STATISTICAL EQUIVALENCE DOES NOT DETERMINE STREAMING MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Can asymptotically equivalent statistical experiments have different streaming-memory complexity? For every fixed $B > 0$, we construct explicit finite realizable learning problems with N concepts on N uniformly sampled points, where $R := \log_2 N$. Their labeled-sample experiments have Le Cam distance at most R^{-B} uniformly over the sample count, while their base Gram matrices and robust-SQ profiles agree exactly up to radius $\Theta(\sqrt{R/\log R})$. Nevertheless, one problem admits exact one-pass learning with $R^{O_B(1)}$ samples and $O(R^2/\log R)$ bits, whereas for every fixed ordered pass count p, the other requires $\Omega_p(R^2)$ bits or $N^{\Omega_p(1)}$ samples, even for improper learners with error $1/64$. Rare field equations and binary parity observations have identical pairwise correlations but different storage costs, separating unrestricted statistical simulation from fixed-pass streaming simulation.

1 INTRODUCTION

Statistical comparison asks whether observations from one problem can replace those from another without changing decision quality. Streaming learning additionally requires the relevant information to fit in a small working memory. We ask whether statistically equivalent learning problems can nevertheless have different streaming-memory complexity.

The candidate characterization of Gonen et al. (2020) uses ordinary statistical-query (SQ) dimension under nearby distributions. Bounding the density ratio gives a scalar robust-SQ profile indexed by the reweighting radius; unlike the base Gram matrix, this profile can expose correlations hidden by the sampling distribution.

Our pair rules out this profile as a characterization of bounded-memory learning. The two problems have identical base Gram matrices and identical robust-SQ profiles on a growing radius interval. More strongly, their complete labeled-sample experiments are asymptotically equivalent, uniformly over the sample count. Still, exact one-pass learning uses $O(R^2/\log R)$ bits on one side, whereas constant-accuracy learning on the other needs $\Omega(R^2)$ bits or exponentially more samples, where $R = \log_2 N$ and N is both the number of concepts and the number of domain points.

Write $E_{s,m}$ and $H_{s,m}$ for the experiments of m independent labeled examples at family index s, with a common target index θ. Their Le Cam distance Δ is the maximum of the two worst-target simulation errors in total variation, using unrestricted target-independent kernels. Write $\mathcal{R}_J(\mu)$ for the maximum ordinary SQ dimension under density ratios in $[1/\mu, \mu]$ on side J. Section 2 gives the precise definitions.

Theorem 1.1 (Statistical equivalence with a memory separation)**.** *For every fixed $B > 0$, there are explicit sequences of Boolean concept classes $\mathcal{C}_s^E, \mathcal{C}_s^H$, on uniform finite domains $\mathcal{X}_s^E, \mathcal{X}_s^H$, with a common concept indexing and*

$$|\mathcal{C}_s^E| = |\mathcal{C}_s^H| = |\mathcal{X}_s^E| = |\mathcal{X}_s^H| = N_s, \qquad R_s := \log_2 N_s \longrightarrow \infty,$$

such that, for all sufficiently large s,

$$\sup_{m \geq 0} \Delta(E_{s,m}, H_{s,m}) \leq R_s^{-B}.$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The easy class has a proper exact one-pass learner, successful with probability at least $2/3$ for every target, using $R_s^{O_B(1)}$ samples and $O(R_s^2/\log R_s)$ bits. For every fixed integer $p \geq 1$, any ordered p-pass learner for the hard class with error at most $1/64$ and per-target success at least $2/3$ must use

$$b = \Omega_p(R_s^2) \qquad \text{or} \qquad m = N_s^{\Omega_p(1)}.$$

The lower bound applies to randomized, nonuniform learners and arbitrary final-state hypotheses.

In addition, the base Gram matrices coincide, distinct concepts have disagreement probability at least $1/16$, and the robust-SQ profiles agree exactly for $1 \leq \mu \leq M_s$, where $M_s = \Theta(\sqrt{R_s/\log R_s})$. Samples and targets have $O(R_s)$-bit encodings. Evaluation and easy learning from encoded examples take polynomial bit time in R_s; exact uniform sampling takes expected polynomial bit time.

Proof sketch. The proof uses two sample-size regimes. Short datasets couple unless a rare row occurs. Long datasets identify their target from common observations, after which an unrestricted kernel resamples either experiment. The rare blocks replace binary parity constraints by projective field equations with the same Gram matrix, while an auxiliary sign code matches the robust-SQ profiles. Field equations admit compact elimination; binary parity retains the fixed-pass quadratic-memory lower bound. The dilution exponent then makes the experiment distance at most R_s^{-B}. □

For every common loss in $[0, 1]$, this experiment comparison bounds the difference in optimal unrestricted Bayes and minimax risks by R_s^{-B}. The precision exponent B is chosen before the family; increasing it raises the easy learner's polynomial sample budget. At polynomial sample budgets in R_s, the memory bounds differ by a factor of order $\log R_s$. Thus asymptotic equivalence of the unrestricted experiments does not determine their streaming-memory complexity.

Proof mechanism. Uniform comparison over all sample counts requires both a rare substitution and a common identifiable component. Exact correlation matching is supplied by projective field-equality observations: a positive label reveals a field equation, while the full block has the same pairwise correlations as a mixture of binary parity and constant rows. The easy learner retains k equations over $\mathbb{F}_{2^s}$ using $O(k^2 s)$ bits. The hard target has binary dimension ks and inherits the $\Omega(k^2 s^2)$ parity threshold. The auxiliary sign code matches the robust-SQ profiles within the smaller memory budget.

