FULLY POLYNOMIAL LEARNING OF POSITIVE RELU SUMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We give a fully polynomial algorithm for learning positive sums of bias-free ReLUs under Gaussian inputs. From labeled examples of a sum with k hidden units in $\mathbb{R}^d$, the algorithm returns, with probability at least $1-\delta$, an efficiently evaluable polynomial of degree $O(\epsilon^{-4/3})$ with relative L_2 error at most ϵ. It uses $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$ labeled samples and arithmetic operations, with the same bound on memory, output size, and evaluation time. The guarantee holds for arbitrary nonnegative weights and ridge vectors, including coincident and antipodal directions. The central step estimates a small, label-observable covariance at each degree. Aligning its factor with the hidden ridge tensors gives a low-rank comparator for tensor-train rounding; balanced degree splits then prevent the rounding loss from accumulating exponentially.

1 INTRODUCTION

A positive ReLU sum has the form $F(x) = \sum_{i=1}^{k} w_i \,\mathrm{ReLU}(v_i^\top x)$, where $\mathrm{ReLU}(t) \coloneqq \max\{t, 0\}$, $w_i \geq 0$, and $v_i \in \mathbb{R}^d$. We observe independent pairs $(X, F(X))$ for $X \sim \mathcal{N}(0, I_d)$ and seek a hypothesis H whose Gaussian L_2 error is at most $\epsilon \|F\|_{L_2}$. The directions may coincide, and their weights may be arbitrarily small. Positive homogeneity allows the analysis to normalize each nonzero v_i while retaining this entire class.

Theorem 1.1 (Fully polynomial learning, informal version of Theorem 5.3)**.** *Every positive sum of k ReLUs in $\mathbb{R}^d$ can be learned from Gaussian labeled examples to relative L_2 error ϵ with probability at least $1-\delta$. The hypothesis is a degree-$O(\epsilon^{-4/3})$ polynomial, and all sample, arithmetic, representation, and evaluation resources are jointly $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$ in the real-arithmetic model, for $0 < \epsilon, \delta \leq 1/2$.*

Earlier algorithms learned this class without a separation condition: Diakonikolas et al. (2020) obtained a proper learner with a width-dependent exponent, and Diakonikolas & Kane (2020) reduced that dependence to quasipolynomial. The fully polynomial question remained open in Diakonikolas & Kane (2024). The implicit-moment learner of Diakonikolas & Kane (2025) made the dependence on d, k polynomial while retaining exponential dependence on accuracy. Table 1 places the resulting bounds on the same noiseless Gaussian problem.

Table 1: Learning positive ReLU sums from noiseless Gaussian examples at constant success probability. The runtimes use L_2 error ϵ after normalizing the total weight; Theorem 5.3 gives the confidence dependence for our learner. Output names the returned hypothesis representation.

Work	Runtime	Output
Diakonikolas et al. (2020)	$\mathrm{poly}(d/\epsilon) + (k/\epsilon)^{O(k^2)}$	k-unit ReLU sum
Diakonikolas & Kane (2020)	$\mathrm{poly}(dk/\epsilon) + (k/\epsilon)^{O(\log^2 k)}$	Larger ReLU sum
Diakonikolas & Kane (2025)	$\mathrm{poly}(d, k) 2^{\mathrm{poly}(1/\epsilon)}$	Succinct polynomial
Our work	$\mathrm{poly}(d, k, 1/\epsilon)$	TT polynomial

Our learner returns a polynomial rather than recovered ReLU units; its Hermite coefficients at each degree are stored as a tensor train, so the hypothesis need not list a high-dimensional coefficient array.

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

The difficulty is representing informative high-degree statistics. In the unit-direction representation, the degree-s Hermite coefficient of the even part of F is a multiple of $\sum_i w_i v_i^{\otimes s}$, while a direct array for this tensor has d^s entries. At degree s, we instead maintain a subspace that approximates the ridge tensors $v_i^{\otimes s}$ in weighted squared distance. Symmetric products of the subspaces at degrees $\lfloor s/2 \rfloor$ and $\lceil s/2 \rceil$ give a candidate space of dimension at most k^2.

The covariance of the ridge tensors restricted to that space can be estimated from labels. A factor of this covariance is close, after an orthogonal alignment, to linear combinations of the hidden ridge tensors. These combinations admit small Hermite-coordinate tensor trains (TTs). Bicriteria rounding therefore compresses the computed factor while increasing its weighted approximation error by a constant factor. Because the degree split is balanced, the number of such increases along a dependency path is logarithmic in s. This is the mechanism behind the joint polynomial bound.

2 Observable moments and compact tensors

Throughout, $\|f\|_{L_2} \coloneqq (\mathbb{E}[f(X)^2])^{1/2}$ for $X \sim \mathcal{N}(0, I_d)$. The norm $\|u\|_2$ is Euclidean for vectors, and $\|A\|_F$ is Frobenius for tensors. Let $h_j \coloneqq \mathrm{He}_j / \sqrt{j!}$ be the normalized probabilists' Hermite polynomial of degree j. For a multi-index α, write $|\alpha| \coloneqq \sum_j \alpha_j$, $\alpha! \coloneqq \prod_j \alpha_j!$, and $h_\alpha(x) \coloneqq \prod_j h_{\alpha_j}(x_j)$; e_j is the jth standard basis vector. For a symmetric order-s tensor A, define $p_A(x) \coloneqq \langle A, \mathcal{H}_s(x) \rangle$, where $\mathcal{H}_s$ is characterized by $\langle v^{\otimes s}, \mathcal{H}_s(x) \rangle = h_s(v^\top x)$ for unit v. Gaussian orthogonality gives $\|p_A\|_{L_2} = \|A\|_F$.

