COMPRESSION SELECTS STRUCTURE: WHY COMPRESSION FAVORS GENERALIZABLE SOLUTIONS IN GROUP-OPERATION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks can fit the same training data yet generalize differently. We study why compression can favor reusable rules, using finite-group operations to test generalization to unseen combinations. We introduce *effective gate complexity* (EGC), measuring the structure needed to reproduce a complete prediction map within a prescribed fixed-depth ReLU architecture. EGC minimizes the number of switching and everywhere-active gates in one-hidden-layer networks with separate operand embeddings and, at arbitrary fixed ReLU depth, records the coordinatewise minimal budgets of affine rank and layerwise ReLU responses retained across equivalent realizations and exact structural reductions. For one-hidden-layer networks, we establish a uniform finite-sample generalization bound in terms of training error and EGC. Every true group law of order Q has EGC $O(\sqrt{Q}\log Q)$, whereas fits that generalize poorly require EGC $\Omega(Q/\log Q)$ with high probability under constant-density random sampling. This separation yields a nonempty compression window for sufficiently large Q and ensures that minimum-EGC fits generalize. We further extend the counting and generalization bounds to multilayer EGC budgets at any fixed ReLU depth. On the stated machine-learning group families at the specified fixed ReLU depths with hidden widths at least logarithmic in Q and bounded width ratios, near-optimality of the quadratically regularized objective controls parameter norms and training error, yielding exact low-EGC realizations of complete prediction maps in the original architecture and hence generalization. Uniformly over an explicit weight-decay interval, an objective-gap tolerance proportional to the decay strength permits sparse sampling with $n = \Theta(Q^{2-\beta})$, yielding held-out error $O(Q^{\beta-1/8})$ with high probability for $0 \le \beta < 1/8$. These guarantees apply throughout the prescribed near-optimal region, without further optimizer-dependent selection. Experiments on modular addition illustrate the connection between lower regularized objective values, smaller verified EGC budgets, and improved held-out accuracy.

1 INTRODUCTION

Modern neural networks can fit the same training data in many ways, ranging from memorizing individual examples to learning reusable rules, yet these solutions can generalize very differently on unseen data [72]. Interpolation alone therefore does not determine generalization [8, 51]. This raises a fundamental selection question: among predictors that fit the observations, what distinguishes those that generalize? One longstanding answer is *compression*. Minimum-description-length (MDL) and Occam-style theories formalize the principle that hypotheses admitting shorter descriptions can generalize under appropriate conditions [58, 9], while modern compression bounds develop related guarantees for neural networks [2, 40, 41]. Such guarantees depend critically on the chosen notion of description: compression can favor the underlying rule only if that rule is compact under the chosen measure while predictors that merely fit the observations are not. We ask when, and why, a *structural notion of compression* produces this separation.

Even if compression identifies a class of generalizable predictors, a separate question remains: why should training produce a predictor in that class? Implicit-bias analyses study optimizer and trajectory effects [62, 25, 42, 11], while explicit regularization changes the objective itself. Quadratic weight

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.

decay favors small-norm solutions and can improve generalization [34], although its effects may depend on the optimizer and parameterization [39, 33, 19, 31]. However, norm and trajectory-based accounts alone do not establish whether every near-optimal solution admits a structurally compressible complete prediction map. We therefore ask whether near-optimality for a quadratically regularized training objective can itself enforce structural compression sufficient for generalization, independently of how that objective value is reached.

In this paper, we study finite-group operations as they make the distinction between memorization and rule reuse explicit [57, 50, 63, 27]. Group laws describe the composition of reversible transformations, such as image rotations and graph-vertex relabelings [12, 45]. The learner observes the products of only a subset of input pairs and predicts the rest. Under independent uniform inputs, either operand alone gives only chance-level accuracy, whereas their pair determines the answer. The finite domain therefore allows us to evaluate every unseen pair and directly test whether compression favors the shared group law over poorly generalizing completions.

We introduce *effective gate complexity* (EGC), which measures the least structural resources needed to reproduce a complete prediction map within a prescribed fixed-depth ReLU architecture. In the one-hidden-layer specialization (Section 2.1), EGC is the fewest output-contributing ReLU gates among all realizations of the same complete prediction map. The count charges both gates that switch across inputs and gates that remain active everywhere, since the latter carry input-affine coefficients. Appendix B formalizes their tradeoff on product groups.

This raises two questions. Can the true group law be realized with substantially lower EGC than every prediction map that fits the observations but generalizes poorly? And can a standard regularizer ensure that its solutions have sufficiently low EGC while maintaining small training error?

Main Contributions. We answer the first question in the one-hidden-layer specialization of EGC and establish a compression window. For every finite group G of order Q, the true group law has EGC $O(\sqrt{Q}\log Q)$. In contrast, under constant-density random sampling, Theorem 3.2 implies that any fit with sufficiently small training error but non-vanishing held-out error must have EGC $\Omega(Q/\log Q)$ with high probability. Hence, these bounds establish a nonempty compression window that contains the true group law while excluding such poorly generalizing fits. Consequently, minimum-EGC fits have vanishing held-out error whenever their training error vanishes as Q grows.

We answer the second question for quadratic weight decay at fixed ReLU depth on the group families specified in Section 4.1. With high probability, every parameter whose regularized objective is sufficiently close to its global minimum has vanishing training error and a complete prediction map admitting an exact low-EGC realization within the original architecture (cf. Theorem 4.1). Consequently, every such near-optimal solution has vanishing held-out error. Theorem 4.2 extends this guarantee to sparse sampling with $n = \Theta(Q^{2-\beta})$ observations for $0 < \beta < 1/8$, uniformly over a polynomial-width interval of weight-decay (WD) strengths, and yields held-out error $O(Q^{\beta-1/8})$. The admissible objective gap is proportional to the WD strength.

Related Works. Compression, MDL, Occam, and PAC–Bayes analyses connect concise descriptions to generalization [58, 9, 46, 15, 2, 75, 40, 41, 68, 38]. Sample compression [7] and post-training pruning or quantization [16] use different reconstruction criteria: we require an implementation in the original architecture to preserve every prediction, including errors. Work on minimal ReLU representations [14], norm and convex formulations [3, 59, 55, 56], and ReLU geometry, identifiability, and expressivity [73, 10, 22, 49, 60, 26, 1, 64, 71, 29] provide complementary notions of complexity.

Our objective-level guarantee differs from optimizer-dependent implicit-bias analyses [62, 24, 25, 42, 11] and from studies of WD dynamics [39, 33, 19, 31]. Appendix I.8 compares the spectral-margin bound of Bartlett et al. [5] with the task-dependent output-row estimates, showing why coarse energy control alone does not yield our polynomial rate.

For modular arithmetic, Morwani et al. [50] characterize full-table margin maximizers. Mohamadi et al. [48] prove partial-table generalization. Tian [65] constructs full-table solutions. These results use quadratic networks, whereas our fixed ReLU architecture admits a guarantee for every near-optimal parameter. Appendix A develops these comparisons and further discusses grokking and algebraic-task theorems [57, 36, 23, 52, 63, 35, 44, 66, 27] and group structure and equivariance [12, 45]. Our proofs use polynomial sign bounds [21, 6], the progression-free-set bound [17], and finite-group representation theory [61].

2 PROBLEM SETUP

Let G be a finite group of order Q, and $X = G \times G$ be the full input domain containing Q^2 labeled input pairs. Throughout the paper, we use $a, b \in G$ to denote group elements. We consider the Q-class classification task on X whose target is the group product $B_G(a, b) = ab \in G$. We draw a training set S uniformly from all subsets of X of size n, where $1 \leq n < Q^2$, and provide the learner with $\{((a, b), B_G(a, b)) : (a, b) \in S\}$. Predictions are evaluated on the held-out set $X \setminus S$. Unless otherwise specified, $n = \lfloor \alpha Q^2 \rfloor$ for a fixed sampling density $\alpha \in (0, 1)$; Section 4.2 additionally considers the sparse-sampling regime $n = \lfloor \alpha_0 Q^{2-\beta} \rfloor$ for fixed $\alpha_0 \in (0, 1)$ and $\beta \in [0, 1/8)$.

2.1 ONE-HIDDEN-LAYER NETWORKS AND EGC

EGC is defined within any prescribed fixed-depth ReLU architecture. We begin with its particularly simple one-hidden-layer scalar form; Section 2.2 gives the general fixed-depth formulation. Fix an ordering of G, and let $e : G \to \mathbb{R}^Q$ denote the standard-basis coordinate map, with $e_a := e(a)$ for any $a \in G$. The network receives the two operands separately as $(e_a, e_b) \in \mathbb{R}^{2Q}$. A one-hidden-layer ReLU network of width m, parameterized by $\theta = (U, V, b_1, W_2)$, has logits

$$F_\theta^m(a, b) = W_2 \sigma(U e_a + V e_b + b_1) = \sum_{j=1}^{m} c_j \, \sigma(u_j(a) + v_j(b) + \beta_j), \tag{2.1}$$

where $U, V \in \mathbb{R}^{m \times Q}$ are learned operand embeddings, $W_2 \in \mathbb{R}^{Q \times m}$, $b_1 \in \mathbb{R}^m$, $\theta \in \Theta_m := \mathbb{R}^{2m \times Q} \times \mathbb{R}^m \times \mathbb{R}^{Q \times m}$, and the ReLU activation function $\sigma(t) = \max\{t, 0\}$ acts coordinatewise. Here $u_j(a) = (U e_a)_j$, $v_j(b) = (V e_b)_j$, $\beta_j = (b_1)_j$, and $c_j = (W_2)_{:j}$. Each summand represents one ReLU gate and its contribution to the output. We assume that the model has no output bias or direct affine input-to-output term. A gate is *effective* if its output contribution is not identically zero on X. For each integer $m \geq 0$, define

$$\mathcal{S}_m := \bigcup_{k=0}^{m} \left\{ F_\theta^k : \theta \in \Theta_k \right\}, \tag{2.2}$$

where $\Theta_0 = \{\varnothing\}$ and $F_\theta^0 : X \to \{\mathbf{0}\}$ denotes the zero map. $\mathcal{S}_m$ consists exactly of the logit maps realizable with at most m effective gates. Furthermore, padding any narrower network with ineffective gates gives the equivalent characterization $\mathcal{S}_m = \{F_\theta^m : \theta \in \Theta_m\}$. Let $\mathcal{S} = \bigcup_{m \in \mathbb{N}_0} \mathcal{S}_m$.

We call $B : X \to G$ a *complete prediction map* as its input domain is the full domain X. Each logit map induces the complete prediction map $B_{F_\theta^m}(a, b) \coloneqq \arg\max_{g \in G} \big(F_\theta^m(a, b)\big)_g$ for every pair $(a, b) \in X$, with ties resolved according to a fixed ordering of G.

Definition 2.1 (Effective Gate Complexity (EGC) for one-hidden-layer ReLU networks)**.** For a complete prediction map $B : X \to G$ realizable by one-hidden-layer models in $\mathcal{S}$, its EGC is

$$\kappa(B) := \min\big\{ m \in \mathbb{N}_0 : \exists F \in \mathcal{S}_m, \; B_F = B \big\}. \tag{2.3}$$

Thus, EGC is the smallest number of effective gates needed to reproduce B on the entire input domain. The minimization preserves prediction labels, but need not preserve weights or logits. It therefore removes redundancy such as dead, duplicated, or canceling gates. The standard-basis vectors merely index two unrestricted learned embedding tables; any fixed full-column-rank categorical encoding gives the same function class after reparameterizing the first layer. The separate operand structure is the substantive choice: each gate uses $O(Q)$ coefficients shared across Q^2 input pairs. With a distinct one-hot vector for each pair, two counted ReLU gates suffice to realize any complete prediction map (Appendix B). Furthermore, an EGC budget $m \in \mathbb{N}_0$ specifies a class of prediction maps $\mathcal{B}_m := \{B : X \to G : \kappa(B) \leq m\}$ for one-hidden-layer networks

Both *switching* and *everywhere-active* gates contribute to EGC. A gate is *switching* if its scalar preactivation $u(a) + v(b) + \beta$ takes both positive and negative values on X, and *everywhere-active* if its scalar preactivation is nonnegative throughout X. Everywhere-active gates implement input-affine functions of the form $u(a) + v(b) + \beta$ and must be counted because the model has no separate affine output term. Counting both types captures the tradeoff between nonlinear switching and affine representation cost; Appendix B develops this tradeoff, and Appendix C.2 gives the corresponding rank-one characterization.

2.2 MULTILAYER NETWORKS AND EGC

We now extend EGC to multilayer ReLU networks. Unlike the one-hidden-layer setting, successive layers compose and reuse features, so we track the retained structure at each layer rather than using a single gate count. Consider a ReLU architecture $\mathcal{A}$ with $L \geq 1$ hidden layers of widths $k_1, \ldots, k_L$. Let $\Theta_{\mathcal{A}}$ be its parameter space. Set $D = L + 1$, $k_0 = 2Q$, and $k_D = Q$. For parameters $\theta = (W_i, b_i)_{i=1}^{D} \in \Theta_{\mathcal{A}}$, the network is

$$h_0 = (e_a, e_b), \qquad h_i = \sigma(W_i h_{i-1} + b_i) \quad (1 \leq i < D), \qquad F_\theta = W_D h_{D-1} + b_D, \tag{2.4}$$

where $W_i \in \mathbb{R}^{k_i \times k_{i-1}}$ and $b_i \in \mathbb{R}^{k_i}$. Here $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius and Euclidean norms. Hidden biases are allowed and b_D is either allowed or fixed to zero as specified by $\mathcal{A}$. Similarly, the induced complete prediction map is $B_\theta(a, b) \coloneqq \arg\max_{g \in G} \big(F_\theta(a, b)\big)_g$ for every pair $(a, b) \in X$.

Removing redundant computation. Before defining EGC, we first simplify a network realization without changing its logits on X. A neuron response may be eliminated if, throughout X, it equals an input-affine function plus a linear combination of retained responses from earlier layers or earlier in a fixed ordering within its layer. Substituting this identity into all subsequent uses removes the response while retaining its contribution. This includes eliminating zero responses, combining proportional responses, and replacing everywhere-active ReLUs by their preactivations. All finite sequences of these reductions, including none, are allowed; Appendix H.1 gives the details.

For a chosen realization and reduction, let t_i be the number of retained ReLU responses from original layer $i = 1, \ldots, L$, and collect them into $g_i : X \to \mathbb{R}^{t_i}$. Substitution collects input-affine terms into the remaining preactivations and output, while allowing retained responses to contribute directly to later layers. After subtracting the mean logit, the first $q = Q - 1$ centered coordinates thus admit

$$F(a, b) = M(a, b) + \sum_{i=1}^{L} C_i g_i(a, b), \quad g_i(a, b) = \sigma\left(U_i(a, b) + \sum_{j<i} A_{ij} g_j(a, b) \right), \; 1 \leq i \leq L. \tag{2.5}$$

Here $U_i : X \to \mathbb{R}^{t_i}$ and $M : X \to \mathbb{R}^q$ are input-affine maps, whose components have the form $u(a) + v(b) + \beta$, and $A_{ij} \in \mathbb{R}^{t_i \times t_j}$ and $C_i \in \mathbb{R}^{q \times t_i}$. Let $M_{\mathrm{coef}} \in \mathbb{R}^{q \times (2Q-1)}$ be the coefficient matrix of M in any basis of the input-affine function space, and define $t_0 \coloneqq \mathrm{rank}(M_{\mathrm{coef}})$; this rank is basis-independent. Predictions use all Q centered logits with the original tie rule, reconstructing the omitted coordinate as minus the sum of the coordinates of F.

These coordinates $t_0, t_1, \ldots, t_L$ play a role analogous to the effective gate count in the one-hidden-layer setting, but measure retained affine rank and ReLU responses separately. In Proposition 3.4, we show that the minimum attainable $t_0 + t_1$ over the shallow model class agrees with $\kappa(B)$ up to a factor of two.

Definition 2.2 (EGC for multilayer ReLU networks)**.** For a realizable complete prediction map $B : X \to G$, let $\mathcal{T}_{\mathcal{A}}(B)$ be the set of vectors $\mathbf{t} = (t_0, \ldots, t_L)$ attainable by the reductions above from any $\theta \in \Theta_{\mathcal{A}}$ satisfying $B_\theta = B$ on X. Define

$$\mathrm{EGC}_{\mathcal{A}}(B) \coloneqq \{\mathbf{t} \in \mathcal{T}_{\mathcal{A}}(B) : \nexists\, \mathbf{s} \in \mathcal{T}_{\mathcal{A}}(B) \text{ with } \mathbf{s} \preceq \mathbf{t} \text{ and } \mathbf{s} \neq \mathbf{t}\},$$

where $\mathbf{s} \preceq \mathbf{t}$ means $s_i \leq t_i$ for all i. Thus, EGC consists of the attainable vectors that cannot be improved in one coordinate without worsening another.

With zero affine budget, two responses in different layers realize prediction maps unattainable with both responses in one layer. Appendix H.2 also gives two incomparable minimal budgets for one map, exhibiting a tradeoff between affine and nonlinear structure. For a budget $\mathbf{z} = (\tau_0, \ldots, \tau_L) \in \mathbb{N}_0^{L+1}$, we use the notation

$$\mathrm{EGC}_{\mathcal{A}}(B) \preceq \mathbf{z} \quad \Longleftrightarrow \quad \exists\, \mathbf{t} \in \mathrm{EGC}_{\mathcal{A}}(B) \text{ such that } t_i \leq \tau_i \quad (0 \leq i \leq L). \tag{2.6}$$

Equivalently, one realization of B in $\mathcal{A}$ admits a reduction satisfying all budget constraints simultaneously. Unlike the shallow definition in Definition 2.1, which allows arbitrary finite widths, this definition keeps $\mathcal{A}$ fixed. Similarly, an EGC budget $\mathbf{z}$ specifies a class of prediction maps $\mathcal{B}_{\mathcal{A}}(\mathbf{z}) \coloneqq \{B : X \to G : \mathrm{EGC}_{\mathcal{A}}(B) \preceq \mathbf{z}\}$ for multilayer networks with fixed architecture $\mathcal{A}$.

2.3 Errors and the Compression Window

Given a uniformly sampled training set S and a complete prediction map $B : X \to G$, let $E_B := \{(a, b) \in X : B(a, b) \neq ab\}$ be the error set. The training and held-out error rates are

$$\mathrm{err}_{\mathrm{in}}(B) = \frac{|E_B \cap S|}{|S|}, \qquad \mathrm{err}_{\mathrm{out}}(B) = \frac{|E_B \setminus S|}{Q^2 - |S|}. \tag{2.7}$$

Exact interpolation corresponds to $\mathrm{err}_{\mathrm{in}}(B) = 0$. For either class of prediction maps $\mathcal{H} \in \{\mathcal{B}_m, \mathcal{B}_{\mathcal{A}}(\mathbf{z})\}$ and tolerances $\varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}} \geq 0$, the corresponding EGC budget lies in the *compression window* if

$$\varnothing \neq \{B \in \mathcal{H} : \mathrm{err}_{\mathrm{in}}(B) \leq \varepsilon_{\mathrm{in}}\} \subseteq \{B \in \mathcal{H} : \mathrm{err}_{\mathrm{out}}(B) < \varepsilon_{\mathrm{out}}\}. \tag{2.8}$$

Thus, the budget admits a fit with the prescribed training accuracy while excluding all such fits with held-out error at least $\varepsilon_{\mathrm{out}}$. Setting $\varepsilon_{\mathrm{in}} = 0$ gives the interpolation window. Appendix B provides the equivalent threshold formulation for the one-hidden-layer setting.

3 Why Compression Forces Generalization

We now show how structural compression, measured by EGC, controls generalization. For one-hidden-layer networks, we bound the number of prediction maps under a given EGC budget and derive a uniform generalization guarantee. We then show that the true group law has sufficiently low EGC to establish a nonempty compression window. Finally, we extend the counting and generalization bounds to multilayer networks.

3.1 Generalization from One-Hidden-Layer EGC

Recall that $\mathcal{B}_m = \{B : X \to G : \kappa(B) \leq m\}$ is the class of complete prediction maps with EGC at most m realizable by the one-hidden-layer models $\mathcal{S}$. We first bound the size of this class.

Theorem 3.1 (Counting prediction maps under an EGC budget)**.** *There is a universal constant C such that, for $m \in \mathbb{N}^*$,*

$$\log |\mathcal{B}_m| \leq CmQ \log(2mQ). \tag{3.1}$$

We sketch the proof here, while Appendix E gives the details, including the treatment of zero signs and the fixed tie rule. Each gate has $O(Q)$ parameters. Conditional on its activation pattern, the network logits are polynomial in the parameters, and their pairwise differences determine the predictions. Counting activation patterns and the signs of these differences yields Theorem 3.1.

Theorem 3.1 bounds the number of complete prediction maps at each EGC budget. Applying a finite-population sampling bound to these classes and summing over budgets gives the following generalization guarantee, including for maps selected after observing the training data.

Theorem 3.2 (Finite-sample generalization from EGC)**.** *Fix $Q \geq 2$, $\delta \in (0, 1)$ and $\alpha \in (0, 1)$ such that $1 \leq n = \lfloor \alpha Q^2 \rfloor < Q^2$. For a uniformly sampled n-element training set, there is a constant $C_\alpha > 0$, depending only on α, such that, with probability at least $1 - \delta$, simultaneously for every complete prediction map B realizable in $\mathcal{S}$,*

$$\mathrm{err}_{\mathrm{out}}(B) \leq C_\alpha \left[\mathrm{err}_{\mathrm{in}}(B) + \frac{(\kappa(B) + 1) \log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2} \right]. \tag{3.2}$$

The bound holds uniformly over B, so it also applies to prediction maps selected after observing S. Its proof combines the counting bound with the fact that uniform sampling is unlikely to miss a large fraction of a fixed prediction map's errors. Appendix D gives the proof and refinements for general sample sizes and exact interpolation.

Conversely, the bound in Theorem 3.2 implies that fitting the training data while generalizing poorly requires large EGC. On the same event, every B with $\mathrm{err}_{\mathrm{in}}(B) \leq \varepsilon_{\mathrm{in}}$ and $\mathrm{err}_{\mathrm{out}}(B) \geq \varepsilon_{\mathrm{out}}$ satisfies

$$\kappa(B) + 1 \geq \frac{Q}{\log(2Q)} \left[\frac{\varepsilon_{\mathrm{out}}}{C_\alpha} - \varepsilon_{\mathrm{in}} - \frac{\log(1/\delta)}{Q^2} \right] \tag{3.3}$$

For fixed α and error tolerances satisfying $\varepsilon_{\mathrm{in}} < \varepsilon_{\mathrm{out}}/C_\alpha$, choosing $\delta = Q^{-2}$ gives $\kappa(B) = \Omega(Q/\log Q)$ with high probability. We next show that the true group law has substantially smaller EGC than this lower bound.

3.2 EGC OF GROUP LAWS AND THE COMPRESSION WINDOW

The preceding result shows that poorly generalizing fits require large EGC. We next prove that the true group law has substantially smaller EGC, so a suitable budget can retain the target while excluding such fits.

Theorem 3.3 (Universal EGC upper bound for group laws)**.** *There is a universal constant $C > 0$ such that, for every finite group G of order $Q \geq 2$, the target group law $B_G : X \to G$ defined by $B_G(a, b) = ab$ satisfies*

$$\kappa(B_G) \leq C\sqrt{Q}\log(2Q). \tag{3.4}$$

Appendix F gives the full proof. The construction uses a finite family of matrix representations, which encode group multiplication as matrix multiplication, with dimensions at most $\sqrt{Q}$. Representations are sampled with probabilities proportional to their squared dimensions. Each copy uses $O(\sqrt{Q})$ real sign products. Averaging $O(\log Q)$ copies gives a positive margin on X, and four ReLU gates per product give the bound.

The compression window. Let $m_G := \lceil C\sqrt{Q}\log(2Q)\rceil$, with C from Theorem 3.3. Since $B_G \in \mathcal{B}_{m_G}$ has zero training error, this budget admits a fit for every tolerance $\varepsilon_{\text{in}} \geq 0$. On the event in Theorem 3.2, every $B \in \mathcal{B}_{m_G}$ with $\text{err}_{\text{in}}(B) \leq \varepsilon_{\text{in}}$ satisfies

$$\text{err}_{\text{out}}(B) \leq C_\alpha \left[\varepsilon_{\text{in}} + \frac{(m_G + 1)\log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2}\right].$$

Whenever the right-hand side is below ε_{out}, m_G therefore lies in the compression window. This gives a sufficient condition at finite Q. For fixed α and $\varepsilon_{\text{out}} > 0$, sufficiently small ε_{in}, and $\delta = Q^{-2}$, the target upper bound $O(\sqrt{Q}\log Q)$ and the bad-fit lower bound $\Omega(Q/\log Q)$ establish a *nonempty compression window* for all sufficiently large Q. Theorem B.3 gives the full threshold statement.

Since B_G is feasible, every minimum-EGC fit $\widehat{B} \in \arg\min_{B:\text{err}_{\text{in}}(B)\leq\varepsilon_{\text{in}}} \kappa(B)$ satisfies $\kappa(\widehat{B}) \leq \kappa(B_G)$ and obeys the preceding bound with m_G replaced by $\kappa(B_G)$. With $\delta = Q^{-2}$, the structural term is $O((\log Q)^2/\sqrt{Q})$ for arbitrary finite groups. For finite abelian groups, Appendix C.4 gives $\kappa(B_G) \leq 2\log_2 Q + 2$, and for cyclic groups $\kappa(B_G) \leq 4$. Their structural terms improve to $O((\log Q)^2/Q)$ and $O(\log Q/Q)$, respectively.

3.3 GENERALIZATION FROM MULTILAYER EGC

We next apply the counting and generalization argument to the joint budgets of fixed-depth EGC. Their one-hidden-layer specialization first recovers the scalar EGC up to a constant factor.

Proposition 3.4 (Comparison with one-hidden-layer EGC)**.** *For a complete prediction map B realizable in $\mathcal{S}$, let $c(B)$ be the minimum attainable $t_0 + t_1$ over all one-hidden-layer realizations of B with arbitrary finite widths and all allowed exact reductions. Then*

$$\tfrac{1}{2}\kappa(B) \leq c(B) \leq \kappa(B). \tag{3.5}$$

Thus, the minimum total structural budget agrees with the one-hidden-layer EGC up to a factor of two. Appendix H.1 provides the proof.

For multilayer networks, we similarly bound the number of prediction maps in $\mathcal{B}_{\mathcal{A}}(\mathbf{z})$ under a structural budget $\mathbf{z} = (\tau_0, \ldots, \tau_L) \in \mathbb{N}_0^{L+1}$. Let $d = 2Q - 1$ be the dimension of the input-affine function space and recall $q = Q - 1$. Define $V(\mathbf{z}) := (d + q)\|\mathbf{z}\|_1 + \sum_{1\leq j<i\leq L} \tau_j\tau_i + \sum_{i=1}^{L} \tau_i + 1$ and $\mathscr{H}_Q(\mathbf{z}) := C_L V(\mathbf{z})\log(2 + C_L Q^2(Q^2 + \sum_{i=1}^{L} \tau_i + 1))$, where $C_L > 0$ depends only on L. Applying polynomial sign bounds to the reduced expression (2.5) yields

$$\log|\mathcal{B}_{\mathcal{A}}(\mathbf{z})| \leq \mathscr{H}_Q(\mathbf{z}), \tag{3.6}$$

uniformly over the original network architecture. Appendix H.3 provides the parameter count underlying $V(\mathbf{z})$ and the full proof. Here $\mathscr{H}_Q(\mathbf{z})$ plays the same role as the shallow counting bound in Theorem 3.1: it controls capacity from an attainable structural budget. Combining this bound in (3.6) with uniform sampling yields the multilayer counterpart of Theorem 3.2.

Theorem 3.5 (Generalization from multilayer EGC). *Fix $Q \geq 2$, $L \in \mathbb{N}^*$, a sampling density $\alpha \in (0,1)$ satisfying $1 \leq \lfloor \alpha Q^2 \rfloor < Q^2$, and a failure probability $\delta \in (0,1)$. Set $n = \lfloor \alpha Q^2 \rfloor$. For a uniformly sampled n-element training set, there is a constant $C_{\alpha,L} > 0$, depending only on α and L, such that, with probability at least $1-\delta$, simultaneously for every architecture $\mathcal{A}$ with L hidden layers, every EGC budget $\mathbf{z}$, and every $B \in \mathcal{B}_{\mathcal{A}}(\mathbf{z})$,*

$$\mathrm{err}_{\mathrm{out}}(B) \leq C_{\alpha,L}\left[\mathrm{err}_{\mathrm{in}}(B) + \frac{\mathscr{H}_Q(\mathbf{z}) + \log(1/\delta)}{Q^2}\right]. \tag{3.7}$$

The uniform guarantee applies even when the architecture, budget, and prediction map are selected after observing S. Provided that $\mathcal{B}_{\mathcal{A}}(\mathbf{z})$ contains a map with training error at most $\varepsilon_{\mathrm{in}}$, the budget $\mathbf{z}$ lies in the *compression window* whenever the right-hand side of (3.7) with $\mathrm{err}_{\mathrm{in}}(B) = \varepsilon_{\mathrm{in}}$ is smaller than $\varepsilon_{\mathrm{out}}$. For fixed α and L, choosing $\delta = Q^{-2}$ shows that vanishing training error and $\mathscr{H}_Q(\mathbf{z}) = o(Q^2)$ imply vanishing held-out error in probability. For cyclic addition, the dependence on the structural budget and sample size is sharp up to a logarithmic factor. Fix the depth, a bound on the hidden-width ratio, a positive affine budget, and $\alpha_0 \in (0,1)$, and let $J = \sum_i \min\{\tau_i, k_i\}$. Under the finite range $C\log(4Q/\delta) \leq J \leq Q$ and $1 \leq n \leq \alpha_0 Q^2$, Corollary E.6 shows that, with probability at least $1-\delta$, the largest held-out error among training-consistent behaviors in the original-architecture budget class lies between $c\min\{1, QJ/n\}$ and $\min\{1, C'[QJ\log Q + \log(2/\delta)]/n\}$. Here c, C, C' depend only on the fixed depth, width ratio, affine budget, and α_0. The next section establishes both small training error and sufficiently small structural budgets from near-optimality of the quadratically regularized objective under the stated conditions.

4 Weight Decay as a Compression Mechanism

Section 3 shows that small training error and low EGC guarantee generalization. We now show that quadratic weight decay enforces both conditions throughout a quantitative near-optimal objective region, at any fixed ReLU depth.

4.1 Compression from Regularized Objective Near-Optimality

To explain generalization in regularized training, we focus on group structures that recur in machine-learning tasks, such as modular arithmetic, permutations, and discrete spatial symmetries [52, 63, 12]. For the full families specified below, including their product and hierarchical compositions [69], we prove that WD near-optimality implies low exact EGC and hence generalization.

Fix $M \in \mathbb{N}^*$ with $M \geq 48$, and let $G \in \mathscr{G}_{\mathrm{ML}}(M)$ be a group of order Q, where the class is defined in Appendix J.5. It contains all finite abelian, symmetric, alternating, and dihedral groups, the specified periodic lattice symmetry groups, and all finite direct and permutation wreath products of these groups. Set $\rho = 1$ for abelian G and $\rho = 1 + 5/D$ otherwise, with $D \geq 2$ and $D \geq 3$, respectively. Fix a ReLU architecture $\mathcal{A}$ as in Section 2.2, with affine depth $D = L + 1$ and hidden widths satisfying $\min_i k_i \geq C_{D,M}\log Q$ and $\frac{\max_i k_i}{\min_i k_i} \leq C_{\mathrm{w}}$, where $C_{D,M} > 0$ is sufficiently large and depends only on D, M, and C_{w} is independent of Q. For abelian G, the width constant depends only on D. Hidden biases are included; the output bias is either included and penalized or fixed to zero, as specified by $\mathcal{A}$. For $\lambda > 0$ and a uniformly sampled training set S of size $n = \lfloor \alpha Q^2 \rfloor$, with fixed $\alpha \in (0,1)$, define

$$J(\theta) = \mathcal{L}_S(F_\theta) + \frac{\lambda}{2}\sum_{i=1}^{D}\left(\|W_i\|_F^2 + \|b_i\|_2^2\right), \quad \Delta(\theta) = J(\theta) - J^*_{\mathcal{A}} = J(\theta) - \min_{\vartheta \in \Theta_{\mathcal{A}}} J(\vartheta), \tag{4.1}$$

where $\mathcal{L}_S(F) = -\frac{1}{n}\sum_{(a,b)\in S}\log\left[\exp\big((F(a,b))_{ab}\big)/\sum_{c\in G}\exp\big((F(a,b))_c\big)\right]$ is the average multi-class cross-entropy. Here $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius and Euclidean norms. The gap $\Delta(\theta)$ measures excess regularized objective value within the fixed architecture $\mathcal{A}$.

Layer norms control approximation error, while training cross-entropy bounds the frequency of small correct-class margins; together, these controls yield exact EGC compression within the original architecture. The following theorem derives these controls from near-optimality of the quadratically regularized objective and applies independent EGC counting to obtain generalization, using Lemma I.4 in Appendix I.7 for $D \geq 3$ and the corresponding shallow argument for $D = 2$.

Theorem 4.1 (Exact compression and generalization throughout a near-optimal region)**.** *Let* $\ell = \log Q$*, fix* $0 < \delta < 1$*, and set*

$$\lambda = Q^{-1/4-15/(8D)}\ell^{-(2+\rho-3/D)}, \qquad \Delta_Q = nQ^{-5/2+1/(8D)}\ell^{-(\rho D+6-3/D)}. \tag{4.2}$$

For every $Q \geq 3$ *satisfying the finite conditions in Appendix J.5, with probability at least* $1-\delta-Q^{-3}-Q^{-4}$*, simultaneously for every allowed architecture* $\mathcal{A}$ *and every* $\theta \in \Theta_{\mathcal{A}}$ *with* $\Delta(\theta) \leq \Delta_Q$*, the complete prediction map has an attainable EGC budget* $\mathbf{z}_\theta = (t_0, \ldots, t_L)$ *in its original architecture satisfying*

$$t_0 \leq 1, \qquad \max_{1\leq i<D} t_i \leq C' \frac{Q^{7/8}}{\log Q}. \tag{4.3}$$

Consequently,

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq CQ^{-1/8} + C\frac{\log(1/\delta)}{Q^2}. \tag{4.4}$$

The constants $C, C' > 0$ *depend only on* $D, M, \alpha, C_{\mathrm{w}}$ *and are independent of the hidden widths; for abelian groups they are independent of* M*.*

The guarantee applies to every global minimizer and every parameter within the prescribed objective gap, regardless of how it is obtained. Here we give a proof sketch. Group structure enters through the parameter norm (a.k.a. energy) required to fit the target. The group law admits a low-loss realization with controlled quadratic energy, which bounds the regularized optimum and hence the energy and training loss throughout the stated near-optimal region. Together with the output-row control derived from the task and objective, these bounds imply exact EGC compression. The independent counting theorem then converts this restriction on complete prediction maps into a generalization guarantee. Appendix L proves a quadratic-energy separation between group labels and independent uniform labels at equal cross-entropy tolerance in a common ReLU architecture.

