MEASUREMENT AS TRAINING: EXPLORING BP-FREE LEARNING METHOD OF NEURAL NETWORKS VIA QUANTUM SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

The training of neural networks, especially large language models, is dominated by iterative optimization driven by backpropagation (BP), whose compute cost grows rapidly with the model and data size, and whose training process suffers from slow convergence, loss spikes, and even sudden collapse. Quantum computing (QC) has been proposed as a promising direction for breaking this bottleneck, yet existing studies merely insert quantum components into the classical pipeline, so the BP outer loop and its cost remain untouched. In this paper, we propose ***m**easurement **a**s **t**raining (MaT)*, a fundamentally different QC-based training scheme that does not require BP at all. MaT loads a coherent prior over discretized network parameters as a quantum superposition, marks the parameters whose full-batch training loss is at most ε through a coherent loss indicator, and applies amplitude amplification to return a certified ε-good parameter. Theoretically, we prove that under aligned split priors the prior mass of the ε-good set is a constant bounded away from zero, which yields $O(1)$ oracle queries up to preparation cost together with a full-batch ε-certificate and a quadratic query advantage over classical accept/reject from the same prior. We then extend the analysis from a single linear layer to deep linear models, ReLU networks, and softmax attention, and the guarantees become progressively more conditional as the architecture grows, with negative results and a certificate gap against mini-batch SGD delimiting the scope of the framework. We hope MaT can inspire more training paradigms beyond classical BP for neural networks.

1 INTRODUCTION

Neural networks, especially large language models (LLMs), have achieved remarkable success across natural language processing and are reshaping many scientific and industrial domains (Vaswani et al., 2017; Brown et al., 2020; OpenAI, 2023; Grattafiori et al., 2024; Yang et al., 2025; Kimi Team, 2025). However, the training of such models is notoriously expensive and fragile. On the one hand, the compute budget required by training grows rapidly with both the model size and the data size (Kaplan et al., 2020), which makes large-scale training a privilege of only a few institutions. The numbers are staggering. For instance, GPT-3 consumed an estimated 3.14×10^{23} FLOPs of training compute (Brown et al., 2020), and the largest Llama 3 model was pre-trained with 3.8×10^{25} FLOPs on up to 16,384 H100 GPUs (Grattafiori et al., 2024). On the other hand, the training process itself suffers from slow convergence, loss spikes, and even sudden collapse. The public logbook of OPT-175B documents repeated divergences that forced the team to roll back to earlier checkpoints and restart training (Zhang et al., 2022), such instabilities appear systematically even in small-scale proxies (Wortsman et al., 2023; Zhai et al., 2023), and dedicated spike-aware optimizers are still being developed in 2025 to keep pre-training stable (Huang et al., 2025), so each failed run burns the same enormous budget again. Such pain points motivate the search for fundamentally different training paradigms.

Quantum computing is a promising direction that deserves careful evaluation. Grover search and its generalization, amplitude amplification (Grover, 1996; Brassard et al., 2002), provably square-root the cost of finding marked items in an unstructured search space. Quantum machine learning (QML) has accordingly been proposed as a vehicle for importing such speedups into learning

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.

tasks (Biamonte et al., 2017; Cerezo et al., 2021; Benedetti et al., 2019; Garg & Ramakrishnan, 2020; Havlíček et al., 2019; Schuld & Killoran, 2019; Liu et al., 2021; Huang et al., 2021; Abbas et al., 2021; Dallaire-Demers & Killoran, 2018; Zoufal et al., 2019; Yang et al., 2022; Li et al., 2022; Zhang et al., 2025; Tomal et al., 2025; Smaldone et al., 2025; Kong et al., 2025; Aizpurua et al., 2026). This line of work has accelerated in the past two years. Quantum-enhanced attention layers and quantized self-attention have been embedded into classical architectures (Tomal et al., 2025; Smaldone et al., 2025), the first hybrid quantum-classical LLM for language generation was demonstrated with variational quantum circuits replacing roughly ten percent of the parameters of a 150M-parameter model (Kong et al., 2025), and trainable unitary adapters executed on a 156-qubit quantum processor were inserted into a frozen 8B LLM (Aizpurua et al., 2026).

However, existing studies on quantum computing for neural network training mostly optimize only a part of the iterative optimization pipeline, and therefore inherit rather than remove the pain points of classical learning methods. This raises the following drawbacks: *First*, they still operate within the classical gradient-based outer loop, so the parameter update remains an iterative backpropagation (BP) procedure whose per-step cost and slow convergence are untouched. Indeed, the most recent systems still insert quantum components only as adapters or replaced sublayers inside an otherwise classical pipeline (Tomal et al., 2025; Kong et al., 2025; Aizpurua et al., 2026). *Second*, when the parameterized model is trained by such an outer loop, the known landscape pathologies of quantum circuits, such as barren plateaus, can make the training itself infeasible (McClean et al., 2018; Larocca et al., 2023; 2025). *Third*, comparisons against classical optimization are rarely *leveled*, since the initialization distribution and the strength of the success criterion usually differ on the two sides, which makes the reported gains difficult to interpret. These limitations result in an awkward situation of training LLMs with quantum computing (QC), where QC seems merely a collection of fixes or buffs on the existing pipeline instead of a significant rebuild of the whole training paradigm.

To overcome these problems, in this paper, we design and analyze a fundamentally different training scheme built on the QC platform, which we call ***m**easurement **a**s **t**raining (MaT)*. The core principle of MaT relies on the power of QC where each parameter within the neural networks (NNs) is initialized in the *superposition* of all possibilities, and the model parameters only *collapse* once they (or their *entangled* variables) are *measured*. In this way, the "trained" NN parameters can be simply obtained once we obtain a good observation. This paradigm significantly differs from the classical learning models. On the one hand, **MaT does not require BP at all, and does not incur any extra backward or gradient computation resources**, meaning that the training procedure can be executed on one hardware as long as it supports the inference. On the other hand, the parameters returned by MaT are certified ε-good parameters of the full-batch training loss, instead of the minimizer of a mini-batch objective in classical training methods, demonstrating **a genuinely sample-parallel treatment of the whole training set** instead of a learning trajectory.

Theoretically, we develop three components that together determine when this scheme works. First, we derive quantitative lower bounds on the volume of the ε-good sets of linear least squares, deep linear models, ReLU networks with frozen activation patterns, and softmax attention. Second, we prove prior-mass theorems showing that under aligned split priors, whose narrow directions follow the loss-sensitive subspace and whose wide directions follow the flat interpolating subspace, the prior mass of the good set is a constant bounded away from zero. Third, we level all comparisons by fixing the prior family and the certificate strength on both sides, which yields a quadratic query advantage over classical rejection sampling from the identical prior with the identical predicate, together with the same conditional output law. We also map the boundaries of the scheme, including ambient uniform search, isotropic Gaussian priors, Euclidean gauge-split priors on curved orbits, and the certificate gap between our full-batch oracle and uncertified mini-batch SGD.

We summarize our contributions as follows:

• **Method:** We propose Measurement as Training (MaT), a QC-based training scheme. To the best of our knowledge, MaT is the first BP-free training strategy of NN parameters on QC, which is provably query-efficient with a full-batch certificate.

• **Theorems:** Starting from a complete analysis of the single linear layer, we extend the guarantees to deep linear models, ReLU networks with cell prior mass at least $1/8$, softmax attention on polynomial boxes, and their combinations.

• **Scope and Experiments:** We delimit the scope of the framework through four negative results and a certificate gap against uncertified mini-batch SGD, and we validate the theory empirically: finite instances check every theorem of the ladder, and a language-model probe on a real corpus shows that training manufactures the prior mass that MaT consumes (Appendices I.1 and I.2).

2 PRELIMINARIES

This section fixes the notation and formalizes both sides of the problem. Section 2.1 recalls the classical training of LLMs, and Sections 2.2–2.3 formalize the quantum query model in which our training scheme operates.

2.1 NOTATION AND CLASSICAL TRAINING

Training Procedure. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a fixed training dataset of N samples, where x_i denotes the input and y_i denotes the target. A neural network with trainable parameters $\theta \in \mathbb{R}^d$ defines a predictor f_θ, and each sample induces a nonnegative per-sample loss

$$\ell_i(\theta) := \ell\big(f_\theta(x_i), y_i\big).$$

The training objective is the full empirical risk

$$\mathcal{L}(\theta) := \sum_{i=1}^{N} \ell_i(\theta). \tag{1}$$

Throughout the paper, $\mathcal{L}$ always denotes this full-batch objective, and a parameter is called ε-good when $\mathcal{L}(\theta) \le \varepsilon$.

Mini-Batch SGD. Modern LLMs are trained by mini-batch SGD and its adaptive variants. At step t, a batch $B_t \subset \{1, \dots, N\}$ is sampled and the update reads

$$\theta_{t+1} = \theta_t - \eta_t \nabla \mathcal{L}_{B_t}(\theta_t), \qquad \mathcal{L}_{B_t}(\theta) := \sum_{i \in B_t} \ell_i(\theta), \tag{2}$$

where η_t is the learning rate and the gradient is computed by backpropagation (BP). Two properties of this loop matter for our discussion. First, the update is driven by the batch objective $\mathcal{L}_{B_t}$ rather than by the full objective $\mathcal{L}$, so termination does not certify a small value of $\mathcal{L}$ (Appendix F). Second, each step costs one backward pass, and a full training run chains T such steps.

2.2 PARAMETER GRID AND THE ε-GOOD SET

We work in a standard quantum query model, where the training data are classical and fixed, and the learnable parameters live in a quantum register. This work aims to clarify when MaT is query-efficient, when it is not, and what guarantees it provides beyond a single SGD trajectory.

Definition 1 (Parameter Box and Grid). *A quantum register of finitely many qubits spans only finitely many computational basis states, so the continuous parameter domain must be discretized into finitely many candidates. Fix dimension $d \in \mathbb{N}$, box radius $R > 0$, and mesh size $\delta \in (0, R]$, where R bounds the region searched and δ controls the granularity. The continuous parameter domain is the cube*

$$\mathcal{B}_R := [-R, R]^d,$$

and the discrete grid is

$$\mathcal{G}_{\delta,R} := \delta\mathbb{Z}^d \cap \mathcal{B}_R.$$

Let $M := |\mathcal{G}_{\delta,R}|$. Encode grid points injectively by bitstrings of length

$$K := \lceil \log_2 M \rceil,$$

and write $\theta \in \{0,1\}^K$ for the corresponding computational-basis label. A superposition of all M labels thus represents every candidate parameter simultaneously, and unused labels are assigned infinite loss so that the search never returns them.

Definition 2 (Training Loss and ε-Good Set)**.** *Producing a parameter that drives the full training loss of Equation 1 to at most ε is the goal of any training method. For $\varepsilon \geq 0$, the continuous and discrete ε-good sets collect exactly these parameters:*

$$\mathcal{S}_\varepsilon := \{\theta \in \mathcal{B}_R : \mathcal{L}(\theta) \leq \varepsilon\},$$
$$\mathcal{S}_\varepsilon^\delta := \mathcal{S}_\varepsilon \cap \mathcal{G}_{\delta,R}.$$

The uniform good-set fraction

$$p_\varepsilon := \frac{|\mathcal{S}_\varepsilon^\delta|}{M}$$

is the probability that a uniformly drawn grid point is ε-good, so classical random search needs about $1/p_\varepsilon$ draws to hit one.

Classical training draws a single initialization vector from a distribution and then improves it step by step, whereas our scheme loads the entire initialization distribution as **one** quantum state and holds all candidates in superposition. Formally, we have the following.

Definition 3 (Coherent Parameter Prior)**.** *A coherent parameter prior is a normalized state*

$$|\Psi_{\text{prior}}\rangle = \sum_{\theta \in \mathcal{G}_{\delta,R}} c_\theta |\theta\rangle$$

together with a (classical or quantum) procedure that prepares it. Measuring the state realizes the distribution $|c_\theta|^2$ over the grid, and the prior mass of the ε-good set,

$$p_\varepsilon^{\text{prior}} := \sum_{\theta \in \mathcal{S}_\varepsilon^\delta} |c_\theta|^2,$$

is the probability that the measured parameter is already ε-good.

This quantity is the central object of our main results. The uniform superposition is the special case $c_\theta = M^{-1/2}$. We also allow a finite label-register purification $\sum_z a_z |z\rangle|\theta(z)\rangle$ prepared by a reversible circuit. Its parameter prior is $\pi(\theta) = \sum_{z:\theta(z)=\theta} |a_z|^2$. Amplification reflects about the full prepared state and marks only the computed parameter. This convention is used for many-to-one chart pushforwards and gives the same conditional parameter law.

Definition 4 (Discrete Gaussian / Spectral Prior)**.** *The most important prior family mirrors the Gaussian initialization [1] familiar from classical deep learning. Fix a center $\theta_0 \in \mathcal{B}_R$, which plays the role of the initialization point, and a width $\sigma > 0$, which controls how far a typical sample strays from it. The discrete Gaussian prior (up to normalization on the grid) uses amplitudes satisfying*

$$|c_\theta|^2 \propto \exp\Big(-\frac{\|\theta - \theta_0\|_2^2}{2\sigma^2}\Big) \qquad (\theta \in \mathcal{G}_{\delta,R}).$$

Remark 1 (Cost Model)**.** *Complexity is measured by the number of calls to the loss oracle (Section 2.3), plus the cost of preparing the initial parameter state when that preparation is nontrivial. Implementing the oracle requires a coherent evaluation of $\mathcal{L}$ and a reversible comparison against ε. We do **not** treat one oracle call as free relative to a classical forward pass.*

2.3 The Oracle Hierarchy

The sampler never computes gradients. Its only access to the loss is a coherent indicator oracle, and the strength of the indicator determines the strength of the returned certificate. We adopt the full-batch hard indicator as the default and treat the weaker variants as boundary objects.

Definition 5 (Full-Batch Hard Indicator)**.** *An ε-indicator oracle is a unitary $U_\mathcal{L}$ on parameters and one ancilla such that for every grid label θ,*

$$U_\mathcal{L}|\theta\rangle|0\rangle = |\theta\rangle|f_\varepsilon(\theta)\rangle, \qquad f_\varepsilon(\theta) := \mathbf{1}[\mathcal{L}(\theta) \leq \varepsilon].$$

The oracle stamps every candidate parameter with its own answer bit $f_\varepsilon(\theta)$, which equals 1 exactly when θ is ε-good, so a single application evaluates the predicate on all grid points in superposition. The default oracle uses the full empirical loss, so that any returned parameter $\hat{\theta}$ satisfies $\mathcal{L}(\hat{\theta}) \leq \varepsilon$.

[1] A spectral prior replaces the Euclidean quadratic form by a quadratic form defined through a fixed positive-definite matrix (for example a diagonal fan-out mirroring classical Xavier- (Glorot & Bengio, 2010) or Kaiming-style (He et al., 2015) coordinate scalings), which shapes the widths differently along different directions and is exactly the freedom that our aligned split priors exploit.

Definition 6 (Mini-Batch and Soft Variants)**.** *For a batch $B \subset \{1, \ldots, N\}$, the mini-batch hard indicator marks θ when the batch loss $\mathcal{L}_B(\theta) \leq \varepsilon$. A soft score oracle coherently encodes a decreasing score of $\mathcal{L}(\theta)$ instead of a Boolean verdict. Both variants and their query theorems are deferred to Appendix D.*

3 Measurement as Training (MaT): Theoretical Analysis

This section defines measurement as training (MaT) and develops its analysis as a ladder of architectures, from a single linear layer through deep linear models, ReLU networks, and softmax attention, with guarantees weakening from aligned linear priors to additional chart and regularity hypotheses. Quantitative refinements of every statement are collected in Appendices B–E.

3.1 The MaT Algorithm and Its Universal Guarantees

While classical random parameter search draws a parameter from a prior and tests whether $\mathcal{L} \leq \varepsilon$, MaT performs the same search coherently:

Definition 7 (Measurement as Training (MaT))**.** *MaT regards the training procedure as multiple rounds of measurement. Specifically, it prepares the coherent prior $|\Psi_{\text{prior}}\rangle$ (Definition 3), applies amplitude amplification with the marked set defined by the ancilla bit of $U_{\mathcal{L}}$, and measures the parameter register. A final predicate test is included, and the algorithm reports a certified parameter only when that flag accepts. The algorithm succeeds if the outcome lies in $\mathcal{S}_\varepsilon^\delta$, and success carries the same certificate on both sides.*
Note. MaT is not random repetition. *The rounds are the coherent reflections of Grover search and its amplitude-amplification generalization (Grover, 1996; Brassard et al., 2002), not independent random draws, and the two mechanisms need different numbers of trials, about $1/\sqrt{p_\varepsilon^{\text{prior}}}$ oracle calls against about $1/p_\varepsilon^{\text{prior}}$ draws, with no unstructured quantum search primitive improving on the former (Appendix H.1). The readout stays random, as in any quantum sampler, and the output distribution is exactly the prior conditioned on the ε-good set, with the certificate on success. The design effort therefore goes into the geometric prior. A prior aligned with the loss geometry makes $p_\varepsilon^{\text{prior}}$ a constant, in which case a single prepare-amplify-measure procedure returns a certified parameter, and constructing such priors is the subject of the alignment theorems in Section 3.2 and of Appendix C.*

We adopt the hard full-batch indicator as the marked predicate because mini-batch SGD does not automatically certify it (Section 2.3).

Three guarantees of MaT hold for **every architecture** and **require no geometric input** beyond the prior mass $p_\varepsilon^{\text{prior}}$. We first fix the cost accounting, including the state-preparation cost that is often hidden in quantum algorithms.

Cost Accounting. We price one call to the hard full-batch indicator oracle $U_{\mathcal{L}}$ as one landscape access, comparable to one classical evaluation of the full-batch predicate f_ε, and we write T_Q for the total number of oracle calls made by MaT. One coherent preparation of $|\Psi_{\text{prior}}\rangle$ costs T_{prep}, and the preparation unitary and its inverse are invoked $O(T_Q)$ times inside amplification, so the total preparation cost is $O(T_Q T_{\text{prep}})$, or $O(T_{\text{prep}})$ when the good mass is bounded below by a constant. Lemma 27 gives a polynomial preparation bound for product discrete priors under a coherent hypothesis, and Remark 18 quantifies the approximation error of the prepared prior.

Theorem 1 (Universal Guarantees)**.** *Given $p_\varepsilon^{\text{prior}} \in (0, 1]$, $U_{\mathcal{L}}$, and a coherent prior-preparation unitary and its inverse, MaT enjoys the following:*
• ***Query Bound.*** *Amplitude amplification returns a sample from the prior conditioned on $\mathcal{S}_\varepsilon^\delta$ with success probability at least $2/3$ using $O(1/\sqrt{p_\varepsilon^{\text{prior}}})$ oracle calls when the mass or a constant-factor lower bound on it is known, and in expectation under the standard adaptive search when the mass is unknown (Brassard et al., 2002; Grover, 1996).*
• ***Certificate.*** *On success, $\mathcal{L}(\hat{\theta}) \leq \varepsilon$.*

• ***Output Law.*** *On success the output law is the prior conditioned on $\mathcal{S}_\varepsilon^\delta$, which is exactly the law realized by classical rejection sampling from the same prior.*

What differs across architectures is only the prior mass, so the analysis that follows focuses on the mass side of the ladder.

3.2 MaT Analysis: From a Single Linear Layer to Deep Architectures

We start with the simplest classical version, namely a single linear layer.

3.2.1 A Single Linear Layer

The Single Linear Layer. The training data consist of N input-target pairs with $X \in \mathbb{R}^{N\times d}$ stacking the inputs x_i, of full row rank ($d \geq N$), and $y = X\theta^\star$ stacking the targets y_i, for some $\theta^\star \in \mathcal{B}_{R/2}$, so the training loss is $\mathcal{L}(\theta) = \|X\theta - y\|_2^2$, where $\sigma_{\min}$ and $\sigma_{\max}$ denote the extreme singular values of X, and we write $r := \sqrt{\varepsilon}/\sigma_{\max}$. Here $\ker(X)$ denotes the null space of the input matrix, namely the set of all exact interpolators, and we call $\ker(X)^\perp$ and $\ker(X)$ the normal and kernel blocks.

Theorem 2 (Constant Prior Mass under Aligned Priors)**.** *Assume $N \geq 16$ and an aligned discrete Gaussian prior (Definition 4) whose center θ_0 sits within $r/2$ of the affine solution set $\theta^\star + \ker(X)$. The following prior families have $p_\varepsilon^{\text{prior}} \geq 1/4$.*

• ***Split Spectral Prior.*** *This is the aligned prior family used throughout the paper. For widths $\sigma_\perp$ and $\sigma_\parallel$ on the normal and kernel blocks, let the prior be the product box and grid in an orthonormal frame of $\ker(X)^\perp$ and $\ker(X)$, as in Definition 9. If $\sigma_\perp \leq r/\sqrt{18N}$, $\sigma_\parallel \in [R/4, R]$, $\delta \leq \min\{\sigma_\perp/(8d), \sigma_\parallel/(8d), r/(4\sqrt{d})\}$, and the ball of radius $4\sigma_\perp\sqrt{N}$ about the projection of θ_0 onto $\ker(X)^\perp$ sits in the box, then conditional samples additionally explore a macroscopic portion of the interpolating chart. When $d > N$, conditioning leaves the law of the kernel block unchanged, so samples explore each kernel coordinate on the scale R (Theorem 19).*

• ***Isotropic Prior.*** *This is the simplest witness that constant mass is possible. If the isotropic prior of width $\sigma \leq r/\sqrt{18N}$ satisfies Assumption 16, it has $p_\varepsilon^{\text{prior}} \geq 1/4$ (Theorem 18). Both priors therefore give $T_Q = O(1)$ oracle calls by Theorem 1.*

The mass bound and the universal guarantees assemble into the complete single-layer analysis.

Theorem 3 (MaT Analysis for a Single Linear Layer)**.** *Fix a realizable linear least-squares problem whose prior center is aligned with the affine solution set, and a split spectral prior that is narrow on the loss-sensitive directions and wide along the solution kernel, which guarantees a prior mass of at least $1/4$ by Theorem 2. Under the cost accounting above, the following hold simultaneously.*

• ***Query Bound.*** *Measurement as training (Definition 7) returns a sample from the prior conditioned on $\mathcal{S}_\varepsilon^\delta$ with success probability at least $2/3$ using $T_Q = O(1)$ calls to $U_\mathcal{L}$, plus T_{prep}, by Theorem 1.*

• ***Certificate.*** *On success the returned $\hat{\theta}$ satisfies $\mathcal{L}(\hat{\theta}) \leq \varepsilon$.*

• ***Matched Classical Baseline.*** *Classical rejection sampling from the identical prior and predicate has $T_{\text{rej}} = \Theta(1)$ expected predicate evaluations in this constant-mass regime, and more generally $T_{\text{rej}} = \Theta(1/p_\varepsilon^{\text{prior}})$ whenever the prior mass is $p_\varepsilon^{\text{prior}}$, with the same conditional output law on success, by Proposition 9.*

Proof sketch. The pipeline has three steps. First, the alignment condition places the prior center within $r/2$ of the affine solution set, and the Chebyshev lower tail keeps the normal fluctuation below $r/3$ with probability at least $7/8$ (the tube geometry is Lemma 1). Second, the typical-ball and discrete transfer bounds applied to the normal block (Lemmas 22 and 23) leave mass at least $1/4$, and product-box independence preserves the marginal law of the kernel block after conditioning, as in Theorem 2. Third, Theorem 1 converts the mass into the query bound and the certificate, and Proposition 9 supplies the matched classical baseline. □

3.2.2 Deep Linear Models and Linear Attention

We then climb to two-layer and three-layer deep linear models, and to fixed-mixing linear attention, where the exact solutions form a curved gauge manifold rather than a flat affine set (Baldi & Hornik, 1989; Kawaguchi, 2016).

The Deep Linear Model. An L-layer deep linear model stacks L parameter matrices and multiplies them, $f_W(x) = W_L \cdots W_1 x$ with $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, where $d_0 = d_{\text{in}}$ and $d_L = d_{\text{out}}$, so $d_1, \ldots, d_{L-1}$ are the hidden widths, collecting all factor entries into $\theta \in \mathbb{R}^D$ with $D = \sum_\ell d_\ell d_{\ell-1}$, and the training loss is $\mathcal{L}(\theta) = \sum_{i=1}^N \|f_W(x_i) - y_i\|_2^2$. We assume an interior zero-loss $\theta^\star$ whose intermediate products have full rank equal to $\operatorname{rank}(W_L^\star \cdots W_1^\star)$, with intermediate widths large enough for a nontrivial gauge.

Theorem 4 (MaT Analysis for Deep Linear Models and Linear Attention)**.** *Equipped with an aligned prior center and the chart-supported tubular prior along the gauge chart, which places the prior on a tube of fixed radius around the curved chart image rather than in the ambient space, with construction and width conditions in Appendix C.1.4, the following hold:*

• ***Two Layers.*** *Under the full-rank chart hypotheses of Lemma 9, $p_\varepsilon^{\text{prior}} \geq 1/4$, conditional samples explore a macroscopic portion of the chart domain, namely the ball on which the gauge chart is bi-Lipschitz, and the three guarantees of Theorem 3 hold verbatim, with T_{prep} including the coherent chart circuit (Remark 13).*

• ***Three Layers and Linear Attention.*** *Under Assumption 4, the same $1/4$ mass holds for the three-layer gauge chart of dimension $d_1^2 + d_2^2$, and for fixed-mixing linear attention, which reduces to the two-layer geometry on the mixed features. Softmax attention is not included.*

• ***More Layers.*** *For more than three layers under equal, monotone-funnel, or unimodal-valley hidden-width regimes, the gauge nondegeneracy obeys $\kappa^{(L)} \geq \sigma_*/(L \operatorname{cond}_*^{L-2})$ and the tubular prior still has mass at least $1/4$ (Corollary 6). Here L is the number of layers, σ_* is the smallest singular value among the factors $W_\ell^\star$, and cond_* is the largest condition number among the intermediate factors (Lemma 15). Multimodal zigzag widths remain open (Remark 5).*

Proof sketch. The pipeline is the same as in the single layer, with one geometric twist. First, the gauge freedom of deep linear models means the exact solutions form a curved manifold rather than a flat set, since multiplying the factors by (G, G^{-1}) preserves the product exactly (Lemma 8), and only $N_0 d_{\text{out}}$ directions are loss-sensitive, where N_0 is the rank of the matrix Z that stacks the training inputs (Lemma 10). Second, because the manifold is curved, the prior must live on a tube around the chart rather than in the ambient space, and in chart coordinates the prior mass is a constant $1/4$ (Theorem 20). Third, with constant mass in hand, Theorem 1 converts it into the query bound and the certificate, and Proposition 9 supplies the matched classical baseline. □

Informal Theorem 1 (Depth Bends the Good Set). Our analysis uncovers the following structural phenomenon. Under the stated rank conditions, the exact-solution set contains a gauge orbit of dimension d_1^2 for two layers and $d_1^2 + d_2^2$ for three layers, and the full zero set can have additional dimensions. Curvature is the whole story: a prior that is wide along the linear tangent can leave the good set, as the scalar counterexample shows (Proposition 6), whereas a prior following an exact-zero chart keeps mass at least $1/4$ under the stated width conditions. The chart-supported tubular prior of Theorem 4 implements exactly this idea, and this is why the ladder survives depth.

3.2.3 ReLU Networks

For ReLU networks, the role of the curved gauge manifold is played by activation patterns.

The ReLU Network. The network is $f_\theta(x) = V\,\sigma(Ux)$ with entrywise ReLU σ, first-layer weights $U \in \mathbb{R}^{m \times d_{\text{in}}}$, second-layer weights $V \in \mathbb{R}^{d_{\text{out}} \times m}$, and hidden width $m \geq N$. The training set is realizable by some interior $\theta^\star = (U^\star, V^\star)$, whose activation pattern $P_\star$ records the signs of the pre-activations $(U^\star x_i)_j$, and the margin $\gamma > 0$ denotes the smallest magnitude among these pre-activations. The network is piecewise linear, and a quantitative margin freezes the pattern, meaning that every U close enough to $U^\star$ induces the same signs on the training set (Montúfar et al., 2014; Pascanu et al., 2013). Inside one frozen pattern the network is linear in V at each fixed U, so a single pattern cell behaves like a linear least-squares problem with a flat interpolating fiber, namely

the affine set of second-layer weights that fit the targets exactly (Lemma 17 in Appendix B), and Σ_{box} denotes the automatic uniform bound on the spectral norm of the hidden-feature matrix across the whole box (Lemma 18).