Normalize nonzero ridge vectors to $\|v_i\|_2 = 1$ in the supporting analysis, and set $W \coloneqq \sum_i w_i$, $M_s \coloneqq \sum_i w_i v_i^{\otimes s}$, and $c_s \coloneqq \mathbb{E}[|G| h_s(G)]$ for $G \sim \mathcal{N}(0,1)$. Positive homogeneity removes this normalization in the final theorem.

Lemma 2.1 (Gaussian moment tests)**.** *For $F(x) = \sum_i w_i \,\mathrm{ReLU}(v_i^\top x)$ with nonnegative weights and unit directions, let $a \coloneqq \sum_i w_i v_i$. Then*

$$\mathbb{E}[F(X)] = \frac{W}{\sqrt{2\pi}}, \qquad \frac{W}{\sqrt{2\pi}} \le \|F\|_{L_2} \le \frac{W}{\sqrt{2}},$$

$$2F(x) = a^\top x + \sum_i w_i |v_i^\top x|,$$

$$a = 2\mathbb{E}[F(X)X].$$

For even $s \ge 2$ and symmetric A of order s,

$$2\mathbb{E}[F(X) p_A(X)] = c_s \langle A, M_s \rangle.$$

If p is a degree-m Gaussian chaos polynomial, then

$$\mathbb{E}[F(X)^2 p(X)^2] \le W^2 (d + 2m) \|p\|_{L_2}^2.$$

Moreover $c_{2j} = \sqrt{2/\pi}(-1)^{j-1}(2j-3)!!/\sqrt{(2j)!}$ for $j \ge 1$, with $(-1)!! = 1$; odd coefficients vanish, and $\sum_{s>T} c_s^2 \le T^{-3/2}$ for every even $T \ge 2$.

Proof. The Gaussian mean and second moment of a unit ReLU are $1/\sqrt{2\pi}$ and $1/2$. Weighted Cauchy–Schwarz bounds $\mathbb{E}[F(X)^2]$ by $W^2/2$, while Jensen gives the lower norm bound. The identity $2\,\mathrm{ReLU}(t) = t + |t|$ gives the decomposition, and rotational invariance gives $\mathbb{E}[X\,\mathrm{ReLU}(v^\top X)] = v/2$. The Hermite generating function factors as

$$\exp(tv^\top x - t^2/2) = \prod_{j=1}^{d} \exp(t v_j x_j - t^2 v_j^2/2).$$

The univariate degree-s projection of $|v^\top X|$ is $c_s h_s(v^\top X)$. Comparing degree-s coefficients in the displayed identity identifies its symmetric-tensor coefficient as $c_s v^{\otimes s}$. The linear part of $2F$ is orthogonal to every even positive degree, proving the label identity.

Expand $p = \sum_{|\alpha|=m} b_\alpha h_\alpha$. The recurrence

$$X_j h_\alpha(X) = \sqrt{\alpha_j + 1}\, h_{\alpha + e_j}(X) + \sqrt{\alpha_j}\, h_{\alpha - e_j}(X)$$

and orthogonality give $\sum_j \|X_j p\|_{L_2}^2 = (d+2m)\|p\|_{L_2}^2$. Since $0 \le F(X) \le W\|X\|_2$, this proves the variance bound.

Let φ be the standard Gaussian density. Since $\mathrm{He}_{2j}\,\varphi = \varphi^{(2j)}$ and the distributional second derivative of $|t|$ is $2\delta_0$, two integrations by parts give

$$\mathbb{E}[|G|\,\mathrm{He}_{2j}(G)] = 2\varphi^{(2j-2)}(0) = \sqrt{2/\pi}(-1)^{j-1}(2j-3)!!.$$

Divide by $\sqrt{(2j)!}$ for c_{2j}. Factorial cancellation and the central-binomial bound yield

$$c_{2j}^2 = \frac{2}{\pi}\frac{\binom{2j-2}{j-1}/4^{j-1}}{(2j)(2j-1)} \le \frac{1}{\pi j^{5/2}}.$$

For $T = 2N$, summing the integral bound $\sum_{j>N} j^{-5/2} \le (2/3)N^{-3/2}$ gives $\sum_{s>T} c_s^2 \le T^{-3/2}$. □

To store high-degree tensors, index an array by $\alpha \in \{0,\dots,s\}^d$. For a symmetric order-$s$ tensor A, let $A_{[\alpha]}$ be its entry on any ordered index string with multiplicities α, and define

$$\mathcal{J}_s(A)_\alpha \coloneqq \begin{cases} \sqrt{s!/\alpha!}\,A_{[\alpha]}, & |\alpha| = s, \\ 0, & |\alpha| \ne s. \end{cases}$$

A tensor train (TT) represents this array by a product of d matrix-valued cores, one for each coordinate. Its TT rank is the largest internal matrix dimension.