4.2 Sparse observations

We extend Theorem 4.1 to sparse sampling and an interval of WD strengths, using the same group class, architecture, objective, and exponent ρ. Let $n = \lfloor \alpha_0 Q^{2-\beta} \rfloor$, where $0 \leq \beta < 1/8$ and $0 < \alpha_0 < 1$. For $\beta > 0$, the observed fraction $n/Q^2 = \Theta(Q^{-\beta})$ vanishes as Q grows.

Theorem 4.2 (Uniform WD interval and sparse sampling through exact EGC)**.** *Fix* $0 < \omega < 1/8 - \beta + 1/(8D)$ *and* $0 < \delta < 1$*, and require*

$$\begin{aligned} \lambda_0 &= Q^{-1/4-15/(8D)}(\log Q)^{-(2+\rho-3/D)}, & \lambda_0 Q^{-\omega} \leq \lambda \leq \lambda_0, \\ \Delta(\theta) &\leq \Delta_{\mathrm{tol}}(\lambda) := n\lambda Q^{2/D-9/4}(\log Q)^{-[\rho(D-1)+4]}. \end{aligned} \tag{4.5}$$

For every Q *satisfying the finite conditions in Appendix G.10 for* $\rho = 1$ *and Appendix J.5 otherwise, with probability at least* $1 - \delta - Q^{-3} - Q^{-4}$*, simultaneously for every allowed architecture and every* λ *in this interval, every such parameter has the exact EGC budget in* (4.3)*. Independent EGC counting then gives*

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq CQ^{\beta-1/8} + C\frac{\log(1/\delta)}{n}. \tag{4.6}$$

The constants $C, C' > 0$ *depend only on* $D, M, \alpha_0, \beta, \omega, C_{\mathrm{w}}$*. The interval has endpoint ratio* Q^ω*, and* $\Delta_{\mathrm{tol}}(\lambda)$ *is linear in* λ*.*

For $0 < \beta < 1/8$, the exact-compression cost grows more slowly than the sample size, so the error bound vanishes at fixed confidence throughout the WD interval despite a decreasing observed fraction. Theorem I.1 gives sharper abelian rates and further objective tolerances.

Corollary J.2 gives a continuous tradeoff between objective accuracy and exact compression within the same interval; its error bound is nonincreasing in λ at fixed objective excess.

4.3 Empirical illustration

We train two-hidden-layer ReLU networks of widths $(64, 64)$ on 30% of $\mathbb{Z}_{127}^2$, pairing ten seeds across $\lambda \in 10^{-4}\{0, 0.5, 1, 2, 5\}$. All fit training perfectly; Figure 1(a) compares held-out accuracy.

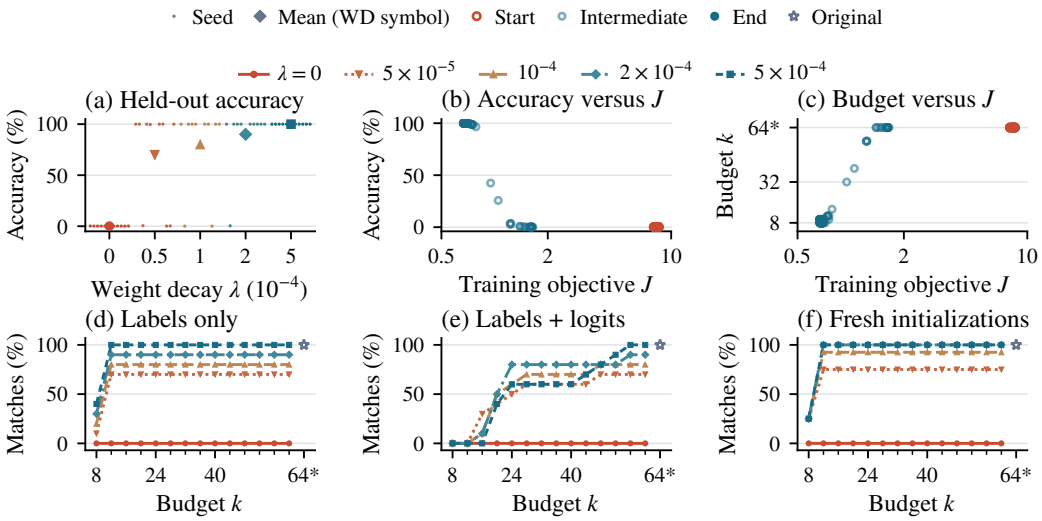

Figure 1: Weight decay, generalization, and exact prediction matching on $\mathbb{Z}_{127}$.

We match each network's complete prediction map at EGC budget $(1, k, k)$, $k \in \mathbb{N}^*$: affine rank at most one and at most k retained nonlinear responses per hidden layer. Exact matching preserves all 16,129 predictions, including errors, within the original architecture. Figure 1(a) displays ten paired-seed accuracies and their means; the WD strengths occupy equally spaced positions on its categorical axis. Panels (b,c) use all ninety checkpoints, with symbols marking starts, intermediate checkpoints, and endpoints. Panels (d–f) give cumulative exact-matching rates across every tested budget on a linear k axis. 64^* marks original realizations.

For (b,c), each unregularized network continues training on its original training set at fixed $\lambda = 5 \cdot 10^{-4}$; all objectives use this strength. Endpoints have lower J, higher accuracy, and verified bounds 8 or 10. Identical complete prediction maps share their best verified bound in (c).

Panel (d) fits labels alone; (e) adds a centered-logit discrepancy. At $k = 12$, label-only matches across the five increasing strengths are $0, 7, 8, 9, 10$ out of ten. Panel (f) repeats training with four fresh initializations per training set, yielding $0, 30, 37, 40, 40$ matches out of forty. At $\lambda = 5 \cdot 10^{-4}$, all ten networks also match by $k = 56$ with the logit loss. Appendix K gives protocols and full records.

5 Conclusion

EGC is the minimum structural complexity of realizing a complete prediction map in the prescribed architecture. On group tasks, the shallow compression window separates compact laws from poorly generalizing fits. For the stated machine-learning group families and fixed ReLU depths, every parameter in the stated near-optimal quadratic-WD region admits an exact low-EGC realization in the original architecture, and independent EGC counting gives generalization. Over the stated WD interval, the admissible gap grows linearly with decay strength; for $n = \Theta(Q^{2-\beta})$, the same mechanism gives $O(Q^{\beta-1/8})$ error when $0 \le \beta < 1/8$.

Thus the regularized objective itself selects a generalizing region: after the gap threshold is reached, no further optimizer-dependent selection is needed. The experiments illustrate this finite-scale connection.

Future work can determine when training trajectories cross this objective-gap threshold. A second direction is to extend architecture-preserving EGC constructions and counting to broader compositional tasks and model families. The prediction-map viewpoint also supports comparing regularizers and algorithms through the maps they select.

AI USE STATEMENT

We used generative AI tools for mathematical exploration, including formulating candidate claims, proposing explicit constructions, and developing technical proof arguments. We also used them to assist with experimental design, implementation and analysis, literature retrieval, and writing. The authors formulated the research questions and introduced effective gate complexity and its extension to deep networks. The viewpoints based on regularized energy and the complexity of complete prediction behavior, together with several proof ideas, also originated with the authors. The authors assessed the AI-generated proposals, checked, revised, and completed the mathematical arguments, verified cited sources, and made the final scientific and mathematical judgments. We reviewed all AI-assisted work and take responsibility for the final text, claims, proofs, code, and experimental results.

REPRODUCIBILITY STATEMENT

Proofs and explicit constructions are provided in Appendices B–J and L. Appendix K describes the task, data splits, architectures, initialization, training settings, random seeds, and reconstruction searches. It also specifies the numerical procedure for verifying agreement with the original network on the complete input domain. The reported comparisons include every seed in each experimental condition.

REFERENCES

[1] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

[2] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, 2018.

[3] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[4] Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764, 2022.

[5] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html.

[6] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[7] Mathieu Bazinet, Valentina Zantedeschi, and Pascal Germain. Sample compression unleashed: New generalization bounds for real valued losses. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3286–3294. PMLR, 2025. URL https://proceedings.mlr.press/v258/bazinet25a.html.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

[10] Marie-Charlotte Brandenburg, Moritz Grillo, and Christoph Hertrich. Decomposition polyhedra of piecewise linear functions. *Discrete & Computational Geometry*, 2026. doi: 10.1007/s00454-026-00875-1.

[11] Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 1305–1338, 2020.

[12] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, 2016.

[13] Francesco D'Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/29496c942ed6e08ecc469f4521ebfff0-Abstract-Conference.html.

[14] Steffen Dereich and Sebastian Kassing. On minimal representations of shallow ReLU networks. *Neural Networks*, 148:121–128, 2022. doi: 10.1016/j.neunet.2022.01.006.

[15] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

[16] Houssam El Cheairi, David Gamarnik, and Rahul Mazumder. Theoretical compression bounds for wide multilayer perceptrons. In *Proceedings of the Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2200–2258. PMLR, 2026. URL https://proceedings.mlr.press/v336/el-cheairi26a.html.

[17] Jordan S. Ellenberg and Dion Gijswijt. On large subsets of $\mathbb{F}_q^n$ with no three-term arithmetic progression. *Annals of Mathematics*, 185(1):339–343, 2017. doi: 10.4007/annals.2017.185.1.8.

[18] Hamza Fawzi and Pablo A. Parrilo. Self-scaled bounds for atomic cone ranks: Applications to nonnegative rank and CP-rank. *Mathematical Programming*, 158(1–2):417–465, 2016. doi: 10.1007/s10107-015-0937-7. URL https://arxiv.org/abs/1404.3240.

[19] Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso A. Poggio. SGD with weight decay secretly minimizes the ranks of your neural networks. In *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 1388–1412. PMLR, 2025. URL https://proceedings.mlr.press/v280/galanti25a.html.

[20] Matteo Gamba, Adrian Chmielewski-Anders, Josephine Sullivan, Hossein Azizpour, and Marten Bjorkman. Are all linear regions created equal? In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6573–6590, 2022.

[21] Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995. URL https://www.cs.ox.ac.uk/people/paul.goldberg/papers/goldbergjerrum.pdf.

[22] Moritz Grillo and Guido Montúfar. Most ReLU networks admit identifiable parameters. *arXiv preprint arXiv:2605.03601*, 2026. URL https://arxiv.org/abs/2605.03601.

[23] Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

[24] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[25] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[26] Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2596–2604, 2019.

[27] Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. *arXiv preprint arXiv:2602.16849*, 2026. URL https://arxiv.org/abs/2602.16849.

[28] Andreas Heinecke and Wen-Liang Hwang. Deep representation with ReLU neural networks. *arXiv preprint arXiv:1903.12384*, 2019. URL https://arxiv.org/abs/1903.12384.

[29] Christoph Hertrich, Amitabh Basu, Marco Di Summa, and Martin Skutella. Towards lower bounds on the depth of ReLU neural networks. *SIAM Journal on Discrete Mathematics*, 37(2): 997–1029, 2023. doi: 10.1137/22M1489332.

[30] Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 5–13, 1993.

[31] Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26673–26699. PMLR, 2025. URL https://proceedings.mlr.press/v267/jacobs25a.html.

[32] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005. doi: 10.1214/009117905000000044.

[33] Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25333–25369. PMLR, 2024. URL https://proceedings.mlr.press/v235/kosson24a.html.

[34] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, pp. 950–957, 1991. URL https://proceedings.neurips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html.

[35] Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. In *Advances in Neural Information Processing Systems*, 2025.

[36] Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34651–34663, 2022.

[37] Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023.

[38] Ziming Liu, Ziqian Zhong, and Max Tegmark. Grokking as compression: A nonlinear complexity perspective. *arXiv preprint arXiv:2310.05918*, 2023.

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

[40] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. Pac-bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31459–31473, 2022.

[41] Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818. PMLR, 2024. URL https://proceedings.mlr.press/v235/lotfi24a.html.

[42] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

[43] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XsHqr9dEGH.

[44] Neil Rohit Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: Grokking modular arithmetic via average gradient outer product. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42834–42856. PMLR, 2025. URL https://proceedings.mlr.press/v267/mallinar25a.html.

[45] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019.

[46] David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

[47] Gavin McCracken, Arthur Ayestas Hilgert, Sihui Wei, Gabriela Moisescu-Pareja, Zhaoyue Wang, and Jonathan Love. Interpreting deep neural networks trained on elementary p groups reveals algorithmic structure. In *Proceedings of the 1st Conference on Topology, Algebra, and Geometry in Data Science*, volume 321 of *Proceedings of Machine Learning Research*, pp. 389–402. PMLR, 2026. URL https://proceedings.mlr.press/v321/mccracken26a.html.

[48] Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? a theoretical analysis on grokking modular addition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35934–35967. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohamadi24a.html.

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

[50] Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M. Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=i9wDX850jR.

[51] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[52] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

[53] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[54] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[55] Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

[56] Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705, 2020.

[57] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[58] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[59] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 2667–2690, 2019.

[60] Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4558–4566, 2018.

[61] Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer, 1977.

[62] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[63] Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46441–46467. PMLR, 2024. URL https://proceedings.mlr.press/v235/stander24a.html.

[64] Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pp. 1517–1539, 2016.

[65] Yuandong Tian. Composing global solutions to reasoning tasks via algebraic objects in neural nets. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/666c3cae2f4d0cd796d2113fb797f38d-Abstract-Conference.html.

[66] Yuandong Tian. Li_2: A framework on dynamics of feature emergence and delayed generalization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7ce5da35e01cfa8d303c2dc71e61a470-Paper-Conference.pdf.

[67] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

[68] Yijun Wan, Melih Barsbey, Abdellatif Zaidi, and Umut Simsekli. Implicit compressibility of overparametrized neural networks trained with heavy-tailed SGD. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49845–49866. PMLR, 2024. URL https://proceedings.mlr.press/v235/wan24a.html.

[69] Renhao Wang, Marjan Albooyeh, and Siamak Ravanbakhsh. Equivariant networks for hierarchical structures. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13806–13817, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9efb1a59d7b58e69996cf0e32cb71098-Paper.pdf.

[70] Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign overfitting and grokking in ReLU networks for XOR cluster data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BxHgpC6FNv.

[71] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

[72] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[73] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5824–5832, 2018.

[74] Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond ReLU to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024. URL `https://jmlr.org/papers/v25/23-0912.html`.

[75] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Extended Literature Review

Role of the references. The counting proofs use the polynomial sign bounds of Goldberg & Jerrum [21, Theorem 2.1] and Bartlett et al. [6, Lemma 17]; the product-group lower bound uses the progression-free-set estimate of Ellenberg & Gijswijt [17]. Spectral-margin bounds provide a norm-based perspective [5]. The other works discussed below provide related results or background.

A.1 Compression and generalization

MDL, Occam, PAC–Bayes, and other compression analyses relate compact descriptions to generalization [58, 9, 46, 15, 2, 75, 40, 41, 68, 7, 30, 38]. Post-training pruning and quantization provide a complementary notion of compression [16]. Sample compression retains selected observations and auxiliary information from which a predictor is reconstructed; its size counts retained data and side information. EGC instead counts jointly realizable structure while preserving every label on the finite input domain, including errors. A guarantee for approximate scores or retained training accuracy alone does not establish this requirement on the complete prediction map.

Interpolation alone does not determine generalization [72, 8, 51]. EGC describes exact complete predictions in the stated model; our contribution combines a task-specific compression window with an explicit regularizer that enforces a sufficient budget under the stated conditions.

A.2 Representation complexity and ReLU geometry

Minimal exact representations [14] and variation-norm or convex formulations [3, 59, 55, 56, 18] measure different aspects of representation cost.

Related geometric and identifiability analyses study piecewise-affine networks [73, 10, 22].

Region, expressivity, and capacity bounds describe network classes [49, 60, 26, 6, 1, 64, 71, 29, 28, 20, 74]. EGC minimizes jointly attainable structural counts for one complete prediction map in its original architecture and activation, allowing logits to change while charging affine contributions. The shallow growth bound accounts for each factor's $O(Q)$ coefficients. Exact function representation fixes real-valued outputs, whereas label equivalence allows scores and margins to change as long as the prescribed class decisions are preserved. Conversely, such freedom does not allow a change of architecture or activation. Counting all regions attainable by a network class therefore answers a different question from minimizing the structure required by one prediction map.

A.3 Regularization, implicit bias, and norm bounds

Implicit-bias analyses connect optimization to margins and norms [53, 54, 62, 42, 24, 25, 11]; weight-decay studies examine decoupling, rotation, low-rank bias, and mirror flow [39, 13, 33, 19, 31]. Low rank of a trained weight matrix is a property of that parameter realization; EGC belongs to the complete prediction map. Connecting the two requires an architecture-preserving construction and a count for the resulting class of prediction maps. Our specified quadratic objective controls exact EGC of the complete prediction map, to which the independent generalization theorem applies, throughout the stated objective-gap region. This separates the statistical consequence of reaching the region from the dynamics needed to reach it.

A.4 Grokking and algebraic-task theorems

Grokking studies examine delayed generalization and hidden progress [57, 36, 23, 52, 37, 4, 70]; algebraic-task studies analyze cosets, superposition, and feature dynamics [63, 35, 44, 66, 27, 47]. Lyu et al. [43] analyze the transition from kernel predictors to KKT solutions of margin or norm problems under large initialization and small WD, with provable generalization examples in diagonal linear networks and matrix completion. Tian [66, Theorem 4] analyze feature emergence through an energy function for independent feature learning and control the persistence of representation-related local maxima under partial sampling for quadratic group networks. Three close theorem-level comparisons clarify our contribution.

Morwani et al. [50, Theorem 5] characterize every maximum $L_{2,3}$-normalized-margin solution on the full prime-modular-addition table: one-hidden-layer quadratic networks, no biases, and width $m \geq 4(p-1)$ have specified neuronal Fourier support. Their general-group result imposes further representation and width conditions. Observing the full table removes the problem of extending predictions from sampled pairs. The finite-regularization question studied here instead concerns all near-optimal completions of a partial table, rather than only maximizers of a scale-normalized margin.

Mohamadi et al. [48, Theorem 4.3] treat randomly observed modular-addition pairs and one-hidden-layer quadratic networks with $h \leq Mp$ and a constant parameter ℓ_∞ bound. For all such predictors, their guarantee is empirical margin loss at margin p plus $\widetilde{O}(\sqrt{p^{5/3}/n})$; a construction realizes the required margins with bounded norm. Proposition 4.2 relates vanishing regularization to maximum normalized margin. This already gives nontrivial partial-data generalization. Its norm and margin requirements specify a family to which the statistical bound applies. Our objective-gap assumption has a different role: it must itself imply sufficient training fit and a small exact prediction-map budget before the independent counting result can be used.

Tian [65, Theorems 1–3] express the full-table projected squared loss of one-hidden-layer quadratic networks on abelian groups through Fourier polynomials, using character orthogonality. Semiring operations construct solutions across widths; Theorem 5 connects global solutions related by semiring multiplication and compares their regularization costs; Theorem 6 studies an infinite-width initialization limit for a simplified loss. Character orthogonality applies to the full-table average; the sampled-table objective requires a separate argument. Our theorem fixes a ReLU architecture at every fixed depth from one hidden layer, uses a random partial table and cross-entropy with finite quadratic WD, and controls every parameter in a quantitative objective-gap region. Under the hidden-width-ratio condition, it derives sufficient training fit and exact compression of each complete prediction map, then applies independent EGC generalization. This gives an objective-level prediction-map guarantee without prescribing Fourier support or a particular optimization trajectory.

A.5 Group structure and equivariance

Equivariant architectures incorporate group actions [12, 45]. Finite-group representation theory [61] provides mathematical background for our group constructions. The target is implemented by an ordinary one-hot-input network below the statistical exclusion scale. Equivariance is not imposed on all candidate predictors: the model can also represent completions that violate the target group law. The statistical separation is therefore substantive. Group structure supplies a compact realization of the target, while counting excludes poorly generalizing fits at that complexity scale; the regularized-objective theorem addresses how such a scale is enforced.

B Structural Definitions and Examples

This section records the explicit shallow model classes, the affine–switching tradeoff, and the scalar compression-window thresholds used in the proofs. The underlying model and equality of complete prediction maps are those of Section 2.

B.1 Properties of the group task

Group multiplication records how reversible operations compose. Rotations and permutations are familiar examples; associativity makes the same rule reusable across successive compositions. Learning the table therefore tests whether a shared combination rule extends to unobserved pairs.

The task also isolates the role of the two operands together. Let U, V be independent uniform elements of G and let $Y = UV$. For all $a, b, y \in G$,

$$\mathbb{P}(Y = y \mid U = a) = \mathbb{P}(V = a^{-1}y) = Q^{-1}, \qquad \mathbb{P}(Y = y \mid V = b) = \mathbb{P}(U = yb^{-1}) = Q^{-1}.$$

Thus either operand alone gives only chance-level population accuracy $1/Q$, whereas the pair determines the label. These are properties of the uniform full table; a finite training subset can have unequal row and column frequencies. Complete evaluation on unseen pairs then directly measures whether prediction follows the shared rule. Candidate predictions range over all complete tables

realizable by the model below, so the group identities belong to the target, rather than being imposed on the predictions. This makes the task a test of whether compression favors reuse of the target rule over fitting entries individually.

B.2 Explicit width classes

For each integer $m \geq 0$, define

$$\mathcal{S}_m := \bigcup_{k=0}^{m} \Bigg\{ (a,b) \longmapsto \sum_{j=1}^{k} c_j \, \sigma(u_j(a) + v_j(b) + \beta_j) : \\ (u_j, v_j, \beta_j, c_j) \in (\mathbb{R}^G)^2 \times \mathbb{R} \times \mathbb{R}^Q, \quad 1 \leq j \leq k \Bigg\}. \tag{B.1}$$

Here $\mathbb{R}^G$ is the set of real-valued functions on G, and the $k = 0$ sum is the zero logit map. Hence $\mathcal{S}_m$ is exactly the class of logits representable by at most m nonzero effective gates, and $\mathcal{S} = \bigcup_{m \geq 0} \mathcal{S}_m$.

The separate-operand representation is what makes the count informative. One affine score has the form

$$\ell(a,b) = u(a) + v(b) + \beta. \tag{B.2}$$

It contains $O(Q)$ operand-specific coefficients reused over Q^2 pairs. With a single pair token $e_{(a,b)}$, one scalar preactivation can instead assign an independent value to every pair. Number the classes by $c \in \{0, \ldots, Q-1\}$ and fix any desired label table $y : X \to \{0, \ldots, Q-1\}$. Two gates then give

$$\begin{aligned} F_c(a,b) &= 2c\,\sigma(y(a,b) + 1) - (c^2 + 2c)\,\sigma(1) \\ &= y(a,b)^2 - (c - y(a,b))^2. \end{aligned}$$

The unique maximizing class is $y(a,b)$. Both gates are everywhere active, and the construction has no output bias or direct input-to-output term. Hence the pair-token class contains all Q^{Q^2} prediction maps at a budget of two gates.

Switching and everywhere-active gates. Each gate in (2.1) sends one scalar affine score in one output direction. If its scalar preactivation is nonnegative everywhere on X, then

$$c_j \sigma(\ell_j(a,b)) = c_j \ell_j(a,b),$$

and we call it an *everywhere-active gate*. Its contribution is globally affine, and its coefficient matrix is the outer product of the output vector c_j and the coefficient vector of ℓ_j, and hence has rank at most one. If ℓ_j takes both positive and negative values, the same rank-one affine contribution is applied only where $\ell_j > 0$; we call it a *switching gate*. An effective gate is therefore either everywhere-active or switching. These are two operating regimes of the same standard ReLU gate, not separately postulated resources.

Counting only switching gates would leave an arbitrary amount of affine information outside the budget. Counting every gate makes the competition visible. For a realization F, let $n_{\mathrm{sw}}(F)$ and $n_{\mathrm{aff}}(F)$ denote its numbers of switching and everywhere-active gates. For each integer $s \geq 0$, define

$$A_s(B) := \min\big\{ n_{\mathrm{aff}}(F) : B_F = B,\ n_{\mathrm{sw}}(F) \leq s \big\}, \tag{B.3}$$

with $A_s(B) = +\infty$ if no such realization exists. The sequence $A_s(B)$ is nonincreasing, and the definition gives

$$\kappa(B) = \min_{s \geq 0} \big\{ s + A_s(B) \big\}. \tag{B.4}$$

Adding switching gates lowers the total budget precisely when the number of everywhere-active gates saved exceeds the number of switches added. Appendix C.1 proves the identity and the affine obstruction $A_0(B_G) = \infty$ for every nontrivial group. Appendix C.2 gives the equivalent description in terms of rank-one affine coefficient matrices.

B.3 Affine obstruction and product examples

Proposition B.1 (Affine obstruction and nonlinear compression)**.** *Let G be a nontrivial finite group of order Q, using the separate categorical operand representation of Section 2.1. No collection of globally affine logits can realize the multiplication map $B_G(a,b) = ab$. Nevertheless, for the universal constant $C > 0$ of Theorem 3.3,*

$$\kappa(B_G) \le C\sqrt{Q}\log(2Q). \tag{B.5}$$

For cyclic addition, at most two switching gates and two everywhere-active gates suffice, so

$$\kappa(B_{\mathbb{Z}_Q}) \le 4. \tag{B.6}$$

For affine logits, summing correct-versus-permuted-class comparisons cancels every row and column contribution, forcing all classes to tie. Nonlinear gates escape this obstruction. The proof and constructions appear in Appendices C and F.

The distinction remains quantitative as the task grows. For d independent binary additions, $G = \mathbb{Z}_2^d$, one has

$$A_s(B_G) = \infty \quad (2s < d), \qquad A_d(B_G) \le 2. \tag{B.7}$$

Thus no finite affine budget replaces the necessary switches, while d switches permit a shared readout with only two everywhere-active gates. This illustrates why EGC charges both contributions: the resulting total is at most $d + 2$, with no free affine storage. Theorem C.3 in Appendix C.3 proves the bounds and extends them to other product groups.

B.4 Scalar compression-window thresholds

Definition B.2 (Fitting and generalization thresholds)**.** For $\varepsilon_{\mathrm{in}} \in [0,1]$ and $\varepsilon_{\mathrm{out}} \in (0,1)$, define

$$K_{\mathrm{low}}(S, \varepsilon_{\mathrm{in}}; G) = \min\{\kappa(B) : \mathrm{err}_{\mathrm{in}}(B) \le \varepsilon_{\mathrm{in}}\}, \tag{B.8}$$

$$K_{\mathrm{high}}(S, \varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}}; G) = \min\{\kappa(B) : \mathrm{err}_{\mathrm{in}}(B) \le \varepsilon_{\mathrm{in}},\ \mathrm{err}_{\mathrm{out}}(B) \ge \varepsilon_{\mathrm{out}}\}. \tag{B.9}$$

Both minima range over complete prediction maps $B : X \to G$; a minimum over the empty set is $+\infty$. An integer $m \ge 0$ lies in the *compression window* at tolerances $(\varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}})$ when

$$K_{\mathrm{low}}(S, \varepsilon_{\mathrm{in}}; G) \le m < K_{\mathrm{high}}(S, \varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}}; G). \tag{B.10}$$

Inside the window, some prediction map with EGC at most m fits the sample within $\varepsilon_{\mathrm{in}}$, and every such fit has held-out error below $\varepsilon_{\mathrm{out}}$. The case $\varepsilon_{\mathrm{in}} = 0$ is the *interpolation compression window*. Only the training requirement is relaxed: EGC still concerns the exact complete prediction map. Theorem B.3 proves a nonempty window in $\mathcal{S}$ when $\varepsilon_{\mathrm{in}}$ is sufficiently small relative to $\varepsilon_{\mathrm{out}}$.

B.5 Uniform Selection Inside the Window

The following finite condition makes the target–bad-completion separation explicit.

Theorem B.3 (Finite compression window and minimum-EGC selection)**.** *Fix a finite group G of order $Q \ge 2$, a density $\alpha \in (0,1)$ such that $1 \le n = \lfloor \alpha Q^2 \rfloor < Q^2$, and $\delta \in (0,1)$. Let $m_G = \lceil C\sqrt{Q}\log(2Q) \rceil$, where C is the universal constant in* (3.4)*; let $C_\alpha > 0$ be the constant from Theorem 3.2, depending only on α. For a uniform sample $S \subseteq X$ of size n, with probability at least $1 - \delta$, simultaneously for every $\varepsilon_{\mathrm{in}} \in [0,1]$,*

$$K_{\mathrm{low}}(S, \varepsilon_{\mathrm{in}}; G) \le \kappa(B_G) \le m_G. \tag{B.11}$$

Moreover, for $\varepsilon_{\mathrm{in}} \in [0,1]$, let

$$\mathfrak{M}_Q(S, \varepsilon_{\mathrm{in}}) = \operatorname*{argmin}_{B : \mathrm{err}_{\mathrm{in}}(B) \le \varepsilon_{\mathrm{in}}} \kappa(B). \tag{B.12}$$

On the same event, uniformly in $\varepsilon_{\mathrm{in}}$,

$$\sup_{B \in \mathfrak{M}_Q(S, \varepsilon_{\mathrm{in}})} \mathrm{err}_{\mathrm{out}}(B) \le C_\alpha \left[\varepsilon_{\mathrm{in}} + \frac{(m_G + 1)\log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2} \right]. \tag{B.13}$$

Moreover, whenever

$$C_\alpha \left[\varepsilon_{\text{in}} + \frac{(m_G + 1)\log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2} \right] < \varepsilon_{\text{out}},$$

one has

$$K_{\text{low}}(S, \varepsilon_{\text{in}}; G) \leq m_G < K_{\text{high}}(S, \varepsilon_{\text{in}}, \varepsilon_{\text{out}}; G). \tag{B.14}$$

Since $m_G = O(\sqrt{Q}\log Q)$, the familiar $O(\sqrt{Q}\log Q)$ versus $\Omega(Q/\log Q)$ separation is the corresponding constant-density scaling consequence. If $\varepsilon_{\text{in}} \to 0$, $\delta_Q \to 0$, and $\log(1/\delta_Q) = o(Q^2)$, the finite bound gives uniform consistency over the minimizing set. Appendix F gives the proof.

C GROUP LAWS: CONSTRUCTIONS AND SWITCHING REQUIREMENTS

Proposition C.1 (Constant-count realization of modular addition)**.** *Let $p \in \mathbb{N}^*$ with $p \geq 2$. Consider modular addition over $\mathbb{Z}_p$, with inputs $x_1, x_2 \in \{0, \ldots, p-1\}$ and label $y = x_1 + x_2 \pmod p$. There is a one-hidden-layer ReLU logit model that classifies the full table correctly with a constant factor count, independent of p. In the architecture without an explicit affine output term, four factors suffice. The modular fold uses at most two switching factors.*

Proof. First form the integer sum $t = x_1 + x_2$, where $t \in \{0, \ldots, 2p-2\}$. This is a linear function of the two one-hot input vectors:

$$t = \sum_{a=0}^{p-1} a\, \mathbf{1}\{x_1 = a\} + \sum_{b=0}^{p-1} b\, \mathbf{1}\{x_2 = b\}.$$

Define the fold

$$q(t) = t - p(t - (p-1))_+ + p(t-p)_+. \tag{C.1}$$

We call q the modular fold. We claim that $q(t) = t \pmod p$ on $\{0, \ldots, 2p-2\}$.

Indeed, if $0 \leq t \leq p-1$, then $(t-(p-1))_+ = 0$ and $(t-p)_+ = 0$, so $q(t) = t$. If $p \leq t \leq 2p-2$, then $(t-(p-1))_+ = t-p+1$ and $(t-p)_+ = t-p$, hence

$$q(t) = t - p(t-p+1) + p(t-p) = t - p.$$

Therefore $q(t) = t \pmod p$ for every valid integer sum t.

For each class $c \in \{0, \ldots, p-1\}$, define the logit

$$F_c(t) = 2c\, q(t) - c^2. \tag{C.2}$$

Let $y = q(t)$. Then for any $c \neq y$,

$$F_y(t) - F_c(t) = 2yq(t) - y^2 - \big(2cq(t) - c^2\big) = (y-c)^2 > 0.$$

Thus the argmax class is exactly $y = q(t) = x_1 + x_2 \pmod p$ on the full table.

It remains to count factors. Expanding the logit expression using the modular fold gives

$$F_c(t) = -c^2 + 2ct - 2pc(t-(p-1))_+ + 2pc(t-p)_+.$$

Since $t = \sigma(t+1) - \sigma(1)$ on the valid domain $0 \leq t \leq 2p-2$, the constant and linear scaffold can be implemented by the two globally active factors $\sigma(1)$ and $\sigma(t+1)$. They are charged to the same budget as the switching factors. Thus four factors suffice:

$$\sigma(1), \qquad \sigma(t+1), \qquad \sigma(t-(p-1)), \qquad \sigma(t-p).$$

The last two are switching factors when $p \geq 3$; at $p = 2$ the last is zero and is omitted. The first two implement the scaffold and are globally affine factors. Hence the complete modular-addition rule has four charged factors, independent of p. □

Proposition C.2 (Globally affine logits cannot represent a group law)**.** *Let G be a nontrivial finite group. Logits of the form*

$$F_c(a,b) = r_c(a) + s_c(b) + \gamma_c, \qquad a, b, c \in G,$$

cannot have ab as their winning class at every input under the fixed class-priority rule.

Proof. Suppose the logits realize the group law. Fix $g \in G$. Correctness gives

$$F_{ab}(a,b) \geq F_{agb}(a,b) \qquad (a,b \in G). \tag{C.3}$$

Sum both sides over a, b. For fixed a, both $b \mapsto ab$ and $b \mapsto agb$ are bijections of G, so the sums of the r terms agree. For fixed b, both $a \mapsto ab$ and $a \mapsto agb$ are bijections, so the sums of the s terms agree. The sums of the class biases agree as well. Hence the two sides of the summed inequality are equal. Since every summand in (C.3) is nonnegative, equality holds at every input. For fixed (a, b), the map $g \mapsto agb$ is a bijection, so all class logits tie. The fixed tie-breaking rule would then return one constant class, contradicting the nontrivial group law. □

Together with the four-factor modular construction above and the universal $O(\sqrt{Q}\log Q)$ construction in Appendix F, this proves Proposition B.1: switching factors reduce the affine-only requirement from impossibility to a finite, and in general sublinear, number of charged rank-one actions.

C.1 The switching–affine tradeoff identity

We derive (B.4) and the accompanying affine obstruction. On the fixed finite domain, every effective factor belongs to exactly one of two classes. If its affine preactivation is nonnegative everywhere, its ReLU output equals that affine preactivation everywhere and the factor is globally affine. If the preactivation takes both positive and negative values, the factor is switching. A preactivation that is never positive gives an ineffective factor and is absent from a minimal realization.

The feasible realizations in the definition of $A_s(B)$ are contained in those for $A_{s+1}(B)$, so $A_{s+1}(B) \leq A_s(B)$. For any s with $A_s(B) < \infty$, choose a realization attaining $A_s(B)$ and let $t \leq s$ be its actual number of switching factors. Its total number of effective factors is $t + A_s(B) \leq s + A_s(B)$, and therefore

$$\kappa(B) \leq \min_{s \geq 0}\{s + A_s(B)\}. \tag{C.4}$$

Conversely, take a realization with exactly $\kappa(B)$ effective factors, and let t and r be its numbers of switching and globally affine factors. Then $t + r = \kappa(B)$ and $A_t(B) \leq r$. Consequently,

$$\min_{s \geq 0}\{s + A_s(B)\} \leq t + A_t(B) \leq t + r = \kappa(B). \tag{C.5}$$

Equations (C.4) and (C.5) prove the identity. For finite values at $s_1 < s_2$, subtracting the candidate totals gives

$$s_2 + A_{s_2}(B) < s_1 + A_{s_1}(B) \quad \Longleftrightarrow \quad A_{s_1}(B) - A_{s_2}(B) > s_2 - s_1. \tag{C.6}$$

If $s = 0$, every effective factor is globally affine, so their sum has logits of the form

$$F_c(a,b) = r_c(a) + s_c(b) + \gamma_c.$$

Proposition C.2 shows that such logits cannot realize a nontrivial group law. Thus $A_0(B_G) = +\infty$. For modular addition, the realization in Proposition C.1 consists of the two switching folds $\sigma(t-(p-1))$ and $\sigma(t-p)$ and the two globally affine scaffold factors $\sigma(1)$ and $\sigma(t+1)$. Hence $A_2(B_{\mathbb{Z}_p}) \leq 2$.

