A PROPER SQ PTAS FOR AGNOSTIC LEARNING OF ARBITRARILY BIASED RELUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agnostic ReLU regression seeks to match the best single ReLU under arbitrary label noise. Under standard Gaussian inputs, we give a deterministic proper statistical query (SQ) polynomial-time approximation scheme (PTAS) for this problem with arbitrary bias. For every fixed $\eta > 0$, the learner returns a ReLU with weight norm at most W and squared loss at most $(1+\eta)\mathrm{OPT}_W + \varepsilon$, where OPT_W is the minimum loss over the same class. This sharpens prior constant-factor guarantees to a factor arbitrarily close to one while preserving the model class and norm constraint. To obtain this guarantee, we reduce the search to $O(\eta^{-2})$ dimensions while preserving the comparator's Gaussian output distribution and controlling its label correlation. For fixed η and a known label second-moment bound, the query count, inverse tolerance, and arithmetic cost are polynomial in the dimension, W, the moment bound, and $1/\varepsilon$. Moreover, the same guarantee extends to unknown finite label second moments without polynomial dependence on label scale: only logarithmically many additional queries are needed to locate the label mean.

1 INTRODUCTION

Rectified linear units (ReLUs) are basic components of neural networks. Networks built from these units have been used for image classification (Krizhevsky et al., 2012) and speech recognition (Maas et al., 2013). Each unit applies an affine score and sets negative values to zero. Learning a single such unit is a basic regression task and a useful setting for studying the computational limits of neural network learning. The central question is how accurately an efficient algorithm can fit this unit when the data contain noise or the model is misspecified.

We study this question in the *agnostic* setting: the labels need not be generated by a ReLU, and their dependence on the inputs is unrestricted. The goal is to approach the smallest expected squared loss achievable by the chosen ReLU class. Gaussian inputs provide a standard benchmark for this problem (Diakonikolas et al., 2020a; Guo & Vijayaraghavan, 2025). They give the inputs a precise geometric structure while leaving the label noise unrestricted.

Formally, let $(X, Y) \sim \mathcal{D}$, with $X \sim \mathcal{N}(0, I_d)$ and $\mathbb{E}Y^2 < \infty$. Write $z_+ = \max\{z, 0\}$ and define

$$f_{w,b}(X) = (w^T X + b)_+, \qquad R(f) = \mathbb{E}(Y - f(X))^2, \qquad N = \mathrm{OPT}_W = \inf_{\|w\| \le W,\ b \in \mathbb{R}} R(f_{w,b}). \tag{1}$$

We seek risk at most $(1+\eta)N + \varepsilon$. Here η controls the multiplicative error and ε is an additive tolerance. This guarantee approaches the best achievable risk, including when N is small. We also require the output to be a single ReLU with weight norm at most W. This requirement is called *proper learning*. It preserves the original model class; an accurate polynomial predictor need not have an equally accurate ReLU replacement.

Allowing any bias b introduces two difficulties. A large negative bias makes the neuron active only in a rare Gaussian tail. An approximation that ignores this region can lose the improvement needed to approach N. A large positive bias allows an arbitrarily large output mean, even with bounded weights. Appending a constant input coordinate does not remove these difficulties: the augmented input is no longer a standard Gaussian.

Prior results identify both the opportunity and its limits. For homogeneous ReLUs, Diakonikolas et al. (2020a, Theorem 19) give an improper polynomial-time approximation scheme (PTAS). The

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.

proper learner of Diakonikolas et al. (2021a, Theorem 27) achieves additive accuracy, but its dimension exponent depends on that accuracy. For arbitrary bias, Guo & Vijayaraghavan (2025) give a proper constant-factor statistical query (SQ) algorithm. They also show that correlational SQ (CSQ) access is too weak even for that guarantee. Stronger approximation schemes face separate barriers. Subexponential-LWE hardness rules out a fully polynomial approximation scheme (FPTAS) (Diakonikolas et al., 2023b). The sparse-parity construction of Goel et al. (2019) also obstructs an efficient PTAS (EPTAS) under parameterized sparse-LPN hardness (Appendix G). These results leave a specific target: a *proper PTAS for arbitrary bias using general SQ queries*.

1.1 Our Results

We achieve this target for arbitrarily biased ReLUs. The first theorem uses a known second-moment bound.

Theorem 1.1 (Deterministic proper SQ PTAS)**.** *Let $X \sim \mathcal{N}(0, I_d)$, $W > 0$, and $0 < \eta, \varepsilon \le 1$. Suppose $\mathbb{E}Y^2 \le 1$ is known and bounded STAT access is available. Algorithm 1 is deterministic and returns a single ReLU $\widehat{f} = f_{\widehat{w},\widehat{b}}$ with $\|\widehat{w}\| \le W$ and*

$$R(\widehat{f}) \le (1+\eta)N + \varepsilon. \tag{2}$$

For each fixed η, its query number, inverse tolerance, and arithmetic cost are $\mathrm{poly}_\eta(d, W, 1/\varepsilon)$. The guarantee holds for every valid adaptive sequence of oracle answers. The construction uses a transverse Hermite degree k and a search dimension at most $m + 2$, where

$$k = O(\eta^{-2}), \qquad m = O(\eta^{-2}). \tag{3}$$

A supplied bound $\mathbb{E}Y^2 \le B$, $B \ge 1$, adds polynomial dependence on B.

Here poly_η allows both constants and polynomial exponents to depend on η; this is a PTAS. In particular, fixing η keeps the dimension exponent independent of ε. The next result removes the need to supply a moment bound and avoids polynomial dependence on the label scale.

Theorem 1.2 (Unknown label scale)**.** *In the model* (1)*, assume only $\mathbb{E}Y^2 < \infty$. Algorithm 2 is deterministic and returns a ReLU with $\|\widehat{w}\| \le W$ attaining* (2)*, with query cost*

$$\mathrm{poly}_\eta(d, W, 1/\varepsilon) + O(\ell), \qquad \ell = 1 + \log(2 + |\mathbb{E}Y| + \sqrt{\mathrm{Var}\, Y}). \tag{4}$$

Inverse tolerance remains $\mathrm{poly}_\eta(d, W, 1/\varepsilon)$, independent of the label moments. The same bound applies to arithmetic-operation counts; bit complexity additionally depends on the lengths of the numbers represented. The conditional degree bounds in Corollary 4.2 also apply.

The logarithmic term pays for locating the label mean with bounded queries. Proposition 5.1 shows that some dependence on numerical location is unavoidable, even for constant labels.

1.2 Main Ideas

The proof first finds useful coordinates and then constructs a proper replacement. For a nonconstant optimum $f_* = a(v^T X - t)_+$, positive-label weights and a deterministic ellipsoid grid give coarse directions in the rare-activation regime. When the signal exceeds the target risk scale, one correlation vector identifies a useful plane. Together these steps produce a polynomial-size list of subspaces L, each of dimension at most two, one satisfying

$$J_L = \mathbb{E}\|\nabla_{L^\perp} f_*\|^2 = a^2 \|P_{L^\perp} v\|^2 \overline{\Phi}(t) = O(N + \varepsilon).$$

A histogram in L and a transverse Hermite expansion then approximate f_* at squared error $O(\eta^2(N + \varepsilon))$. A derivative matrix weighted by $1/j$ at degree j isolates $O(\eta^{-2})$ directions: the comparator's gradients occupy one transverse direction, and the residual risk controls the remaining spectrum. A reflection produces a single ReLU in the resulting subspace while preserving its weight norm and Gaussian output distribution. Controlling the change in label correlation gives relative error. Preservation of the mean also supports the candidate filter that handles unknown label scale.

1.3 Related Work

Table 1 places the result between the known approximation algorithms and the relevant impossibility results. We distinguish computational assumptions from unconditional oracle lower bounds.

Table 1: Algorithms and barriers for ReLU regression. Inputs are Gaussian; the first algorithm also covers subgaussian marginals. N is the optimum for each row's comparator class. The first two algorithms assume bounded labels; ours permits finite second moment. The FPTAS and EPTAS consequences use the reductions in Appendix G, with $W = 1$ and $\mathbb{E}Y^2 \leq 1$.

Work	Setting	Guarantee or obstruction
Algorithms		
Diakonikolas et al. (2020a)	Homogeneous	Improper $(1+\eta)N + \varepsilon$; PTAS
Diakonikolas et al. (2021a)	Homogeneous comparator	$N + \varepsilon$; affine ReLU output; accuracy-dependent exponent
Guo & Vijayaraghavan (2025)	Arbitrary bias	Proper $CN + \varepsilon$; polynomial SQ
This work	Arbitrary bias	Proper $(1+\eta)N + \varepsilon$; polynomial SQ for fixed η
Barriers		
Diakonikolas et al. (2023b)	General algorithms	No FPTAS under subexponential-LWE hardness
Goel et al. (2019)	Sparse-parity reduction	No EPTAS under parameterized sparse-LPN hardness
Diakonikolas et al. (2020b; 2021b)	General SQ	No SQ EPTAS with EPTAS query count and inverse tolerance
Guo & Vijayaraghavan (2025)	CSQ; arbitrary bias	No polynomial-resource $CN + \varepsilon$ learner for fixed C

Oracle and distributional limitations. Diakonikolas et al. (2020b) and, independently, Goel et al. (2020) establish general-SQ lower bounds for Gaussian agnostic ReLU learning; Diakonikolas et al. (2021b) strengthen and relate these bounds to polynomial approximation. In particular, an SQ EPTAS with both query count and inverse tolerance of EPTAS form is unconditionally excluded (Appendix G.3). Separately, the arbitrary-bias CSQ lower bound of Guo & Vijayaraghavan (2025, Theorem 2) already excludes a constant-factor approximation. Our positive-label weights use nonlinear label information available to general SQ queries. These oracle bounds are not unconditional hardness results for all sample-based algorithms. Without a marginal assumption, Diakonikolas et al. (2022a) exclude even constant-factor approximation by general SQ algorithms, and by arbitrary algorithms under a random-XOR refutation assumption.

Further algorithmic context. Polynomial methods give dimension-efficient learners on the unit sphere (Goel et al., 2017); gradient and surrogate-loss methods give guarantees for homogeneous or restricted-bias neurons (Frei et al., 2020; Diakonikolas et al., 2022b; Awasthi et al., 2023; Wang et al., 2023). Gaussian smoothing and related methods yield constant-factor guarantees for broader monotone activations (Zarifis et al., 2025; Wang et al., 2025). Pointwise label queries provide a different access model (Diakonikolas et al., 2025). Our contribution combines arbitrary bias, proper norm-feasible output, and factor $1 + \eta$ in general bounded STAT; it does not claim an optimal dependence on η.

Our proof builds on polynomial regression (Kalai et al., 2008), derivative-based dimension reduction (Diakonikolas et al., 2021a; Constantine et al., 2014), weighted mean lists (Charikar et al., 2017; Diakonikolas et al., 2018a), and deterministic covering (Dadush, 2013). Unlike the thresholded second moments in Guo & Vijayaraghavan (2025), our list retains positive label magnitudes as weights. Bounded-query implementation uses standard SQ mean-estimation tools (Kearns, 1998; Feldman et al., 2017; Feldman, 2017).

2 PRELIMINARIES

Notation and learning model. All expectations refer to the joint law $\mathcal{D}$ unless a marginal is specified. Labels may depend arbitrarily on X. Vector norms are Euclidean, P_L is orthogonal projection onto a subspace L, and $\nabla_{L^\perp} f = P_{L^\perp} \nabla f$. For random variables, $\langle f, g \rangle = \mathbb{E}[fg]$ and

$\|f\| = (\mathbb{E}f^2)^{1/2}$. Write $z_- = (-z)_+$ and let ϕ and $\overline{\Phi}$ be the standard Gaussian density and upper tail. The feasible class in (1) contains zero and every nonnegative constant. Properness requires the output to belong to this same class, including the bound $\|w\| \le W$.

Comparator parameterization and optimality. An optimum exists (Appendix A). A nonconstant optimum can be written as $f_* = a(v^T X - t)_+$, where $a = \|w_*\| \in (0, W]$, $v = w_*/a$, and $t = -b_*/a$. Thus $\|v\| = 1$ and its activation probability is $p = \overline{\Phi}(t)$. Set $B = \mathbb{E}Y^2$ and $A = \mathbb{E}f_*^2$. Scaling down is feasible even when $a = W$, while bias perturbations are unrestricted. The resulting first-order conditions are

$$\mathbb{E}Y f_* \ge A, \qquad \mathbb{E}[Y \mathbf{1}_{v^T X > t}] = \mathbb{E}f_*. \tag{5}$$

Consequently $A \le B$ and $(b_*)_+ \le \mathbb{E}f_* \le \sqrt{B}$. If $a < W$, two-sided scaling also gives $\mathbb{E}Y f_* = A$; we use this stronger identity only under that condition. These distinctions let the analysis cover optima on the weight boundary.