Lemma 2.2 (Hermite-coordinate TT operations)**.** *The map $\mathcal{J}_s$ preserves Frobenius norms and satisfies $p_A(x) = \sum_{|\alpha|=s} \mathcal{J}_s(A)_\alpha h_\alpha(x)$. A linear combination of k tensors $v_i^{\otimes s}$ has a Hermite-coordinate TT of rank at most $k(s+1)$. Given TTs for symmetric tensors A, B of degrees a, b and ranks R_A, R_B, one can compute a TT for $\mathrm{Sym}(A \otimes B)$ of rank at most $R_A R_B$. TT inner products, Gram matrices, and evaluation of $p_A(x)$ use arithmetic time polynomial in $d, a+b, R_A, R_B$.*

Proof. There are $s!/\alpha!$ ordered entries of multiplicity α, so

$$\|A\|_F^2 = \sum_{|\alpha|=s} \frac{s!}{\alpha!}|A_{[\alpha]}|^2 = \|\mathcal{J}_s(A)\|_F^2\,.$$

Comparing coefficients in the Hermite generating function from Lemma 2.1 gives the formula for p_A. For products, compare coefficients in $\langle A, z^{\otimes a}\rangle\langle B, z^{\otimes b}\rangle = \langle \mathrm{Sym}(A\otimes B), z^{\otimes(a+b)}\rangle$. For $\gamma = \alpha + \beta$ this yields

$$\mathcal{J}_{a+b}(\mathrm{Sym}(A\otimes B))_\gamma = \sqrt{\frac{a!b!}{(a+b)!}} \sum_{\alpha+\beta=\gamma} \sqrt{\frac{\gamma!}{\alpha!\beta!}}\,\mathcal{J}_a(A)_\alpha \mathcal{J}_b(B)_\beta.$$

The weight factors coordinatewise. For input cores G_j, H_j, use

$$K_j(t) \coloneqq \sum_{u+v=t} \sqrt{\frac{t!}{u!v!}}\,G_j(u) \otimes H_j(v),$$

and put the global factor in the first core. Its TT rank is at most $R_A R_B$. For a coherent power,

$$\mathcal{J}_s(v^{\otimes s})_\alpha = \sqrt{s!}\prod_j \frac{v_j^{\alpha_j}}{\sqrt{\alpha_j!}}\mathbf{1}_{\{|\alpha|=s\}}.$$

A core records partial degree $0,\dots,s$, giving TT rank $s+1$; direct sums of k such TTs have rank at most $k(s+1)$. Pairing cores and multiplying their transfer matrices computes inner products. Multiplying each core by $(h_0(x_j),\dots,h_s(x_j))$ evaluates $p_A(x)$ in $O(d(s+1)R_A^2)$ operations; the univariate Hermite values take $O(ds)$ operations by their three-term recurrence. □

Lemma 2.3 (Gaussian regression estimate). *Fix $C \in \mathbb{R}^{m\times N}$ and an $N \times \ell$ matrix U with $U^\top U = I_\ell$, where $0 \le \ell \le r$. Let $G \in \mathbb{R}^{N\times h}$ have independent $\mathcal{N}(0, 1/h)$ entries. If $0 < \alpha, \beta < 1/2$ and $h \ge 100(r+1)^2/(\alpha^2\beta)$, then, with probability at least $1-\beta$, every $D \in \mathbb{R}^{m\times\ell}$ simultaneously satisfies*

$$(1-\alpha)\left\|DU^\top - C\right\|_F^2 \le \left\|(DU^\top - C)G\right\|_F^2 \le (1+\alpha)\left\|DU^\top - C\right\|_F^2.$$

The bound holds conditionally when C, U depend on earlier randomness and G is fresh.

Proof. Put $E \coloneqq C(I - UU^\top)$ and $A \coloneqq D - CU$. Then $EU = 0$ and $\left\|DU^\top - C\right\|_F^2 = \|A\|_F^2 + \|E\|_F^2$. Gaussian second and fourth moments give

$$\begin{aligned}\mathbb{E}[\left\|U^\top GG^\top U - I\right\|_F^2] &= (\ell^2+\ell)/h,\\ \mathbb{E}[\left\|EGG^\top U\right\|_F^2] &= \ell\left\|E\right\|_F^2/h,\\ \mathrm{Var}(\|EG\|_F^2) &\le 2\left\|E\right\|_F^4/h.\end{aligned}$$

Markov's and Chebyshev's inequalities bound the Gram deviation, the cross term relative to $\|E\|_F$, and the residual-energy deviation relative to $\|E\|_F^2$ by $\alpha/3$ each, outside an event of probability at most $9(\ell^2+2\ell+2)/(h\alpha^2) \le \beta$. Zero E makes the last two conditions automatic. On this event the sketched squared residual differs from $\|A\|_F^2 + \|E\|_F^2$ by at most

$$\frac{\alpha}{3}(\|A\|_F^2 + 2\|A\|_F\|E\|_F + \|E\|_F^2) \le \frac{2\alpha}{3}(\|A\|_F^2 + \|E\|_F^2).$$

All three bounds depend on C, U, G and hold uniformly in A, hence in D. Conditioning leaves the Gaussian calculation unchanged. □

The TT representation follows Oseledets (2011); the Gaussian right-to-left sweep is given by Huber et al. (2017), and Mahankali et al. (2024, Appendix B.1) analyze its bicriteria approximation. The following compact-input version is the form used in our learner.

Proposition 2.4 (Compact TT rounding). *Given a q-mode TT A with local dimensions at most n and input rank at most R, a target rank $r \ge 1$, and $0 < p < 1$, a randomized algorithm outputs a TT $\widehat{A}$ such that, with probability at least $1-p$,*

$$\left\|A - \widehat{A}\right\|_F \le \frac{3}{2}\min_{B:\mathrm{rank}_{\mathrm{TT}}(B)\le r}\|A - B\|_F.$$

The minimum is the best error among TT tensors of rank at most r. The output rank is at most $2\cdot 10^4 q^3(r+1)^2$, and time, storage, and Gaussian randomness are bounded by $\mathrm{poly}(q, n, R, r, \log(1/p))$. For degree-$s$ Hermite-coordinate arrays, restricting the output to total degree s preserves the error bound and multiplies its TT rank by at most $s+1$.