C.2 Rank-one characterization

Write $\widetilde{x} = (e_a, e_b, 1)$ and $\ell_j(a,b) = r_j^\top \widetilde{x}$. On $\{\ell_j > 0\}$ the jth effective gate equals $(c_j r_j^\top)\widetilde{x}$, and outside that set it is zero. Its coefficient matrix has rank one. Conversely, a rank-one matrix $cr^\top$ applied on the positive side of $r^\top \widetilde{x}$ is exactly $c\sigma(r^\top \widetilde{x})$. Thus the shortest such decomposition is the shortest gate list in Definition 2.1. Whether its affine form changes sign on the finite domain distinguishes switching from everywhere-active gates.

C.3 Switching requirements on product groups

The following theorem proves (B.7) and records its product-group extensions in the model of (2.1). The lower bounds do not restrict the operand profiles to individual coordinates. We identify $\mathbb{Z}_p^d$ with the additive group of the vector space $\mathbb{F}_p^d$. Write $[t]_+ = \max\{t, 0\}$.

Theorem C.3 (Switching requirements and affine compression)**.** *For $G = \mathbb{Z}_2^d$, every $d \geq 1$ satisfies*

$$A_s(B_G) = \infty \quad (2s < d), \qquad A_d(B_G) \leq 2. \tag{C.7}$$

For each $p \in \{3, 5\}$ there are constants $c_p > 0$ and $C_p \geq 0$, independent of d, such that for $G = \mathbb{Z}_p^d$,

$$A_s(B_G) = \infty \quad (s < c_p d - C_p), \qquad A_{2d}(B_G) \leq 2. \tag{C.8}$$

Consequently, for each $p \in \{2, 3, 5\}$, both the minimum feasible switching budget and $\kappa(B_{\mathbb{Z}_p^d})$ are $\Theta(d)$. The lower bounds allow arbitrary whole-operand profiles and any finite number of everywhere-active gates.

Lemma C.4 (A product set with fixed activation decisions)**.** *For arbitrary finite nonempty operand sets U, V and r forms $\ell_j(a, b) = u_j(a) + v_j(b) + \beta_j$, there are $E \subseteq U$ and $H \subseteq V$ such that*

$$|E| \geq |U|2^{-r}, \qquad |H| \geq |V|2^{-r}, \tag{C.9}$$

and every $[\ell_j]_+$ is affine or identically zero on $E \times H$.

Proof. Start with $E = U, H = V$ and process the forms successively. Choose medians m_u, m_v of the current row and column values. At least half of each set lies on either weak side of its median. If $m_u + m_v + \beta_j \geq 0$, retain the upper halves in both operands; otherwise retain both lower halves. The current form is respectively nonnegative or negative throughout the retained product. Later restrictions preserve all preceding sign decisions. Each step loses at most half of each operand set, proving (C.9). Zero preactivations cause no exception, since ReLU agrees with its affine argument on a nonnegative set and vanishes on a nonpositive set. □

On the product supplied by this lemma, a realization with r switches and any finite affine budget has logits $L_c(a, b) = R_c(a) + S_c(b) + \gamma_c$. We may allow arbitrary such logits when proving impossibility: this enlarges the class being excluded, and does not give an uncharged term to any upper-bound construction.

Lemma C.5 (Affine obstruction on parallel cosets)**.** *Let K be a nontrivial subgroup of a finite abelian group. Separated affine logits cannot classify addition correctly on $(a_0 + K) \times (b_0 + K)$ under the fixed class-priority rule.*

Proof. Translate the input and output labels to write the target as $x + y$, $x, y \in K$. Fix a nonzero $h \in K$ and sum the comparisons of $x + y$ against $x + y + h$ over $K \times K$. Every row, column, and bias contribution cancels because the two label lists in each row and column are permutations of the same output coset. Each comparison is nonnegative. At least one is strict: on every cycle of the permutation $c \mapsto c + h$, some comparison has a lower-priority true label and a higher-priority competing label. Correct prediction requires a strict advantage there. The sum cannot therefore equal zero, a contradiction. □

For the binary lower bound, put $G = \mathbb{F}_2^d$ and $Q = 2^d$. If $|E||H| > Q$, two distinct pairs $(a, b), (a', b') \in E \times H$ have the same sum. Both operands differ, and $h = a' - a = b' - b \neq 0$. The rows $a + \{0, h\}$ and columns $b + \{0, h\}$ form parallel cosets of an order-two subgroup. The preceding lemma forbids affine classification on this subtable. Consequently any realization with r switches satisfies $Q^2 4^{-r} \leq |E||H| \leq Q$, whence $2r \geq d$. This proves the first assertion of (C.7).

For $p = 3, 5$, let $\mu_p(d)$ be the maximum relative size of a subset of $\mathbb{F}_p^d$ with no three distinct points $x, x + h, x + 2h$, $h \neq 0$. Such three points are a nontrivial three-term arithmetic progression. We show that every realization with r switches satisfies

$$4^{-r} \leq \mu_p(d), \qquad p \in \{3, 5\}. \tag{C.10}$$

For $p = 3$, partition $E \times H$ by the sum $c = a + b$, and set $C_c = \{a \in E : c - a \in H\}$. Then $\sum_c |C_c| = |E||H|$. If $|E||H| > \mu_3(d)Q^2$, some C_c contains a nontrivial progression. In characteristic three it is an entire coset of an order-three subgroup; its corresponding columns form a parallel coset. Lemma C.5 gives a contradiction. Together with (C.9), this proves (C.10) for $p = 3$.

For $p = 5$, a three-point progression is not a full subgroup coset. Instead consider the following two label tables:

$$T = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 2 & 4 \\ 2 & 3 & 0 \end{pmatrix}, \qquad T' = \begin{pmatrix} 1 & 3 & 0 \\ 2 & 1 & 4 \\ 0 & 2 & 3 \end{pmatrix}. \tag{C.11}$$

The first is addition on rows $0, 1, 2$ and columns $0, 1, 3$ in $\mathbb{F}_5$. Each row and each column has the same label multiplicities in T and T'. Thus, for separated affine logits,

$$\sum_{i,j} \big(L_{T_{ij}}(i,j) - L_{T'_{ij}}(i,j)\big) = 0. \tag{C.12}$$

Correct classification of T makes all summands nonnegative. Its first row compares 0 to 1, 1 to 3, and 3 to 0, so at least one is strict under every priority order. This contradicts (C.12). Translating the operands and replacing each index k by kh for any nonzero $h \in \mathbb{F}_5^d$ preserves the contradiction, since the five output labels stay distinct.

Now use the weighted sums $c = 2a + b$ and define $C_c = \{a \in E : c - 2a \in H\}$. Again $\sum_c |C_c| = |E||H|$. If this sum exceeds $\mu_5(d)Q^2$, one C_c contains $x, x+h, x+2h$. The corresponding columns are $c - 2x, c - 2x - 2h, c - 2x - 4h$, namely a translate of $\{0, 3h, h\}$. Reordering the columns gives exactly the impossible table (C.11). This proves (C.10) for $p = 5$ as well.

To evaluate the bound, Theorem 4 of Ellenberg & Gijswijt [17] implies

$$\mu_p(d) \le \frac{3M_{p,d}}{p^d}, \qquad M_{p,d} = \#\left\{\nu \in \{0, \dots, p-1\}^d : \sum_i \nu_i \le \frac{(p-1)d}{3}\right\}. \tag{C.13}$$

Its coefficients are $1, -2, 1$, all nonzero for $p = 3, 5$; the excluded solutions are precisely the nonconstant progressions just defined. For $0 < t < 1$, every exponent counted in $M_{p,d}$ satisfies $t^{\sum_i \nu_i} \ge t^{(p-1)d/3}$, so

$$M_{p,d} \le \left[t^{-(p-1)/3}(1 + t + \cdots + t^{p-1})\right]^d.$$

Taking $t = 1/2$ for $p = 3$ and $t = 2/3$ for $p = 5$ gives

$$\begin{aligned} \mu_p(d) &\le 3\, 2^{-\tau_p d}, \\ \tau_3 = \tfrac{1}{3}\log_2(432/343) > 0, &\qquad \tau_5 = \tfrac{1}{3}\log_2(13122000/9393931) > 0. \end{aligned} \tag{C.14}$$

Indeed the cubes of the two normalized brackets are respectively $343/432$ and $9393931/13122000$. It follows that $2r \ge \tau_p d - \log_2 3$. Thus one may take $c_p = \tau_p/2$ and $C_p = (\log_2 3)/2$ in (C.8). No affine budget enters these inequalities.

C.4 A SHARED AFFINE READOUT FOR FINITE ABELIAN GROUPS

Proposition C.6 (Two everywhere-active gates after coordinate folds)**.** *Let $G = \prod_{i=1}^r \mathbb{Z}_{n_i}$ with $n_i \ge 2$, and let b_2 be the number of factors with $n_i = 2$. In the unchanged model,*

$$A_{2r-b_2}(B_G) \le 2, \qquad \kappa(B_G) \le 2r - b_2 + 2. \tag{C.15}$$

Proof. Represent coordinate a_i by its integer in $\{0, \dots, n_i - 1\}$. The inputs remain the two whole-operand one-hot vectors; taking their coordinate values is a choice of permitted operand profiles. Put

$$M_1 = 1, \quad M_i = \prod_{j<i} n_j, \qquad C(a) = \sum_i M_i a_i, \qquad t_i = a_i + b_i.$$

The positional code C bijects G with $\{0, \dots, Q-1\}$: recover successive digits by taking a remainder modulo n_1, dividing the difference by n_1, and repeating. Its maximum is $\sum_i M_i(n_i - 1) = Q - 1$. Use the modular folds

$$h_{i,1} = [t_i - (n_i - 1)]_+, \qquad h_{i,2} = [t_i - n_i]_+, \qquad q_i = t_i - n_i h_{i,1} + n_i h_{i,2}.$$

For $t_i < n_i$, both hinges vanish. For integer $t_i \ge n_i$, their difference is one. Hence q_i is exactly the representative of $a_i + b_i \pmod{n_i}$. The first hinge switches for every n_i; the second switches when $n_i \ge 3$ and is zero when $n_i = 2$.

Set $\alpha = C(a) + C(b)$ and $q = \alpha - \sum_i M_i n_i h_{i,1} + \sum_i M_i n_i h_{i,2} = C(a+b)$. The two globally affine features are $g_1 = [\alpha + 1]_+$ and $g_0 = [1]_+$. For class c, define

$$F_c = 2C(c)g_1 - \big(C(c)^2 + 2C(c)\big)g_0 \\ - \sum_i 2C(c)M_i n_i h_{i,1} + \sum_i 2C(c)M_i n_i h_{i,2}. \tag{C.16}$$

Every summand is a counted gate. Expanding yields

$$F_c = 2C(c)q - C(c)^2 = q^2 - \big(C(c) - q\big)^2.$$

The correct class uniquely maximizes the score, with advantage at least one over every other class. There are $2r - b_2$ switching gates and two everywhere-active gates, proving (C.15). □

Taking every $n_i = 2$ or every $n_i = p \geq 3$ proves the upper bounds in Theorem C.3. Together with the lower bounds, this proves its $\Theta(d)$ conclusions. More explicitly, the minimum feasible switching budget lies between $\lceil d/2 \rceil$ and d for $p = 2$, and between $\max\{1, \lceil (\tau_p d - \log_2 3)/2 \rceil\}$ and $2d$ for $p = 3, 5$. The lower bound one follows from the affine obstruction.

All coefficients in (C.16) are integers of magnitude at most $2Q^2$, so each has $O(\log Q)$ bits. The construction uses length-Q operand and output profiles and establishes the stated gate-count bound, with increasing nonlinear necessity and a constant achievable affine budget.

D Sampling bounds with training-error tolerance

Lemma D.1 (Finite-class bound with training errors)**.** *Let X have size N, let S be a uniform n-element subset with $1 \leq n < N$, and fix a target map and a nonempty finite prediction class $\mathcal{H}$ independently of S. For $\delta \in (0, 1)$, with probability at least $1 - \delta$, every $B \in \mathcal{H}$ satisfies*

$$\mathrm{err}_{\mathrm{out}}(B) \leq \frac{2N - n}{N - n} \mathrm{err}_{\mathrm{in}}(B) + \frac{8N}{n(N - n)} \log \frac{|\mathcal{H}|}{\delta}. \tag{D.1}$$

Here errors are measured against the fixed target, on S and $X \setminus S$.

Proof. For a fixed B, write E for its error set, $e = |E|$, and $Z = |E \cap S|$. The hypergeometric variable Z has mean $\mu = ne/N$ and satisfies the standard lower-tail inequality

$$\Pr(Z \leq \mu/2) \leq e^{-\mu/8}. \tag{D.2}$$

For clarity, this bound follows from sampling without replacement: for $t > 0$, the mean product of n distinct entries of a nonnegative list is at most the nth power of its arithmetic mean. Indeed, at a fixed sum, averaging two entries cannot decrease the sum of all such products; its maximum is attained when all entries are equal. Applying this to the N numbers $e^{-t\mathbf{1}_E(x)}$ gives $\mathbb{E}e^{-tZ} \leq (1 - e/N + (e/N)e^{-t})^n \leq \exp\{\mu(e^{-t} - 1)\}$. Markov's inequality with $t = \log 2$ then gives $\Pr(Z \leq \mu/2) \leq \exp\{-\mu(1 - \log 2)/2\} \leq e^{-\mu/8}$.

Put $L = \log(|\mathcal{H}|/\delta)$. If $e > 2NZ/n + 8NL/n$, then $Z < \mu/2$ and $\mu > 8L$. Thus this event has probability at most e^{-L} for each fixed B. A union bound gives, simultaneously for all $B \in \mathcal{H}$,

$$e \leq \frac{2N}{n} Z + \frac{8N}{n} L.$$

Subtract Z, divide by $N - n$, and use $Z/n = \mathrm{err}_{\mathrm{in}}(B)$ to obtain (D.1). □

Proof of Theorem 3.2. Put $N = Q^2$ and $n = \lfloor \alpha Q^2 \rfloor$. For each $0 \leq m \leq Q$, apply Lemma D.1 to $\mathcal{B}_m$ with failure probability $\delta_m = \delta/[(m+1)(m+2)]$. These probabilities sum to at most δ. The growth theorem gives

$$\log |\mathcal{B}_m| \leq C(m+1)Q \log(2Q).$$

For $m = 0$ the zero logit table and fixed tie-breaking rule give one prediction map, so the same bound holds. Combining Lemma D.1 with this growth bound gives, on one event of probability at least $1 - \delta$, every B with $m = \kappa(B) \le Q$ the finite inequality

$$\mathrm{err}_{\mathrm{out}}(B) \le \frac{2N - n}{N - n}\,\mathrm{err}_{\mathrm{in}}(B) + \frac{8N}{n(N-n)}\left[C(m+1)Q\log(2Q) + \log\frac{(m+1)(m+2)}{\delta}\right]. \tag{D.3}$$

Since $n \ge 1$, $\lfloor \alpha Q^2 \rfloor \ge \alpha Q^2/2$, while $Q^2 - n \ge (1-\alpha)Q^2$. Therefore (D.3) implies that every prediction map with $\kappa(B) \le Q$ satisfies

$$\mathrm{err}_{\mathrm{out}}(B) \le C_\alpha\left[\mathrm{err}_{\mathrm{in}}(B) + \frac{(\kappa(B)+1)\log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2}\right]. \tag{D.4}$$

We used $m = \kappa(B)$ after forming the common event; no independence of the selected prediction map and the sample is required. The bound is trivial for $\kappa(B) > Q$ after increasing C_α. This is the finite bound in (3.2). The constant depends only on α, not on Q, the group law, or the selected prediction map.

□

At zero training error the exact avoidance calculation has a sharper constant. The interpolation results use this sharper constant.

Theorem D.2 (Finite-growth exclusion of high-error interpolants)**.** *Let X be a finite public domain of size N, and let $S \subseteq X$ be sampled uniformly without replacement, with $|S| = n$, where $1 \le n < N$. For any deterministic finite class of prediction maps $\mathcal{H}$, fixed target prediction map $B_\star$, and $\eta > 0$,*

$$\begin{aligned}&\Pr(\exists B \in \mathcal{H} : B|_S = B_\star|_S,\ |\{x : B(x) \ne B_\star(x)\}| \ge \eta(N - |S|)) \\ &\qquad \le |\mathcal{H}| \exp\left\{-\frac{n\eta(N-n)}{N}\right\}.\end{aligned} \tag{D.5}$$

In particular, for a $Q \times Q$ operation table the right-hand side applies with $\mathcal{H} = \mathcal{B}_m$ for a group G of order Q, and $\log|\mathcal{H}|$ bounded by (3.1)*.*

Proof. For a fixed map with e erroneous inputs, the probability that a uniform sample of size n avoids them is $\binom{N-e}{n}/\binom{N}{n} \le \exp(-ne/N)$ (and is zero if $N - e < n$). If $e \ge \eta(N-n)$, this is at most $\exp\{-\frac{n\eta(N-n)}{N}\}$. Summing over the deterministic class $\mathcal{H}$ proves the assertion. □

E GATE-COUNTING AND GROWTH-FUNCTION BOUND

This appendix proves the counting bound used in Section 3.1. We first write the proof with an arbitrary p-element set labelled as $\mathbb{Z}_p$; neither addition nor any other algebraic property is used. Relabelling p as Q therefore gives the statement for every Q-element operand and label set.

For the gate-counting bounds below, $p, m \in \mathbb{N}^*$ with $p \ge 2$. The zero-gate class contains only one complete prediction map.

Counting convention. For the counting theorem, we use the no-free-affine convention. We consider one-hidden-layer ReLU logits of the form

$$\begin{aligned}F_\theta(a,b) &= \sum_{j=1}^{m} z_j(a,b)\,g_j(a,b)\,w_j, \\ z_j(a,b) &= u_j(a) + v_j(b) + \beta_j, \\ g_j(a,b) &= \mathbf{1}\{z_j(a,b) > 0\}.\end{aligned}$$

Every constant or input-linear scaffold, including a class-dependent affine logit map on the full $2p$-dimensional one-hot input, is implemented by hidden ReLU factors and counted in m. In particular, the four-factor construction in Appendix C follows this convention. We therefore count any unrestricted class-dependent affine logit map on the full $2p$-dimensional one-hot input within the budget because it carries $O(p^2)$ degrees of freedom.

Lemma E.1 (Growth-function bound)**.** *Let $q, D, P \in \mathbb{N}^*$, and let q real polynomials of degree at most D be defined on $\mathbb{R}^P$. The number of sign patterns these polynomials induce is at most*

$$\left(\frac{CDq}{P}\right)^P$$

whenever $q \geq P$, for a universal constant $C > 0$. Consequently, its logarithm is at most $C'P\log(C'Dq)$ for another universal constant $C' > 0$. For $q < P$, the elementary bound 3^q, including zero signs, gives the same logarithmic estimate.

Proposition E.2 (Gate-counting bound for modular addition)**.** *Let $\mathcal{R}_{p,m}$ be the set of prediction maps over $\mathbb{Z}_p^2$ realizable by one-hidden-layer ReLU logits with gate count at most m, under the counting convention above. Then there is a universal constant $C > 0$ such that*

$$\log|\mathcal{R}_{p,m}| \leq Cmp\log(2mp).$$

Proof. Counting m potential gates only enlarges the class of predictors with gate count at most m. It is therefore enough to count prediction maps realized by logits of the form

$$\begin{aligned} F_\theta(a,b) &= \sum_{j=1}^{m} z_j(a,b)g_j(a,b)w_j, \\ z_j(a,b) &= u_j(a) + v_j(b) + \beta_j, \\ g_j(a,b) &= \mathbf{1}\{z_j(a,b) > 0\}. \end{aligned}$$

The parameter vector is

$$\theta = \left\{ \begin{array}{l} u_j(a), v_j(b), \beta_j, w_j(c) : \\ j = 1,\dots,m,\ a,b,c \in \mathbb{Z}_p \end{array} \right\}.$$

Let

$$P_m = m(2p+1) + mp.$$

We call P_m the number of real parameters. It satisfies $P_m \leq C_0 mp$ for a universal constant C_0.

We first count possible gate patterns. Define

$$\mathcal{Z} = \{z_j(a,b) : j = 1,\dots,m,\ (a,b) \in \mathbb{Z}_p^2\}.$$

We call $\mathcal{Z}$ the gate-preactivation family. It contains

$$q_Z = mp^2$$

linear polynomials in the parameters. By Lemma E.1, the number of sign patterns induced by $\mathcal{Z}$ is at most

$$\exp\{C_1 P_m \log(C_1 q_Z)\} \leq \exp\{C_2 mp\log(C_2 mp)\}.$$

Fix one attainable gate sign pattern. On the corresponding parameter cell, each gate value

$$\gamma_{j,a,b} = \mathbf{1}\{z_j(a,b) > 0\}$$

is constant. For this fixed pattern $\Gamma = (\gamma_{j,a,b})$, define

$$M^\Gamma_{a,b,c,d}(\theta) = \sum_{j=1}^{m} \gamma_{j,a,b}\big(u_j(a) + v_j(b) + \beta_j\big)\big(w_j(c) - w_j(d)\big), \qquad c \neq d.$$

We call $M^\Gamma_{a,b,c,d}$ the pairwise margin polynomial inside the gate cell. Since $\gamma_{j,a,b}$ is fixed on this cell, $M^\Gamma_{a,b,c,d}$ is a polynomial of degree at most 2 in θ.

For a fixed gate pattern Γ, the prediction map is determined by the signs of the margins

$$\mathcal{M}_\Gamma = \{M^\Gamma_{a,b,c,d} : (a,b) \in \mathbb{Z}_p^2,\ c,d \in \mathbb{Z}_p,\ c \neq d\},$$

together with the fixed deterministic tie-breaking rule. The family $\mathcal{M}_\Gamma$ contains at most

$$q_M \leq p^2 \cdot p(p-1) \leq p^4$$

polynomials of degree at most 2. Applying Lemma E.1 to this family gives at most

$$\exp\{C_3 P_m \log(C_3 p)\} \le \exp\{C_4 mp \log(C_4 p)\}$$

possible margin sign patterns inside the fixed gate cell. When $q_M < P_m$, the trivial bound 3^{q_M}, including zero signs, is smaller than $\exp\{CP_m\}$ after increasing the universal constant, so the same displayed upper bound still applies.

Combining the number of gate sign patterns with the number of margin sign patterns inside each gate cell gives

$$|\mathcal{R}_{p,m}| \le \exp\{C_2 mp \log(C_2 mp)\} \exp\{C_4 mp \log(C_4 p)\}.$$

Therefore, after increasing the universal constant,

$$\log |\mathcal{R}_{p,m}| \le Cmp \log(2mp).$$

This proves the stated explicit logarithmic-factor bound. □

Corollary E.3 (Cardinality-uniform EGC growth)**.** *For every set A of size $Q \ge 2$, the class $\mathcal{B}_m(A) = \{B : A^2 \to A : \kappa(B) \le m\}$ satisfies*

$$\log |\mathcal{B}_m(A)| \le CmQ \log(2mQ).$$

Proof. Label the operands and classes by $\mathbb{Z}_Q$ and apply Proposition E.2 with $p = Q$. Its proof uses only the numbers of row, column, and output coefficients and is invariant under this relabelling. □

Corollary E.4 (Zero-training-error exclusion from gate counting)**.** *Let $S \subseteq \mathbb{Z}_p^2$ be sampled uniformly without replacement with $|S| = n$, where $1 \le n < p^2$. For every $\eta > 0$,*

$$\Pr(K_{\text{high}}(S, 0, \eta; \mathbb{Z}_p) \le m) \le \exp\left\{ Cmp \log(2mp) - \frac{n\eta(p^2 - n)}{p^2} \right\}.$$

Proof. If $K_{\text{high}}(S, 0, \eta; \mathbb{Z}_p) \le m$, then there exists $B \in \mathcal{R}_{p,m}$ such that

$$E_B \cap S = \varnothing, \qquad |E_B| \ge \eta(p^2 - n).$$

For a fixed B with $|E_B| = s$, Appendix D gives

$$\Pr(E_B \cap S = \varnothing) \le \exp(-ns/p^2).$$

Thus every fixed B with $|E_B| \ge \eta(p^2 - n)$ avoids the training set with probability at most

$$\exp\{-\frac{n\eta(p^2 - n)}{p^2}\}.$$

Taking a union bound over all $B \in \mathcal{R}_{p,m}$ gives

$$\Pr(K_{\text{high}}(S, 0, \eta; \mathbb{Z}_p) \le m) \le |\mathcal{R}_{p,m}| \exp\{-\frac{n\eta(p^2 - n)}{p^2}\}.$$

Using Proposition E.2 yields

$$\Pr(K_{\text{high}}(S, 0, \eta; \mathbb{Z}_p) \le m) \le \exp\left\{ Cmp \log(2mp) - \frac{n\eta(p^2 - n)}{p^2} \right\}.$$

□

E.1 Sharpness of independent EGC generalization

Proposition E.5 (Sharpness of generalization from an EGC budget)**.** *Consider the cyclic-group task on $\mathbb{Z}_Q$ and fix an original ReLU architecture $\mathcal{A}$ with $L \ge 1$ hidden layers of widths $k_1, \dots, k_L \in \mathbb{N}^*$, hidden biases, and either original output-bias convention. Put $k_{\min} = \min_i k_i$ and $k_{\max} = \max_i k_i$. Fix a positive integer affine budget $\bar{a}$ and a nonlinear budget $\boldsymbol{\tau} = (\tau_1, \dots, \tau_L) \in \mathbb{N}_0^L$ independently of*

the training sample. Let S be a uniformly sampled n-element subset of $\mathbb{Z}_Q^2$. Replace τ_i by $\min\{\tau_i, k_i\}$, put $J = \sum_i \tau_i$, and choose i_* with $\tau_{i_*} = \max_i \tau_i$. For $1 \le n < Q^2$ and $0 < \delta < 1$, set

$$\alpha = \frac{n}{Q^2}, \qquad b_* = \min\{\tau_{i_*}, \min_i k_i - 1\}, \qquad t = \log\frac{2Q}{\delta}, \qquad s = \min\left\{Q, \left\lfloor \frac{b_*}{24\alpha} \right\rfloor\right\}.$$

If

$$b_* \ge 24t + 18, \qquad (1-\alpha)sQ \ge 8\log(2/\delta),$$

then, with probability at least $1-\delta$, there exists $B \in \mathcal{B}_{\mathcal{A}}(\bar{a}, \boldsymbol{\tau})$ such that

$$\mathrm{err}_{\mathrm{in}}(B) = 0, \qquad \mathrm{err}_{\mathrm{out}}(B) \ge \frac{s}{2Q} \ge \frac{1}{96}\min\left\{1, \frac{b_*}{\alpha Q}\right\}.$$

If $k_{\max} \le C_w k_{\min}$, the latter bound is at least

$$\frac{1}{192 L C_w}\min\left\{1, \frac{J}{\alpha Q}\right\}.$$

Proof. Before sampling, fix the first s rows R. Write $y(a,b) = (a+b) \bmod Q$ and

$$z(a,b) = a + b + \mathbf{1}_R(a), \qquad y'(a,b) = z(a,b) \bmod Q.$$

On $R \times \mathbb{Z}_Q$, $y' = y + 1 \bmod Q$ differs from y; elsewhere it agrees with y. Since z is an integer between zero and $2Q-1$, three ReLU responses give an exact representation:

$$g_0 = \sigma(z+1), \quad g_+ = \sigma(z-Q+1), \quad g_- = \sigma(z-Q), \qquad y' = g_0 - 1 - Qg_+ + Qg_-.$$

Let $E = S \cap (R \times \mathbb{Z}_Q)$, and let d be its maximum column degree. Ordering the edges within each column partitions E into d sets, each containing at most one edge in any column. Every nonnegative table supported on one such set has a three-ReLU representation. Indeed, give row a the integer code $w_a = a + 1$. For a column with edge $(a(b), b)$, set $u_b = w_{a(b)}$ and let $v_b \ge 0$ be the desired edge value; for an empty column set $u_b = -1, v_b = 0$. For an integer $M > \max_b v_b$,

$$v_b \mathbf{1}_{\{w_a = u_b\}} = \sigma(M(w_a - u_b) + v_b) + \sigma(M(w_a - u_b) - v_b) - 2\sigma(M(w_a - u_b)).$$

The identity follows at a zero code difference, and at every nonzero integer code difference all three terms are either linear or zero. Each preactivation is a sum of one-operand profiles and is thus legal under the original separate one-hot inputs.

Apply this representation to the two nonnegative tables

$$W = \mathbf{1}_E, \qquad U = (y - y' + 1)\mathbf{1}_E.$$

The nonzero values of U equal Q, so $M = 2Q + 2$ suffices. Together they use at most $6d$ responses. The integer-valued table

$$v = y' + U - W$$

equals y on S and outside $R \times \mathbb{Z}_Q$, and equals $y' \ne y$ on $(R \times \mathbb{Z}_Q) \setminus S$. The class scores $F_c = 2cv - c^2$, $0 \le c < Q$, uniquely select class v, with winning margin at least one.

To realize all $h = 3 + 6d$ responses in original layer i_*, write them as $\mathbf{g} = \sigma(Ax + \gamma)$, where $x = (e_a, e_b)$. Their linear combination has the form $v = \zeta^\top \mathbf{g} - 1$. If $i_* > 1$, choose a constant vector q such that $p = Ax + \gamma + q \ge \mathbf{1}$ throughout the finite input set. The first layer computes $(p, 1)$, intermediate layers transmit it, and layer i_* computes $(\sigma(p-q), 1) = (\mathbf{g}, 1)$. If $i_* = 1$, compute $(\mathbf{g}, 1)$ directly. If a later hidden layer is present, it computes $(\sigma(\zeta^\top \mathbf{g} - 1), 1) = (v, 1)$, which subsequent layers transmit; the output weights for class c are $(2c, -c^2)$. If $i_* = L$, the output weights on $(\mathbf{g}, 1)$ are instead $(2c\zeta^\top, -2c - c^2)$. All output biases may therefore be zero, as required under either original output-bias convention. Thus each layer uses at most $h + 1$ original positions. Eliminating the stable affine and nonnegative transmission responses by the permitted exact reductions leaves

$$t_0 \le 1, \qquad t_{i_*} \le h, \qquad t_i = 0 \quad (i \ne i_*).$$

The only remaining input-affine output contribution is constant, so its coefficient matrix has rank at most one.

For completeness, a hypergeometric variable X with mean μ satisfies $\mathbb{E}e^{\theta X} \le \exp(\mu(e^\theta - 1))$ for every real θ. Indeed, its exponential moment is an elementary symmetric mean of weights e^θ and 1, bounded by the corresponding power of their arithmetic mean. Each column degree in E has this distribution with mean $\mu = \alpha s$. Taking $\theta = \log 2$ gives

$$\Pr\{X \ge 2\mu + 2t\} \le \exp((1 - 2\log 2)\mu - 2t\log 2) \le e^{-t}.$$

A union over the Q columns therefore gives

$$\Pr\{d > \lceil 2\alpha s + 2t \rceil\} \le Qe^{-t} \le \delta/2.$$

On the complementary event,

$$h \le 9 + 12\alpha s + 12t \le 9 + b_*/2 + 12t \le b_*.$$

Consequently $h+1 \le \min_i k_i$ and $h \le \tau_{i_*}$, so the constructed behavior lies in the original architecture and budget.

The number of unobserved points in the selected rows is hypergeometric with mean $\mu_R = (1-\alpha)sQ$. The same exponential-moment estimate at $\theta = -\log 2$ gives

$$\Pr\left\{|(R \times \mathbb{Z}_Q) \setminus S| < (1-\alpha)sQ/2\right\} \le e^{-(1-\log 2)\mu_R/2} \le e^{-\mu_R/8} \le \delta/2.$$

All these unobserved points are errors of the constructed behavior, and they are its only errors. Dividing by $Q^2 - n$ proves the $s/(2Q)$ bound. Since $b_*/(24\alpha) > 2$, rounding gives

$$s \ge \min\{Q, b_*/(48\alpha)\} \ge \frac{1}{48}\min\{Q, b_*/\alpha\}.$$

Finally, under the width-ratio condition, $\tau_{i_*} \ge J/L$ and $k_{\min} - 1 \ge k_{\min}/2 \ge J/(2LC_w)$; the first finite condition ensures $k_{\min} \ge 2$. Hence $b_* \ge J/(2LC_w)$, proving the final bound. □

Corollary E.6 (Finite-sample budget range)**.** *Consider the cyclic-group task and original model of Proposition E.5. Fix an integer $L \ge 1$, a constant $C_w \ge 1$, a positive integer $\bar{a}$, and $0 < \alpha_0 < 1$. There are constants $c, C, C' > 0$, depending only on these fixed parameters, such that the following holds for every $Q \ge 2$. For any original architecture with hidden widths $k_1, \dots, k_L \in \mathbb{N}^*$, put $k_{\min} = \min_i k_i$ and $k_{\max} = \max_i k_i$, and assume $k_{\max} \le C_w k_{\min}$. For any sample-independent nonlinear budget $\boldsymbol{\tau} \in \mathbb{N}_0^L$ with effective total $J = \sum_i \min\{\tau_i, k_i\}$, and any $0 < \delta < 1$, suppose*

$$C\log(4Q/\delta) \le J \le Q, \qquad 1 \le n \le \alpha_0 Q^2, \qquad \alpha = n/Q^2.$$

For a uniformly sampled n-element training set S, with probability at least $1 - \delta$,

$$c\min\left\{1, \frac{J}{\alpha Q}\right\} \le \sup_{\substack{B \in \mathcal{B}_{\mathcal{A}}(\bar{a}, \boldsymbol{\tau}) \\ \mathrm{err}_{\mathrm{in}}(B) = 0}} \mathrm{err}_{\mathrm{out}}(B) \le \min\left\{1, C'\left[\frac{J\log Q}{\alpha Q} + \frac{\log(2/\delta)}{n}\right]\right\}.$$

Proof. Take $C = 768LC_w/(1-\alpha_0)$ and apply the proposition with failure probability $\delta/2$, so $t = \log(4Q/\delta)$. The width ratio gives $b_* \ge J/(2LC_w) \ge 384t/(1-\alpha_0)$, which implies the first finite condition. Since $b_* \le J \le Q$, the rounding bound in the proof gives $s \ge b_*/48$, and hence

$$(1-\alpha)sQ \ge (1-\alpha_0)b_*Q/48 \ge 8Qt \ge 8\log(4/\delta).$$

This verifies the second finite condition. The upper bound follows from the independent behavior count and Lemma D.1, also with failure probability $\delta/2$; for $J \le Q$ and fixed $\bar{a}$, its logarithmic behavior count is at most $C_{L,\bar{a}}QJ\log Q$. A union bound completes the proof. □

Comparison with the independent upper bound. Within the finite budget range of Corollary E.6, the upper and lower bounds match in their polynomial dependence on the budget and sampling density, up to one logarithmic factor. For example, let $m = Q^{7/8}/\log Q$, take original hidden widths $k_i = \lceil C_1 m \rceil$, and fix layer budgets $\tau_i = \lfloor C_2 m \rfloor$, where $0 < C_2 < C_1$. These are among the architectures permitted by the weight-decay theorem, and their effective total budget satisfies $J \asymp m$. For $\alpha \asymp Q^{-\beta}$, $0 \le \beta < 1/8$, and $\delta = Q^{-2}$, the finite conditions above hold for sufficiently large Q. The resulting budget classes therefore contain training-consistent behaviors with held-out error at least $cQ^{\beta - 1/8}/\log Q$. Consequently a guarantee based only on this budget and training consistency, valid uniformly over all the allowed architectures, cannot improve the polynomial error exponent. The regularized objective may select a smaller subset of these behaviors.