Relation to prior work. The SQ simulation in Steinhardt et al. (2016) provides a direction from statistical queries to limited-memory learning. The nearby-distribution bounds in Gonen et al. (2020) give sufficient and necessary conditions with different quantitative dependencies. Our theorem gives two problems with the same profile over a growing interval, the same base Gram matrix, and asymptotically equivalent sample experiments, but different streaming-memory requirements. The profile uses ordinary correlation-packing SQ dimension. The general statistical-query characterizations of Feldman (2017) use richer dimensions.

Quadratic-memory parity lower bounds appear in Raz (2016), and extractor-based bounds in Garg et al. (2018). We use the ordered constant-pass theorem of Lyu et al. (2023). Our reduction reuses each auxiliary choice across passes, preserves branching-program width, and converts a low-error improper hypothesis into exact parity recovery.

The computational deficiency of Akdemir (2025) restricts comparison kernels by polynomial time. We instead measure working memory. Corollary 6.1 shows that converting hard samples into a stream usable by the easy learner requires quadratic memory or exponentially many source samples, although an unrestricted comparison kernel has vanishing error.

2 Learning model and statistical comparison

All logarithms are base two. A finite realizable learning problem consists of a known class $\mathcal{C} \subseteq \{-1, 1\}^{\mathcal{X}}$ and a known distribution P on $\mathcal{X}$. Under target $c \in \mathcal{C}$, examples $(X, c(X))$ are independent with $X \sim P$. An output $h\colon \mathcal{X} \to \{-1, 1\}$ has error $\mathrm{err}_P(h, c) := P[h(X) \neq c(X)]$. A proper exact learner returns the index of c; an improper learner may return any such h.

A learner with b bits of working memory has at most 2^b states at each layer. On each of p ordered passes it reads the same m examples in their original order, carrying its state between passes. Layer-dependent transitions may be nonuniform and may use fresh randomness. All information retained

across sample accesses, including output buffers, belongs to the b-bit state, and the final hypothesis is a fixed function of the final state. This is the branching-program convention used for the lower bound in Lyu et al. (2023). The sample count is m, and the number of sample accesses is pm.

Definition 2.1 (Ordinary SQ dimension and robust profile)**.** For a distribution Q on $\mathcal{X}$, let $\mathrm{SQdim}_Q(\mathcal{C})$ be the largest integer r for which distinct $c_1, \dots, c_r \in \mathcal{C}$ satisfy

$$|\mathbb{E}_Q[c_i(X)c_j(X)]| \le 1/r \qquad (i \ne j).$$

For full-support P and $\mu \ge 1$, define the nearby-distribution profile

$$\mathcal{R}_{\mathcal{C},P}(\mu) \coloneqq \max_{Q:P(x)/\mu \le Q(x) \le \mu P(x)\ \forall x} \mathrm{SQdim}_Q(\mathcal{C}).$$

This is the ordinary-SQ nearby-distribution quantity of Gonen et al. (2020). The base Gram matrix has entries $G_{c,c'} \coloneqq \mathbb{E}_P[c(X)c'(X)]$.

An experiment J_m is the family of full labeled-transcript distributions $P^{(m)}_{J,\theta}$, indexed by a target $\theta \in \Theta$. A target prior may be used to evaluate Bayes risk, while the learning guarantee is required for every target.

Definition 2.2 (Le Cam comparison)**.** For experiments E_m, H_m with the same target set, define

$$\delta(E_m, H_m) \coloneqq \inf_{\mathsf{K}} \sup_{\theta \in \Theta} \|\mathsf{K}P^{(m)}_{E,\theta} - P^{(m)}_{H,\theta}\|_{\mathrm{TV}},$$

$$\Delta(E_m, H_m) \coloneqq \max\{\delta(E_m, H_m), \delta(H_m, E_m)\}.$$

Here $\|P - Q\|_{\mathrm{TV}} \coloneqq \frac{1}{2}\sum_z |P(z) - Q(z)|$, and K ranges over target-independent Markov kernels. This is the standard deficiency convention, as in Ray & Schmidt-Hieber (2019). Kernels may depend on the family index and sample count; their time and memory are unrestricted.

3 When rare substitutions preserve a statistical experiment

The statistical comparison uses two finite conditions. Both experiments have a recognizable rare block of target-independent probability ζ, and their complementary rows have identical probabilities and labels under a fixed identification. Moreover, every pair of targets disagrees on a common row with sampling probability at least $1/16$. Replacing each rare observation and its label by a marker produces a transcript W_m with the same conditional law on both sides.

Lemma 3.1 (Uniform comparison of full experiments)**.** *Under these conditions, with $N \coloneqq |\Theta| \ge 2$ and $R \coloneqq \log N$, put $p_m \coloneqq \min\{1, (N-1)(15/16)^m\}$. Then*

$$\Delta(E_m, H_m) \le \min\{1 - (1-\zeta)^m, p_m\},$$

$$\sup_{m \ge 0} \Delta(E_m, H_m) \le \max\{65R\zeta, N^{-4}\}.$$

For any common action space and loss in $[0,1]$, the optimal Bayes risks under each common prior, and the minimax risks, differ by at most the corresponding deficiency bound.

Proof. A fixed wrong target agrees with all retained observations with probability at most $(15/16)^m$. A decoder that selects the first consistent target therefore has worst-target error at most p_m, by a union bound over the $N-1$ alternatives.