Proof. For $q = 1$, return A. Otherwise fix a best TT-rank-r tensor B; a minimizer exists because the rank constraints on prefix unfoldings are closed. Sweep from mode q to mode 2. At cut j, the already chosen suffix is an isometry. For the current unfolding C and an orthonormal basis U of the corresponding prefix of B, the squared error of a feasible comparator has the form

$$c + \left\|DU^\top - C\right\|_F^2,$$

where c is the energy discarded by earlier suffix projections and $\mathrm{rank}(U) \le r$. Set $\alpha \coloneqq 1/(8q)$, $\beta \coloneqq 1/(3q)$, and $h \coloneqq 19200q^3(r+1)^2$. Draw a fresh $N \times h$ Gaussian matrix G on the prefix and let Q span the columns of CG. On the event of Lemma 2.3, $U^\top G$ has full row rank. The sketched least-squares minimizer

$$\widehat{D} \coloneqq CG(U^\top G)^\dagger$$

has columns in Q. Comparing its sketched residual with that of the best D and applying the two sides of Lemma 2.3 gives

$$\left\|\widehat{D}U^\top - C\right\|_F^2 \le \frac{1+\alpha}{1-\alpha}\min_D\left\|DU^\top - C\right\|_F^2.$$

The discarded energy c is charged once, so the full candidate error grows by at most $\kappa := (1+\alpha)/(1-\alpha)$. Express $\widehat{D}$ in the basis Q and absorb its coefficients into comparator core $j-1$ at cut j. This makes the new suffix feasible while leaving comparator cores through $j-2$ intact. Induction over the cuts gives squared-error multiplier at most $\kappa^{q-1} \le e^{1/2} < 9/4$; the final first core is the exact projection onto the chosen suffix. The $q-1$ conditional failure probabilities sum to less than $1/3$. Thus one sweep attains the stated $3/2$ norm factor with probability above $2/3$.

For a TT input, first left-orthogonalize its cores. At each cut the untouched prefix gives a factorization $CL^\top$ with $L^\top L = I$ and at most R columns. Conditional on earlier suffix choices, $L^\top G$ has the same law as a fresh $\mathcal{N}(0, 1/h)$ matrix Z with at most R rows. To compute the sketch, multiply input core A_j by the current suffix matrix D_{j+1} and reshape to $C_j \in \mathbb{R}^{(n_j t_{j+1}) \times \rho_{j-1}}$. Draw $Z_j \in \mathbb{R}^{\rho_{j-1} \times h}$, form a thin basis Q_j of $C_j Z_j$, and set $D_j := C_j^\top Q_j$; the reshaped $Q_j^\top$ is the next output core, and the first input core multiplied by D_2 closes the train. Every matrix has polynomial dimensions. An empty image returns zero cores of rank one. Iterated conditioning reproduces the joint distribution of the dense sweep, including rank-deficient cases. Independent sweeps followed by exact TT residual comparison reduce failure to p in $\lceil \log(1/p)/\log 3 \rceil$ repetitions. Each output rank is at most $h \le 2 \cdot 10^4 q^3 (r+1)^2$. Finally, entrywise multiplication by the rank-$(s+1)$ total-degree mask projects onto the homogeneous support. Since the input already has that support, projection preserves the error bound. □

3 THE LEARNER

The algorithm uses a common-batch median-of-means estimator. For J scalar tests $\psi_1, \dots, \psi_J$ of variance at most V, let $\mathrm{MoM}(\psi; V, t, \rho)$ draw fresh examples in the least odd number $b \ge 2\log_2(2J/\rho)$ of blocks, each of size $\lceil \max\{1, 8V/t^2\} \rceil$, and return the coordinatewise median of their block averages. Chebyshev's inequality makes a block fail in any fixed coordinate with probability at most $1/8$; the median and a union bound give simultaneous coordinate error t with probability at least $1-\rho$. These bounds also hold after conditioning on all earlier stages.

The two learners separate scale from shape. Algorithm 1 estimates total weight and calls Algorithm 2 on normalized labels; the latter builds a compact basis at each degree before estimating the final polynomial coefficients.

Algorithm 1 Relative-error learner

Require: Independent samples $(X, F(X))$ with $X \sim \mathcal{N}(0, I_d)$; d, k and $0 < \epsilon, \delta \le 1/2$.
Ensure: A polynomial H with $\|H - F\|_{L_2} \le \epsilon \|F\|_{L_2}$, with probability at least $1-\delta$.

1: Draw $\lceil \log_2(4/\delta) \rceil$ examples; if every label is zero, return the zero polynomial.
2: Draw the least odd number $b \ge 2\log_2(4/\delta)$ of fresh blocks of $\lceil 32\pi \rceil$ labels each.
3: Set $\widehat{W} \leftarrow \sqrt{2\pi}\,\mathrm{median}$(block means); if $\widehat{W} \le 0$, return zero.
4: Run Algorithm 2 on fresh labels $F(X)/\widehat{W}$ with $\eta \leftarrow \epsilon/(2\sqrt{2\pi})$ and failure parameter $p \leftarrow \delta/2$; return $\widehat{W}$ times its output.

For the bounded-mass learner below, choose the least power of two $T \ge 2$ with $T^3 \ge 10^4 \eta^{-4}$. Set

$$\nu := \frac{\eta^2}{10^8 (T+1)^{12}}, \qquad \tau := \nu/4, \qquad \xi := \theta := \nu/(4k),$$

$$\rho := \frac{p}{10(k+3)(T+1)}, \quad V_T := 16(d+4T) \max\left\{1, \max_{\substack{2 \le j \le 2T \\ j \text{ even}}} c_j^{-2}\right\}.$$

The explicit formula in Lemma 2.1 computes the required c_j by polynomially many arithmetic operations. The same lemma shows that V_T bounds the variances of all normalized moment and coefficient tests below. Every tensor basis is stored as an orthonormal list of Hermite-coordinate TTs. A positive rank-k truncation of a symmetric matrix keeps its k largest positive eigenvalues. A singular-value threshold retains left singular vectors with squared singular values at least the stated threshold; Gram matrices of TT columns determine these vectors and their TT combinations.

