.$$

Increasing the fixed threshold for s makes this at least four and also makes the overflow term at most $1/10$. Equations (6)–(8) yield

$$\mathbb{E}\,\mathrm{err}_D(\widehat{h}, h^\star) \geq c\frac{r}{n}\log(e + m/r).$$

If $m \geq n$, choose $\theta = 1$ and set $s = n/r$. Then $T = n \leq m$ deterministically. Use the same window (8). Since $2s/k \leq L/4$,

$$\mu \geq \frac{8s}{L}(e+s)^{-1/4},$$

which is at least four for large fixed s. The same substitution gives $c(r/n)\log(e + n/r)$. When $(m \wedge n)/r$

is bounded, the ordinary clean lower bound $c(1 \wedge d/n)$ supplies the theorem. It also handles $d > n/4$. Finally, $r = \lfloor d/2 \rfloor$ is comparable to d. Zero-probability free coordinates give exact dimension when d is odd.

The construction above is randomized only to expose its posterior symmetry. For any fixed learner on the finite instance, its risk averaged over adversary seeds obeys the same lower bound. Some seed therefore defines a deterministic exact-budget adversary with at least that risk. This completes the proof.

5 The Confidence Law

We first consider $d \geq 2$. An ERM on the augmented sample is consistent with the hidden clean sample. The realizable VC tail therefore gives

$$\mathbb{P}\{\mathrm{err}_D(\widehat{h}, h^\star) > \varepsilon\} \leq C_1 \left(\frac{C_2}{\varepsilon}\right)^d e^{-c_1 n\varepsilon}.$$

This proves the upper bound in Theorem 3.

For the $d \log(1/\varepsilon)$ lower term, use $r = \lfloor d/2 \rfloor$ projective blocks, exact budget $m = n$, and a prime-plane support size $k \asymp 1/\varepsilon$. If

$$n \leq c\frac{d}{\varepsilon} \log \frac{1}{\varepsilon}$$

for a sufficiently small universal c, Lemma 6 gives

$$\mu \geq k \exp\left(-\frac{2n}{rk}\right) \geq k\varepsilon^{1/4}.$$

After universal endpoint adjustments, $\mu \geq 16$. If U is the number of unsuccessful blocks, then $\mathbb{E}U \leq r/(1+\mu) \leq r/16$. Markov's inequality shows that $J = r - U \geq r/2$ with probability at least $7/8$.

Conditional on the final sample, the selected flags, and all paired samples, the J successful orientation bits remain mutually independent and fair by Lemma 4. For any fixed prediction vector, the number of selected checkers mislabeled by the learner is therefore $\mathrm{Bin}(J, 1/2)$. With a universal positive probability it is at least $J/4$. The same statement holds for randomized learners after conditioning on their coins. On the intersection of these events,

$$\mathrm{err}_D(\widehat{h}, h^\star) \geq \frac{J}{4rk} \geq \frac{1}{8k} \geq c'\varepsilon.$$

Rescaling ε by the universal c' proves the constant-confidence lower term.

For the $\log(1/\delta)$ term, take two concepts that disagree on a point of mass 2ε and agree on a common point carrying the remaining mass. Under a fair target prior, absence of the rare point makes the two labeled samples identical, so one target has failure probability at least

$$\frac{1}{2}(1-2\varepsilon)^n.$$

Appending common points gives any desired exact budget. Adding zero-mass shattered points gives exact VC dimension d. Hence $n \gtrsim \varepsilon^{-1} \log(1/\delta)$ is necessary. The maximum of the constant-confidence and rare-point lower terms is at least half their sum, proving the lower bound in Theorem 3.

6 Why Dimension One Is Different

The consensus rule of Mehrotra (2026) admits a stronger analysis than its leave-one-out expectation. Fix $f \in \mathcal{H}$ and transform $\mathcal{H}$ to $\mathcal{C} = \{h \triangle f : h \in \mathcal{H}\}$, where $\triangle$ is pointwise XOR. Then $0 \in \mathcal{C}$. Given an augmented sample T, let $\mathcal{V}(T)$ be its transformed version space and predict

$$\gamma_T(x) = \mathbf{1}\{c(x) = 1 \text{ for every } c \in \mathcal{V}(T)\},$$
$$\widehat{h}_T = f \triangle \gamma_T.$$