For an early-sample simulation, preserve the common observations and map each rare observation to any destination rare observation. Couple the two transcripts on the event that all rows are common, giving error at most $1-(1-\zeta)^m$ in each direction. For a second simulation, decode the target from W_m and independently generate m fresh destination samples under that estimate. For each true target, the output law contains its exact destination law with weight at least $1 - p_m$, so the remaining mixture weight bounds the total variation error. Both rules define target-independent kernels.

Set $m_0 \coloneqq \lceil 64R \rceil$. Since $(15/16)^{64} < 1/32$, we have $p_{m_0} \le N^{-4}$. For $m \le m_0$, the early bound is at most $m\zeta \le 65R\zeta$. For $m \ge m_0$, the decoding bound is at most N^{-4}.

Composing a decision rule with either simulation kernel changes every target's expected loss by at most its total variation error. Taking the infimum over rules after averaging over a prior, or after taking the supremum over targets, gives the Bayes or minimax conclusion in both directions. □

Exact correlation matching. With a common complement and rare mass $\zeta > 0$, linearity gives

$$G_E - G_H = \zeta\big(G_{E,\mathrm{exc}} - G_{H,\mathrm{exc}}\big),$$

where the two matrices on the right use the conditional rare-block distributions. Thus rarity controls statistical distance, while exact base-Gram equality follows from matching the conditional rare-block Gram matrices. The next section constructs this match and separately aligns the robust-SQ profiles.

4 A CORRELATION-PRESERVING REPLACEMENT FOR PARITY

The experiment comparison uses rare rows and an identifiable common part. We choose the rare rows so that field equations and binary parity have identical pairwise correlations, then add an auxiliary sign code that matches the robust-SQ profiles within the easy learner's memory budget.

Fix an integer $a \geq 5$, independently of the growing index s. Choose

$$q \coloneqq 2^s, \qquad K \coloneqq 8q, \qquad D \coloneqq K^a,$$
$$k \coloneqq K^2, \qquad n \coloneqq ks, \qquad d \coloneqq q^k.$$

Here q controls the bits per field equation, K controls the profile window, and D dilutes the rare rows. Throughout, $s \geq 2$ is sufficiently large. Fix a binary basis identifying $\mathbb{F}_q^k$ with $\mathbb{F}_2^n$ and define $\chi_u(x) \coloneqq (-1)^{u \cdot x}$ in binary coordinates.

4.1 MATCHING THE EQUATION BLOCKS

Let $\mathcal{A}$ contain the nonzero vectors of $\mathbb{F}_q^k$ normalized to have first nonzero coordinate one, and put $A \coloneqq q^2(d-1)/(4(q-1))$. An equality label $+1$ at (x, y) reveals the field equation $\langle x, u\rangle = y$. The next identity allows these labels to replace a parity block while preserving its entire Gram matrix.

Lemma 4.1 (Projective equality and diluted parity)**.** *For $q = 2^s \geq 4$, $d = q^k$, and $A = q^2(d-1)/(4(q-1))$, consider the equality functions $e_u(a, y) \coloneqq 2\mathbf{1}\{y = \langle a, u\rangle\} - 1$ on $\mathcal{A} \times \mathbb{F}_q$, with $q/4$ copies of each row. Compare them to functions on A rows consisting of the d binary parity rows and $A - d$ constant-$+1$ rows. Both families have Gram entry 1 when $u = u'$ and $1 - d/A$ otherwise.*

Proof. There are $(d-1)/(q-1)$ projective directions and hence A copied equality rows. For $u \neq u'$, the kernel of the nonzero linear form $a \mapsto \langle a, u - u'\rangle$ contains q^{k-1} vectors. Therefore exactly

$$\frac{q^k - q^{k-1}}{q-1} = q^{k-1}$$

projective directions give distinct field evaluations. In each such direction, exactly two choices of y make e_u and $e_{u'}$ disagree. Including copies, the disagreement fraction is

$$\frac{2q^{k-1}(q/4)}{A} = \frac{d}{2A}.$$

Distinct binary parity functions disagree on exactly half their d rows and never on the constant rows. Their disagreement fraction is also $d/(2A)$. Converting disagreement to correlation proves the claim; identical indices agree on every row. □

4.2 A RECOVERABLE CODE FOR THE ROBUST-SQ PROFILE

A common constant block of probability one half constrains the SQ dimension under every bounded density ratio. Balanced sign groups attain that constraint independently of the equation block. We multiply every remaining label by an auxiliary sign function built from those groups. An explicit character code supplies enough separated indices to balance the concept and domain cardinalities and admits an efficient decoder.

Put $C \coloneqq K(K+1)/2 - 1$. For each $2 \leq t \leq K$, let Ω_t contain the vectors in $\{-1, 1\}^t$ whose coordinate sum is zero for even t and has absolute value one for odd t. Put $\mathcal{Z} \coloneqq \prod_{t=2}^{K} \Omega_t$, $T \coloneqq |\mathcal{Z}|$, and $\mathcal{V} \coloneqq \{(t, j) : 2 \leq t \leq K, 1 \leq j \leq t\}$. Thus $|\mathcal{V}| = C$ and $\log T = \Theta(K^2)$. For $v = (t, j)$, write $g_v(z) \coloneqq z_j^{(t)}$.