F UNIVERSAL FINITE-GROUP TARGET CONSTRUCTIONS

This section proves the target bound used in Theorem 3.3. The construction uses only the public separate one-hot inputs and belongs exactly to the fixed model $\mathcal{S}$ of Definition 2.1. Consequently every table depending on one operand alone is a legal affine score of that operand, while every factor in the construction remains charged.

F.1 A UNIVERSAL SQUARE-ROOT CONSTRUCTION

Proof of Theorem 3.3. Assume $Q \geq 2$. A unitary representation assigns a unitary matrix $\pi(g) \in \mathbb{C}^{m\times m}$ to each $g \in G$, with $\pi(gh) = \pi(g)\pi(h)$. It is irreducible when these matrices have no common nonzero proper invariant subspace. Choose one representative from each equivalence class under unitary changes of basis, and sample a representation π of dimension m with probability m^2/Q. These probabilities sum to one, and $m \leq \sqrt{Q}$. Conditional on π, let U, V be independent uniform vectors on the complex sphere of radius $\sqrt{m}$. For a vector Z with this distribution put $s(z) = \operatorname{sgn} \Re z + i \operatorname{sgn} \Im z, \qquad \gamma_m = 2\mathbb{E}|\Re Z_1|$. Coordinate sign symmetries and rotational invariance give

$$\mathbb{E}[\overline{Z}_i\, s(Z_k)] = \gamma_m \mathbf{1}_{\{i=k\}}, \qquad \sqrt{2/3} \leq \gamma_m \leq \sqrt{2}. \tag{F.1}$$

For the bounds, $\mathbb{E}(\Re Z_1)^2 = 1/2$ and $\mathbb{E}(\Re Z_1)^4 = 3m/[4(m+1)] \leq 3/4$. Cauchy–Schwarz and moment interpolation give $\mathbb{E}|X| \leq (\mathbb{E}X^2)^{1/2}$ and $\mathbb{E}|X| \geq (\mathbb{E}X^2)^{3/2}/(\mathbb{E}X^4)^{1/2}$. No independence between sphere coordinates is used.

Define the operand features and output feature $A(a) = U^*\pi(a), \quad B(b) = \pi(b)V, \quad \eta(a,b) = m^{-1/2}\sum_{k=1}^m s(A_k(a))s(B_k(b)), \quad \zeta(c) = \frac{U^*\pi(c)V}{\gamma_m^2\sqrt{m}}$. For fixed π, a, b, c, the vectors $A(a)$ and $B(b)$ are independent uniform sphere vectors, and $U^*\pi(c)V = A(a)\pi(a)^{-1}\pi(c)\pi(b)^{-1}B(b)$. Applying (F.1) twice and cyclically permuting the trace gives $\mathbb{E}[\overline{\zeta(c)}\eta(a,b) \mid \pi] = \frac{1}{m}\operatorname{tr}\pi(c^{-1}ab)$. The regular-character identity therefore yields

$$\mathbb{E}[\overline{\zeta(c)}\eta(a,b)] = \frac{1}{Q}\sum_{\pi} m_\pi \operatorname{tr}\pi(c^{-1}ab) = \mathbf{1}_{\{c=ab\}}. \tag{F.2}$$

The concentration constants are uniform in the representation dimension. For $m \geq 2$, a sphere coordinate satisfies $\mathbb{P}\{|Z_1| > u\} = (1-u^2/m)^{m-1} \leq e^{-u^2/2}$ for $0 \leq u < \sqrt{m}$; for $m = 1$ its modulus is one. Conditional on V, $U^*\pi(c)V/\sqrt{m}$ has this same coordinate distribution. Hence $\zeta(c)$ is uniformly sub-Gaussian. Conditional on $A(a)$ and the absolute values of the real and imaginary coordinates of $B(b)$, their signs are independent Rademacher variables. The real and imaginary parts of $\eta(a,b)$ are normalized sums of $2m$ such signs with coefficients of modulus at most one, and so are uniformly sub-Gaussian as well.

Consequently $Y = \Re(\overline{\zeta}\,\eta)$ satisfies $\mathbb{E}|Y - \mathbb{E}Y|^p \leq p!C^p$ for all integers $p \geq 2$. Indeed Cauchy–Schwarz combines the two sub-Gaussian moment bounds; independence of ζ and η is not required. Bernstein's inequality for d independent copies gives $\mathbb{P}\left\{\left|d^{-1}\sum_{j=1}^d Y_j - \mathbb{E}Y\right| > 1/8\right\} \leq 2e^{-cd}$. The same argument bounds $d^{-1}\sum_j |\zeta_j(c)|^2$ by an absolute constant with failure probability $2e^{-cd}$. Take $d = \lceil C_0 \log Q\rceil$ with C_0 sufficiently large and union bound over the Q^3 triples and Q output classes. Some deterministic choice of these features then satisfies

$$F_c^0(a,b) = \frac{1}{d}\sum_{j=1}^d \Re\big(\overline{\zeta_j(c)}\eta_j(a,b)\big), \qquad \max_{a,b,c}|F_c^0(a,b) - \mathbf{1}_{\{c=ab\}}| \leq 1/8. \tag{F.3}$$

Its complete-table winning margin is at least $3/4$.

It remains to realize these logits in the model of Section 2. Every real or imaginary sign of $A_k(a)$ or $B_k(b)$ is a public one-operand lookup and hence a legal affine one-hot score. For $x, y \in \{-1, 1\}$,

$$xy = \tfrac{1}{2}[\sigma(x+y) + \sigma(-x-y) - \sigma(x-y) - \sigma(-x+y)]. \tag{F.4}$$

Expanding one complex product uses four real sign products, hence at most sixteen ReLU factors. Taking the specified linear readout realizes F^0 with no free affine output and with at most

$16\sum_{j=1}^{d} m_j \le 16d\sqrt{Q} \le C\sqrt{Q}\log Q$ factors. Rescale the readout by $4/3$ if unit margin is desired. This realization therefore upper-bounds the EGC of the complete group law in the same model class used throughout the paper. Finally $\sqrt{Q}\log Q = o(Q/(\log Q)^A)$ for every fixed A, which gives all the asserted asymptotic target bounds. □

F.2 Uniform selection with a training-error tolerance

Proof of the window and selection in Theorem B.3. Work on the event of Theorem 3.2, and fix $\varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}} \in [0,1]$. If $\mathrm{err}_{\mathrm{in}}(B) \le \varepsilon_{\mathrm{in}}$ and $\mathrm{err}_{\mathrm{out}}(B) \ge \varepsilon_{\mathrm{out}}$, then

$$\kappa(B) \ge \frac{Q}{C_\alpha \log(2Q)} \left[\varepsilon_{\mathrm{out}} - C_\alpha \varepsilon_{\mathrm{in}} - \frac{C_\alpha \log(1/\delta)}{Q^2} \right] - 1. \tag{F.5}$$

Under the sufficient condition in Theorem B.3, the right-hand side is strictly larger than m_G. Hence $m_G < K_{\mathrm{high}}(S, \varepsilon_{\mathrm{in}}, \varepsilon_{\mathrm{out}}; G)$, also when the defining set is empty.

The target construction gives $\kappa(B_G) \le m_G$. Since B_G has zero training error, it is feasible for every $\varepsilon_{\mathrm{in}} \ge 0$, proving (B.11). Together with the preceding lower bound, this also proves (B.14).

For $\varepsilon_{\mathrm{in}} \in [0,1]$, the minimizing set $\mathfrak{M}_Q(S, \varepsilon_{\mathrm{in}})$ is nonempty because complete prediction maps form a finite set and B_G is feasible. Every minimizer has EGC at most m_G. Applying (3.2) gives, on the same event and uniformly over $\varepsilon_{\mathrm{in}}$ and all minimizers,

$$\sup_{B \in \mathfrak{M}_Q(S, \varepsilon_{\mathrm{in}})} \mathrm{err}_{\mathrm{out}}(B) \le C_\alpha \left[\varepsilon_{\mathrm{in}} + \frac{(m_G+1)\log(2Q)}{Q} + \frac{\log(1/\delta)}{Q^2} \right]. \tag{F.6}$$

This proves (B.13); no choice of a particular minimizer or uniqueness is used. □

G Sparse Sampling through Exact EGC Compression

This appendix proves Theorem 4.2. The order of the argument is essential: we first construct, in the original architecture, an exact realization of the complete prediction map; only then do we apply the independent EGC generalization theorem. No full-table error bound for the trained network is used to obtain the EGC bound.

Throughout, G is a finite abelian group of order Q, $X = G^2$, $\ell = \log Q$, and S is sampled uniformly without replacement from X. Write $n = |S|$ and $\alpha_Q = n/Q^2$. The affine depth $D \ge 2$ is fixed. We first establish a pointwise result for $D \ge 3$. Section G.9 supplies the one-hidden-layer estimates, and Section G.10 proves Theorem 4.2 for every $D \ge 2$ with common budgets throughout the decay interval. The hidden widths satisfy $k_i \ge C_D \ell$ and $\max_i k_i / \min_i k_i \le C_{\mathrm{w}}$, as in Section 4.1. All constructions use the same one-hot inputs, ReLU activation, connections, widths, and output-bias convention as the original model. For a displayed network F, $R(F)$ denotes the penalty evaluated at its specified parameters; $\mathbb{E}_X$ denotes the uniform average on X.

G.1 Objective bounds used in the construction

Write $\mu = \lambda Q^{2/D}$ and $\nu = 2 - 1/D$, and let $\overline{\Delta}$ bound the original objective excess. We retain μ and $\overline{\Delta}$ in the estimates below. The finite conditions in Section G.10 and the proof of Corollary G.7 derive the required energy bounds from the original objective before applying these estimates.

G.2 Sample event and objective control

Every row, column, and set $\{(a,b) : ab = c\}$, called a label fiber, contains Q inputs. Standard hypergeometric bounds and a union bound show that, with probability $1 - O(Q^{-4})$, each sampled row, column, and label fiber has between constant multiples of n/Q points, and the label frequencies have their usual $O(\sqrt{\ell/(nQ)})$ deviations. In particular, for every matrix A,

$$\frac{1}{n} \sum_{(a,b) \in S} \|A(e_a, e_b)\|^2 \le \frac{C}{Q} \|A\|_F^2. \tag{G.1}$$

Indeed, writing $A = [A_1, A_2]$ and using $\|A_1 e_a + A_2 e_b\|^2 \le 2\|A_1 e_a\|^2 + 2\|A_2 e_b\|^2$, each column is counted at most Cn/Q times. This event is independent of the learned parameters.

The group-law comparator of Lemma I.2, scaled to margin 4ℓ, gives for every training set

$$R(\theta) \le C_D Q^{2/D}\ell + \frac{Q^{-3} + \overline{\Delta}}{\lambda}, \qquad \mathcal{L}_S(F_\theta) \le C_D \mu\ell + Q^{-3} + \overline{\Delta}.$$

In the application to Corollary G.7, its nonsaturated branch proves $\overline{\Delta}/(\mu\ell) \le \alpha_Q/\ell^{D+2}$, and the finite conditions bound $Q^{-3}/(\mu\ell)$. Thus the following intermediate estimates use

$$R(\theta) \le E_0 := C_D Q^{2/D}\ell, \qquad \mathcal{L}_S(F_\theta) \le C_D\mu\ell. \tag{G.2}$$

Every weight matrix and bias consequently has norm at most $q = C_D Q^{1/D}\sqrt{\ell}$. The saturated branch is treated by an explicit table realization and does not use this energy bound.

G.3 A GROUP-LAW COMPARATOR ON ANY SUBRECTANGLE

Lemma G.1 (Subrectangle comparator)**.** *For nonempty $I, J \subseteq G$, put $v = |I| + |J|$. There is a depth-D ReLU score network $G_{I,J}$ using at most $C_D\ell$ positions per hidden layer such that*

$$\begin{aligned} G_{I,J}(a,b) &= 0 && ((a,b) \notin I \times J),\\ G_{I,J,ab}(a,b) - G_{I,J,c}(a,b) &\ge 1 && ((a,b) \in I \times J,\ c \ne ab),\\ R(G_{I,J}) &\le C_D Q^{2/D}\ell^{1-2/D}(v/Q)^{1/D}. \end{aligned} \tag{G.3}$$

The construction uses no output bias.

Proof. A character is a homomorphism $\chi : G \to \{z \in \mathbb{C} : |z| = 1\}$. Let e be the group identity. Choose $M = \lceil C\ell \rceil$ characters so that, for every $g \ne e$, $M^{-1}\sum_{j=1}^{M} \operatorname{Re}\chi_j(g) \le 1/2$; character orthogonality, Hoeffding's inequality, and a union over $G \setminus \{e\}$ give a fixed choice. The scores

$$T_c(a,b) = \frac{1}{M}\sum_{j=1}^{M} \operatorname{Re}\big(\chi_j(a)\chi_j(b)\overline{\chi_j(c)}\big)$$

therefore separate the correct class by a constant. Each complex product is a fixed linear combination of four real products. On $[-1,1]^2$ these products have constant-accuracy ReLU representations, for example by a fixed Riemann sum of

$$xy = \frac{1}{2}\int_0^2 [\sigma(x+y-t) + \sigma(-x-y-t) - \sigma(x-y-t) - \sigma(-x+y-t)]\,dt.$$

Multiplying by a fixed constant gives unit margin.

To restrict a factor $\sigma(u_a + w_b - t)$ to $I \times J$, replace it by

$$\sigma(\mathbf{1}_I(a)(u_a + 4) + \mathbf{1}_J(b)(w_b + 4) - 8 - t)\,.$$

It agrees with the original factor on $I \times J$ and is zero elsewhere. The first-layer squared norm is $O((v+1)\ell)$, the output squared norm is $O(Q/\ell)$, and each intervening identity block has squared norm $O(\ell)$. Positive homogeneity balances the D blocks, including the first-layer biases, and yields (G.3). □

G.4 FREEING POSITIONS WITHOUT CHANGING THE MODEL CLASS

Let $k = \min_i k_i$ and $r = C_D\lceil\ell\rceil$, where the fixed constant accommodates Lemma G.1 and the constant transmissions below.

Lemma G.2 (Free positions)**.** *Suppose* (G.2) *holds and $k \ge 2r$. In every hidden layer one can delete r positions and obtain a score network $\widetilde{F}$ using the remaining original positions and at most $D-1$ constant positions per layer, with*

$$\left(\frac{1}{n}\sum_S \|\widetilde{F} - F\|^2\right)^{1/2} \le C_D\frac{\sqrt{Q}\,\ell^{D/2+1}}{k}, \qquad R(\widetilde{F}) \le R(F) + C_D Q^{2/D}\ell(\ell/k)^{2/D}. \tag{G.4}$$

No stationarity condition is used.

Proof. Use $0 \in \mathbb{R}^{2Q}$ only as an auxiliary reference input and write $h_j^0 = h_j(0)$. From (G.2) and (G.1),

$$\|h_j^0\| \le C_D q^j, \qquad \left(\frac{1}{n}\sum_S \|h_j(x) - h_j^0\|^2\right)^{1/2} \le C_D q^j/\sqrt{Q}.$$

At layer j, let c_i be column i of the next matrix, z_i the current response, and

$$u_i = \left(n^{-1}\sum_S |z_i(x) - z_i(0)|^2\right)^{1/2}, \qquad A_i = \|c_i\|\left(u_i + |z_i(0)|/\sqrt{Q}\right).$$

Cauchy–Schwarz gives $\sum_i A_i \le C_D q^{j+1}/\sqrt{Q}$. Delete the r smallest A_i. Their nonconstant contribution to the next preactivation is at most $C_D(r/k_j)q^{j+1}/\sqrt{Q}$, while their constant contribution $d_{j+1} = \sum c_i z_i(0)$ has norm at most $C_D(r/k_j)q^{j+1}$. Propagating the former through the remaining matrices and summing over the fixed number of layers gives the first bound in (G.4).

The constant d_j is not treated as a free bias. A separate positive constant chain of length j transmits it to layer j, with the original bias unchanged. Balancing the scalar links and final vector gives penalty $(j/2)\|d_j\|^{2/j} \le C_D q^2 (r/k)^{2/D}$. These chains use at most $D-1$ positions per layer and remain valid when the output bias is fixed to zero. Summing over j proves the second bound. □

The freed positions contain all constant chains and the network $G_{I,J}$. Hence $\widetilde{F} + G_{I,J}$ is a legal parameter in the same original architecture.

G.5 Local small margins from the original objective

Let $\mathcal{A}_F = \{(a,b) : \mathrm{gap}_F(a,b) \le 1/2\}$, where the gap is the difference between the largest and second-largest scores. Adding $G_{I,J}$ cannot increase the cross-entropy on $I \times J$ and reduces it by a fixed constant on sampled points in $\mathcal{A}_F \cap (I \times J)$. The cross-entropy gradient with respect to the score vector has norm at most $\sqrt{2}$. Comparing the original objective values of F and $\widetilde{F} + G_{I,J}$ therefore yields, for every I, J,

$$\frac{|S \cap \mathcal{A}_F \cap (I \times J)|}{n} \le C_D\left[\mu \ell^{1-2/D}(v/Q)^{1/D} + e_k + \overline{\Delta}\right], \tag{G.5}$$

where

$$e_k = \frac{\sqrt{Q}\,\ell^{D/2+1}}{k} + \mu\ell(\ell/k)^{2/D}. \tag{G.6}$$

The factor $(v/Q)^{1/D}$ is the gain over a comparison based only on total training loss. It follows from the paid penalty of the subrectangle comparator and is not an assumption on the trained network.

G.6 Row–column approximation and local score covers

The original readout at arbitrary sampling density. Write $\alpha = n/Q^2$, $\ell = \log Q$, and let n_c be the sample count of class c. On the common row, column, and class-fiber event,

$$n_c \ge \alpha Q/2, \qquad \pi_c := n_c/n \le C/Q, \qquad |\pi_c - \pi_d| \le C\sqrt{\ell/(\alpha Q^3)}.$$

The event follows from hypergeometric concentration when $\alpha Q \ge C\ell$. Use the original-objective comparison

$$R(\theta) \le E_0, \qquad J(\theta)/\lambda \le C_D E_0, \qquad E_0 = C_D Q^{2/D}\ell.$$

Fix all original hidden parameters and minimize only the original readout. Denote the original and conditionally minimizing parameters by Ω, Ω^0, including a bias coordinate precisely when the original model has an output bias. Strong convexity and the definition of the original objective gap Δ give

$$\|\Omega - \Omega^0\|_F^2 \le 2\Delta/\lambda, \qquad \|\Omega^0\|_F^2 \le C_D E_0.$$

The first inequality uses that the conditional minimum is no smaller than the minimum over the original architecture.

Let $z(x)$ be the hidden response, augmented by a constant one only in the original biased-output model, and put

$$m_c = n_c^{-1} \sum_{\substack{x \in S \\ B_G(x)=c}} z(x), \qquad P_* = \prod_{i<D} \|W_i\|_F.$$

Every group-label fiber uses each operand at most once. Subtract the coordinate means from each of the two blocks of W_1. Their squared Frobenius norms do not increase, and hence, with $x_0 = Q^{-1}(\mathbf{1}, \mathbf{1})$,

$$\frac{1}{n_c} \sum_{\substack{x \in S \\ B_G(x)=c}} \|W_1(x - x_0)\|_2^2 \le \frac{2\|W_1\|_F^2}{n_c}.$$

The subsequent layers are Lipschitz with product constant $\prod_{2 \le i < D} \|W_i\|_F$. Their biases cancel when comparing responses at two inputs. Therefore

$$\|m_c - z(x_0)\|_2 \le \sqrt{2} P_* / \sqrt{n_c}, \qquad \max_{c,d} \|m_c - m_d\|_2 \le C_D P_* / \sqrt{\alpha Q}.$$

The augmented constant coordinate, when present, has zero difference. The recursion $\|h_i(x_0)\|_2 \le \|W_i\|_F \|h_{i-1}(x_0)\|_2 + \|b_i\|_2$ and $E_0 \ge 1$ also imply

$$\max_c \|m_c\|_2 \le C_D E_0^{(D-1)/2}, \qquad P_*^2 \le C_D E_0^{D-1}.$$

Let ω_c^0 be a row of Ω^0, and let $p_c^0(x)$ be its softmax probability. Subtracting its two stationarity equations gives

$$\lambda(\omega_c^0 - \omega_d^0) + \frac{1}{n} \sum_{x \in S} (p_c^0(x) - p_d^0(x)) z(x) = \pi_c m_c - \pi_d m_d.$$

Pair with $v_{cd} = \omega_c^0 - \omega_d^0$. Since $v_{cd}^{\mathsf{T}} z = \log(p_c^0 / p_d^0)$, each summand on the left is nonnegative after pairing. Thus

$$\|v_{cd}\|_2 \le \lambda^{-1} \left[\pi_c \|m_c - m_d\|_2 + |\pi_c - \pi_d| \, \|m_d\|_2 \right].$$

Choose d with $\|\omega_d^0\|_2 \le \|\Omega^0\|_F / \sqrt{Q}$, then transfer back to Ω. For the actual original weight rows this proves

$$s_F := \max_c \|W_{D,c:}\|_2^2 \le C_D \left[\frac{E_0}{Q} + \frac{E_0^{D-1} \ell}{\alpha \lambda^2 Q^3} + \frac{\Delta}{\lambda} \right]. \tag{G.7}$$

This bound holds in both original output-bias conventions. At fixed nominal density, use the actual n/Q^2, which is at least half the nominal density under the stated finite conditions. Its reciprocal is absorbed into the allowed fixed-density constant.

Lemma G.3 (Layerwise approximation with row and column control)**.** *Let $Q \ge 2$, $D \ge 3$, and $L = D - 1$. Consider an original fully connected ReLU architecture with positive hidden widths k_i, hidden biases, and output bias either allowed or fixed to zero. On $X = G^2$, write*

$$h_0(a, b) = (e_a, e_b), \qquad h_i = \sigma(W_i h_{i-1} + b_i), \qquad F = W_D h_L + b_D.$$

Set

$$P_i = \prod_{j \le i} \|W_j\|_F, \quad P_* = P_L, \quad P = P_* \|W_D\|_F,$$

$$s_F = \max_c \|W_{D,c:}\|_2^2, \quad T = P_* \sqrt{s_F}, \quad \ell = \log Q.$$

For every integer $K \ge 1$, there is a score network H in this same architecture using at most $\min(k_i, K)$ positions for $i < L$. The last hidden layer uses at most $\min(k_L, K)$ positions when an output bias is allowed and at most $\min(k_L, K + 1)$ otherwise. For

$$e(a, b) = \max_{c,d} |(F_c - F_d)(a, b) - (H_c - H_d)(a, b)|,$$

there are nonnegative r_a, s_b satisfying

$$e(a, b)^2 \le r_a + s_b, \qquad \sum_a r_a + \sum_b s_b \le C_D \left(\frac{T^2 \ell}{K} + \frac{P^2 \ell^2}{K^2} \right). \tag{G.8}$$

The same H has a centered analytical representation

$$H(x) = F(x_0) + U\psi_L(x), \qquad x_0 = Q^{-1}(\mathbf{1}, \mathbf{1}), \qquad \|U\|_F \le 2P, \tag{G.9}$$

with at most $\min(k_i, K)$ *centered coordinates in each layer. These coordinates have the form*

$$\begin{aligned} \psi_{i,j} &= a_{i,j}\big[\sigma(v_{i,j}^{\mathsf{T}}\psi_{i-1} + \beta_{i,j}) - \sigma(\beta_{i,j})\big], \\ a_{i,j} \ge 0, \quad \sum_j a_{i,j}^2 \le 1, \quad \|v_{i,j}\|_2 \le 1, \quad \psi_0 &= h_0 - x_0. \end{aligned} \tag{G.10}$$

There are nonnegative R_a, S_b, common to all these layers, such that

$$\|\psi_i(a,b)\|_2^2 \le R_a + S_b, \qquad \sum_a R_a + \sum_b S_b \le 2, \qquad \mathbb{E}_X\|\psi_i\|_2^2 \le Q^{-1} \quad (1 \le i \le L). \tag{G.11}$$

The reference value $F(x_0)$ is the value of the original network on the centered one-hot mean. The factorization (G.9)–(G.10) *is analytical; the score network H has the original connection and output-bias conventions.*

Proof. For a vector-valued table V, define

$$\|V\|_{\mathrm{rc}}^2 = \min_{\substack{r_a, s_b \ge 0 \\ \|V(a,b)\|_2^2 \le r_a + s_b}} \left(\sum_a r_a + \sum_b s_b \right).$$

The minimum is attained. Its square root is a norm: for two feasible bounds, apply

$$(\sqrt{A} + \sqrt{B})^2 \le (1+t)A + (1+t^{-1})B$$

pointwise and choose $t = \|W\|_{\mathrm{rc}}/\|V\|_{\mathrm{rc}}$; zero cases follow directly. The norm is monotone under pointwise domination of Euclidean norms.

A sampling estimate. Put $N = Q^2$. For deterministic vector tables V_j and independent random signs ε_j,

$$\mathbb{E}_\varepsilon \left\| \sum_j \varepsilon_j V_j \right\|_{\mathrm{rc}}^2 \le B_Q \sum_j \|V_j\|_{\mathrm{rc}}^2, \qquad B_Q = 6 + 4\log N. \tag{G.12}$$

To see this, choose optimal bounds r_{ja}, s_{jb} for each table. At a fixed input x, put $v_x = \sum_j \|V_j(x)\|_2^2$. The expected Euclidean norm of the sign sum is at most $\sqrt{v_x}$. The bounded-differences inequality gives an excess of at most $\sqrt{2v_x t}$ with failure probability at most e^{-t}. If $v_x = 0$, the sum is identically zero. A union over the N inputs with $t = \log N + u$ bounds all squared norms by

$$(2 + 4\log N + 4u)\left(\sum_j r_{ja} + \sum_j s_{jb} \right)$$

with failure probability at most e^{-u}. Integrating the resulting bound proves (G.12). Symmetrization with independent copies, followed by (G.12), gives for independent copies $Z_1, \dots, Z_K$ of a random vector table Z,

$$\mathbb{E}\left\| K^{-1}\sum_{t=1}^K Z_t - \mathbb{E}Z \right\|_{\mathrm{rc}}^2 \le \frac{4B_Q}{K}\,\mathbb{E}\|Z\|_{\mathrm{rc}}^2. \tag{G.13}$$

Indeed Jensen's inequality replaces $\mathbb{E}Z$ by an independent copy; the differences can be symmetrized, and the squared triangle inequality bounds their sign sum by twice each of the two sign sums.

Centered construction and dimensions. Before processing hidden layer i, its preactivation is

$$A_i\psi_{i-1} + \beta_i, \qquad A_i \in \mathbb{R}^{k_i \times m_{i-1}}, \quad \psi_{i-1} = h'_{i-1} - q_{i-1} \in \mathbb{R}^{m_{i-1}},$$

and subsequent matrices $W_{i+1}, \dots, W_D$ are still original. Initially $m_0 = 2Q$, $q_0 = x_0$, $A_1 = W_1$, and $\beta_1 = b_1 + W_1 x_0$. All changes preserve the preactivations at x_0.

For $i < L$, write the rows of A_i as a_j^T and the columns of $C = W_{i+1}$ as c_j. Discard zero outgoing columns and include constant contributions from zero incoming rows in $\beta_{i+1} = b_{i+1} + C\sigma(\beta_i)$. On the remaining indices put

$$r_j = \|a_j\|_2, \quad M_i = \sum_j r_j \|c_j\|_2, \quad p_j = \frac{r_j\|c_j\|_2}{M_i}, \quad f_j(\psi) = \frac{\sigma(\beta_{i,j} + a_j^\mathsf{T}\psi) - \sigma(\beta_{i,j})}{r_j}.$$

Then

$$b_{i+1} + C\sigma(A_i\psi + \beta_i) = \beta_{i+1} + \sum_j r_j c_j f_j(\psi). \tag{G.14}$$

If $M_i = 0$, the current score map is constant and equals $F(x_0)$; one constant final hidden response implements it with zero output bias, or the output bias implements it when allowed. The same convention applies if the final nonconstant output vanishes. Its analytical factorization has $U = 0$ and zero centered features.

If $k_i \le K$, retain all remaining terms and set

$$(\psi_i)_j = \sqrt{p_j} f_j(\psi_{i-1}), \qquad (A_{i+1})_{:j} = \frac{r_j c_j}{\sqrt{p_j}}.$$

If $k_i > K$, draw K independent indices J_t with probabilities p_j, and set

$$(\psi_i)_t = K^{-1/2} f_{J_t}(\psi_{i-1}), \qquad (A_{i+1})_{:t} = \frac{r_{J_t} c_{J_t}}{\sqrt{K}\, p_{J_t}}. \tag{G.15}$$

Repeated indices occupy separate original positions. In either case $m_i \le \min(k_i, K)$ and

$$\|A_{i+1}\|_F^2 = M_i^2 \le \|A_i\|_F^2 \|W_{i+1}\|_F^2, \qquad \|A_i\|_F \le P_i. \tag{G.16}$$

These bounds hold for every sampling outcome.

For each retained or sampled response with multiplier ρ_j equal to $\sqrt{p_j}$ or $K^{-1/2}$, let V_i have row $\rho_j a_j^\mathsf{T}/r_j$, and let $\widetilde{\beta}_i$ have entry $\rho_j \beta_{i,j}/r_j$. The actual layer is

$$h'_i = \sigma\big(V_i h'_{i-1} + \widetilde{\beta}_i - V_i q_{i-1}\big), \qquad q_i = \sigma(\widetilde{\beta}_i), \qquad \psi_i = h'_i - q_i. \tag{G.17}$$

Thus its incoming matrix is $V_i \in \mathbb{R}^{m_i \times m_{i-1}}$, its bias is the displayed vector, and A_{i+1}, not V_i, is the updated next-preactivation matrix. Unused original positions and connections are zero.

Bounds valid for every outcome. The response multipliers have squared sum one, and $|f_j(\psi)| \le \|\psi\|_2$. At the first layer, subtract the means of the two blocks of a_j/r_j, obtaining u_j, v_j. Then

$$\sum_a u_{ja}^2 + \sum_b v_{jb}^2 \le 1, \qquad |f_j(a,b)|^2 \le 2(u_{ja}^2 + v_{jb}^2).$$

A retained weighted sum or a sampled average gives R_a, S_b with total at most two. Subsequent normalized layers contract the feature norm pointwise, so the same bounds hold for all layers. Also

$$\mathbb{E}_X |w^\mathsf{T}(h_0 - x_0)|^2 \le Q^{-1}\|w\|_2^2$$

by the covariance of two independent uniform one-hot vectors. Applying this bound separately to each normalized first-layer row, and then contracting the feature norms, gives the last part of (G.11). This is a directional covariance calculation, not a bound $\mathbb{E}_X\|h_0 - x_0\|_2^2 \le Q^{-1}$.

Nonfinal error and deterministic choices. For a sampled layer the single-sample next-preactivation table is

$$Z_j = \frac{r_j c_j}{p_j} f_j.$$

Its row and column norm is at most $\sqrt{2}M_i$, using the individual first-layer bound when $i = 1$ and (G.11) otherwise. Equation (G.13) gives a sampling outcome for which the change δ_i in the next preactivation satisfies

$$\|\delta_i\|_{\mathrm{rc}}^2 \le 8B_Q M_i^2/K.$$

The unchanged subsequent layers amplify its class-difference error by at most

$$2\sqrt{s_F} \prod_{j=i+2}^{L} \|W_j\|_F.$$

Together with (G.16), this bounds the row and column norm squared of that layer's score-difference change by $CT^2\ell/K$. Fix such an outcome and continue. All matrix, feature, reference-value, and position bounds hold for every outcome, so they remain valid at the next step. A retained layer incurs no change. No random event over trained parameters is needed for these choices.

The final hidden layer. Put $A = A_L$, $C = W_D$, and let

$$A_2 = \sum_j r_j^2 \le P_*^2, \qquad M = \sum_j r_j \|C_{:j}\|_2 \le P,$$

where only nonconstant contributing rows are included in the sums. If $k_L > K$, sample K terms with

$$p_j = \frac{r_j^2}{2A_2} + \frac{r_j\|C_{:j}\|_2}{2M}. \tag{G.18}$$

For a fixed $\psi = \psi_{L-1}(x)$, the single-sample class output $Z_c = r_j C_{cj} f_j(\psi)/p_j$ satisfies

$$|Z_c| \le 2M\|\psi\|_2, \quad |\mathbb{E}Z_c| \le M\|\psi\|_2, \quad \mathrm{Var}(Z_c) \le 2A_2\|C_{c:}\|_2^2\|\psi\|_2^2.$$

Scalar Bernstein and a union bound over all NQ input–class pairs, with $t = \log(4NQ)$, give an outcome with class-difference error bounded pointwise by

$$4\left(T\sqrt{\frac{t}{K}} + \frac{Pt}{K}\right)\|\psi(x)\|_2. \tag{G.19}$$

The total failure probability is at most $1/2$, so such an outcome exists. By (G.11), the squared row and column norm in (G.19) is at most $C(T^2\ell/K + P^2\ell^2/K^2)$. Every outcome also gives

$$\|U\|_F^2 = \frac{1}{K}\sum_{t=1}^{K} \frac{r_{J_t}^2\|C_{:J_t}\|_2^2}{p_{J_t}^2} \le 4M^2 \le 4P^2. \tag{G.20}$$

Its features use multipliers $K^{-1/2}$, so (G.10)–(G.11) still hold. If $k_L \le K$, keep the raw responses and original readout and output bias, so the computation is retained exactly. The analytical normalization with $p_j = r_j\|C_{:j}\|_2/M$ gives $\psi_{L,j} = \sqrt{p_j} f_j$ and $U_{:j} = r_j C_{:j}/\sqrt{p_j}$, whence $\|U\|_F^2 = M^2 \le P^2$.

Final constants and total error. Let $\beta_D = b_D + W_D\sigma(\beta_L) = F(x_0)$. For a sampled final layer, the physical responses h'_L from (G.17) give scores

$$\beta_D + U\psi_L = Uh'_L + (\beta_D - Uq_L).$$

Use the displayed constant as the output bias when permitted. Otherwise put $\sigma(1) = 1$ in one extra original final position and use output column $\beta_D - Uq_L$. Sampling occurred only when $k_L > K$, so $K + 1 \le k_L$. If $k_L \le K$, keep the raw responses $\sigma(A_L\psi_{L-1} + \beta_L)$ and original readout W_D. The normalized ψ_L, U above are then an analytical factorization of those same scores, with the original b_D; no additional position or output bias is introduced.