Fix the transformed target $c^\star$. For $c \in \mathcal{C}$, define its one-sided disagreement range

$$A_c = \{x : c^\star(x) = 1,\ c(x) = 0\}.$$

Lemma 7 (Nested disagreements)**.** *If* $\mathrm{VCdim}(\mathcal{H}) \leq 1$, *then* $\{A_c : c \in \mathcal{C}\}$ *is linearly ordered by inclusion.*

Proof. If A_c and A_b were incomparable, take $x \in A_c \setminus A_b$ and $y \in A_b \setminus A_c$. On $\{x, y\}$, the concepts $0, c^\star, c, b$ realize $00, 11, 01, 10$. This shatters the pair. □

Proposition 8 (Exact exponential tail)**.** *For every class of VC dimension at most one, every target and distribution, every budget, and every adaptive monotone adversary,*

$$\mathbb{P}\{\mathrm{err}_D(\widehat{h}_T, h^\star) > \varepsilon\} \leq (1-\varepsilon)^n.$$

Proof. The consensus rule never errs where $c^\star = 0$. Its error set is

$$E_T = \bigcup_{c \in \mathcal{V}(T)} A_c. \tag{9}$$

Honest insertion only shrinks the clean version space, so E_T is contained in the same union formed from concepts consistent with the clean sample S. Every range in that larger union misses S. If the finite chain contains a range of mass greater than ε, let A_ε be its smallest such range. Some range of mass greater than ε misses S exactly when A_ε misses S. This has probability $(1 - D(A_\varepsilon))^n < (1-\varepsilon)^n$. If no such range exists, the failure event is empty. □

A two-concept rare-point class matches the exponent up to universal constants. Proposition 8 proves the dimension-one part of Theorem 3. It also shows why the negative high-confidence result for arbitrary valid one-inclusion orientations (Aden-Ali et al., 2023) does not apply: the fixed-reference consensus rule has nested one-sided error ranges, a property absent from a generic orientation.

7 Discussion

Reciprocal completion converts conditional distributional symmetry into an exact honest insertion attack. It reduces the projective lower-bound budget to the number of informative clean observations and, after dilution, yields a logarithmic interpolation in the prescribed budget. Whether the lower curve in Theorem 1 is the exact finite-budget rate remains open. The current upper envelope (2) leaves a gap between logarithmic and linear dependence on m/d.

The confidence law has no such gap. At dimension one, insertion cannot enlarge the nested clean error envelope. At dimension two, reciprocal projective ambiguity already restores the ERM logarithm at constant confidence. The same transition persists for every smaller failure probability through the independent classical confidence obstruction.

Acknowledgments

References

Aden-Ali, I., Cherapanamjeri, Y., Shetty, A., and Zhivotovskiy, N. (2023). The one-inclusion graph algorithm is not always optimal. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 72–88. PMLR.

Aden-Ali, I., Høgsgaard, M. M., Larsen, K. G., and Zhivotovskiy, N. (2024). Majority-of-three: The simplest optimal learner? In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 22–45. PMLR.

Blum, A., Hanneke, S., Qian, J., and Shao, H. (2021). Robust learning under clean-label attack. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 591–634. PMLR.

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

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

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

Hirschfeld, J. W. P. (1998). *Projective Geometries over Finite Fields*. Oxford University Press, 2 edition.

Larsen, K. G. (2023). Bagging is an optimal PAC learner. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 450–468. PMLR.

Larsen, K. G., Pabbaraju, C., and Shetty, A. (2026). Learning with monotone adversarial corruptions. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–18. PMLR.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. (2018). Learning under p-tampering attacks. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 572–596. PMLR.

Mehrotra, A. (2026). Optimal rates for learning with monotone adversaries. *arXiv preprint arXiv:2608.06337*.

A Formal Minimax Definitions

For a finite class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, target $h^\star$, and distribution D, write $D_{h^\star}$ for the law of $(X, h^\star(X))$ when $X \sim D$. An exact-budget randomized monotone adversary is a kernel from $(\mathcal{X} \times \{0,1\})^n$ to length-m labeled sequences. On every clean sequence labeled by $h^\star$, its output labels must also agree with $h^\star$. The final sequence is uniformly shuffled. For a learner L, define

$$\mathfrak{r}(L; \mathcal{H}, n, m) = \sup_{h^\star, D, A} \mathbb{E}\, \mathrm{err}_D(L(T), h^\star).$$

The supremum includes randomized adaptive adversaries. Let

$$R_{\mathcal{H}}(n,m) = \inf_L \mathfrak{r}(L; \mathcal{H}, n, m), \qquad R(n,m,d) = \sup_{\mathcal{H}:\, \mathrm{VCdim}(\mathcal{H})=d} R_{\mathcal{H}}(n,m).$$

All lower-bound classes are finite. The class and learner are known before the adversary is chosen.