Use a fixed explicit bias-at-most-$1/2$ list $a_1, \dots, a_L \in \mathbb{F}_2^\ell$ with $L = \ell^{O(1)}$ from Naor & Naor (1993, Theorem 3.2). Its coordinates can be evaluated in polynomial time in $\log \ell$ from the list index. Choose the least integer $\ell \geq 2$ with $H := 8DTL \leq 2^\ell$, and let $\mathcal{W}$ be the first H binary vectors of length ℓ. The profile-code functions are

$$f_{v,w}(z, r) := g_v(z)(-1)^{w \cdot a_r}, \qquad (v, w) \in \mathcal{V} \times \mathcal{W}.$$

A target is $\theta = (u, v, w) \in \mathbb{F}_q^k \times \mathcal{V} \times \mathcal{W}$, giving $N := dCH = 8DdCTL$ target indices.

4.3 The exact profile from the common constant block

The balanced-group correlations attain the restriction imposed by a half-mass constant block.

Lemma 4.2 (Exact robust-SQ profile)**.** *Suppose $P(A_0) = 1/2$ and every concept equals $+1$ on A_0. For each $2 \leq t \leq K$, suppose the class contains t concepts whose pairwise correlations conditional on A_0^c equal $-1/(t-1)$ for even t and $-1/t$ for odd t. Define*

$$\tau_t := \begin{cases} \max\{1, t^2/(4t-2)\}, & t \text{ even}, \\ (t+1)/4, & t \text{ odd}, \end{cases}$$

$$r_K(\mu) := \max\{t \in \{2, \dots, K\} : \tau_t \leq \mu\}.$$

Then $\mathcal{R}_{\mathcal{C},P}(\mu) = r_K(\mu)$ for $1 \leq \mu < \tau_{K+1}$, and $\mathcal{R}_{\mathcal{C},P}(\mu) < 4\mu$ for every $\mu \geq 1$.

Proof. Let $c_1, \dots, c_r$ be an SQ witness under an admissible Q, with $r \geq 2$, and put $S := \sum_i c_i$. The correlation constraint gives $\mathbb{E}_Q S^2 \leq 2r - 1$. On A_0, $S^2 = r^2$ and $Q(A_0) \geq 1/(2\mu)$. Hence even r requires $r^2/(2\mu) \leq 2r - 1$. For odd r, $S^2 \geq 1$ on A_0^c, giving

$$1 + \frac{r^2 - 1}{2\mu} \leq \mathbb{E}_Q S^2 \leq 2r - 1.$$

These conditions are $\mu \geq \tau_r$ and imply $r < 4\mu$. The sequence $(\tau_t)_{t \geq 2}$ is nondecreasing, with $\tau_2 = \tau_3 = 1$ and strict increase thereafter.

For $t \geq 4$, use density $1/\tau_t$ on A_0 and $2 - 1/\tau_t$ on A_0^c. Both values lie in $[1/\tau_t, \tau_t]$, and the resulting correlation is

$$\frac{1}{2\tau_t} + \left(1 - \frac{1}{2\tau_t}\right) \rho_t = \frac{1}{t},$$

where ρ_t is the stated conditional correlation. For $t = 2, 3$, the base distribution gives correlations 0 and $1/3$. These witnesses and the upper constraint prove the claim. □

4.4 Assembling the common and rare rows

Each domain is the disjoint union in table 1, with uniform probability on its N copy-indexed rows. Block membership and every row coordinate are observed. The profile-code factor in every nonconstant label is $f := f_{v,w}(z, r)$.

Table 1: The paired domains. Every nonconstant row includes $(z, r) \in \mathcal{Z} \times [L]$. Multiplicities distinguish rows with identical labels.

Block	Number of rows	Label on easy / hard side
Constant	$N/2$	$+1$ / $+1$
Shared parity	$DdCTL = N/8$	$\chi_u(a)f$ / $\chi_u(a)f$
Profile code	$2DdCTL = N/4$	f / f
Rare	ATL	field equality / diluted parity
Balancing	$(DdC - A)TL$	f / f

The shared parity block contains DC copies of each (a, z, r) with $a \in \mathbb{F}_2^n$. The rare easy block contains $q/4$ copies of each (a, y, z, r) with $a \in \mathcal{A}$ and $y \in \mathbb{F}_q$, labeled $(2\mathbf{1}\{y = \langle a, u \rangle\} - 1)f$. The rare hard block contains one copy of each binary parity row (a, z, r), plus $A - d$ rows labeled f for each (z, r). The profile-code and balancing multiplicities per (z, r) are $2DdC$ and $DdC - A$.

Proposition 4.3 (Matched correlations and common identifiability)**.** *For every fixed $a \geq 5$, the construction has N distinct concepts and N uniformly sampled domain points on each side. The base Gram matrices coincide, and the robust-SQ profiles agree for $1 \leq \mu \leq K/8$, with $\mathcal{R}_J(\mu) < 4\mu$ for all $\mu \geq 1$. The common rows separate every target pair with probability at least $1/16$. The rare-block probability obeys*

$$\zeta \coloneqq \frac{A}{8DdC} = \frac{q^2(d-1)}{32(q-1)DdC}, \qquad \frac{q}{32DC} \leq \zeta \leq \frac{q}{24DC}.$$

Moreover, $\ell, \log H = \Theta(K^2)$ and $R \coloneqq \log N = \Theta(K^2 s)$. Rows and targets have $O(R)$-bit encodings, evaluation takes polynomial bit time in R, and exact uniform sampling takes expected polynomial bit time.