Statistical queries and resource conventions. A bounded $\mathrm{STAT}(\tau)$ query specifies a computable function $q(X, Y) \in [-1, 1]$ and receives any value within τ of $\mathbb{E}q$ (Kearns, 1998). Queries may be adaptive, and our guarantees hold for every valid sequence of answers. General SQ queries may depend nonlinearly on Y, as our positive-label weights do; CSQ access is restricted to label correlations with functions of X. With unbounded labels, correlations must be implemented by bounded queries. We use clipping and Gaussian feature moments for this purpose (Appendix C.4). Query count, inverse tolerance, and arithmetic cost, including query evaluation to the stated precision, are counted separately. Bit complexity additionally charges the lengths of represented numbers.

Gaussian projections. For orthonormal coordinates $Z = P_L X$ and $G = P_{L^\perp} X$, the two Gaussian components are independent. The normalized probabilists' Hermite polynomials satisfy $\mathbb{E}[H_\alpha(G) H_\beta(G)] = \mathbf{1}_{\alpha=\beta}$ and $\partial_i H_\alpha = \sqrt{\alpha_i} H_{\alpha - e_i}$. If $f(Z, G) = \sum_\alpha c_\alpha(Z) H_\alpha(G)$ has square-integrable weak derivatives, then

$$\mathbb{E}\|\nabla_G f\|^2 = \sum_\alpha |\alpha| \mathbb{E}c_\alpha(Z)^2, \qquad \|f - P^G_{\le k} f\|^2 \le \frac{\mathbb{E}\|\nabla_G f\|^2}{k+1}. \tag{6}$$

Here $P^G_{\le k}$ retains all transverse Hermite terms of total degree at most k. These identities, and Gaussian Poincare $\mathrm{Var}\, f \le \mathbb{E}\|\nabla f\|^2$, explain why transverse energy controls approximation (O'Donnell, 2014). ReLU derivatives are understood in the weak sense; Appendix A records the required scalar moments.

Throughout, $0 < \eta, \varepsilon \le 1$, $e = \varepsilon/C_0$, and $n = N + e$, with C_0 a sufficiently large absolute constant. Other absolute constants c, C may change between displays. Sections 3–4 and 5.1 use $B \le 1$. A supplied bound $B \ge 1$ is handled by replacing (Y, W, ε) with $(Y/\sqrt{B}, W/\sqrt{B}, \varepsilon/B)$ and rescaling the output. Section 5.2 removes this supplied bound.

3 Finding a Subspace at the Relative-error Scale

We seek transverse derivative energy $O(N + e)$. This suffices for relative-error Hermite approximation.

Use the parameterization and quantities a, v, t, p, A, B, n from Section 2. Unless zero already satisfies $B \le (1 + \eta/4)N + e$, we also have

$$A/B \ge c\eta^2, \qquad A \ge c\eta e. \tag{7}$$

Indeed, $B - N \le 2\sqrt{AB}$, whereas failure of zero gives $B - N \ge c(\eta B + e)$. Appendix A gives the details. A constant optimum is handled by a constant grid.

For a subspace L, write $J = J_L = a^2 \|P_{L^\perp} v\|^2 p$.

Lemma 3.1 (Useful subspaces)**.** *A list of subspaces L, each of dimension at most two, can be constructed deterministically with $\mathrm{poly}_\eta(d, W, 1/e)$ bounded SQ resources. Unless zero is sufficient, one choice satisfies*

$$J \le Cn. \tag{8}$$

3.1 A USEFUL MEAN WEIGHTED BY POSITIVE LABELS

Lemma 3.2. *Suppose the optimum has $t \geq 1$ and $A/B \geq \Delta > 0$. For $a/(64t) \leq \tau \leq a/(32t)$, set $Y_\tau = Y\mathbf{1}_{Y\geq\tau}$ and $I = \{Y \geq \tau,\ v^T X > t\}$. The weighted masses $s = \mathbb{E}Y_\tau$ and $s_I = \mathbb{E}[Y\mathbf{1}_I]$ satisfy*

$$s_I \geq \frac{ap}{4t}, \qquad s \leq \frac{B}{\tau}, \qquad \frac{s_I}{s} \geq \alpha := \frac{\Delta}{512}. \tag{9}$$

The weighted mean $\mu_I = \mathbb{E}[YX\mathbf{1}_I]/s_I$ obeys

$$\|\mu_I - tv\| \leq \frac{\sqrt{Bp}}{s_I} \leq \frac{C}{\sqrt{\Delta}}. \tag{10}$$

Bias optimality forces enough positive-label mass on the activation event. A dimension-free Gaussian second-moment bound then controls its weighted mean by Cauchy–Schwarz. Appendix B gives the proof, using only the label second moment.

3.2 A DETERMINISTIC GRID FOR ALL WEIGHTED MEANS

Lemma 3.3. *Let $\psi \geq 0$ have $\mathbb{E}\psi^2 \leq 1$ and $s = \mathbb{E}\psi \geq s_{\min} > 0$. Given $0 < \alpha \leq 1$, bounded SQ estimates of s and $\mathbb{E}[\psi XX^T]$ suffice to construct a deterministic list of $O(s_{\min}^{-1/2})$ points. For every event I with $\mathbb{E}[\psi\mathbf{1}_I] \geq \alpha s$, the list contains a point within $C/\sqrt{\alpha}$ of $\mu_I = \mathbb{E}[\psi X\mathbf{1}_I]/\mathbb{E}[\psi\mathbf{1}_I]$. All resources are polynomial in $d, 1/s_{\min}, 1/\alpha$.*

Proof sketch. Weighted Jensen confines the admissible means to an ellipsoid. The Gaussian trace bound (34) controls its volume after removing low-eigenvalue directions; a lattice grid and prefix enumeration give the stated coverage and list size. Appendix B gives the construction and precision analysis. □

Applied to $\psi = Y_\tau$, the two lemmas give a point z with $\|z - tv\| \leq C/\sqrt{\Delta}$. For $A \leq n$, enumerate dyadic Δ from $1/4$ to below $c\eta^2$. One choice yields $J \leq Cn$, using either the axis of z or $L = \{0\}$ (Appendix B.3).

3.3 ONE CORRELATION VECTOR WHEN THE SIGNAL IS LARGER

If $A > n$, then $B \leq (\sqrt{A} + \sqrt{N})^2 \leq 4A$. The list with $\Delta = 1/4$ therefore has an absolute error bound. For t bounded above, the first-moment direction suffices:

$$\mathbb{E}[YX] = apv + r_1, \qquad \|r_1\| \leq \sqrt{N}. \tag{11}$$

Estimating this vector to error $\sqrt{e}$ gives $J \leq 2n/p = O(n)$. For large t, the list gives $\angle(u, v) \leq K/t$, with an absolute constant K.

Choose a threshold guess $|T - t| \leq 1/(t+1)$ and let $p_T = \overline{\Phi}(T)$. Estimate the single vector

$$g(u) = \mathbb{E}[YP_{u^\perp}X\mathbf{1}_{u^TX>T}] \quad \text{to error } \sqrt{ep_T}, \qquad H = \operatorname{span}(u, \widehat{g}(u)). \tag{12}$$

Projected correlations also appear in gradient analyses (Diakonikolas et al., 2022b; Guo & Vijayaraghavan, 2025). Here we use the plane H directly. Gaussian integration by parts, with $Y = f_* + \xi$, gives

$$g(u) = aq(u)P_{u^\perp}v + r(u), \quad q(u) = \mathbb{P}(u^TX > T, v^TX > t), \quad \|r(u)\| \leq \sqrt{Np_T}. \tag{13}$$

The angle bound implies $q(u) \geq c_K p_T$ and $p_T \asymp p$. Projecting (13) onto $H^\perp$ eliminates the observed vector and yields

$$aq(u)\|P_{H^\perp}v\| \leq \sqrt{Np_T} + \sqrt{ep_T}, \qquad a^2p\|P_{H^\perp}v\|^2 \leq C_K n. \tag{14}$$

Take $L = H$. Equation (14) is exactly $J_L \leq C_K n$, proving (8) when $A > n$. No direction search inside the plane is needed.

Appendix C gives the joint-tail bound, threshold grid, and stable construction for small $\widehat{g}$. Each vector estimate uses d scalar bounded queries after clipping. For the weighted lists, (9) and (7) supply $s_{\min} = c\eta e/(W+1)$, and a dyadic grid contains a valid label threshold. The queries depend nonlinearly on Y, consistent with the CSQ lower bound of Guo & Vijayaraghavan (2025). This completes the proof of Lemma 3.1.

4 PROPER DIMENSION REDUCTION AT RELATIVE ERROR

A good subspace controls approximation error. To obtain a proper output, we move the comparator into a larger computable subspace while preserving its Gaussian distribution. Fix a good L, put $r_0 = \dim L \le 2$, and write $Z = P_L X$, $G = P_{L^\perp} X$ in orthonormal coordinates. Partition each coordinate of Z into intervals and an outer cell. Let P project onto functions constant on each product cell in Z and polynomial in G, retaining *all* transverse Hermite polynomials of total degree at most k. This histogram space includes constants. When $r_0 = 0$, omit the histogram. Gaussian Sobolev identities give

$$\|(I - P)f_*\| \le \theta\sqrt{n}, \qquad \mathbb{E}\|\nabla_G P f_*\|^2 \le J, \tag{15}$$

for $k \ge C/\eta^2$ and $\theta = c\eta$. Appendix D gives finite parameters and bounded-query estimates. Crucially, P commutes with every transverse orthogonal rotation.

Equation (6) controls the transverse tail by $J/(k+1)$; Lipschitz continuity controls histogram error. Appendix D.1 gives the full approximation argument.

Lemma 4.1 (Proper replacement with relative error)**.** *Suppose* (15) *holds, and $\widehat{g} \in \operatorname{im} P$ estimates $g = PY$ with $\|\widehat{g} - g\| \le \delta$. Write $\widehat{g}_j$ for its transverse Hermite component of degree j, and form*

$$M_g = \sum_{j=1}^{k} \frac{1}{j} \mathbb{E}[\nabla_G \widehat{g}_j \, \nabla_G \widehat{g}_j^T]. \tag{16}$$

For an integer $m \ge 2$, let U span eigenvectors for the m largest eigenvalues, breaking ties arbitrarily and retaining all directions if fewer are available. Let λ be the mth largest eigenvalue, or zero if all transverse directions are retained. There is a feasible ReLU f' whose weight lies in $L \oplus U$ such that

$$\lambda \le \min\left\{ \frac{\operatorname{tr}(M_g)}{m}, \frac{(\sqrt{N} + \delta)^2}{m - 1} \right\}, \tag{17}$$

$$R(f') - N \le 4\big((\sqrt{\lambda} + \delta)\sqrt{J} + \theta\sqrt{nN} + \theta^2 n\big). \tag{18}$$

The search dimension is at most $m + r_0 \le m + 2$.

Proof sketch. The gradients of Pf_* share one transverse direction, so the reciprocal-degree weights bound the remaining spectrum by the residual risk. Moving the comparator's discarded norm into the last retained direction preserves its Gaussian output distribution. A reflection commuting with P and weighted Cauchy–Schwarz bound the change in projected label correlation by $2\sqrt{\lambda J}$. Projection and coefficient errors give (18). Appendix D.2 gives the spectral calculation, reflection construction, and exact risk identity. □

For every listed subspace, use the same parameters:

$$k = m = \lceil C/\eta^2 \rceil, \qquad \delta^2 = c\eta^2 e. \tag{19}$$

For the useful choice, (8) and (17) give

$$\lambda \le \frac{(\sqrt{N} + \delta)^2}{m - 1} \le \frac{Cn}{m - 1}. \tag{20}$$

Thus the replacement error is at most $(\eta/8)n$ for sufficiently large C and small c, in dimension at most $m + 2$. This proves (3). The algorithm always retains at most m directions.

Corollary 4.2 (Smaller degrees under additional promises)**.** *In Theorem 1.1, keep $m = O(\eta^{-2})$. If some optimum has weight norm strictly below W, then $k = O(\eta^{-5/3})$ suffices. If some optimum f_* has nonnegative bias or $\mathbb{E} f_*^2 \ge N + \varepsilon$, then $k = O(\eta^{-4/3})$ suffices.*

Appendix D.3 proves the corollary using Lemma D.1 and, for an interior optimum, $B = N + A$. The spectral rank and proper-replacement argument are unchanged.

5 A FINITE BOUNDED-STAT LEARNER

5.1 KNOWN LABEL SCALE

We describe the normalized construction. For each projection, keep the m largest eigenvalue directions of (16) and form one feasible ReLU net. The rank bound holds for every projection; the useful one gives the risk guarantee.

Algorithm 1 Deterministic proper SQ approximation scheme, $\mathbb{E}Y^2 \le 1$

Require: Bounded STAT access; d, W, η, ε.