Theorem 5 (MaT Analysis for ReLU Networks)**.** *The following hold:*

• ***Prior Mass on a Pattern Cell.*** *Fix a realizable ReLU pattern cell satisfying Assumption 19 and the cell-aligned split prior of Definition 12, with the loss-dependent U-width of Theorem 21, tight enough to keep U inside one stable pattern ball, a tight width on the loss-sensitive directions of V, and a wide width along the flat V-fiber. Then $p_\varepsilon^{\mathrm{prior}} \geq 1/8$, upgraded to $1/4$ under the additional width condition in that theorem. These constants also bound the mass of the good parameters inside the stated cell.*

• ***Sampler Guarantees.*** *Under the same conditions all guarantees of Theorem 3 hold verbatim.*

• ***Conditional Packing.*** *Data and width assumptions alone give a diluted-margin set of relative volume at least $1/2$ and an automatic Σ_{box} (Lemma 19). Under the volume-retention hypothesis, which asks that well-conditioned fibers keep a constant fraction of the diluted-margin volume inside the box (Assumption 10), a stable cover of relative volume at least $1/4$ follows (Theorem 14), while realizability and width alone do not imply volume retention (Remark 8). The grid packing of Theorem 13 also requires these box-room and residual-rounding conditions.*

3.2.4 Softmax Attention

For softmax attention, finally, even the geometric half needs an explicit hypothesis, and the sampler half is available only conditionally. We collect the trainable weights of a standard single-head residual attention block of head width d_h with bilinear pre-softmax scores into $\theta \in \mathbb{R}^D$, stack the training residuals into a $C^{1,1}$ map $F(\theta) \in \mathbb{R}^n$ of dimension $n = Nd_{\mathrm{out}}$ (Lipschitz Jacobian), so that $\mathcal{L}(\theta) = \|F(\theta)\|_2^2$, and call a parameter $\theta^\star$ with $F(\theta^\star) = 0$ a realizable zero (Assumption 12).

Theorem 6 (Softmax Attention: Geometry and a Conditional Sampler)**.** *The following hold:*

• ***Macroscopic Tube under a Hypothesis.*** *Under Assumption 15 which polynomially bounds the search box and local geometry at $\theta^\star$, the good set contains a tube around the exact solutions whose radius r_0, the inverse-function-theorem ratio of the Jacobian floor $\kappa_{\min}$ (the smallest singular value of $DF(\theta^\star)$) to the derivative scale, is an inverse-polynomial fraction of R (Theorem 17).*

• ***Conditional Sampler Guarantee.*** *The tube is quantitative, not only geometric. Under the same assumption, together with the loss tolerance $\sqrt{\varepsilon} \leq \kappa_{\min} r_0/4$ and a mesh condition $\delta \leq r/(4\sqrt{D})$, the uniform fraction of the box that is ε-good obeys $p_\varepsilon \geq C_{D,n} \left(\frac{5}{8}\right)^{D-n} \left(\frac{r_0}{4R\sqrt{D-n}}\right)^{D-n} \left(\frac{r}{R}\right)^{n}$, where r is the normal radius of the tube, supplied by Lemma 20, and $C_{D,n}$ is the explicit packing constant of Lemma 6. Hence the universal guarantees of Theorem 1 hold for the uniform prior with $T_Q = O(1/\sqrt{p_\varepsilon})$ oracle calls, the full-batch certificate, and the conditional output law, against a matched classical rejection cost of $\Theta(1/p_\varepsilon)$, which is not necessarily constant.*

• ***Boundaries.*** *No constant-mass softmax prior is constructed, so the $O(1)$ query bound remains open here. For fixed dimensions, vanilla bilinear scores prevent a radius proportional to R on which all logits stay bounded as $R \to \infty$ (Proposition 2), and scale-invariant scores on super-polynomial boxes remain open (Conjecture 1).*

3.2.5 Compositions of Analyzed Blocks

We close the ladder with compositions of the blocks above. The full statements and proofs are in Appendix B.1.4.

Theorem 7 (MaT Analysis for Compositions)**.** *The following hold.*

• ***Linear Stacks.*** *Series compositions of linear layers and fixed-mixing linear-attention layers are deeper linear models, so the more-layers item of Theorem 4 applies to the composite under all the corresponding factor-rank, width, chart, box, and prior-width hypotheses, with the gauge dimensions counted at the hidden layers of the composite.*

• ***Residual Linear Stacks.*** *In a residual stack of linear blocks, the gauge freedom of each block leaves the block map, and hence the end-to-end map and the loss, exactly invariant, so the exact-*

zero set contains the product of the per-block gauge orbits, and the orbit dimension adds over blocks under per-block nondegeneracy (Lemma 21).

• ***Attention Plus an MLP.*** *A block of the form $x \mapsto V\,\sigma\big(U(x + A_\theta x)\big)$ whose attention part A_θ is bilinear in trained factors (W_O, W_V) and vanishes when $W_O = 0$, as in fixed-mixing linear attention, inherits the ReLU cell guarantee on the zero-attention slice. Under the cell prior of Theorem 5 for (U, V) and a discrete Gaussian of width $\sigma_O \geq \delta$ centered at 0 on the $d_{\text{in}} \times d_h$ output factor W_O, the composite prior mass is at least $\frac{1}{8}\big(\frac{\delta}{4\sigma_O}\big)^{d_{\text{in}} d_h}$, with the certificate carried over unchanged (Proposition 3).*

4 Related Work

Near-term quantum machine learning typically trains parameterized circuits with a classical outer loop (Biamonte et al., 2017; Cerezo et al., 2021; Benedetti et al., 2019), and the landscape obstacles of that iterative paradigm, most prominently barren plateaus, have been analyzed from the original discoveries (McClean et al., 2018; Larocca et al., 2023) to the recent comprehensive review (Larocca et al., 2025). On the algorithmic side, Grover search and amplitude amplification (Grover, 1996; Brassard et al., 2002) provide query speedups for finding marked items, and quantum-parameterized circuits with coherent cost oracles (Shyamsundar, 2021; Peters & Shyamsundar, 2021) are the closest algorithmic precursors of our approach. For language models specifically, quantum self-attention networks and quantum-enhanced attention layers have been proposed for text classification (Li et al., 2022; Yang et al., 2022; Tomal et al., 2025), quantized self-attention has been applied to molecular generation (Smaldone et al., 2025), surveys now systematize the emerging quantum Transformer designs (Zhang et al., 2025), a hybrid quantum-classical LLM for language generation has been demonstrated (Kong et al., 2025), and trainable unitary adapters executed on real quantum hardware have been inserted into a frozen LLM (Aizpurua et al., 2026). All of these methods, however, still train through the classical BP outer loop and treat the quantum part as a replaced sublayer or an adapter. Compared with them, MaT removes the iterative loop entirely and rebuilds training as a single certified sampling step, contributing the prior-mass geometry, a leveled initialization comparison, and a full-batch certificate analysis.

5 Discussion: No-Free Lunch in Training Efficiency?

Quantum no-free-lunch theorems (Poland et al., 2020; Sharma et al., 2022) formalize that no advantage holds uniformly across problems, so a speedup must be bought with problem structure. Here the structure is the alignment between the prior and the loss geometry: the alignment theorems treat it as an input to the sampler and state sufficient structural conditions for a constant mass, while obtaining such priors, for example from a cheap warm start, remains a separate problem the guarantees take as given. Dropping the alignment has an explicit price, from the exponential decay of the uniform prior to the curved-orbit and bilinear-softmax obstructions (Appendices C and B). All comparisons are leveled against same-prior baselines with the full-batch certificate priced on both sides (Appendix F), the preparation cost T_{prep} is reported separately, and the framework offers a disciplined account of when the constant-mass regime holds, not an unconditional claim of quantum advantage. The certificate concerns the full-batch training loss, and generalization to held-out data is decided by the learning problem, not the sampler.

6 Conclusion

We have studied measurement as training, a certificate-bearing quantum sampler that draws neural network parameters from a coherent prior. The single linear layer establishes constant prior mass, $O(1)$ oracle queries up to preparation cost, and a full-batch ε-certificate for aligned split priors. The analysis then extends to a deep linear tubular prior, a ReLU cell prior of mass at least $1/8$ under the volume-retention hypothesis, and macroscopic softmax geometry on polynomial boxes, with progressively weaker guarantees. Four negative results and the certificate gap against mini-batch SGD delimit the scope of these guarantees (Appendix F). Finite-instance validation and a real-corpus probe accompany the theory in Appendix I. Sufficient conditions for ReLU volume retention and scale-invariant softmax beyond polynomial boxes remain open problems for future work.

AI USE STATEMENT

In this work, generative AI tools were used for drafting and editing text, for assisting with the preparation and verification of mathematical proofs, and for literature verification against original sources, which are tasks with required disclosure. Generative AI tools were also used to run and visualize the empirical validation in Appendices I.1 and I.2, which is a task with required disclosure. We have not used generative AI tools for generating synthetic data sets or for formulating the research contributions. We also used generative AI tools for readability edits and reference formatting, which are tasks with recommended disclosure. We have reviewed all AI-assisted work. All AI-assisted text was checked and edited by the authors, and all mathematical statements and proofs were verified line by line by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This work is theoretical, so reproducibility of the proofs rests on complete statements, explicit constants, and full proofs. Every theorem in Section 3 states its hypotheses in full, either inline or by reference to the appendix assumptions that define them, and the complete proofs of the supporting lemmas are collected in Appendices B–E, with all packing constants, chart radii, and preparation costs given in closed form. Unresolved extensions and necessary extra hypotheses are identified explicitly, including sufficient conditions for ReLU volume retention and scale-invariant softmax beyond polynomial boxes. The proved-versus-open split is recorded in Appendix H.1 and in the nonlinear checklist of Appendix B, and the comparison conventions used to read the query bounds are recorded in Appendix F. The source code of our empirical analysis is available at: `https://anonymous.4open.science/r/matQC/`.

REFERENCES

Amira Abbas et al. The power of quantum neural networks. *Nature Computational Science*, 1: 403–409, 2021.

Borja Aizpurua, Sukhbinder Singh, Augustine Kshetrimayum, Saeed S. Jahromi, and Roman Orus. Quantum-enhanced large language models on quantum hardware via Cayley unitary adapters. *arXiv preprint arXiv:2605.05914*, 2026.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Keith Ball. Volumes of sections of cubes and related problems. In *Geometric Aspects of Functional Analysis*, volume 1376 of *Lecture Notes in Mathematics*, pp. 251–260. Springer, 1989.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, 1997.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Information*, volume 305 of *Contemporary Mathematics*, pp. 53–74. American Mathematical Society, 2002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pierre-Luc Dallaire-Demers and Nathan Killoran. Quantum generative adversarial networks. *Physical Review A*, 98(1):012324, 2018.

Siddhant Garg and Goutham Ramakrishnan. Advances in quantum deep learning: An overview. *arXiv preprint arXiv:2005.04316*, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lov Grover and Terry Rudolph. Creating superpositions that correspond to efficiently integrable probability distributions. *arXiv preprint quant-ph/0208112*, 2002.

Lov K Grover. A fast quantum mechanical algorithm for database search. *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 212–219, 1996.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning, lecture 6a: Overview of mini-batch gradient descent. Coursera lecture slides, 2012.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, et al. Power of data in quantum machine learning. *Nature Communications*, 12:2631, 2021.

Tianjin Huang et al. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *International Conference on Learning Representations*, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Andrej Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165, 2020.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, 2016.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Alexei Kitaev and William A. Webb. Wavefunction preparation and resampling using a quantum computer. *arXiv preprint arXiv:0801.0342*, 2008.

Desheng Kong, Xiangshuo Cui, Jiaying Jin, Jing Xu, and Donglin Wang. Hybrid quantum transformer for language generation. *arXiv preprint arXiv:2511.10653*, 2025.

Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.

Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and M. Cerezo. Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 7:174–189, 2025.

Guangxi Li, Xuanqiang Zhao, and Xin Wang. Quantum self-attention neural networks for text classification. *arXiv preprint arXiv:2205.05625*, 2022.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17:1013–1017, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, 2018.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 27, 2014.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Razvan Pascanu, Guido Montúfar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Evan Peters and Prasanth Shyamsundar. Qarameterized circuits: Quantum parameters for QML. QHack Open Hackathon, 2021. See also `https://peterse.github.io/groveropt/`.

Kyle Poland, Kerstin Beer, and Tobias J Osborne. No free lunch for quantum machine learning. *arXiv preprint arXiv:2003.14103*, 2020.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

Kunal Sharma, Marco Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J Coles. Reformulation of the no-free-lunch theorem for entangled datasets. *Physical Review Letters*, 128 (7):070501, 2022.

Prasanth Shyamsundar. Non-boolean quantum amplitude amplification and quantum mean estimation. *arXiv preprint arXiv:2102.04975*, 2021.

Anthony M. Smaldone et al. A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation. *arXiv preprint arXiv:2502.19214*, 2025.

S. M. Yousuf Iqbal Tomal, Abdullah Al Shafin, Debojit Bhattacharjee, MD. Khairul Amin, and Rafiad Sadat Shahir. Quantum-enhanced attention mechanism in NLP: A hybrid classical-quantum approach. *arXiv preprint arXiv:2501.15630*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mitchell Wortsman et al. Small-scale proxies for large-scale Transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chao-Han Huck Yang, Jun Qi, et al. When BERT meets quantum temporal convolution learning for text classification in heterogeneous computing. *arXiv preprint arXiv:2203.03550*, 2022.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pp. 40770–40803, 2023.

Hui Zhang, Qinglin Zhao, Mengchu Zhou, Li Feng, Dusit Niyato, Shenggen Zheng, and Lin Chen. A survey of quantum transformers: Architectures, challenges and outlooks. *arXiv preprint arXiv:2504.03192*, 2025.

Susan Zhang, Stephen Roller, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. Quantum generative adversarial networks for learning and loading random distributions. *npj Quantum Information*, 5:103, 2019.

A Appendix Outline

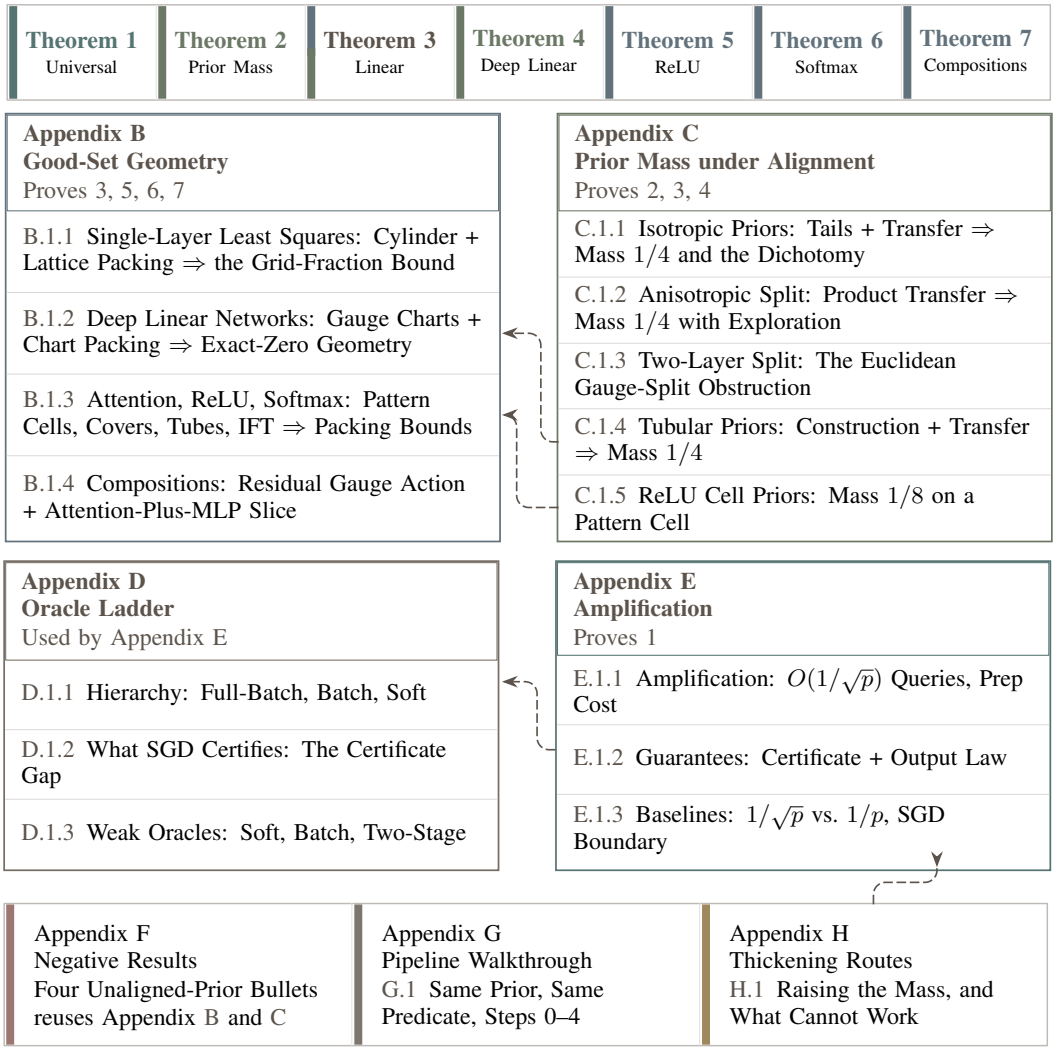

Figure 1: **A Map of the Appendix for the Convenience of Navigation**. Each panel is one appendix part, and each line is one subsection together with the result it contains. The top strip names the seven main-text theorems of the ladder in theorem order. Every panel can be read on its own. Within each part, the prose opens with a roadmap, and each proof states its idea before the details. 1. The bar color marks the appendix part that proves the main-text theorem drawn in that color. A split bar means the proof is split across both parts. The geometry part proves Theorems 3, 5, 6, and 7, the prior-mass part proves Theorems 2, 3, and 4, and the complexity part proves Theorem 1. 2. A dashed arrow points to the block whose result is reused. (a) C.1.4 reuses B.1.2. (b) C.1.5 reuses B.1.3. (c) Appendix E reuses Appendix D. (d) Appendix H reuses Appendix E. (e) Appendix F reuses Appendix B and Appendix C. 3. The supporting parts without their own main-text theorems are Appendix D, which prices the oracle variants and the certificate gap, Appendix G, which walks through the same-prior pipeline, and Appendix H, which collects the optimization routes discussed in the no-free-lunch section.

B Full Geometry Proofs

B.1 Geometry of ε-Good Sets

This appendix develops the geometry half of the argument, namely lower bounds on the uniform good-set fraction p_ε. The blocks follow the architecture ladder of the main text, and each block isolates the geometric mechanism that the corresponding prior-mass theorem uses. For the single linear layer the good set is an explicit cylinder, and a short packing pipeline counts its grid points. For deep linear models the exact solutions form curved gauge charts, and the packing counts a flat chart domain first and pushes the count through the chart. For ReLU networks the role of the chart is played by frozen activation patterns, with cylinder packings inside one pattern, finite unions of patterns, and packings under a polynomial-volume stable cover. For softmax attention the geometric guarantee is a local inverse-function-theorem tube under explicit nondegeneracy assumptions, together with an obstruction showing why the search box must stay polynomial, and a closing block records the composition statements that reduce residual linear stacks and attention-plus-MLP blocks to the analyzed ones. Statements whose hypotheses we could not verify from data or architecture alone are labeled as conjectures.

B.1.1 Single-Layer Linear Least Squares

The single-layer block is self-contained and needs no curvature input. Lemma 1 characterizes the good set through its normal component, and Lemma 2 exhibits an inscribed cylinder. Lemma 4 turns the Euclidean inner volume of that cylinder into a grid count. The argument uses the volume of Voronoi cells of $\delta\mathbb{Z}^d$, so it does not require $\ker(X)$ to be aligned with the coordinate axes. Lemma 6 converts the volume into a fraction, and Theorem 8 assembles the pipeline. Lemma 7 is the matching upper bound, used only for the uniform-prior comparison with gradient descent.

Assumption 1 (Realizable Linear Regression)**.** *Let $X \in \mathbb{R}^{N\times d}$ have full row rank N, with $d \geq N$, and let $y \in \mathbb{R}^N$ lie in the range of X, that is, $y = X\theta^\star$ for some $\theta^\star \in \mathcal{B}_{R/2}$. Write*

$$\sigma_{\min} := \sigma_{\min}(X) > 0, \qquad \sigma_{\max} := \sigma_{\max}(X),$$

for the smallest and largest singular values of X. Define

$$\mathcal{L}(\theta) := \|X\theta - y\|_2^2.$$

Lemma 1 (Normal-Space Characterization)**.** *For the linear instance of Section 3.2.1, let $\mathcal{K} := \ker(X)$ and $\mathcal{K}^\perp = \mathrm{row}(X)$. For $v = \theta - \theta^\star$ write $v = v_\parallel + v_\perp$ with $v_\parallel \in \mathcal{K}$ and $v_\perp \in \mathcal{K}^\perp$. Then*

$$\mathcal{L}(\theta) = \|Xv_\perp\|_2^2,$$

and

$$\sigma_{\min}\|v_\perp\|_2 \leq \|Xv_\perp\|_2 \leq \sigma_{\max}\|v_\perp\|_2.$$

Consequently,

$$\|v_\perp\|_2 \leq \frac{\sqrt{\varepsilon}}{\sigma_{\max}} \implies \mathcal{L}(\theta) \leq \varepsilon,$$

$$\mathcal{L}(\theta) \leq \varepsilon \implies \|v_\perp\|_2 \leq \frac{\sqrt{\varepsilon}}{\sigma_{\min}}.$$

Proof. Since $y = X\theta^\star$, we have $\mathcal{L}(\theta) = \|X(\theta - \theta^\star)\|_2^2 = \|Xv\|_2^2$. For $v_\parallel \in \ker(X)$ one has $Xv_\parallel = 0$, hence $Xv = Xv_\perp$. The singular-value inequalities on $\mathcal{K}^\perp$ are the definition of $\sigma_{\min}$ and $\sigma_{\max}$ restricted to the row space. Squaring yields the two implications. □

Lemma 2 (Continuous Tube Contains a Cylinder)**.** *Keep the linear instance of Section 3.2.1. Suppose*

$$r := \frac{\sqrt{\varepsilon}}{\sigma_{\max}} > 0$$

and that the parallel slice is nonempty:

$$\mathcal{A} := \big(\theta^\star + \mathcal{K}\big) \cap \mathcal{B}_{R-r} \neq \emptyset.$$

Then the Minkowski sum

$$\mathcal{C} := \mathcal{A} + \left(B_2^N(r) \cap \mathcal{K}^\perp\right)$$

satisfies $\mathcal{C} \subset \mathcal{S}_\varepsilon \cap \mathcal{B}_R$*, where* $B_2^N(r)$ *denotes the Euclidean ball of radius* r *after identifying* $\mathcal{K}^\perp \cong \mathbb{R}^N$*.*

Proof. Take $\theta = a + w$ with $a \in \mathcal{A}$ and $w \in \mathcal{K}^\perp$, $\|w\|_2 \le r$. Then $a - \theta^\star \in \mathcal{K}$, so the normal component of $\theta - \theta^\star$ is exactly w. Lemma 1 gives $\mathcal{L}(\theta) \le \varepsilon$. The radius-$r$ enlargement stays inside $\mathcal{B}_R$ by construction of $\mathcal{A}$. □

Lemma 3 (Inscribed-Cube Lattice Count)**.** *Let* $k \ge 1$ *and* $\rho > 0$*. After any orthonormal identification* $\mathbb{R}^k \cong \mathrm{span}$*, the Euclidean ball* $B_2^k(\rho)$ *contains the axis-aligned cube of half-side* $\rho/\sqrt{k}$*. If a* δ*-grid in those coordinates meets*

$$\delta \le \frac{\rho}{2\sqrt{k}},$$

then that cube contains at least

$$\left(\frac{\rho}{2\delta\sqrt{k}}\right)^k$$

grid points.

Proof. The cube $\left[-\rho/\sqrt{k}, \rho/\sqrt{k}\right]^k$ sits inside $B_2^k(\rho)$ because a vector with every coordinate at most $\rho/\sqrt{k}$ in absolute value has Euclidean norm at most ρ. Each coordinate interval has length $2\rho/\sqrt{k}$. The number of $\delta\mathbb{Z}$ points in an interval of length s is at least $s/\delta - 1$. The hypothesis $\delta \le \rho/(2\sqrt{k})$ yields $2\rho/(\delta\sqrt{k}) \ge 4$, hence $s/\delta - 1 \ge 2\rho/(\delta\sqrt{k}) - 1 \ge \rho/(2\delta\sqrt{k})$. The product over k coordinates is the claim. A global shift of the lattice by less than δ is absorbed by the same -1 per axis. □

Lemma 4 (Grid Count from Inner Volume)**.** *Let* $h := \delta\sqrt{d}/2$ *and let* $S \subset \mathcal{B}_R$*. Write* $S^{-2h} := \{x : B_2(x, 2h) \subset S\}$*. Then*

$$|S \cap \mathcal{G}_{\delta,R}| \ge \frac{\mathrm{vol}_d(S^{-2h})}{\delta^d}.$$

The right-hand side is interpreted as 0 *when* S^{-2h} *is empty.*

Proof. The Voronoi cells of the lattice $\delta\mathbb{Z}^d$ partition $\mathbb{R}^d$ up to boundaries of measure zero, and each cell has volume δ^d. Take $x \in S^{-2h}$ and let θ be a closest lattice point. Then $\|\theta - x\|_2 \le h$. The ball $B_2(x, 2h)$ is contained in S and hence in $\mathcal{B}_R$, so θ lies in that ball and therefore in $S \cap \mathcal{G}_{\delta,R}$. The Voronoi cells of these lattice points cover S^{-2h}. The cells are disjoint and each has volume δ^d, which is the claim. □

Lemma 5 (Inner Volume of the Linear Cylinder)**.** *Assume* $\delta \le r/(4\sqrt{d})$*, and write* $h := \delta\sqrt{d}/2$*, so* $h \le r/8$*. Let* $\mathcal{A}_{2h}$ *be the set of points of* $\mathcal{A}$ *whose intrinsic* $2h$*-neighborhood inside the affine subspace* $\theta^\star + \ker(X)$ *is still contained in* $\mathcal{A}$*. Then*

$$|\mathcal{S}_\varepsilon^\delta| \ge \mathrm{vol}_{d-N}(\mathcal{A}_{2h})\, v_N \left(\frac{3r}{4}\right)^N \Big/ \delta^d,$$

where $v_N := \pi^{N/2}/\Gamma(N/2+1)$ *is the volume of the Euclidean unit ball in* $\mathbb{R}^N$*, and* $v_0 := 1$*. The factor* $\mathrm{vol}_{d-N}(\mathcal{A}_{2h})$ *is Hausdorff measure on that affine subspace. Orthogonality of* $\ker(X)$ *and* $\ker(X)^\perp$ *fixes the product volume, independently of the angle between* $\ker(X)$ *and the coordinate axes.*

Proof. Lemma 2 gives the cylinder $\mathcal{C} = \mathcal{A} + (B_2^N(r) \cap \ker(X)^\perp)$ inside $\mathcal{S}_\varepsilon \cap \mathcal{B}_R$. Let $x = a + w$ with $a \in \mathcal{A}_{2h}$ and $w \in \ker(X)^\perp$, $\|w\|_2 \le r - 2h$. For any y with $\|y - x\|_2 \le 2h$, split $y - x$ into tangential and normal components. Each has norm at most $2h$, so the tangential piece stays in $\mathcal{A}$ and the normal piece stays inside the ball of radius r. Thus $x \in \mathcal{C}^{-2h}$. The hypothesis $h \le r/8$ gives $r - 2h \ge 3r/4$. The two subspaces are orthogonal, so the d-volume of this product is the displayed product of Hausdorff measure and Euclidean ball volume. Lemma 4 converts that volume into a grid count. □

Lemma 6 (Explicit Packing Factor). *Let $n \geq 1$ and $D \geq n$. Write $v_n := \pi^{n/2}/\Gamma(n/2+1)$ and*

$$C_{D,n} := \frac{2^{D-n}\, v_n\, (3/4)^n}{3^D}.$$

If $\delta \leq \min\{r/(4\sqrt{D}), R\}$ and an inner set supplies d-volume at least

$$c_0\, (2R)^{D-n}\, v_n \left(\frac{3r}{4}\right)^n$$

with $c_0 \in (0,1]$, then the good-set fraction on a grid of cardinality $M \leq (3R/\delta)^D$ satisfies

$$p \geq C_{D,n}\, c_0 \left(\frac{r}{R}\right)^n.$$

Proof. Lemma 4 gives at least that volume divided by δ^D grid points. Divide by $M \leq (3R/\delta)^D$. The powers of δ cancel. The remaining factor is

$$c_0\, 2^{D-n}\, R^{D-n}\, v_n\, (3r/4)^n \Big/ (3^D R^D) = C_{D,n}\, c_0\, (r/R)^n.$$

□

Theorem 8 (Good-Set Fraction for Linear Least Squares). *For the linear instance of Section 3.2.1, set $r = \sqrt{\varepsilon}/\sigma_{\max}$ and assume $\mathcal{A} \neq \emptyset$. Let $h := \delta\sqrt{d}/2$ and let $\mathcal{A}_{2h}$ be the intrinsic $2h$-interior of $\mathcal{A}$ from Lemma 5. Assume*

$$\mathrm{vol}_{d-N}(\mathcal{A}_{2h}) \geq c_0 (2R)^{d-N}$$

for a stated factor $c_0 \in (0,1]$ that may depend on the orientation, the location of the affine slice, r/R, and the erosion scale. If $\delta \leq \min\{r/(4\sqrt{d}), R\}$, then

$$p_\varepsilon \geq C_{d,N}\, c_0 \left(\frac{r}{R}\right)^N = C_{d,N}\, c_0 \left(\frac{\sqrt{\varepsilon}}{\sigma_{\max} R}\right)^N$$

with the explicit factor $C_{d,N} = 2^{d-N}\, v_N\, (3/4)^N / 3^d$ from Lemma 6, where $v_N = \pi^{N/2}/\Gamma(N/2+1)$.