The PAC quantity $N(d, \varepsilon, \delta)$ is the least n such that every class $\mathcal{H}$ with $\mathrm{VCdim}(\mathcal{H}) \le d$ has, for each m, a learner satisfying

$$\sup_{h^\star, D, A} \mathbb{P}\{\mathrm{err}_D(L(T), h^\star) > \varepsilon\} \le \delta.$$

Allowing a different learner for each known m agrees with the convention of Mehrotra (2026).

B Reciprocal Completion in Full Detail

We state a finite version sufficient for the lower bound. Let $\mathcal{F}$ be a finite flag set. For each $f \in \mathcal{F}$, let h_f^0, h_f^1 be paired targets. Conditional on an informative count t, orientation $b \in \{0,1\}$ has a clean law $P_{f,t}^b$ on labeled multisets. A selection kernel applied to the clean multiset outputs either a flag or $\perp$. Suppose the joint law of the selected flag and clean multiset can be written

$$\mathbb{P}\{F = f, S = s \mid b, t\} = \pi_{f,t} Q_{f,t}^b(s), \tag{10}$$

where $\pi_{f,t}$ does not depend on b. Assume every multiset in the support of $Q_{f,t}^0$ or $Q_{f,t}^1$ contains only domain points on which h_f^0 and h_f^1 agree.

Lemma 9 (Exact reciprocal law)**.** *On the event $F = f$, orientation zero can append $S^1 \sim Q_{f,t}^1$ and orientation one can append $S^0 \sim Q_{f,t}^0$. For every final labeled multiset w,*

$$\mathbb{P}\{S^0 \uplus S^1 = w, F = f \mid b = 0, t\} = \mathbb{P}\{S^0 \uplus S^1 = w, F = f \mid b = 1, t\}.$$

Both completions are monotone.

Proof. For orientation zero, expand the left side as

$$\pi_{f,t} \sum_{s^0 \uplus s^1 = w} Q_{f,t}^0(s^0) Q_{f,t}^1(s^1).$$

Orientation one gives the same expression with the two factors written in reverse order. The agreement-support assumption makes every appended label equal to the active target label. □

The lemma remains true after an independent uniform shuffle. It also tensorizes across blocks conditional on their count vector. A useful generative description is to draw f, S^0, S^1 first and draw the orientation bit last. The final multiset is then independent of that bit.

Fixing the adversary's coins. The lower bound first averages over rankings and conditional completion samples. Since the class, domain, and clean-sample space are finite, one random seed specifies a deterministic output for every possible clean input. For any fixed randomized learner, the final expected loss or failure probability is an average over these seeds, the target prior, and the clean sample. There is a seed and target with loss at least the average. This gives a deterministic adversary without changing the learner's coins.

C The Projective Block

Let $(\mathcal{P}, \mathcal{L})$ be a projective plane of prime order q. Put $K = q^2 + q + 1$ and $k = q + 1$. Every point lies on k lines, every line contains k points, and each point-line pair has a unique incidence relation (Hirschfeld, 1998). For every flag (r, M) introduce a checker c_{rM} and selector s_{rM}. The class

$$\mathcal{H}_q = \{u_p : p \in \mathcal{P}\} \dot{\cup} \{v_L : L \in \mathcal{L}\}$$

has labels

$$\begin{aligned} u_p(c_{rM}) &= \mathbf{1}\{p \in M,\ p \neq r\}, & v_L(c_{rM}) &= \mathbf{1}\{L = M\}, \\ u_p(s_{rM}) &= \mathbf{1}\{p = r\}, & v_L(s_{rM}) &= \mathbf{1}\{L = M\}. \end{aligned}$$

The selector formula implies that exactly u_r and v_M label s_{rM} by one.

Lemma 10 (Projective block dimension)**.** $\mathrm{VCdim}(\mathcal{H}_q) = 2$.