The successive class differences telescope. The pointwise final error is at most the sum of the layerwise errors. The triangle inequality in the row and column norm, and the fixed number of layers, give (G.8). The same constructed network has (G.9)–(G.11) by the deterministic properties just established. □

Mean-square and exceptional-set consequences. For the same network, summing the bounds in (G.8) over G^2 gives

$$\mathbb{E}_X e^2 \le \frac{1}{Q}\left(\sum_a r_a + \sum_b s_b\right) \le \frac{C_D}{Q}\left(\frac{P_*^2 s_F \log(2Q)}{K} + \frac{P^2 \log^2(2Q)}{K^2}\right). \tag{G.21}$$

For any $\varepsilon > 0$, all inputs with $e > \varepsilon$ lie in the rows with $r_a > \varepsilon^2/2$ or columns with $s_b > \varepsilon^2/2$. Their total number is at most

$$\frac{2C_D}{\varepsilon^2}\left(\frac{T^2\ell}{K} + \frac{P^2\ell^2}{K^2}\right). \tag{G.22}$$

Row and column control from the original objective. Let $D \ge 3$, $\alpha = n/Q^2$, and $\mu = \lambda Q^{2/D} \le 1$. With P, T as in Lemma G.3, the energy bound and (G.7) give

$$P^2 \le C_D Q^2 \ell^D, \qquad T^2 \le C_D\left[\frac{Q\ell^{2D-1}}{\alpha\mu^2} + \frac{\Delta Q^2 \ell^{D-1}}{\mu}\right].$$

The lemma and its exceptional-set consequence, at precision $1/8$, therefore cover all approximation failures with at most

$$h_e \le C_D\left[\frac{Q\ell^{2D}}{\alpha\mu^2 K} + \frac{\Delta Q^2\ell^D}{\mu K} + \frac{Q^2\ell^{D+2}}{K^2}\right] \tag{G.23}$$

rows and columns. The same center has the readout and centered-feature bounds required in the following local covering argument.

Lemma G.4 (Local score covering)**.** *Let $K, h_0 \in \mathbb{N}^*$. Consider the centered score representations constructed above, with at most K responses per hidden layer, the deterministic bounds*

$$\|U\|_F \le 2P_0, \qquad \|F(x_0)\|_\infty \le B_0, \qquad P_0^2 = C_D Q^2 \ell^D, \qquad B_0 \le C_D Q^{C_D},$$

and the response bounds (G.11)*. Delete rows with $R_a > 1/h_0$ and columns with $S_b > 1/h_0$; there are at most $2h_0$ of them. For every nonempty $I, J \subseteq G$, a deterministic score grid $\mathcal{N}_{I,J}$ covers, to coordinatewise precision $1/64$, the restrictions of all such scores for which $I \times J$ avoids the deleted rows and columns. With $v = |I| + |J|$, its size satisfies*

$$\log|\mathcal{N}_{I,J}| \le C_D\ell\left[vK + K^2 + Q + \frac{P_0^2}{h_0}\right]. \tag{G.24}$$

The grid is independent of the training sample and of λ.

Proof. Two elementary cases allow arbitrary budgets in the statement. If $h_0 \ge (1024P_0)^2$, then on every retained rectangle

$$\|U\psi_L(x)\|_\infty \le 2\sqrt{2}P_0/\sqrt{h_0} < 1/256.$$

Quantizing only the reference score to precision $1/1024$ therefore gives the required cover at logarithmic cost $C_D Q\ell$. Otherwise, if $K \ge Q^2$, quantize all $Q|I||J|$ score entries directly. They lie in a polynomially bounded interval, and

$$Q|I||J| \le Q^2(|I| + |J|)/2 \le (|I| + |J|)K.$$

This costs at most $C_D(|I| + |J|)K\ell$. It remains to consider $K < Q^2$ and $h_0 < (1024P_0)^2$.

Each centered response can be written as

$$a_j[\sigma(\beta_j + v_j^\top \psi) - \sigma(\beta_j)], \qquad \sum_j a_j^2 \le 1, \qquad \|v_j\|_2 \le 1.$$

All intermediate response norms are at most $\sqrt{2}$. Replacing β_j by its projection onto $[-2, 2]$ preserves this centered response on all inputs: outside that interval the ReLU is either zero throughout or linear throughout. This replacement parameterizes the same score table in (G.9). It does not change the already constructed network or its output-bias convention.

On $I \times J$, only v first-layer input coordinates vary. The remaining reference contributions can be included in a bounded first-layer constant. The resulting centered feature representation has at most

$C_D(vK + K^2 + 1)$ parameters, all in a fixed bounded interval. After multiplying the last response by $\sqrt{h_0}$, put

$$\Phi(x) = \sqrt{h_0}\psi_L(x), \qquad V = U/\sqrt{h_0}, \qquad U_0 = 2P_0/\sqrt{h_0}.$$

On the retained rectangle, $\|\Phi(x)\|_2 \le \sqrt{2}$ and $\|V\|_F \le U_0$.

Let $q_0 = 1/1024$ and $\tau = q_0/(1 + U_0)$. Write a centered layer as $\sigma(Az + b) - c$. Its normalized matrix has Frobenius norm at most one. Quantize each coefficient with error at most

$$\zeta = \frac{\tau}{8\,2^L(1+K)\sqrt{1+h_0}}.$$

At the first layer, each input uses only two variable coefficients, a bias and a subtracted constant, giving vector error $e_1 \le 4\sqrt{K}\zeta \le 4K\zeta$. At later layers the matrix perturbation has operator norm at most $K\zeta \le 1$. Using the bound $\|z\|_2 \le \sqrt{2}$ for the unquantized response gives

$$e_i \le 2e_{i-1} + 4K\zeta.$$

Consequently $\sqrt{h_0}e_L \le 4(2^L - 1)K\sqrt{h_0}\zeta < \tau$. All budgets and $1/\tau$ are polynomially bounded in Q, with fixed-depth constants. The logarithmic grid size is therefore at most $C_D(vK + K^2 + 1)\ell$. Keep only grid features whose norm is at most two at all inputs in the rectangle; the nearest feature to every admissible Φ is retained.

Fix one such grid feature table f_x. All matrices $\|V\|_F \le U_0$ admit a finite simultaneous output cover. Start with $V' = 0$. Whenever some pair (c, x) satisfies $|(V - V')_{c:}f_x| > q_0$, update

$$V' \leftarrow V' + \frac{q_0}{4}\,\mathrm{sgn}((V - V')_{c:}f_x)e_c f_x^\top.$$

Because $\|f_x\|_2 \le 2$, this decreases $\|V - V'\|_F^2$ by at least $q_0^2/4$. At most $\lceil 4U_0^2/q_0^2 \rceil$ updates are therefore possible. Each update is specified by one class, one rectangle input and one sign. All update strings of this bounded length constitute a fixed finite cover, with logarithmic size at most

$$C\left(1 + \frac{P_0^2}{h_0}\right)\ell.$$

This construction accounts for all output rows through one Frobenius potential.

Finally quantize the Q coordinates of $F(x_0)$ to precision q_0. Their logarithmic cost is $C_D Q\ell$. The errors from feature approximation, output approximation, and constant quantization total at most $3q_0 < 1/64$. Combining the three counts proves (G.24). □

A common sampling event for all rectangles. For each grid score f, let

$$A_f = \{x \in I \times J : \mathrm{gap}_f(x) \le 1/4\}.$$

The hypergeometric lower-tail bound implies

$$|A_f| \le \frac{2}{\alpha}|S \cap A_f| + \frac{C_D\ell}{\alpha}\left[vK + K^2 + Q + \frac{P_0^2}{h_0}\right] \tag{G.25}$$

on an event of probability at least $1 - Q^{-3}$, simultaneously for all these grids and rectangles. For completeness, allocate failure probability

$$\frac{Q^{-3}}{(Q+1)^2\binom{Q}{|I|}\binom{Q}{|J|}\,|\mathcal{N}_{I,J}|}$$

to each grid element. The allocation costs at most $C(v + 1)\ell + \log|\mathcal{N}_{I,J}|$, which (G.24) absorbs. If a fixed set has more than $2|S \cap A_f|/\alpha + 8t/\alpha$ elements, its sample count is below half its mean and its mean exceeds $8t$; the hypergeometric tail is then at most e^{-t}. This proves (G.25) and its stated failure probability.

After selecting the actual H, restrict to rectangles outside both deletion sets. A $1/64$-close grid changes any class difference by at most $1/32$. Hence the gap-at-most-$1/8$ set of H is contained in the grid gap-at-most-$1/4$ set. On that grid set the original score gap is at most

$$1/4 + 1/32 + 1/8 = 13/32 < 1/2.$$

Substituting the original-objective subrectangle comparison into (G.25) therefore gives the claimed local small-gap estimate. All grids use only the common public budgets and energy bounds, so this event applies to sample-selected centers and to every λ in the interval.

G.7 One score approximation and all subrectangles

Put

$$K, h_0 \in \{1, \dots, Q\}, \qquad P_0^2 = CQ^2\ell^D. \tag{G.26}$$

Conditional minimization of the original last layer, its strong convexity, and the rowwise layer approximation of Lemma G.3 produce a common size-$O(K)$ center H. Outside strips comprising at most

$$h_e \le C_D \left[\frac{Q\ell^{2D}}{\alpha_Q \mu^2 K} + \frac{\overline{\Delta} Q^2 \ell^D}{\mu K} + \frac{Q^2 \ell^{D+2}}{K^2} \right] \tag{G.27}$$

rows and columns, all class differences of F and H differ by at most $1/8$. The first two terms follow respectively from the conditional readout row bound and the strong-convexity displacement $\sqrt{2\overline{\Delta}/\lambda}$; the third is the accumulated deterministic layerwise approximation error. At every sampling step, the response norm and the next matrix norm are bounded for each outcome, whereas the squared approximation error is bounded in sampling expectation. Thus one may select the layers successively without losing the conditions needed later.

Delete also the h_0 largest rows and columns and denote the union of all strips by $\mathcal{E}$. On a fixed subrectangle $I \times J \subseteq X \setminus \mathcal{E}$, the same center has a score grid of precision $1/64$ whose logarithmic size is at most

$$C_D \ell \left[vK + K^2 + Q + P_0^2/h_0 \right]. \tag{G.28}$$

The grid and all subrectangles are defined before sampling. Hypergeometric lower tails, followed by a union over these choices, give one event of probability $1 - O(Q^{-3})$ on which every grid score f satisfies

$$|\mathcal{A}_f| \le \frac{2}{\alpha_Q} |S \cap \mathcal{A}_f| + \frac{C_D \ell}{\alpha_Q} \left[vK + K^2 + Q + P_0^2/h_0 \right], \tag{G.29}$$

where $\mathcal{A}_f$ is its gap-at-most-$1/4$ set in that subrectangle.

Let $\mathcal{A} = \{x \notin \mathcal{E} : \mathrm{gap}_H(x) \le 1/8\}$. The $1/64$ grid and the class-difference approximation place $\mathcal{A} \cap (I \times J)$ inside a grid small-gap set whose corresponding original points lie in $\mathcal{A}_F$. Substituting (G.5) into (G.29) yields

$$|\mathcal{A} \cap (I \times J)| \le C_D \left[Q^2 \mu \ell^{1-2/D} (v/Q)^{1/D} + Q^2 (e_k + \overline{\Delta}) + \frac{\ell}{\alpha_Q} \left(vK + K^2 + Q + P_0^2/h_0 \right) \right]. \tag{G.30}$$

This is a structural small-margin count, not a full-table true-error bound.

G.8 Degeneracy and exact recovery

View $\mathcal{A}$ as a bipartite graph between the two operands, and let d_A be its degeneracy, the maximum minimum degree over all nonempty induced subgraphs. If an induced subgraph has minimum degree $d > 0$ and v vertices, it has at least $dv/2$ edges and $v \ge 2d$. Applying (G.30) to its vertex sets and separating the possible dominant terms gives

$$d_A \le C_D \left[\frac{K\ell}{\alpha_Q} + Q(\mu \ell^{1-2/D})^{1/\nu} + Q\sqrt{e_k + \overline{\Delta}} + \sqrt{\frac{\ell}{\alpha_Q} \left(K^2 + Q + P_0^2/h_0 \right)} \right]. \tag{G.31}$$

In particular, the first nonlinear term follows by solving $d \le CQ^{2-1/D} \mu \ell^{1-2/D} d^{-(D-1)/D}$. Thus no small-degeneracy condition is assumed.

For the wide-architecture applications, $k > m$ gives

$$e_k \le C_D \left[\frac{\sqrt{Q}\, \ell^{D/2+1}}{m} + \mu \ell (\ell/m)^{2/D} \right]. \tag{G.32}$$

The finite conditions ensure $m \ge C_D \ell$, so Lemma G.2 applies in that branch.

A graph of degeneracy d_A splits into at most $2d_A$ star forests, each a disjoint union of stars. Each forest represents arbitrary nonnegative values on its edges using three first-layer ReLU factors. The strips use one nonnegative star per row or column. Since every disagreement between B_F and B_H lies

in $\mathcal{E} \cup \mathcal{A}$, encode the original class on these sets, pass two nonnegative correction responses through the later layers, and add the usual quadratic class code at the output. This exactly recovers B_F on all of X, including its errors and tie choices. The construction uses at most $C_D(K + h_e + h_0 + d_A + 1)$ positions in each hidden layer. The applications below bound this quantity from the original objective and verify the original widths. Independent EGC counting is applied only after this exact preservation of every prediction. We next derive the one-hidden-layer estimates; the common finite conditions and the arbitrary-excess result then complete both depth cases.

G.9 The one-hidden-layer sparse estimate

Write the abelian group additively in this subsection. The fiber, atomic, and sampling estimates below apply to both fixed-density and sparse arguments. The finite-width comparator used in the atomic estimates is proved in Section I.3.

Lemma G.5 (Uniform fiber concentration at sparse sampling density)**.** *Let $\alpha = n/Q^2$, $\ell = \log Q$, and suppose $\alpha Q \geq C\ell$. For normalized scalar atoms*

$$h(a, b) = \sigma(u_a + v_b + b_0), \qquad \|u\|_2^2 + \|v\|_2^2 + b_0^2 \leq 1,$$

write

$$n_y = |\{(a, b) \in S : a + b = y\}|, \quad s_y(h) = \sum_a h(a, y - a), \quad \widehat{s}_y(h) = \frac{Q}{n_y} \sum_{(a,y-a)\in S} h(a, y - a).$$

There is an event of probability at least $1 - Q^{-4}$, independent of the network, its width, and λ, on which the row, column and fiber counts have their usual relative concentration bounds, and

$$\sup_h \|\Pi(\widehat{s}(h) - s(h))\|_2 \leq \frac{C\sqrt{Q\ell}}{\alpha}, \qquad \Pi = I - Q^{-1}\mathbf{1}\mathbf{1}^\top. \tag{G.33}$$

Proof. Hypergeometric concentration gives $n_y \in [\alpha Q/2, 2\alpha Q]$ and relative deviations at most $C\sqrt{\ell/(\alpha Q)}$, simultaneously with the analogous row and column bounds. For example, relative threshold $8\sqrt{\ell/(\alpha Q)}$ is at most $1/2$ when $\alpha Q \geq 256\ell$, and the union over the $3Q$ counts has failure probability less than $Q^{-4}/2$.

Condition on the fiber counts. Sampling within different fibers is then independent, with a uniform n_y-element subset in fiber y. First use independent Bernoulli variables $B_{a,y}$ of means $p_y = n_y/Q$, and let $\xi_{a,y} = B_{a,y}/p_y - 1$. Their ranges have lengths at most $2/\alpha$. The matrix $M_{b,y} = \xi_{y-b,y}$ satisfies, for fixed unit vectors,

$$\Pr(|x^\top M z| > t) \leq 2e^{-\alpha^2 t^2/2}.$$

Two $1/4$-nets give $\|M\|_{\mathrm{op}} \leq 40\sqrt{Q\ell}/\alpha$ with failure probability exponentially small in $Q\ell$.

For a unit $z \perp \mathbf{1}$, set

$$F_a(t) = \sum_y z_y \xi_{a,y} \sigma(t + v_{y-a}).$$

The fundamental theorem of calculus gives

$$\left|\sum_a F_a(b_0 + u_a)\right| \leq \left|\sum_a F_a(b_0)\right| + \|u\|_2 \left(\sum_a \sup_t |F_a'(t)|^2\right)^{1/2}. \tag{G.34}$$

For a fixed ordering of the v_b, each derivative is one of at most $Q + 1$ prefix sums. For fixed z, let $W_a = \sup_t |F_a'(t)|^2$. Hoeffding's inequality yields

$$\Pr(W_a > s) \leq 2(Q + 1)e^{-\alpha^2 s/2}, \qquad \mathbb{E}e^{\alpha^2 W_a/4} \leq CQ.$$

The W_a are independent in the Bernoulli model. Their exponential moments, a union over the $Q!$ possible orderings, and a $1/2$-net of the unit z-sphere give

$$\sup_{\text{orderings},\, \|z\|_2=1} \sum_a \sup_t |F_a'(t)|^2 \leq CQ\ell/\alpha^2. \tag{G.35}$$

For each fixed ordering the square root on the left is a seminorm in z, so the net extends to the whole sphere. By choosing fixed numerical constants in these estimates, their total failure probability can be made at most $e^{-10Q\ell}$.

Condition now on each Bernoulli column having sum n_y. Since n_y is a mode of $\mathrm{Bin}(Q, n_y/Q)$, the probability of the conditioning event is at least $(Q+1)^{-Q}$. The preceding exponentially small failures therefore remain below $Q^{-4}/2$ after conditioning, after increasing the constants if needed. This is exactly the required product of uniform fixed-size fiber samples. On this event $\sum_a \xi_{a,y} = 0$. Hence the first term in (G.34) equals

$$\left|(\sigma(b_0+v)-\sigma(b_0)\mathbf{1})^\top Mz\right| \le \|v\|_2 \|M\|_{\mathrm{op}}.$$

Together with (G.35), this proves (G.33). The event is uniform over all normalized atoms; no union over trained networks is used. □

The actual shallow atoms under the original objective. Let $D = 2$, $\mu = \lambda Q$, and let Δ be the excess objective value in the original width-k architecture. For the fixed-density and sparse applications, the original comparison network and their stated objective-gap bounds give

$$R(\theta) \le CQ\ell, \qquad \mathcal{L}_S(F_\theta) \le C\mu\ell. \tag{G.36}$$

We use the finite-width comparison from Appendix I.3 only in the wide branch, where $k \ge C\ell^2$. Recall its auxiliary full-table minimizer, averaged over the group symmetries, and write its common softmax probabilities as p_g. For a probability vector r, let $H(r) = -\sum_g r_g \log r_g$. For a vector-valued table Y, define $\Omega^*(Y) = \sup_h \|\sum_{x\in X} h(x)Y(x)\|_2$, where h ranges over the normalized scalar atoms above. For

$$\epsilon = 1 - p_0, \qquad d_g = \mathbf{1}_{\{g=0\}} - p_g, \qquad Y_c(a,b) = \frac{d_{c-a-b}}{\lambda Q^2},$$

the auxiliary atomic dual and the group-character construction give

$$\epsilon \le C\lambda Q, \qquad \Omega^*(Y) \le 1, \qquad J^*_{\mathcal{A}} \le H(p) + C\lambda Q\ell^2/k. \tag{G.37}$$

The last comparison is valid for every fixed training set: its proof bounds the expected cross-entropy of the sampled comparator at each input before averaging over that set. It does not require a constant sampling fraction.

On the event of Lemma G.5, define

$$\widehat{Y}(x) = \mathbf{1}_S(x)\frac{Q}{n_{B_G(x)}}Y(x), \qquad \tau = C_\tau\frac{\sqrt{\ell/Q}}{\alpha}, \qquad \widetilde{Y} = \frac{\widehat{Y}}{1+\tau}.$$

Take C_τ as a fixed sufficiently large constant. Since $\sum_g d_g = 0$, the convolution pairing only uses the projected fiber error in (G.33). Young's convolution inequality and $\|d\|_1 = 2\epsilon \le C\lambda Q$ give

$$\Omega^*(\widehat{Y} - Y) \le \tau, \qquad \Omega^*(\widetilde{Y}) \le 1.$$

The same τ dominates the relative count error. For $\tau \le 1/8$, increasing C_τ first if necessary, write $y = B_G(x)$ for each $x \in S$ and put

$$a_y = \frac{n}{Qn_y(1+\tau)} \in [1-C\tau, 1], \qquad q_x = (1-a_y)e_y + a_y p^y = e_y - n\lambda\widetilde{Y}(x), \qquad p^y_c = p_{c-y}.$$

The fiber weights have equal total mass, so $\sum_{x\in S}\widetilde{Y}(x) = 0$. Both applications below have sufficiently small μ under their respective finite conditions, so $\epsilon \le 1/4$. Writing r for the conditional distribution of the incorrect classes,

$$H((1-a)e_0 + ap) = H((a\epsilon, 1-a\epsilon)) + a\epsilon H(r)$$

and differentiating in $a \in [1/2, 1]$ show

$$0 \le H(p) - \frac{1}{n}\sum_{x\in S} H(q_x) \le C\tau\epsilon\log(2Q/\epsilon) \le C\lambda Q\tau\ell. \tag{G.38}$$

The last bound uses $\log(1/\mu) = O(\ell)$, which holds for both original polynomial decay ranges used below. Here $x \mapsto x\log(2Q/x)$ is increasing for the relevant small values of x, so its value at $\epsilon \le C\lambda Q$ is bounded by its value at $C\lambda Q$; no lower bound on ϵ is needed.

Normalize the actual finite set of network atoms:

$$F_\theta = b_2 + \sum_j t_j c_j h_j, \qquad \|c_j\|_2 = 1, \qquad M = \sum_j t_j \le R(\theta).$$

Here t_j is the product of the norm of the incoming affine row and the norm of its outgoing column. Define

$$\delta_j = 1 - \langle \widetilde{Y}, c_j h_j \rangle \ge 0, \qquad D_\theta = \sum_j t_j \delta_j.$$

The cross-entropy identity gives exactly

$$J(\theta) - \frac{1}{n}\sum_{x\in S} H(q_x) = \frac{1}{n}\sum_{x\in S} \mathrm{KL}(q_x \| \operatorname{softmax} F_\theta(x)) + \lambda D_\theta + \lambda(R(\theta) - M). \tag{G.39}$$

The output-bias pairing vanishes, and its original quadratic cost is included in the last term. All terms on the right are nonnegative. Equations (G.37)–(G.39) therefore imply

$$D_\theta \le C\left[\frac{\sqrt{Q}\,\ell^{3/2}}{\alpha} + \frac{Q\ell^2}{k} + \frac{\Delta}{\lambda}\right]. \tag{G.40}$$

The coordinate bounds also retain the sampling fraction. Erase an input coordinate of magnitude s from a normalized atom. The change is supported on one row or column. Its vector of weighted fiber sums has norm at most $Cs\sqrt{Q}/\alpha$. Convolution with d, followed by division by $\lambda Q^2(1+\tau)$, changes the dual pairing by at most $Cs/(\alpha\sqrt{Q})$. The erased input has norm at most $\sqrt{1-s^2}$. Dual feasibility consequently implies

$$1 - \delta_j \le \sqrt{1-s^2} + \frac{Cs}{\alpha\sqrt{Q}}, \qquad s^2 \le C\left(\delta_j + \frac{1}{\alpha^2 Q}\right). \tag{G.41}$$

For $w_j = \sum_x h_j(x)\widetilde{Y}(x)$, subtract the constant $\sigma(b_{0,j})$. On every group fiber the squared sum of the remaining input variation is at most two. Its weighted absolute sum is at most $C\sqrt{Q}/\alpha$, giving

$$\|w_j\|_2 \le 1, \qquad \|w_j\|_\infty \le \frac{C}{\alpha\sqrt{Q}}, \qquad \|c_j - w_j\|_2^2 \le 2\delta_j.$$

Together with (G.41), this bounds every input coordinate and every class-coefficient difference by $C\sqrt{\delta_j + 1/(\alpha^2 Q)}$. Thus the actual, unscaled atoms satisfy

$$T := \sum_j t_j\left(\delta_j + \frac{1}{\alpha^2 Q}\right) \le C\left[\frac{\sqrt{Q}\,\ell^{3/2}}{\alpha} + \frac{\ell}{\alpha^2} + \frac{Q\ell^2}{k} + \frac{\Delta Q}{\mu}\right]. \tag{G.42}$$

Lemma G.6 (Shallow approximation with simultaneous norm control)**.** *Let $Q \ge 3$, $\ell = \log Q$, and consider a finite decomposition*

$$F(a,b) = b_2 + \sum_j t_j c_j h_j(a,b), \qquad h_j(a,b) = \sigma(u_{j,a} + v_{j,b} + b_j),$$

where $t_j > 0$, $\|c_j\|_2 = 1$, and the squared norm of each incoming affine row is at most one. Suppose the positive numbers s_j satisfy

$$|u_{j,a}|,\ |v_{j,b}|,\ |c_{j,c} - c_{j,d}| \le C_1\sqrt{s_j}.$$

Set $M = \sum_j t_j$, $T = \sum_j t_j s_j$, and $x_0 = Q^{-1}(\mathbf{1},\mathbf{1})$. There is a constant C, depending only on C_1, such that for every $v > 0$ and integer $K \ge C(v+T)\ell$, one can construct a score table H for which

$$\max_{c,d} |(F_c - F_d) - (H_c - H_d)| \le 1/8$$

outside at most CMT/v rows and columns. The same table has an analytical representation

$$H = F(x_0) + U\psi, \qquad \|U\|_F \le 4M, \qquad \|\psi(a,b)\|_2^2 \le R_a + S_b, \qquad \sum_a R_a + \sum_b S_b \le 2.$$

It is realized by K ordinary ReLUs and an output bias, or by $K+1$ ordinary ReLUs with zero output bias. Thus it belongs to the original single-hidden-layer architecture whenever these positions are available. The empty decomposition is retained exactly.

Proof. Subtract the mean from each of the two input coefficient blocks, obtaining u'_j, v'_j. Their squared norms do not increase and their coordinate bounds increase by at most a factor of two. This is an identity for the original scalar response, with its reference preactivation adjusted accordingly. Put

$$U_a = \sum_j t_j (u'_{j,a})^2, \qquad V_b = \sum_j t_j (v'_{j,b})^2, \qquad \sum_a U_a + \sum_b V_b \le M.$$

For a nonempty decomposition, sample J with probabilities $\pi_j = t_j s_j / T$, and write

$$Z(x) = \frac{t_J}{\pi_J} c_J [h_J(x) - h_J(x_0)].$$

The input and output coordinate bounds imply, for fixed constants B, V,

$$|Z_{cd}(x)| \le BT, \qquad \operatorname{Var} Z_{cd}(x) \le VT(U_a + V_b).$$

In particular, the second moment is bounded by

$$C \sum_j \frac{t_j^2}{\pi_j} s_j ((u'_{j,a})^2 + (v'_{j,b})^2) = CT(U_a + V_b).$$

Delete rows with $U_a > \varepsilon_0 v/T$ and columns with $V_b > \varepsilon_0 v/T$, for a fixed $\varepsilon_0 > 0$. Their number is at most $MT/(\varepsilon_0 v)$. Elsewhere the variance is at most $2V\varepsilon_0 v$, while a centered sample has absolute value at most $2BT$. Bernstein's inequality consequently gives a fixed C_0 with

$$\Pr\left(\left| K^{-1} \sum_{r=1}^{K} Z_{r,cd}(x) - \mathbb{E} Z_{cd}(x) \right| > 1/8 \right) \le 2 \exp\left[-\frac{K}{C_0(v+T)} \right].$$

Taking $C \ge 6C_0$, the union over at most Q^4 input–class-pair choices has probability at most $2Q^{-2} < 1/4$.

For these same draws let

$$w_r = \frac{t_{J_r}}{K \pi_{J_r}}, \qquad M_H = \sum_r w_r.$$

Since $\mathbb{E} M_H = M$, Markov's inequality gives $\Pr(M_H > 4M) \le 1/4$. An outcome therefore exists that satisfies both the uniform approximation and $M_H \le 4M$. For that outcome set

$$H = F(x_0) + \sum_r w_r c_{J_r} [h_{J_r} - h_{J_r}(x_0)], \quad \psi_r = \sqrt{\frac{w_r}{M_H}} [h_{J_r} - h_{J_r}(x_0)], \quad U_{:r} = \sqrt{w_r M_H}\, c_{J_r}.$$

Then $\|U\|_F = M_H$. The ReLU Lipschitz bound gives the stated feature bounds with

$$R_a = \frac{2}{M_H} \sum_r w_r (u'_{J_r,a})^2, \qquad S_b = \frac{2}{M_H} \sum_r w_r (v'_{J_r,b})^2.$$

Their total is at most two.

To realize these scores, retain the sampled original scalar ReLUs and assign outgoing columns $w_r c_{J_r}$. The remaining constant is

$$F(x_0) - \sum_r w_r c_{J_r} h_{J_r}(x_0).$$

Use the original output bias when allowed. Otherwise one constant hidden response $\sigma(1) = 1$ with this output column suffices. No bound on the quadratic penalty of the constructed parameters is required; its analytical readout and feature bounds have just been established. □

One shallow center with both approximation and norm control. Use the actual atoms with $s_j = \delta_j + 1/(\alpha^2 Q)$. Equations (G.41)–(G.42) give the coordinate bounds in Lemma G.6. Put

$$m = Q^{7/8}/\ell, \quad v_0 = \alpha m/\ell^2, \quad K = \lceil C_s \alpha m/\ell \rceil, \quad h_0 = \lceil m \rceil. \tag{G.43}$$

The estimates in the next paragraph give $T \le C_T v_0$ with a fixed constant C_T. If C is the sampling constant in the lemma, choose $C_s \ge C(1 + C_T)$. Then $K \ge C(v_0 + T)\ell$. The same sampled center has class-difference error at most $1/8$ outside at most

$$h_e \le CMT/v_0 \le \frac{CQ\ell^3 T}{\alpha m} \tag{G.44}$$

rows and columns, and it has an analytical representation

$$H = F_\theta(x_0) + U\psi, \qquad \|U\|_F \le 4M \le CQ\ell,$$
$$\|\psi(a,b)\|_2^2 \le R_a + S_b, \qquad \sum_a R_a + \sum_b S_b \le 2.$$

The empty decomposition is retained exactly. In the wide branch $k > C_* m$, choose C_* after C_s. The finite condition $\alpha m/\ell \ge 2$ implies $K + 1 \le (C_s + 1)\alpha m/\ell \le (C_s + 1)m$, so the original width also accommodates the constant response when the output bias is fixed to zero. The narrow branch retains the original implementation. The local score-covering lemma applies to this same center with $P_0^2 = CQ^2\ell^2$ and no intermediate K^2 parameters. Together with the common sampling argument, this gives the shallow instance of the score grid and its sampling event.

The enlarged objective tolerance. In the wide branch $k > C_* m$, divide (G.44) by m and substitute (G.42). The four contributions are bounded by

$$C\left[\frac{Q^{-1/4}\ell^{13/2}}{\alpha^2} + \frac{Q^{-3/4}\ell^6}{\alpha^3} + \frac{Q^{-5/8}\ell^8}{\alpha} + \frac{\Delta Q^{1/4}\ell^5}{\alpha\mu}\right]. \tag{G.45}$$

For $D = 2$, the stated tolerance is $\Delta_{\mathrm{tol}} = \alpha\mu Q^{-1/4}\ell^{-5}$, so the last ratio is at most one. Furthermore (G.44)'s expression satisfies the exact relation

$$\frac{T}{v_0} = \frac{m}{Q\ell}\,\frac{Q\ell^3 T}{\alpha m^2}, \qquad \frac{m}{Q\ell} = Q^{-1/8}\ell^{-2}. \tag{G.46}$$

The finite conditions therefore give $T \le C_T v_0$ with a fixed constant. This verifies the bound used to choose C_s in the common sampling application above. The last term of (G.50), $Q^{-1/8}/\ell^2$, controls T/v_0 through (G.46). The finite-width comparison follows instead from the third term of (G.51). For $D = 2$, that condition gives

$$Q^{1/8} \ge C_D\ell^5, \qquad m/\ell^2 = (Q^{1/8})^7/\ell^3 \ge C_D^7\ell^{32}.$$

Thus $k > C_* m$ implies $k \ge C\ell^2$. The auxiliary requirements $\alpha Q \ge C\ell$, $\tau \le 1/8$, and small μ follow from the same finite conditions after their fixed constants are chosen. In particular, the first term of (G.50) controls $\ell/(\alpha^2 Q)$. No additional width assumption is imposed on the original model.

G.10 Uniform decay interval and finite-size conditions

We state the common finite conditions, then prove a bound for arbitrary objective excess. The uniform interval theorem follows by substituting its objective tolerance. Use the task, architecture, objective, and sampling assumptions of that theorem, and write

$$\ell = \log Q, \quad \alpha = n/Q^2, \quad \nu = 2 - 1/D, \quad a_D = 3 - 3/D, \quad \eta_0 = \nu/8, \quad \eta_1 = \eta_0 + \omega.$$

Thus $0 \le \beta < 1/8$, $0 < \alpha_0 < 1$, $n = \lfloor \alpha_0 Q^{2-\beta} \rfloor$, and $\eta_1 < 3/8 - \beta$. For each admissible λ, put $\mu = Q^{2/D}\lambda$ and let Δ denote excess above the minimum of that same regularized objective in the original architecture. The interval and objective tolerance are

$$Q^{-\eta_1}\ell^{-a_D} \le \mu \le Q^{-\eta_0}\ell^{-a_D}, \qquad \Delta_{\mathrm{tol}} := \alpha\mu Q^{-1/4}\ell^{-(D+3)}. \tag{G.47}$$

Fix the constants in the approximation, shallow sampling, and local score-covering estimates. Choose the fixed constant C_s large enough for the shallow Bernstein estimate, and set

$$m = Q^{7/8}/\ell, \qquad K = \lceil C_s\alpha m/\ell \rceil, \qquad h_0 = \lceil m \rceil. \tag{G.48}$$

In particular, $K, h_0 \in \mathbb{N}^*$. These budgets are independent of λ. If $\alpha m/\ell \ge 2$, then $C_s\alpha m/\ell \le K \le (C_s + 1/2)\alpha m/\ell$.