Proof. Lemma 5 supplies the inner volume of the cylinder, with normal radius at least $3r/4$. Lemma 6 converts that volume into the displayed fraction. The identification $r = \sqrt{\varepsilon}/\sigma_{\max}$ is Lemma 1. □

Lemma 7 (Uniform Fraction Upper Bound for the Linear Tube). *Keep the linear instance of Section 3.2.1, and write $\rho := \sqrt{\varepsilon}/\sigma_{\min}$ and $h := \delta\sqrt{d}/2$. If $\delta \leq R$, then*

$$p_\varepsilon \leq v_N \Big(\frac{\sqrt{2}\,(\rho+h)}{R}\Big)^N \big(2(1+h/R)\big)^{d-N},$$

where $v_N = \pi^{N/2}/\Gamma(N/2+1)$. If the right-hand side exceeds 1, the trivial bound $p_\varepsilon \leq 1$ is the one in force.

Proof. Lemma 1 puts $\mathcal{S}_\varepsilon$ inside the tube of normal radius ρ about $\theta^\star + \ker(X)$. Every grid point of $\mathcal{S}_\varepsilon^\delta$ has its Voronoi cell of $\delta\mathbb{Z}^d$ inside the Euclidean h-neighborhood of that tube. Those cells are disjoint and each has volume δ^d, so

$$|\mathcal{S}_\varepsilon^\delta| \leq \mathrm{vol}_d(T_{\rho+h} \cap [-R-h, R+h]^d)/\delta^d.$$

The cube $[-R-h, R+h]^d$ is the cube $[-1/2, 1/2]^d$ scaled by $2(R+h)$. By the theorem of Ball, every central section of $[-1/2,1/2]^d$ of codimension N has $(d-N)$-volume at most $2^{N/2}$ (Ball, 1989). For a symmetric convex body, the Brunn Minkowski inequality makes the central section the largest parallel section, so the same bound applies to the affine slices parallel to $\ker(X)$. Scaling multiplies $(d-N)$-volumes by $(2(R+h))^{d-N}$. Integrating the normal ball of radius $\rho+h$ therefore bounds the tube volume by

$$v_N(\rho+h)^N\, 2^{N/2}\, (2(R+h))^{d-N}.$$

On each axis the number of points of $\delta\mathbb{Z} \cap [-R, R]$ is $2\lfloor R/\delta \rfloor + 1 \geq R/\delta$ once $\delta \leq R$, so $M \geq (R/\delta)^d$. Dividing the grid count by M produces

$$p_\varepsilon \leq v_N(\rho+h)^N\, 2^{N/2}\, 2^{d-N}(R+h)^{d-N}/R^d.$$

The identity $2^{N/2}(\rho+h)^N/R^N = (\sqrt{2}(\rho+h)/R)^N$ and the parallel factor $(2(R+h)/R)^{d-N}$ give the claim. □

Remark 2 (The Linear Case Is a Warmup). *Lemma 7 shows that ambient uniform search sees a good-set fraction that decays as a power of $\sqrt{\varepsilon}/R$ times $N^{-N/2}$, hence faster than exponentially in the data size N when the condition number and the ratio d/N stay bounded. Theorem 8 is the matching lower bound of the same shape, up to the explicit factor $C_{d,N}$. For realizable linear least squares a classical direct solver computes an exact interpolator in $\mathrm{poly}(N, d)$ arithmetic operations. Thus the linear ambient-grid case is a geometric warmup. It does not by itself witness a win against Gaussian elimination or against SGD. Stronger p_ε requires either a restricted search chart (coordinates along $\ker(X)$) or a nonlinear model whose sublevel sets are thick in ambient space.*

B.1.2 Deep Linear Networks

Depth replaces the flat affine solution set of the single layer by curved gauge orbits, so the cylinder argument needs a new first step. For two layers, Lemma 8 exhibits a d_1^2-dimensional exact-zero chart, and Lemma 9 gives that chart a macroscopic nondegenerate radius ρ_0. Lemma 11 then counts grid points in the flat chart domain and pushes them onto the curved image, and Theorem 9 converts the count into a good-set fraction whose exponent is the data rank $N_0 d_{\mathrm{out}}$ rather than the ambient dimension. The three-layer chart, its volume, and the factor-wise bound $\kappa^{(L)}$ repeat the same template for more layers.

Definition 8 (Deep Linear Model). *For widths $d_0 = d_{\mathrm{in}}, d_1, \ldots, d_L = d_{\mathrm{out}}$ let*

$$f_W(x) = W_L W_{L-1} \cdots W_1 x, \qquad W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}.$$

Collect all entries of $(W_L, \ldots, W_1)$ into a vector $\theta \in \mathbb{R}^D$ with

$$D := \sum_{\ell=1}^{L} d_\ell d_{\ell-1}.$$

With training pairs $\{(x_i, y_i)\}_{i=1}^N$ define

$$\mathcal{L}(\theta) := \sum_{i=1}^{N} \|f_W(x_i) - y_i\|_2^2.$$

Assumption 2 (Realizable Deep Linear Target). *There exists $\theta^\star$ in the interior of $\mathcal{B}_{R/2}$ with $\mathcal{L}(\theta^\star) = 0$. Write $W^{\mathrm{eff},\star} := W_L^\star \cdots W_1^\star$. Assume every intermediate width satisfies $d_\ell \geq \min(d_{\mathrm{in}}, d_{\mathrm{out}})$ and that each partial product along $\theta^\star$ has full rank equal to $\mathrm{rank}(W^{\mathrm{eff},\star})$.*

Lemma 8 (Two-Layer Gauge Chart). *Specialize to $L = 2$ with factors $A \in \mathbb{R}^{d_2 \times d_1}$ and $B \in \mathbb{R}^{d_1 \times d_0}$, so $D = d_2 d_1 + d_1 d_0$. Fix a full-rank factorization $A^\star B^\star = W^{\mathrm{eff},\star}$ with $\mathrm{rank}(W^{\mathrm{eff},\star}) = r \leq d_1$. Then for every $G \in \mathrm{GL}(d_1)$ small enough that $(A^\star G, G^{-1} B^\star)$ remains inside $\mathcal{B}_R$,*

$$\mathcal{L}(A^\star G, G^{-1} B^\star) = \mathcal{L}(A^\star, B^\star) = 0.$$

If $A^\star$ has full column rank and $B^\star$ has full row rank (hence $d_1 \leq \min(d_{\mathrm{in}}, d_{\mathrm{out}})$), the differential of $H \mapsto (A^\star(I + H), (I + H)^{-1} B^\star)$ at $H = 0$ is injective on $\mathbb{R}^{d_1 \times d_1}$, and the local orbit is a smooth embedded disk of dimension d_1^2. Without those rank conditions the construction still contains a local embedded disk of dimension at least r^2. Applying a quantitative theorem to a reduced chart also requires its own nondegeneracy and volume constants, and is not an automatic dimension substitution.

Proof. $(A^\star G)(G^{-1} B^\star) = A^\star B^\star$, so the end-to-end map is invariant and the training loss, which depends on the factors only through $f_W(x) = ABx$, is unchanged. The map $H \mapsto (A^\star(I+H), (I+H)^{-1} B^\star)$ is smooth for $\|H\|_F$ small. Its differential at $H = 0$ is $(A^\star H, -H B^\star)$. Full column rank of $A^\star$ and full row rank of $B^\star$ force $H = 0$ from $A^\star H = 0$ and $H B^\star = 0$, hence injectivity on $\mathbb{R}^{d_1 \times d_1}$. If only $\mathrm{rank}(W^{\mathrm{eff},\star}) = r$, the same argument on the active r-dimensional factors yields an r^2-dimensional embedded disk of exact zeros. □

Lemma 9 (Two-Layer Bi-Lipschitz Gauge Chart). *Under the full column and full row rank conditions of Lemma 8, write*

$$\varphi(H) := \big(A^\star(I + H),\ (I + H)^{-1} B^\star\big)$$

for $\|H\|_F < 1/2$, and

$$\kappa := \sigma_{\min}\big(D\varphi(0)\big) \geq \min\big\{\sigma_{\min}(A^\star), \sigma_{\min}(B^\star)\big\} > 0.$$

The differential is $D\varphi(H)[\dot{H}] = \big(A^\star\dot{H},\ -(I+H)^{-1}\dot{H}\,(I+H)^{-1}B^\star\big)$. Let $\mathrm{Lip}_{D\varphi}$ be a Lipschitz constant of $D\varphi$ on $\|H\|_F \leq 1/4$, and let Lip_φ be a Lipschitz constant of φ on that ball. Both are finite because φ is rational on a compact invertibility region. Set

$$\rho_0 := \min\Big\{\frac{1}{4},\ \frac{\mathrm{dist}(\theta^\star, \partial\mathcal{B}_R)}{2(1+\mathrm{Lip}_\varphi)},\ \frac{\kappa}{4\,\mathrm{Lip}_{D\varphi}}\Big\}.$$

Then φ is bi-Lipschitz from the Euclidean ball of radius ρ_0 in $\mathbb{R}^{d_1^2}$ onto its image, with lower constant $\ell \geq \kappa/2$ and a finite upper constant L_φ. Writing v_G for the d_1^2-dimensional volume of the image divided by $R^{d_1^2}$,

$$v_G \geq \Big(\frac{\kappa\rho_0}{2R\sqrt{d_1^2}}\Big)^{d_1^2}.$$

The radius ρ_0 does not shrink with ε. It is the macroscopic chart scale used by the tubular prior of Definition 11.

Proof. On the ball of radius ρ_0 one has $\|D\varphi(H) - D\varphi(0)\| \leq \mathrm{Lip}_{D\varphi}\rho_0 \leq \kappa/4$, hence $\sigma_{\min}(D\varphi(H)) \geq \kappa - \kappa/4 = 3\kappa/4$. Integrating the difference $D\varphi(H) - D\varphi(0)$ along a segment gives $\|\varphi(H) - \varphi(K) - D\varphi(0)(H-K)\| \leq (\kappa/4)\|H-K\|$. The triangle inequality yields a lower constant $3\kappa/4$, hence also $\ell \geq \kappa/2$. The upper constant L_φ is $\|D\varphi(0)\| + \kappa/4$. The domain ball contains a cube of half-side $\rho_0/\sqrt{d_1^2}$, whose volume is $(2\rho_0/\sqrt{d_1^2})^{d_1^2}$. Multiplying by $\ell^{d_1^2}$ lower-bounds the volume of the image. Dividing by $R^{d_1^2}$ produces the displayed v_G. The box constraint is the middle term in ρ_0. The singular-value lower bound on $D\varphi(0)$ is the same end-factor argument as Lemma 13: $\|D\varphi(0)[H]\|_F^2 \geq \|A^\star H\|_F^2 + \|HB^\star\|_F^2$. □

Lemma 10 (Effective Data Constraints for Deep Linear)**.** *Let $Z \in \mathbb{R}^{N\times d_{\mathrm{in}}}$ collect the training inputs as rows and let $Y \in \mathbb{R}^{N\times d_{\mathrm{out}}}$ collect the targets. Then*

$$\mathcal{L}(\theta) = \|ZW^{\mathrm{eff},\top} - Y\|_F^2 = \|Z(W^{\mathrm{eff}} - W^{\mathrm{eff},\star})^\top\|_F^2.$$

If $\mathrm{rank}(Z) = N_0 \leq N$, the linear map $W^{\mathrm{eff}} \mapsto Z(W^{\mathrm{eff}})^\top$ has rank $N_0 d_{\mathrm{out}}$. Thus the end-to-end constraint contributing to small loss has rank at most $N_0 d_{\mathrm{out}}$.

Proof. Expand each residual $W^{\mathrm{eff}}x_i^\top$ in Frobenius form. The stated matrix identity follows by stacking. The rank claim is the standard identity $\mathrm{rank}((Z\otimes I)\,\mathrm{vec}(W^{\mathrm{eff},\top})) = \mathrm{rank}(Z)\,d_{\mathrm{out}}$. □

Assumption 3 (Macroscopic Complementary Interpolating Cube)**.** *Let $n := N_0 d_{\mathrm{out}}$ as in Lemma 10 and let k be the dimension of a gauge chart through a realizable interior zero $\theta^\star$. Write $u := D - n - k$ and assume $u \geq 0$. Assume there exists a u-dimensional affine cube $\mathcal{Q} \subset \{\theta \in \mathcal{B}_R : \mathcal{L}(\theta) = 0\}$ of side $2R$, contained in a complementary exact-zero slice through the gauge disk (a lift of a data-kernel slice in W^{eff}, or a local constant-rank slice of $F^{-1}(0)$ that reaches the box faces). In particular the product of the truncated gauge disk with $\mathcal{Q}$ is an embedded interpolating set of $(D-n)$-dimensional volume at least $v_G R^k (2R)^u$.*

Lemma 11 (Flat-Domain Volume Pushed by a Bi-Lipschitz Chart)**.** *Let $k \geq 1$, $n \geq 1$, and $u := D-n-k \geq 0$. Let $\varphi : B_2^k(\rho_0) \to \mathbb{R}^D$ be bi-Lipschitz onto its image with lower constant $\ell > 0$, and let $\mathrm{Lip}_{D\varphi}$ be a Lipschitz constant of $D\varphi$ on that ball. Assume $\mathcal{L}(\varphi(H)) = 0$ for every domain point. Assume a Lipschitz tube of normal radius r about $\mathrm{im}(\varphi)$ sits inside $\mathcal{S}_\varepsilon$. Assume a complementary exact-zero cube of side $2R$ and dimension u as in Assumption 3, allowing $u = 0$ to contribute a factor 1. Assume*

$$\delta \leq \min\Big\{\frac{r}{4\sqrt{D}},\ R\Big\} \qquad \textit{and} \qquad r \leq \min\Big\{R/4,\ \frac{\ell^2}{2\,\mathrm{Lip}_{D\varphi}}\Big\}.$$

For the full $(D-n)$-dimensional product manifold, assumed C^2 on the interior of its parameter domain, impose the following tube conditions explicitly. The core Σ obtained from $\|H\| \leq \rho_0/2$ and

the intrinsically r-eroded complementary cube has volume at least $\ell^k v_k(\rho_0/2)^k(3R/2)^u$, where $h = \delta\sqrt{D}/2$. Its second fundamental form has norm at most $\mathrm{Lip}_{D\varphi}/\ell^2$, including the complementary directions. The normal map $(m, \nu) \mapsto m+\nu$ on this core is injective for $\|\nu\| \le r$, and its entire ambient r-neighborhood is contained in $\mathcal{B}_R \cap \mathcal{S}_\varepsilon$. These are additional geometric hypotheses. A curvature bound on the gauge chart alone does not establish them for its product with a complementary slice, or rule out global overlap of normal fibers.

Write $v_k := \pi^{k/2}/\Gamma(k/2+1)$ and

$$c_\varphi := \Big(\frac{5}{8}\Big)^{D-n}\Big(\frac{3}{4}\Big)^u v_k\Big(\frac{\ell\rho_0}{4R}\Big)^k.$$

Then

$$p_\varepsilon \ge C_{D,n}\, c_\varphi \left(\frac{r}{R}\right)^n.$$

The count is the inner Euclidean volume of a tube about the chart image, divided by the volume of a grid cell. It does not treat a curved chart as a coordinate-aligned product of grid cubes.

Proof. Let Σ be the image under φ of the domain ball of radius $\rho_0/2$, crossed with the exact-zero cube after intrinsic erosion by r, where $h := \delta\sqrt{D}/2 \le r/8$. Bi-Lipschitz control gives k-volume at least $\ell^k v_k(\rho_0/2)^k$. The eroded cube keeps side at least $3R/2$, because $2r \le R/2$, so its u-volume is at least $(3R/2)^u = (3/4)^u(2R)^u$. The product has $(D-n)$-volume at least

$$\ell^k v_k(\rho_0/2)^k\,(3/4)^u(2R)^u.$$

Dividing by $(2R)^{D-n}$ produces the factor $v_k(3/4)^u(\ell\rho_0/(4R))^k$. On this piece the second fundamental form satisfies $\|\mathrm{II}\| \le \mathrm{Lip}_{D\varphi}/\ell^2$ by the full-product curvature hypothesis. For the gauge factor alone this follows by pulling back arc-length with norm at most $1/\ell$, but the complementary directions require the extra stated assumption. The hypothesis on r gives $\|\mathrm{II}\|\cdot(3r/4) \le 3/8$. By the explicit injectivity and full-product curvature hypotheses, the normal map has tangential Jacobian at least $(5/8)^{D-n}$ on the normal ball of radius $3r/4$, while the normal directions themselves contribute Jacobian 1. The resulting D-volume is at least $c_\varphi(2R)^{D-n}v_n(3r/4)^n$. Lemma 6 converts that volume into the displayed fraction. Every point counted in the radius-$3r/4$ tube has its radius-$2h$ ball in the ambient r-neighborhood of its base point, because $2h \le r/4$. Thus the counted volume is inner volume in $\mathcal{B}_R \cap \mathcal{S}_\varepsilon$, as required by the lattice lemma. □

Theorem 9 (Quantitative Grid Fraction for Two-Layer Linear)**.** *Under the deep-linear assumptions of Section 3.2.2 with $L = 2$, under the full column/row rank conditions of Lemma 8 that make the orbit dimension d_1^2, and under Assumption 3 with $k = d_1^2$, let $N_0 = \mathrm{rank}(Z)$ as in Lemma 10, and set*

$$r := c_r\frac{\sqrt{\varepsilon}}{\mathrm{Lip}},$$

where Lip *is a Lipschitz constant of $\theta \mapsto \mathcal{L}(\theta)^{1/2}$ on $\mathcal{B}_R$ (finite on the compact box) and $c_r \in (0,1]$ is absolute. Let $k := d_1^2$ and let (ρ_0, ℓ) be the bi-Lipschitz radius and lower constant of Lemma 9. Write $n := N_0 d_{\mathrm{out}}$, $u := D - n - k$, and*

$$c_\varphi := \Big(\frac{5}{8}\Big)^{D-n}\Big(\frac{3}{4}\Big)^u v_k\Big(\frac{\ell\rho_0}{4R}\Big)^k$$

with $v_k = \pi^{k/2}/\Gamma(k/2+1)$, as in Lemma 11. If the hypotheses of Lemma 11 hold, then

$$p_\varepsilon \ge C_{D,n}\, c_\varphi \left(\frac{r}{R}\right)^n = C_{D,n}\, c_\varphi \left(\frac{c_r\sqrt{\varepsilon}}{\mathrm{Lip}\,R}\right)^n$$

with the explicit factor $C_{D,n}$ of Lemma 6. Gauge freedom contributes the pushed-chart prefactor c_φ. The exponent is governed by the data-side rank n rather than by the ambient factor dimension D. If Assumption 3 fails and the complementary exact-zero slice exists only at scale ρ_{fib}, the same packing yields an extra $(\rho_{\mathrm{fib}}/R)^u$ and the displayed bound is not claimed.

Proof. Lemma 9 supplies a bi-Lipschitz exact-zero chart of dimension $k = d_1^2$ on the ball of radius ρ_0 in H. Lemma 10 bounds the residual map by rank n, so a Lipschitz tube of radius r in that normal

space sits in $\mathcal{S}_\varepsilon$. Assumption 3 supplies the complementary exact-zero cube, and the additional full-product tube hypotheses are those of Lemma 11. Lemma 11 packs the flat H-domain first, pushes by φ, and converts the count through Lemma 6. Directions that are not exact zeros are not filled at scale R. They are either absorbed into the n-dimensional Lipschitz tube or omitted, which only weakens a lower bound. □

Remark 3 (Does Two-Layer Linear Beat SGD?)**.** *Theorem 9 improves the single-layer ambient picture by inserting v_G and by tying the exponent to $N_0 d_{\text{out}}$ instead of a naive ambient codimension. If $N_0 d_{\text{out}}$ remains $\Theta(N)$, the $(\sqrt{\varepsilon}/R)^{\Theta(N)}$ decay still prevents a win against $T_{\text{SGD}} = O(\text{poly}(N)\log(1/\varepsilon))$ under ambient uniform search. A win requires either v_G accompanied by a restricted chart that removes the ambient denominator, or a much smaller effective exponent.*

Lemma 12 (Layerwise Gauge Action)**.** *Keep the deep linear model of Section 3.2.2. For invertible matrices $G_\ell \in \text{GL}(d_\ell)$ with $\ell = 1, \dots, L-1$, the replacement*

$$W_\ell \mapsto G_\ell^{-1} W_\ell G_{\ell-1}, \qquad G_0 := I_{d_0}, \quad G_L := I_{d_L},$$

leaves $W_L \cdots W_1$ unchanged. Consequently the training loss is invariant along this GL *action (Baldi & Hornik, 1989; Kawaguchi, 2016).*

Proof. The product telescopes:

$$(G_L^{-1} W_L G_{L-1})(G_{L-1}^{-1} W_{L-1} G_{L-2}) \cdots (G_1^{-1} W_1 G_0) = W_L \cdots W_1.$$

The loss depends on the factors only through that product (Lemma 10). □

Assumption 4 (Three-Layer Gauge Nondegeneracy)**.** *Keep the deep-linear assumptions of Section 3.2.2 with $L = 3$. Write factors $(W_3^\star, W_2^\star, W_1^\star)$ at a realizable interior zero $\theta^\star$. Assume $W_1^\star$ has full row rank and $W_3^\star$ has full column rank, and that each intermediate width satisfies $d_1 \le d_{\text{in}}$ and $d_2 \le d_{\text{out}}$. Assume the truncated orbit constructed below remains inside $\mathcal{B}_R$.*

Lemma 13 (Three-Layer Gauge Chart)**.** *Under Assumption 4 define*

$$\varphi(H_1, H_2) := \big(W_3^\star(I + H_2),\ (I + H_2)^{-1} W_2^\star (I + H_1),\ (I + H_1)^{-1} W_1^\star\big)$$

for $\|H_\ell\|_F < 1/2$. Then $\mathcal{L}(\varphi(H_1, H_2)) = 0$. The differential at the origin is the linear map

$$D\varphi(0)[H_1, H_2] = \big(W_3^\star H_2,\ -H_2 W_2^\star + W_2^\star H_1,\ -H_1 W_1^\star\big),$$

which is injective, and

$$\sigma_{\min}\big(D\varphi(0)\big) \ge \min\big\{\sigma_{\min}(W_1^\star), \sigma_{\min}(W_3^\star)\big\} > 0.$$

The local orbit is therefore a smooth embedded disk of dimension $k := d_1^2 + d_2^2$.

Proof. Loss vanishing is Lemma 12 with $G_\ell = I + H_\ell$. Differentiating the chart at the origin is a direct expansion using $d(I + H)^{-1}\big|_{H=0} = -dH$. For injectivity, $D\varphi(0)[H_1, H_2] = 0$ implies $W_3^\star H_2 = 0$ and $H_1 W_1^\star = 0$. Full column rank of $W_3^\star$ forces $H_2 = 0$. Full row rank of $W_1^\star$ forces $H_1 = 0$. For the singular-value bound, drop the nonnegative middle Frobenius term:

$$\|D\varphi(0)[H_1, H_2]\|_F^2 \ge \|W_3^\star H_2\|_F^2 + \|H_1 W_1^\star\|_F^2 \ge \sigma_{\min}(W_3^\star)^2 \|H_2\|_F^2 + \sigma_{\min}(W_1^\star)^2 \|H_1\|_F^2.$$

The last inequality uses full column rank of $W_3^\star$ and full row rank of $W_1^\star$. Taking square roots yields the claimed $\sigma_{\min}$. Injective differential plus smoothness of φ on $\|H_\ell\|_F < 1/2$ gives an embedded disk of dimension k. □

Lemma 14 (Three-Layer Truncated Gauge Volume)**.** *Under Assumption 4 let $k = d_1^2 + d_2^2$ and $\kappa := \sigma_{\min}(D\varphi(0))$. Let $\text{Lip}_{D\varphi}$ be a Lipschitz constant of $D\varphi$ on $\|H_\ell\|_F \le 1/4$ (finite because φ is rational and the domain is compact). Set*

$$\rho_0 := \min\Big\{\frac{1}{4},\ \frac{\text{dist}(\theta^\star, \partial\mathcal{B}_R)}{2(1 + \text{Lip}_\varphi)},\ \frac{\kappa}{4\,\text{Lip}_{D\varphi}}\Big\}$$

with Lip_φ a Lipschitz constant of φ on that ball. Then φ is bi-Lipschitz from the Euclidean ball of radius ρ_0 in $\mathbb{R}^k$ onto its image, with lower constant $\ell \ge \kappa/2$. Writing $v_G^{(3)}$ for the k-dimensional volume of the image divided by R^k,

$$v_G^{(3)} \ge \Big(\frac{\kappa\rho_0}{2R\sqrt{k}}\Big)^k.$$

Proof. On the ball of radius ρ_0 one has $\|D\varphi(H) - D\varphi(0)\| \le \mathrm{Lip}_{D\varphi}\rho_0 \le \kappa/4$, hence $\sigma_{\min}(D\varphi(H)) \ge \kappa - \kappa/4 = 3\kappa/4$. Integrating the difference from the fixed map $D\varphi(0)$ gives $\|\varphi(H) - \varphi(K) - D\varphi(0)(H-K)\| \le (\kappa/4)\|H-K\|$. Thus a lower constant $3\kappa/4$, and in particular $\kappa/2$, is valid. The domain ball contains a cube of half-side $\rho_0/\sqrt{k}$, whose k-volume is $(2\rho_0/\sqrt{k})^k$. Multiplying by ℓ^k lower-bounds the volume of the image. Dividing by R^k produces the displayed $v_G^{(3)}$. The box constraint is the middle term in ρ_0. □

Theorem 10 (Quantitative Grid Fraction for Three-Layer Linear)**.** *Under the deep-linear assumptions of Section 3.2.2 and 4, and under Assumption 3 with $k = d_1^2 + d_2^2$, let $n := N_0 d_{\mathrm{out}}$ as in Lemma 10 and set $r := c_r\sqrt{\varepsilon}/\mathrm{Lip}$ as in Theorem 9. Let (ρ_0, ℓ) be the bi-Lipschitz pair of Lemma 14, and set $u := D - n - k$ and*

$$c_\varphi^{(3)} := \Big(\frac{5}{8}\Big)^{D-n}\Big(\frac{3}{4}\Big)^{u} v_k \Big(\frac{\ell\rho_0}{4R}\Big)^k$$

with $v_k = \pi^{k/2}/\Gamma(k/2+1)$, as in Lemma 11. If the hypotheses of Lemma 11 hold, then

$$p_\varepsilon \ge C_{D,n}\, c_\varphi^{(3)} \left(\frac{r}{R}\right)^n$$

with $k = d_1^2 + d_2^2$ and the explicit $C_{D,n}$ of Lemma 6. The volume fraction $v_G^{(3)}$ of Lemma 14 remains a valid continuous lower bound. The grid count uses domain-then-push packing rather than a flat ambient cube on the curved image.