1: Set $e = \varepsilon/C_0$, $\delta^2 = c\eta^2 e$, and $k, m = \lceil C/\eta^2 \rceil$; use a smaller k from Corollary 4.2 when its condition is promised.
2: Estimate the first-moment axis.
3: For each $\Delta \in \mathcal{D}_\eta$, form ellipsoid-list axes using $Y_\tau = Y\mathbf{1}_{Y \ge \tau}$. For $\Delta = 1/4$, also add the planes from (12) over the threshold grid.
4: Form a degree-k histogram–Hermite projection for every listed axis or plane, and for $L = \{0\}$. Start the candidate set with zero and a constant grid.
5: **for** each projection P **do**
6: Estimate PY to L^2 accuracy δ and compute M_g in (16).
7: Retain the m largest eigenvalue directions of M_g, or all directions if fewer are available.
8: Enumerate the feasible ReLU net in $L \oplus U$ and add it to the candidate set.
9: **end for**
10: Estimate the true risk difference of every candidate against zero to error ce, and return a candidate with smallest estimate.

Each of the $r_0 \le 2$ coordinate histograms has interval widths between $h/2$ and h, with

$$h = \frac{c\eta\sqrt{e}}{W+1}, \qquad R = \sqrt{C \log \frac{C(W+1)}{\eta^2 e}}, \tag{21}$$

on $[-R, R]$, and one outer cell. The smallest product-cell mass $\pi_{\min}$ remains inverse polynomial because $r_0 \le 2$. With D orthonormal features, estimate each $\mathbb{E}[Y\varphi]$ to accuracy $s = \delta/\sqrt{D}$ by one bounded query $\mathrm{clip}_T(Y\varphi)/T$. Hölder gives

$$|\mathbb{E}[Y\varphi] - \mathbb{E}[\mathrm{clip}_T(Y\varphi)]| \le \left(\frac{B^2 \mathbb{E}\varphi^4}{T}\right)^{1/3}. \tag{22}$$

Only $\mathbb{E}Y^2 \le B$ is needed: the label may depend on the feature. Appendices C.4 and D.4 give explicit clipping levels and a sharper inverse-tolerance bound using higher Gaussian moments. Form $\widehat{g}$ in the original basis and compute its derivative matrix from $\partial_j H_\alpha = \sqrt{\alpha_j} H_{\alpha - e_j}$.

In each retained subspace, enumerate a net of feasible weights and biases. The useful bias interval is

$$-CW\sqrt{\log \frac{C(W+1)}{\eta e}} \le b \le 1. \tag{23}$$

An L^2 prediction net of accuracy $c\eta\sqrt{e}$ follows from a parameter net because ReLU is 1-Lipschitz. Project weight-net points back into the norm ball. The dimension depends only on η, so the net has polynomial size for fixed η. Add a constant grid and zero.

Validate each actual ReLU through $\mathbb{E}f^2 - 2\mathbb{E}Yf$ to accuracy ce. The Gaussian norm term is known; the correlation uses (22). This compares true risk differences.

Finite precision. All label thresholds and scalar bisection endpoints are exact dyadic rationals. Clipping, polynomial evaluation, coordinate approximation, and eigenspace computation use inverse-polynomial accuracy. Appendix D.5 bounds the Rayleigh quotient outside the first $m-1$ eigenvectors under matrix perturbations; no eigenvalue gap is required.

Proof sketch of Theorem 1.1. Zero and the constant grid handle their respective cases. Otherwise Lemma 3.1 supplies a useful L, and Lemma 4.1 gives a feasible replacement of risk $N + (\eta/8)(N +$

e) in dimension at most $m+2$. Rounding to the parameter net adds $O(\eta)N + O(e)$; bounded-query comparison adds $O(e)$. The feature count is $d^{O(k)}$ times a polynomial histogram factor, and the net size has exponent $O(m+2)$. Since k, m depend only on η, all resources are polynomial for fixed η. Appendix D.7 gives the full risk and resource calculation. □

5.2 Unknown Label Scale and Numerical Location

We now prove Theorem 1.2 using the candidate construction of Algorithm 1. Algorithm 2 below restricts candidate means: if the label scale exceeds the attempted bound, every retained candidate already has small relative excess risk. Unknown moments distinguish proof cases, not algorithmic branches.

Let $\mu = \mathbb{E}Y$, $V = \operatorname{Var} Y$, $c_\mu = \mu_+$, and $B_0 = R(c_\mu) = V + \mu_-^2$. For a nonnegative mean ν and a real center z, write

$$D_z(\nu) = (\nu - z)^2 - z_-^2 = (\nu - z_+)^2 + 2z_-\nu. \tag{24}$$

This is the increase in squared mean loss over the best nonnegative constant. Gaussian Poincare gives $\operatorname{Var} f \le W^2$ for every feasible ReLU. Hence, with $\nu = \mathbb{E}f$,

$$R(f) = B_0 + \operatorname{Var} f - 2\operatorname{Cov}(Y, f) + D_\mu(\nu). \tag{25}$$

Two consequences are

$$N \ge (\sqrt{B_0} - W)_+^2, \qquad D_\mu(\mathbb{E}f_*) \le 2W\sqrt{V}. \tag{26}$$

The first follows by the triangle inequality; the second compares f_* with c_μ and uses $|\operatorname{Cov}(Y, f)| \le W\sqrt{V}$ (Appendix E.2).

For fixed $\nu \ge 0$, $z \mapsto \sqrt{D_z(\nu)}$ is 1-Lipschitz (Appendix E.2). Thus

$$\left|\sqrt{D_\mu(\nu)} - \sqrt{D_z(\nu)}\right| \le |z - \mu|. \tag{27}$$

Set $U_0 = W + 1$, $C_* = 512$, and

$$S = C_*^2 U_0^2/\eta^2, \qquad D_0 = 4\sqrt{U_0\sqrt{S}}, \qquad \gamma = e/S. \tag{28}$$

Estimate $\widehat{\mu}$ with $|\widehat{\mu} - \mu|^2 \le \gamma V + e$, put $\widehat{c} = (\widehat{\mu})_+$, and only retain candidates satisfying

$$D_{\widehat{\mu}}(\mathbb{E}f) \le D_0^2. \tag{29}$$

Always include $\widehat{c}$, for which $D_{\widehat{\mu}}(\widehat{c}) = 0$. Every retained candidate also satisfies $|\mathbb{E}f - \widehat{c}| \le D_0$. Appendix E.4 gives a conservative finite-precision test valid for every $\widehat{\mu}$.

Proof sketch of Theorem 1.2. If $B_0 \le S$, mean estimation has squared error at most $2e$ and the centered moment bound $2S$ holds. For $\widehat{c} \le 4\sqrt{S}$, apply the normalized construction with bound $32S$; otherwise use a centered first-moment direction and bounded bias offsets. Distribution preservation ensures that a useful nearby candidate survives the mean-loss filter.

If $B_0 > S$, (26) and the filter bound *every* retained candidate's risk by $(1 + \eta/4)N$, so resource caps remain valid without a variance test. Appendix E.1 gives both risk calculations; Appendices E.3–E.4 verify retention, precision, and resource caps.

Lemma F.1 supplies the mean estimate (55) and logarithmic query cost, completing Theorem 1.2; Appendix F gives the bounded-query implementation (Feldman, 2017).

The next bound shows that numerical location has a cost.

Proposition 5.1 (Numerical location cannot be free)**.** *For constant labels $Y = M \in [0, M_{\max}]$, an algorithm using at most Q bounded STAT queries, each of tolerance at least τ, and achieving squared risk at most ε with constant success probability, must satisfy*

$$Q \log(C/\tau) \ge \log \frac{M_{\max}}{C\sqrt{\varepsilon}}.$$

Proof sketch. Round oracle answers to a finite grid and count the distinguishable label intervals. Appendix F.1 gives the transcript-counting proof, including randomized algorithms. □

Algorithm 2 Proper SQ learner without a supplied label moment bound

Require: Bounded STAT access; d, W, η, ε; promise $\mathbb{E}Y^2 < \infty$.

1: Choose $e = \varepsilon/C_0$ and S, D_0, γ from (28).
2: Use Lemma F.1 with $p_0 = c\gamma$ to obtain a dyadic $\widehat{\mu}$; set $\widehat{c} = (\widehat{\mu})_+$ and initialize the candidate set to $\{\widehat{c}\}$.
3: **if** $\widehat{c} \le 4\sqrt{S}$ **then**
4: Run the candidate-generation part of Algorithm 1, after rescaling with the attempted bound $\mathbb{E}Y^2 \le 32S$; do not select a candidate yet.
5: **else**
6: With centered labels $Y - \widehat{c}$ and attempted second-moment bound $2S$, estimate the first-moment axis and perform the partial-projection and spectral construction.
7: Enumerate feasible weights in the resulting subspace and biases $b = \widehat{c} + s$ with $s \in [-D_0 - W - 1, D_0 + W + 1]$.
8: **end if**
9: Retain only candidates that pass the finite test (54); always keep $\widehat{c}$.
10: Using the attempted bound $2S$, compare $\mathbb{E}h^2 - 2\mathbb{E}[(Y - \widehat{c})h]$ for $h = f - \widehat{c}$ to accuracy ce. Return a minimizing retained ReLU f.

After mean estimation, cap every subroutine using only S, D_0, d, W, η, e. Discard infeasible or over-budget branches; if comparison exceeds its cap, return any retained candidate. Appendix E.4 proves that the fallback is harmless when the attempted bounds are false and unnecessary when they are valid.

6 Conclusion

We establish a proper SQ PTAS for agnostic ReLU regression with arbitrary bias under Gaussian inputs. By combining relative-error dimension reduction with a distribution-preserving ReLU replacement, the learner achieves risk at most $(1+\eta)\mathrm{OPT}_W + \varepsilon$ within the original norm-constrained class. Thus, it sharpens the approximation factor from a constant to arbitrarily close to one while retaining properness. Moreover, the extension to unknown label scale removes the need for a supplied moment bound without introducing polynomial dependence on the label moments.

Although the algorithm is polynomial for fixed η, its accuracy-dependent exponents limit scalability. It enumerates $d^{O(k)}$ Hermite features and searches a parameter net in $O(\eta^{-2})$ dimensions, with $k = O(\eta^{-2})$ in general. Corollary 4.2 lowers k under additional promises, but leaves the search dimension unchanged. Hence, further progress calls for sharper approximation or spectral bounds and a way to avoid exhaustive search while retaining the risk guarantee. These improvements remain subject to known barriers: subexponential-LWE hardness excludes an FPTAS, and parameterized sparse-LPN hardness excludes an EPTAS (Appendix G). Within these limits, the best dependence on η in the PTAS regime remains unresolved.

Beyond computational efficiency, extending the result to non-Gaussian inputs requires new structural tools. Our partial Hermite approximation relies on independence between orthogonal Gaussian coordinates, whereas the proper replacement uses rotation invariance to preserve the comparator's output distribution. Since general isotropic inputs need not satisfy either property, the projection and risk arguments require substitutes for both steps. Moreover, even constant-factor approximation is SQ-hard without a marginal assumption (Diakonikolas et al., 2022a). The next step is therefore to identify weaker, explicit input assumptions that support these substitutes and allow relative approximation with a proper output.

AI USE STATEMENT

OpenAI Codex was used to retrieve and compare literature, develop and revise mathematical claims and proofs, write and run local numerical checks, and draft and typeset this manuscript. The same AI-assisted workflow checked identities, source claims, and selected numerical cases. ChatGPT was additionally used to reorganize the exposition. The authors are responsible for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices A–F give the analytic estimates, query tolerances, finite grids, and numerical precision conventions. In particular, Appendices D.2, D.7, and E provide the detailed arguments for the proper replacement and the two main theorems.

REFERENCES

Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Agnostic learning of general ReLU activation using gradient descent. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.02711.

Stéphane Boucheron and Maud Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12, 2012. doi: 10.1214/ECP.v17-2210. URL https://arxiv.org/abs/1207.7209.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017. URL https://theory.stanford.edu/~valiant/papers/csv_stoc17.pdf.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://doi.org/10.1137/130916138.

Daniel Dadush. A deterministic polynomial space construction for ϵ-nets under any norm. *arXiv preprint arXiv:1311.6671*, 2013. URL https://arxiv.org/abs/1311.6671.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018a. URL https://arxiv.org/abs/1711.07211.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1061–1073. ACM, 2018b. URL https://arxiv.org/abs/1707.01242.

Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R. Klivans, and Mahdi Soltanolkotabi. Approximation schemes for ReLU regression. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1452–1485. PMLR, 2020a. URL https://proceedings.mlr.press/v125/diakonikolas20b.html.

Ilias Diakonikolas, Daniel M. Kane, and Nikos Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in Neural Information Processing Systems*, volume 33, 2020b. URL https://papers.nips.cc/paper/2020/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under Gaussian marginals. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1522–1551. PMLR, 2021a. URL https://proceedings.mlr.press/v134/diakonikolas21b.html.

Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1552–1584. PMLR, 2021b. URL https://proceedings.mlr.press/v134/diakonikolas21c.html.