Proof. The integer $q/4$ makes A integral. Since $d \geq q \geq 4$,

$$\frac{q}{4} \leq \frac{A}{d} = \frac{q^2}{4(q-1)}\left(1 - \frac{1}{d}\right) < \frac{q}{3}.$$

Hence $A \geq d$ and, for our sufficiently large parameters with $DC > q/3$, $DdC - A > 0$. All multiplicities are nonnegative integers, and all listed domain rows have positive probability. The four nonconstant block sizes sum to $4DdCTL = N/2$, so each domain has exactly N points. There are $dCH = N$ target indices by construction.

The common blocks have identical labels and row counts. On either rare block, the label factors into a function of the field or parity row and the profile-code function. Independence of this row from (z, r) lets the correlations factor. Lemma 4.1 identifies the first factor in the two experiments for every u, u', so the base Gram matrices agree for every target pair.

At each (z, r) the nonconstant row count is $4DdC$, independently of (z, r), so its conditional law is uniform. Fix u and w, and for any $2 \leq t \leq K$ consider the t concepts indexed by $v = (t, 1), \ldots, (t, t)$. On each nonconstant row, all factors except the corresponding g_v are shared and cancel in pairwise products. Uniformity on Ω_t gives correlation $-1/(t-1)$ for even t and $-1/t$ for odd t, as calculated in lemma A.1. The common constant block and lemma 4.2 give the same exact profile up to τ_{K+1} and the all-radius upper bound. Since $K/8 < \tau_{K+1}$, this includes the claimed interval.

If two targets have different profile-code indices, the profile-code block distinguishes them with probability at least $(1/4)(1/4) = 1/16$. If the indices agree and $u \neq u'$, their shared parity labels disagree with probability $(1/8)(1/2) = 1/16$. Thus the concepts are distinct and the common experiment has the required separation. Division of the bounds on A/d by $8DC$ gives the bounds on ζ.

The central binomial coefficient is the largest of the $t+1$ coefficients in $(1+x)^t$. Therefore $2^t/(t+1) \leq |\Omega_t| \leq 2^t$, and summing over $2 \leq t \leq K$ gives $C - O(K \log K) \leq \log T \leq C$. Since L is polynomial in ℓ, the least admissible length satisfies $\ell = \log(DT) + O(\log\log(DT)) = \Theta(K^2)$, and hence $\log H = \Theta(K^2)$. With $k = K^2$ and $s = \Theta(\log K)$,

$$R = ks + \log C + \log H = \Theta(K^2 s).$$

A target index uses $n + O(\log C) + \ell = O(R)$ bits. A row needs $O(n+s)$ bits for its field or binary coordinates, C bits for z, and $O(\log L + \log(DdC))$ bits for the list and copy indices; each term is $O(R)$. Represent $\mathbb{F}_q$ using the first irreducible polynomial of degree s in a fixed order. Deterministic search and trial division take $2^{O(s)} = \mathrm{poly}(R)$ bit operations, after which evaluation uses field arithmetic, binary dot products, and generated profile-code rows.

For exact uniform sampling, draw a uniform $\lceil \log N \rceil$-bit integer and reject it if it is at least N. Each trial succeeds with probability greater than $1/2$. Unrank the accepted integer by block and copy index; balanced slices use binomial coefficients, and a projective direction is specified by its first nonzero coordinate and suffix. The ranks and coefficients have $O(R)$ bits, so sampling takes expected polynomial bit time in R. □

5 THE MEMORY SEPARATION

The two rare blocks have the same Gram matrix, but their labels expose different constraints. On the easy side a positive equality label reveals an equation over $\mathbb{F}_q$. Storing k such equations costs $O(k^2 s)$ bits. On the hard side the corresponding information arrives as binary parity constraints in dimension $n = ks$, where the quadratic memory threshold is $\Theta(k^2 s^2)$. The profile-code index can be recovered within the smaller budget.

Although H grows exponentially with ℓ, the learner does not search its CH possible profile-code indices. For a candidate v, each labeled row (z_i, r_i, y_i) gives the binary linear constraint $w \cdot a_{r_i} = \mathbf{1}\{y_i g_v(z_i) = -1\}$. The coefficient matrix is the same for every v, so the learner solves C systems successively. With $O(\ell + \log(CH))$ independent rows, small bias makes that matrix full rank and pairwise separation excludes all wrong indices with probability at least $11/12$. The resulting decoder uses $O(K^4)$ bits, as proved in lemma A.1.

5.1 ACCUMULATING THE EASY CONSTRAINTS

Proposition 5.1 (One-pass exact learner)**.** *For every fixed dilution exponent $a \geq 5$, the easy class has a proper exact one-pass learner with per-target success at least $2/3$, using*

$$m_E = O(DCk + \log H) = O(K^{a+4}),$$
$$b_E = O(k^2 s + K^4) = O(K^4 s).$$

Its bit running time, given encoded examples, is polynomial in R.

Proof. Use two phases with deterministic sample budgets. In the first phase take $8r_0$ examples, with r_0 as in lemma A.1, and retain the first r_0 profile-code examples. Their count before truncation has binomial mean $2r_0$. The lower-tail bound $\mathbb{P}[X < r_0] \leq \exp(-r_0/4)$ is at most $1/12$ for our choice of constant. Conditional on enough profile-code rows, their coordinates are independent and uniform on $\mathcal{Z} \times [L]$. The recovery guarantee in lemma A.1 therefore makes first-phase failure at most $1/6$. Once the index has been found, discard its examples and retain only (v, w).