Proof. Lemma 13 supplies a k-dimensional exact-zero chart. Lemma 14 supplies the bi-Lipschitz pair (ρ_0, ℓ) with $\ell \ge \kappa/2$. Lemma 10 bounds the residual map $\theta \mapsto Z(W^{\mathrm{eff}})^\top$ by rank n, so a Lipschitz tube of radius r in a normal space of dimension at most n stays inside $\mathcal{S}_\varepsilon$. Assumption 3 supplies the complementary cube. Lemma 11 packs the flat (H_1, H_2)-domain, pushes by φ, and converts the count through Lemma 6. Remaining fiber directions inside $\ker(DF)$ that are not packed can only add grid mass. □

Remark 4 (Three-Layer versus SGD)**.** *As in Remark 3, the exponent $n = N_0 d_{\mathrm{out}}$ can remain $\Theta(N)$. Ambient uniform search therefore need not beat SGD. The theorem instantiates $v_G^{(3)}$ and $C_{D,n}$. It does not create an unconditional first-order gap.*

Lemma 15 (Factor-Wise Lower Bound on $\kappa^{(L)}$)**.** *Keep the deep linear model of Section 3.2.2 with $L > 3$. Assume $W_1^\star$ has full row rank and $W_L^\star$ has full column rank. Write*

$$\sigma_\ell := \sigma_{\min}(W_\ell^\star), \qquad \Sigma_\ell := \sigma_{\max}(W_\ell^\star), \qquad \mathrm{cond}_\ell := \frac{\Sigma_\ell}{\sigma_\ell},$$

and

$$\sigma_* := \min_{1\le\ell\le L} \sigma_\ell, \qquad \mathrm{cond}_* := \max_{2\le\ell\le L-1} \mathrm{cond}_\ell.$$

Let $\varphi^{(L)}$ be the multiplicative gauge chart

$$(H_1, \dots, H_{L-1}) \mapsto \big(W_L^\star(I + H_{L-1}),\, (I+H_{L-1})^{-1}W_{L-1}^\star(I+H_{L-2}),\, \dots,\, (I+H_1)^{-1}W_1^\star\big)$$

on $\|H_\ell\|_F < 1/2$, with differential at the origin

$$D\varphi^{(L)}(0)[H] = \big(W_L^\star H_{L-1},\, W_{L-1}^\star H_{L-2} - H_{L-1}W_{L-1}^\star,\, \dots,\, W_2^\star H_1 - H_2 W_2^\star,\, -H_1 W_1^\star\big).$$

Assume one of the following width regimes.

1. *Equal hidden widths: $d_1 = \cdots = d_{L-1}$, and each intermediate $W_\ell^\star$ for $\ell = 2, \dots, L-1$ is square and invertible.*
2. *Nonincreasing funnel: $d_1 \ge d_2 \ge \cdots \ge d_{L-1}$, and each intermediate $W_\ell^\star$ for $\ell = 2, \dots, L-1$ has full row rank (so a right inverse exists).*
3. *Nondecreasing funnel: $d_1 \le d_2 \le \cdots \le d_{L-1}$, and each intermediate $W_\ell^\star$ for $\ell = 2, \dots, L-1$ has full column rank (so a left inverse exists).*

4. *Unimodal valley: there exists an index $k \in \{1, \dots, L-1\}$ with $d_1 \geq \cdots \geq d_k \leq \cdots \leq d_{L-1}$, each $W_\ell^\star$ for $\ell = 2, \dots, k$ has full row rank, and each $W_\ell^\star$ for $\ell = k+1, \dots, L-1$ has full column rank (with $W_L^\star$ still full column rank as above).*

Then $D\varphi^{(L)}(0)$ is injective and

$$\kappa^{(L)} := \sigma_{\min}\big(D\varphi^{(L)}(0)\big) \geq \frac{\sigma_*}{L\,\mathrm{cond}_*^{L-2}}.$$

For $L = 3$ the tighter end-factor bound of Lemma 13 remains available and does not need an invertible middle factor. Outside the four width regimes above, $\kappa^{(L)}$ is left as an explicit hypothesis. In particular a width sequence with two or more strict direction changes is not covered by the inverse-chain argument below.

Proof. Write $H = (H_1, \dots, H_{L-1})$ and

$$E_1 := -H_1 W_1^\star, \qquad R_\ell := W_\ell^\star H_{\ell-1} - H_\ell W_\ell^\star \quad (\ell = 2, \dots, L-1),$$

together with the unused end block $W_L^\star H_{L-1}$. Then

$$\big\|D\varphi^{(L)}(0)[H]\big\|_F^2 \geq \|E_1\|_F^2 + \sum_{\ell=2}^{L-1} \|R_\ell\|_F^2.$$

Full row rank of $W_1^\star$ gives $\|H_1\|_F \leq \|E_1\|_F/\sigma_1$.

Under item 1 or item 2, each intermediate $W_\ell^\star$ admits a right inverse of operator norm at most $1/\sigma_\ell$. Rearranging $R_\ell = W_\ell^\star H_{\ell-1} - H_\ell W_\ell^\star$ and multiplying on the right by that inverse produces

$$\|H_\ell\|_F \leq \mathrm{cond}_\ell \|H_{\ell-1}\|_F + \frac{\|R_\ell\|_F}{\sigma_\ell}.$$

(The square invertible case is the special case of a two-sided inverse.) Iterating and using $\mathrm{cond}_\ell \leq \mathrm{cond}_*$ and $\sigma_\ell \geq \sigma_*$ produces

$$\|H_\ell\|_F \leq \mathrm{cond}_*^{L-2}\Big(\frac{\|E_1\|_F}{\sigma_*} + \frac{1}{\sigma_*}\sum_{j=2}^{L-1} \|R_j\|_F\Big)$$

for every $\ell \leq L-1$. Taking the Euclidean norm over the $L-1$ blocks and applying Cauchy-Schwarz to the $L-1$ terms $\|E_1\| + \sum \|R_j\|$ yields

$$\|H\|_F \leq (L-1)\,\mathrm{cond}_*^{L-2}\,\sigma_*^{-1}\big\|D\varphi^{(L)}(0)[H]\big\|_F.$$

Since $L - 1 \leq L$ for $L \geq 2$, the claimed $\sigma_{\min}$ bound follows. If $D\varphi^{(L)}(0)[H] = 0$ then $E_1 = 0$ and every $R_\ell = 0$, so $H = 0$ by the same recurrence. This is injectivity.

Under item 3, reconstruct from the top instead. Full column rank of $W_L^\star$ gives $\|H_{L-1}\|_F \leq \|W_L^\star H_{L-1}\|_F/\sigma_L$. Each intermediate full column rank factor admits a left inverse of operator norm at most $1/\sigma_\ell$, and the identity $W_\ell^\star H_{\ell-1} = R_\ell + H_\ell W_\ell^\star$ rearranges to

$$\|H_{\ell-1}\|_F \leq \mathrm{cond}_\ell \|H_\ell\|_F + \frac{\|R_\ell\|_F}{\sigma_\ell}.$$

Iterating downward from H_{L-1} to H_1 yields the same $(L-1)\mathrm{cond}_*^{L-2}\sigma_*^{-1}$ comparison after Cauchy-Schwarz, and the same injectivity argument.

Under item 4, split the chain at the valley index k. On the left arm $\ell = 1, \dots, k$ the widths are nonincreasing, so the right-inverse recurrence of items 1–2 reconstructs $H_1, \dots, H_k$ from $(E_1, R_2, \dots, R_k)$ with the same $\mathrm{cond}_*^{L-2}\sigma_*^{-1}$ factors. On the right arm $\ell = k, \dots, L-1$ the widths are nondecreasing in the layer index, so the left-inverse recurrence of item 3 reconstructs $H_{L-1}, \dots, H_k$ from $(W_L^\star H_{L-1}, R_{L-1}, \dots, R_{k+1})$. The two arms share the block H_k, and the combined Euclidean estimate over all $L-1$ blocks again yields

$$\|H\|_F \leq (L-1)\,\mathrm{cond}_*^{L-2}\,\sigma_*^{-1}\big\|D\varphi^{(L)}(0)[H]\big\|_F.$$

If $D\varphi^{(L)}(0)[H] = 0$ then both arms force every block to vanish, which is injectivity. A width pattern with two or more strict direction changes leaves an intermediate segment that is reached by neither a pure right-inverse arm from the bottom nor a pure left-inverse arm from the top, so this argument stops. □

Remark 5 (Obstruction beyond Unimodal Valleys). *The reconstruction uses a single right-inverse arm, a single left-inverse arm, or one of each joined at a unique valley. Equal and monotone funnels are the degenerate unimodal cases. A multimodal zigzag (two or more strict direction changes) breaks both one-sided inverse chains before they cover every block. Closing $\kappa^{(L)} > 0$ in that class needs a different chart or extra intermediate rank structure. We do not claim it here.*

Proposition 1 (Inductive Template for $L > 3$). *For $L > 3$ keep the deep-linear assumptions of Section 3.2.2, Assumption 3 with $k_L = \sum_{\ell=1}^{L-1} d_\ell^2$, and the layerwise action of Lemma 12. Write $\varphi^{(L)}$ for the chart*

$$(H_1, \dots, H_{L-1}) \mapsto \big(W_L^\star(I + H_{L-1}),\, (I + H_{L-1})^{-1} W_{L-1}^\star (I + H_{L-2}),\, \dots,\, (I + H_1)^{-1} W_1^\star\big)$$

on $\|H_\ell\|_F < 1/2$. Under any of the four width regimes of Lemma 15, $D\varphi^{(L)}(0)$ is injective and the same domain-then-push packing as Theorem 10 yields

$$p_\varepsilon \geq C_{D,n}\, c_\varphi^{(L)} \left(\frac{r}{R}\right)^n, \qquad c_\varphi^{(L)} := \left(\frac{5}{8}\right)^{D-n} \left(\frac{3}{4}\right)^{D-n-k_L} v_{k_L} \left(\frac{\ell^{(L)} \rho_0^{(L)}}{4R}\right)^{k_L},$$

with $v_{k_L} = \pi^{k_L/2}/\Gamma(k_L/2+1)$, as in Lemma 11, where $n = N_0 d_{\mathrm{out}}$, under all the tube, injectivity, curvature, box, and mesh hypotheses of that lemma, and

$$v_G^{(L)} \geq \left(\frac{\kappa^{(L)} \rho_0^{(L)}}{2R\sqrt{k_L}}\right)^{k_L}, \qquad \kappa^{(L)} \geq \frac{\sigma_*}{L\, \mathrm{cond}_*^{L-2}}, \qquad \ell^{(L)} \geq \kappa^{(L)}/2,$$

where $\rho_0^{(L)}$ is the analogue of Lemma 14 (local invertibility of $I + H_\ell$, box, and $D\varphi$ perturbation) and σ_*, cond_* are as in Lemma 15. Outside those width regimes the same template holds with $\kappa^{(L)}$ left as an explicit hypothesis.

Proof. Lemma 15 supplies injectivity and the displayed factor-wise lower bound on $\kappa^{(L)}$ under equal, nonincreasing, nondecreasing, or unimodal-valley hidden widths with the matching full-rank intermediate hypotheses. Copy the bi-Lipschitz volume argument of Lemma 14 with $(\kappa^{(L)}, \rho_0^{(L)}, k_L)$ in place of (κ, ρ_0, k), then copy the domain-then-push packing of Lemma 11. If the widths have two or more strict direction changes or some intermediate factor fails the corresponding full-rank hypothesis, the reconstruction step of Lemma 15 is not available and $\kappa^{(L)}$ remains an explicit hypothesis (Remark 5). □

B.1.3 Attention, ReLU, and Softmax

The nonlinear blocks inherit the packing pipeline but replace the exact charts with structures that freeze part of the nonlinearity. For ReLU networks a quantitative activation margin freezes the pattern, and inside one frozen pattern the network is linear in the second-layer weights, so the single-layer cylinder argument returns with a pattern prefactor. For linear attention with fixed mixing, the model is exactly a two-layer deep linear model on mixed features, so the deep-linear block applies verbatim. For softmax attention nothing freezes, and the guarantee drops to a local inverse-function-theorem tube whose radius is macroscopic only under explicit scale-compatible bounds.

Assumption 5 (Wide Two-Layer ReLU). *Let $f_\theta(x) = V\sigma(Ux)$ with σ entrywise ReLU, $U \in \mathbb{R}^{m \times d_{\mathrm{in}}}$, $V \in \mathbb{R}^{d_{\mathrm{out}} \times m}$, and $m \geq N$. Assume the training set is realizable by some $\theta^\star = (U^\star, V^\star)$ in the interior of the box. Write $h_i^\star := \sigma(U^\star x_i)$ and let $H^\star \in \mathbb{R}^{N \times m}$ be the matrix with rows $(h_i^\star)^\top$.*

Assumption 6 (Pattern Stability at a Realizable ReLU Point). *Keep Assumption 5. Assume a quantitative margin*

$$\gamma := \min_{i,j} |(U^\star x_i)_j| > 0$$

and write $X_{\max} := \max_i \|x_i\|_2$. Assume $H^\star$ has full row rank N (possible once $m \geq N$ and the active hidden features spanning the samples are linearly independent). Let $P_\star$ be the activation pattern of $U^\star$ on the training set, namely the array $\mathrm{sign}(U^\star x_i)_j \in \{\pm 1\}$.

Lemma 16 (Fixed-Pattern ReLU Cell Contains a Ball). *Under Assumption 6, every first-layer matrix U with $\|U - U^\star\|_F \leq \rho_{\mathrm{pat}}$ for*

$$\rho_{\mathrm{pat}} := \frac{\gamma}{2X_{\max}}$$

induces the same training-set activation pattern $P_\star$. On that ball the network is the masked bilinear map $(U, V) \mapsto \big(V \,\mathrm{diag}(a_i) U x_i\big)_{i=1}^N$ with a fixed mask $a_i \in \{0,1\}^m$ determined by $P_\star$.

Proof. For each pair (i, j),

$$\big|(Ux_i)_j - (U^\star x_i)_j\big| \le \|U - U^\star\|_F \|x_i\|_2 \le \rho_{\mathrm{pat}} X_{\max} = \frac{\gamma}{2}.$$

The margin γ therefore prevents any coordinate from crossing zero, so the sign pattern is constant. ReLU with a frozen sign pattern is linear in U on each sample, with inactive units zeroed by the mask. □

Assumption 7 (Uniform V-Fiber Volume in a Pattern Ball)**.** *Keep Assumptions 5 and 6. Set*

$$\rho_U := \min\Big\{\frac{\gamma}{2X_{\max}},\ \frac{\sigma_{\min}(H^\star)}{2X_{\max}\sqrt{N}}\Big\}.$$

Assume this U-ball is contained in the U-box. Put $r_V = \sqrt{\varepsilon}/(2\sigma_{\max}(H^\star))$ and assume $r_V \le R/4$. For each U in the closed ball $\|U - U^\star\|_F \le \rho_U$, write $\mathcal{A}_V(U)$ for the affine interpolating fiber in V intersected with the smaller box $[-R + r_V, R - r_V]^{md_{\mathrm{out}}}$, and write $\kappa_V := (m - N)d_{\mathrm{out}}$ for the dimension of that fiber (full row rank N of $H(U)$). Assume

$$c_{0,V} := \inf_{\|U - U^\star\|_F \le \rho_U} \frac{\mathrm{vol}_{\kappa_V}\big(\mathcal{A}_V(U)\big)}{(2R)^{\kappa_V}} \in (0, 1].$$

This is the relative volume of the fiber with room for the normal tube. It is not assumed to equal 1 *unless each fiber meets the box in a full-radius cube.*

For the ReLU grid-packing results below, assume $R/\delta \in \mathbb{N}$ and impose the additional rounding condition

$$\delta \le \frac{\sqrt{\varepsilon}}{4\sqrt{D}\,L_R}, \qquad L_R := \sqrt{N} X_{\max} R \left(\sqrt{md_{\mathrm{in}}} + \sqrt{md_{\mathrm{out}}}\right).$$

The training residual is L_R-Lipschitz on the box: expand the difference $H(U)V^\top - H(U')V'^\top$, use the nonexpansiveness of ReLU, and bound the two weight norms by their box bounds. Aligned box endpoints ensure that nearest lattice rounding stays in the box and moves a point by at most $h = \delta\sqrt{D}/2$. This condition controls the change of the residual when both U and V are rounded.

Lemma 17 (Fixed-Pattern ReLU Cylinder)**.** *Under Assumptions 5, 6, and 7, write $n := Nd_{\mathrm{out}}$, $D := md_{\mathrm{in}} + d_{\mathrm{out}}m$, and keep ρ_U as in Assumption 7. On the ball $\|U - U^\star\|_F \le \rho_U$ the activation pattern is $P_\star$ by Lemma 16, and $\sigma_{\min}(H(U)) \ge \sigma_{\min}(H^\star)/2$, so every such U admits an exact interpolating V (full row rank linear least squares in V). Let $r_V := \sqrt{\varepsilon}/(2\sigma_{\max}(H^\star))$. If $\delta \le \min\{r_V/(4\sqrt{D}), \rho_U/(2\sqrt{D})\}$ and the packing is nonvacuous, then*

$$p_\varepsilon \ge C_{D,n}\, c_{\mathrm{pat}}\, c_{0,V} \left(\frac{r_V}{R}\right)^n,$$

where $C_{D,n}$ is Lemma 6,

$$c_{\mathrm{pat}} \ge \Big(\frac{\rho_U}{2R\sqrt{md_{\mathrm{in}}}}\Big)^{md_{\mathrm{in}}}$$

is the relative volume of the inscribed pattern ball in U-space, and $c_{0,V}$ is the V-fiber factor of Assumption 7. The V-fiber is affine, hence flat. No unit-factor identification of $c_{0,V}$ with 1 *is used. This lemma does not union over activation patterns.*

Proof. On the pattern cell the network is linear in V at each frozen U. The hidden-feature matrix satisfies $\|H(U) - H^\star\|_F \le \sqrt{N}\,\rho_U X_{\max} \le \sigma_{\min}(H^\star)/2$, hence $\sigma_{\min}(H(U)) \ge \sigma_{\min}(H^\star)/2$ and each such U has an affine interpolating fiber in V of dimension κ_V. In particular one must re-solve V as U moves. Keeping $V = V^\star$ while moving U can produce a residual of size $\Theta(\|V^\star\|_F\gamma)$, which need not be $\le \sqrt{\varepsilon}$. After re-solving, Lemma 1 applied to the linear map in V with design $H(U)$ gives an ε-tube of normal radius at least r_V in an n-dimensional residual space, because $\sigma_{\max}(H(U)) \le \sigma_{\max}(H^\star) + \sigma_{\min}(H^\star)/2 \le 2\sigma_{\max}(H^\star)$ once $\sigma_{\min}(H^\star) \le \sigma_{\max}(H^\star)$. Use U in the ball of radius $\rho_U/2$ and, at each such U, use the trimmed interpolating fiber plus its orthogonal V-normal ball of radius $3r_V/4$. The U-ball has relative volume at least $(\rho_U/(2R\sqrt{md_{\mathrm{in}}}))^{md_{\mathrm{in}}}$. The trimmed fiber ensures that the whole normal ball stays in the box. Fubini's theorem gives volume at least

$$c_{\mathrm{pat}} c_{0,V} (2R)^{D-n} v_n (3r_V/4)^n.$$

Each point of this set has residual norm at most $3\sqrt{\varepsilon}/4$. Its nearest grid point has residual norm at most $3\sqrt{\varepsilon}/4 + L_R h \le 7\sqrt{\varepsilon}/8$ and remains in the box. The U rounding displacement is at most $\rho_U/4$, so the rounded point remains in the same pattern ball. The lattice cells of these good points cover the continuous set and each has volume δ^D. Divide by $M \le (3R/\delta)^D$ to obtain the stated constant $C_{D,n}$. No inner-volume estimate for a moving fiber is required. □

Remark 6 (What a Single Cell Contributes)**.** *Lemma 17 closes the fixed-pattern cylinder after resolving V. The prefactor c_{pat} is the volume of an inscribed ball in one activation cell. It is typically exponentially small in md_{in} and is not an overparameterization gain. The factor $c_{0,V}$ records how much of the interpolating V-fiber sits in the box. It may be 1 when every fiber reaches the box faces, and we keep the factor explicit in every bound.*

Assumption 8 (Finite Stable Pattern Class)**.** *Keep Assumption 5. Let $\mathcal{P}$ be a finite indexed family of activation patterns, $|\mathcal{P}| = P$. For each $\alpha \in \mathcal{P}$ assume there is a realizable interpolator $\theta_\alpha = (U_\alpha, V_\alpha)$ in the interior of the box with margin $\gamma_\alpha > 0$ and full-row-rank hidden matrix H_α, satisfying Assumption 6 at that point. Write $\rho_{U,\alpha}$, $c_{\mathrm{pat},\alpha}$, and $c_{0,V,\alpha}$ for the quantities of Lemma 17 and Assumption 7 at θ_α, and write $\mathcal{B}_\alpha$ for the closed U-ball of radius $\rho_{U,\alpha}$ about U_α. Let χ be the overlap multiplicity of these balls, that is, every U lies in at most χ of the balls $\mathcal{B}_\alpha$.*

Theorem 11 (Finite ReLU Pattern Union with Overlap Control)**.** *Under Assumption 8, set $r_V := \min_\alpha \sqrt{\varepsilon}/(2\sigma_{\max}(H_\alpha))$. If every cell satisfies all the mesh and box conditions of Lemma 17, then*

$$p_\varepsilon \ge \frac{1}{\chi}\, C_{D,n} \left(\frac{r_V}{R}\right)^n \sum_{\alpha \in \mathcal{P}} c_{\mathrm{pat},\alpha}\, c_{0,V,\alpha}.$$

If the U-balls are pairwise disjoint then $\chi = 1$ and the prefactor is the sum of the one-cell volumes times V-fiber factors. This beats the single-pattern bound of Lemma 17 by the factor $(\sum_\alpha c_{\mathrm{pat},\alpha} c_{0,V,\alpha})/(\chi\, c_{\mathrm{pat},\alpha_0} c_{0,V,\alpha_0})$ at any reference cell α_0. It does not by itself cancel the exponential-in-md_{in} inscribed-ball penalty.

Proof. Lemma 17 supplies a grid packing of size at least $C_{D,n} c_{\mathrm{pat},\alpha} c_{0,V,\alpha} (r_V/R)^n M$ in each pattern cell, up to the global $C_{D,n}$ conversion of Lemma 6 applied with the same n. Summing over α counts each grid point at most χ times because a point $\theta = (U, V)$ can lie in at most χ of the U-balls. Dividing by χ yields the claim. Disjoint balls are the case $\chi = 1$. □

Theorem 12 (Permutation Orbit of a Separated ReLU Interpolator)**.** *Keep Assumptions 5, 6, and 7 at a single interpolator $(U^\star, V^\star)$. Assume the hidden rows of $U^\star$ are pairwise separated:*

$$\min_{j \ne j'} \|U^\star_{j,\cdot} - U^\star_{j',\cdot}\|_2 > 2\rho_U,$$

with ρ_U as in Lemma 17. Let $\mathcal{P}_{\mathrm{perm}}$ be the $m!$ indexed cell centers obtained by permuting hidden units, with interpolators $(\Pi U^\star, V^\star \Pi^\top)$. Then the corresponding U-balls are pairwise disjoint, so Theorem 11 applies with $P = m!$ and $\chi = 1$. In particular

$$p_\varepsilon \ge C_{D,n}\, m!\, c_{\mathrm{pat}}\, c_{0,V} \left(\frac{r_V}{R}\right)^n.$$

All one-cell box and mesh hypotheses are retained when applying the union theorem. The factorial factor improves this lower bound, but does not by itself guarantee a polynomial prefactor. One may use the conservative value $c_{\mathrm{pat}} = (\rho_U/(2R\sqrt{md_{\mathrm{in}}}))^{md_{\mathrm{in}}}$. Different indexed centers may have the same activation pattern.

Proof. The map $(U, V) \mapsto (\Pi U, V\Pi^\top)$ is a Frobenius isometry and preserves realizability and singular values of H. For $\Pi \ne \Pi'$,

$$\|\Pi U^\star - \Pi' U^\star\|_F = \|(\Pi^{-1}\Pi')U^\star - U^\star\|_F \ge \min_{j \ne j'} \|U^\star_{j,\cdot} - U^\star_{j',\cdot}\|_2 > 2\rho_U,$$

because a non-identity permutation moves at least one row. The closed balls of radius ρ_U are therefore disjoint. Apply Theorem 11 with equal one-cell constants. □

Assumption 9 (Polynomial-Volume Stable ReLU Cover)**.** *Keep Assumption 5. Fix constants $C < \infty$, $\gamma_0 > 0$, and $\Sigma < \infty$. Let Ω_U be a measurable subset of the U-cube $[-R, R]^{md_{\mathrm{in}}}$ with*

$$\frac{\mathrm{vol}(\Omega_U)}{(2R)^{md_{\mathrm{in}}}} \ge m^{-C}.$$

Assume that for every $U \in \Omega_U$ the training activation pattern of U has margin at least γ_0, the hidden matrix $H(U)$ has full row rank N, and $\sigma_{\max}(H(U)) \le \Sigma$. Write $c_{0,V}(U) \in (0,1]$ for the relative volume of the interpolating V-fiber at that U inside $[-R+r_V, R-r_V]^{md_{\text{out}}}$, where $r_V = \sqrt{\varepsilon}/(2\Sigma) \le R/4$, in the sense of Assumption 7, and assume $c_{0,V,*} := \operatorname{essinf}_{U\in\Omega_U} c_{0,V}(U) > 0$.

Theorem 13 (ReLU Packing under a Polynomial-Volume Stable Cover)**.** *Under Assumption 9 set $r_V := \sqrt{\varepsilon}/(2\Sigma)$. If $\delta \le r_V/(4\sqrt{D})$ and the packing is nonvacuous, then*

$$p_\varepsilon \ge C_{D,n}\, c_{0,V,*}\, m^{-C} \left(\frac{r_V}{R}\right)^n$$

with $n = Nd_{\text{out}}$ and the explicit $C_{D,n}$ of Lemma 6. The U-prefactor is polynomial in m. No inscribed-ball comparison is required because Ω_U is packed directly.