Ilias Diakonikolas, Daniel M. Kane, Pasin Manurangsi, and Lisheng Ren. Hardness of learning a single neuron with adversarial label noise. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8199–8213. PMLR, 2022a. URL https://proceedings.mlr.press/v151/diakonikolas22a.html.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning a single neuron with adversarial label noise via gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4313–4361. PMLR, 2022b. URL https://proceedings.mlr.press/v178/diakonikolas22c.html.

Ilias Diakonikolas, Jelena Diakonikolas, Daniel M. Kane, Puqian Wang, and Nikos Zarifis. Near-optimal bounds for learning Gaussian halfspaces with random classification noise. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21793–21825. Curran Associates, Inc., 2023a. URL https://papers.neurips.cc/paper_files/paper/2023/hash/44c150733f9c5b6f98cb0caad0c664c7-Abstract-Conference.html.

Ilias Diakonikolas, Daniel M. Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and ReLU regression under Gaussian marginals. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7922–7938. PMLR, 2023b. URL https://proceedings.mlr.press/v202/diakonikolas23b.html.

Ilias Diakonikolas, Daniel M. Kane, and Mingchen Ma. Robust regression of general ReLUs with queries. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4911. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d7d0a3e9479e2cb252b040a1022d8425-Abstract-Conference.html.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Vitaly Feldman. Dealing with range anxiety in mean estimation via statistical queries. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pp. 629–640. PMLR, 2017. URL https://proceedings.mlr.press/v76/feldman17b.html.

Vitaly Feldman, Cristóbal Guzmán, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2017. URL https://arxiv.org/abs/1512.09170.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/3a37abdeefe1dab1b30f7c5c7e581b93-Abstract.html.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1004–1042. PMLR, 2017. URL https://proceedings.mlr.press/v65/goel17a.html.

Surbhi Goel, Sushrut Karmalkar, and Adam Klivans. Time/accuracy tradeoffs for learning a ReLU with respect to Gaussian marginals. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/067a26d87265ea39030f5bd82408ce7c-Abstract.html.

Surbhi Goel, Aravind Gollakota, and Adam Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/17257e81a344982579af1ae6415a7b8c-Abstract.html.

Anxin Guo and Aravindan Vijayaraghavan. Agnostic learning of arbitrary ReLU activation under Gaussian marginals. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2592–2631. PMLR, 2025. URL https://proceedings.mlr.press/v291/guo25a.html.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013. doi: 10.1017/CBO9781139020411. URL https://doi.org/10.1017/CBO9781139020411.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. doi: 10.1137/060649057. URL https://doi.org/10.1137/060649057.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45 (6):983–1006, 1998. doi: 10.1145/293347.293351. URL https://doi.org/10.1145/293347.293351.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. URL https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. URL https://arxiv.org/abs/2105.10386.

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning a single neuron via sharpness. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36541–36577. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23aq.html.

Puqian Wang, Nikos Zarifis, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning monotone single-index models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2508.04670.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning monotone generalized linear models via data augmentation. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5921–5990. PMLR, 2025. URL https://proceedings.mlr.press/v291/zarifis25a.html.

A GAUSSIAN IDENTITIES AND COMPARATOR REDUCTIONS

We use normalized probabilists' Hermite polynomials, so $\mathbb{E}H_\alpha(G)H_\beta(G) = \mathbf{1}_{\alpha=\beta}$ and $\partial_j H_\alpha = \sqrt{\alpha_j}H_{\alpha-e_j}$. For a Gaussian Sobolev function $f = \sum_\alpha f_\alpha H_\alpha$,

$$\mathbb{E}\|\nabla f\|^2 = \sum_\alpha |\alpha| f_\alpha^2. \tag{30}$$

All derivative statements about ReLU use its weak derivative. Gaussian Poincare (O'Donnell, 2014) follows by omitting the constant coefficient in this identity. Conditional versions follow by first fixing independent coordinates and then integrating.

A.1 EXISTENCE AND USEFUL SIGNAL

The risk is continuous on bounded parameter sets. To see this, the 1-Lipschitz property gives $\|f_{w,b} - f_{w',b'}\|_2 \le (\|w-w'\|^2+|b-b'|^2)^{1/2}$, and $Y \in L^2$. Uniformly over $\|w\| \le W$, $\|(w^TX+b)_+\|_2 \to 0$ as $b \to -\infty$. Thus if the risk infimum is below $B = \mathbb{E}Y^2$, a minimizing sequence cannot have bias diverging negatively. A bounded-risk sequence has $\|f\|_2 \le \sqrt{R(f)} + \sqrt{B}$, and $b_+ \le \mathbb{E}f \le \|f\|_2$, which bounds its positive biases. Compactness then gives an optimum. If the infimum equals B, zero attains it.

For an optimum f_*, sf_* is feasible for $0 \le s \le 1$. The derivative from the left at $s = 1$ of $B + s^2A - 2s\mathbb{E}Yf_*$ is nonpositive. Hence $\mathbb{E}Yf_* \ge A$, and Cauchy–Schwarz gives $A \le B$. If its weight norm is $a < W$, scaling is feasible on both sides of one, so

$$\mathbb{E}Yf_* = A, \qquad B = N + A. \tag{31}$$

For a nonconstant optimum, the bias is unconstrained. Since the Gaussian activation boundary has probability zero and the ReLU is 1-Lipschitz in its bias, differentiation in L^2 gives

$$\partial_b R(f_*) = -2\mathbb{E}[(Y - f_*)\mathbf{1}_{v^TX>t}] = 0.$$

Indeed, the bias difference quotient converges in L^2 to the activation indicator by dominated convergence; pair it with $Y - f_* \in L^2$. This proves the second identity in (5) even when $a = W$.

Suppose $B > (1 + \eta/4)N + e$ and let $G_0 = B - N$. Rearranging and using $G_0 = 2\mathbb{E}Yf_* - A \le 2\sqrt{BA}$ gives

$$G_0 > \frac{\eta B/4 + e}{1 + \eta/4}, \qquad A \ge \frac{G_0^2}{4B} \ge c\eta^2 B + c\eta e.$$

If $B = 0$, zero is already optimal. These facts prove (7) in the only case where it is needed. For a constant optimum, its value is $\mu_+ \le \sqrt{B}$ and it is included to the required accuracy in the constant grid.

A.2 ONE-DIMENSIONAL EXCESS MOMENTS

Let $Z \sim \mathcal{N}(0,1)$, $p = \overline{\Phi}(t)$, $\phi(t) = (2\pi)^{-1/2}e^{-t^2/2}$, and

$$A_t = \mathbb{E}(Z-t)_+^2 = (1+t^2)p - t\phi(t), \qquad \mu_t = \mathbb{E}(Z-t)_+ = \phi(t) - tp.$$

For $t \ge 0$, put $q_t = \mu_t/p$ and $s_t = A_t/p$. Direct substitution gives $s_t = 1 - tq_t$, while conditional Jensen gives $q_t^2 \le s_t$. Consequently

$$1 \le (t^2+2)s_t - s_t^2 \le (t^2+2)s_t, \qquad \frac{p}{t^2+2} \le A_t \le p.$$

Integration by parts also gives

$$p = A_t + t\mu_t, \qquad 2\mu_t = tA_t + \mathbb{E}(Z-t)_+^3 \ge tA_t.$$

Thus for $t \ge 1$, $\mu_t \ge tp/(t^2+2) \ge p/(3t)$, as used in the weighted mass bound.

For $t \ge 1$, the excess tail bound $\overline{\Phi}(t+x)/p \le e^{-tx}$ yields $\mu_t \le p/t$ and $A_t \le 2p/t^2 \le 1/t^2$. For $t \in [0,1]$, bounded-interval estimates combine with this to give $A_t \le Cp/(t+1)^2$. If $t \le 0$, the

degree-one Hermite coefficient of $(Z-t)_+$ is p, and hence $A_t \geq \mathrm{Var}((Z-t)_+) \geq p^2 \geq p/2$. In all cases,

$$a^2 p \leq C(1+t_+^2)A. \tag{32}$$

We will also use

$$\phi(t)^2 A_t \leq 2p^3. \tag{33}$$

The Gaussian hazard rate $h(t) = \phi(t)/\overline{\Phi}(t)$ is increasing (Boucheron & Thomas, 2012): indeed $h'(t) = h(t)\mu_t/p > 0$. Hence $\overline{\Phi}(t+x) \leq pe^{-h(t)x}$ for $x \geq 0$, and

$$A_t = 2\int_0^\infty x\overline{\Phi}(t+x)\,dx \leq \frac{2p}{h(t)^2},$$

which proves (33) for every real t.

For $t \geq 1$, (7) and $a \leq W$ imply $p \geq A/W^2 \geq c\eta e/W^2$, since $A_t \leq p$. The Gaussian Chernoff bound then gives $t^2 \leq C\log(C(W+1)/(\eta e))$. Also $a/t \geq \sqrt{A}$. These imply both the useful negative-bias interval in (23) and the existence of a dyadic $\tau \in [a/(64t), a/(32t)]$ in a known range $[c\sqrt{\eta e}, W]$.

B Weighted mean lists for rare activations

B.1 Constants in the weighted mean bound

Proof. Let $\mu_t = \mathbb{E}(Z-t)_+$ and $A_t = \mathbb{E}(Z-t)_+^2$. The scalar bounds $\mu_t \geq p/(3t)$ and $A_t \leq 2p/t^2$ are proved in Appendix A.2. Bias optimality in (5) gives

$$s_I \geq \mathbb{E}[Y\mathbf{1}_{v^\tau X > t}] - \tau p = a\mu_t - \tau p \geq ap/(4t).$$

Also $Y_\tau \leq Y^2/\tau$. Combining these inequalities with $A \leq 2a^2p/t^2$ proves (9). Gaussian independence gives the matrix identity

$$\mathbb{E}[(X-tv)(X-tv)^T\mathbf{1}_{v^\tau X>t}] = A_t vv^T + pP_{v^\perp} \preceq pI_d.$$

For any unit z, Cauchy–Schwarz therefore bounds $|\mathbb{E}[Yz^T(X-tv)\mathbf{1}_I]|$ by $\sqrt{Bp}$. Divide by s_I and use $B \leq A/\Delta \leq 2a^2p/(t^2\Delta)$ to obtain (10). The argument allows negative labels outside I and uses only their second moment. □

For the stated constants, the mass bounds give

$$\frac{s_I}{s} \geq \frac{ap}{4t}\frac{a}{64Bt} = \frac{a^2p}{256Bt^2} \geq \frac{A}{512B} \geq \frac{\Delta}{512}.$$

The dimension-free Cauchy–Schwarz estimate in that proof yields

$$\|\mu_I - tv\| \leq \frac{\sqrt{Bp}}{s_I} \leq \frac{4t\sqrt{B}}{a\sqrt{p}} \leq 4\sqrt{2/\Delta}.$$

The activation event is used only for analysis. The algorithm queries the weight Y_τ and never needs to identify I.

B.2 Deterministic coverage and enumeration

For $M = \mathbb{E}[\psi XX^T]/s$ and a rank-r orthogonal projection P_H, the key estimate is

$$\mathrm{tr}(P_H M) \leq 2r\log 2 + 8\log(1/s). \tag{34}$$

Write $M = \mathbb{E}[\psi XX^T]/s$ and $\mathbb{E}_\psi h = \mathbb{E}[\psi h]/s$. For a rank-$r$ orthogonal projection P_H, Cauchy–Schwarz gives $\mathbb{E}[\psi\exp(\|P_H X\|^2/8)] \leq 2^{r/4}$. Jensen under $\mathbb{E}_\psi$ yields

$$\tfrac{1}{8}\mathrm{tr}(P_H M) \leq \log \mathbb{E}_\psi \exp(\|P_H X\|^2/8) \leq (r/4)\log 2 + \log(1/s),$$

proving (34). For every vector z and every admissible event I, conditional Jensen also gives

$$(z^T\mu_I)^2 \le \frac{\mathbb{E}[\psi(z^TX)^2\mathbf{1}_I]}{\mathbb{E}[\psi\mathbf{1}_I]} \le \alpha^{-1}z^TMz.$$

Estimate M so that $M \preceq \widetilde{M} \preceq M + 3I$, and let H span eigenvectors with eigenvalues $\lambda_i > 64$. With $r = \dim H$ and $c_0 = 2\log 2 + 3$, (34) implies

$$\sum_{i\in H}(\lambda_i - c_0) \le 8\log(1/s), \qquad r \le \frac{8\log(1/s)}{64 - c_0}. \tag{35}$$

The mean bound gives $\mu_I\mu_I^T \preceq \widetilde{M}/\alpha$. Thus $\|P_{H^\perp}\mu_I\| \le 8/\sqrt{\alpha}$ and $P_H\mu_I$ belongs to $E_H = \{z : \sum_i z_i^2/\lambda_i \le 1/\alpha\}$. For $r > 0$, put $R_0 = 1/\sqrt{\alpha}$ and $q = \lceil\sqrt{r}\rceil$, and form

$$\mathcal{L} = \frac{5}{2q\sqrt{\alpha}}\mathbb{Z}^r \cap \frac{5}{4}E_H. \tag{36}$$

Rounding to the grid $(5R_0/(2q))\mathbb{Z}^r$ changes Euclidean norm by at most $5R_0/4$ and ellipsoidal norm by at most $5R_0/32$, since $\lambda_i > 64$. Thus a rounded point of E_H lies in $(37/32)E_H$ and is included in (36). The discarded component is orthogonal to the grid. Rounding the final output to error $R_0/4$ therefore leaves distance at most $(\sqrt{64 + 25/16} + 1/4)R_0 < 9R_0$ to every covered mean.