Condition on correct first-phase recovery; the fresh second-phase samples remain independent. On a rare easy row (a, y, z, r) compute the unmasked label $y_{\mathrm{obs}} f_{v,w}(z, r)$. Keep the equation $\langle a, u \rangle = y$ only when this sign is $+1$. Conditional on any direction a, the observed field value is uniform before the label is revealed, so a positive occurs with probability $1/q$. Thus each example supplies an equation with probability $\zeta/q \geq 1/(32DC)$. The successful directions are independent and uniform on $\mathcal{A}$, including when $u = 0$.

Take $256DCk$ examples in this phase. The mean number of positives is at least $8k$, so the probability of fewer than $2k$ is at most $\exp(-9k/4)$ by the binomial lower-tail bound. For any nonzero $v' \in \mathbb{F}_q^k$, a uniform projective direction is orthogonal to v' with probability

$$\frac{q^{k-1} - 1}{q^k - 1} < \frac{1}{q}.$$

If the first $2k$ positive directions fail to span, some nonzero v' is orthogonal to all of them. A union bound makes this probability less than $(q^k - 1)q^{-2k} < q^{-k}$. For sufficiently large s, $\exp(-9k/4) + q^{-k} \leq 1/6$. The two phases therefore recover the exact target with probability at least $2/3$.

Process equations immediately into a row-echelon basis, storing at most k rows of $k+1$ field elements. This needs $k(k+1)s$ bits. First-phase storage is $O(r_0(C + \log L) + r_0\ell + \ell^2) = O(K^4)$. Field setup, row decoding, counters, target parameters, and generated small-bias rows fit within the same bound. Profile-code recovery solves C systems; the subsequent field elimination and concept evaluation take polynomial bit time in R. The total sample count is $8r_0 + 256DCk$. □

5.2 TRANSFERRING THE PARITY LOWER BOUND ACROSS PASSES

The shared block makes the target identifiable from a full dataset. To retain hardness through several passes, its simulation must also reproduce the same transformed row whenever the learner revisits a stream position.

Theorem 5.2 (Constant-pass parity lower bound, Lyu et al., 2023)**.** *For every fixed integer $p \geq 1$, there are constants $c_p, d_p, e_p > 0$ and $n_0(p)$ such that the following holds for all $n \geq n_0(p)$. Let U be uniform in $\mathbb{F}_2^n$, independently of $A_1, \ldots, A_m$, which are independent and uniform in $\mathbb{F}_2^n$. A deterministic ordered p-pass branching program of width at most $2^{c_p n^2}$, reading the same stream $(A_i, (-1)^{U \cdot A_i})_{i=1}^m$ in the same order on each pass, has exact-recovery probability at most $2^{-e_p n}$ if $m \leq 2^{d_p n}$.*

This is the parity consequence stated in Lyu et al. (2023, Theorem 2). Its formal general theorem is Theorem 6 in Section 8.3; the ordered multipass model is defined in Section 3.4, Definitions 2–4. For $p = 1$, one may append an inert second pass. The constants can depend on p, which remains fixed independently of n.

The following uniqueness fact converts an improper prediction rule into an exact parity index.

Lemma 5.3 (Final-output uniqueness)**.** *If $\{\phi_u\}_u$ is an orthonormal family in $L_2(P)$ and h takes values in $\{-1, 1\}$, then $\sum_u \langle h, \phi_u \rangle_P^2 \leq 1$. In particular, at most one u has $\langle h, \phi_u \rangle_P \geq 3/4$.*

Proof. Orthogonal projection onto the span of the family has squared norm $\sum_u \langle h, \phi_u \rangle_P^2$, at most $\|h\|_2^2 = 1$. Two correlations of size $3/4$ would contribute $2(3/4)^2 > 1$. □

Proposition 5.4 (Hardness under every fixed ordered pass count)**.** *For each fixed $a \geq 5$, fixed $p \geq 1$, and fixed $\rho > 0$, any randomized, nonuniform ordered p-pass learner that outputs an arbitrary hypothesis with hard-side error at most $1/64$ and per-target success at least ρ must use $b = \Omega_p(R^2)$ or $m = 2^{\Omega_p(R)}$.*

Proof. Fix a profile-code index $c_0 = (v_0, w_0)$ and reveal it to the learner. Let U be uniform in $\mathbb{F}_2^n$, identified with $\mathbb{F}_q^k$ through the fixed basis. We transform a uniform parity stream of length m into a hard-side stream with target (U, c_0).

For every stream position i, draw an auxiliary variable R_i, independently across positions and independently of $U, A_1, \ldots, A_m$, encoding a block choice, (z, r) when needed, and a copy index. Its law is target-independent. Split the rare hard block into its parity and constant parts. The resulting six block probabilities are

$$\frac{1}{2}, \quad \frac{1}{8}, \quad \frac{1}{4}, \quad \frac{1}{8DC}, \quad \frac{A-d}{8DdC}, \quad \frac{DdC-A}{8DdC},$$

for the constant, shared parity, profile-code, rare parity, rare constant, and balancing blocks, respectively. They sum to one by the row counts.

Given a parity example (a, b), shared and rare parity choices output the corresponding hard row with binary coordinate a and label $bf_{c_0}(z, r)$. The other choices output a row with known label $+1$ or $f_{c_0}(z, r)$ and consume (a, b). Conditional on U, the transformed stream is independent and uniform on the hard domain: each block is selected with its exact mass, and its remaining coordinates are uniform. Reuse the same R_i on every pass through position i, so the learner sees the same transformed dataset on every ordered pass.