Proof. This is Lemma 3.7 of Mehrotra (2026). We include the incidence argument. Associate each domain point x with its line M_x, and write $S_x = \{p : u_p(x) = 1\}$. Thus $S_{s_{rM}} = \{r\}$ and $S_{c_{rM}} = M \setminus \{r\}$. If three points have distinct associated lines, the column concepts realize the zero trace and the three singleton traces. Shattering would then require row traces 110 and 111, so $S_{x_1} \cap S_{x_2}$ would contain two distinct points. This is impossible because two projective lines meet once. If exactly two associated lines agree, say $M_{x_1} = M_{x_2} \neq M_{x_3}$, the column concepts realize $000, 110, 001$. Shattering would require both directed differences $S_{x_1} \setminus S_{x_2}$ and $S_{x_2} \setminus S_{x_1}$ to contain at least two points, which is impossible for sets that are either singletons or a common line with one point removed. If all three associated lines agree, a selector is positive under only one row concept, while the common column concept contributes only 111. If all three are distinct checkers, the row traces are contained in $\{000, 011, 101, 110, 111\}$. Hence no triple is shattered. For the lower bound, fix a point p and distinct lines L, M through it. The pair s_{pL}, s_{pM} has traces $11, 10, 01, 00$ under u_p, v_L, v_M, u_r, respectively, for any $r \neq p$. Therefore $\mathrm{VCdim}(\mathcal{H}_q) = 2$. □

For $p \in \mathcal{P}$, let D_p^{r} be uniform on the row $\{c_{pM} : M \ni p\}$. For $L \in \mathcal{L}$, let D_L^{c} be uniform on the column $\{c_{rL} : r \in L\}$. The target is u_p under D_p^{r} and v_L under D_L^{c}. Every row label is zero and every column label is one. For $f = (p, L)$, these targets disagree only at c_{pL} on the union of the two supports.

Draw an independent uniform ordering of a support's k checkers. If some checker is absent, select the first absent one. Conditional on t draws, let ρ_t be the probability that some checker is absent. Permutation symmetry gives, for each incident $f = (p, L)$,

$$\mathbb{P}\{F = f \mid u_p, t\} = \mathbb{P}\{F = f \mid v_L, t\} = \frac{\rho_t}{k}.$$

The conditional row and column samples both omit c_{pL}. They therefore satisfy Lemma 9.

D Proof of the Prescribed-Budget Bound

We prove the nontrivial range first. Assume $2 \leq d \leq n/4$ and set $r = \lfloor d/2 \rfloor$. Take r disjoint tagged copies of $\mathcal{H}_q$ and their Cartesian product. If d is odd, add one zero-probability point whose label is an independent coordinate. Lemma 10 and additivity across disjoint product domains give exact VC dimension d.

Add a common point a labeled zero by every concept. Draw an independent fair row-column orientation and the corresponding row point or column line in each block. The clean distribution assigns total mass θ to the projective blocks, uniformly over blocks and then over the active support's k checkers. It assigns mass $1 - \theta$ to a.

Let T_i be the informative count in block i and $T = \sum_i T_i$. When $T \leq m$, apply reciprocal completion in every block with a selected missing checker. A successful block uses exactly T_i insertions. Leave unsuccessful blocks uncompleted. Append copies of a until exactly m examples have been inserted. When $T > m$, append m copies of a. This defines an honest exact-budget adversary on every clean input.

Let J be the number of successful blocks on the branch $T \leq m$. Conditional on the final sample and all nonorientation latent variables, each successful orientation is a fair bit. At its selected checker the paired labels

are opposite and the test mass is $\theta/(rk)$. Therefore

$$\mathbb{E}\,\mathrm{err}_D(\widehat{h}, h^\star) \geq \frac{\theta}{2rk}\mathbb{E}J. \tag{11}$$

For a fixed block, let Z count unseen support checkers. Each checker has unconditional probability $p = \theta/(rk)$. Thus

$$\mathbb{E}Z = k(1-p)^n =: \mu,$$

and

$$\mathbb{E}Z^2 = \mu + k(k-1)(1-2p)^n \leq \mu + \mu^2.$$

Paley–Zygmund yields $\mathbb{P}(Z > 0) \geq \mu/(1+\mu)$. Since $p \leq 1/2$,

$$\mu \geq k\exp\left(-\frac{2n\theta}{rk}\right). \tag{12}$$

D.1 The range $m < n$

Set $\theta = m/(8n)$ and $s = m/r$. The informative count has law $T \sim \mathrm{Bin}(n, \theta)$ and mean $m/8$. A multiplicative Chernoff bound gives

$$\mathbb{P}(T > m) \leq \exp(-c_0 m) \tag{13}$$

for a universal $c_0 > 1$.