The grid cubes have disjoint interiors, volume $(5R_0/(2q))^r$, and ellipsoidal radius at most $5R_0/32$. Cubes centered in $(5/4)E_H$ therefore lie in $(45/32)E_H \subset (3/2)E_H$. If v_r denotes the Euclidean unit-ball volume, integration of $e^{-\|x\|^2/2}$ over the radius-$\sqrt{r}$ ball gives $v_r \le (2\pi\mathrm{e}/r)^{r/2}$. Consequently

$$|\mathcal{L}| \le (3q/5)^r v_r \prod_{i\in H}\sqrt{\lambda_i} \le \prod_{i\in H} 5\sqrt{\lambda_i}.$$

Here $(6/5)\sqrt{2\pi\mathrm{e}} < 5$. For $x \ge 64$, the gap $(x - c_0)/16 - \log(5\sqrt{x})$ is positive at 64 and has derivative $1/16 - 1/(2x) > 0$. Equation (35) therefore gives

$$|\mathcal{L}| \le \prod_{i\in H} 5\sqrt{\lambda_i} \le \exp\left[\frac{1}{16}\sum_{i\in H}(\lambda_i - c_0)\right] \le 1/\sqrt{s}. \tag{37}$$

For enumeration, use integer coordinates $k_i = 2qz_i/(5R_0)$. A prefix $(k_1, \dots, k_j)$ is kept precisely when

$$\sum_{i=1}^{j} k_i^2/\lambda_i \le q^2/4.$$

The next integers form an interval whose endpoints follow from this quadratic inequality. Every retained prefix extends by zeros, so the recursion has at most $(r+1)|\mathcal{L}|$ nodes. This gives polynomial enumeration without inspecting the surrounding box.

Use rational upper approximations of the eigenvalues, reserving enough matrix accuracy to preserve $M \preceq \widetilde{M} \preceq M + 3I$. Prefix comparisons are then exact rational operations. Approximate the final scale R_0 and orthogonal map to the rounding budget above; the strict gaps between the ellipsoid radii allow inverse-polynomial precision. For $r = 0$, use $\{0\}$. Coverage holds for all admissible I under every valid moment estimate.

B.3 Weighted moment estimates and list coverage

Both $s = \mathbb{E}\psi$ and $\mathbb{E}[\psi X_iX_j]$ admit the clipping estimate (39), with label ψ, second-moment bound one, and feature 1 or X_iX_j. The feature fourth moments are at most 105. Thus prescribed inverse-polynomial entry errors require only bounded queries with inverse-polynomial tolerance, even when the label fourth moment is infinite.

The unnormalized matrix has norm at most $\mathbb{E}[\psi\|X\|^2] \le \sqrt{d(d+2)} \le d + 1$. Numerator entry errors $cs_{\min}/d$ and denominator error $cs_{\min}^2/(d+1)$ therefore give constant operator error for $s \ge s_{\min}$. Retain thresholds whose estimated mass is at least $s_{\min}/2$, also requiring mass error at most

$s_{\min}/8$. Every useful threshold survives, and every retained mass is at least a constant fraction of $s_{\min}$. Adjust the list-size cap and moment-error constants to this lower bound. Label thresholds themselves are exact dyadic rationals, so atoms create no additional error.

For a useful label threshold and $\psi = Y_\tau$, (9) and (7) give

$$s \geq s_I \geq \frac{ap}{4t} \geq \frac{At}{8a} \geq \frac{c\eta e}{W+1} =: s_{\min}. \tag{38}$$

The dyadic threshold range in Appendix A.2 has logarithmic size. With $\alpha = \Delta/512$, a useful grid point obeys

$$\|z - tv\| \leq C/\sqrt{\Delta} + C/\sqrt{\alpha} \leq K_\Delta = C/\sqrt{\Delta}.$$

For $t \geq 2K_\Delta$, its normalized axis has

$$J \leq 4K_\Delta^2 a^2 p/t^2 \leq CK_\Delta^2 A.$$

For smaller t, (32) bounds the total gradient energy by $CK_\Delta^2 A$. If $A \leq n$, enumerate dyadic Δ from $1/4$ down to the first value below $c\eta^2$, with c chosen below the absolute constant in (7); call this set $\mathcal{D}_\eta$. Since $B \leq (\sqrt{A} + \sqrt{N})^2 \leq 4n$, one choice has $\Delta \leq A/B$ and $A \leq 8\Delta n$. This enumeration requires no knowledge of A, B, or N. For that choice,

$$J \leq C\Delta^{-1} A \leq Cn,$$

proving (8). A zero grid point is replaced by an arbitrary axis; the useful large-t point is nonzero. Include these axes and $L = \{0\}$ in the subspace list.

C A USEFUL PLANE AND BOUNDED QUERIES

This section gives the remaining estimates and finite parameters for the plane in Section 3.3.

C.1 THE FIRST-MOMENT DIRECTION

Gaussian integration by parts and Bessel's inequality give

$$q_1 = \mathbb{E}[YX] = apv + r_1, \qquad \|r_1\| = \|\mathbb{E}[(Y - f_*)X]\| \leq \sqrt{N}.$$

If $\|\widehat{q}_1 - q_1\| \leq \sqrt{e}$ and $u = \widehat{q}_1/\|\widehat{q}_1\|$, then

$$ap\|P_{u^\perp} v\| \leq \sqrt{N} + \sqrt{e}, \qquad J \leq 2(N+e)/p.$$

If $\widehat{q}_1 = 0$, the same bound holds for any axis. Thus $J \leq C_{T_0} n$ for any fixed upper bound $t \leq T_0$.

C.2 ONE VECTOR IDENTIFIES A USEFUL PLANE

Let $K \geq 1$ be fixed, $t \geq 4K$, and $\angle(u, v) \leq K/t$. Choose $|T - t| \leq 1/(t+1)$ and put $p_T = \overline{\Phi}(T)$. Integration by parts in any direction perpendicular to u gives

$$g(u) = aq(u)P_{u^\perp} v + r(u), \qquad q(u) = \mathbb{P}(u^T X > T, v^T X > t).$$

For every unit $z \perp u$, independence of $z^T X$ and $u^T X$ gives

$$|z^T r(u)| = |\mathbb{E}[\xi z^T X \mathbf{1}_{u^T X > T}]| \leq \sqrt{N}\sqrt{\mathbb{E}[(z^T X)^2 \mathbf{1}_{u^T X > T}]} = \sqrt{N p_T}.$$

Taking the supremum proves the vector error bound without a dimension factor.

For the joint-tail bound, write $Z = u^T X$ and $v^T X = \cos\theta\, Z + \sin\theta\, H_0$, with independent standard normals Z, H_0. The event

$$Z > \max\{T, t/\cos\theta\}, \qquad H_0 \geq 0$$

implies joint activation. Since $\theta \leq 1/4$, its first threshold is at most $T + C(K^2+1)/t$. The Gaussian hazard bound $\phi(x)/\overline{\Phi}(x) \leq x + 1$ for $x \geq 0$ gives $q(u) \geq c_K p_T$. It also gives $p_T \asymp p$.

Substitution in (14) proves $J_H \leq C_K n$, including $\dim H = 1$.

C.3 Finite thresholds and stable coordinates

When $A > n$, the $\Delta = 1/4$ list gives $\|z - tv\| \le K_*$ for an absolute K_*. For $t \ge 16K_*$ its axis obeys $\angle(z, v) \le 4K_*/t$, so the preceding proof applies with $K = 4K_*$. The first-moment direction handles $t \le 16K_*$. In both cases $J \le Cn$, proving (8).

Appendix A.2 gives the useful-range bounds

$$p \ge p_{\min} = c\eta e/(W+1)^2, \qquad t \le T_{\max} = \sqrt{C \log(C(W+1)/(\eta e))}.$$

Use a dyadic mesh on $[1, T_{\max} + 1]$ with spacing at most $1/(10(T_{\max} + 1))$. A threshold satisfies $|T - t| \le 1/(t+1)$. Discard guesses with $p_T < cp_{\min}$; the useful guess survives for small enough c.

For every listed axis and threshold, estimate the coordinates of $\mathbb{E}[YX\mathbf{1}_{u^TX>T}]$ to error $c\sqrt{ep_T/d}$ and project onto $u^\perp$. Each feature $X_i\mathbf{1}_{u^TX>T}$ has fourth moment at most three, so the clipping bound below gives inverse-polynomial STAT tolerance.

If $\|\widehat{g}\| \le \sqrt{ep_T}$, use $L = \mathrm{span}(u)$: the vector identity gives $aq(u)\|P_{u^\perp}v\| \le (\sqrt{N} + 2\sqrt{e})\sqrt{p_T}$ and hence $J_L \le C_K n$. Otherwise $(u, \widehat{g}/\|\widehat{g}\|)$ is an orthonormal frame for $L = H$. Every normalization divides by at least an inverse-polynomial quantity. Appendix D gives sufficient accuracy for this frame. Each axis and threshold produce one subspace.

A direction perturbation ω changes a threshold event with probability $O(\sqrt{\omega})$, by splitting at distance $\sqrt{\omega}$ from its boundary. Cauchy–Schwarz twice bounds the resulting coordinate-correlation error by $C\sqrt{B}\,\omega^{1/8}$. Inverse-polynomial precision therefore suffices for all coordinate operations, with smaller constants in the estimation budget. Guesses that are not useful may still generate candidates; their counts are bounded in advance.

C.4 One clipping bound for every correlation

Write $\mathrm{clip}_T(z) = \max\{-T, \min\{z, T\}\}$. For any $r > 1$,

$$|\mathbb{E}Z - \mathbb{E}\,\mathrm{clip}_T(Z)| \le \mathbb{E}|Z|^r / T^{r-1}.$$

Take $Z = YH$, $r = 4/3$, and apply Hölder. If $\mathbb{E}Y^2 \le B$ and $\mathbb{E}H^4 \le V_4$, then

$$|\mathbb{E}YH - \mathbb{E}\,\mathrm{clip}_T(YH)| \le (B^2V_4/T)^{1/3}. \tag{39}$$

Choose $T \ge 8B^2V_4/s^3$ and query $\mathrm{clip}_T(YH)/T$ to tolerance $s/(2T)$. The clipping bias and query error are each at most $s/2$. No independence or label fourth moment is required. This applies to all correlations in the algorithm; for Hermite coefficients, the higher-moment bound below permits smaller T.

D Finite projection, precision, and proper search

D.1 Partial Hermite approximation

For fixed $Z = z$, expand $f_*(z, G)$ in the transverse Hermite basis. Its squared gradient energy, averaged over Z, is $J = a^2\|P_{L^\perp}v\|^2 p$. Equation (30) gives

$$\|f_* - P^G_{\le k} f_*\|^2 \le J/(k+1), \qquad \mathbb{E}\|\nabla_G P^G_{\le k} f_*\|^2 \le J.$$

For the full-space projection the same formulas hold with total gradient energy a^2p.

Let $c_\alpha(z) = \mathbb{E}_G[f_*(z, G)H_\alpha(G)]$. The vector of coefficients for $|\alpha| \le k$ is W-Lipschitz as a map of z into Euclidean space, because orthogonal projection contracts the $L^2(G)$ difference and $a\|P_Lv\| \le W$. Replacing it by its conditional average in a box of side at most h adds squared error at most $r_0W^2h^2$, where $r_0 = \dim L \le 2$.

For any cell with an outer coordinate, compare to the zero coefficient vector. Since $b_+ \le 1$ in the normalized model,

$$f_*(X) \le a|v^TX| + 1.$$

Write $v^T X = \beta^T Z + \sqrt{1-\|\beta\|^2} G_1$, with β the coordinates of $P_L v$. Conditional integration and a union bound over at most two coordinates give

$$\mathbb{E}[f_*^2 \mathbf{1}_{\max_i |Z_i|>R}] \le C(W^2+1)(R^2+1)\overline{\Phi}(R).$$

The parameters (21), with sufficiently large C and small c, make the sum of histogram errors at most $c\eta^2 e$. Transverse truncation and histogram projection commute and their error components are orthogonal. Taking $k \ge C/\eta^2$ proves the first inequality of (15). The histogram averages over independent coordinates and contracts the transverse Sobolev energy, proving the second.

D.2 Proof of the proper replacement lemma

We prove Lemma 4.1 under the projection bounds (15). The construction preserves the comparator's weight norm and output distribution, while controlling its correlation with the labels.