Algorithm 2 Bounded-mass learner, $W \in [1/2, 2]$

Require: Fresh Gaussian labeled examples; d, k and $0 < \eta, p \le 1/2$; parameters $T, \nu, \tau, \xi, \theta, \rho, V_T$ above.

Ensure: A polynomial of degree at most T with L_2 error at most η, with probability at least $1 - p$.

1: Estimate each entry of $K_1 = M_2$ from the tests $2Fp_{\mathrm{Sym}(e_i \otimes e_j)}/c_2$ using MoM with coordinate tolerance ξ/d and failure ρ; symmetrize.
2: Take the positive rank-k truncation, then keep eigenvalues at least θ; store its orthonormal eigenvectors as degree-one Hermite-coordinate TTs U_1.
3: **for** $s = 2, \dots, T$ **do**
4: $(a, b) \leftarrow (\lfloor s/2 \rfloor, \lceil s/2 \rceil)$; set $U_s \leftarrow \textsc{Merge}(U_a, U_b, s, \tau, \xi, \theta, V_T, \rho)$ by Algorithm 3.
5: **end for**
6: Estimate $b_0 \coloneqq 2\mathbb{E}[F(X)]$ to error $\eta/10$ and $a \coloneqq 2\mathbb{E}[F(X)X]$ to Euclidean error $\eta/10$ using fresh batches.
7: **for** even $s \in \{2, \dots, T\}$ **do**
8: Estimate each coordinate of $P_s M_s$ in U_s using MoM on tests $2Fp_A/c_s$, with variance bound V_T, tolerance $\eta/(10\sqrt{kT})$, and failure ρ. Store $\widehat{M}_s$ as the estimated linear combination of the basis TTs, or zero if U_s is empty.
9: **end for**
10: **return** $H(x) \leftarrow \big(\widehat{a}^\top x + \widehat{b}_0 + \sum_{2 \le s \le T,\ s \text{ even}} c_s p_{\widehat{M}_s}(x)\big)/2$.

For a degree-s orthogonal projector P, define the weighted loss

$$L_s(P) \coloneqq \sum_i w_i \left\|(I - P)v_i^{\otimes s}\right\|_F^2, \qquad K_s \coloneqq \sum_i w_i v_i^{\otimes s}(v_i^{\otimes s})^\top.$$

The merge estimates K_s only on a space spanned by products of the two child bases.

Algorithm 3 Balanced subspace merge

Require: Degree-a, b TT bases U_a, U_b of dimension at most k, $s = a + b$; $\tau, \xi, \theta, V_T, \rho$; fresh labeled examples.

Ensure: An orthonormal degree-s TT basis U_s of dimension at most k.

1: Form all $\mathrm{Sym}(A \otimes B)$ for $A \in U_a$, $B \in U_b$; from their Gram matrix retain squared singular values at least τ, obtaining a candidate basis $U = (A_1, \dots, A_m)$.
2: **if** $m = 0$ **then**
3: **return** the empty basis.
4: **end if**
5: Estimate each $B_{ij} = 2\mathbb{E}[F(X)p_{\mathrm{Sym}(A_i \otimes A_j)}(X)]/c_{2s}$ using MoM with variance bound V_T, coordinate tolerance ξ/m, and failure $\rho/2$; symmetrize.
6: Take the positive rank-k truncation $\widehat{B}$, and form $\widehat{Z}\widehat{Z}^\top = \widehat{B}$ with k columns, padding with zeros.
7: **for** $j = 1, \dots, k$ **do**
8: Form $Z_j = U\widehat{Z}_j$ as a TT; apply Proposition 2.4 with target rank $k(s + 1)$, failure $\rho/(2k)$, and a total-degree-s mask, obtaining C_j.
9: **end for**
10: From the Gram matrix of $C = (C_1, \dots, C_k)$, retain left singular vectors with squared singular values at least θ; return their TT basis.

4 Why a balanced merge works

The factor-alignment viewpoint is related to the Bures–Wasserstein distance studied by Bhatia et al. (2017). Here it turns an observable covariance estimate into a nearby low-TT-rank factor. The following inequality applies to rectangular factors and singular covariances.

Lemma 4.1 (Factor alignment). *For $X, Z \in \mathbb{R}^{m \times k}$, some orthogonal $O \in \mathbb{R}^{k \times k}$ satisfies*

$$\|Z - XO\|_F^2 \le \left\|ZZ^\top - XX^\top\right\|_*.$$

Proof. Put $A \coloneqq ZZ^\top$ and $B \coloneqq XX^\top$. Singular-value decomposition gives

$$\min_{O^\top O=I} \|Z - XO\|_F^2 = \operatorname{tr} A + \operatorname{tr} B - 2\left\|Z^\top X\right\|_* = \min_{U^\top U=I} \left\|A^{1/2} - B^{1/2}U\right\|_F^2 \le \left\|A^{1/2} - B^{1/2}\right\|_F^2 .$$