Average the learner's success over U, the uniform parity stream, the auxiliary choices, and its random coins. The per-target guarantee makes this average at least ρ. Hence some fixing of $(R_1, \ldots, R_m)$ and the learner coins has average success at least ρ over the original independent uniform $U, A_1, \ldots, A_m$. Compile this fixing into the layer transitions of a deterministic nonuniform parity branching program. Its width is at most 2^b, its stream length is m, and its input parity stream retains the independent uniform law required by theorem 5.2.

It remains to relabel final states by parity indices. On the shared block, define

$$\phi_u(a, z, r, j) \coloneqq \chi_u(a) f_{c_0}(z, r),$$

where j is the copy index and the inner product uses the uniform conditional law on that block. Since $f_{c_0}^2 = 1$ and binary parities are orthogonal, the ϕ_u are orthonormal. If the final hypothesis h has total hard-side error at most $1/64$, its conditional error on the block of mass $1/8$ is at most $1/8$. Consequently,

$$\langle h, \phi_U \rangle \geq 1 - 2/8 = 3/4.$$

By lemma 5.3, each final hypothesis identifies at most one such u. Assign that u to the final state when it exists, and an arbitrary index otherwise. This is a nonuniform relabeling of the output map and adds no states. Whenever the original learner succeeds, the relabeled program recovers U exactly.

The resulting success at least ρ contradicts theorem 5.2 when both $b \le c_p n^2$ and $m \le 2^{d_p n}$, for sufficiently large s. Since $n = ks = \Theta(R)$, the desired resource alternative follows. □

6 Statistical transfer and its memory cost

The matched construction satisfies lemma 3.1, while propositions 5.1 and 5.4 give opposite memory bounds. Choosing the rare-block mass completes theorem 1.1.

Proof. Fix $B > 0$ and choose an integer $a > \max\{5, 2B + 1\}$. Use the construction in section 4 with this fixed exponent. By proposition 4.3, $q = K/8$, $R = \Theta(K^2 s)$, and $\zeta = \Theta(K^{-a-1})$. Since $s = \Theta(\log K) = \Theta(\log R)$, set

$$M_s := K/8 = \Theta\left(\sqrt{\frac{R}{\log R}}\right).$$

The Gram and profile claims follow from proposition 4.3.

The statistical comparison gives

$$\begin{aligned}\sup_{m \ge 0} \Delta(E_{s,m}, H_{s,m}) &\le \max\{65R\zeta, 2^{-4R}\} \\ &= O(sK^{1-a}) + 2^{-4R} = o(R^{-B}).\end{aligned}$$

The last equality uses $R = \Theta(K^2 s)$ and $a > 2B + 1$. Hence the discrepancy is at most R^{-B} for all sufficiently large s.

By proposition 5.1, $m_E = O(K^{a+4}) = R^{O_B(1)}$ and

$$b_E = O(K^4 s) = O(R^2/\log R).$$

The hard-side alternative follows from proposition 5.4 with $\rho = 2/3$, for each fixed p. This proves theorem 1.1. □

Lemma 3.1 supplies unrestricted kernels transferring datasets between the two problems with arbitrarily small fixed inverse-polynomial error. At a length sufficient for the easy learner, a streaming implementation of that transfer would solve the hard problem.

Corollary 6.1 (Memory required for streaming simulation)**.** *Fix a family from theorem 1.1, and let $t := m_E$ be a sufficient sample count for proposition 5.1. Suppose a target-independent simulator reads m hard examples in p ordered passes, for fixed p, and streams t easy examples once to the learner with worst-target total variation error at most a fixed $\eta < 2/3$. All reusable information is stored in its b_S-bit state. Then $b_S = \Omega_p(R^2)$ or $m = 2^{\Omega_p(R)}$.*

Proof. Compose the simulator with the easy learner, using independent coins. Total variation gives exact target recovery with probability at least $\rho := 2/3 - \eta > 0$ for every target. The composite uses p passes and at most $b_S + b_E + O(\log t + R)$ bits. Applying proposition 5.4 with this fixed ρ makes the composite state $\Omega_p(R^2)$ unless $m = 2^{\Omega_p(R)}$. Since $b_E = o(R^2)$ and $\log t + R = o(R^2)$, the same bound follows for b_S. □

In particular, take $m = t = R^{O_B(1)}$. An unrestricted kernel transfers the hard dataset to an easy dataset with error at most R^{-B}, whereas every fixed-pass streaming simulator with any fixed error below $2/3$ requires $\Omega_p(R^2)$ bits.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, were used extensively in developing the theoretical construction and extensions, drafting and checking proof arguments, retrieving and comparing literature, writing deterministic verification code, and drafting and editing the manuscript. AI assistance included proposing mathematical claims and proof ingredients. The verification work included explicit derivations, checks of imported results against primary sources, independent AI-assisted proof checks, and finite arithmetic checks. The authors bear responsibility for the claims, citations, and final submitted text.

REFERENCES

Deniz Akdemir. Approximate computation via le Cam simulability. arXiv:2512.24860, 2025. URL https://arxiv.org/abs/2512.24860.

Vitaly Feldman. A general characterization of the statistical query complexity. In *Proceedings of the 30th Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 785–830, 2017. URL https://proceedings.mlr.press/v65/feldman17c.html.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018. URL https://arxiv.org/abs/1708.02639.