Suppose $s \geq s_0$ for a sufficiently large universal s_0. Write $L = \log(e+s)$ and $B = \lceil 8s/L \rceil$. Bertrand's postulate supplies a prime q with $B < q < 2B$. Hence the projective support size $k = q + 1$ satisfies

$$\frac{8s}{L} < k \leq \frac{20s}{L}. \tag{14}$$

Increasing s_0 absorbs the ceiling. Equations (12) and (14) give

$$\mu \geq k\exp\left(-\frac{s}{4k}\right) \geq \frac{8s}{L}(e+s)^{-1/32}.$$

Thus every block has a missing checker with probability at least $4/5$. Since J can lose at most r on the overflow event,

$$\mathbb{E}J \geq r\left(\frac{4}{5} - \mathbb{P}(T > m)\right) \geq \frac{7r}{10}$$

after one further increase of s_0. Substituting into (11) and using the upper bound on k gives

$$\mathbb{E}\,\mathrm{err}_D(\widehat{h}, h^\star) \geq c\frac{rL}{n}.$$

D.2 The range $m \geq n$

Set $\theta = 1$, so $T = n \leq m$ deterministically, and let $s = n/r$. Choose k by (14). Now

$$\mu \geq k\exp(-2s/k) \geq \frac{8s}{L}(e+s)^{-1/4}.$$

This exceeds four for $s \geq s_0$. Equations (11) and (14) again give

$$\mathbb{E}\,\mathrm{err}_D(\widehat{h}, h^\star) \geq c\frac{r}{n}\log(e + n/r).$$

D.3 Completing all regimes

If $(m \wedge n)/r < s_0$, then $\log(1 + (m \wedge n)/d)$ is universally bounded. The standard clean lower bound is $c[1 \wedge d/n]$ (Ehrenfeucht et al., 1989). It remains valid at exact budget m by adding m copies of the common point. If $d > n/4$, the same lower bound is a positive constant. Finally, $r \geq d/3$ for $d \geq 2$, apart from harmless adjustment at $d = 2$, and the logarithms with denominators r and d are comparable. These observations prove the prescribed-budget theorem in the main paper.

E Constant-Confidence Projective Lower Bound

We give the probability upgrade used for the $d\log(1/\varepsilon)$ term. Take exact budget $m = n$, set $\theta = 1$, and choose a projective support size

$$\frac{a}{\varepsilon} \le k \le \frac{2a}{\varepsilon}$$

for a sufficiently small fixed $a > 0$. Bertrand's postulate again supplies such a prime order after universal endpoint adjustments.

Suppose

$$n \le c\frac{d}{\varepsilon}\log\frac{1}{\varepsilon}$$

for a sufficiently small universal c. With $r = \lfloor d/2 \rfloor$, equation (12) gives

$$\mu \ge k\exp\left(-\frac{2n}{rk}\right) \ge k\varepsilon^{1/4}.$$

For small enough ε, this is larger than a fixed large constant. Endpoint cases are absorbed into the constants. Hence the expected number of unsuccessful blocks is at most $r/16$. Markov's inequality shows that at least $r/2$ blocks are successful with probability at least $7/8$.

Conditional on the final sample and on the successful latent flags and paired samples, their orientation bits remain independent and fair. For every deterministic prediction vector, the number of selected checkers mislabeled by the learner is therefore binomial with parameters J and $1/2$. With universal positive probability it is at least $J/4$. The same conclusion holds for randomized learners after conditioning on their coins. On this event,

$$\mathrm{err}_D(\widehat{h}, h^\star) \ge \frac{J}{4rk} \ge \frac{1}{8k} \ge c'\varepsilon.$$

Rescaling ε by c' proves the constant-confidence lower bound used in the high-confidence theorem.