Proof of Lemma 4.1. If all transverse directions are retained, take $f' = f_*$. Otherwise write $\xi = Y - f_*$ and $\zeta = \widehat{g} - Pf_*$, so $\|\zeta\| \le \sqrt{N} + \delta$. Choose a transverse unit vector u parallel to $P_{L^\perp} v$ when this vector is nonzero. Since $(Pf_*)_j = c_j(Z) H_j(u^T G)$, its transverse gradient is parallel to u. The following rank-one estimate is a direct form of low-rank approximation (Eckart & Young, 1936):

$$\sum_{i\ge 2} \lambda_i(M_g) \le \operatorname{tr}(P_{u^\perp} M_g) = \sum_{j=1}^{k} \frac{\mathbb{E}\|P_{u^\perp} \nabla_G \zeta_j\|^2}{j} \le \|\zeta\|^2 \le (\sqrt{N}+\delta)^2. \tag{40}$$

Here the eigenvalues are in decreasing order. The first inequality uses $u^T M_g u \le \lambda_1(M_g)$; the Hermite derivative identity bounds the sum by $\|\zeta\|^2$. Since the terms with $2 \le i \le m$ are at least λ, this proves (17); the trace bound is immediate.

Let U_- span the first $m-1$ eigenvectors, put $V_0 = L \oplus U_-$, and let e_0 be the last retained unit eigenvector. Set

$$v' = P_{V_0} v + \|P_{V_0^\perp} v\| e_0, \qquad f' = a(v'^T X - t)_+, \qquad h = f' - f_*. \tag{41}$$

Then $\|v'\| = 1$ and $P_L v' = P_L v$. If $v' = v$, there is nothing to prove. Otherwise set $r = (v - v')/\|v - v'\|$ and $S = I_d - 2rr^T$. This reflection fixes V_0, commutes with P, and satisfies $f_* \circ S = f'$. Subtract the common conditional mean given Z and use Gaussian Poincare and the triangle inequality to obtain

$$\|h\|^2 \le 4J, \qquad \|(I-P)h\|^2 \le 4\theta^2 n. \tag{42}$$

Move the reflection to the other factor in the inner product. For each j, $\widehat{g}_j \circ S - \widehat{g}_j$ has conditional mean zero given $P_{r^\perp} X$. Conditional Gaussian Poincare and the triangle inequality for its derivative give

$$\|\widehat{g}_j \circ S - \widehat{g}_j\|^2 \le 4\mathbb{E}(\partial_r \widehat{g}_j)^2.$$

Reflection preserves transverse degrees and the histogram cells. Orthogonality and weighted Cauchy–Schwarz now give

$$\begin{aligned} |\langle \widehat{g}, h\rangle| &= \left| \sum_{j=1}^{k} \langle \widehat{g}_j \circ S - \widehat{g}_j, (Pf_*)_j \rangle \right| \\ &\le 2 \left(\sum_{j=1}^{k} \frac{\mathbb{E}(\partial_r \widehat{g}_j)^2}{j} \right)^{1/2} \left(\sum_{j=1}^{k} j \|(Pf_*)_j\|^2 \right)^{1/2} \le 2\sqrt{\lambda J}. \end{aligned} \tag{43}$$

The first sum is $r^T M_g r \le \lambda$, since $r \perp U_-$. The second is $\mathbb{E}\|\nabla_G P f_*\|^2 \le J$ by (15).

Put $h_\perp = (I-P)h$. Reflection preserves both the full norm and the projection error of f_*. Hence

$$-2\langle (I-P)f_*, h_\perp \rangle = \|h_\perp\|^2,$$

which gives the exact risk decomposition

$$R(f') - N = -2\langle \widehat{g}, h\rangle - 2\langle g - \widehat{g}, h\rangle - 2\langle \xi, h_\perp\rangle + \|h_\perp\|^2. \tag{44}$$

Use (43), $\|g - \widehat{g}\| \le \delta$, $\|\xi\| = \sqrt{N}$, and (42) to obtain (18). □

D.3 A SHARPER HERMITE TAIL

Lemma D.1. *For $f_* = a(v^T X - t)_+$, put $\sigma = \|P_{L^\perp} v\|$, $p = \overline{\Phi}(t)$, $A = \mathbb{E} f_*^2$, and $J = a^2\sigma^2 p$. For every integer $k \geq 1$,*

$$\|f_* - P^G_{\leq k} f_*\|^2 \leq \min\left\{ \frac{J}{k+1}, \frac{Ca^2\sigma^3\phi(t)}{(k+1)^{3/2}} \right\} \leq \frac{CJ^{3/2}}{\sqrt{A}\,(k+1)^{3/2}}. \tag{45}$$

The last bound is interpreted as zero if $f_* = 0$.

Proof. The Sobolev identity gives the first bound. For the second, couple two copies of G with correlation $\rho = k/(k+1)$, keeping Z fixed. Their activation coordinates U, V have correlation $\cos\alpha = 1 - \sigma^2/(k+1)$, so $\alpha \leq C\sigma/\sqrt{k+1}$. Then

$$\|f_* - P^G_{\leq k} f_*\|^2 \leq \frac{a^2\sigma^2}{k+1} \mathbb{P}(\mathbf{1}_{U>t} \neq \mathbf{1}_{V>t}) \leq \frac{Ca^2\sigma^3\phi(t)}{(k+1)^{3/2}}. \tag{46}$$

For the first inequality, the squared difference of the two gradients multiplies the squared degree-j norm by $2j(1-\rho^{j-1}) \geq k+1$ for $j > k$. The second uses the Gaussian halfspace bound $\mathbb{P}(\mathbf{1}_{U>t} \neq \mathbf{1}_{V>t}) \leq (\alpha/\pi)e^{-t^2/2}$ (Diakonikolas et al., 2023a, Fact B.9); see also Diakonikolas et al. (2018b, Lemma 4.2). Indeed, at fixed radius the two active arcs differ on angular length at most 2α, and they agree below radius $|t|$, whose exceedance probability is $e^{-t^2/2}$. Finally, (33) gives $a^2\sigma^3\phi(t) \leq CJ^{3/2}/\sqrt{A}$. □

For a useful subspace, $J \leq Cn$. If $a < W$ and zero is insufficient, (31) gives $A = B - N > \eta N/4 + e \geq \eta n/4$. Hence

$$\|f_* - P^G_{\leq k} f_*\|^2 \leq \frac{Cn}{\sqrt{\eta}\,(k+1)^{3/2}},$$

which permits $k \geq C\eta^{-5/3}$. If $A \geq n$, the factor $\sqrt{\eta}$ is unnecessary. The same holds when $b_* \geq 0$, since $t \leq 0$ gives $a^2\sigma^3\phi(t) \leq \sqrt{2/\pi}\,J$. These two conditions permit $k \geq C\eta^{-4/3}$. In each case the squared tail is at most $c\eta^2 n$; adding the histogram error $c\eta^2 e$ proves (15).

D.4 FEATURE DIMENSION AND BOUNDED COEFFICIENT ESTIMATES

Partition each copy of $[-R, R]$ into equal intervals with widths between $h/2$ and h, adding one outer cell. Take their product across the r_0 coordinates. Then

$$N_Z = O(1 + R/h)^{r_0}, \qquad \pi_{\min} \geq \min\{(h/2)\phi(R), 2\overline{\Phi}(R)\}^{r_0}.$$

Thus $\pi_{\min}$ is inverse polynomial in $W, 1/e, 1/\eta$. The projection space has

$$D = N_Z \binom{d - r_0 + k}{k}$$

features. For $r_0 = 0$, set $N_Z = \pi_{\min} = 1$; when $r_0 = d$, the transverse factor is one.

The features are $\varphi_{j,\alpha}(Z, G) = \mathbf{1}_{Z \in C_j} H_\alpha(G)/\sqrt{\pi_j}$. Gaussian hypercontractivity (O'Donnell, 2014) and independence give, for every integer $q \geq 2$,

$$\mathbb{E}|\varphi_{j,\alpha}|^{2q} \leq (2q-1)^{kq} \pi_j^{1-q} =: V_{2q}.$$

These moments give the following inverse tolerance for estimating one coefficient to accuracy s:

$$\tau^{-1} \leq \frac{CB}{\pi_{\min} s^2} \left[C\left(1 + \frac{\log(2 + B/s^2)}{k}\right) \right]^k. \tag{47}$$

For $r = 2q/(q+1)$, Hölder and the clipping inequality above yield

$$|\mathbb{E}[Y\varphi] - \mathbb{E}\,\mathrm{clip}_T(Y\varphi)| \leq \left(\frac{B^q V_{2q}}{T^{q-1}} \right)^{1/(q+1)}.$$

Assume $B \geq 1$ and $0 < s \leq 1$, increasing a supplied bound if necessary. Choose

$$q = 1 + \left\lceil \frac{\log(2 + B/s^2)}{k} \right\rceil, \qquad T = \frac{CB}{\pi_j s}(Cq)^k, \qquad \tau = \frac{s}{2T}.$$

For a sufficiently large absolute C, the clipping bias is at most $s/2$: use $(B/s^2)^{1/(q-1)} \leq \mathrm{e}^k$ and $(2q-1)^{1/(q-1)} \leq 3$. Since $q \leq 2 + \log(2 + B/s^2)/k$, this choice proves (47). Rounding T upward to a dyadic number within a factor two preserves the bound. With $s = \delta/\sqrt{D}$, orthonormality gives $\|\widehat{g} - PY\| \leq \delta$. Centering labels and functions by a known constant is allowed because constants belong to the projection; transverse derivative matrices are unchanged.

The matrix (16) can be computed algebraically. Index histogram cells by q: if $\widehat{g} = \sum_{q,\alpha} c_{q,\alpha}\varphi_{q,\alpha}$, define $d^{(i)}_{q,\beta} = \sqrt{\beta_i + 1}\, c_{q,\beta+e_i}$. Then

$$(M_g)_{i\ell} = \sum_{q,\, |\beta| \leq k-1} \frac{d^{(i)}_{q,\beta} d^{(\ell)}_{q,\beta}}{|\beta| + 1}.$$

The denominator is the degree before differentiation. Thus the matrix requires only finite arithmetic on the estimated coefficients; no additional queries are needed.

D.5 Coordinate and spectral precision

The exact projected space contains all total-degree-at-most-k transverse polynomials. Removing an arbitrary subset of monomials would invalidate rotation commutation and is not part of this algorithm.

Here is a conservative continuity bound for numerical coordinates. Suppose orthogonal coordinate transforms differ in operator norm by $\omega < 1$. For a fixed histogram cell, its indicator can change only when one of at most two coordinates is close to an endpoint or its perturbation is large. Gaussian density boundedness and Chebyshev's inequality, splitting at distance $\sqrt{\omega}$ and taking a union bound, give

$$\mathbb{P}(\text{cell indicator changes}) \leq C\sqrt{\omega}.$$

Cauchy–Schwarz with Gaussian fourth moments controls the polynomial on this event. The degree-k Gaussian chaos is also continuous under small orthogonal rotations with L^2 perturbation at most $Ck\omega$ for a normalized feature. Consequently the squared L^2 change of one normalized histogram–Hermite feature is bounded by

$$C\pi_{\min}^{-1}\big(3^k\omega^{1/4} + k^2\omega^2\big). \tag{48}$$

For coefficient norm at most $\sqrt{B}$ and D features, choose

$$\omega \leq \left(\frac{c\delta^2\pi_{\min}}{BD(1 + 3^k + k^2)}\right)^4, \qquad W^2\omega^2 \leq e.$$

Then the function change is absorbed into the coefficient error budget, and a perturbed subspace $\widetilde{L}$ still obeys $J_{\widetilde{L}} \leq 2J_L + CW^2\omega^2$. For fixed η these are inverse-polynomial requirements. One may represent the ideal nearby orthogonal frame implicitly by a sequence of plane rotations and evaluate it to this precision.

Let $\overline{M}_g$ be symmetric with $\|\overline{M}_g - M_g\|_{\mathrm{op}} \leq \nu$, including the error of the eigensolver. When $m < d - r_0$, retain its m largest eigenvalue directions and let $\overline{U}_-$ span the first $m-1$. The variational characterization of eigenvalues (Horn & Johnson, 2013, Sections 4.2–4.3) gives, for every transverse unit $r \perp \overline{U}_-$,

$$r^T M_g r \leq \lambda_m(\overline{M}_g) + \nu \leq \lambda + 2\nu \leq \frac{(\sqrt{N} + \delta)^2}{m - 1} + 2\nu. \tag{49}$$

Here $\lambda_m(\overline{M}_g)$ denotes the mth eigenvalue in decreasing order; its eigenvector is the norm-preserving direction e_0. Thus the proof of Lemma 4.1 applies with $\lambda + 2\nu$ in place of λ. Taking $\nu = c\eta^2 e$ preserves the risk bound since $e \leq n$. No separation between adjacent eigenvalues is needed, and the retained dimension is always at most m. Gaussian probabilities and moments are evaluated to the corresponding inverse-polynomial absolute errors.