Let $D \coloneqq A^{1/2} - B^{1/2}$ and let S be its matrix sign. In an eigenbasis of D, positivity of $A^{1/2}$ and $B^{1/2}$ gives $|D_{jj}| \le (A^{1/2} + B^{1/2})_{jj}$. Consequently

$$\|D\|_F^2 \le \operatorname{tr}\big(|D|(A^{1/2} + B^{1/2})\big) = \operatorname{tr}\big(S(A - B)\big) \le \|A - B\|_* .$$

□

Lemma 4.2 (One balanced merge)**.** *Let $a, b \ge 1$ be integers, $s = a + b$, $W \in [1/2, 2]$, and let $T \ge s$ be the degree parameter of Algorithm 2. Suppose the child projectors P_a, P_b have rank at most k and orthonormal Hermite-coordinate TT bases. For any $\tau, \xi, \theta > 0$ and $0 < \rho < 1$, Algorithm 3 with the variance bound V_T from Section 3 returns, with probability at least $1 - \rho$, a projector P_s of rank at most k satisfying*

$$L_s(P_s) \le 200\big(L_a(P_a) + L_b(P_b) + \tau + k\xi + k\theta\big).$$

Its basis has TT rank polynomial in d, k, s, independently of the input TT ranks; its sample and arithmetic costs are polynomial in the input ranks, d, k, T, the inverse tolerances, and $\log(1/\rho)$.

Proof. Let $\mathcal{C}$ map the tensor product of the child coefficient spaces to symmetric order-s tensors by $\mathcal{C}(u \otimes v) = \operatorname{Sym}(U_a u \otimes U_b v)$. It has at most k^2 input coordinates. Let Q project onto the left singular directions of $\mathcal{C}$ whose squared singular values are at least τ. For each hidden atom, the coefficient vector of $P_a v_i^{\otimes a} \otimes P_b v_i^{\otimes b}$ has norm at most one. The discarded singular directions therefore contribute at most $\sqrt{\tau}$ to its image. Symmetrization is norm nonincreasing, and expanding the difference of two tensor products yields

$$\left\|v_i^{\otimes s} - \operatorname{Sym}(P_a v_i^{\otimes a} \otimes P_b v_i^{\otimes b})\right\|_F \le \left\|(I - P_a)v_i^{\otimes a}\right\|_F + \left\|(I - P_b)v_i^{\otimes b}\right\|_F .$$

Squaring the resulting three-term bound and summing with w_i gives

$$L_Q \coloneqq L_s(Q) \le 3\big(L_a(P_a) + L_b(P_b) + W\tau\big) \le 3\big(L_a(P_a) + L_b(P_b) + 2\tau\big).$$

If $Q = 0$, the algorithm returns zero and this bound already proves the claim.

Otherwise write $U \coloneqq (A_1, \dots, A_m)$ for its orthonormal basis, where $m \le k^2$. Its restricted covariance is $B \coloneqq U^\top K_s U$. For every pair of columns, Lemma 2.1 gives

$$B_{j\ell} = \langle \operatorname{Sym}(A_j \otimes A_\ell), M_{2s}\rangle = \frac{2}{c_{2s}} \mathbb{E}\big[F(X) p_{\operatorname{Sym}(A_j \otimes A_\ell)}(X)\big].$$

The test tensor has Frobenius norm at most one. Since $|c_{2s}|^{-1}$ and its second moment are polynomial in s, d, W by Lemma 2.1, a fresh common batch estimates all entries to ξ/m. The resulting symmetric matrix $\widetilde{B}$ satisfies $\|\widetilde{B} - B\|_{\mathrm{op}} \le \xi$ except with probability $\rho/2$.

Let $\widehat{B}$ be the positive rank-k truncation of $\widetilde{B}$. The true B is positive semidefinite of rank at most k, so the spectral tail removed from $\widetilde{B}$ has operator norm at most ξ. Hence $\|\widehat{B} - B\|_{\mathrm{op}} \le 2\xi$; as the difference has rank at most $2k$,

$$\left\|\widehat{B} - B\right\|_* \le 4k\xi.$$

Choose an $m \times k$ factor $\widehat{Z}$ with $\widehat{Z}\widehat{Z}^\top = \widehat{B}$, padding it by zero columns. Let X have columns $\sqrt{w_i} U^\top v_i^{\otimes s}$, padding if the target has fewer than k terms; then $XX^\top = B$. By Lemma 4.1, an orthogonal O obeys $\left\|\widehat{Z} - XO\right\|_F^2 \le 4k\xi$.

Embed the computed factor as $Z = U\widehat{Z}$. Let Y be the k columns obtained by applying O to $[\sqrt{w_i} v_i^{\otimes s}]_i$. Replacing full atoms by their projections has squared cost L_Q. The squared triangle inequality therefore gives

$$E \coloneqq \|Z - Y\|_F^2 \le 2L_Q + 8k\xi.$$

Every column of Y is a combination of at most k coherent powers and has Hermite-coordinate TT rank at most $k(s+1)$ by Lemma 2.2. Apply Proposition 2.4 to the columns of Z and denote their degree-s rounded factor by C. With total rounding failure at most $\rho/2$,

$$\|Z - C\|_F^2 \le \frac{9}{4}E, \qquad \|Y - C\|_F^2 \le \frac{25}{4}E.$$

The second bound follows from $\|Y - C\|_F \le \|Y - Z\|_F + \|Z - C\|_F$.