F High-Confidence Upper and Lower Bounds

F.1 Dimension at least two

Every ERM on the augmented sample is consistent with the clean sample. A standard realizable VC bound states that, for universal constants $C_i, c_i > 0$,

$$\mathbb{P}\{\mathrm{err}_D(\widehat{h}, h^\star) > \varepsilon\} \le C_1(C_2/\varepsilon)^d e^{-c_1 n\varepsilon}.$$

This follows from the double-sampling argument and Sauer's lemma (Blumer et al., 1989). Taking

$$n \ge \frac{C}{\varepsilon}\left[d\log(1/\varepsilon) + \log(1/\delta)\right]$$

makes the right side at most δ, uniformly in the adversary and budget.

The preceding section proves the $d\log(1/\varepsilon)/\varepsilon$ lower term at a fixed failure probability. For the confidence term, use two concepts that agree at a common point and disagree at a point of mass 2ε. If the rare point is absent from the clean sample, the observed clean labels are identical under the two targets. Under a fair target prior, every learner is wrong at the rare point with conditional probability at least $1/2$. Therefore one target has failure probability at least

$$\frac{1}{2}(1-2\varepsilon)^n.$$

Appending common points makes this a monotone exact-budget instance. Adding $d-1$ zero-mass points and all their labelings gives exact VC dimension d without changing the experiment. Thus $n \ge c\varepsilon^{-1}\log(1/\delta)$ is necessary. The maximum of this term and the constant-confidence projective term is at least half their sum.

F.2 Dimension one

We prove the exact consensus tail with finite classes. The standard separability argument extends it to measurable infinite classes. Fix $f \in \mathcal{H}$ and let $\mathcal{C} = \{h \triangle f : h \in \mathcal{H}\}$. This preserves VC dimension and places the zero concept in $\mathcal{C}$. Fix transformed target $c^\star$ and define $A_c = \{x : c^\star(x) = 1, c(x) = 0\}$.

Lemma 11 (Chain property)**.** *The family $\{A_c : c \in \mathcal{C}\}$ is a chain under inclusion.*

Proof. If A_c, A_b are incomparable, choose $x \in A_c \setminus A_b$ and $y \in A_b \setminus A_c$. Since both points lie in the support of $c^\star$, the four concepts $0, c^\star, c, b$ have traces $00, 11, 01, 10$. This contradicts VC dimension one. □

For an augmented sample T, the transformed consensus rule is one only when every $c \in \mathcal{V}(T)$ is one. Since $c^\star \in \mathcal{V}(T)$, it is correct whenever $c^\star(x) = 0$. Its error set when $c^\star(x) = 1$ is

$$E_T = \bigcup_{c \in \mathcal{V}(T)} A_c.$$

Let S be the hidden clean sample. Monotonicity gives $\mathcal{V}(T) \subseteq \mathcal{V}(S)$, so $E_T \subseteq E_S$. Every A_c with $c \in \mathcal{V}(S)$ misses S.

For a finite chain, let A_ε be its smallest member of mass greater than ε, if one exists. The event that any range of mass greater than ε misses S is exactly the event that A_ε misses S. Its probability is at most $(1-\varepsilon)^n$. Consequently

$$\mathbb{P}\{\mathrm{err}_D(\widehat{h}_T, h^\star) > \varepsilon\} \le (1-\varepsilon)^n.$$

The same two-concept rare-point construction used above has VC dimension one and supplies the matching lower exponent.

G Computational Verification

The accompanying source archive contains three independent checks.

1. `verify_reciprocal_completion.py` builds the Fano plane, verifies exact VC dimension two for the checker-selector class, enumerates every informative count multiset for $t = 0, \ldots, 6$, and compares the full rational distribution of final labeled multisets under both reciprocal orientations. Every atom agrees exactly.
2. `verify_vc1_nesting.py` enumerates all transformed binary classes containing zero on domains of size at most four. For all 512 classes of VC dimension at most one and all 1,980 target choices, every pair of one-sided disagreement ranges is nested.
3. `verify_budget_scaling.py` evaluates the explicit prime-order choice, second-moment coupon bound, exact binomial overflow probability, and resulting risk over a grid of n, d, m. The smallest certified ratio to $(d/n)\log(1 + (m \wedge n)/d)$ on the nontrivial grid is positive.

These checks are not used as proofs. They test the finite symmetry, the exceptional combinatorial lemma, and the parameter choices separately.