D.6 SEARCHING FEASIBLE PARAMETERS

The replacement (41) preserves a, b, the comparator's norm, and its one-dimensional distribution. Thus (23) applies to f'. In a subspace of dimension at most $m+2$, choose Euclidean weight accuracy $\rho_0 = c\eta\sqrt{e}$ and bias accuracy ρ_0. A lattice grid at spacing $\rho_0/\sqrt{m+2}$ followed by projection onto the radius-W ball is a finite feasible net with size $(1 + CW/\rho_0)^{O(m+2)}$; combine it with the bias grid. The net contains f_{net} with $\|f_{\mathrm{net}} - f'\| \le C\rho_0$.

The risk comparison in the proof of Theorem 1.1, with $s = c\eta$, shows that this net adds $O(\eta)N + O(e)$ after allocating a small fraction of the multiplicative budget to Lemma 4.1. A feasible candidate in the range (23) has fourth moment at most $C(W^4+1)$, since $f(X) \le W|v^T X| + 1$. Thus validation of $\mathbb{E}Yf$ uses (39). For $w \neq 0$, $\mathbb{E}f^2 = \|w\|^2 A_{-b/\|w\|}$; constants are evaluated directly. Choosing the smallest estimated risk difference loses at most twice its error. With all internal constants sufficiently small and $e = \varepsilon/C_0$, the guarantee follows.

Every listed subspace and projection is enumerated with a known polynomial cap, and each projection produces one search subspace of dimension at most $m + 2$. The largest dimension-dependent feature count is $d^{O(k)}$, and the parameter search depends on m but not on d or $\log(1/e)$ in its exponent. The weighted mean list adds only a fixed power of $1/s_{\min}$. This gives the asserted poly_η resource form.

D.7 COMPLETION OF THE NORMALIZED THEOREM

Completion of Theorem 1.1. If zero is sufficient or an optimum is constant, the candidate set already contains an adequate approximation. Otherwise Lemma 3.1 supplies a useful projection, and Lemma 4.1 gives $R(f') \le N + (\eta/8)(N + e)$ in its search subspace of dimension at most $m + 2$. A net point f_0 has $\|f_0 - f'\| \le c\eta\sqrt{e}$. The inequality

$$R(f_0) \le (1 + s)R(f') + (1 + 1/s)\|f_0 - f'\|^2, \qquad s = \eta/8,$$

gives $R(f_0) \le (1+\eta/2)N + Ce$. Risk comparison adds $O(e)$; a sufficiently small internal e proves (2). All these bounds hold for every valid adaptive sequence of query answers.

For fixed η, both k and m are constants. The feature count is $d^{O(k)}$ times the polynomial histogram size, and the feasible net has size $(1 + CW/(\eta\sqrt{e}))^{O(m+2)}$ times a polynomial bias-grid factor. All list sizes, clipping levels, inverse tolerances, and remaining grids are polynomial in the input bounds. This proves the resource claim. □

E REMOVING POLYNOMIAL DEPENDENCE ON LABEL SCALE

This appendix proves the reduction underlying Theorem 1.2, assuming the normalized construction and the mean estimator proved in Appendix F. Write $\mu = \mathbb{E}Y$, $V = \mathrm{Var}\, Y$, $c_\mu = \mu_+$, $\mu_- = \max\{-\mu, 0\}$, and $B_0 = V + \mu_-^2$.

Algorithm 2 in Section 5.2 gives the procedure. The unknown quantity B_0 separates proof cases, not algorithmic branches: the only branch uses the computed nonnegative center.

E.1 PROOF OF THE UNKNOWN-SCALE REDUCTION

If $B_0 \le S$, then $|\widehat{\mu} - \mu|^2 \le 2e$. Projection onto the nonnegative constants and (27) give

$$\begin{aligned} \mathbb{E}(Y - \widehat{c})^2 &\le B_0 + |\widehat{\mu} - \mu|^2 \le 2S, \\ \sqrt{D_{\widehat{\mu}}(\mathbb{E}f_*)} &\le \sqrt{2W\sqrt{S}} + \sqrt{2e} < D_0/2. \end{aligned} \tag{50}$$

Thus centered labels satisfy the attempted moment bound, and the optimum has slack for rotation and rounding.

If $B_0 > S$, every retained f obeys $D_\mu(\mathbb{E}f) \le 2D_0^2 + 2\gamma B_0 + 2e$ by (27). Equations (25)–(28) then give

$$\begin{aligned} R(f) &\le B_0 + W^2 + 2W\sqrt{B_0} + 2D_0^2 + 2\gamma B_0 + 2e \\ &\le (1 + \eta/4)N. \end{aligned} \tag{51}$$

Fixed resource caps therefore suffice: valid moment bounds preserve a good candidate, while false bounds still leave every retained output good. No variance test is needed.

Candidate generation branches on $\widehat{c}$, not B_0. When $B_0 \le S$ and $\widehat{c} \le 4\sqrt{S}$, the normalized construction applies with $\mathbb{E}Y^2 \le 32S$. Still assuming $B_0 \le S$, if $\widehat{c} > 4\sqrt{S}$, a nonconstant optimum has positive bias and $p \ge 1/2$. One centered correlation $\mathbb{E}[(Y - \widehat{c})X]$ gives $J = O(N + e)$; apply the proper reduction to centered functions and search bounded bias offsets $b - \widehat{c}$. The envelope (53) bounds their moments independently of $\widehat{c}$. Appendices E.3–E.4 prove that the filter retains a good candidate in either branch.

E.2 Scalar estimates

Gaussian Poincare gives $\operatorname{Var} f \le W^2$. Writing $\nu = \mathbb{E}f$ yields the risk decomposition (25) and

$$\|Y - f\| \ge \|Y - \nu\| - \|f - \nu\| \ge \sqrt{B_0} - W.$$

The nonnegative constant $c_\mu = \mu_+$ is feasible, so $R(f_*) \le B_0$. Dropping $\operatorname{Var} f_*$ from the decomposition and using $|\operatorname{Cov}(Y, f_*)| \le W\sqrt{V}$ proves the second inequality of (26).

For $\nu \ge 0$, the function in (24) satisfies

$$\sqrt{D_z(\nu)} = \begin{cases} |\nu - z|, & z \ge 0, \\ \sqrt{\nu^2 - 2z\nu}, & z < 0. \end{cases}$$

It is continuous at zero. On the negative half-line its derivative has absolute value $\nu/\sqrt{\nu^2 - 2z\nu} \le 1$ when $\nu > 0$; for $\nu = 0$ it is constant there. The positive half-line is immediate. This proves (27).

Projection onto the nonnegative constants gives, with $\widehat{c} = (\widehat{\mu})_+$,

$$R(\widehat{c}) \le B_0 + |\widehat{\mu} - \mu|^2. \tag{52}$$

Indeed, $(\widehat{\mu} - \mu)^2 - D_\mu((\widehat{\mu})_+) = ((\widehat{\mu})_- - \mu_-)^2 + 2(\widehat{\mu})_-\mu_+ \ge 0$.

For $B_0 \le S$, the last inequality in (50) follows from

$$\sqrt{2W\sqrt{S}} \le D_0/\sqrt{8}, \qquad \sqrt{2e} \le \sqrt{2}, \qquad D_0 \ge 4\sqrt{512}.$$

For $B_0 > S$, put $x = U_0/\sqrt{B_0} \le \eta/C_*$. The first line of (51), together with $\gamma \le \eta/C_*$, gives

$$\frac{R(f)}{N} \le \frac{1 + 36\eta/C_* + 3\eta^2/C_*^2}{(1 - \eta/C_*)^2} \le 1 + \eta/4.$$

This proves the second line for $C_* = 512$ and $0 < \eta \le 1$, uniformly over all retained candidates, even when the moment assumptions used to generate them are false.

E.3 Candidate generation in the small-B_0 analysis

Candidate generation branches on the observable center $\widehat{c}$. Suppose first that $\widehat{c} \le 4\sqrt{S}$. If $\mu < 0$ then $\mu^2 \le B_0 \le S$; otherwise $\mu \le \widehat{c} + \sqrt{2e} \le 5\sqrt{S}$. Thus $\mathbb{E}Y^2 \le 32S$. Generate all candidates of the normalized construction with this known bound, but filter their means before validation.

The constant $\widehat{c}$ is always retained. If the best nonnegative constant already reaches the tighter target, (52) gives the desired bound after allocating $O(e)$ slack. Otherwise zero also fails that target, so the useful-signal analysis applies. In that case generate all projections unconditionally over the useful parameter ranges; the good replacement preserves a, b and hence ν_*. It satisfies $D_{\widehat{\mu}}(\nu_*) < D_0^2/4$, and the net changes its mean by at most its L^2 error. The finite filter below therefore retains the good candidate, even when zero is outside the retained interval.

Suppose instead that $\widehat{c} > 4\sqrt{S}$. A constant optimum is handled by $\widehat{c}$. For a nonconstant optimum,

$$\nu_* \ge \widehat{c} - \sqrt{2e} - \sqrt{2W\sqrt{S}} > W.$$

If its bias were nonpositive, its mean would be at most $W/\sqrt{2\pi}$, a contradiction. Hence its bias is positive and $p \geq 1/2$. Estimate

$$q_1 = \mathbb{E}[(Y - \widehat{c})X] = apv + r_1, \qquad \|r_1\| \leq \sqrt{N},$$

to error $c\sqrt{e}$, using the known centered second moment $2S$. Its direction gives $J \leq C(N + e)$ as in Appendix C. The positive-bias branch therefore needs only the first-moment direction.

For any feasible ReLU and its own unit weight direction v, averaging the Lipschitz difference against an independent standard normal Z' gives

$$|f(X) - \mathbb{E}f| \leq W\mathbb{E}[|v^T X - Z'| \mid X] \leq W(|v^T X| + \sqrt{2/\pi}). \tag{53}$$

Constants obey the same bound with zero left side. Apply the partial projection to $\widetilde{Y} = Y - \widehat{c}$ and $\widetilde{f}_* = f_* - \widehat{c}$. Its envelope is

$$|\widetilde{f}_*(X)| \leq W(|v^T X| + \sqrt{2/\pi}) + D_0.$$

The histogram proof uses this polynomial bound in place of $W|v^T X| + 1$; its radius and minimum mass remain polynomially controlled. Constant subtraction changes only degree zero, leaving (16), its rank-one estimate, J, and the projection error unchanged. The replacement preserves $\mathbb{E}f$ and $\mathbb{E}f^2$, hence also $\mathbb{E}(f - \widehat{c})^2$. Thus Lemma 4.1 applies to the centered functions with the common parameters (19).

Search weights in the same norm ball and write the bias as $b = \widehat{c} + s$. Positive-bias ReLU satisfies $0 \leq \mathbb{E}f - b \leq W/\sqrt{2\pi}$, so the target and its replacement have $s \in [-D_0 - W, D_0]$. A grid over $[-D_0 - W - 1, D_0 + W + 1]$ suffices. Its resolution depends on W, η, e and not on $\widehat{c}$. Centered prediction should be represented by the identity

$$(\widehat{c} + s + w^T X)_+ - \widehat{c} = \max\{s + w^T X, -\widehat{c}\},$$

which avoids unnecessary subtraction of two large, nearly equal numbers.

E.4 Finite filtering and validation

Always retain the constant $\widehat{c}$. For any other candidate, estimate $r = \mathbb{E}f - \widehat{c}$ by a dyadic $\widehat{r}$ with $|\widehat{r} - r| \leq \kappa$, choosing dyadic $\kappa \in [c\eta\sqrt{e}/2, c\eta\sqrt{e}]$. Use the test

$$\overline{r} = |\widehat{r}| + \kappa, \qquad \overline{r}(\overline{r} + 2(\widehat{\mu})_-) \leq T, \qquad D_0^2/2 \leq T \leq D_0^2, \tag{54}$$

where T is dyadic. The mean estimator can also return dyadic $\widehat{\mu}$, so the comparison is exact. Since $(\widehat{\mu})_- \widehat{c} = 0$ and $r \geq 0$ whenever $(\widehat{\mu})_- > 0$,

$$D_{\widehat{\mu}}(\mathbb{E}f) = |r|(|r| + 2(\widehat{\mu})_-) \leq \overline{r}(\overline{r} + 2(\widehat{\mu})_-).$$

Thus every accepted candidate satisfies (29), for any sign or size of $\widehat{\mu}$.

If $B_0 \leq S$, then $(\widehat{\mu})_- \leq \mu_- + \sqrt{2e} \leq 2\sqrt{S}$. The good replacement has $|r_*| < D_0/2$ and $D_{\widehat{\mu}}(\nu_*) < D_0^2/4$. A net of L^2 accuracy $\rho \leq c\eta\sqrt{e}$ and the mean estimate give $\overline{r} \leq |r_*| + q$, where $q = \rho + 2\kappa$. Hence

$$\overline{r}(\overline{r} + 2(\widehat{\mu})_-) \leq D_{\widehat{\mu}}(\nu_*) + (D_0 + 4\sqrt{S})q + q^2 < D_0^2/2 \leq T$$

for small absolute c. The good net point therefore survives the same test.