Let P_s retain the left singular vectors of C with squared singular values at least θ. At most k directions are discarded, so $\|(I - P_s)C\|_F^2 \le k\theta$. Orthogonal invariance gives $L_s(P_s) = \|(I - P_s)Y\|_F^2$, and hence

$$\begin{aligned} L_s(P_s) &\le 2\|Y - C\|_F^2 + 2\|(I - P_s)C\|_F^2 \\ &\le 13E + 2k\theta \\ &\le 78\big(L_a(P_a) + L_b(P_b)\big) + 156\tau + 104k\xi + 2k\theta, \end{aligned}$$

which is below the stated bound. The Gram matrices have sizes at most k^2 and k; Lemma 2.2 and Proposition 2.4 give the TT costs and the output rank bound. The two failure budgets sum to ρ. □

5 Accuracy and resource bounds

First we bound the loss of the retained subspace at each degree. The balanced recurrence keeps that loss polynomial in the degree; Hermite orthogonality then converts it into function error, and the rank bounds control storage and evaluation.

Lemma 5.1 (Balanced loss recurrence)**.** *Suppose $C \ge 1$, $\nu > 0$, $L_1 \le C\nu$, and $L_s \le C(L_{\lfloor s/2 \rfloor} + L_{\lceil s/2 \rceil} + \nu)$ for every $s \ge 2$. Then $L_s \le 4C^2\nu s^{1+\log_2 C}$.*

Proof. Expand the recurrence to degree-one leaves. The tree has s leaves, fewer than $2s$ nodes, and depth at most $\lceil \log_2 s \rceil$. Each node contributes at most $C\nu$ multiplied by C for each ancestor. Thus $L_s \le 2sC^{1+\lceil \log_2 s \rceil}\nu \le 2C^2\nu s^{1+\log_2 C}$; the stated bound also covers $s = 1$. □

Proposition 5.2 (Learning at bounded total weight)**.** *For a positive ReLU sum with unit directions and $W \in [1/2, 2]$, and for $0 < \eta, p \le 1/2$, Algorithm 2 returns, with probability at least $1 - p$, a Hermite-coordinate TT polynomial H of degree at most $T = O(\eta^{-4/3})$ satisfying $\|H - F\|_{L_2} \le \eta$. Its labeled sample count is at most*

$$O\left(V_T\left[\frac{Tk^6 + d^2k^2}{\nu^2} + \frac{kT^2 + d}{\eta^2}\right]\log\frac{2dkT}{\rho}\right).$$

Its arithmetic time, working memory, output size, evaluation time and memory, and auxiliary Gaussian draws are jointly polynomial in $d, k, 1/\eta, \log(1/p)$.

Proof. The base estimate of $K_1 = M_2$ has operator error at most ξ because every entry is estimated within ξ/d. Its positive rank-k truncation differs from K_1 in nuclear norm by at most $4k\xi$, by the spectral argument in Lemma 4.2. Removing squared eigenvalues below θ gives

$$L_1(P_1) = \operatorname{tr}((I - P_1)K_1) \le 4k\xi + k\theta = 5\nu/4.$$

The bound of Lemma 4.2 and $\tau + k\xi + k\theta = 3\nu/4$ yield $L_s \le 200(L_{\lfloor s/2 \rfloor} + L_{\lceil s/2 \rceil} + \nu)$. By Lemma 5.1, $1 + \log_2 200 < 9$, and the choice of ν,

$$\max_{s \le T} L_s \le 4 \cdot 200^2 \nu T^{1+\log_2 200} \le \eta^2/100.$$

Write $g(x) \coloneqq \sum_i w_i |v_i^\top x|$ and $a \coloneqq \sum_i w_i v_i$. By Lemma 2.1, $2F(x) = a^\top x + g(x)$, the constant Hermite coefficient of g is $b_0 \coloneqq 2\mathbb{E}[F(X)]$, and its degree-$s$ coefficient is $c_s M_s$ for positive even s.

Weighted Cauchy–Schwarz gives $\|(I-P_s)M_s\|_F^2 \le W L_s(P_s) \le 2L_s(P_s)$. Denote by $\widehat{g}$ the even polynomial returned in Algorithm 2. Orthogonality of Hermite degrees and of P_s and $I-P_s$ gives

$$\|g-\widehat{g}\|_{L_2}^2 \le \frac{\eta^2}{100} + W^2 \sum_{s>T} c_s^2 + 2\max_{s\le T} L_s + \frac{\eta^2}{100T} \sum_{\substack{2\le s\le T \\ s \text{ even}}} c_s^2$$
$$\le \frac{\eta^2}{100} + 4T^{-3/2} + \frac{2\eta^2}{100} + \frac{\eta^2}{100T} \le \frac{8\eta^2}{100}.$$

The linear and even errors are orthogonal; the linear estimate has Euclidean error at most $\eta/10$. Consequently

$$\|H-F\|_{L_2}^2 = \frac{1}{4}\big(\|\widehat{a}-a\|_2^2 + \|\widehat{g}-g\|_{L_2}^2\big) \le \frac{9\eta^2}{400} \le \eta^2.$$

Each covariance batch and output-coordinate batch uses fresh examples conditional on the learned bases. There are at most T calls of each type, two additional coefficient batches, and at most kT rounding calls within the merges. The per-call budget $\rho = p/[10(k+3)(T+1)]$ makes their total failure probability below p by conditioning and a union bound.

For the sample bound, a covariance on m coordinates uses m^2 tests at tolerance ξ/m. The median-of-means bound gives $O(V_T m^2 \xi^{-2}\log(2m^2/\rho))$ examples. The base has $m=d$, and the other at most T covariances have $m \le k^2$. Since $\xi = \nu/(4k)$, these contribute the first bracketed term. Each degree uses at most k coefficient tests at tolerance $\eta/(10\sqrt{kT})$; summing their batches gives $O(V_T kT^2\eta^{-2}\log(2k/\rho))$. The scalar and d linear tests contribute $O(V_T d\eta^{-2}\log(2d/\rho))$. These bounds yield the displayed expression.