Finite conditions. The constants denoted by C_D below are fixed sufficiently large, after the preceding constants, and may depend on the fixed depth, sampling constants, width-ratio bound, β, and ω, but not on the absolute widths, λ, the sample, or the parameters. Require $Q \ge 3$, $\alpha m/\ell \ge 2$, and $m \ge C_D\ell$. For $D \ge 3$, require

$$C_D\left[\frac{Q^{-3/4+2\eta_1}\ell^{2D+3+2a_D}}{\alpha^2} + \frac{Q^{-5/8}\ell^{D+7}}{\alpha^2}\right] \le 1. \tag{G.49}$$

For $D = 2$, require

$$C_D \left[\frac{Q^{-1/4}\ell^{13/2}}{\alpha^2} + \frac{Q^{-3/4}\ell^6}{\alpha^3} + \frac{Q^{-5/8}\ell^8}{\alpha} + \frac{Q^{-1/8}}{\ell^2} \right] \le 1. \tag{G.50}$$

For both depth cases, require

$$\begin{aligned} C_D \Big[& Q^{-3+\eta_1}\ell^{a_D-1} + \alpha Q^{-1/4}\ell^{-(D+4)} + Q^{-1/8}\ell^{D/2+4} + Q^{-13/(8D)}\ell^{7/D} \\ & + \alpha Q^{-\eta_0}\ell^{-(a_D+D+1)} + \frac{\alpha}{\ell} + \frac{Q^{-3/4}\ell^3}{\alpha} + \frac{Q^{-5/8}\ell^{D+4}}{\alpha} + Q^{1/8-\eta_0}\ell^{1-a_D} \Big] \le 1. \end{aligned} \tag{G.51}$$

Every term tends to zero in the stated parameter range. The conditions therefore hold beyond a threshold depending only on the fixed quantities specified above. We use their displayed finite form in the proof.

Auxiliary conditions. The condition $\alpha m/\ell \ge 2$ and the term $Q^{-1/8}\ell^{D/2+4}$ in (G.51) imply

$$\alpha Q \ge 2Q^{1/8}\ell^2 \ge 2C_D\ell^{D/2+6}.$$

They ensure the required row, column, and label-fiber concentration conditions. They also imply $m \ge C\ell^2$ for every fixed constant needed in the finite-width comparison, after C_D is fixed large enough. In the shallow case this can be seen from

$$Q^{1/8} \ge C_D\ell^5, \qquad \frac{m}{\ell^2} = \frac{(Q^{1/8})^7}{\ell^3} \ge C_D^7\ell^{32}.$$

Furthermore,

$$\frac{\ell}{\alpha^2 Q} = \frac{Q^{-1/4}\ell^{13/2}}{\alpha^2} Q^{-3/4}\ell^{-11/2},$$

so the shallow concentration parameter $\tau = C_\tau\sqrt{\ell/Q}/\alpha$ is at most $1/8$. Writing $A_Q = Q^{-1/8}\ell^{D/2+4}$, we also have

$$\mu \le Q^{-\nu/8}\ell^{-a_D} = A_Q^\nu \ell^{-a_D-\nu(D/2+4)} \le C_D^{-\nu}.$$

This supplies $\mu \le 1$, including the smaller fixed upper bound needed by the shallow entropy comparison.

The condition $C_D Q^{-1/8}\ell^{D/2+4} \le 1$ and $\ell \ge 1$ also imply $\ell \ge 8\log C_D$. Choosing the fixed C_D sufficiently large therefore supplies every fixed lower bound on ℓ required in Corollary G.7 below. That corollary proves exact compression for arbitrary objective excess; substituting $\Delta \le \Delta_{\mathrm{tol}}$ will prove Theorem 4.2.

G.11 Dependence on optimization accuracy

The same decay interval also gives a continuous dependence of exact compression on the excess objective value. Use the task, model, sampling assumptions, and decay interval of Section G.10. Keep its notation $\ell = \log Q$, $\alpha = n/Q^2$, $\mu = \lambda Q^{2/D}$, ν, a_D, and η_0, η_1. For a parameter vector θ, let $\Delta \ge 0$ denote its excess above the minimum of the same regularized objective in its original architecture.

Corollary G.7 (Optimization accuracy and exact compression)**.** *There is a constant C, depending only on the fixed depth, sampling constants, width-ratio bound, β, and ω, with the following property. Suppose the finite conditions in Section G.10 hold with sufficiently large fixed constants, and $\ell \ge C$. Set*

$$m_0 = \frac{Q^{7/8}}{\ell}, \qquad m_\theta = \min\left\{ Q,\, m_0 + Q\sqrt{\frac{\Delta\ell^{D+1}}{\alpha\mu}} \right\}.$$

For every $\delta \in (0,1)$, on one sample event of probability at least $1 - \delta - Q^{-3} - Q^{-4}$, the following statements hold simultaneously for all allowed original architectures, both output-bias conventions,

all decay strengths in the interval, and all parameter vectors θ. The original complete prediction map has an exactly equivalent realization in the same architecture with joint EGC coordinates

$$t_0 \le 1, \qquad \max_{1\le i<D} t_i \le C m_\theta.$$

Its held-out error satisfies

$$\begin{aligned}\mathrm{err}_{\mathrm{out}}(B_\theta) \le \min\Bigg\{1, \frac{2-\alpha}{1-\alpha}\,\mathrm{err}_{\mathrm{in}}(B_\theta) \\ + \frac{CQm_\theta\ell + 8\log(1/\delta)}{(1-\alpha)n}\Bigg\},\end{aligned} \tag{G.52}$$

and hence

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \le \min\left\{1,\, C\left[\frac{Q^{-1/8}}{\alpha} + \sqrt{\frac{\Delta\ell^{D+3}}{\alpha^3\mu}}\right] + C\frac{\log(1/\delta)}{n}\right\}. \tag{G.53}$$

Proof. Write

$$\widehat{m} = m_0 + Q\sqrt{\frac{\Delta\ell^{D+1}}{\alpha\mu}}, \qquad m = m_\theta.$$

First suppose $\widehat{m} < Q$. Then $m = \widehat{m} \ge m_0$ and

$$\frac{\Delta Q^2\ell^{D+1}}{\alpha\mu m^2} \le 1, \qquad \frac{\Delta}{\mu\ell} \le \frac{\alpha}{\ell^{D+2}}\left(\frac{m}{Q}\right)^2 \le \frac{\alpha}{\ell^{D+2}}. \tag{G.54}$$

Lemma I.2, scaled to margin 4ℓ, has loss at most Q^{-3} on every training set and penalty at most $C_D Q^{2/D}\ell$. Comparison in the original architecture therefore gives $J(\theta) \le C_D\mu\ell + Q^{-3} + \Delta$. The first term of (G.51) and (G.54) imply the common bounds

$$R(\theta) \le C_D Q^{2/D}\ell, \qquad \mathcal{L}_S(F_\theta) \le C_D\mu\ell.$$

The constant in these estimates is independent of the actual Δ and λ within this branch.

The approximate score with a variable budget. Let k be the smallest original hidden width. The case $k \le C_* m$ already has the stated EGC budget by the original width-ratio bound. For $k > C_* m$, set

$$K = \lceil C_s\alpha m/\ell \rceil, \qquad h_0 = \lceil m \rceil.$$

Here C_s is the fixed shallow sampling constant. The original condition $\alpha m_0/\ell \ge 2$ gives

$$C_s\alpha m/\ell \le K \le (C_s + 1/2)\alpha m/\ell.$$

With $\ell \ge 2(C_s+1)$, both K and h_0 are integers between 1 and Q. All bounds below use only $k > m$; C_* is chosen at the end.

For $D \ge 3$, the actual readout comparison is valid for arbitrary excess Δ: its strong-convexity remainder is at most $2\Delta/\lambda$. Thus the row-and-column estimate (G.23), valid under both output-bias conventions, gives

$$\frac{h_e}{m} \le C_D\left[\frac{Q\ell^{2D+1}}{\alpha^2\mu^2m^2} + \frac{\Delta Q^2\ell^{D+1}}{\alpha\mu m^2} + \frac{Q^2\ell^{D+4}}{\alpha^2 m^3}\right].$$

The first and third terms decrease with m and are controlled at m_0 by (G.49). The middle term is at most one by (G.54). Consequently the same approximate score has $h_e \le C_D m$ exceptional strips.

For $D = 2$, the original atom estimate (G.42) gives

$$T \le C\left[\frac{\sqrt{Q}\,\ell^{3/2}}{\alpha} + \frac{\ell}{\alpha^2} + \frac{Q\ell^2}{k} + \frac{\Delta Q}{\mu}\right].$$

Using $k > m$, define and bound

$$A_T := \frac{Q\ell^3 T}{\alpha m^2} \le C\left[\frac{Q^{3/2}\ell^{9/2}}{\alpha^2 m^2} + \frac{Q\ell^4}{\alpha^3 m^2} + \frac{Q^2\ell^5}{\alpha m^3} + \frac{\Delta Q^2\ell^3}{\alpha\mu m^2}\right] \le C_T.$$

The first three terms decrease with m and at m_0 are the first three terms of (G.50); the last term is at most one by (G.54). For $v_0 = \alpha m/\ell^2$ we consequently have

$$\frac{T}{v_0} = \frac{m}{Q\ell} A_T \le \frac{C_T}{\ell} \le 1.$$

Choose $C_s \ge 2C$, where C is the constant in Lemma G.6. Then $K \ge C(v_0 + T)\ell$, so that lemma supplies one center with class-difference error at most $1/8$ outside

$$h_e \le CMT/v_0 \le CQ\ell^3 T/(\alpha m) \le Cm$$

rows and columns, and simultaneously supplies the readout norm and row–column domination needed in the local score cover. It retains the empty decomposition exactly and uses a charged constant position when the original output bias is absent. Thus all hypotheses of the common score-covering argument hold for this same center.

One sample event for all budgets. The common energy bound places reference scores in the same polynomial range as in Lemma G.4. For each pair of integers $1 \le K, h_0 \le Q$ and each nonempty subrectangle $I \times J$, let $\mathcal{N}_{K,h_0,I,J}$ be that lemma's $1/64$ score grid. For $v = |I| + |J|$, it satisfies

$$\log |\mathcal{N}_{K,h_0,I,J}| \le C_D \ell \left[vK + K^2 + Q + \frac{Q^2 \ell^D}{h_0} \right].$$

Assign each grid element failure probability

$$\frac{Q^{-3}}{Q^2(Q+1)^2 \binom{Q}{|I|}\binom{Q}{|J|} |\mathcal{N}_{K,h_0,I,J}|}.$$

Summing over budgets, rectangles, and grid elements gives total failure probability at most Q^{-3}. The additional $2\log Q$ from budget choices is absorbed by the displayed score-grid bound. Hence the original subrectangle sampling inequality holds simultaneously for all these budgets. The center and the deleted strips can then be chosen using the sample. The row, column, label-fiber, and atom events have total failure probability at most Q^{-4}, as before.

The small-gap graph and exact recovery. The original objective comparison on subrectangles retains the actual excess Δ. On the preceding event it yields (G.31), using (G.32), with the present K, h_0, m, Δ. The term $K\ell/\alpha$ is $O(m)$, while the upper interval endpoint gives $Q(\mu\ell^{1-2/D})^{1/\nu} \le m_0 \le m$. After squaring and dividing by m^2, the remaining contributions are bounded by fixed multiples of

$$Q^{-1/8}\ell^{D/2+4}, \quad Q^{-13/(8D)}\ell^{7/D}, \quad \alpha\mu\ell^{-(D+1)}, \quad \frac{\alpha}{\ell}, \quad \frac{Q^{-3/4}\ell^3}{\alpha}, \quad \frac{Q^{-5/8}\ell^{D+4}}{\alpha}.$$

The third bound uses (G.54); all others decrease as m increases from m_0. They are controlled by (G.51), so $d_A \le C_D m$. Lemma I.3 and its recovery construction now preserve all original predictions using $C'_D m$ positions per hidden layer. The affine output rank is at most one. After fixing $C_* > \max\{1, C'_D\}$, this fits the original architecture in the wide case; the narrower case was already covered by the original realization.

Budgets reaching Q. If $\widehat{m} \ge Q$, let $Y(a,b) \in \{0, \dots, Q-1\}$ be the original label table. The Q nonnegative responses

$$h_r(a,b) = \sigma\big(Y(r,b) + (Q-1)\mathbf{1}_{\{a=r\}} - (Q-1)\big) = Y(r,b)\mathbf{1}_{\{a=r\}}$$

sum to $Y(a,b)$. With one further constant response, scores $F_c = 2cY - c^2$ realize the specified table with strict margin at least one. They use the original linear readout for $D = 2$; for $D \ge 3$, the following hidden layers transmit the nonnegative sum and the constant one. This fits when the smallest original width is at least $Q + 1$. Otherwise every original hidden width is at most $C_w Q$, and the original realization gives the desired joint budget. Neither case requires an output bias. This proves exact compression for every Δ.

Independent generalization. For each integer budget $1 \le j \le \lceil CQ \rceil$, the union of its EGC classes over all allowed original architectures has the same reduced-expression count. At the budget above, its logarithm is at most $C_D[Q(m+1)+m^2]\ell \le C_D Qm\ell$. Allocate failure probability $\delta/\lceil CQ \rceil$ to each budget. The resulting $O(\log Q)$ term is absorbed by the same count. After exact recovery, Lemma D.1 yields (G.52). In the nonsaturated case, $\mathrm{err}_{\mathrm{in}}(B_\theta) \le C_D \mu \ell$, which the original finite conditions absorb into $CQ^{-1/8}/\alpha$. Moreover,

$$\frac{\widehat{m}\ell}{\alpha Q} = \frac{Q^{-1/8}}{\alpha} + \sqrt{\frac{\Delta \ell^{D+3}}{\alpha^3 \mu}}.$$

This proves (G.53). If $\widehat{m} \ge Q$, its right-hand side is one after truncation; no small-energy estimate is used in that case. All probability events were chosen before λ, Δ, θ, with total failure probability at most $\delta + Q^{-3} + Q^{-4}$. □

At fixed objective excess and with the remaining parameters fixed, the upper bound in (G.53) is nonincreasing in λ throughout the interval; its excess-dependent term scales as $\lambda^{-1/2}$.

For example, an excess at most $\alpha^3 \mu \ell^{-(D+5)}$ gives error

$$C\left[\frac{Q^{-1/8}}{\alpha} + \frac{1}{\ell}\right] + C\frac{\log(1/\delta)}{n}.$$

Its ratio to the tolerance in (G.47) is $\alpha^2 Q^{1/4}/\ell^2$, which grows without bound for each fixed $\beta < 1/8$. Thus slower error decay permits asymptotically larger objective excess throughout the same decay interval.

Proof of Theorem 4.2. For the abelian case $\rho = 1$, at the theorem's tolerance,

$$Q\sqrt{\frac{\Delta_{\mathrm{tol}} \ell^{D+1}}{\alpha \mu}} = \frac{Q^{7/8}}{\ell} = m_0.$$

Corollary G.7 therefore gives $m_\theta \le 2m_0$, and hence $t_0 \le 1$ and $\max_i t_i \le CQ^{7/8}/\ell$ in the original architecture. In its independently counted error bound,

$$\sqrt{\frac{\Delta_{\mathrm{tol}} \ell^{D+3}}{\alpha^3 \mu}} = \frac{Q^{-1/8}}{\alpha} \le C_{\alpha_0} Q^{\beta - 1/8},$$

where $\alpha = n/Q^2 \ge (\alpha_0/2)Q^{-\beta}$ under the finite conditions. This gives (4.6) on the same event of probability at least $1 - \delta - Q^{-3} - Q^{-4}$, simultaneously over the full decay interval and all original architectures. The tolerance is linear in λ by its definition. For $\rho = 1 + 5/D$, Corollary J.2 gives the stated budget and error bound under (4.5). □

H Structural Counting in the Original Architecture

This appendix proves the structural results in Sections 2.2 and 3.3. Here $X = G^2$, $N = Q^2$, and the original ReLU architecture has $L \ge 1$ hidden layers. Hidden biases are present; the output bias is either present or fixed to zero, as prescribed by the architecture. The fixed priority rule resolves ties. The structural results apply to arbitrary target labels; the fixed-density and sparse optimization arguments are given in Appendices I and G, respectively.

H.1 Exact expressions and the structural definition

Let $\Theta_{\mathcal{A}}$ be the parameter set of this architecture. Subtracting the common logit leaves predictions unchanged. Represent centered logits by their first $q = Q - 1$ coordinates; the final logit is minus their sum. An input-affine function on X has the form $u(a) + v(b) + \beta$, and these functions form a vector space of dimension $d = 2Q - 1$. As in the main text, vector-valued input-affine terms are written directly as $U(a, b)$ or $M(a, b)$. Whenever coefficients or ranks are used, choose any basis of this space; the resulting ranks do not depend on that choice. The original network input remains (e_a, e_b).

Start with an actual $\theta \in \Theta_{\mathcal{A}}$. One may substitute an everywhere-active ReLU by its preactivation, remove zero contributions, combine proportional responses, and substitute an exact relation

$$g_v(a,b) = \ell_v(a,b) + \sum_{w \in I_v} \beta_{vw} g_w(a,b), \qquad (a,b) \in X, \tag{H.1}$$

Here $\ell_v(a,b) = u_v(a) + v_v(b) + \beta_v$ is input-affine, and I_v contains retained responses from earlier original layers, or responses preceding v in a chosen order within its original layer. The relation must hold on all of X. Substitute its right-hand side into every later consumer and then delete g_v. Original units in the same layer never consume g_v, so this operation creates no nonlinear dependence within that layer. It is optional; the expressions range over all allowed choices and orders. The identity $\sigma(x_i + y_i - 1) + \sigma(x_i - y_i) = x_i$ for binary operand indicators illustrates a full-table cancellation that proportionality alone misses.

After finitely many such substitutions and removal of terms with no output effect, the centered score map has the form

$$\begin{aligned} g_i &= \sigma\left(U_i(a,b) + \sum_{j<i} A_{ij} g_j\right), \quad g_i \in \mathbb{R}^{t_i}, \\ F &= M(a,b) + \sum_{i=1}^{L} C_i g_i. \end{aligned} \tag{H.2}$$

Here $F \in \mathbb{R}^q$ gives the first q centered logits, while $U_i : X \to \mathbb{R}^{t_i}$ and $M : X \to \mathbb{R}^q$ are input-affine. In any basis of the input-affine space, their coefficient matrices have sizes $t_i \times d$ and $q \times d$; $A_{ij} \in \mathbb{R}^{t_i \times t_j}$ and $C_i \in \mathbb{R}^{q \times t_i}$. The notation t_0 means the basis-independent rank of the latter coefficient matrix. Predictions use the corresponding Q logits and the fixed tie rule. Every retained g_i is still computed by the original activation from earlier original layers. The displayed direct input and earlier-layer terms result from substitution; the behavior-equivalent network itself must still lie in $\Theta_{\mathcal{A}}$.

The structural coordinates are t_0 and the retained dimensions t_i. For $\mathbf{z} = (\tau_0, \tau_1, \ldots, \tau_L) \in \mathbb{N}_0^{L+1}$, $\mathrm{EGC}_{\mathcal{A}}(B) \preceq \mathbf{z}$ means that some $\theta \in \Theta_{\mathcal{A}}$ with $B_\theta = B$ on X admits one allowed exact reduction (H.2) with

$$t_0 \le \tau_0, \qquad t_i \le \tau_i \quad (1 \le i \le L). \tag{H.3}$$

One expression must attain all inequalities. The coordinatewise minimal attainable vectors form a nonempty set for every realizable behavior. Minimality is taken over these $L+1$ coordinates. This permits tradeoffs without separately minimizing coordinates and combining incompatible realizations. The definition uses neither the sample nor the target labels, loss, regularizer, margins, or numerical precision.

Comparison with Section 2.1. For one hidden layer, let $c(B)$ be the least $t_0 + t_1$ over allowed reductions and all finite widths in the model of that section. A list of m paid factors gives $t_0 + t_1 \le m$: each everywhere-active factor contributes at most one affine rank. Conversely, a reduced realization can be expanded into at most $t_1 + 2t_0$ paid factors, since each affine rank-one summand is the difference of two ReLU gates. Consequently

$$\tfrac{1}{2}\kappa(B) \le c(B) \le \kappa(B). \tag{H.4}$$

This proves Proposition 3.4. The zero-logit default behavior has both values zero. This comparison uses the all-finite-width model of Section 2.

H.2 Layer allocation changes attainable behaviors

The distribution of retained responses across layers can change the behavior class even at a fixed total budget. The following example uses zero affine budget. For $i \in \{0, \ldots, L\}$, let $\mathbf{e}_i \in \mathbb{R}^{L+1}$ be the unit vector in structural coordinate i, with coordinate zero corresponding to the affine rank.

Proposition H.1 (Layer allocation at zero affine budget)**.** *Let $Q \ge 4$, and let $\mathcal{A}$ be a prescribed fully connected ReLU architecture with $L \ge 2$ hidden layers, each of width at least two, and with*

hidden biases. The output bias may be allowed or fixed to zero. For every $1 \le r < s \le L$ *and every* $1 \le j \le L$*,*

$$\mathcal{B}_{\mathcal{A}}(2\mathbf{e}_j) \subsetneq \mathcal{B}_{\mathcal{A}}(\mathbf{e}_r + \mathbf{e}_s). \tag{H.5}$$

All the classes on the left are equal. Thus both sides have zero affine budget and total retained-response budget two.

Proof. First consider any allowed reduced expression with zero affine budget and at most two retained responses, both in one layer. Its class comparisons can be represented by

$$F_c(a,b) = C_{c1}\sigma(u(a,b)) + C_{c2}\sigma(v(a,b)), \qquad c \in G,$$

where $u(a,b) = u_1(a) + u_2(b)$ and $v(a,b) = v_1(a) + v_2(b)$. The zero affine budget removes every remaining class-dependent affine term, including constants. A common logit shift has no effect on predictions. This representation therefore includes all equivalent realizations with this budget, even when the architecture allows an output bias.

Let c_0 be the first class in the fixed tie-breaking order. Where both responses vanish, the prediction is c_0. Elsewhere, divide the scores by their positive common factor $\sigma(u) + \sigma(v)$ and put

$$z = \frac{\sigma(u)}{\sigma(u) + \sigma(v)} \in [0,1].$$

The scores become $C_{c1}z + C_{c2}(1-z)$. The ratios won by each class form an interval, with endpoint inclusion determined by the fixed tie rule. Suppose two distinct classes $c_1, c_2 \neq c_0$ occur in the complete behavior. The winning interval of c_1 cannot contain both endpoints of $[0,1]$: it would then contain the whole interval, leaving no ratio won by c_2.

Only finitely many ratios occur on the input domain. Place strict thresholds between consecutive distinct observed ratios. The inputs predicted as c_1 are consequently described by one of the following pairs, with thresholds in $(0,1)$ and $l < h$ in the last case:

$$\begin{aligned} &v > 0, && hv - (1-h)u > 0; \\ &u > 0, && (1-l)u - lv > 0; \\ &(1-l)u - lv > 0, && hv - (1-h)u > 0. \end{aligned} \tag{H.6}$$

The first two pairs describe intervals containing respectively 0 or 1; the last describes an interior interval. In the last case, multiply the first inequality by h and the second by l and add to obtain $(h-l)u > 0$. Multiplying instead by $1-h$ and $1-l$ gives $(h-l)v > 0$. Thus removing the ReLUs in that case is valid. Every pair excludes points where both responses vanish. Since the thresholds are chosen between actual ratios, this reasoning also covers ties and an interval containing only one observed ratio. The class-c_1 inputs are therefore the intersection of two sets of the form $\{(a,b) : f_1(a) + f_2(b) > 0\}$.

Such an intersection cannot contain all six off-diagonal entries of a 3×3 subtable while excluding all three diagonal entries. Indeed, one strict additive inequality containing all the off-diagonal entries can exclude at most one diagonal entry. If it excluded both (a_i, b_i) and (a_j, b_j), then

$$f_1(a_i) + f_2(b_i) + f_1(a_j) + f_2(b_j) \le 0,$$

whereas inclusion of (a_i, b_j) and (a_j, b_i) would make the same sum strictly positive. Two inequalities can therefore exclude at most two diagonal entries.

We now construct a behavior violating this restriction with two responses in different layers. Choose distinct rows a_1, a_2, a_3, distinct columns b_1, b_2, b_3, and distinct classes c_0, c_1, c_2, with c_0 still first in the fixed tie rule. Define one-operand coefficient tables by

$$w_{b_i} = i, \qquad u_{a_i} = i, \qquad v_{a_i} = -i \quad (i = 1,2,3),$$

and set $w_b = 1$ on the remaining columns and $u_a = 4$, $v_a = -10$ on the remaining rows. Put

$$u(a,b) = u_a - w_b, \qquad v(a,b) = v_a + w_b, \qquad g_1 = \sigma(u), \qquad g_2 = \sigma(v + 2g_1),$$

and take the output scores

$$F_{c_0} = 0, \qquad F_{c_1} = 2g_2, \qquad F_{c_2} = g_1, \qquad F_c = -g_1 - g_2 \quad (c \notin \{c_0, c_1, c_2\}). \tag{H.7}$$

On the selected 3×3 subtable, $g_2 = |i - j|$. Its diagonal is labeled c_0, and every off-diagonal entry is labeled c_1. On all remaining rows, $g_1 = 4 - w_b > 0$ and $v + 2g_1 = -2 - w_b < 0$, so the prediction is c_2. The preceding restriction rules out every single-layer two-response realization of this complete behavior. For any $r < s$, this construction uses two original positions per layer. Choose $C > \max_X |u|$. Before layer r, carry $(u + C, v + 10)$, whose coordinates are positive. At r, subtract C in the first preactivation, producing $(g_1, v + 10)$. At s, use

$$(z_1, z_2) \longmapsto (\sigma(z_1), \sigma(2z_1 + z_2 - 10)) = (g_1, g_2).$$

Every intervening and subsequent layer uses I_2 and zero bias. If $r = 1$, the first layer directly computes $(\sigma(u), v + 10)$ from the displayed one-operand profiles. The output coefficients are exactly those in (H.7), with zero bias; unused positions and connections are zero. The permitted reductions remove the affine transmission $v + 10$ and all nonnegative identity transmissions, leaving precisely the two responses at r, s.

Finally, every expression $C_1\sigma(u) + C_2\sigma(v)$ with additive u, v is realizable with both responses in any prescribed layer j: carry $(u + C_u, v + C_v) > 0$ beforehand, subtract both shifts at j, and transmit the nonnegative pair afterwards. To separate the responses between $r < s$, subtract only C_u at r and only C_v at s; the first layer directly forms the required state if $r = 1$. All later transitions are I_2 with the specified bias subtractions, and the original output columns are C_1, C_2, with zero bias. Hence all concentrated-layer classes are equal and contained in every split-layer class. The preceding counterexample proves strictness. □

A tradeoff with the affine coordinate. The same complete behavior admits an affine contribution of rank one and only one retained response. Set $V = v + 2\sigma(u)$ and use scores

$$F_{c_1} = V, \qquad F_{c_2} = -V, \qquad F_c = 0 \quad (c \notin \{c_1, c_2\}).$$

On the selected rows, $V = |i - w_b|$; on every remaining row, $V = -2 - w_b < 0$. These scores therefore give exactly the preceding prediction table, including its ties.

Use the same two-position transport just constructed, but at layer s replace its transition by

$$(z_1, z_2) \longmapsto (\sigma(2z_1 + z_2 - 10), \sigma(-2z_1 - z_2 + 10)) = (\sigma(V), \sigma(-V)).$$

Transmit this nonnegative pair afterwards. Output rows $(1, -1)$ and $(-1, 1)$ give $V, -V$; all other rows and the output bias are zero. Thus the realization uses only original adjacent-layer connections, including when $r = 1$.

Order the negative response at s before the positive response and substitute the identity

$$\sigma(V) = v + 2\sigma(u) + \sigma(-V).$$

The coefficients of $\sigma(-V)$ then cancel in every output. Only the response at layer r and the input-affine contribution v remain, so for this behavior B,

$$B \in \mathcal{B}_{\mathcal{A}}(\mathbf{e}_0 + \mathbf{e}_r) \cap \mathcal{B}_{\mathcal{A}}(\mathbf{e}_r + \mathbf{e}_s).$$

Both displayed budgets are coordinatewise minimal. With zero affine budget, one nonnegative response can predict at most the default class and one fixed other class, whereas B takes three labels. With no retained responses, every class-score difference has the input-additive form $f(a) + g(b)$, regardless of its affine rank. For the c_1–c_0 difference, the first two diagonal entries would be nonpositive and the two off-diagonal entries strictly positive, because c_0 has first priority. This contradicts

$$[f(a_1) + g(b_1)] + [f(a_2) + g(b_2)] = [f(a_1) + g(b_2)] + [f(a_2) + g(b_1)].$$

Hence neither budget can be decreased in any coordinate without increasing another. This gives two incomparable minimal structural budgets for one complete prediction map. The strict distinction between layer allocations in Proposition H.1 uses zero affine budget.

H.3 Independent counting and generalization

Write $\mathcal{B}_{\mathcal{A}}(\mathbf{z})$ for the complete behaviors satisfying (H.3), and use the same direct counting parameters as in the main text:

$$W_{\mathrm{dir}}(\mathbf{z}) = (d+q)(\tau_0 + \sum_{i=1}^{L} \tau_i) + \sum_{1 \le j < i \le L} \tau_j \tau_i,$$
$$V(\mathbf{z}) = W_{\mathrm{dir}}(\mathbf{z}) + \sum_{i=1}^{L} \tau_i + 1, \tag{H.8}$$
$$\mathscr{H}_Q(\mathbf{z}) = C_L V(\mathbf{z}) \log\big(2 + C_L N(Q^2 + \sum_{i=1}^{L} \tau_i + 1)\big).$$

These quantities are derived bounds, not the definition of EGC.

Lemma H.2 (Finite behavior count)**.** *For a constant depending only on L, $\log |\mathcal{B}_{\mathcal{A}}(\mathbf{z})| \le \mathscr{H}_Q(\mathbf{z})$. The bound is uniform over all original widths. For the score expressions themselves, the same argument counts any fixed thresholds of all class differences and hence their top-two-margin sets.*

Proof. Pad the retained feature counts to τ_i in (H.2) and factor M through τ_0 dimensions. The two factors of M use $(d+q)\tau_0$ entries, the matrices U_i, C_i use $(d+q)\tau_i$ entries, and each A_{ij} uses $\tau_i \tau_j$ entries. Thus W_{dir} parameters cover every allowed expression. Conditional on earlier activation signs, all preactivations and scores are polynomials in these parameters of degree bounded in terms of L. At layer i there are $N\tau_i$ sign tests and at the output at most NQ^2 class comparisons. The polynomial sign bound of Bartlett et al. [6, Lemma 17], applied successively to the fixed number of layers, bounds the logarithm of the number of outcomes by $C_L(W_{\mathrm{dir}} + 1)\log(2 + C_L N(Q^2 + \sum_{i=1}^{L} \tau_i + 1))$. If there are fewer tests than variables, the elementary bound by the number of sign strings suffices. Three-valued signs are covered by adjoining one positive variable ϵ and testing $p + \epsilon$ and $p - \epsilon$: each finite sign table is realized with sufficiently small ϵ. This preserves zero preactivations and fixed winner ties. The choices of deleted relations and original positions need not be enumerated: every resulting expression is already in this continuously parameterized containing class. Fixed comparison thresholds only change polynomial constants. The harmless $\sum_{i=1}^{L} \tau_i$ in (H.8) also covers degenerate budgets. This proves the bound independently of the physical widths. □

For each integer budget $\mathbf{z} = (z_0, \dots, z_L) = (\tau_0, \tau_1, \dots, \tau_L)$ assign the weight

$$\pi_0(\mathbf{z}) = \prod_{j=0}^{L} \frac{1}{(z_j + 1)(z_j + 2)}, \qquad \sum_{\mathbf{z} \in \mathbb{N}_0^{L+1}} \pi_0(\mathbf{z}) = 1. \tag{H.9}$$

Let $e_S(B)$ and $e_X(B)$ count errors on S and X, respectively, relative to the fixed target table. Thus $e_S(B) = n\,\mathrm{err}_{\mathrm{in}}(B)$ and $\mathrm{err}_{\mathrm{out}}(B) = (e_X(B) - e_S(B))/(N - n)$.

Theorem H.3 (Generalization from an EGC budget)**.** *For $0 < \delta < 1$, arbitrary fixed target labels, and a uniform sample of $1 \le n < N$ positions without replacement, with probability at least $1 - \delta$, simultaneously for every budget and every $B \in \mathcal{B}_{\mathcal{A}}(\mathbf{z})$,*

$$\frac{e_X(B)}{N} \le \frac{2e_S(B)}{n} + \frac{8}{n}\left[\mathscr{H}_Q(\mathbf{z}) + \log\frac{1}{\delta\pi_0(\mathbf{z})}\right]. \tag{H.10}$$

The event can also be taken simultaneously over original architectures of the given depth.

Proof. For a fixed error set of size e, the sample count is hypergeometric with mean ne/N. Its multiplicative lower-tail bound gives $\mathbb{P}\{e_S \le ne/(2N)\} \le \exp(-ne/(8N))$. Failure of (H.10) implies this event and $ne/(8N) > \mathscr{H}_Q(\mathbf{z}) + \log(1/(\delta\pi_0(\mathbf{z})))$. Apply Lemma H.2, take a union over the finite behavior class, and then sum the failure probabilities $\delta\pi_0(\mathbf{z})$. The containing class used in that lemma is independent of physical width, so no union over unbounded architecture sizes is necessary. Finally $\mathrm{err}_{\mathrm{out}} \le e_X/(N - n)$ and $\log(1/\pi_0(\mathbf{z})) \le 2(L+1)\log(2 + \|\mathbf{z}\|_1)$ give (3.7) at fixed density. □

The counting argument has no coefficient bound or margin assumption. Margins will be used below to construct a behavior-equivalent expression, but are not needed when this theorem is applied to its EGC budget.

I Exact Compression and Generalization at Fixed Sampling Density

For finite abelian groups, a further range of regularization exponents is available. Use the architecture and objective of Section 4.1 with $D \geq 2$, $\min_i k_i \geq C_D \log Q$, the same width-ratio bound, and fixed sampling density $\alpha \in (0,1)$.

Theorem I.1 (Exact compression and generalization throughout a near-optimal region)**.** *Fix $0 < \eta < 1/2$, $0 < u < \min\{\eta, 1/2 - \eta\}$, $0 < \delta < 1$ and set $\lambda = Q^{-(2/D+\eta)}$. For every $Q \geq 3$ satisfying* (I.4)*, with probability at least $1 - \delta - 4Q^{-3}$, simultaneously for every allowed architecture $\mathcal{A}$ and every $\theta \in \Theta_{\mathcal{A}}$ with*

$$\Delta(\theta) \leq \Delta_Q := Q^{-\eta-2u}(\log Q)^{1-D} \tag{I.1}$$

has an attainable EGC budget $\mathbf{z}_\theta$ in its original architecture satisfying

$$V(\mathbf{z}_\theta) \leq C' Q^{2-u/2}, \qquad \mathscr{H}_Q(\mathbf{z}_\theta) \leq C' Q^{2-u/2} \log Q, \qquad \mathrm{err}_{\mathrm{in}}(B_\theta) \leq C' Q^{-\eta} \log Q. \tag{I.2}$$

Consequently,

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq C Q^{-u/2} \log Q + C_\alpha \frac{\log(1/\delta)}{Q^2}. \tag{I.3}$$

The constants $C, C' > 0$ depend only on $D, \alpha, \eta, u, C_{\mathrm{w}}$, and are independent of the hidden widths.

We prove Theorem I.1 for every affine depth $D = L + 1 \geq 2$. The common estimates in Sections G.6 and G.9 bound prediction disagreement here, followed by exact recovery and independent EGC generalization. Write $x = (a, b) \in X = G^2$, $N = Q^2$, $n = \lfloor \alpha N \rfloor$, and $\ell = \log Q$; $\mathbb{E}_X$ denotes the uniform average on X. For any score map F, its correct-class margin is $\Gamma_F(a,b) = F_{ab}(a,b) - \max_{c \neq ab} F_c(a,b)$. For constructed networks, $R(F)$ and $J_S(F)$ mean the penalty and sampled objective evaluated at the displayed parameters. The architecture has hidden biases; its output bias is either penalized in the stated objective or fixed to zero, as specified before optimization. All constructions respect that choice. Constants may depend on D, α, η, u and the stated width-ratio bound C_{w}. They are independent of the absolute widths. The original parameters need not be balanced, stationary, or minimum-norm representations of their scores.