The required accuracy κ does not depend on $\widehat{\mu}$. In the bounded-center branch the parameter grid bounds $|b|$. In the other branch, $|\max\{s + w^T X, -\widehat{c}\}| \leq |s + w^T X|$ and s is bounded. A truncated one-dimensional Gaussian integral therefore computes $\widehat{r}$ with polynomial cost in the known parameters, even for large $\widehat{c}$.

For each retained candidate set $h = f - \widehat{c}$. The envelope gives $\mathbb{E}h^4 \leq C(W + D_0)^4$. Validate $Q(h) = \mathbb{E}h^2 - 2\mathbb{E}[(Y - \widehat{c})h]$ to accuracy ce using (39) and the known centered moment bound $2S$ in the small-B_0 analysis. Gaussian norm terms are evaluated in centered form as well. The good branch then gives the required risk.

All subroutines have fixed resource caps computed from S, D_0, d, W, η, e. Discard any branch that exceeds its cap or returns infeasible parameters. If the actual $B_0 > S$, the moment promises supplied

to these subroutines need not be true, but the cap still guarantees termination. The constant candidate remains, and every possible retained output is already good by (51). This justifies running the same algorithm without an SQ test for the unknown variance. The only remaining moment-dependent work is scalar mean location.

The conditional degrees also apply. When $B_0 \le S$, rescaling in the bounded-center branch preserves each condition in Theorem 1.1, using an internal additive accuracy no larger than the requested one. In the other branch, $b_* > 0$ and centering changes only degree zero, so $k = O(\eta^{-4/3})$ suffices, and therefore so does $O(\eta^{-5/3})$. When $B_0 > S$, the bound for every retained output is independent of the chosen degree.

F A BOUNDED-QUERY MEAN ESTIMATOR

Approximate medians and truncation are standard in SQ mean estimation (Feldman, 2017). Here the search requires no supplied moment bound.

Lemma F.1. *For $\mathbb{E}Y^2 < \infty$ and $0 < p_0 < 1/16$, bounded STAT queries produce $\widehat{\mu}$ satisfying*

$$|\widehat{\mu} - \mu|^2 \le Cp_0V + Cp_0^2, \tag{55}$$

using inverse tolerance $O(1/p_0)$ and

$$O\left(1 + \log(2 + |\mu| + \sqrt{V/p_0})\right)$$

queries and arithmetic operations, without a supplied variance bound.

Proof. Let $F(y) = \mathbb{P}(Y \le y)$. Query CDF values to tolerance $1/32$. Double Q, starting at one, until the returned values at $-Q, Q$ are below $1/4$ and above $3/4$, respectively. Bisect this interval at level $1/2$ until its width is at most one: a returned midpoint value below $1/2$ replaces the left endpoint; otherwise it replaces the right endpoint. The final endpoints satisfy

$$F(\text{left}) \le 17/32, \qquad F(\text{right}) \ge 15/32.$$

Chebyshev's inequality gives $Q \le C(1 + |\mu| + \sqrt{V})$ at termination and bounds the midpoint z_0 by $|z_0 - \mu| \le 2\sqrt{V} + 1$.

Now double R, starting at one, until the query for $\mathbf{1}_{|Y-z_0|>R}$, with tolerance $p_0/4$, returns at most $3p_0/4$. At termination,

$$\mathbb{P}(|Y - z_0| > R) \le p_0, \qquad R \le C(1 + \sqrt{V/p_0}). \tag{56}$$

Indeed, Chebyshev guarantees a successful query once $R \ge |z_0 - \mu| + \sqrt{2V/p_0}$. These statements also hold for atoms and $V = 0$.

Set $Z_R = \text{clip}_{[-R,R]}(Y - z_0)$. Cauchy–Schwarz gives

$$|z_0 + \mathbb{E}Z_R - \mu| \le \mathbb{E}[|Y - z_0|\mathbf{1}_{|Y-z_0|>R}] \le \sqrt{p_0V} + p_0|z_0 - \mu|.$$

Query $Z_R/R \in [-1, 1]$ to tolerance p_0, rescale, and add z_0. The query error is at most p_0R. Round the answer to a dyadic grid of spacing at most p_0. Then

$$|\widehat{\mu} - \mu| \le \sqrt{p_0V} + p_0(|z_0 - \mu| + R + 1) \le C(\sqrt{p_0V} + p_0),$$

which proves the error bound. Doubling and bisection take $O(1 + \log Q + \log R)$ queries and arithmetic operations. Their thresholds are exact dyadic rationals; the final rounding adds $O(1 + \log(1/p_0))$ fractional bits, whose cost is counted separately. □

Use $p_0 = c\gamma = ce/S$ in Lemma F.1. Since $S \ge 1$ and $e \le 1$, its bound implies $|\widehat{\mu} - \mu|^2 \le \gamma V + e$ for small c. Its query count is $O(\ell) + \text{poly}_\eta(W, 1/e)$ and inverse tolerance is polynomial in $W, 1/e$ for fixed η. All subsequent multi-dimensional queries in Appendix E use scale-independent precision bounds. This proves the resource form of Theorem 1.2 within the stated arithmetic convention.

F.1 Numerical location lower bound

Proof of Proposition 5.1. A valid oracle may round answers to a grid of spacing τ. There are at most $(C/\tau)^Q$ answer transcripts. A fixed output covers an interval of constant labels of length at most $2\sqrt{\varepsilon}$, since Jensen implies $|\mathbb{E}f - M| \le \sqrt{\varepsilon}$. For a randomized algorithm, integrate this coverage bound over a uniform M and then average over its random seed. □

G Consequences of existing hardness results

This appendix makes explicit the implications used in Section 1.3. The reductions below concern $W = 1$ and labels of second moment at most one, so the unknown-scale extension is not needed. They are consequences of existing hardness constructions, not new hardness assumptions established by this paper.

G.1 From additive-error hardness to exclusion of an FPTAS

Let $\delta \in (0, 1)$. Since zero is feasible, $N \le 1$. A learner satisfying (2), run with $\eta = \varepsilon = \delta/2$, obeys

$$R(\widehat{f}) \le N + \frac{\delta}{2}N + \frac{\delta}{2} \le N + \delta.$$

Thus an FPTAS for (2) would be a $\mathrm{poly}(d, 1/\delta)$ additive-error learner. For a bounded STAT implementation, a query of tolerance τ can be simulated with fresh samples by Hoeffding's inequality. A union bound over the capped number of adaptive queries gives polynomial sampling overhead in the query count and inverse tolerance; efficient query evaluation is part of our resource convention.

Under their subexponential-LWE assumption, Diakonikolas et al. (2023b, Theorem 1.4) give, for any constants $c > 0$ and $0 < \alpha < 1/2$ and accuracy $\delta \le (\log d)^{-(2+c)}$, a time lower bound of

$$\min\left\{ d^{\Omega((\delta \log^2 d)^{-\alpha})},\ 2^{d^{0.99}} \right\}.$$

For example, taking $\delta = (\log d)^{-3}$ and $0 < c < 1$ contradicts $\mathrm{poly}(d, 1/\delta)$ time. Their construction has labels in $\{-1, 1\}$ and a comparator of weight norm at most one. The null case has independent balanced labels, so allowing an arbitrary output hypothesis does not remove the obstruction. This excludes an FPTAS for our model. Because the stated accuracy range depends on d, this argument alone does not exclude an EPTAS with a completely unrestricted prefactor $F(1/\eta)$.

G.2 The EPTAS consequence of sparse-parity hardness

For precision, the parameterized sparse-LPN assumption used here is the following: at a fixed noise rate $\nu \in (0, 1/2)$, there is no algorithm with running time $H(k)d^C$, for any computable function H and constant C independent of k, that recovers the support of a k-sparse parity from uniform Boolean examples whose labels are independently flipped with probability ν. This is the parameterized obstruction associated with the $d^{\Omega(k)}$ sparse-noisy-parity conjecture in Goel et al. (2019, Assumption 1). We state it explicitly to specify the uniformity needed when the prefactor is arbitrary.

We adapt their Gaussian-lift construction, rather than applying their proper homogeneous-output theorem to a different output class. Let U be uniform on $\{-1, 1\}^d$, let S be the hidden parity support, and let $Y \in \{-1, 1\}$ be its noisy label. Set $X_i = U_i|G_i|$, with independent standard Gaussian G_i. Then $X \sim \mathcal{N}(0, I_d)$ and

$$\mathbb{E}[Y \mid X] = (1 - 2\nu) \prod_{i \in S} \mathrm{sign}(X_i).$$

It suffices to consider $k = |S| \equiv 2 \pmod 4$: any other sparsity can be padded by at most three fresh uniform bits, multiplying the label by their product and adding these known coordinates to the support.

The correlation estimate in Goel et al. (2019, Lemma 1), after constant normalization of its weight vector, supplies a unit vector v supported on S and a known lower bound $0 < \rho_k = 2^{-O(k)} \le 1$

such that

$$r(X) = (v^T X)_+, \qquad \mathbb{E}[Yr(X)] \geq \rho_k, \qquad \mathbb{E}r(X)^2 = \tfrac{1}{2}.$$

The factor $1 - 2\nu$ is absorbed into ρ_k. The feasible homogeneous ReLU $f = \rho_k r$ therefore satisfies

$$R(f) = 1 + \tfrac{1}{2}\rho_k^2 - 2\rho_k \mathbb{E}[Yr] \leq 1 - \tfrac{3}{2}\rho_k^2 \leq 1 - g_k, \qquad g_k = \rho_k^2.$$

It is also a feasible comparator in the arbitrary-bias class. If one drops a coordinate $j \in S$, the label becomes a balanced sign independent of all remaining coordinates. In that case *every* hypothesis h has risk $1 + \mathbb{E}h^2 \geq 1$. If $j \notin S$, the correlated comparator still exists and the optimum remains at most $1 - g_k$.

Run a hypothetical EPTAS on each of these d coordinate-deleted distributions with $\eta = \varepsilon = g_k/4$. In the latter case its output has risk at most

$$(1 + g_k/4)(1 - g_k) + g_k/4 \leq 1 - g_k/2.$$

Clipping the output to $[-1, 1]$ for validation cannot increase squared loss against $Y \in \{-1, 1\}$ and bounds that loss by four. Fresh validation samples distinguish risk at most $1 - g_k/2$ from risk at least one to accuracy $g_k/8$, using $O(g_k^{-2} \log d)$ samples per candidate. Repetition and validation amplify success to recover the entire support with constant probability.

An EPTAS costing $F(1/\eta)\,\mathrm{poly}(d, 1/\varepsilon)$ would make the total cost $H(k)\,\mathrm{poly}(d)$, since $g_k^{-1} = 2^{O(k)}$. This contradicts the stated parameterized assumption. The argument allows improper outputs and therefore covers the proper affine output of this paper. If all bounded-SQ resources had the EPTAS form, the sample simulation in Appendix G.1 would give the same contradiction. This establishes a conditional EPTAS barrier; it neither identifies the optimal η exponent nor equates sparse-LPN hardness with the LWE assumption.

G.3 The unconditional SQ barrier

For completeness, the moment-matching construction of Diakonikolas et al. (2020b, Proposition 4.1 and Section 4) also gives a direct normalized version of the SQ obstruction. For each sufficiently large integer s, it provides a Boolean function φ_s whose first s Gaussian moments vanish and whose correlation with $r(Z) = Z_+$ is at least $\rho_s = s^{-O(1)} > 0$. Decrease this known lower bound if necessary so that $\rho_s \leq 1$. For labels $Y = \varphi_s(v^T X)$ in a hidden unit direction, the feasible comparator $f = \rho_s(v^T X)_+$ has risk at most $1 - \rho_s^2$, by the calculation in Appendix G.2.

The statistical-dimension argument in that work makes this family require $d^{\Omega(s)}$ bounded queries at tolerance $d^{-\Omega(s)}$ to distinguish it from independent balanced labels, for sufficiently large d relative to s. A learner with $\eta = \varepsilon = \rho_s^2/4$ would output risk at most $1 - \rho_s^2/2$ on the family; under independent balanced labels every hypothesis has risk at least one. Clipping its output to $[-1, 1]$ allows this gap to be tested by one additional bounded risk query.

If both query count and inverse tolerance were bounded by $F(1/\eta)(d/\varepsilon)^C$ for a constant C independent of η, fixing a sufficiently large s and then taking d large would violate this query/tolerance tradeoff: the prefactor and ρ_s^{-1} are then constants, while the lower-bound exponent grows with s. Hence an SQ EPTAS of this resource form is impossible without a computational assumption. The stronger polynomial-approximation bounds of Diakonikolas et al. (2021b) further quantify the oracle tradeoffs; we do not identify their additive-accuracy exponents with a tight exponent for the relative parameter η in our upper bound.