Proof. For every $U \in \Omega_U$, add the normal V-ball of radius $3r_V/4$ to the trimmed interpolating fiber. This lies in the box and has residual norm at most $3\sqrt{\varepsilon}/8$. By Fubini, the resulting set has volume at least $m^{-C} c_{0,V,*}(2R)^{D-n} v_n (3r_V/4)^n$. The rounding condition preceding Lemma 17 ensures that every nearest grid point is still good. Its pattern need not equal that of the continuous point, since ReLU is globally Lipschitz. Cover this set by the cells of those good points and divide by $\delta^D M \le (3R)^D$. This proof needs no regularity or erosion assumption on the measurable set Ω_U. □

Corollary 1 (Inscribed-Ball Covers Are a Special Case)**.** *Keep Assumption 8 and write $c_{0,V,*} := \min_\alpha c_{0,V,\alpha}$. If the union of the stable U-balls satisfies*

$$\frac{\operatorname{vol}\big(\bigcup_\alpha \mathcal{B}_\alpha\big)}{(2R)^{md_{\text{in}}}} \ge m^{-C}$$

for an absolute C, then Assumption 9 holds with $\Omega_U = \bigcup_\alpha \mathcal{B}_\alpha$ (up to null sets). Theorem 13 applies under its rounding conditions, using $\Sigma = 2\max_\alpha \sigma_{\max}(H_\alpha)$, so that the common normal radius is no larger than any one-cell radius. The direct-volume argument uses the actual union volume. Replacing each ball by an inscribed cube may give a smaller prefactor.

Proof. Every point of $\mathcal{B}_\alpha$ inherits margin, rank, and $\sigma_{\max}$ control from the center θ_α after shrinking radii as in Lemma 17. The essential infimum of $c_{0,V}$ on the union is at least $c_{0,V,*}$. Apply Theorem 13. □

Lemma 18 (Automatic $\sigma_{\max}$ Bound on the U-Cube)**.** *Under Assumption 5, write $X_1 := \max_i \|x_i\|_1$. For every U in the U-cube $[-R,R]^{md_{\text{in}}}$,*

$$\sigma_{\max}\big(H(U)\big) \le \Sigma_{\text{box}} := \sqrt{Nm}\, R\, X_1.$$

Proof. Each entry of $H(U)$ satisfies $0 \le H(U)_{ij} = \sigma((Ux_i)_j) \le |(Ux_i)_j| \le R\|x_i\|_1$, because every entry of U is at most R in absolute value. Hence $\|H(U)\|_F \le \sqrt{Nm}\, R\, X_1$, which dominates $\sigma_{\max}(H(U))$. □

Lemma 19 (Diluted-Margin Set Has Constant Relative Volume)**.** *Keep Assumption 5 and assume $x_i \ne 0$ for every training index i. Write $x_{\min} := \min_i \|x_i\|_\infty > 0$ and, for $m \ge 1$,*

$$\gamma_0 := \frac{R\, x_{\min}}{2mN}.$$

Let Ω_{marg} be the set of U in the U-cube whose training-set activation margin is at least γ_0. Then

$$\frac{\operatorname{vol}(\Omega_{\text{marg}})}{(2R)^{md_{\text{in}}}} \ge \frac{1}{2}.$$

Proof. Factor the U-cube as a product of m row cubes $[-R,R]^{d_{\text{in}}}$. For a single row u and a single sample x_i, let k be a coordinate attaining $\|x_i\|_\infty$. The slab $\{|\langle u, x_i\rangle| < \gamma_0\}$ has, at every fixed value of the other coordinates, a u_k-width at most $2\gamma_0/\|x_i\|_\infty$. Therefore its relative volume in the row cube is at most $\gamma_0/(Rx_{\min})$. A union bound over the N samples shows that the fraction of a row cube with margin less than γ_0 is at most $N\gamma_0/(Rx_{\min}) = 1/(2m)$. The complementary fraction is at least $1 - 1/(2m)$. Raising to the m-th power and using $(1-1/(2m))^m \ge 1/2$ for every $m \ge 1$ yields the claim. The product structure counts every pattern. No inscribed Frobenius ball is used. □

Assumption 10 (Volume Retention for ReLU V-Fibers)**.** *Keep Assumption 5 and the notation of Lemma 19. Write $Y \in \mathbb{R}^{N\times d_{\text{out}}}$ for the matrix of training targets, assume $\|Y\|_F > 0$, and set*

$$\tau_R := \frac{2\|Y\|_F}{R}.$$

Assume

$$\text{vol}\big(\Omega_{\text{marg}} \cap \{\sigma_{\min}(H) \geq \tau_R\}\big) \geq \tfrac{1}{2}\text{vol}(\Omega_{\text{marg}}).$$

Under this hypothesis every U with $\sigma_{\min}(H(U)) \geq \tau_R$ admits a min-norm interpolating V of Frobenius norm at most $R/2$, because $\|V_{\text{mn}}\|_F \leq \|Y\|_F/\sigma_{\min}(H(U)) \leq R/2$.

Theorem 14 (ReLU Stable Cover from Data, Width, and Volume Retention)**.** *Under Assumption 5, the non-vanishing data hypothesis of Lemma 19, and Assumption 10, the polynomial-volume stable-cover Assumption 9 holds with*

$$\Omega_U := \Omega_{\text{marg}} \cap \{\sigma_{\min}(H) \geq \tau_R\}, \qquad \Sigma = \Sigma_{\text{box}}, \qquad \gamma_0 \text{ as in Lemma 19.}$$

The set Ω_U has relative volume at least $1/4$ (for $m \geq 2$, any C with $m^{-C} \leq 1/4$ is admissible, and for $m = 1$ use the volume bound $1/4$ directly). Moreover

$$c_{0,V,*} \geq v_{\kappa_V}\Big(\frac{1}{8}\Big)^{\kappa_V} > 0, \qquad \kappa_V = (m-N)d_{\text{out}},$$

where v_{κ_V} is the volume of the Euclidean unit ball in dimension κ_V. Consequently, under $r_V \leq R/4$ and all its rounding conditions, Theorem 13 yields

$$p_\varepsilon \geq C_{D,n}\, c_{0,V,*}\, \frac{1}{4}\Big(\frac{r_V}{R}\Big)^n$$

with $r_V = \sqrt{\varepsilon}/(2\Sigma_{\text{box}})$.

Proof. Lemma 19 gives $\text{vol}(\Omega_{\text{marg}})/(2R)^{md_{\text{in}}} \geq 1/2$. Assumption 10 keeps at least half of that volume inside $\{\sigma_{\min} \geq \tau_R\}$, so the displayed Ω_U has relative volume at least $1/4$. This is stronger than a polynomial m^{-C} lower bound. On Ω_U the margin is at least γ_0, so each pattern is frozen. Lemma 18 gives $\sigma_{\max}(H) \leq \Sigma_{\text{box}}$. Full row rank holds because $\sigma_{\min} \geq \tau_R > 0$. For the fiber floor, the min-norm interpolator V_{mn} satisfies $\|V_{\text{mn}}\|_F \leq \|Y\|_F/\tau_R \leq R/2$. The interpolating fiber is the affine flat $V_{\text{mn}}+\{\Delta : \Delta H(U)^\top = 0\}$ of dimension κ_V. The Frobenius ball of radius $R/4$ about V_{mn} in that flat lies in the smaller cube $[-R+r_V, R-r_V]^{md_{\text{out}}}$, because $\|V_{\text{mn}}+\Delta\|_\infty \leq 3R/4 \leq R-r_V$. Its volume is $v_{\kappa_V}(R/4)^{\kappa_V}$, giving $c_{0,V}(U) \geq v_{\kappa_V}(1/8)^{\kappa_V}$. Taking the essential infimum over Ω_U yields the claimed $c_{0,V,*}$. All ingredients of Assumption 9 are therefore verified (with the volume lower bound $1/4$ in place of m^{-C}), and Theorem 13 applies. □

Remark 7 (What Remains If the Box Is Tight)**.** *Lemmas 18 and 19 verify the Σ and polynomial-volume margin parts of Assumption 9 from data and width alone, with no volume-retention hypothesis. The remaining obstruction for a tight box is a uniform positive fiber floor on a polynomial-volume full-rank subset of Ω_{marg}. Inscribed Frobenius balls cannot close that gap by themselves. Even a ball of radius $\Theta(R)$ has relative volume $(O(1/\sqrt{md_{\text{in}}}))^{md_{\text{in}}}$ in the cube.*

Remark 8 (Volume Retention Requires Additional Hypotheses)**.** *Automatic volume retention is false under realizability and $m \geq N$ alone. Take $N = m = d_{\text{in}} = d_{\text{out}} = 1$, $x_1 = 1$, $y_1 = 0.81R^2$, and the interior interpolator $U^\star = V^\star = 0.9R$. Here $\tau_R = 1.62R$, whereas $H(U) = \max\{U, 0\} \leq R$ on the box. Thus the retained set is empty. Theorem 14 remains conditional on volume retention. Identifying additional data, target-size, and box conditions that imply it is an open problem.*

Assumption 11 (Fixed-Mixing Linear Attention)**.** *Let training tokens for each sample i be stacked as a matrix $X_i \in \mathbb{R}^{T\times d_{\text{in}}}$. A linear-attention surrogate with a* fixed *mixing matrix $P \in \mathbb{R}^{T\times T}$ is*

$$f_{W_O,W_V}(X_i) = W_O\big(PX_iW_V\big)^\top.$$

Collect $\theta = (W_O, W_V)$ with $W_O \in \mathbb{R}^{d_{\text{out}}\times d_h}$ and $W_V \in \mathbb{R}^{d_{\text{in}}\times d_h}$. Assume realizability by some interior $\theta^\star$. Softmax attention is a different map and is not included here.

Corollary 2 (Linear Attention Reduces to Two-Layer Deep Linear)**.** *Under Assumption 11, define mixed features $Z_i := PX_i$. After vectorization, f is the two-layer product $W_O W_V^\top$ acting on those features. Theorem 9 applies when its rank, complementary-slice, and tube hypotheses hold, with the data matrix Z formed from $\{Z_i\}$ in place of the original inputs, with exponent $n = N_0 d_{\text{out}}$ for $N_0 = \operatorname{rank}(Z)$, and with gauge dimension d_h^2. The chart-supported tubular prior of Definition 11 and Theorem 20 apply under their chart rank, width, and box hypotheses, because the gauge chart is the two-layer chart of Lemma 8 on the mixed features. Softmax attention is a different map. The Euclidean gauge-split obstruction of Proposition 6 still applies if one places a wide linear-tangent Gaussian on (W_O, W_V) instead of a tubular prior.*

Proof. $W_O(Z_i W_V)^\top = W_O W_V^\top Z_i^\top$. This is the deep linear model of Section 3.2.2 with $L = 2$ and input features Z_i. Realizability and the two-layer gauge chart transfer unchanged. The tubular prior is a statement about that chart and does not use the original unmixed tokens. □

Lemma 20 (Local Full-Rank Tube around the Zero Set)**.** *Let $F : \mathbb{R}^D \to \mathbb{R}^n$ be $C^{1,1}$ on $B_2(\theta^\star, r_0) \subset \mathcal{B}_R$, with $F(\theta^\star) = 0$ and $DF(\theta^\star)$ of full row rank. Let Λ bound the Lipschitz constant of DF, and put $\kappa_{\min} = \sigma_{\min}(DF(\theta^\star))$, $\kappa_{\max} = \sigma_{\max}(DF(\theta^\star))$. Assume*

$$r_0 \le \frac{\kappa_{\min}}{4\Lambda}, \qquad \sqrt{\varepsilon} \le \frac{\kappa_{\min} r_0}{4}.$$

The zero set contains a $C^{1,1}$ graph disk M of dimension $k = D - n$ inside $B_2(\theta^\star, r_0/2)$, parametrized over the radius-$r_0/4$ ball in $\ker DF(\theta^\star)$. Its ambient radius-r neighborhood lies in the good set, where

$$r = \frac{\sqrt{\varepsilon}}{2K}, \qquad K = \kappa_{\max} + \kappa_{\min}/4.$$

For $\delta \le \min\{r/(4\sqrt{D}), R\}$,

$$p_\varepsilon \ge C_{D,n} c_{\text{loc}} (r/R)^n, \qquad c_{\text{loc}} \ge (5/8)^k \left(\frac{r_0}{4R\sqrt{k}}\right)^k.$$

For $k = 0$ the parallel factor is interpreted as 1.

Proof. Translate the zero to the origin and split coordinates as (u, v) in $\ker DF(0)$ and its orthogonal complement. Let A be the restriction of $DF(0)$ to the latter space, so $\|A^{-1}\| = 1/\kappa_{\min}$. For $\|u\| \le r_0/4$, the map $v \mapsto v - A^{-1}F(u, v)$ is a contraction of factor at most $1/4$ on $\|v\| \le r_0/8$. Its value at zero has norm at most $\Lambda\|u\|^2/(2\kappa_{\min}) \le r_0/128$, so it maps that ball into itself. Its unique fixed point $g(u)$ satisfies $\|g(u)\| \le r_0/96$. The implicit-function theorem gives a $C^{1,1}$ graph $M = \{(u, g(u)) : \|u\| \le r_0/4\}$. The derivative of F has norm at most K throughout the larger ball. Since $r \le r_0/8$, every radius-r ball about a graph point stays in that ball and has residual norm at most $Kr = \sqrt{\varepsilon}/2$.

Now use fixed normal coordinates and form $S_0 = \{(u, g(u) + w) : \|u\| \le r_0/4,\ \|w\| \le 3r/4\}$. For $h = \delta\sqrt{D}/2 \le r/8$, each point of S_0 has its $2h$-ball inside the radius-r ball about the same graph point. Thus $S_0 \subset \mathcal{S}_\varepsilon^{-2h}$. Fubini gives its volume exactly as $v_k(r_0/4)^k v_n(3r/4)^n$. The base ball contains the cube of half-side $r_0/(4\sqrt{k})$, so its relative volume is at least $(r_0/(4R\sqrt{k}))^k$, which also implies the stated weaker bound with $(5/8)^k$. Apply Lemma 6. This graph argument does not infer a global injectivity radius from pointwise curvature alone. □

Assumption 12 (Softmax Attention as a C^1 Residual Map)**.** *Let the model contain standard single-head residual attention with scores $S = XW_Q W_K^\top X^\top/\sqrt{d_h}$, possibly followed by an MLP, with trainable weights collected in $\theta \in \mathbb{R}^D$. Assume the training residual map is $C^{1,1}$ on the neighborhood under consideration. Smooth activations with locally bounded second derivatives suffice, as does restricting a ReLU MLP to a stable activation region. Let $F(\theta)$ stack training residuals so that $\mathcal{L}(\theta) = \|F(\theta)\|_2^2$, with $n = N d_{\text{out}}$.*

Assumption 13 (Bounded Pre-Softmax Logits)**.** *Keep Assumption 12 at a realizable interior zero $\theta^\star$ with $DF(\theta^\star)$ of full row rank n. Let T be the number of tokens per sample. Assume every pre-softmax score at $\theta^\star$ is bounded by $B \ge 1$ in absolute value, and let $r_{\text{logit}} > 0$ be small enough that every parameter in $B_2(\theta^\star, r_{\text{logit}})$ has pre-softmax scores bounded by $2B$. Write Λ_B for a Lipschitz constant of DF on that ball, which is an explicit additional bound on the full parameter*

residual map. It depends on data, weight norms, and MLP derivatives as well as softmax derivatives. Bounded logits alone do not bound it. Write $\kappa_{\min} := \sigma_{\min}(DF(\theta^\star))$.

Theorem 15 (Quantitative Local Softmax Tube under a Logit Bound)**.** *Under Assumption 13 set*

$$r_0 := \min\Big\{r_{\text{logit}},\ \frac{\kappa_{\min}}{4\Lambda_B},\ \operatorname{dist}(\theta^\star, \partial\mathcal{B}_R)\Big\}.$$

If $\sqrt{\varepsilon} \le \kappa_{\min} r_0/4$*, then Lemma 20 applies verbatim with this* r_0 *and with* $\Lambda = \Lambda_B$*. The resulting* p_ε *lower bound is a sufficient local tube estimate. This choice of radius is not asserted to be maximal. The bound is not a macroscopic chart of radius* $\Theta(R)$ *unless* $r_0 = \Omega(R/\text{poly})$.

Proof. The explicit $C^{1,1}$ and derivative-bound assumptions give a Lipschitz constant Λ_B on the stated ball. The choice of r_0 is exactly the hypothesis of Lemma 20. These sufficient bounds do not rule out a larger valid chart. □

Assumption 14 (Scale-Compatible Macroscopic Softmax Chart)**.** *Keep Assumption 13. Fix positive finite constants* $C_B, C_\kappa, C_\Lambda, C_{\text{logit}}, C_\partial$*, with* $C_B \ge 1$*, and write*

$$\text{Poly} := (2 + T\, d_{\text{in}}\, d_{\text{out}}\, d_h\, N)^{C_B}.$$

Assume the following four bounds.

1. *Logit height:* $B \le C_B \log \text{Poly}$.
2. *Jacobian floor:* $\kappa_{\min} \ge R/\big(C_\kappa \text{Poly}\big)$.
3. *Jacobian Lip:* $\Lambda_B \le C_\Lambda \text{Poly}$ *as a separate hypothesis on the full residual map.*
4. *Score flatness and interior room:* $r_{\text{logit}} \ge R/\big(C_{\text{logit}} \text{Poly}\big)$ *and* $\operatorname{dist}(\theta^\star, \partial\mathcal{B}_R) \ge R/\big(C_\partial \text{Poly}\big)$.

Item 4 is the geometric content that forces logits to remain $2B$*-bounded on a Euclidean ball of radius* $\Omega(R/\text{poly})$*. It is not implied by item 1 alone when score Lipschitz constants grow with R.*

Theorem 16 (Softmax Macroscopic Inverse-Function-Theorem Radius under Scale-Compatible Bounds)**.** *Under Assumption 14 the radius of Theorem 15 satisfies*

$$r_0 \ge \frac{R}{C_\star\, \text{Poly}^2}$$

with $C_\star := 4\max\{C_{\text{logit}}, C_\partial, 4C_\kappa C_\Lambda\}$*. If in addition* $\sqrt{\varepsilon} \le \kappa_{\min} r_0/4$*, then Lemma 20 yields*

$$p_\varepsilon \ge C_{D,n}\, c_{\text{loc}} \Big(\frac{r}{R}\Big)^n, \qquad c_{\text{loc}} \ge \Big(\frac{5}{8}\Big)^{D-n} \Big(\frac{r_0}{4R\sqrt{D-n}}\Big)^{D-n},$$

so the chart radius is an inverse-polynomial fraction of R. The volume prefactor raises this fraction to dimension $D - n$ *and can still be exponentially small in that dimension. Linear attention with fixed mixing is not part of this theorem. It is Corollary 2.*

Proof. Theorem 15 defines

$$r_0 = \min\Big\{r_{\text{logit}},\ \frac{\kappa_{\min}}{4\Lambda_B},\ \operatorname{dist}(\theta^\star, \partial\mathcal{B}_R)\Big\}.$$

The first and third arguments are at least $R/(C_{\text{logit}}\text{Poly})$ and $R/(C_\partial\text{Poly})$. The middle argument is at least $R/(4C_\kappa C_\Lambda \text{Poly}^2)$. Since $\text{Poly} \ge 1$, the stated common lower bound follows. The packing claim is Lemma 20. □

Proposition 2 (Bilinear Score Obstruction on Large Boxes)**.** *Keep Assumption 12 with bilinear self-attention scores* $xW_QW_K^\top x^\top$ *(or these scores divided by* $\sqrt{d_h}$*). Assume the diagonal score of the token used below is among the bounded scores. Assume some training token* x *satisfies* $\|x\|_2 \ge x_* > 0$*. Let* $\theta^\star$ *be any realizable zero whose pre-softmax scores are at most* B *in absolute*

value, and let r_{logit} be the radius of a Euclidean ball about $\theta^\star$ on which every pre-softmax score stays at most $2B$ in absolute value. Then

$$r_{\mathrm{logit}} \le \sqrt{6}\,\frac{d_h^{1/4}\sqrt{B}}{x_*}.$$

If item 1 of Assumption 14 holds, then $B = O(\log \mathrm{Poly})$ and $r_{\mathrm{logit}} = O(\sqrt{\log \mathrm{Poly}})$. The lower bound $r_{\mathrm{logit}} \ge R/(C_{\mathrm{logit}}\mathrm{Poly})$ in item 4 then fails for every fixed choice of the constants once R is large, because Poly *does not grow with R.*

Proof. Write $s(W_Q, W_K)$ for the pre-softmax score on the token x, including the factor $1/\sqrt{d_h}$ when that normalization is present. Let $\hat{u} := x^\top/\|x\|_2$ and let $e \in \mathbb{R}^{d_h}$ be a unit vector. The path $W_Q(t) = W_Q^\star + t\hat{u}e^\top, W_K(t) = W_K^\star + t\hat{u}e^\top$ changes (W_Q, W_K) at Frobenius speed $\sqrt{2}$ and leaves every other parameter fixed. Expanding the bilinear score along this path gives $s(t) = s(0) + tL + t^2 a$ with $a \ge x_*^2/\sqrt{d_h}$ (the denominator is 1 when the score is unnormalized, which only increases a). Replacing t by $-t$ flips the sign of L and preserves the quadratic term. Choose the sign so that $tL \ge 0$ for $t > 0$. Then $s(t) - s(0) \ge t^2 a$. At displacement $r = t\sqrt{2}$ the increase is at least $r^2 a/2$. The center has $|s(0)| \le B$, and the ball keeps $|s| \le 2B$, so the increase is at most $3B$. Therefore $r_{\mathrm{logit}}^2 x_*^2/(2\sqrt{d_h}) \le 3B$, which is the displayed bound. Item 1 replaces B by $O(\log \mathrm{Poly})$. The resulting upper bound stays independent of R, while item 4 asks for a radius that grows linearly with R. □

Assumption 15 (Polynomial Search Box with Quantitative Local Bounds)**.** *Keep Assumption 13. Let $P \ge 2$ be a fixed polynomial in the widths, token count, and sample count, with $R \le P$. For a constant $a \ge 1$ independent of these dimensions, assume*

$$\kappa_{\min} \ge P^{-a}, \qquad \Lambda_B \le P^a, \qquad r_{\mathrm{logit}} \ge P^{-a}, \qquad \mathrm{dist}(\theta^\star, \partial\mathcal{B}_R) \ge P^{-a}.$$

These quantitative assumptions are stronger than mere positivity of local constants and include control of the full residual derivatives.

Theorem 17 (Softmax Macroscopic Inverse-Function-Theorem Tube on a Polynomial Box)**.** *Under Assumption 15, the radius of Theorem 15 satisfies*

$$r_0 \ge \frac{1}{4P^{2a}} \ge \frac{R}{4P^{2a+1}}.$$

Under its loss-tolerance and mesh conditions, Lemma 20 therefore supplies the displayed local packing bound with a chart radius that is an inverse-polynomial fraction of the search radius. It does not supply a polynomial lower bound on the volume prefactor when the chart dimension grows.

Proof. The three entries in the minimum defining r_0 are bounded below by P^{-a}, $1/(4P^{2a})$, and P^{-a}, respectively. Since $P \ge 2$ and $a \ge 1$, their minimum is at least $1/(4P^{2a})$. Use $R \le P$ for the second inequality and apply the local-tube lemma. □

Conjecture 1 (Scale-Invariant Softmax beyond Polynomial Boxes)**.** *Lemma 20 and Theorem 15 give a local tube under the stated derivative bounds. Theorem 17 gives an inverse-polynomial radius when the search box and all the local derivative and nondegeneracy scales obey quantitative polynomial bounds. Proposition 2 blocks the same conclusion for vanilla bilinear scores as $R \to \infty$. The remaining gap is an architecture with scale-invariant scores (for example cosine attention or LayerNorm before the QK product) for which r_{logit} and $\kappa_{\min}$ stay $\Omega(R/\mathrm{poly})$ on large boxes. Linear attention with fixed mixing is not part of this conjecture. It is Corollary 2.*

B.1.4 Compositions of Analyzed Blocks

Real architectures stack the blocks analyzed above, and this subsubsection records the two composition statements that the ladder supports without new geometric input, together with the boundaries of what composition gives. The first statement says that gauge freedom acts within each block of a residual linear stack, so the exact-zero orbits of the blocks multiply.

Lemma 21 (Blockwise Gauge Action through a Residual Linear Stack)**.** *Let the network be a residual stack of B linear blocks, where block b applies $x \mapsto x + A_b x$ with $A_b := W_{b,L_b} \cdots W_{b,1}$ a product of factors $W_{b,\ell} \in \mathbb{R}^{d_{b,\ell} \times d_{b,\ell-1}}$ satisfying $d_{b,0} = d_{b,L_b} = d$ for a common residual width d, so the end-to-end map is $(I + A_B) \cdots (I + A_1)$. For any invertible $G_{b,\ell} \in \mathrm{GL}(d_{b,\ell})$, $\ell = 1, \ldots, L_b - 1$, with $G_{b,0} = G_{b,L_b} = I$, the replacement*

$$W_{b,\ell} \mapsto G_{b,\ell}^{-1} W_{b,\ell} G_{b,\ell-1}$$

leaves every A_b, and hence the end-to-end map and the training loss, exactly unchanged. If moreover the differential of each block gauge action is injective at a realizable zero $\theta^\star$, as happens for two-layer blocks under the end-factor rank conditions of Lemma 8, or for deeper blocks under the full hypotheses of Lemma 15, then the exact-zero set of the stack contains a product of per-block gauge orbits through $\theta^\star$ of total dimension

$$\sum_{b=1}^{B} \sum_{\ell=1}^{L_b - 1} d_{b,\ell}^2.$$

Proof. Inside block b the replacement telescopes, because the product reads

$$G_{b,L_b}^{-1} W_{b,L_b} G_{b,L_b-1} \cdot G_{b,L_b-1}^{-1} W_{b,L_b-1} G_{b,L_b-2} \cdots G_{b,1}^{-1} W_{b,1} G_{b,0} = W_{b,L_b} \cdots W_{b,1} = A_b,$$

using $G_{b,0} = G_{b,L_b} = I$. The composite network is the function $x \mapsto (I + A_B) \cdots (I + A_1)x$, which depends on the parameters only through the block maps $(A_1, \ldots, A_B)$, so any joint choice of the gauge matrices, across all blocks at once, preserves the end-to-end map and hence the loss on the training set. For the dimension count, the blocks occupy disjoint parameter coordinates, so the differential of the joint gauge action at $\theta^\star$ is block-diagonal with diagonal entries the per-block gauge differentials, and its rank is the sum of the per-block ranks. Under the per-block injectivity hypothesis each per-block rank equals the full gauge dimension, which is the two-layer computation of Lemma 8 (the differential $(A^\star H, -HB^\star)$ has trivial kernel under the end-factor rank conditions), so the joint orbit dimension is the displayed sum. □

The second statement says that an attention block followed by an MLP inherits the ReLU guarantee on the slice where the attention map vanishes, at the price of an explicit slice factor.

Proposition 3 (Attention-Plus-MLP Inheritance on the Zero-Attention Slice)**.** *Let the composite block be $f_\theta(x) = V\,\sigma\big(U(x + A_\theta x)\big)$ on training inputs $\{x_i\}_{i=1}^N$, where the attention part A_θ is bilinear in trained factors (W_O, W_V) and satisfies $A_\theta = 0$ whenever $W_O = 0$, as in fixed-mixing linear attention, whose single-token form is $A_\theta = W_O W_V^\top P$ with fixed P. Let the prior factorize across the (U, V)-block, W_O, and W_V, with the (U, V)-part of mass at least $m_0 > 0$ on the single-block ReLU ε-good set (for example $m_0 = 1/8$ under Theorem 21), and with the W_O-part a discrete Gaussian of width $\sigma_O \geq \delta$ centered at 0. Then*

$$p_\varepsilon^{\mathrm{prior}} \;\geq\; m_0 \left(\frac{\delta}{4\sigma_O}\right)^{d_O}, \qquad d_O := d_{\mathrm{in}}\, d_h,$$

where $W_O \in \mathbb{R}^{d_{\mathrm{in}} \times d_h}$ maps the attention output back to the input width required by the residual sum, d_O is its number of entries, and on the zero-attention slice the composite loss equals the single-block loss, so the certificate of the (U, V)-part carries over unchanged.