For the remaining resources, set $Q(d,r) \coloneqq 2\cdot 10^4 d^3(r+1)^2$ and $B_T \coloneqq \max\{2, k(T+1)Q(d,k(T+1))\}$. Every rounded factor column has TT rank at most $(T+1)Q(d,k(T+1))$; a retained basis vector combines at most k such columns, so its rank is at most B_T. Child products have rank at most B_T^2, and the factors supplied to rounding have rank at most $k^4 B_T^2$. These bounds hold inductively after every merge. Lemma 2.2 and Proposition 2.4 make every Gram matrix, label test, factor operation, residual comparison, and output evaluation polynomial in these ranks. There are T stages, and rounding uses $O(\log(1/\rho))$ repetitions. Finally, $T = O(\eta^{-4/3})$, ν^{-1} is polynomial in $1/\eta$, and $V_T = O((d+T)T^{5/2})$ by Lemma 2.1. This proves all stated resource bounds. □

Theorem 5.3 (Fully polynomial learning, formal version of Theorem 1.1)**.** *Let $d,k \ge 1$ be integers and let $0 < \epsilon,\delta \le 1/2$. For every $F(x) = \sum_{i=1}^k w_i \,\mathrm{ReLU}(v_i^\top x)$ with $w_i \ge 0$ and arbitrary $v_i \in \mathbb{R}^d$, Algorithm 1 draws independent $(X,F(X))$ with $X \sim \mathcal{N}(0,I_d)$ and, with probability at least $1-\delta$, returns a polynomial H such that $\|H-F\|_{L_2} \le \epsilon\|F\|_{L_2}$. Its degree is $O(\epsilon^{-4/3})$, and its Hermite coefficients are represented by TTs. In the real-arithmetic model, its labeled sample count, preprocessing time, working memory, output size, evaluation time and memory, and auxiliary randomness are jointly bounded by* $\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))$*, with universal exponents.*

Proof. Discard zero directions and replace each remaining (w_i,v_i) by $(w_i\|v_i\|_2, v_i/\|v_i\|_2)$. Positive homogeneity preserves F. Let W be the total weight of this unit-direction representation. If $W=0$, every label is zero and Algorithm 1 returns the exact zero function. For $W>0$, one positive-weight summand shows $\Pr[F(X)>0] \ge 1/2$, so the pilot all-zero event has probability at most $\delta/4$.

By Lemma 2.1, $\mathbb{E}[F(X)] = W/\sqrt{2\pi}$ and $\mathbb{E}[F(X)^2]/\mathbb{E}[F(X)]^2 \le \pi$. Chebyshev's inequality on each scale-estimation block, followed by the median bound, gives $W/2 \le \widehat{W} \le 2W$ except with probability $\delta/4$. Conditional on these events, $F/\widehat{W}$ has total weight in $[1/2,2]$. Apply Proposition 5.2 to fresh normalized labels with $\eta = \epsilon/(2\sqrt{2\pi})$ and $p=\delta/2$, obtaining $\widetilde{H}$. For $H = \widehat{W}\widetilde{H}$,

$$\|H-F\|_{L_2} \le \widehat{W}\eta \le \frac{\epsilon W}{\sqrt{2\pi}} \le \epsilon\|F\|_{L_2}.$$

The three failure probabilities sum to δ. The scale estimate uses $O(\log(1/\delta))$ examples; rescaling labels and output preserves the degree and joint polynomial resource bounds in Proposition 5.2. □

AI USE STATEMENT

Generative AI tools assisted with research ideation and literature discovery; formulation of the theoretical model, mathematical claims, and proof ingredients; drafting and revising proofs and manuscript sections; and creating, implementing, and interpreting synthetic numerical checks. The authors take responsibility for the final text, mathematical claims, and references.

REFERENCES

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. arXiv:1712.01504, 2017. URL https://arxiv.org/abs/1712.01504.

Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *Proceedings of the 61st IEEE Symposium on Foundations of Computer Science*, pp. 184–195, 2020. doi: 10.1109/FOCS46700.2020.00026. URL https://arxiv.org/abs/2012.07774.

Ilias Diakonikolas and Daniel M. Kane. Efficiently learning one-hidden-layer ReLU networks via Schur polynomials. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1364–1378, 2024. URL https://proceedings.mlr.press/v247/diakonikolas24c.html.

Ilias Diakonikolas and Daniel M. Kane. Implicit high-order moment tensor estimation and learning latent variable models. In *Proceedings of the 66th IEEE Symposium on Foundations of Computer Science*, 2025. doi: 10.1109/FOCS63196.2025.00064. URL https://arxiv.org/abs/2411.15669v2.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1514–1539, 2020. URL https://proceedings.mlr.press/v125/diakonikolas20d.html.

Benjamin Huber, Reinhold Schneider, and Sebastian Wolf. A randomized tensor train singular value decomposition. In *Compressed Sensing and its Applications*, pp. 261–290. Birkhäuser, 2017. doi: 10.1007/978-3-319-69802-1_9. URL https://arxiv.org/abs/1710.08513.

Arvind V. Mahankali, David P. Woodruff, and Ziyu Zhang. Near-linear time and fixed-parameter tractable algorithms for tensor decompositions. In *15th Innovations in Theoretical Computer Science Conference*, volume 287 of *Leibniz International Proceedings in Informatics*, pp. 79:1–79:23, 2024. doi: 10.4230/LIPIcs.ITCS.2024.79. URL https://arxiv.org/abs/2207.07417v3. Appendix B.1 of the linked full version.

Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5): 2295–2317, 2011. doi: 10.1137/090752286.