For the finite statement, take $\zeta = (\eta - u)/2$ and suppose $Q \geq 3$ satisfies

$$\begin{aligned} C_{\mathrm{fin}} \Big[& Q^{-2u}\ell^{-D} + Q^{-(\eta-u)/2}\ell + Q^{-u} + Q^{-1} + Q^{-1/2}\ell^{1/2} + Q^{u-1}\ell^D \\ & + \begin{cases} Q^{2\eta+2u-1}\ell^{3D-2} + Q^{(3u-1)/2}\ell^{3D/2}, & D \geq 3, \\ Q^{-\eta} + Q^{u-1}\ell^3 + Q^{-1/2+2u}\ell^{5/2} & \\ \qquad + Q^{-1+3u}\ell^4 + Q^{-1+2u}\ell^2, & D = 2 \end{cases} \Big] \leq 1. \end{aligned} \tag{I.4}$$

Here C_{fin} is a fixed constant depending only on $D, \alpha, \eta, u, C_{\mathrm{w}}$. The inequality collects the deterministic estimates used below. The common fiber-concentration lemma supplies the sampling probability bound.

I.1 Comparison, energy, and margins

Lemma I.2 (A comparison network at every depth)**.** *For a finite abelian group, every architecture with $D \geq 2$ and $k_i \geq C_D \ell$ contains a network with zero output bias, full-table margin at least 3ℓ, and*

$$\mathcal{L}_S(F_G) \leq Q^{-2}, \qquad R(F_G) \leq C_D Q^{2/D} \ell. \tag{I.5}$$

Proof. Choose $m = C \log Q$ characters independently and uniformly. Orthogonality, scalar concentration, and a union over fewer than Q^2 pairs give a deterministic choice for which $\phi(a) = m^{-1/2}(\mathrm{Re}\,\chi_j(a), \mathrm{Im}\,\chi_j(a))_{j=1}^m$ has norm one and $|\langle \phi(a), \phi(b) \rangle| \leq 1/4$ for $a \neq b$. Character values are one-operand coefficient profiles. The identity $\chi_j(ab) = \chi_j(a)\chi_j(b)$ computes each product coordinate from a constant number of real products on $[-1,1]^2$.

A piecewise-linear interpolation of t^2 on $[-2,2]$ with mesh h has error at most $h^2/4$ and is an affine function plus ReLU hinges. Substitution into $xy = ((x+y)^2 - (x-y)^2)/4$ approximates each product to any fixed accuracy with a constant number of factors. Choose the accuracy so that $\|\widetilde{\phi}(a,b) - \phi(ab)\| \le 1/8$. Then $\langle \phi(c), \widetilde{\phi}(a,b)\rangle$ has margin at least $1/2$. Affine and constant terms are implemented by positive shifted factors or positive-negative pairs. The resulting one-hidden-layer network has $s = O(m)$ nonnegative responses and a linear readout.

Each factor has incoming norm $O(\sqrt{Q})$, including its bias, and output-column norm $O(\sqrt{Q}/m)$. Their product sum is $O(Q)$, so individual positive rescalings make both block Frobenius norms $O(\sqrt{Q})$. Insert $D-2$ identity maps on the nonnegative responses, each of norm $\sqrt{s}$. The product of all block norms is $O_D(Qm^{(D-2)/2})$. Scale the output to margin 3ℓ and balance the D block scales. Positive homogeneity, with zero intermediate biases, preserves the scores and gives energy at most $C_D\{Qm^{(D-2)/2}\ell\}^{2/D} \le C_D Q^{2/D}\ell$. For $D = 2$ no identity map is inserted. Padding by zero positions fits the given widths, and $(Q-1)e^{-3\ell} \le Q^{-2}$ bounds the loss on every input. □

Let $\Delta = J(\theta) - J^*_{\mathcal{A}}$, $\epsilon_\Delta = \sqrt{\Delta Q^\eta \ell^{D-1}}$, and $b_\Delta = Q^{-u} + \epsilon_\Delta$. The theorem assumes $\Delta \le \Delta_Q$ from (I.1), so $\epsilon_\Delta \le Q^{-u}$ and $b_\Delta \le 2Q^{-u}$. Comparing in the same architecture gives

$$R(\theta) \le E_0 = C_D Q^{2/D}\ell, \qquad \mathcal{L}_S(F_\theta) \le C_D Q^{-\eta}\ell, \qquad \prod_{i=1}^{D} \|W_i\|_F \le C_D Q \ell^{D/2}. \tag{I.6}$$

Indeed $\Delta/\lambda \le Q^{2/D}\ell^{1-D}$, and the product bound follows by the arithmetic-geometric mean inequality. Set $\gamma = \zeta\ell$ and $f = F_\theta/\gamma$. This positive rescaling preserves every prediction. If Γ_f denotes the correct-class margin, then

$$\frac{|\{x \in S : \Gamma_f(x) < 1\}|}{n} \le 2e^\gamma \mathcal{L}_S(F_\theta) \le CQ^{-u}. \tag{I.7}$$

The last inequality follows from (I.4). Lemma H.2 counts both behaviors and fixed top-two-margin sets of networks using at most $K+1$ positions in each layer. For $K \le Q$ their logarithmic count is bounded by

$$\Xi_Q(K) = C_D[Q(K+1) + (K+1)^2]\ell. \tag{I.8}$$

We next construct such a network H inside the original architecture, with class-difference error at most $1/8$ outside a set E, and $|E|, \Xi_Q(K) \le CNb_\Delta$. These estimates are derived from the objective; they are not assumptions about the trained parameters.

I.2 The readout estimate with multiple hidden layers

Fixed-density specialization of the readout estimate. Let $n = \lfloor \alpha Q^2 \rfloor$ for the fixed nominal density $0 < \alpha < 1$, and put $\widehat{\alpha} = n/Q^2$. The readout estimate in Section G.6 gives on the same row, column, and label-fiber event

$$s_F \le C_D \left[\frac{E_0}{Q} + \frac{E_0^{D-1}\log Q}{\widehat{\alpha}\lambda^2 Q^3} + \frac{\Delta}{\lambda} \right].$$

When $\alpha Q^2 \ge 2$, one has $\widehat{\alpha} \ge \alpha/2$, so this becomes

$$s_F \le C_{D,\alpha} \left[\frac{E_0}{Q} + \frac{E_0^{D-1}\log Q}{\lambda^2 Q^3} + \frac{\Delta}{\lambda} \right]. \tag{I.9}$$

This uses the original readout in either output-bias convention. Write $\gamma = \zeta\ell$, with $\zeta = (\eta - u)/2$ as above. The layer-product bound and the original readout estimate imply

$$P^2 \le CQ^2\ell^D, \qquad T^2 \le C\left[Q\ell^D + Q^{1+2\eta}\ell^{2D-1} + \Delta Q^{2+\eta}\ell^{D-1}\right].$$

Apply Lemma I.4 with failure probability Q^{-3}. Using $\log(1+e^{-\gamma}) \ge e^{-\gamma}/2$ and $\sqrt{x+y+z} \le \sqrt{x} + \sqrt{y} + \sqrt{z}$, it gives an exact attainable budget $t_0 \le 1$, $t_i \le m$ with

$$\begin{aligned} \frac{m^2}{Q^2} \le C\Bigg[& Q^{-(\eta+u)/2}\ell + \frac{\ell}{Q} + Q^{-1/2}\ell^{D/2} + Q^{\eta-1/2}\ell^{D-1/2} \\ & + \sqrt{\Delta Q^\eta \ell^{D-1}} + Q^{-1/3}\ell^{(D+2)/3} + \frac{\ell}{Q^2} \Bigg]. \end{aligned} \tag{I.10}$$

Each term is at most CQ^{-u} under the original finite conditions. Indeed the loss term uses $Q^{-(\eta-u)/2}\ell \le C$ and the gap term uses $\Delta \le \Delta_Q$. For the norm terms, put

$$A_Q = Q^{2\eta+2u-1}\ell^{3D-2}, \qquad B_Q = Q^{(3u-1)/2}\ell^{3D/2}.$$

After multiplication by Q^u, the two terms involving T are bounded by $\sqrt{A_Q}$, since $D \ge 3$, $\eta > 0$, and $\ell \ge 1$. The remaining norm term is bounded by $B_Q^{2/3}$ because $(D+2)/3 \le D$. The two residual logarithmic terms use $Q^{u-1}\ell^D \le C$. Thus $m \le CQ^{1-u/2}$ and

$$V(1, m, \dots, m) \le C_D[Q(m+1) + m^2] \le CQ^{2-u/2}, \qquad \mathscr{H}_Q(1, m, \dots, m) \le CQ^{2-u/2}\ell.$$

This proves the exact-compression part for $D \ge 3$ with the original parameter range and tolerance. The final independent EGC bound below then applies together with (I.6).

I.3 A FINITE-WIDTH COMPARATOR FOR ONE HIDDEN LAYER

For this subsection $D = 2$ and $k = k_1$ is the single original hidden width; the same notation is used in the following fiber-concentration argument. For $w = (u, v, \beta) \in \mathbb{R}^{2Q+1}$, set $h_w(a, b) = \sigma(u_a + v_b + \beta)$. A normalized atom is ch_w with $\|w\|_2 \le 1$ and $c \in \mathbb{R}^Q$, $\|c\|_2 = 1$. Let $\Omega(F)$ be the minimum total nonnegative mass of atoms representing the full score table. The atom set is compact, and the ReLU $\sigma(\mathbf{1}_{a=a_0} + \mathbf{1}_{b=b_0} - 1)$ isolates an input. Thus the atoms span the finite table space, their symmetric convex hull defines a norm, and Carathéodory's theorem gives an attaining finite decomposition. Positive homogeneity balances each factor to quadratic cost $\|w\|\|c\|$. Consequently Ω equals the infimum of the original penalty over all finite widths without a separate output bias. This equivalence is used for an auxiliary full-table comparison, not to replace the fixed-width training problem. For a vector-valued table $Y : X \to \mathbb{R}^Q$, the dual norm uses the unnormalized pairing, summing over all $x \in X$:

$$\Omega^*(Y) = \sup_{\|w\| \le 1} \left\| \sum_x h_w(x) Y(x) \right\|_2 . \tag{I.11}$$

Write the abelian group additively. For $Z_{*,c}(a, b) = \mathbf{1}_{c=a+b} - Q^{-1}$, character orthogonality and the exact identity

$$xy = \frac{1}{2}\int_0^2 [\sigma(x+y-t) + \sigma(-x-y-t) - \sigma(x-y-t) - \sigma(-x+y-t)]\, dt \tag{I.12}$$

give $\Omega(Z_*) \le 32Q$. Indeed $Z_{*,c} = Q^{-1}\sum_{\chi \ne 1} \mathrm{Re}[\chi(a)\chi(b)\overline{\chi(c)}]$. Each real product has atomic mass at most $8Q$: an incoming norm is at most $2\sqrt{Q}$, an output norm at most $\sqrt{Q}$, and the four integrals have total coefficient mass four. Expanding the complex product uses four real products. Compactness converts the integral to a finite atom representation without increasing mass.

Minimize $\mathcal{L}_X(F) + \lambda\Omega(F)$ and average a minimizer over translations $(a, b, c) \mapsto (a+s, b+t, c+s+t)$ and over its common logit. This gives a centered minimizer $\bar{F}_c(a, b) = f_0(c - a - b)$. Put $p_g = e^{f_0(g)} / \sum_h e^{f_0(h)}$, $\epsilon = 1 - p_0$, and $d_g = \mathbf{1}_{g=0} - p_g$. The directional derivative along Z_* and full-table optimality imply

$$0 < \epsilon \le 32\lambda Q, \qquad Y_c(a, b) = \frac{d_{c-a-b}}{\lambda Q^2}, \qquad \Omega^*(Y) \le 1, \qquad \langle Y, \bar{F}\rangle = \Omega(\bar{F}) =: \bar{M}. \tag{I.13}$$

For a probability vector r, write $H(r) = -\sum_g r_g \log r_g$ for its Shannon entropy. The full-table optimum is $H(p)$ by the entropy identity. The comparison tZ_*, with $t = O(\ell)$, gives $\bar{M} \le CQ\ell$. For $D = 2$ and $\lambda = Q^{-(1+\eta)}$, condition (I.4) gives $\epsilon \le 1/4$.

Every positive-mass atom in a norm-attaining decomposition of $\bar{F}$ saturates the pairing with Y. Removing an input coefficient u_a of magnitude s leaves incoming norm at most $\sqrt{1-s^2}$. Each class occurs once on that row, and $\|d\|_1 = 2\epsilon \le 64\lambda Q$, so the change in the output pairing has norm at most $Cs/\sqrt{Q}$. Saturation gives $1 \le \sqrt{1-s^2} + Cs/\sqrt{Q}$, whence $|u_a| \le C/\sqrt{Q}$; the column bound is identical. On a label fiber, $\sum(h - \sigma(\beta))^2 \le 2$. Convolution with d then bounds $\|\sum_x h(x)Y(x)\|_\infty \le C/\sqrt{Q}$. Saturation and dual feasibility imply that this vector equals c, so

$\|c\|_\infty \le C/\sqrt{Q}$. Finally Y annihilates every input-affine table. A saturated atom cannot be everywhere affine or zero, and therefore $|\beta| \le \|u\|_\infty + \|v\|_\infty$. In particular

$$\|h\|_\infty \le C/\sqrt{Q}, \qquad \|ch\|_\infty \le C/Q.$$

Sample k atoms with probabilities $t_j/\bar{M}$, and set $F^{(k)} = (\bar{M}/k)\sum_{r=1}^{k} c_{J_r} h_{J_r}$. Scaling each side of each factor by $\sqrt{\bar{M}/k}$ gives exactly $R(F^{(k)}) = \bar{M}$ in every outcome, using k original positions and zero output bias. Each coordinate of $\bar{M} c_J h_J$ is bounded by $B = C\ell$. Hoeffding's exponential-moment bound yields

$$\mathbb{E}\exp\{(H_c^{(k)} - \bar{F}_c) - (H_y^{(k)} - \bar{F}_y)\} \le e^{2B^2/k}.$$

Apply Jensen's inequality to the logarithm in cross-entropy. Since the total incorrect softmax mass of $\bar{F}$ is ϵ at every input, the expected excess loss is at most $\log[1 + \epsilon(e^{2B^2/k} - 1)] \le C\epsilon\ell^2/k$ when $k \ge C\ell^2$. Averaging over any given S proves the existence of a comparison network in the original width with

$$J_S(F^{(k)}) \le H(p) + C\lambda Q\ell^2/k. \tag{I.14}$$

This comparison is valid with either original output-bias convention.

I.4 Uniform fiber concentration and the one-hidden-layer approximation

Application at fixed sampling density. Let $\widehat{\alpha} = n/Q^2$ be the actual sampling fraction. The finite conditions give $\widehat{\alpha} \ge \alpha/2$, $\widehat{\alpha}Q \ge C\ell$, and $C_\tau\sqrt{\ell/Q}/\widehat{\alpha} \le 1/8$, where α is the fixed nominal density in Theorem 4.1. Apply Lemma G.5 with $\widehat{\alpha}$ in place of its sampling fraction. Put

$$\mu = Q^{-\eta}, \qquad \gamma = \zeta\ell, \qquad \epsilon_\Delta^2 = \Delta Q^\eta \ell, \qquad b_\Delta = Q^{-u} + \epsilon_\Delta.$$

Set $f = F_\theta/\gamma$. If $k \le A_0 Q b_\Delta/\ell$, retain f itself and take $E = \varnothing$. Equation (I.8) already gives the required count. In the remaining branch, $k > A_0 Q b_\Delta/\ell$, condition (I.4) implies $k \ge C\ell^2$. The finite-width comparator and the actual-atom estimates (G.36)–(G.42) therefore apply, with sampling fraction $\widehat{\alpha}$. The energy estimate is (I.6); the finite conditions ensure that μ is sufficiently small, and $\log(1/\mu) = \eta\ell$. Thus the common entropy estimate applies. Only the original width-k objective is used to obtain the atom bounds. On their common event, of probability at least $1 - Q^{-4}$, the factors $1/\widehat{\alpha}$ and $1/\widehat{\alpha}^2$ are absorbed into constants depending on α.

For f, both atomic masses are divided by γ; the s_j and sampling probabilities are unchanged. At fixed density the estimates become

$$M_f \le C_\alpha Q, \qquad T_f \le C_\alpha\left[1 + \sqrt{Q\ell} + \frac{Q\ell}{k} + \frac{Q\epsilon_\Delta^2}{\ell^2}\right].$$

Set $v = Q b_\Delta/\ell^2$ and apply Lemma G.6 to f. The original finite conditions bound the ratios of the four terms in the display to v, respectively, by fixed multiples of

$$Q^{u-1}\ell^2, \qquad Q^{-1/2+u}\ell^{5/2}, \qquad A_0^{-1}Q^{-1+2u}\ell^4, \qquad b_\Delta.$$

They are controlled by the stated shallow finite conditions and $\epsilon_\Delta \le Q^{-u}$. Taking $K = \lceil C(v+T_f)\ell\rceil$ therefore gives

$$K + 1 \le C_\alpha Q b_\Delta/\ell, \qquad h_e \le C_\alpha\left[\frac{\sqrt{Q}\,\ell^{5/2}}{b_\Delta} + \frac{\ell^2}{b_\Delta} + \frac{Q\ell^3}{k b_\Delta} + \frac{Q\epsilon_\Delta^2}{b_\Delta}\right] \le C_\alpha Q b_\Delta. \tag{I.15}$$

The four normalized terms are controlled by $Q^{-1/2+2u}\ell^{5/2}$, $Q^{-1+2u}\ell^2$, $A_0^{-1}Q^{-1+3u}\ell^4$, and $\epsilon_\Delta^2/b_\Delta^2 \le 1$. First use $A_0 \ge 1$ to bound the displayed A_0^{-1} terms by one, so the constants in T_f/v and K are independent of A_0. Choose A_0 larger than the resulting position constant. Then $K + 1 \le k$, including the original zero-output-bias convention. The finite condition $C_{\mathrm{fin}}Q^{u-1}\ell^3 \le 1$ also gives $Q b_\Delta/\ell \ge C_{\mathrm{fin}}\ell^2$, which absorbs rounding and the constant hidden response. The removed rows and columns contain at most $Q h_e$ inputs. In both width branches, $|E|, \Xi_Q(K) \le C N b_\Delta$, as required for the common prediction-disagreement argument.

I.5 Prediction disagreement for the one-hidden-layer case

The following disagreement and recovery steps finish the remaining case $D = 2$. The multi-hidden-layer budget was obtained directly from Lemma I.4.

The preceding constructions cover all $D \geq 2$. Let $A_H = \{x : \mathrm{gap}_H(x) \leq 1/2\}$, where gap is the difference between the largest two scores. For every $K \leq Q$, count all such sets using (I.8). The hypergeometric lower tail $\Pr(|A \cap S| \leq n|A|/(2N)) \leq e^{-n|A|/(8N)}$ gives, on one event of probability at least $1 - Q^{-3}$,

$$|A_H| \leq \frac{2N}{n}|A_H \cap S| + \frac{8N}{n}[\Xi_Q(K) + C_D \log Q]$$

simultaneously for every approximation. Outside E, a training margin of f at least one gives a margin of H at least $7/8$; thus $A_H \cap S \subseteq \{\Gamma_f < 1\} \cup E$. Equations (I.7) and the proved sizes of E and the family imply $|A_H| \leq CNb_\Delta$. Outside E, differing predictions force $\mathrm{gap}_H \leq 1/8$, also when the original winner is selected by a tie rule. Consequently

$$r := |\{B_\theta \neq B_H\}| \leq |E| + |A_H| \leq CNb_\Delta. \tag{I.16}$$

This argument uses no full-table true-error bound. The parameter-dependent exceptional set needs no separate concentration event because its full size has already been bounded.

The event covers all parameters and allowed widths simultaneously, because the containing score families do not depend on physical width. We next recover the original complete behavior exactly before applying the independent EGC theorem.

I.6 Common exact recovery and the remaining widths

Lemma I.3 (Scalar tables on a sparse support)**.** *Two real tables supported on the same r input pairs admit representations with at most $10\lceil\sqrt{r}\rceil$ first-layer ReLUs and two merged input-affine remainders. If both tables are nonnegative, at most $8\lceil\sqrt{r}\rceil$ first-layer ReLUs suffice, with no affine remainder. For $r = 0$, no term is needed.*

Proof. Assume $r \geq 1$ and put $d = \lceil\sqrt{r}\rceil$. Remove all rows containing more than d support points. Their number h satisfies $h(d + 1) \leq r$, hence $h \leq d$. Order the remaining support points within each row. Putting each row's j-th point in group j partitions the remaining support into at most d groups, each with at most one point in every row. Column degrees need not be bounded.

For a nonnegative table, one removed row is represented by

$$y_b \mathbf{1}_{\{a=a_0\}} = \sigma\big(y_b + M(\mathbf{1}_{\{a=a_0\}} - 1)\big), \qquad M > \max_b y_b.$$

For a group with at most one point in each row, choose distinct integer column codes $v_b \in \{0, \ldots, Q-1\}$. On a nonempty row, let u_a equal its selected column code and let y_a be its table value. On an empty row put $u_a = -1, y_a = 0$. Then $z = u_a - v_b$ vanishes exactly on the group and has $|z| \geq 1$ elsewhere. For $M > \max_a y_a$,

$$y_a \mathbf{1}_{\{z=0\}} = \sigma(Mz + y_a) + \sigma(Mz - y_a) - 2\sigma(Mz). \tag{I.17}$$

At $z = 0$ this equals y_a; for $z \geq 1$ the linear terms cancel, and for $z \leq -1$ all terms vanish. Every preactivation is affine in the separately encoded operands. Thus each nonnegative table uses at most $h + 3d \leq 4d$ ReLUs.

For a signed table, a removed row instead has representation

$$y_b \mathbf{1}_{\{a=a_0\}} = \sigma\big(y_b + c + M(\mathbf{1}_{\{a=a_0\}} - 1)\big) - c\mathbf{1}_{\{a=a_0\}},$$

where $c > \max_b |y_b|$ and $M > c + \max_b |y_b|$. Its final term is input-affine. For the same one-point-per-row groups, with $M > \max_a |y_a|$, use

$$\begin{aligned} y_a \mathbf{1}_{\{z=0\}} = {} & \sigma(2Mz + M + y_a) - \sigma(2Mz - M + y_a) \\ & - \sigma(2Mz + M) + \sigma(2Mz - M). \end{aligned}$$

The cases $z = 0, z \geq 1, z \leq -1$ verify the identity directly. Each signed table therefore uses at most $h + 4d \leq 5d$ ReLUs and one merged input-affine remainder. Applying the constructions to both tables proves the claims. □

Recovery from strips and a graph of bounded degeneracy. The nonnegative formulas in Lemma I.3 apply with rows and columns interchanged. A graph of degeneracy d_A can be oriented along a degeneracy ordering so that every vertex has outdegree at most d_A. Group edges by the side containing their tail and by their outgoing-edge number. There are at most $2d_A$ groups, each with degree at most one on one side. Consequently two nonnegative tables supported on s row or column strips and this graph use at most $2s + 12d_A$ first-layer ReLUs. Assign overlaps to a single strip or group before specifying the table values.

Preserving the original complete prediction map. Let $D_0 = \{B_\theta \neq B_H\}$, encode classes as $0, \dots, Q-1$, and form

$$u = B_\theta \mathbf{1}_{D_0}, \qquad w = \mathbf{1}_{D_0}.$$

Both tables are nonnegative. Either support bound above supplies their first-layer ReLU representations. With one hidden layer, connect them directly to the original linear readout. With at least two hidden layers, use two positions in the second layer to form u, w and two positions per later layer to transmit them. Their nonnegativity makes these ReLUs exact identities on the entire input domain. Retain the center H and set unused cross-connections to zero. The correction has no input-affine remainder or output-bias term.

Choose $A > 1 + \max_{x,c,d} |H_c(x) - H_d(x)|$ and set

$$\widetilde{F}_c = H_c + A(2cu - c^2 w). \tag{I.18}$$

Outside D_0, the scores are unchanged. On D_0, if $b = B_\theta(x)$, then for every $c \neq b$,

$$\widetilde{F}_b - \widetilde{F}_c = H_b - H_c + A(b-c)^2 > 0.$$

Thus every original prediction is restored, including incorrect predictions and the prescribed winner at a tie. In the reduced expression, the stable nonnegative transmissions can be substituted into subsequent layers, so new nonlinear responses occur only in the first original layer. The input-affine output rank does not increase. The existing wide- and narrow-architecture arguments apply with these position counts.

For one allowed expression of this network, substitute the stable transmissions and merge the affine terms. At most $10\lceil\sqrt{r}\rceil$ new nonlinear factors arise in the first original layer, none in later layers, and the affine output rank does not increase. The direct count therefore gives

$$V(\mathbf{z}_\theta) \le C_D[Q(K + \sqrt{r} + 1) + (K+1)^2 + K\sqrt{r}] \le CQ^2\sqrt{b_\Delta}. \tag{I.19}$$

Here $K \le CQb_\Delta/\ell$ and $r \le CQ^2 b_\Delta$; for b_Δ bounded away from zero the same estimates hold after increasing the constant. The recovery did not first replace the predictions by true labels.

Let $h_\Delta = K + 10\lceil\sqrt{r}\rceil + C \le CQ\sqrt{b_\Delta}$. If every original width is at least h_Δ, the construction fits. Otherwise, a hidden-width ratio bounded by C_{w} makes all original widths less than $C_{\mathrm{w}} h_\Delta$. Use the original realization itself, with budget $(1, k_1, \dots, k_L)$, whose direct count is at most $C_D[Q(1 + \sum_i k_i) + \sum_{j<i} k_j k_i] \le CQ^2\sqrt{b_\Delta}$. Thus both cases obey (I.19); for one hidden layer the ratio is automatically one. All selected coordinates are polynomially bounded in Q, giving

$$\mathscr{H}_Q(\mathbf{z}_\theta) \le CQ^2\sqrt{b_\Delta}\ell, \qquad \log(1/\pi_0(\mathbf{z}_\theta)) = O_D(\ell). \tag{I.20}$$

Only now apply the independent EGC theorem to the recovered behavior. The training bound in (I.6) gives held-out error at most $C\sqrt{b_\Delta}\ell + C_\alpha \log(1/\delta)/Q^2$. This rate follows from the exact EGC budget of the original behavior. The fiber event, the common small-gap event, and the independent EGC event have total failure probability at most $3Q^{-3} + Q^{-3} + \delta$. Substituting $\Delta \le Q^{-\eta-2u}\ell^{1-D}$ gives $b_\Delta \le 2Q^{-u}$ and all asserted near-optimal rates, including global minimizers at $\Delta = 0$.

I.7 Energy and exact behavior complexity

The proof below uses the common layerwise approximation, finite counting, and supported-table construction. It requires no optimization assumption.

Use the architecture and sampling conventions of Section 4.1.

We first connect parameter norms and training loss to exact EGC compression. The following lemma uses products of layer norms to control approximation error and training cross-entropy to control the frequency of small prediction margins.

Lemma I.4 (Exact EGC bounds from parameter norms and training loss)**.** *Assume $D \geq 3$ and $n \geq 1$. Let $P = \prod_{i=1}^{D} \|W_i\|_F$ and $T = (\prod_{i<D} \|W_i\|_F) \max_c \|W_{D,c:}\|_2$, where $W_{D,c:}$ is the output-weight row corresponding to class c. For $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every parameter θ in every allowed architecture $\mathcal{A}$ and every margin threshold $\gamma > 0$, the complete prediction map B_θ admits an attainable EGC budget $(t_0, \ldots, t_L)$ satisfying $t_0 \leq 1$ and $\max_{1 \leq i < D} t_i \leq m \in \mathbb{N}$, where*

$$m \leq C \min\left\{Q, \left[\frac{Q^2 \mathcal{L}_S(F_\theta)}{\log(1+e^{-\gamma})} + \log\frac{1}{\delta} + Q\log Q\left(1 + \frac{T}{\gamma} + \left(\frac{P^2 \log Q}{\gamma^2}\right)^{1/3}\right)\right]^{1/2}\right\}. \tag{I.21}$$

Here C depends only on $D, \alpha, C_{\mathrm{w}}$. All predictions are preserved within the original architecture.

We write $P_\theta = P$ and $T_\theta = T$. Appendix I.7 provides the proof, which applies to any fixed labeling of X. The norm quantities P_θ and T_θ control the structural cost of approximation, while training cross-entropy bounds the fraction of observed inputs whose true-label logit exceeds the largest competing logit by at most γ. These controls yield an exact realization of the complete prediction map with the stated EGC budget.

We prove Lemma I.4 under the same original-architecture and fixed-density sampling conventions.

Proof. For an integer $1 \leq K \leq Q$, apply Lemma G.3 to F/γ. It supplies an actual center H, using at most $K+1$ positions per layer, whose class-difference error e satisfies

$$\mathbb{E}_X e^2 \leq V_K := \frac{C_D}{Q\gamma^2}\left(\frac{T^2\ell}{K} + \frac{P^2\ell^2}{K^2}\right).$$

The positive output rescaling respects either original output-bias convention. Put $E = \{e > 1/4\}$ and $A_H = \{\mathrm{gap}_H \leq 1/2\}$. Lemma H.2 counts the sets A_H by $\Xi_Q(K)$ in (I.8). The hypergeometric lower tail and a union over $K = 1, \ldots, Q$ give one event of probability at least $1-\delta$ on which

$$|A_H| \leq \frac{2Q^2}{n}|A_H \cap S| + \frac{8Q^2}{n}\left[\Xi_Q(K) + \log\frac{Q}{\delta}\right]$$

for every center. Their unrestricted real-parameter containing class already includes every rescaled center, so the event covers all $\gamma > 0$.

Let Γ_F be the correct-class margin for the fixed labels. Cross-entropy gives

$$\frac{|\{\Gamma_F < \gamma\} \cap S|}{n} \leq \frac{\mathcal{L}_S(F)}{\log(1+e^{-\gamma})}.$$

Outside E, $\Gamma_F \geq \gamma$ implies a correct-class margin of at least $3/4$ for H, and $\mathrm{gap}_H > 1/2$ implies $B_F = B_H$. Thus $A_H \cap S \subseteq \{\Gamma_F < \gamma\} \cup E$ and $\{B_F \neq B_H\} \subseteq A_H \cup E$. Put $E_K = \max\{\mathbb{E}_X e^2, n^{-1}\sum_{x \in S} e(x)^2\}$. Markov's inequality gives $|E| \leq 16Q^2 E_K$ and $|E \cap S|/n \leq 16 E_K$. The preceding estimates yield

$$r := |\{B_F \neq B_H\}| \leq C\left[\frac{Q^2\mathcal{L}_S(F)}{\log(1+e^{-\gamma})} + Q^2 E_K + \frac{Q^2}{n}\left(\Xi_Q(K) + \log\frac{Q}{\delta}\right)\right].$$

This inequality counts disagreement with the approximate center, without using the true error of the original behavior.

Lemma I.3 and (I.18) restore the original predictions using at most $h = K + 3 + 8\lceil\sqrt{r}\rceil$ positions per layer. If all original widths are at least h, embed that construction; otherwise the width ratio bounds every original width by $C_{\mathrm{w}}h$ and the original realization suffices. Retaining the actual responses in either realization leaves at most a constant input-affine output, so $t_0 \leq 1$. Hence $m \leq Ch$ gives the general estimate

$$m^2 \leq C\left[K^2 + \frac{Q^2\mathcal{L}_S(F)}{\log(1+e^{-\gamma})} + Q^2 E_K + \frac{Q^2}{n}\left(\Xi_Q(K) + \log\frac{Q}{\delta}\right)\right]. \tag{I.22}$$

All steps through this estimate hold for any fixed $1 \leq n < Q^2$; its constant depends only on D, C_{w}, and the event covers all $1 \leq K \leq Q$ and all original widths. At fixed sampling density, $E_K \leq (Q^2/n)V_K \leq 2V_K/\alpha$. Using $K^2 \leq QK$ now gives

$$m^2 \leq C\left[\frac{Q^2\mathcal{L}_S(F)}{\log(1+e^{-\gamma})} + Q\ell\left(K + 1 + \frac{T^2}{\gamma^2 K} + \frac{P^2\ell}{\gamma^2 K^2}\right) + \log\frac{Q}{\delta}\right].$$

The same width argument, using only $K \le Q$ and $r \le Q^2$, always gives $m \le CQ$. Put $q_* = \max\{1, T/\gamma, (P^2\ell/\gamma^2)^{1/3}\}$. When $q_* \le Q$, choose $K = \lceil q_* \rceil$ and absorb $\log Q$ into $Q\ell$. When $q_* > Q$, the energy term in (I.21) is at least a constant times Q^2, and use the CQ bound. This proves the claim. □

The distinction between P and T is substantive: $T^2/P^2 = \max_c \|W_{D,c:}\|_2^2/\|W_D\|_F^2$ when $P > 0$. The original-objective readout estimate (I.9) controls the maximum row contribution in addition to the total energy. Training cross-entropy supplies the complementary control of small correct-class margins.

I.8 What a spectral margin bound gives from the energy estimate

We compare with the covering theorem of Bartlett et al. [5, Theorem 3.3], rather than directly applying their i.i.d. generalization theorem to our sample without replacement. The calculation below treats biases, keeps the additional norm factor, and derives the required finite-population margin bound. Its conclusion concerns this particular use of the coarse energy estimate, not all norm-based analyses.

Biases and deterministic norm bounds. Fix an architecture and a deterministic energy cap $E_0 = C_D Q^{2/D}\ell$, where $\ell = \log Q$. Let $\mathcal{F}$ be the class of score maps with $R(\theta) \le E_0$ in this architecture and $\mathcal{F}_X = \{(F(x))_{x\in X} : F \in \mathcal{F}\}$. Write $\mathcal{N}(\mathcal{H}, \varepsilon, d)$ for the least number of radius-ε balls in metric d covering $\mathcal{H}$. For each such parameter, augment the input by a constant coordinate: $z_0 = (h_0, 1)$. Its norm is $\sqrt{3}$. For $i < D$ use

$$\widetilde{W}_i = \begin{pmatrix} W_i & b_i \\ 0 & 1 \end{pmatrix}, \qquad \widetilde{W}_D = (W_D\ b_D).$$

Coordinatewise ReLU leaves the appended coordinate equal to one; the final activation is the identity. This is only a way to apply a norm theorem to the original scores, not a change to the architectures allowed by EGC. It also covers $b_D = 0$. Set $r_i = k_i + 1$ for $i < D$, $r_D = Q$, $W = \max\{2Q + 1, r_1, \dots, r_D\}$, and $T = (2E_0 + 1)^{1/2}$. Then

$$\|\widetilde{W}_i\|_{\mathrm{op}} \le \|\widetilde{W}_i\|_F \le T, \qquad \|\widetilde{W}_i^\top\|_{2,1} \le \sqrt{r_i}\,T.$$

Here $\|A^\top\|_{2,1}$ is the sum of the Euclidean norms of the rows of A. The last inequality is Cauchy–Schwarz. These caps depend only on the architecture and E_0, not on the sampled training set.