Proof. If W_O sits at the zero grid point, then $A_\theta = 0$ and $f_\theta(x) = V\sigma(Ux)$, so the composite training loss coincides with the single-block ReLU loss on the same targets, and every prior outcome whose (U, V)-block lies in the single-block good set is ε-good for the composite, regardless of W_V. The prior factorizes, so the mass of this event is the product of the (U, V)-mass, which is at least m_0, the W_V-mass, which is 1, and the W_O-point mass at zero. For one coordinate, the box-truncated discrete Gaussian point mass at 0 is at least $1/S$, where $S := \sum_{k \in \mathbb{Z}} e^{-(k\delta)^2/(2\sigma_O^2)}$. The summand $f(t) := \exp(-(t\delta)^2/(2\sigma_O^2))$ is even and decreasing on $[0, \infty)$. For every integer $k \geq 1$, $f(k) \leq \int_{k-1}^{k} f(t)\,dt$, so the positive indices contribute at most $\int_0^\infty f(t)\,dt$. The negative indices contribute the same amount. The half-line integral equals $\sigma_O\sqrt{2\pi}/(2\delta)$, and therefore

$$S \leq 1 + \frac{\sigma_O \sqrt{2\pi}}{\delta}.$$

The hypothesis $\sigma_O \geq \delta$ gives $1 \leq \sigma_O/\delta$, so $S \leq (1+\sqrt{2\pi})\sigma_O/\delta$. Since $1+\sqrt{2\pi} \leq 4$, one has $S \leq 4\sigma_O/\delta$. The per-coordinate mass is at least $\delta/(4\sigma_O)$, and the d_O coordinates give the displayed factor. □

Remark 9 (What Composition Does Not Give)**.** *The two statements above are reductions, not free lunches. The slice factor of Proposition 3 decays exponentially in the attention output size d_O, so the inheritance guarantee is a statement about a promised slice rather than about typical parameters, and widening the analysis to participating attention is open. Constant prior mass for a composed architecture does not follow from constant mass per block, because the loss of a composition couples the blocks, and stacks of three or more ReLU layers are not covered by the pattern-cell analysis, which freezes one pattern of a single hidden layer. Softmax attention inside a composition inherits neither statement, since its block map is neither a product of linear factors nor piecewise linear in the frozen-pattern sense used here.*

Proposition 4 (Nonlinear Checklist)**.** *The status of the nonlinear blocks splits into two open directions, one open width regime, and a proved list, and we record them together so that the proved-versus-open boundary of the appendix is explicit.*

The open side is short.

1. *Sufficient additional conditions for ReLU volume retention (Remark 8). Realizability and width alone do not imply it, so Theorem 14 remains conditional.*
2. *Scale-invariant softmax scores on super-polynomial boxes (Conjecture 1). Settling it lets the macroscopic inverse-function-theorem tube of Theorem 16 apply beyond Theorem 17.*
3. *Factor-wise $\kappa^{(L)}$ for multimodal zigzag widths (Remark 5) is the remaining open width regime.*

On the proved side, the ReLU block contains the fixed-pattern cylinder with an explicit V-fiber factor $c_{0,V}$, finite-pattern unions with overlap multiplicity χ, and permutation-orbit unions, together with the cell-supported ReLU prior of constant mass (Theorem 21), ReLU packing under a polynomial-volume stable cover (Theorem 13), the diluted-margin volume and the automatic Σ_{box} bound (Lemmas 19 and 18), and the stable-cover theorem under volume retention (Theorem 14). The softmax block contains local full-rank tubes, the bounded-logit local tube, the macroscopic tube under scale-compatible bounds (Theorem 16), the macroscopic inverse-function-theorem tube on polynomial boxes (Theorem 17), and the bilinear large-box obstruction (Proposition 2). Linear attention with fixed mixing, and the chart-supported tubular prior for two-layer and three-layer deep linear models, are proved as well, as is the tubular prior for $L > 3$ under the four width regimes (Corollary 6). The composition reductions are proved on their stated slices, namely the blockwise gauge action through residual linear stacks (Lemma 21) and the attention-plus-MLP inheritance (Proposition 3, with the boundaries in Remark 9). Factor-wise $\kappa^{(L)}$ is proved for equal, monotone funnel, and unimodal-valley hidden widths (Lemma 15).

Closing either direction still requires settling the corresponding open problem, and then a comparison of $1/\sqrt{p_\varepsilon}$ against $T_{\mathrm{SGD}}(\varepsilon)$. No unconditional win against SGD is claimed.

C Full Prior-Mass Proofs

C.1 Prior Mass under Alignment

Proposition 11 warns that a Gaussian prior does not automatically yield polynomial good-set mass. This subsection proves the positive results under explicit alignment, together with a dichotomy that clarifies the isotropic case. The blocks follow the ladder of the main text. For the single linear layer we first treat the isotropic Gaussian, which gives constant mass under strong alignment and collapses super-polynomially outside a narrow width window, with a transition window not covered by these sufficient bounds. We then split the width across $\ker(X)^\perp$ and $\ker(X)$, which keeps the constant mass and adds macroscopic exploration of the interpolating chart. For two-layer deep linear models we show that the Euclidean analogue of the split prior is not constant-mass, and we construct the chart-supported tubular prior that is. For ReLU networks we place the split prior inside one frozen activation pattern cell.

C.1.1 Isotropic Gaussian Priors for Linear Least Squares

We begin with the most familiar prior family and record the alignment conditions that make it work.

Assumption 16 (Aligned Isotropic Prior)**.** *Keep the linear instance of Section 3.2.1 and* $r := \sqrt{\varepsilon}/\sigma_{\max}$*. Let* $\theta_0 \in \mathcal{B}_R$ *satisfy*

$$\operatorname{dist}(\theta_0, \theta^\star + \ker(X)) \le \frac{r}{2},$$

so the center of the prior lies within half the tube radius of the affine solution set. Let the prior be the discrete Gaussian of Definition 4 with width $\sigma > 0$*, and assume the mesh obeys*

$$\delta \le \min\Big\{\frac{\sigma}{8d}, \frac{r}{4\sqrt{d}}\Big\}.$$

Assume the Euclidean typical ball of radius $4\sigma\sqrt{d}$ *about* θ_0 *sits inside* $\mathcal{B}_R$*.*

Three elementary estimates carry every mass computation in this subsection. The first controls the Gaussian tail outside a typical ball.

Lemma 22 (Typical-Ball Gaussian Tail)**.** *Let* $g \sim \mathcal{N}(0, \sigma^2 I_d)$*. Then*

$$\Pr\big(\|g\|_2 > 4\sigma\sqrt{d}\big) \le e^{-d}.$$

Proof. Write $g = \sigma\xi$ with $\xi \sim \mathcal{N}(0, I_d)$, so the event is $\|\xi\|_2^2 > 16d$. A standard χ^2 tail bound gives $\Pr(\|\xi\|_2^2 \ge d + 2\sqrt{dt} + 2t) \le e^{-t}$. Taking $t = d$ yields deviation $d + 2d + 2d = 5d < 16d$, so the probability of a $16d$ deviation is at most e^{-d}. □

The second transfers mass between the continuous Gaussian and its discretization on the grid.

Lemma 23 (Discrete Gaussian Mass Transfer)**.** *Let* $\mu = \mathcal{N}(\theta_0, \sigma^2 I_d)$ *and let* π_δ *be its point-weight discretization on* $\mathcal{G}_{\delta,R}$*. Put* $h = \delta\sqrt{d}/2$ *and* $B_{\mathrm{typ}} = B_2(\theta_0, 4\sigma\sqrt{d})$*. If* $\delta \le \min\{\sigma/(8d), R\}$*, then every measurable* $A \subset \mathcal{B}_R$ *satisfies*

$$\pi_\delta(A \cap \mathcal{G}_{\delta,R}) \ge \frac{3}{5}\mu(A^{-h} \cap B_{\mathrm{typ}}) \ge \frac{1}{2}\mu(A^{-h} \cap B_{\mathrm{typ}}).$$

Here $A^{-h} = \{x : B_2(x, h) \subset A\}$*. The same comparison holds on the untruncated lattice.*

Proof. Write f for the normalized Gaussian density. If $x \in B_{\mathrm{typ}}$ and θ is its nearest lattice point, then

$$|\log f(x) - \log f(\theta)| \le 4\sqrt{d}\, h/\sigma + h^2/(2\sigma^2) < 1/3.$$

For a one-dimensional Gaussian density f_j, the Riemann-sum error is at most $\delta \int |f_j'| = 2\delta/(\sigma\sqrt{2\pi})$. Consequently the full-lattice partition function satisfies

$$\delta^d \sum_{\theta \in \delta\mathbb{Z}^d} f(\theta) \le \left(1 + \frac{2\delta}{\sigma\sqrt{2\pi}}\right)^d \le e^{1/(4\sqrt{2\pi})}.$$

The box partition function is no larger. Cells meeting $A^{-h} \cap B_{\mathrm{typ}}$ have their centers in $A \cap \mathcal{B}_R$. Integrating over those cells gives a numerator at least $e^{-1/3}\mu(A^{-h} \cap B_{\mathrm{typ}})$ after multiplication by δ^d. Division proves the result, since $e^{-1/3-1/(4\sqrt{2\pi})} > 3/5$. □

The third is a Chebyshev lower tail for the norm of a standard Gaussian vector, which prices the probability that the normal fluctuation stays small.

Lemma 24 (Chi Lower Tail via Chebyshev)**.** *Let* $\xi \sim \mathcal{N}(0, I_N)$*. Then for every* $N \ge 1$*,*

$$\Pr\big(\|\xi\|_2^2 > \tfrac{3}{2}N\big) \le \frac{8}{N}.$$

Equivalently,

$$\Pr\big(\|\xi\|_2 \le \sqrt{\tfrac{3}{2}N}\big) \ge 1 - \frac{8}{N}.$$

Proof. We have $\mathbb{E}[\|\xi\|_2^2] = N$ and $\mathrm{Var}(\|\xi\|_2^2) = 2N$. Chebyshev gives

$$\Pr\big(|\|\xi\|_2^2 - N| \geq \tfrac{1}{2}N\big) \leq \frac{2N}{(N/2)^2} = \frac{8}{N},$$

and the event $\|\xi\|_2^2 > \frac{3}{2}N$ is contained in that deviation event. □

The three estimates assemble into the constant-mass statement for the isotropic prior.

Theorem 18 (Constant Prior Mass under Strong Isotropic Alignment)**.** *Keep all conditions of Assumption 16, and choose*

$$\sigma \leq \frac{r}{\sqrt{18N}}.$$

Also retain the typical-ball-in-box condition of Assumption 16. Then for all $N \geq 16$,

$$p_\varepsilon^{\mathrm{prior}} \geq \frac{1}{4}.$$

Consequently Theorem 1 uses only $O(1)$ loss-oracle queries.

Proof. Write continuous Gaussian measure $\mu = \mathcal{N}(\theta_0, \sigma^2 I_d)$ and decompose fluctuations into the $\ker(X)$ and $\ker(X)^\perp$ components. If $g \sim \mathcal{N}(0, \sigma^2 I_d)$, the normal component is distributed as $\sigma\xi$ with $\xi \sim \mathcal{N}(0, I_N)$ after an orthonormal identification of $\ker(X)^\perp$. The alignment condition places the mean within $r/2$ of the affine solution set. On the event $\|\xi\|_2^2 \leq 2N$ one has $\|\sigma\xi\|_2 \leq r/3$, because $\sigma\sqrt{2N} \leq r\sqrt{2/18} = r/3$. Chebyshev uses $\mathrm{Var}(\|\xi\|_2^2) = 2N$ and deviation N, so

$$\mu\big(\{\|P_\perp(g)\|_2 \leq r/3\}\big) \geq 1 - \frac{2}{N} \geq \frac{7}{8} \qquad (N \geq 16).$$

Intersect this event with $\|g\|_2 \leq 3\sigma\sqrt{d}$. The chi-square bound used in Lemma 22 also bounds the excluded probability by e^{-d}, since $9d > 5d$. The radius-$4\sigma\sqrt{d}$ box assumption leaves an h margin on this smaller ball. Mean distance, normal fluctuation, and rounding margin sum to at most $r/2 + r/3 + r/8 < r$. Thus the intersection is in $\mathcal{S}_\varepsilon^{-h}$. The intersection therefore has continuous mass at least $7/8 - e^{-16}$. Lemma 23 transfers at least half of that mass onto the discrete grid. Half of $7/8 - e^{-16}$ is at least $1/4$, since $3/16 > (1/2)e^{-16}$. The typical-ball-in-box hypothesis of Theorem 2 keeps the comparison inside $\mathcal{B}_R$. □

The same machinery also shows where the isotropic family stops working, and the stated constants and truncation conditions matter.

Proposition 5 (Narrow and Broad Isotropic Regimes)**.** *Under Assumption 16, widths $\sigma \leq r/\sqrt{18N}$ give mass at least $1/4$ for $N \geq 16$. For the broad-width regime assume in addition $d = O(N)$ and bounded $\sigma_{\max}(X)/\sigma_{\min}(X)$. If $\sigma\sqrt{N}/r \to \infty$, then the discrete good-set mass is smaller than every fixed negative power of N. These two sufficient regimes do not exclude polynomially small mass in the transition window $\sigma = \Theta(r/\sqrt{N})$.*

Proof. The narrow-width assertion is Theorem 18. Put $b = \sqrt{\varepsilon}/\sigma_{\min}(X)$ and $h = \delta\sqrt{d}/2$. The good set is contained in the normal tube of radius b, by Lemma 1. For a cell centered at θ,

$$e^{-\|\theta-\theta_0\|^2/(2\sigma^2)} \leq e^{h^2/(2\sigma^2)} e^{-\|x-\theta_0\|^2/(4\sigma^2)}.$$

The cell union of good points lies in the normal tube of radius $b + h$. Its mass under $\mathcal{N}(\theta_0, 2\sigma^2 I_d)$ is at most $(e(b+h)^2/(2N\sigma^2))^{N/2}$, by normal-ball volume times the maximum projected Gaussian density. Meanwhile the discrete partition function, multiplied by δ^d and normalized as a density, is at least $e^{-1/3}(1 - e^{-d})$: integrate over the radius-$3\sigma\sqrt{d}$ ball, whose nearest lattice points lie in the box. Hence

$$p_\varepsilon^{\mathrm{prior}} \leq 3\,2^{d/2} \left(\frac{e(b+h)^2}{2N\sigma^2}\right)^{N/2}.$$

Here $h/\sigma \leq 1/(16\sqrt{d})$ and $b = O(r)$. Since $d = O(N)$ and $(b+h)/(\sigma\sqrt{N}) \to 0$, the right side is $o(N^{-c})$ for every fixed $c > 0$. □

Remark 10 (Leveled Reading of the Constant-Mass Regime)**.** *When Theorem 2 applies, both measurement as training and classical rejection sampling succeed with constant probability, so the quadratic gap is constant versus constant. Same-prior SGD is likewise easy, because the initialization already lies in an $O(r)$ normal neighborhood of the interpolating set. The comparison is therefore leveled, and it does not by itself exhibit a quantum advantage over SGD.*

C.1.2 Anisotropic Spectral Priors

The broad-width bound limits simultaneous loss control and exploration by one isotropic width. The fix is to put different scales on the normal and parallel spaces, which is exactly the freedom that a spectral prior provides.

Definition 9 (Split Spectral Prior for Linear Models)**.** *Let $P_\perp$ project onto $\ker(X)^\perp$ and $P_\parallel$ onto $\ker(X)$. A split spectral prior has density (up to grid normalization) proportional to*

$$\exp\Big(-\frac{\|P_\perp(\theta-\theta_0)\|_2^2}{2\sigma_\perp^2}-\frac{\|P_\parallel(\theta-\theta_0)\|_2^2}{2\sigma_\parallel^2}\Big),$$

with a narrow width $\sigma_\perp$ on the loss-sensitive directions and a wide width $\sigma_\parallel$ along the interpolating kernel. The box $\mathcal{B}_R$ and the grid $\mathcal{G}_{\delta,R}$ are taken in an orthonormal frame that identifies $\ker(X)^\perp$ and $\ker(X)$ with coordinate subspaces. In that frame the prior is a product discrete Gaussian on the product mesh $\delta\mathbb{Z}^d\cap[-R,R]^d$. Singular values of the design are unchanged by this orthogonal change of parameter coordinates.

The mass-transfer lemma extends to product Gaussians with arbitrary per-coordinate widths, which covers the split prior after rotating into its eigenbasis.

Lemma 25 (Product Discrete Gaussian Mass Transfer)**.** *In an orthonormal coordinate frame let $\mu=\mathcal{N}(\theta_0,\mathrm{diag}(\sigma_1^2,\dots,\sigma_d^2))$ and let π_δ be its point-weight discretization on the product grid in $[-R,R]^d$. Write $h=\delta\sqrt{d}/2$ and*

$$E=\left\{x:\sum_i(x_i-\theta_{0,i})^2/\sigma_i^2\le 16d\right\}.$$

If $\delta\le\min\{\min_i\sigma_i/(8d),R\}$, every measurable $A\subset\mathcal{B}_R$ satisfies

$$\pi_\delta(A\cap\mathcal{G}_{\delta,R})\ge\frac{3}{5}\mu(A^{-h}\cap E)\ge\frac{1}{2}\mu(A^{-h}\cap E).$$

The same assertion holds on the untruncated product lattice.

Proof. Whitening gives a cell radius at most $1/(16\sqrt{d})$, so the log-density comparison with a nearest grid point is less than $1/3$ on E. For the normalization, apply the one-dimensional Riemann-sum bound from Lemma 23 to each width. The product is at most $\exp(\sum_i 2\delta/(\sigma_i\sqrt{2\pi}))\le e^{1/(4\sqrt{2\pi})}$. The same cell integration proves the claim. This argument bounds the complete partition function and does not extrapolate a typical-set density estimate into the tails. □

The split prior now gives constant mass and, unlike the isotropic prior, keeps exploring the chart after conditioning.

Theorem 19 (Constant Mass with Parallel Exploration)**.** *Keep the alignment conditions on the center θ_0 from Theorem 2, and choose*

$$\sigma_\perp\le\frac{r}{\sqrt{18N}}\qquad\text{and}\qquad\sigma_\parallel\in\Big[\frac{R}{4},R\Big].$$

In the adapted frame, assume the normal-coordinate ball of radius $4\sigma_\perp\sqrt{N}$ about $\theta_{0,\perp}$ lies in $[-R,R]^N$. No typical-ball condition is imposed on the wide kernel block. Then for $N\ge16$ and

$$\delta\le\min\Big\{\frac{\sigma_\perp}{8d},\frac{\sigma_\parallel}{8d},\frac{r}{4\sqrt{d}}\Big\},$$

$$p_\varepsilon^{\mathrm{prior}}\ge\frac{1}{4}.$$

If $d>N$, conditioning on $\mathcal{S}_\varepsilon^\delta$ leaves the kernel marginal unchanged, with coordinatewise exploration on scale R.

Proof. The product grid and box make the truncated normal and kernel blocks independent. The predicate depends only on the normal block. Apply Theorem 18 in dimension N, using the stated normal box margin and normal mesh. The full prior therefore has the same good mass, at least $1/4$. Conditioning on success leaves the kernel marginal exactly unchanged. If $d > N$, each kernel coordinate has width in $[R/4, R]$ and center in $[-R, R]$. Its truncated discrete Gaussian assigns a positive, dimension-independent mass to intervals of length proportional to R inside that coordinate interval. Thus exploration is retained coordinatewise on scale R. For $d = N$ there is no kernel exploration assertion. □

Remark 11 (What the Split Prior Buys)**.** *Theorem 19 separates two roles that isotropic priors entangle: (i) putting enough mass on ε-good parameters, and (ii) exploring many global interpolators along the degenerate chart. Relative to same-prior rejection sampling, Theorem 1 still gives a quadratic improvement whenever the success probability is a small constant (or smaller under weaker alignment). Relative to SGD, advantage remains conditional on Proposition 10. The structural gain that does not require beating SGD is Theorem 1, namely that repeated samples probe the distribution of global ε-solutions along the chart.*

Corollary 3 (Query Bound under Aligned Priors)**.** *Whenever $p_\varepsilon^{\mathrm{prior}} \geq 1/4$ by Theorem 18 or Theorem 19,*

$$T_Q(\varepsilon) = O(1)$$

loss-oracle calls suffice for constant success probability, up to the preparation cost T_{prep}.

C.1.3 Split Spectral Priors for Two-Layer Deep Linear Models

The linear analysis places one narrow width on every loss-sensitive direction and a wide width on the flat solution kernel. In a two-layer deep linear model the solution set is a curved gauge orbit rather than a flat affine set, and the following negative result shows that the naive Euclidean analogue of the split prior fails. This failure is what forces the tubular construction of the next subsubsection.

Assumption 17 (Aligned Two-Layer Center)**.** *Keep the deep-linear assumptions of Section 3.2.2 with $L = 2$ and the data rank N_0 from Lemma 10. Let $\theta^\star = (A^\star, B^\star)$ be a realizable zero-loss factor pair, and let θ_0 lie within Euclidean distance $r/2$ of the gauge orbit of $\theta^\star$ from Lemma 8, where $r := c_r\sqrt{\varepsilon}/\mathrm{Lip}$ as in Theorem 9.*

Definition 10 (Gauge-Split Prior)**.** *At θ_0, split tangent directions into*

1. *gauge / fiber directions spanning the local d_1^2-dimensional orbit,*
2. *data-normal directions spanning an at most $N_0 d_{\mathrm{out}}$-dimensional space controlling the loss to first order,*
3. *residual ambient directions.*

A gauge-split spectral prior uses widths $(\sigma_{\mathrm{gauge}}, \sigma_{\mathrm{normal}}, \sigma_{\mathrm{res}})$ on these blocks (and is discretized on $\mathcal{G}_{\delta,R}$).

Proposition 6 (Euclidean Two-Layer Gauge-Split Is Not Constant-Mass)**.** *Under Assumption 17, the Euclidean block-Gaussian of Definition 10 with*

$$\sigma_{\mathrm{normal}} \leq \frac{r}{\sqrt{12\, N_0 d_{\mathrm{out}}}}, \qquad \sigma_{\mathrm{gauge}} \in \left[\tfrac{R}{4}, R\right], \qquad \sigma_{\mathrm{res}} \leq \sigma_{\mathrm{normal}},$$

and $\delta \leq \min\{\sigma_{\mathrm{normal}}, \sigma_{\mathrm{gauge}}\}/(8D)$, does not imply

$$p_\varepsilon^{\mathrm{prior}} \geq \frac{1}{4}.$$

Proof. A scalar counterexample suffices. Take $A^\star = B^\star = 1$, one training input $x = 1$ with target $y = 1$, and $R = 4$. Write the orthonormal tangent and normal displacements as

$$A = 1 + (s + t)/\sqrt{2}, \qquad B = 1 + (s - t)/\sqrt{2}.$$

Then $AB - 1 = \sqrt{2}s + (s^2 - t^2)/2$. Take tangent width $\sigma_t = R/4 = 1$, normal width $\sigma_s = r/\sqrt{12}$, and $\delta = \varepsilon^2$ for sufficiently small ε. These satisfy the stated bounds. The box-truncated discrete

prior converges, as $\varepsilon \downarrow 0$, to a continuous truncated Gaussian on the line $s = 0$: this follows from Riemann sums in the scaled coordinates $(s/\sigma_s, t)$, since $\delta/\sigma_s \to 0$. On that line the residual is $-t^2/2$, which vanishes only at $t = 0$, a set of limiting probability zero. For any $a > 0$, outside $|t| < a$ the residual stays bounded away from zero as $s \to 0$. First take the small-ε limit and then $a \downarrow 0$ to obtain $p_\varepsilon^{\text{prior}} \to 0$. Thus the hypotheses do not imply the proposed $1/4$ bound. This counterexample does not assert a quadratic lower curvature bound in every gauge direction of every factorization. □

Remark 12 (Certificate Remains Full-Batch). *Proposition 6 concerns prior mass only. When a two-layer prior is supported on a tubular neighborhood of the gauge chart, the returned sample is still certified by the hard full-batch indicator, and that construction is Theorem 20. The tubular theorem accompanies the main-text analysis Theorem 3 rather than rewriting it, and Proposition 6 explains why the detour through the tube is necessary.*

The same curvature obstruction reaches the three-layer chart.

Corollary 4 (Three-Layer Euclidean Gauge-Split Is Not Constant-Mass). *Under Assumption 4 and the three-layer chart of Lemma 13, form a Euclidean gauge-split prior as in Definition 10 with gauge block of dimension $k = d_1^2 + d_2^2$, data-normal block of dimension at most $n = N_0 d_{\text{out}}$, and residual block. If*

$$\sigma_{\text{normal}} \le \frac{r}{\sqrt{12n}}, \qquad \sigma_{\text{gauge}} \in \left[\tfrac{R}{4}, R\right], \qquad \sigma_{\text{res}} \le \sigma_{\text{normal}},$$

with $r = c_r\sqrt{\varepsilon}/\text{Lip}$ and $\delta \le \min\{\sigma_{\text{normal}}, \sigma_{\text{gauge}}\}/(8D)$, these hypotheses do not imply $p_\varepsilon^{\text{prior}} \ge 1/4$. The assertion is a failure of a uniform guarantee, rather than a curvature lower bound for every three-layer orbit.

Proof. Take three scalar factors equal to 1, input and target equal to 1, and $R = 4$. The tangent plane at the interpolator consists of displacements $(a, b, -a - b)$. The product on this plane is $(1+a)(1+b)(1-a-b)$, a nonconstant polynomial. Its level set at 1 has two-dimensional Lebesgue measure zero. A tangent Gaussian of fixed positive width has a density on the plane, while a normal Gaussian of width $r/\sqrt{12}$ concentrates on it as $\varepsilon \downarrow 0$. With $\delta = \varepsilon^2$, the same scaled Riemann-sum argument as Proposition 6 gives good mass tending to zero. □

C.1.4 Chart-Supported Tubular Priors

The Euclidean obstruction of Proposition 6 is a statement about a linear-tangent Gaussian. The constant-mass construction that works places mass on a tube around the image of the nonlinear gauge chart, not on the tangent space at a single point. The price is that the prior must be built in chart coordinates.

Definition 11 (Chart-Supported Tubular Prior). *Keep Assumption 17 and the chart $\varphi(H) = (A^\star(I + H), (I + H)^{-1}B^\star)$ of Lemma 9, with bi-Lipschitz radius ρ_0 and lower constant ℓ. Write $k := d_1^2$ and $n := N_0 d_{\text{out}}$. Let $\Omega_\star := \{H : \|H\|_F \le \rho_0\}$ be the chart domain. A chart-supported tubular prior is the pushforward, under the nearest-grid map of*

$$\psi(H, \nu) := \varphi(H) + \nu,$$

of a product discrete Gaussian on the coordinates (H, ν), where $\nu \in \mathbb{R}^n$ is a displacement in a fixed orthonormal n-plane in the ambient parameter space, with $n \le D$. If the residual Jacobian has rank n, its row space is one possible choice. The proof only uses the displacement norm. The H-marginal is that discrete Gaussian of width σ_H conditioned on $\Omega_\star$. Both coordinate lattices have mesh δ. The ν-marginal is a product discrete Gaussian of width σ_{n} on $\delta\mathbb{Z}^n \cap [-r, r]^n$, so the support is finite. The nearest-grid map selects a nearest point of $\mathcal{G}_{\delta,R}$, clipping outside the box if necessary. The good events used below lie far enough inside the box that their rounding displacement is at most $\delta\sqrt{D}/2$. Residual coordinates orthogonal to the chart tangent and to this n-plane are omitted from the support (equivalently, given width zero). The prior is therefore supported on a tubular neighborhood of $\text{im}(\varphi)$ of width controlled by σ_{n}. Loss control on the tube uses the ambient Lipschitz constant of $\mathcal{L}^{1/2}$, so ν need not track the rotated normal bundle at $\varphi(H)$.