Alon Gonen, Shachar Lovett, and Michal Moshkovitz. Towards a combinatorial characterization of bounded-memory learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/665d5cbb82b5785d9f344c46417c6c36-Abstract.html.

Xin Lyu, Avishay Tal, Hongxun Wu, and Junzhao Yang. Tight time-space lower bounds for constant-pass learning. In *64th Annual IEEE Symposium on Foundations of Computer Science*, 2023. URL https://arxiv.org/abs/2310.08070.

Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, 1993. doi: 10.1137/0222053.

Kolyan Ray and Johannes Schmidt-Hieber. Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 55(4):2195–2208, 2019. doi: 10.1214/18-AIHP946. URL https://arxiv.org/abs/1802.03425.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *57th Annual IEEE Symposium on Foundations of Computer Science*, 2016. URL https://eccc.weizmann.ac.il/report/2016/019/.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1490–1516, 2016. URL https://proceedings.mlr.press/v49/steinhardt16.html.

A PROFILE-CODE RECOVERY DETAILS

We prove separation and recovery for the auxiliary sign code used by propositions 4.3 and 5.1.

A.1 BALANCED GROUPS AND PROFILE-CODE RECOVERY

All coordinates below are uniform on their stated finite sets, independently across factors.

Lemma A.1 (Profile-code separation and recovery)**.** *The functions $f_{v,w}$ are distinct and have pairwise disagreement probability at least $1/4$ on $\mathcal{Z} \times [L]$. Moreover, for a sufficiently large absolute constant c_0, from $r_0 := \lceil c_0(\log(CH) + \ell + 1) \rceil$ independent labeled profile-code samples, the true index is the unique consistent member of $\mathcal{V} \times \mathcal{W}$ with probability at least $11/12$. It can then be found by solving C binary linear systems, with $O(r_0(C + \log L) + r_0\ell + \ell^2)$ bits of workspace and polynomial time in these parameters.*

Proof. For a uniform vector in Ω_t, sign symmetry gives zero coordinate means. Exchangeability and the specified sum give, for $i \neq j$,

$$\mathbb{E}[z_i^{(t)} z_j^{(t)}] = \frac{\mathbb{E}[(\sum_j z_j^{(t)})^2] - t}{t(t-1)} = \begin{cases} -1/(t-1), & t \text{ even}, \\ -1/t, & t \text{ odd}. \end{cases}$$

Coordinates in different groups have correlation zero by independence. Consequently, if $v \neq v'$ and $w = w'$, the correlation of $f_{v,w}$ and $f_{v',w'}$ is nonpositive. If $w \neq w'$, independence of z and r gives

$$\mathbb{E}[f_{v,w} f_{v',w'}] = \mathbb{E}[g_v g_{v'}] \cdot \frac{1}{L} \sum_{r=1}^{L} (-1)^{(w+w') \cdot a_r}.$$

The second factor has absolute value at most $1/2$ by small bias, and the first has absolute value at most one. In either case the correlation is at most $1/2$, which is equivalent to disagreement at least $1/4$.

Let the stored observations be (z_i, r_i, y_i). For each $v \in \mathcal{V}$, form the binary system

$$w \cdot a_{r_i} = \mathbf{1}\{y_i g_v(z_i) = -1\}, \qquad 1 \leq i \leq r_0.$$

For each nonzero $\Delta \in \mathbb{F}_2^\ell$, small bias gives $\mathbb{P}[\Delta \cdot a_r = 0] \leq 3/4$. Thus the common coefficient matrix fails to have full rank with probability at most $(2^\ell - 1)(3/4)^{r_0}$. A fixed wrong profile-code index survives all samples with probability at most $(3/4)^{r_0}$, so the probability that any wrong member survives is at most $(CH - 1)(3/4)^{r_0}$. The chosen constant c_0 makes their sum at most $1/12$.

On the complementary event, each consistent system has a unique solution, and exactly one solution belongs to $\mathcal{W}$. Solve the C systems successively, checking membership by the binary index bound $w < H$. The samples occupy $O(r_0(C + \log L))$ bits. Row generation and Gaussian elimination need $O(r_0\ell + \ell^2)$ further bits. Coordinates of a_r are generated from the explicit list. □

B PARITY EXTRACTOR VERIFICATION

The extractor hypothesis used to specialize the general lower bound to theorem 5.2 follows from the next calculation. Norms and expectations on $\mathbb{F}_2^n$ are uniform.

Lemma B.1 (Parity extractor parameters)**.** *Put $\kappa := \lfloor n/8 \rfloor$. For every nonzero nonnegative function f on $\mathbb{F}_2^n$ with $\|f\|_2/\|f\|_1 \leq 2^\kappa$, the fraction of $a \in \mathbb{F}_2^n$ satisfying*

$$\frac{|\mathbb{E}_u[(-1)^{a \cdot u} f(u)]|}{\|f\|_1} \geq 2^{-\kappa}$$

is at most $2^{-\kappa}$.

Proof. Parseval gives

$$\sum_{a \in \mathbb{F}_2^n} |\mathbb{E}_u[(-1)^{a \cdot u} f(u)]|^2 = \|f\|_2^2.$$

Each specified row contributes at least $2^{-2\kappa}\|f\|_1^2$. Dividing the resulting count by 2^n bounds its fraction by

$$2^{2\kappa - n} \frac{\|f\|_2^2}{\|f\|_1^2} \leq 2^{4\kappa - n} \leq 2^{-\kappa}.$$

The last inequality uses $5\kappa \leq n$. □