Assumption 18 (Tubular Prior Widths). *Keep Definition 11 and $r := c_r\sqrt{\varepsilon}/\text{Lip}$ as in Theorem 9. Assume $n \ge 16$,*

$$\sigma_{\text{n}} \le \frac{r}{\sqrt{12n}}, \qquad \sigma_H \in \left[\frac{\rho_0}{4}, \rho_0\right],$$

and

$$\delta \le \min\Big\{ \frac{\sigma_{\mathrm{n}}}{8D}, \frac{\sigma_H}{8D}, \frac{r}{4\sqrt{D}}, \frac{\ell\rho_0}{16\sqrt{Dk}} \Big\}.$$

Assume $\varphi(\Omega_\star)$ sits in the interior of $\mathcal{B}_R$ by a margin of r.

The transfer lemma now compares masses in chart coordinates rather than in ambient coordinates.

Lemma 26 (Discrete Normal Mass in Chart Coordinates)**.** *Under Assumption 18, the finite discrete normal marginal of Definition 11 satisfies*

$$\Pr(\|\nu\|_2 \le r/2) \ge 1/4.$$

This bound is uniform in the chart coordinate H and requires no injectivity of the nearest-grid pushforward.

Proof. Set $h_n = \delta\sqrt{n}/2 \le r/8$ and $A = B_2^n(r/2) \subset [-r, r]^n$. For $\xi \sim \mathcal{N}(0, I_n)$, the event $\|\xi\|^2 \le 3n/2$ has probability at least $1 - 8/n \ge 1/2$. On it $\|\sigma_{\mathrm{n}}\xi\| \le r/(2\sqrt{2}) < 3r/8 \le r/2 - h_n$. Intersecting with the radius-$4\sigma_{\mathrm{n}}\sqrt{n}$ typical ball removes at most e^{-n}. Lemma 23, applied in dimension n with box radius r, gives a discrete probability at least $(3/5)(1/2 - e^{-16}) > 1/4$. The H marginal is independent and already normalized on its finite chart domain. □

Theorem 20 (Constant Mass for the Two-Layer Tubular Prior)**.** *Under Assumptions 17 and 18, the chart-supported tubular prior of Definition 11 satisfies*

$$p_\varepsilon^{\mathrm{prior}} \ge \frac{1}{4}.$$

Conditional samples explore a macroscopic portion of the bi-Lipschitz chart domain $\Omega_\star$, and the radius ρ_0 does not shrink with ε. This theorem does not repair Proposition 6: a Euclidean linear-tangent Gaussian of width $\Theta(R)$ remains insufficient.

Proof. Every $\varphi(H)$ is an exact interpolator. For every chart coordinate, the event $\|\nu\| \le r/2$ has probability at least $1/4$ by Lemma 26. On this event the rounded parameter $\hat\theta$ stays in the box and

$$\mathcal{L}(\hat\theta)^{1/2} \le \mathrm{Lip}(r/2 + \delta\sqrt{D}/2) \le (5/8)\mathrm{Lip}\, r \le \sqrt{\varepsilon}.$$

Integrate this bound over the exact discrete H law. For any subset E of the chart coordinates, the same uniform estimate gives $\Pr(H \in E \mid \text{success}) \ge \Pr(H \in E)/4$. Thus a set carrying a fixed positive chart mass retains a fixed positive mass after success. The width and domain radius of the H marginal do not shrink with ε. □

Remark 13 (Preparation Cost of a Tubular Prior)**.** *The Gaussian on H conditioned to a Euclidean ball is not a product distribution. Its coherent preparation, including this conditioning, is part of T_{prep} and is not covered by the product lemma. After preparing the finite coordinate state, compute the rounded parameter reversibly while retaining (H, ν) as a label register. Amplify on this enlarged state, with the predicate evaluated on the computed parameter, and measure the parameter register on success. Summing over labels gives exactly the intended pushforward law even when distinct labels round to the same parameter. An in-place unitary on parameters alone cannot implement an arbitrary many-to-one map. Chart evaluation, numerical precision, and conditioning costs must all be included. The chart and its known exact interpolator are inputs to this theorem, and their discovery is not supplied by the query bound.*

The corollaries lift the construction to three layers, to every larger depth under the width regimes of Lemma 15, and to fixed-mixing linear attention.

Corollary 5 (Three-Layer Tubular Prior)**.** *Under Assumption 4 and the three-layer chart of Lemma 13, form the chart-supported tubular prior of Definition 11 with (H_1, H_2) in place of H, with $k = d_1^2 + d_2^2$, and with (ρ_0, ℓ) from Lemma 14. If the width and mesh conditions of Assumption 18 hold with this k, then*

$$p_\varepsilon^{\mathrm{prior}} \ge \frac{1}{4}.$$

The Euclidean three-layer obstruction of Corollary 4 is unchanged.

Proof. Replace φ in the proof of Theorem 20 by the three-layer chart. Loss vanishing on the chart is Lemma 12. Bi-Lipschitz control on $\Omega_\star$ is Lemma 14. The uniform discrete normal-mass argument does not use the number of gauge blocks beyond the dimension k in the mesh condition. □

Corollary 6 (Tubular Prior for $L > 3$)**.** *For $L > 3$, keep the four width regimes of Lemma 15 and the chart $\varphi^{(L)}$ of Proposition 1. Let $k_L = \sum_{\ell=1}^{L-1} d_\ell^2$, and let $(\rho_0^{(L)}, \ell^{(L)})$ be the bi-Lipschitz radius and lower constant produced by that proposition, so $\ell^{(L)} \geq \kappa^{(L)}/2$. Form the chart-supported tubular prior of Definition 11 with $(H_1, \ldots, H_{L-1})$ in place of H. If the width and mesh conditions of Assumption 18 hold with $(k_L, \rho_0^{(L)}, \ell^{(L)})$ in place of (k, ρ_0, ℓ), then*

$$p_\varepsilon^{\text{prior}} \geq \frac{1}{4}.$$

Outside those width regimes the same bound holds whenever $\kappa^{(L)} > 0$ is supplied as a hypothesis. Multimodal zigzag widths remain open, as in Remark 5.

Proof. Proposition 1 gives injectivity of $D\varphi^{(L)}(0)$ and the bi-Lipschitz pair $(\rho_0^{(L)}, \ell^{(L)})$ on the chart domain. Replace φ in the proof of Theorem 20 by $\varphi^{(L)}$. Loss vanishing on the chart is Lemma 12. The uniform normal-mass argument uses exact interpolation, the box margin, and the stated mesh, and does not require a general nonlinear mass-transfer lemma. □

Corollary 7 (Linear-Attention Tubular Prior)**.** *Under Assumption 11, the two-layer tubular prior of Definition 11 and Theorem 20 apply, under their full rank, width, and box hypotheses, to the factorization $(W_O, W_V^\top)$ acting on the mixed features $Z_i = PX_i$. Softmax attention is not included.*

Proof. Corollary 2 identifies the model with $L = 2$ deep linear geometry on the mixed features $\{Z_i\}$, so the gauge chart and the tubular prior transfer unchanged. □

C.1.5 Cell-Supported ReLU Priors

After a pattern is frozen, the second-layer map in V is linear and the interpolating V-fiber is flat. A split prior in that cell is therefore in the same family as Theorem 19, not in the curved gauge-split family. The only new feature is that the design matrix of the linear problem now depends on the first-layer weights U, so the U-block needs its own concentration estimate.

Assumption 19 (Aligned ReLU Cell Center)**.** *Keep Assumptions 5, 6, and 7. Let $\theta_0 = (U^\star, V_0)$ with V_0 an exact interpolator at $U^\star$, and put $r_V = \sqrt{\varepsilon}/(2\sigma_{\max}(H^\star))$. Use the product box in the fixed orthonormal V-normal and V-kernel frame at $U^\star$. Require the radius-$4\sigma_U\sqrt{md_{\text{in}}}$ ball about $U^\star$ to lie in the U-box and the radius- $4\sigma_{V,\perp}\sqrt{n}$ ball about $V_{0,\perp}$ to lie in the normal-coordinate box. The wide kernel block is truncated independently to its coordinate box, with no typical-ball containment assumption.*

Definition 12 (Cell-Supported ReLU Split Prior)**.** *Split coordinates into the U-block, the data-normal block in V of dimension $n = Nd_{\text{out}}$, and the V-kernel block of dimension $\kappa_V = (m - N)d_{\text{out}}$. A cell-supported ReLU split prior is the product discrete Gaussian with widths $(\sigma_U, \sigma_{V,\perp}, \sigma_{V,\|})$ on those blocks, centered at θ_0. All V projectors are fixed at $U^\star$, not recomputed at the sampled U. The grid is the product mesh of Lemma 25 in an orthonormal frame of these three blocks, so the V-normal and V-kernel directions are coordinate subspaces of the V-factor.*

Theorem 21 (Constant Mass on a ReLU Pattern Cell)**.** *Under Assumption 19, put $a = md_{\text{in}}$, $b = md_{\text{out}}$, $n = Nd_{\text{out}}$, and*

$$L_U = R\sqrt{b}\,\sqrt{N}X_{\max}, \qquad a_U = \min\{\rho_U, \sqrt{\varepsilon}/(4L_U)\}.$$

Choose

$$\sigma_U \leq \frac{a_U}{\sqrt{18a}}, \qquad \sigma_{V,\perp} \leq \frac{r_V}{\sqrt{18n}}, \qquad \sigma_{V,\|} \in [R/4, R],$$

with $\delta \leq \min\{\sigma_U, \sigma_{V,\perp}, \sigma_{V,\|}\}/(8D)$. For $a, n \geq 16$, the prior puts mass at least $1/8$ on parameters that are both ε-good and in the pattern cell $P_\star$. If also $\sigma_U \leq a_U/(12\sqrt{a})$, the lower bound is $1/4$. For the original loss-only predicate, this bounds total good mass but does not assert that every successful sample stays in $P_\star$. To enforce that support exactly, include the cell test in the predicate.

Proof. Let $E_U = \{\|U - U^\star\| \le a_U/2\}$ and $E_V = \{\|V_\perp - V_{0,\perp}\| \le r_V/2\}$. In each block, Chebyshev gives probability at least $7/8$ to a fluctuation of size at most $a_U/3$ or $r_V/3$, respectively. The block rounding radius is smaller than one eighth of that block's scale. Intersect with its radius-$3\sigma\sqrt{\text{dimension}}$ ball. The assumed radius-$4\sigma\sqrt{\text{dimension}}$ box margin makes the intersection an inner set, and its omitted tail is at most e^{-16}. Lemma 23 gives each discrete block probability at least $(7/8-e^{-16})/2$. The blocks remain independent after product-box truncation, and their product exceeds $1/8$.

For any V in its box, $\|V\|_F \le R\sqrt{b}$. Since V_0 interpolates at $U^\star$ and the kernel component is annihilated by $H^\star$, on $E_U \cap E_V$ the residual obeys

$$\|H(U)V^\top - Y\|_F \le \sigma_{\max}(H^\star)\, r_V/2 + L_U a_U/2 \le 3\sqrt{\varepsilon}/8.$$

Also $\|U - U^\star\| \le \rho_U/2$ fixes the activation pattern. Under the additional U-width bound, the entire radius- $3\sigma_U\sqrt{a}$ ball is in E_U with the required rounding margin. Using the sharper $3/5$ transfer factor in both blocks gives probability at least $(3/5)^2(1-e^{-16})(7/8-e^{-16}) > 1/4$. The estimate is uniform in the truncated kernel coordinates, so any kernel-coordinate subset retains at least the stated constant times its prior mass after conditioning on success. □

Remark 14 (ReLU Cell Prior versus Ambient Union)**.** *Theorem 21 is the ReLU analogue of the linear split prior and of the deep-linear tubular prior. Ambient uniform search still sees only the small prefactor $c_{\text{pat}}c_{0,V}$. Constant prior mass does not require a polynomial cell union. It requires a prior supported in one stable cell, or a mixture of finitely many such cells with weights that do not cancel the $1/4$ (if each component has good mass at least $1/4$, any mixture also has good mass at least $1/4$).*

D Oracle Ladder Details

D.1 Oracle Strength and the SGD Certificate Gap

The hard full-batch indicator is intentionally strong, and this subsection records the weaker oracle variants together with their query theorems. The point of the comparison is to explain why the hard oracle remains the primary object of the paper. It supplies a training-set certificate that standard mini-batch SGD does not provide, and every weakening of the oracle weakens exactly that certificate.

D.1.1 The Oracle Hierarchy

We first restate the hierarchy of Definitions 5 and 6 from the main text in one place, so that the variants can be compared line by line.

Definition 13 (Full-Batch Hard Indicator)**.** *The primary oracle of Definition 5 uses the full empirical loss $\mathcal{L}$ on all N training samples, so the ancilla bit marks θ exactly when $\mathcal{L}(\theta) \le \varepsilon$. Success therefore implies $\mathcal{L}(\hat{\theta}) \le \varepsilon$ by Theorem 1, which is the certificate carried by every main theorem.*

Definition 14 (Mini-Batch Hard Indicator)**.** *For a batch $B \subset \{1, \dots, N\}$, let $\mathcal{L}_B$ denote the empirical loss restricted to B. The batch indicator marks θ if and only if $\mathcal{L}_B(\theta) \le \varepsilon$. The marked set grows, because fitting one batch is easier than fitting all N samples, and the certificate shrinks accordingly.*

Definition 15 (Soft Score Oracle)**.** *Fix a temperature $\tau > 0$ and a decreasing score $s_\tau : \mathbb{R}_{\ge 0} \to (0, 1]$, for example $s_\tau(t) = e^{-t/\tau}$ or $s_\tau(t) = (1 + t/\tau)^{-1}$. A soft oracle coherently encodes a phase or amplitude determined by $s_\tau(\mathcal{L}(\theta))$ instead of a Boolean verdict, which enables non-Boolean amplitude amplification (Shyamsundar, 2021; Peters & Shyamsundar, 2021). The score grades every parameter by its loss, so no parameter is ever strictly excluded, and the returned sample carries no finite-loss guarantee.*

Definition 16 (Two-Stage Certify)**.** *The two-stage scheme runs soft or batch amplification to produce a candidate, and then calls the full-batch hard indicator once to accept or reject the candidate. The complexity splits into a search cost plus one certification call per candidate, and the full-batch certificate holds only when the final call accepts.*

Remark 15 (Design Choice)**.** *We keep Definition 5 as the default theorem target because it matches the scientific claim "return a parameter with training loss $\le \varepsilon$". Soft and batch oracles are retained*

as boundary variants for complexity analyses in which the certificate is intentionally weakened, and Theorem 25 shows how a wrapper can restore the certificate at an explicit cost.

D.1.2 What Mini-Batch SGD Actually Guarantees

The certificate gap has two halves. Mini-batch SGD does not produce a full-batch certificate, and verifying such a certificate classically is not free. We record the two halves as explicit propositions.

Proposition 7 (Certificate Gap)**.** *Termination of mini-batch SGD need not imply $\mathcal{L}(\theta_T) \le \varepsilon$. A full-batch loss evaluation suffices to certify the predicate. A full-batch gradient alone does not certify it for a general loss. Without structural information, a deterministic certificate based on individual nonnegative loss values may require inspecting every previously unknown term at the candidate. No universally necessary additional pass after training is asserted.*

Proof. Take $\ell_1(\theta) = \theta^2$ and $\ell_2(\theta) = (\theta - 1)^2$. At $\theta = 0$, training only on the first term makes no update, while the full loss is 1. At $\theta = 1/2$ the full gradient is zero but the full loss is $1/2$. Finally, if an unqueried nonnegative term has no known upper bound, it can be changed to exceed ε without changing the observed values. Conversely, earlier evaluations or valid structural bounds may already establish the certificate. □

Proposition 8 (Leveled Certificate Comparison)**.** *Under the hard full-batch oracle model, measurement as training returns a parameter already equipped with the predicate $\mathcal{L} \le \varepsilon$ on success. A fair classical baseline must account for how it certifies the same predicate, charging any required full-batch check or other valid certificate. Comparing hard-oracle quantum query complexity only against uncertified mini-batch SGD therefore understates the classical cost of matching the guarantee.*

Proof. Combine Theorem 1 with Proposition 7. □

Remark 16 (Soft or Batch Oracles Remove This Gap)**.** *If one switches the primary object to the mini-batch or the soft variant of Definition 6, the certificate advantage disappears and the comparison against mini-batch SGD becomes closer in strength. That switch is available inside our hierarchy. It should be announced explicitly when used, rather than silently replacing the hard full-batch claim.*

D.1.3 Query Theorems for Soft, Mini-Batch, and Two-Stage Oracles

The default analysis uses Definition 5, and this subsection records what the weaker oracles certify instead. The common theme of the three theorems below is negative in the same way. None of them upgrades a mini-batch or graded predicate into a full-batch ε-certificate.

Definition 17 (Amplitude-Encoded Soft Oracle)**.** *Fix $\tau > 0$ and a nonincreasing score $s_\tau : \mathbb{R}_{\ge 0} \to [0, 1]$. An amplitude-encoded soft oracle is a unitary satisfying*

$$U_{\mathrm{soft}}|\theta\rangle|0\rangle = |\theta\rangle\Big(\sqrt{s_\tau(\mathcal{L}(\theta))}\,|1\rangle + \sqrt{1 - s_\tau(\mathcal{L}(\theta))}\,|0\rangle\Big).$$

The ancilla amplitude on $|1\rangle$ is large exactly when the loss is small, so the score is encoded in amplitudes. The prior soft mass is

$$q_\tau := \sum_{\theta \in \mathcal{G}_{\delta,R}} |c_\theta|^2\, s_\tau(\mathcal{L}(\theta)),$$

which is the expected score under the prior and plays the role that $p_\varepsilon^{\mathrm{prior}}$ plays for the hard oracle.

Theorem 22 (Soft Amplitude-Encoding Query Bound)**.** *Assume $q_\tau \in (0, 1]$ and access to U_{soft} from Definition 17. Amplitude amplification about the prepared joint state $U_{\mathrm{soft}}|\Psi_{\mathrm{prior}}\rangle|0\rangle$ and the ancilla-$|1\rangle$ subspace outputs a sample from the prior reweighted by $s_\tau(\mathcal{L})$ with success probability at least $2/3$, using*

$$O\Big(\frac{1}{\sqrt{q_\tau}}\Big)$$

calls to U_{soft} and its inverse, up to a universal constant. If $s_\tau(t) = \mathbf{1}[t \le \varepsilon]$ then $q_\tau = p_\varepsilon^{\mathrm{prior}}$ and the statement reduces to Theorem 1. For a strictly soft score, success does not imply $\mathcal{L}(\hat{\theta}) \le \varepsilon$.

Proof. The overlap of the prepared state $U_{\text{soft}}|\Psi_{\text{prior}}\rangle|0\rangle$ with the ancilla-$|1\rangle$ subspace equals $\sqrt{q_\tau}$, and the remainder of the argument is Theorem 1 with that overlap in place of $\sqrt{p_\varepsilon^{\text{prior}}}$. The output law is the prior reweighted by s_τ because the ancilla amplitude on $|1\rangle$ is $\sqrt{s_\tau(\mathcal{L}(\theta))}$. A strictly soft score can be positive when $\mathcal{L} > \varepsilon$, so the marked subspace is not contained in $\mathcal{S}_\varepsilon^\delta$ and the certificate fails. □

Definition 18 (Phase-Encoded Soft Oracle)**.** *A phase-encoded soft oracle acts as*

$$U_{\text{phase}}|\theta\rangle = e^{i\pi(1-s_\tau(\mathcal{L}(\theta)))}|\theta\rangle,$$

so the same score is encoded as a phase rotation rather than as an amplitude.

Theorem 23 (Phase Encoding through a Controlled Interference Test)**.** *Assume controlled access to U_{phase} of Definition 18 and its inverse. Define*

$$w(s) = \sin^2(\pi s/2), \qquad q_{\text{phase}} = \sum_\theta |c_\theta|^2 w(s_\tau(\mathcal{L}(\theta))).$$

A controlled interference test followed by amplitude amplification samples the prior reweighted by $w \circ s_\tau$ with constant success using $O(q_{\text{phase}}^{-1/2})$ controlled phase-oracle calls, provided $q_{\text{phase}} > 0$. Amplitude estimation estimates q_{phase} to additive error η with constant success using $O(1/\eta)$ such calls. These bounds concern q_{phase}, not the mean score q_τ. For Boolean scores the weights coincide. For a general soft score, the output has no hard full-batch certificate without an additional check.

Proof. Prepare an ancilla in $|+\rangle$, apply controlled U_{phase}, then a Hadamard to the ancilla. Given θ, the probability of ancilla outcome 0 is

$$\left|\frac{1 + e^{i\pi(1-s)}}{2}\right|^2 = \cos^2(\pi(1-s)/2) = \sin^2(\pi s/2), \qquad s = s_\tau(\mathcal{L}(\theta)).$$

Apply amplitude amplification and estimation to this prepared state and the ancilla-0 predicate (Brassard et al., 2002). The original mean cannot be substituted into this calculation: for small positive s, $w(s) = \Theta(s^2)$, rather than $\Theta(s)$. Positive weights outside the good set still prevent a hard certificate. □

Theorem 24 (Mini-Batch Hard Indicator: Certified versus Not)**.** *Let $B \subset \{1, \dots, N\}$ be a nonempty batch and let $\mathcal{L}_B(\theta) := \sum_{i \in B} \|f_\theta(x_i) - y_i\|_2^2$ when the training loss is a sum of per-sample squares as in Definition 2 and the deep linear model of Section 3.2.2. Write*

$$p_B := \sum_{\theta:\, \mathcal{L}_B(\theta) \le \varepsilon} |c_\theta|^2.$$

Assume $p_B > 0$. Amplitude amplification with the mini-batch indicator of Definition 6 returns a sample from the prior conditioned on $\mathcal{L}_B \le \varepsilon$, with success probability at least $2/3$, using $O(1/\sqrt{p_B})$ batch-oracle calls. On success one has $\mathcal{L}_B(\hat{\theta}) \le \varepsilon$. One need not have $\mathcal{L}(\hat{\theta}) \le \varepsilon$. Because $\mathcal{L}_B \le \mathcal{L}$ pointwise for a sum-of-squares loss, the mini-batch good set contains $\mathcal{S}_\varepsilon^\delta$, hence $p_B \ge p_\varepsilon^{\text{prior}}$, but the inclusion is strict in general. Mini-batch amplification alone does not yield a full-batch ε-certificate.

Proof. The query bound is Theorem 1 with the marked set $\{\mathcal{L}_B \le \varepsilon\}$, and the certificate is the defining predicate of that marked set. For a sum of nonnegative per-sample terms, $\mathcal{L}_B(\theta) \le \mathcal{L}(\theta)$ holds identically, so every full-batch ε-good point is mini-batch ε-good. A point can fit batch B while retaining large loss on $\{1, \dots, N\} \setminus B$, so the converse inclusion fails. The query complexity $O(1/\sqrt{p_B})$ can therefore be smaller than $O(1/\sqrt{p_\varepsilon^{\text{prior}}})$ precisely because the marked set is larger and the certificate is weaker. □

Theorem 25 (Two-Stage Search Then Certify)**.** *Assume $p_\varepsilon^{\text{prior}} > 0$. Run Stage 1 as either Theorem 22 or Theorem 24, retaining a candidate $\tilde{\theta}$ only when its success flag accepts. Repeat Stage 1*

after a failed flag. Run Stage 2 as a single call to the full-batch hard indicator $U_{\mathcal{L}}$. If Stage 2 accepts, then $\mathcal{L}(\tilde{\theta}) \le \varepsilon$. If Stage 1 is mini-batch with success probability $\Theta(1)$ per run, the probability that a successful Stage 1 candidate is accepted at Stage 2 equals

$$\alpha := \frac{p_\varepsilon^{\mathrm{prior}}}{p_B} \in (0, 1].$$

The expected number of full-batch certification calls until the first accept is $1/\alpha$. The expected number of mini-batch oracle calls is $O(\sqrt{p_B}/p_\varepsilon^{\mathrm{prior}})$. This recovers a full-batch certificate only on Stage 2 acceptance. It does not produce a full-batch ε-certificate from mini-batch queries alone, and it does not improve on $O(1/\sqrt{p_\varepsilon^{\mathrm{prior}}})$ full-batch queries of Theorem 1 when both oracles are priced equally.

Proof. Stage 2 applies Definition 5 to a single classical candidate, so acceptance is exactly $\mathcal{L} \le \varepsilon$. For mini-batch Stage 1, a successful candidate is distributed as the prior conditioned on $\mathcal{L}_B \le \varepsilon$, and the conditional probability of $\mathcal{L} \le \varepsilon$ is $p_\varepsilon^{\mathrm{prior}}/p_B$. Independent repetition until Stage 2 accepts is a geometric random variable with that success probability, and multiplying by the $O(1/\sqrt{p_B})$ Stage 1 cost per candidate gives $O(\sqrt{p_B}/p_\varepsilon^{\mathrm{prior}})$. Since $p_B \ge p_\varepsilon^{\mathrm{prior}}$, this upper-bound expression is no smaller than the direct full-batch bound. For the standard amplification implementation, whose cost per candidate is $\Theta(p_B^{-1/2})$, the total is $\Theta(\sqrt{p_B}/p_\varepsilon^{\mathrm{prior}})$. Comparing upper bounds alone would not prove a lower bound for every possible Stage 1 algorithm. The procedure is useful when a batch oracle is cheaper than $U_{\mathcal{L}}$ in a circuit-cost model that distinguishes the two. It is not a query reduction inside the single-oracle cost accounting of Section 3.1. □

Remark 17 (Default Remains Hard Full-Batch)**.** *The default analysis continues to use the hard full-batch indicator. Soft and mini-batch theorems are variants with weaker certificates, and two-stage certification is a wrapper that restores the full-batch predicate at an explicit acceptance-probability cost.*

E Amplification Proofs and Extended Comparisons

E.1 Amplification, Guarantees, and Leveled Comparisons

This subsection records the universal machinery that every architecture block of the main text reuses. Theorem 26 gives the query bound under an arbitrary prior, Lemma 27 prices the preparation of the prior state, Theorems 27 and 28 state the training-loss guarantee and the sampling view of ε-solutions, and the leveled comparisons of Section E.1.3 fix what a fair classical baseline is.

E.1.1 Amplitude Amplification under a Prior

Theorem 26 (Amplitude Amplification under a Prior)**.** *Let $p_\varepsilon^{\mathrm{prior}} \in (0, 1]$ be as in Definition 3, and assume access to $U_{\mathcal{L}}$ and a coherent prior-preparation unitary A and its inverse. A known success mass, or a constant-factor lower bound, gives the bounded-run interpretation below. With unknown positive mass, the standard adaptive search gives the same expected query order. Using amplitude amplification (Brassard et al., 2002; Grover, 1996) about $|\Psi_{\mathrm{prior}}\rangle$ and about the marked subspace, there is a quantum algorithm that outputs a sample from the conditional distribution*

$$\Pr[\theta] \propto |c_\theta|^2 \mathbf{1}[\theta \in \mathcal{S}_\varepsilon^\delta]$$

with success probability at least $2/3$ using

$$O\Big(\frac{1}{\sqrt{p_\varepsilon^{\mathrm{prior}}}}\Big)$$

calls to $U_{\mathcal{L}}$ and $U_{\mathcal{L}}^\dagger$, up to a universal constant. The uniform case is $p_\varepsilon^{\mathrm{prior}} = p_\varepsilon$.

Proof. The initial overlap of $|\Psi_{\mathrm{prior}}\rangle$ with the marked subspace equals $\sqrt{p_\varepsilon^{\mathrm{prior}}}$, because the amplitude mass of the prior on $\mathcal{S}_\varepsilon^\delta$ is exactly the prior mass $p_\varepsilon^{\mathrm{prior}}$. Standard amplitude amplification raises

the marked amplitude to $\Omega(1)$ with $O(1/\sqrt{p_\varepsilon^{\text{prior}}})$ reflections, and each reflection uses one oracle call and one inverse. A final predicate test heralds success and its cost is included. Conditioned on that flag, the measured parameter has the prior law conditioned on membership in $\mathcal{S}_\varepsilon^\delta$. The same argument applies to a prior purified by a retained label register, since the marked predicate depends only on the parameter. □

Lemma 27 (Product Discrete Gaussian Preparation)**.** *Let $n = \lceil \log_2(2R/\delta + 1) \rceil$. Suppose the finite prior factors in a specified coordinate frame. Assume a reversible procedure computes the conditional interval-mass ratios of each one-dimensional discrete Gaussian to b bits, coherently from a prefix, with gate cost $T_{\text{CDF}}(n, b)$. Then a preparation unitary with state-vector error at most η can be implemented with cost*

$$T_{\text{prep}} = O\left(Dn\left[T_{\text{CDF}}(n, b) + \text{poly}(b)\right] + T_{\text{coord}}\right), \qquad b = O(\log(Dn/\eta)).$$

Here T_{coord} includes reversible coordinate evaluation with any necessary input labels retained. A polynomial preparation bound requires a polynomial bound on T_{CDF} and input bit lengths. The preparation and its inverse are used $O(T_Q)$ times in amplification.

Proof. At each of the n prefix levels, reversibly compute the conditional mass ratio, rotate the next qubit, and uncompute the workspace. The resulting binary tree prepares the point-weight distribution when the ratios are exact. Allocate state-vector error $O(\eta/(Dn))$ per rotation and use the triangle inequality over the Dn operations. The extra precision needed for square roots near zero changes b only by a constant factor. This is the sequential-rotation construction (Grover & Rudolph, 2002; Kitaev & Webb, 2008) under the stated coherent ratio-computation hypothesis. A continuous Gaussian integral expressed using erf is not the exact cumulative sum of a point-weight discrete Gaussian. Nor does tabulating exponentially many classical angles give a polynomial coherent circuit. These costs must be supplied by T_{CDF}. Finally the reflection is $A(2|0\rangle\langle 0| - I)A^\dagger$, so the preparation cost is repeated in each amplification iteration. □

Remark 18 (State-Preparation Cost)**.** *The lemma is conditional on coherent discrete mass-ratio computation. Widths and centers contribute to arithmetic precision, and ball conditioning of a multivariate Gaussian is a separate cost. The exact query and output-law statements refer to the actual prepared prior. If its parameter law is within total variation η of a target law of good mass p, then its good mass is at least $p - \eta$ and the conditional-law error is at most $2\eta/p$. Thus choosing $\eta \le p/2$ preserves the query order, but does not give the exact target law. Total preparation work is $O(T_Q T_{\text{prep}})$, which is $O(T_{\text{prep}})$ in the constant-mass regime.*

E.1.2 Training-Loss Guarantee and Solution Distribution

Theorem 27 (ε-Optimality on Success)**.** *Conditional on success of Definition 7, the returned parameter $\hat{\theta}$ satisfies $\mathcal{L}(\hat{\theta}) \le \varepsilon$. Only outputs whose final predicate flag accepts are reported as certified parameters. Other runs are reported as failures.*

Proof. By construction the marked subspace is supported on $\mathcal{S}_\varepsilon^\delta$, and every computational basis state of that subspace satisfies $\mathcal{L} \le \varepsilon$ by the definition of the oracle. Success means that the measured basis state lies in this subspace, which proves the claim. □

Theorem 28 (Sampling Global ε-Solutions)**.** *On success, the output of Theorem 1 is distributed as the prior $\{|c_\theta|^2\}$ conditioned on $\mathcal{S}_\varepsilon^\delta$. Repeating the algorithm yields i.i.d. samples from that conditional distribution (up to abort events of constant probability).*

Proof. This is immediate from the post-amplification state before measurement in the proof of Theorem 1, whose parameter amplitudes are the prior amplitudes restricted to $\mathcal{S}_\varepsilon^\delta$ and renormalized. Independent runs share no randomness, so the samples are independent. □

Remark 19 (Connectivity versus Sampling)**.** *Classical analyses of overparameterized models often establish path-connectivity of interpolating sets, and SGD produces one trajectory endpoint. Measurement as training does not newly prove connectivity. Its distinct guarantee is access to the* measure *of ε-solutions under a stated prior through Theorem 1, which is a different geometric quantity from the existence of one path.*

E.1.3 Leveled Classical Baselines

All comparisons below price one landscape access on both sides, in the cost model of Section 3.1. We first name the three cost functionals.

Definition 19 (Unified Landscape-Access Cost)**.** *Let $T_Q(\varepsilon)$ be the number of $U_{\mathcal{L}}$ calls in Theorem 1. Let $T_{\mathrm{rej}}(\varepsilon)$ be the expected number of prior samples needed by classical rejection sampling to hit $\mathcal{S}_\varepsilon^\delta$ under the same $\{|c_\theta|^2\}$. Let $T_{\mathrm{SGD}}(\varepsilon)$ be the number of gradient steps a stated first-order method needs to reach training loss $\le \varepsilon$ from an initialization drawn from the same prior family, counting one step as one landscape access.*

Proposition 9 (Quadratic Win against Same-Prior Rejection Sampling)**.** *Under the cost model of Section 3.1, $T_{\mathrm{rej}}(\varepsilon) = \Theta(1/p_\varepsilon^{\mathrm{prior}})$ while $T_Q(\varepsilon) = O(1/\sqrt{p_\varepsilon^{\mathrm{prior}}})$. Thus measurement as training is quadratically faster than classical rejection sampling from the identical prior, up to the preparation cost in Section 3.1.*

Proof. Rejection sampling succeeds with probability $p_\varepsilon^{\mathrm{prior}}$ per independent draw, so the number of draws until the first success is geometric and has expectation $\Theta(1/p_\varepsilon^{\mathrm{prior}})$. The quantum count is Theorem 1. □

Proposition 10 (Conditional Win against Same-Prior SGD)**.** *In the cost model of Section 3.1, measurement as training uses fewer landscape accesses than SGD whenever*

$$\frac{C}{\sqrt{p_\varepsilon^{\mathrm{prior}}}} < T_{\mathrm{SGD}}(\varepsilon)$$

for the absolute constant C from Theorem 1. Equivalently,

$$p_\varepsilon^{\mathrm{prior}} > \frac{C^2}{T_{\mathrm{SGD}}(\varepsilon)^2}.$$

This is a boundary condition, not an unconditional claim.

Proof. Immediate from Theorem 1 and the cost model of Section 3.1. □

Proposition 11 (Prior Mass Is Not Automatic)**.** *For a discrete Gaussian on a fixed finite grid, let J be the set of grid points nearest to its center θ_0. As $\sigma \downarrow 0$, the good mass tends to $|J \cap \mathcal{S}_\varepsilon^\delta|/|J|$. In particular it vanishes if all nearest grid points are bad, including the case of a grid-valued bad center. As $\sigma \to \infty$ the prior converges to the uniform grid law. These limits show that constant mass is not automatic. They do not assert that the sufficient alignment conditions in this paper are necessary.*

Proof. On a finite grid, divide all weights by the largest weight. Weights at nearest points are equal to one, and every other weight tends to zero as $\sigma \downarrow 0$. As $\sigma \to \infty$, all weights tend to one. Normalize in each case. □

Corollary 8 (Uniform-Prior Amplification versus Full-Batch GD)**.** *For the linear model, fix $\kappa = \sigma_{\max}^2/\sigma_{\min}^2$, assume $d = O(N)$, $\sqrt{\varepsilon}/(\sigma_{\max} R) \le 1/2$, and $\delta \le r/(4\sqrt{d})$. Standard amplitude amplification from the uniform prior, using reflections about that prior and the good subspace, requires*

$$T_Q = \Omega\left((cN^{1/2}\kappa^{-1/2})^{N/2}/2^{O(d)}\right)$$

oracle calls for constant success probability, for an absolute $c > 0$. For comparison, continuous full-batch gradient descent with step $1/(2\sigma_{\max}^2)$ reaches loss ε in $O(1 + \kappa \log \max\{1, \mathcal{L}(\theta_0)/\varepsilon\})$ steps. If this logarithm grows at most polynomially in N, the displayed amplification cost is asymptotically larger. This is a comparison with continuous loss minimization. It is not an instance-wise black-box lower bound for every algorithm that can exploit the linear structure, or a claim that GD has the same conditional output law.

Proof. For the standard two-reflection search, after t iterations the marked amplitude has magnitude at most $(2t+1)\sqrt{p_\varepsilon}$, by the two-dimensional rotation formula. Constant success therefore requires $t = \Omega(p_\varepsilon^{-1/2})$ when $p_\varepsilon \to 0$. Put $\rho = \sqrt{\varepsilon}/\sigma_{\min}$ and $h = \delta\sqrt{d}/2$. Then $\rho + h \le 9\sqrt{\varepsilon}/(8\sigma_{\min})$ and $h/R \le 1/16$. Lemma 7, together with $v_N \le (2\pi e/N)^{N/2}$, gives $p_\varepsilon \le (C\sqrt{\kappa}/\sqrt{N})^N 2^{O(d)}$. This yields the claimed amplification lower bound. For GD, the residual obeys $e_{t+1} = (I - XX^\top/\sigma_{\max}^2)e_t$. Thus its norm contracts by at most $1 - 1/\kappa$, with one step sufficient when $\kappa = 1$. This proves the loss bound without assuming strong convexity in the kernel directions. The stated growth condition then makes the GD bound polynomial while the amplification bound is super-polynomial. □

Remark 20 (Scope of the Comparisons)**.** *The theory therefore splits into three layers of comparison:*

1. *versus same-prior rejection sampling, where Theorem 1 always gives a quadratic improvement whenever $p_\varepsilon^{\text{prior}} > 0$,*
2. *versus same-prior SGD, where Proposition 10 may hold or fail depending on prior alignment and geometry,*
3. *versus ambient uniform search, where Corollary 8 exhibits an explicit negative case.*

Restricted gauge charts are deferred to a later revision.

F Negative Results for Unaligned Priors

The ladder has sharp boundaries, and this section collects the negative results for unaligned priors. The certificate gap itself is recorded in Appendix D.1.2, and the boundary condition for comparing against same-prior SGD is Proposition 10.

Theorem 29 (Negative Results for Unaligned Priors)**.** *The following hold.*

• ***Gaussian Mass Is Not Automatic.*** *Constant Gaussian mass is not automatic. On a fixed finite grid, the small-width limit is supported on the nearest grid points to the center, while the large-width limit is uniform (Proposition 11). The alignment conditions in Section 3.2 are sufficient, not necessary.*

• ***Uniform-Prior Tube Bounds.*** *For the single linear layer above, with $r = \sqrt{\varepsilon}/\sigma_{\max}$ and a parallel slice whose intrinsic $2h$-interior has relative volume at least $c_0 \in (0, 1]$, where $h = \delta\sqrt{d}/2$, where c_0 is the fraction of the box volume occupied by the interpolating directions, if $\delta \le r/(4\sqrt{d})$ then the uniform good-set fraction is*

$$p_\varepsilon \ge C_{d,N}\, c_0 \Big(\frac{\sqrt{\varepsilon}}{\sigma_{\max} R}\Big)^N$$

with the explicit factor $C_{d,N} = 2^{d-N}\, v_N\, (3/4)^N/3^d$ and $v_N = \pi^{N/2}/\Gamma(N/2+1)$ (Appendix B). The matching upper bound is Lemma 7.

• ***Narrow and Broad Isotropic Regimes.*** *Under Assumption 16, $\sigma \le r/\sqrt{18N}$ gives constant mass. Under the additional bounded aspect-ratio and condition-number hypotheses of Proposition 5, $\sigma = \omega(r/\sqrt{N})$ gives super-polynomially small mass. Polynomial mass in the transition window is not excluded.*

• ***Uniform Amplification versus GD.*** *Under the hypotheses of Corollary 8, standard uniform-prior amplitude amplification needs $\Omega((cN^{1/2}\kappa^{-1/2})^{N/2}/2^{O(d)})$ calls, while continuous full-batch GD takes $O(1 + \kappa \log \max\{1, \mathcal{L}(\theta_0)/\varepsilon\})$ steps. The former is larger when the latter logarithm grows at most polynomially in N. This does not lower-bound every algorithm for a structured linear instance.*

The full step-by-step pipeline comparison against classical random parameters is recorded in Appendix G.1.

Proof. The four items are respectively Proposition 11, Theorem 8 with Lemma 7, Proposition 5, and Corollary 8, with their hypotheses retained. □

G Full Random-Parameter Pipeline

G.1 Pipeline Comparison with Classical Random Parameters

This subsection is expository. It walks measurement as training and classical random parameter search through the same pipeline, under the *same* prior and the *same* full-batch ε-predicate, so that the shared certificate and the quadratic query gap become visible step by step. The classical baseline of this comparison is random search, not gradient training, and the goal is to display the whole process rather than to claim dominance over SGD.

G.1.1 Shared Inputs

Both procedures receive the same three objects. First, the discrete grid $\mathcal{G}_{\delta,R}$ together with the prior $\{|c_\theta|^2\}$ of Definitions 1 and 3. Second, the hard full-batch indicator predicate $f_\varepsilon(\theta) = \mathbf{1}[\mathcal{L}(\theta) \le \varepsilon]$. Third, a failure budget, that is, a constant target success probability such as $2/3$. The two sides then realize the prior differently. The classical side draws independent parameters from the prior and tests each draw against f_ε, whereas the quantum side prepares the prior once as the coherent state $|\Psi_{\mathrm{prior}}\rangle$, amplifies, and measures.

G.1.2 Step-by-Step Comparison

Step 0: Initialization. The classical side samples $\theta \sim \{|c_\theta|^2\}$ from a classical random number generator, once per attempt. The quantum side prepares $|\Psi_{\mathrm{prior}}\rangle$, whose preparation cost T_{prep} is reported separately (Section 3.1).

Step 1: Ask Whether the Parameter Is ε-Good. The classical side evaluates $\mathcal{L}(\theta)$, or equivalently the indicator $f_\varepsilon(\theta)$, once per draw. The quantum side applies $U_{\mathcal{L}}$ in superposition, once inside each amplitude-amplification reflection.

Step 2: Accept or Retry. Classical rejection sampling repeats Steps 0 and 1 until the draw satisfies $f_\varepsilon(\theta) = 1$, and the expected number of indicator evaluations is $T_{\mathrm{rej}} = \Theta(1/p_\varepsilon^{\mathrm{prior}})$ by Proposition 9. Quantum amplitude amplification replaces the blind retries by reflections about the prior state and about the marked set, and the query count drops to $T_Q = O(1/\sqrt{p_\varepsilon^{\mathrm{prior}}})$ by Theorem 1.

Step 3: Certificate. On acceptance, both procedures output a grid point of $\mathcal{S}_\varepsilon^\delta$, so both outputs satisfy $\mathcal{L}(\hat{\theta}) \le \varepsilon$ (Theorem 1). The predicate is identical on the two sides, and only the search mechanism differs.

Step 4: Optional Repetition for Distributional Probes. Repeating either procedure yields additional independent samples from the same conditional distribution on $\mathcal{S}_\varepsilon^\delta$ (Theorem 1 on the quantum side, and independent accepted draws on the classical side).

G.1.3 What This Comparison Is and Is Not

Acceptance-rejection from the prior conditioned on $f_\varepsilon = 1$ implements the conditional distribution by definition, and Theorem 1 implements the same conditional distribution after amplification, so the two pipelines of Section G.1.2 agree on the acceptance set, on the output law, and on nothing else being claimed.

Remark 21 (Not an SGD Claim)**.** *The pipeline above does not compare against iterative first-order training. That comparison remains the conditional boundary of Proposition 10 together with the certificate discussion of Section D.1.2. Here the classical baseline is only random parameters plus the same indicator.*

Remark 22 (How to Read the Quadratic Gap)**.** *When $p_\varepsilon^{\mathrm{prior}}$ is a constant, both query counts are $O(1)$ and the pipeline mainly shows the matching certificates. When $p_\varepsilon^{\mathrm{prior}}$ is small, the same diagram exhibits where amplitude amplification shortens the retry loop of classical random search, still without any reference to SGD.*

H THICKENING ROUTES

H.1 THICKENING p_ε: OPTIMIZATION ROUTES

The query bound of Theorem 1 leaves little room on the search side, because amplitude amplification is already optimal up to constants among unstructured searches for marked items. Optimization therefore has to attack the mass itself, either by increasing p_ε or by changing the measure against which p_ε is defined. This subsection lists the routes in decreasing order of how directly they attack the SGD boundary of Proposition 10, and closes with the moves that provably do not help.

H.1.1 RESTRICTED CHARTS

The most direct route searches only on a subgrid supported near an identified zero-loss manifold, for example the kernel chart of linear models or the gauge chart of Lemma 8. If the algorithm is promised a chart of intrinsic dimension k on which the loss is identically zero, then every chart point is good, the mass inside the chart equals 1, and $T_Q = O(1)$. The classical caveat is identical. Once the chart is known, a classical draw from the chart also succeeds immediately, so a quantum advantage must then be argued for chart discovery or for charts that are only approximately known. We stress that chart-supported tubular priors (Definition 11) are a different object. They place mass in a tube around $\mathrm{im}(\varphi)$ without promising that the search stays on the exact-zero manifold, and their analysis is Theorem 4. Algorithmic restriction to a promised chart remains deferred.

H.1.2 NON-UNIFORM INITIAL STATES

The formal objects are the coherent prior of Definition 3 and the discrete Gaussian of Definition 4. On the sampler side no new machinery is needed, since Theorem 1 and Proposition 9 already give the leveled quadratic comparison for any prior. The remaining task is geometric, namely proving nontrivial lower bounds on $p_\varepsilon^{\mathrm{prior}}$ under alignment assumptions, and Proposition 11 shows that such bounds are not automatic.

H.1.3 ENLARGE THE GOOD SET

The second route enlarges the good set itself, and three concrete moves are available. Raising ε weakens the predicate and grows $\mathcal{S}_\varepsilon^\delta$ on every architecture. Increasing hidden width changes the ReLU fiber dimensions and the packing prefactors, but these theorems do not prove that the normalized good-set mass increases monotonically with width. Increasing gauge dimension similarly changes the chart prefactor. The two-layer exponent $n = N_0 d_{\mathrm{out}}$ is fixed by the data constraints and does not decrease merely because d_1 increases. All three moves are classical geometry, and they are the primary knobs for pushing the prior mass above the boundary value $C^2/T_{\mathrm{SGD}}(\varepsilon)^2$ of Proposition 10.

H.1.4 HYBRID CLASSICAL REDUCTION

A fourth route splits the work between the two platforms. One solves a subset of the variables classically, for example by closing a linear system in free kernel coordinates or by pinning an activation pattern, and then runs measurement as training on the residual discrete block. The query bound of Theorem 1 applies to the residual block alone, whose good-set fraction can be far larger than the ambient fraction.

H.1.5 WHAT WILL NOT THICKEN p_ε

Replacing amplitude amplification by another unstructured quantum search primitive does not improve the $1/\sqrt{p_\varepsilon}$ dependence, since that dependence is already optimal up to constants (Bennett et al., 1997). Weak-measurement narratives and continuous-time analogues may change constants or implementation details, but they do not remove the need for a thick good set under the measure that defines p_ε.

I EMPIRICAL VALIDATION

This part collects the two empirical checks of the paper. Subsection I.1 validates the theorems of the ladder on finite instances, and Subsection I.2 probes the method on a real-language corpus.

I.1 FINITE-INSTANCE VALIDATION OF THE THEOREMS

The proofs above are the guarantees. Figure 2 reports the finite-instance empirical validation of the theorems that the ladder establishes in the main text.

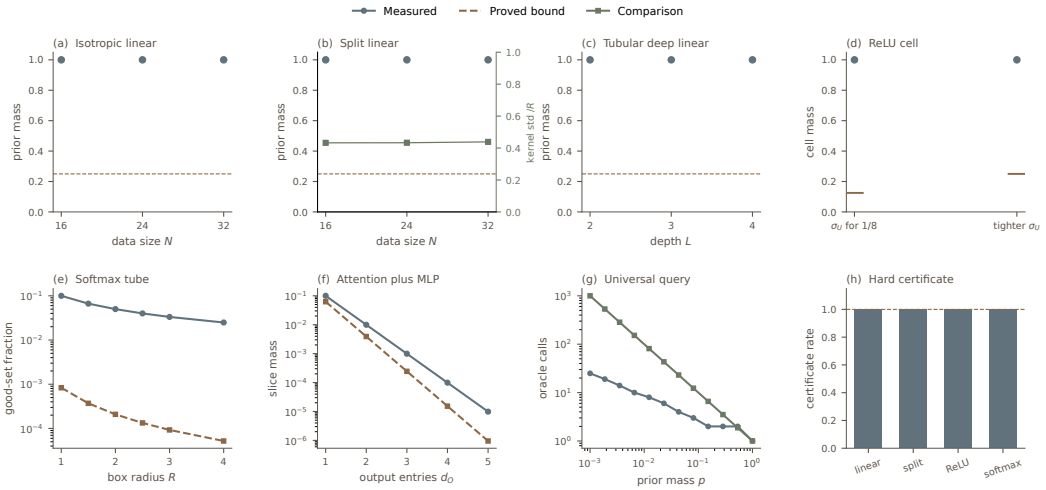

Figure 2: Finite-instance empirical validation of the theorems of the ladder, one panel per theorem. Each panel instantiates only the hypotheses of its theorem, dashed lines and square markers are proved bounds, and whiskers are 99 percent Clopper-Pearson intervals. *Every measured mass lies on or above its proved constant, both conditional bounds are respected as they decay, and every reported output is certified.* **(a), (b) Single Linear Layer.** The isotropic and split prior masses are 1 at $N \in \{16, 24, 32\}$, above the proved $1/4$, and the kernel standard deviation confirms exploration on the stated width scale (Theorem 2). **(c) Deep Linear Models.** The tubular prior mass stays at 1 for $L = 2, 3, 4$, above the proved $1/4$ (Theorem 4). **(d) ReLU Networks.** The measured masses are 1 at both stated widths, above the proved $1/8$ and $1/4$ of Theorem 5. **(e), (f) Conditional Bounds.** The softmax good-set fraction decays with the box radius and the composition slice factor decays with the output dimension, and both stay above their packing bounds, which is the quantitative content of Theorems 6 and 7. **(g) Quadratic Query Advantage.** Oracle calls that reach success probability $2/3$ scale as $p^{-1/2}$ for amplification against p^{-1} for same-prior rejection, the separation of Theorem 1. **(h) Certificate.** On four explicit priors, the probability that a reported output satisfies the hard predicate is 1, so the sampler reports only certified parameters.

I.2 A LANGUAGE-MODEL PROBE ON REAL TEXT

Subsection I.1 validates the theorems on finite instances. This subsection probes the method on real text, where the predicate is the mean next-character cross-entropy, so the squared-loss theorems do not apply. The mass that MaT consumes is manufactured by training: a prior centered at the trained SGD solution has mass one at both dotted thresholds in every cell of the grid, so the schedule accepts on the first draw and returns a certified output, while a prior centered at initialization has an empty marked set. Figure 3 shows the run, and the caption records the full setup.

Two measurements pin the fact down. The initialization-centered prior, measured at depth 2, is never accepted in its draws. A pure gauge move leaves the loss unchanged up to rounding, while a normal displacement of the same Euclidean size moves it, so the flat directions that the aligned priors exploit are present in this model.

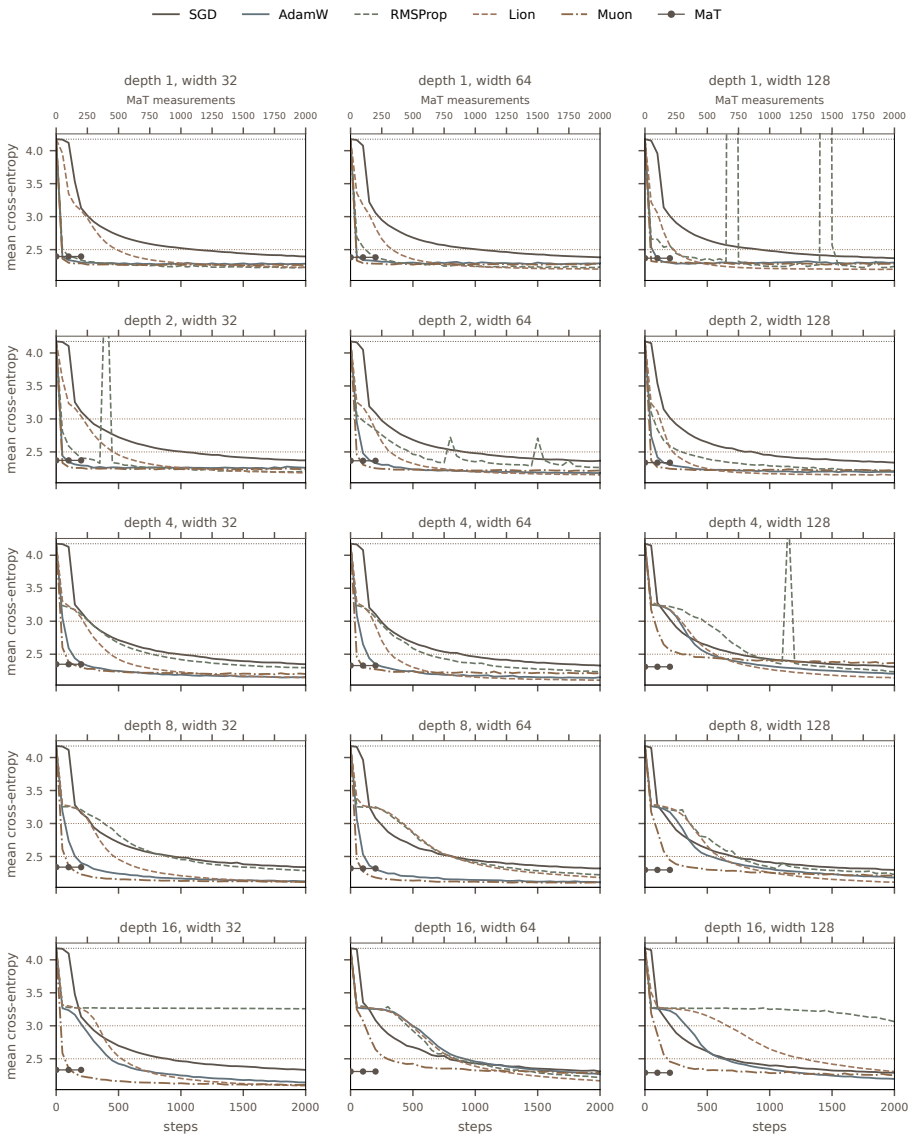

Figure 3: Empirical validation of MaT on a real corpus, one panel per depth and width, with columns for widths and rows for depths. **Setup.** The corpus is Tiny Shakespeare (Karpathy, 2015), 65 symbols, first 32,768 characters, 511 non-overlapping windows of length 64. The model is a fixed-mixing linear attention network (Katharopoulos et al., 2020) with an untied embedding and unembedding, depths 1 to 16, and widths 32 to 128. **Curves.** Five optimizers, SGD, AdamW (Loshchilov & Hutter, 2019), RMSProp (Hinton et al., 2012), Lion (Chen et al., 2023), and Muon (Jordan et al., 2024), train from one shared initialization per cell, with the step size chosen from a fixed candidate list by the lowest full-batch loss after 80 steps, read on the bottom axis of 2,000 optimizer steps. One step reads 32 windows and one draw reads all 511, so the two axes are not matched in compute. **Marks.** The prior is centered at the SGD solution of the cell: each layer is multiplied by a random gauge factor $G = I + (0.25/\sqrt{\text{width}})\, Z$ and compensated by G^{-1} on the output factor, and every trained entry receives independent normal noise of scale 10^{-3}. Marks are the running minimum of the 200 independent draws on the top axis, at counts 1, 100, and 200, one full-slice predicate evaluation per draw. *In every cell the prior mass is one at both dotted thresholds: the first draw is accepted and sits at the SGD loss, later draws lower the minimum by less than 10^{-3}, and the flat cloud never undercuts the other optimizers.* The vertical axis stops just above where training starts, so the RMSProp spike at depth 1 and width 128 leaves the frame.