With zero reference matrices, unit activation Lipschitz constants, spectral caps $s_i = T$, and group-norm caps $a_i = \sqrt{r_i}T$, the cited covering theorem gives, on the full input table,

$$\log \mathcal{N}(\mathcal{F}_X, \varepsilon, \|\cdot\|_F) \le \frac{3N\log(2W^2)}{\varepsilon^2}\,\mathcal{R}^2, \qquad \mathcal{R} = T^D\left(\sum_{i=1}^{D} r_i^{1/3}\right)^{3/2}, \qquad N = Q^2. \quad \text{(I.23)}$$

This retains the correction factor $(\sum_i (a_i/s_i)^{2/3})^{3/2}$ of that theorem. In particular it is not a product-of-Frobenius-norms bound with that factor silently omitted.

A margin bound for a uniform subset. For $\gamma > 0$, let $q_\gamma(F, x)$ equal one when the correct-class margin is nonpositive, zero when it is at least γ, and interpolate linearly in between. This loss belongs to $[0, 1]$ and is $2/\gamma$-Lipschitz as a function of the score vector in Euclidean norm. Write $\mathcal{G} = \{(q_\gamma(F, x))_{x\in X} : F \in \mathcal{F}\}$ and $d_N(g, h) = N^{-1/2}\|g - h\|_2$. Equation (I.23) implies

$$\log \mathcal{N}(\mathcal{G}, \varepsilon, d_N) \le \frac{12\mathcal{R}^2\log(2W^2)}{\gamma^2\varepsilon^2}. \quad \text{(I.24)}$$

We give the sampling step explicitly. Let B_x be independent Bernoulli variables of mean $p = n/N$, where $n = \lfloor \alpha N \rfloor$, and define

$$Z_g = \frac{1}{n}\sum_{x\in X}(B_x - p)g_x.$$

Hoeffding's exponential-moment inequality bounds increments by

$$\mathbb{E}\exp\{t(Z_g - Z_h)\} \le \exp\left\{\frac{t^2 N}{8n^2}d_N(g, h)^2\right\}.$$

Take nets at distances 2^{-j}, $1 \le j \le J = \lceil \log_2 N \rceil$, and telescope from one fixed $g_0 \in \mathcal{G}$. The exponential-moment bound and the elementary maximum bound for finitely many subgaussian variables show that the expected supremum of level-j increments is at most

$$\frac{C_\alpha}{\sqrt{N}}\, 2^{-j} \sqrt{\log \mathcal{N}(\mathcal{G}, 2^{-j}, d_N) + \log \mathcal{N}(\mathcal{G}, 2^{-(j-1)}, d_N)} \le \frac{C_\alpha \mathcal{R}\sqrt{\log(2W^2)}}{\gamma\sqrt{N}}.$$

One may use net centers in $\mathcal{G}$ by doubling their radii. The anchor has expected absolute value at most $C_\alpha/\sqrt{N}$. The last approximation error is deterministically at most $\|B - p\|_2 \|g - g_J\|_2/n \le C_\alpha/N$. Summing the J levels therefore gives

$$\mathbb{E} \sup_{g \in \mathcal{G}} |Z_g| \le C_\alpha \left[\frac{\mathcal{R}\sqrt{\log(2W^2)}\log(2N)}{\gamma\sqrt{N}} + \frac{1}{\sqrt{N}} \right].$$

Changing one B_x changes this supremum by at most $1/n$. The bounded-differences inequality adds $C_\alpha\sqrt{\log(1/\delta')/N}$ with failure probability δ'. Finally, Stirling's formula gives $\Pr(\sum_x B_x = n) \ge c_\alpha/\sqrt{N}$. Conditional on this event, the selected set is exactly a uniform n-subset and $Z_g = \overline{g}_S - \overline{g}_X$. Choose $\delta' = c_\alpha\delta/\sqrt{N}$ and divide the unconditional failure probability by the conditioning probability. Using $\overline{g}_X = p\overline{g}_S + (1-p)\overline{g}_{X\setminus S}$ proves, with probability at least $1-\delta$, simultaneously for this norm-bounded class,

$$\mathrm{err}_{\mathrm{out}}(B_F) \le \frac{1}{n}\sum_{x \in S} q_\gamma(F, x) + C_\alpha \left[\frac{\mathcal{R}\sqrt{\log(2W^2)}\log(2N)}{\gamma\sqrt{N}} + \sqrt{\frac{\log(2N/\delta)}{N}} \right]. \tag{I.25}$$

The constant absorbs $1/(1-p)$, which is bounded for fixed $\alpha \in (0,1)$. No i.i.d. approximation to sampling without replacement was used.

Substitution of the objective comparison. For $\gamma = \zeta\ell$ as in Appendix I, the empirical margin loss is at most $2e^\gamma \mathcal{L}_S(F) = o(Q^{-u})$. The energy comparison supplies the chosen deterministic cap $T^D = \Theta_D(Q\ell^{D/2})$. Substitution in the complexity term of (I.25) yields the upper bound

$$C_{D,\alpha,\zeta} \left(\sum_{i=1}^{D} r_i^{1/3} \right)^{3/2} \ell^{D/2-1} \sqrt{\log(2W^2)}\log(2Q^2). \tag{I.26}$$

The energy-only substitution stays at logarithmic scale even before the width-dependent correction factor is included, whereas our theorem obtains polynomial decay from the task-dependent label-fiber estimate below. This comparison concerns the stated energy estimate within this particular bound.

Our additional information is the label-fiber estimate for the conditional readout minimizer and its transfer to the actual readout by the objective gap. With s_F as in (I.9), it gives

$$s_F \le C E_0 \left[Q^{-1} + Q^{2\eta-1}\ell^{D-1} + \frac{\Delta Q^\eta}{\ell} \right].$$

At $\Delta \le \Delta_Q$, all three terms in brackets tend to zero. Lemma G.3 uses $P_*^2 s_F$, rather than only the coarse product cap, in the leading sampling variance, and controls the maximum output difference at logarithmic class cost. The shallow case obtains its stronger coefficient estimates from the sample-dual identity. These estimates supply the task-dependent information absent from the energy-only substitution above.

J Exact compression beyond abelian groups

The output-row estimate and exact recovery argument use the fact that each group-label fiber contains each operand once. They therefore apply to noncommutative groups as well. The additional step is to construct a target network with controlled quadratic penalty in the prescribed architecture. We give one construction for finite affine actions with a fixed quotient in each block, then apply the existing compression and independent generalization estimates.

J.1 Finite affine actions and the resulting guarantee

Let G be a finite group of order Q. For fixed $r_* \in \{0, 1, \dots\}$ and $C_* \geq 1$, consider

$$A = \prod_{\nu=1}^{b} V_\nu, \qquad b \in \mathbb{N}^*, \qquad V_\nu = (\mathbb{Z}/q_\nu\mathbb{Z})^{d_\nu}, \qquad q_\nu \in \{2, 3, \dots\}, \quad d_\nu \in \mathbb{N}^*.$$

In each block, let U_ν be the span of at most r_* vectors from a basis of V_ν; this is a free direct summand. Identical pairs (V_ν, U_ν) are numbered copies with fixed identifications. Assume that G acts by affine maps

$$\Phi_g(x) = c_g + \Pi_g\big(\bigoplus_\nu \Lambda_{g,\nu}\big)x, \qquad \Phi_{gh} = \Phi_g \circ \Phi_h, \quad \Phi_e = I, \tag{J.1}$$

where $c_g \in A$, Π_g only permutes identical copies, and each $\Lambda_{g,\nu}$ is an automorphism satisfying

$$(\Lambda_{g,\nu} - I)V_\nu \subseteq U_\nu. \tag{J.2}$$

Thus the induced linear action on V_ν/U_ν is the identity. In a basis adapted to U_ν, its matrix is $\left(\begin{smallmatrix} B & C \\ 0 & I \end{smallmatrix}\right)$: it permits changes within the specified summand and shears into that summand. The number of blocks and their quotient dimensions may grow. The block permutation includes no further internal automorphism.

Require a point $x_0 \in A$ with trivial stabilizer, meaning that $\Phi_g(x_0) = x_0$ implies $g = e$, and require

$$|A| \leq Q^{C_*}. \tag{J.3}$$

Its orbit $x_g = \Phi_g(x_0)$ then consists of Q distinct points. The size condition bounds the number of coordinate bits by a fixed multiple of the information needed to distinguish these elements. The rank condition separately specifies how the linear action can change those coordinates.

Choose compatible adapted bases and expand A into cyclic coordinates $A = \prod_{i=1}^{m} \mathbb{Z}/q_i\mathbb{Z}$. Write $\ell_i = \lceil \log_2 q_i \rceil$ and $\Phi_g(x) = c_g + P_g M_g x$. The matrices M_g have a common support $E \subseteq \{1, \dots, m\}^2$ consisting of the diagonals and the rows corresponding to the summands U_ν. Hence

$$\mathsf{L} := \sum_i \ell_i + \sum_{(i,j)\in E} \ell_j \leq (r_* + 2) \sum_\nu d_\nu \lceil \log_2 q_\nu \rceil \leq 2C_*(r_* + 2) \log_2 Q. \tag{J.4}$$

This support bound follows from the fixed-quotient action.

Theorem J.1 (WD exact compression for finite affine group actions)**.** *Fix an affine depth $D \in \{3, 4, \dots\}$, $C_{\mathrm{w}} \geq 1$, $0 < \alpha_0 < 1$, $0 \leq \beta < 1/16$, and $0 < \delta < 1$. Let G satisfy* (J.1)*–*(J.3) *with fixed r_*, C_*. Use separate one-hot inputs, the original fully connected ReLU architecture, and the objective* (4.1)*, with either original output-bias convention. There is a constant $c_0 > 0$, depending only on D, r_*, C_*, such that the conclusion holds for all hidden widths satisfying*

$$\min_i k_i \geq c_0 \log Q, \qquad \max_i k_i / \min_i k_i \leq C_{\mathrm{w}}.$$

Set $n = \lfloor \alpha_0 Q^{2-\beta} \rfloor$, and assume

$$Q \geq 3, \qquad \frac{n}{Q} \geq 16 \log(6Q^5), \qquad (\log Q)^{D+7} \leq Q^{3/128}. \tag{J.5}$$

For a uniformly sampled n-element training set without replacement, with probability at least $1 - \delta - Q^{-3} - Q^{-4}$, simultaneously for every such architecture and

$$\lambda = Q^{-2/D - 1/(32D) - \eta}, \qquad \frac{3}{16} \leq \eta \leq \frac{9}{32} - \frac{\beta}{2}, \tag{J.6}$$

every original parameter with $\Delta(\theta) \leq Q^{-3/4}$ admits an exact realization of its complete prediction map in the same architecture with EGC budget

$$t_0 \leq 1, \qquad \max_{1 \leq i < D} t_i \leq CQ^{15/16}. \tag{J.7}$$

Independent EGC counting then gives

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq CQ^{\beta - 1/16} \log Q + C\frac{\log(1/\delta)}{n}. \tag{J.8}$$

Here Δ is measured from the minimum of the same original objective at the displayed λ. The constant $C > 0$ may depend on $D, r_*, C_*, C_{\mathrm{w}}, \alpha_0, \beta$, but not on Q, δ, λ or the absolute hidden widths.

The proof below constructs the target using two hidden layers. The orbit coordinates describe that construction; the learning inputs remain the two original one-hot vectors.

J.2 A TARGET NETWORK WITH CONTROLLED QUADRATIC PENALTY

We first separate the finite orbit, then compute its separating features from the two operands. Put $\ell = \log Q$ and fix $\xi = 1/256$. All constants in this subsection depend only on D, r_*, C_*.

Separating the orbit. Define

$$\psi(t) = \operatorname{dist}(t, \mathbb{Z}), \qquad \varphi(t) = (\psi(t), \psi(t + 1/4)).$$

For $d(t, s) = \operatorname{dist}(t - s, \mathbb{Z})$,

$$\|\varphi(t) - \varphi(s)\|_\infty \geq d(t, s)/2. \tag{J.9}$$

Indeed, $(\varphi_1 + \varphi_2, \varphi_1 - \varphi_2)$ traverses the boundary of a square of side $1/2$, at speed 2. Its chord distance in ℓ_1 is at least half the shorter boundary distance: equality of the two distances holds for points on the same or adjacent sides, and opposite sides have chord distance at least $1/2$ while the shorter boundary distance is at most 1. Finally $|\Delta(\varphi_1 + \varphi_2)| + |\Delta(\varphi_1 - \varphi_2)| = 2\|\Delta\varphi\|_\infty$.

Set $\varepsilon = Q^{-\xi/4}/16$ and $r_0 = \lceil 12/\xi \rceil$. In each cyclic coordinate, keep the multiplier 1, and sample r_0 additional multipliers independently and uniformly modulo q_i. Concatenate the corresponding values of $\varphi(\kappa x_{g,i}/q_i)$ into $y_g \in [0, 1/2]^N$, $N = 2(r_0 + 1)m$. For distinct g, h, choose a differing coordinate and let h_i be the additive order of $x_{g,i} - x_{h,i}$. If the multiplier 1 does not already separate the two values by ε, (J.9) gives $h_i^{-1} < 2\varepsilon$. A uniform multiplier is uniform on a cyclic subgroup of order h_i, so its circular difference has probability at most $4\varepsilon + h_i^{-1} < 6\varepsilon$ of lying within 2ε of zero. Thus failure for some pair has probability at most

$$\frac{Q^2}{2}(6\varepsilon)^{r_0} \leq \frac{1}{2}(3/8)^{r_0} Q^{-1} < 1.$$

Fix a choice with $\|y_g - y_h\|_\infty \geq \varepsilon$ for all $g \neq h$. This argument also covers composite moduli.

Set $r_1 = \lceil 12/\xi \rceil$ and $\zeta = Q^{-\xi/4}/8$. Choose the entries of $V \in [0, 1]^{r_1 \times N}$ independently and uniformly. Conditional on all but one differing coordinate, each row has probability at most 2ζ of giving a difference smaller than $\varepsilon\zeta$. The union bound over pairs is at most $(Q^2/2)(2\zeta)^{r_1} < 1$. Hence a fixed V gives $z_g = V y_g$ with

$$0 \leq z_{g,t} \leq N/2, \qquad \|z_g - z_h\|_2^2 \geq \mu_Q := Q^{-\xi}/16384 \quad (g \neq h). \tag{J.10}$$

The existence choices here precede training-set sampling.

Computing the product features in two hidden layers. Index the columns of V by (i, k, s), where $0 \leq k \leq r_0$, $\kappa_{i0} = 1$, and $s \in \{0, 1/4\}$. Then

$$z_{g,t} = \sum_{i,k,s} V_{t,i,k,s} \psi(\kappa_{ik} x_{g,i}/q_i + s).$$

Write $P_a e_h = e_{\pi_a(h)}$; this permutation preserves moduli. From the action law,

$$x_{ab,\pi_a(h)} = c_{a,\pi_a(h)} + \sum_{j:(h,j)\in E} (M_a)_{hj} x_{b,j} \pmod{q_h}.$$

Fix t, h, k, s and put

$$\begin{aligned} w(a) &= V_{t,\pi_a(h),k,s}, \\ d(a) &= \frac{[\kappa_{\pi_a(h),k} c_{a,\pi_a(h)}]_{q_h}}{q_h} + s, \\ v_{jl}(a) &= \frac{[\kappa_{\pi_a(h),k} (M_a)_{hj} 2^l]_{q_h}}{q_h}, \qquad 0 \leq l < \ell_j, \end{aligned} \tag{J.11}$$

where $[u]_q \in \{0, \ldots, q - 1\}$ denotes the residue. Let $d_l(x_{b,j}) \in \{0, 1\}$ be its l-th binary digit. The first hidden layer computes

$$\sigma(w(a)), \quad \sigma(w(a)d(a)), \quad \sigma(w(a)v_{jl}(a) + d_l(x_{b,j}) - 1) = w(a)v_{jl}(a)d_l(x_{b,j}). \tag{J.12}$$

Every function of a in these preactivations is a coefficient of the first one-hot block, and each digit of b is a coefficient of the second block. Their entries and biases are bounded by absolute constants, since $0 \le w \le 1$, $0 \le v_{jl} < 1$, and $0 \le d < 5/4$.

Set

$$s_h = \sum_{j:(h,j)\in E} \ell_j, \quad H_h = s_h + 2, \quad T_h = d(a) + \sum_{j,l} v_{jl}(a) d_l(x_{b,j}).$$

Then $0 \le T_h < H_h$ and

$$T_h \equiv \kappa_{\pi_a(h),k} x_{ab,\pi_a(h)}/q_h + s \pmod 1.$$

The sum of the latter two types of first-layer responses is wT_h. The second layer forms $\sigma(wT_h - jw/2)$, $0 \le j \le 2H_h$. It uses the first response to supply $jw/2$, with zero bias. For $0 \le t \le H_h$,

$$\psi(t) = \sigma(t) + 2\sum_{j=1}^{2H_h} (-1)^j \sigma(t - j/2).$$

Positive homogeneity therefore reconstructs $w\psi(T_h)$, also when $w = 0$. Summing over h, k, s gives exactly $z_{ab,t}$. The number of breakpoints depends on the number of coordinate bits, rather than the numerical values of the moduli.

Include one constant response $\sigma(1)$ in the first layer and pass it through the later layers. The output scores are

$$F_c^0(a,b) = 2z_c^\mathsf{T} z_{ab} - \|z_c\|_2^2. \tag{J.13}$$

The correct-class margin is at least μ_Q. At $D = 3$ the linear output directly combines the second-layer responses. At $D \ge 4$, the third hidden layer forms the nonnegative z_{ab} and the constant response; subsequent hidden layers pass them unchanged. All biases after the first layer are zero, and no input-to-output connection is added.

Width, biases, and quadratic penalty. Equation (J.4) gives $\sum_h (s_h + 2) \le 2\mathsf{L} \le C\ell$. Since r_0, r_1 are fixed, every layer uses at most $C\ell$ positions and embeds by zero padding in the original widths. Write $A_j = \|W_j\|_F^2 + \|b_j\|_2^2$ for this construction. The first-layer entries just described give $A_1 \le CQ\ell$. For each fixed t, h, k, s, there are $2H_h + 1$ second-layer rows, with at most $s_h + 1$ unit coefficients and a coefficient of size at most H_h on w. Thus their squared norm sum is at most CH_h^3. Later aggregation uses coefficients $1, \pm 2$ and identity transmissions. The output coefficients in (J.13) are $O(\ell)$, with a constant-column coefficient $O(\ell^2)$ for each class. Consequently

$$\begin{gathered} A_1 \le CQ\ell, \qquad A_2 \le C\ell^3, \qquad A_j \le C\ell \ (3 \le j < D), \qquad A_D \le CQ\ell^4, \\ \prod_{j=1}^{D} A_j \le CQ^2\ell^{D+5}. \end{gathered} \tag{J.14}$$

The output bound also holds at $D = 3$, when the second-layer responses are combined directly.

Scale the layers by positive numbers whose product is $t = 3\ell/\mu_Q$, scaling the first bias with its layer. Choose these numbers so that all scaled A_j are equal. The resulting target network F_G has margin at least 3ℓ, cross-entropy at most $\log(1 + (Q-1)Q^{-3}) \le Q^{-2}$ on every training set, and

$$R(F_G) = \frac{D}{2}\left(t^2 \prod_j A_j\right)^{1/D} \le CQ^{2/D+1/(128D)}\ell^{(D+7)/D} \le CQ^{2/D+1/(32D)}. \tag{J.15}$$

The last step uses (J.5). The constant response supplies the output intercept even when the original output bias is fixed to zero.

J.3 Exact compression of every near-optimal behavior

We now prove Theorem J.1. Put $\alpha = n/Q^2$, $\tau = 1/(32D)$, and retain $\ell = \log Q$. Comparison with (J.15) in the original objective gives, throughout (J.6),

$$R(\theta) \le E_0 = CQ^{2/D+\tau}, \qquad J(\theta)/\lambda \le C_D E_0, \qquad \mathcal{L}_S(F_\theta) \le CQ^{-\eta}. \tag{J.16}$$

Here C is enlarged so that $E_0 \ge 1$. The terms Q^{-2} and $\Delta \le Q^{-3/4}$ are absorbed using the two endpoints of the same interval.

A common sampling event and output-row control. Every row, column, and group-label fiber has size Q. For a fixed such fiber T and its sample count $Z = |S \cap T|$, put $\mu = n/Q$ and $u = \log(6Q^5)$. For $v \in \mathbb{R}$, the hypergeometric moment-generating function is at most $(1-p+pe^v)^n \le \exp(\mu(e^v - 1))$, with $p = 1/Q$: the normalized elementary symmetric mean of the positive numbers $e^{v\mathbf{1}_T(x)}$ is at most their arithmetic mean to power n. Exponential Markov bounds give, for $t \ge 0$,

$$\Pr(Z - \mu \ge t) \le e^{-t^2/[2(\mu + t/3)]}, \qquad \Pr(Z - \mu \le -t) \le e^{-t^2/(2\mu)}.$$

With $t = \sqrt{2\mu u} + 2u/3$ and a union bound over $3Q$ fibers, failure probability is at most $6Qe^{-u} = Q^{-4}$. Since $\mu \ge 16u$, all their counts lie in $[\mu/2, 3\mu/2]$. For the class frequencies π_c this also gives

$$\frac{1}{2Q} \le \pi_c \le \frac{3}{2Q}, \qquad |\pi_c - \pi_d| \le 4\sqrt{\frac{u}{nQ}}, \qquad u \le 8\ell.$$

This event is independent of the architecture and λ.

The calculation in Appendix G.6 uses only these count bounds, the group-label fiber property, and the two bounds involving $E_0 \ge 1$ in (J.16). It therefore gives (G.7) with the present E_0:

$$s_F := \max_c \|W_{D,c:}\|_2^2 \le C_D \left[\frac{E_0}{Q} + \frac{E_0^{D-1}\ell}{\alpha\lambda^2 Q^3} + \frac{\Delta}{\lambda}\right]. \tag{J.17}$$

The conditional readout minimum is taken within the original architecture, including a bias variable exactly when that variable is present there; its strong convexity bounds the squared distance to the original readout by $2\Delta/\lambda$.

One approximation for the sample and complete table. Set $\gamma = \ell/16$, $f = F_\theta/\gamma$, and $K = \lceil Q^{3/4} \rceil$. Lemma G.3 constructs a single network H within the original architecture, using at most $\min(k_i, K+1)$ positions per layer. Its row and column domination controls the mean squared maximum class-difference error under both the uniform measure ν on G^2 and the empirical measure $\widehat{\nu}_S$. Let E_K be the larger of these two mean squared errors. For the empirical measure this follows by summing the same row and column bounds against the row and column counts just obtained. Equations (G.8) and (J.17) give in either measure

$$\begin{aligned} E_K &\le \frac{C_D}{\gamma^2 Q}\left[\frac{E_0^{D-1} s_F \ell}{K} + \frac{E_0^D \ell^2}{K^2}\right] \\ &\le C\left[\frac{Q^{D\tau}}{K\ell} + \frac{Q^{2D\tau+2\eta}}{\alpha K} + \frac{\Delta Q^{1+D\tau+\eta}}{K\ell} + \frac{Q^{1+D\tau}}{K^2}\right]. \end{aligned} \tag{J.18}$$

Since $n \ge \alpha_0 Q^{2-\beta}/2$, the four terms are bounded by constant multiples of

$$Q^{-23/32}/\ell, \quad Q^{\beta+2\eta-11/16}, \quad Q^{\eta-15/32}/\ell, \quad Q^{-15/32}.$$

Each is at most $CQ^{-1/8}$ on the full interval, including its endpoints. Moreover, $\log(1+e^{-\gamma}) \ge e^{-\gamma}/2$ gives

$$\frac{\mathcal{L}_S(F_\theta)}{\log(1+e^{-\gamma})} \le 2e^{\gamma}\mathcal{L}_S(F_\theta) \le CQ^{1/16-\eta} \le CQ^{-1/8}. \tag{J.19}$$

Apply the sample-size-dependent estimate (I.22) from the proof of Lemma I.4, with auxiliary failure probability Q^{-3}. Here $\Xi_Q(K) \le C_D Q^{7/4}\ell$ and

$$\frac{Q^2}{n}[\Xi_Q(K) + 4\ell] \le CQ^{7/4+\beta}\ell \le CQ^{15/8}.$$

Indeed the theorem's finite condition gives $\ell \le Q^{3/[128(D+7)]} \le Q^{1/8-\beta}$. Together with $K^2 \le CQ^{3/2}$ and $E_K \le CQ^{-1/8}$, this yields $m^2 \le CQ^{15/8}$. Thus $h := m \le CQ^{15/16}$ and $t_0 \le 1$, $t_i \le h$ in the original architecture, as claimed in (J.7). The cited estimate includes the full supported-table correction, both original-width cases, the prescribed bias convention, and preservation of every prediction and tie choice.

Finally the independent counting bound is at most $C_D(Qh + h^2)\ell \le CQ^{31/16}\ell$, including the logarithmic budget weight, while $\mathrm{err}_{\mathrm{in}}(B_\theta) \le CQ^{-3/16}$ by (J.16). The arbitrary-sample-size statement of Theorem H.3 therefore gives (J.8). The held-out conversion costs at most $(1-\alpha_0)^{-1}$. Its failure probability is δ; combining the three sampling events gives the stated $1 - \delta - Q^{-3} - Q^{-4}$. All three events concern fixed containing classes or fixed fibers, so the same event covers the entire regularization interval and all allowed architectures. □

J.4 Examples and closure under group constructions

Abelian, symmetric, and Heisenberg groups. Every finite abelian group acts freely on itself by translations; its cyclic factors give one-dimensional blocks with $U = 0$, so $(r_*, C_*) = (0, 1)$. For S_m, $m \geq 2$, permute the coordinates of $A = (\mathbb{Z}/m\mathbb{Z})^m$ and take $x_0 = (0, 1, \ldots, m-1)$. Distinct coordinates make the stabilizer trivial. Also $m^m \leq (m!)^4$: for $m \geq 4$, use $m! \geq (m/2)^{m/2}$; check $m = 2, 3$ directly. Thus $(r_*, C_*) = (0, 4)$ works uniformly. For $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ with p prime and $s \geq 1$, the Heisenberg law is

$$(x, y, z)(x', y', z') = (x + x', y + y', z + z' + x \cdot y'), \qquad x, y \in \mathbb{F}_p^s, \quad z \in \mathbb{F}_p.$$

Its left action on $A = \mathbb{F}_p^{2s+1}$ is affine, with free zero orbit; the linear part differs from the identity only in the last coordinate. Taking that coordinate for U gives $(r_*, C_*) = (1, 1)$.

Matrix groups of bounded dimension over prime fields. Fix $R \in \mathbb{N}^*$. For $2 \leq d \leq R$, let $G = \mathrm{GL}_d(\mathbb{F}_p)$ act on $A = \mathrm{Mat}_d(\mathbb{F}_p)$ by $X \mapsto gX$. The point I_d is free, and $U = A$ has rank at most R^2. The order formula gives

$$|G| = p^{d^2} \prod_{j=1}^{d} (1 - p^{-j}) \geq p^{d^2}/4,$$

since the first factor is at least $1/2$ and the remaining product is at least $1 - \sum_{j \geq 2} 2^{-j} = 1/2$. As $p^{d^2} \geq 16$, $|A| \leq |G|^2$; thus $(r_*, C_*) = (R^2, 2)$. The case $d = 1$ is abelian.

Direct products and permutation wreath products. Direct products use product environments and reference points, with unchanged uniform constants. More generally, let H, K have actions Φ^H, Φ^K, free points x_H, x_K, and constants $(r_H, C_H), (r_K, C_K)$. For any permutation action $\pi : K \to S_m$, the group $H^m \rtimes_\pi K$ has law

$$(h, k)(h', k') = ((h_i h'_{\pi_k^{-1}(i)})_i, kk')$$

and acts on $A_H^m \times A_K$ by

$$((u_i)_i, v) \longmapsto ((\Phi_{h_i}^H(u_{\pi_k^{-1}(i)}))_i, \Phi_k^K(v)).$$

Composition gives the displayed law. Fixing $((x_H)_i, x_K)$ first forces $k = e$, then every $h_i = e$. The rank bound is $\max(r_H, r_K)$, and $|A_H|^m |A_K| \leq (|H|^m |K|)^{\max(C_H, C_K)}$. Thus repeated wreath products retain the same constants; starting from abelian groups gives $(r_*, C_*) = (0, 1)$.

Groups specified by a finite graph. Let p be an odd prime and orient a finite simple graph's s edges as $e = (i, j)$ with $i < j$, on r vertices. On $\mathbb{F}_p^r \times \mathbb{F}_p^s$ use

$$(v, w)(v', w') = (v + v', w + w' + f(v, v')), \qquad f(v, v')_e = v_i v'_j.$$

Bilinearity gives associativity, identity zero, and inverse $(-v, -w + f(v, v))$. On $A = \mathbb{F}_p^r \times \prod_{e=(i,j)} \mathbb{F}_p^2$, the action is

$$(u, (z_e, y_e)_e) \longmapsto (v + u, (w_e + z_e + v_i y_e, v_j + y_e)_e).$$

Composition supplies the same central term $v_i v'_j$. Each edge block has $U_e = \mathrm{span}(1, 0)$; all other blocks have $U = 0$. The zero orbit records (v, w) and is free. Since $|A| = p^{r+2s} \leq |G|^2$, the constants $(1, 2)$ apply without restricting graph size or degree.

J.5 Groups associated with learning tasks

Fix $M \in \mathbb{N}^*$ with $M \geq 48$. Let $\mathscr{B}_M$ contain all finite abelian groups, symmetric groups S_m, alternating groups A_m, dihedral groups D_m of order $2m$ ($m \in \mathbb{N}^*$), and groups $A_{\boldsymbol{L}} \rtimes P$, where $A_{\boldsymbol{L}} = \prod_{j=1}^{d} \mathbb{Z}/L_j\mathbb{Z}$, $d \in \{2, 3\}$, $L_j \in \mathbb{N}^*$, and P is a finite lattice rotation or reflection group preserving the periods, with $|P| \leq M$. Let $\mathscr{G}_{\mathrm{ML}}(M)$ be the closure of $\mathscr{B}_M$ under finite direct

products and permutation wreath products $H \wr_\pi K = H^t \rtimes_\pi K$, where $t \in \mathbb{N}^*$ and $\pi : K \to S_t$ permutes the copies of H.

Modular addition and permutation multiplication provide direct group-law learning tasks [52, 63]. Discrete geometric symmetries and hierarchical data motivate the lattice and wreath-product constructions [12, 69]. The guarantee below concerns multiplication-table learning for these groups with the architecture and objective of Section 4.1.

The same sampling and error exponents hold for these groups. A single exponent ρ records the additional logarithmic cost of the target realization; the abelian choice retains the stronger logarithmic factors of Corollary G.7.

Corollary J.2 (A common WD interval and compression rate)**.** *Take $\rho = 1$ for a finite abelian group with $D \geq 2$, or $\rho = 1 + 5/D$ for $G \in \mathscr{G}_{\mathrm{ML}}(M)$ with $D \geq 3$. Let $Q = |G| \geq 3$, and use the architecture and objective of Section 4.1, with $\min_i k_i \geq C_{D,M} \log Q$ and the same width-ratio bound, where $C_{D,M} > 0$ is sufficiently large. Fix $\alpha_0 \in (0,1)$, $0 \leq \beta < 1/8$, $0 < \omega < 1/8 - \beta + 1/(8D)$, and $\delta \in (0,1)$. Write $n = \lfloor \alpha_0 Q^{2-\beta} \rfloor$, $\alpha = n/Q^2$, $\ell = \log Q$, and*

$$\begin{aligned} \lambda_0 &= Q^{-1/4-15/(8D)} \ell^{-(2+\rho-3/D)}, \quad \lambda_0 Q^{-\omega} \leq \lambda \leq \lambda_0, \\ \Delta_{\mathrm{tol}}(\lambda) &= n\lambda Q^{2/D-9/4} \ell^{-[\rho(D-1)+4]}. \end{aligned} \tag{J.20}$$

For every Q satisfying the finite conditions below, with probability at least $1 - \delta - Q^{-3} - Q^{-4}$, simultaneously over all allowed architectures, both output-bias conventions, and all λ in this interval, every parameter with $\Delta(\theta) \leq \Delta_{\mathrm{tol}}(\lambda)$ has an attainable exact EGC budget in its original architecture satisfying

$$t_0 \leq 1, \qquad \max_{1 \leq i < D} t_i \leq C \frac{Q^{7/8}}{\ell}, \qquad \mathrm{err}_{\mathrm{out}}(B_\theta) \leq C Q^{\beta - 1/8} + C \frac{\log(1/\delta)}{n}. \tag{J.21}$$

More generally, for any objective excess $\Delta = \Delta(\theta) \geq 0$, put $\mu = \lambda Q^{2/D}$ and

$$m_\theta = \min \left\{ Q, \frac{Q^{7/8}}{\ell} + Q \sqrt{\frac{\Delta \ell^{\rho(D-1)+2}}{\alpha \mu}} \right\}. \tag{J.22}$$

On the same event, $t_0 \leq 1$, $\max_i t_i \leq C m_\theta$, and

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq \min \left\{ 1, C \left[\frac{Q^{-1/8}}{\alpha} + \sqrt{\frac{\Delta \ell^{\rho(D-1)+4}}{\alpha^3 \mu}} \right] + C \frac{\log(1/\delta)}{n} \right\}. \tag{J.23}$$

Here $\Delta(\theta)$ is measured for the specified λ in the original architecture. All constants depend only on $D, M, C_{\mathrm{w}}, \alpha_0, \beta, \omega$; the abelian branch is independent of M.

Thus the larger group class has the same polynomial compression and generalization rates, with a different logarithmic dependence in the sufficient regularization conditions. At fixed objective excess, the excess-dependent term in (J.23) decreases as $\lambda^{-1/2}$ throughout the interval. The proof first preserves the complete prediction map, including errors, and then applies the independent EGC generalization bound.

Proof. For $\rho = 1$, the statement follows from Corollary G.7, under its original finite conditions in Appendix G.10. It remains to prove the $\rho = 1 + 5/D$ branch. We first verify the group condition, then establish the improved comparison energy and the resulting compression estimate.

A common affine action. The preceding constructions give $(r_*, C_*) = (0, 1)$ for abelian groups and $(0, 4)$ for symmetric groups. Restricting the symmetric-group action to a nontrivial A_m retains a free orbit, and $m^m \leq (m!)^4 = (2|A_m|)^4 \leq |A_m|^8$, so $(0, 8)$ suffices.

If $B \leq G$ is abelian of index $s \leq M$, choose left-coset representatives t_j and write $g t_j = t_{\pi_g(j)} a_{g,j}$ with $a_{g,j} \in B$. In target coordinates $b_{g,i} = a_{g, \pi_g^{-1}(i)}$, multiplication gives $b_{gh,i} = b_{g,i} b_{h, \pi_g^{-1}(i)}$. Thus $g \mapsto (b_g, \pi_g)$ embeds G into $B \wr S_s$: an element in the kernel fixes every t_j and hence is the identity. For $s \geq 2$, let this wreath product act on $A = B^s \times (\mathbb{Z}/s\mathbb{Z})^s$ by $(y, z) \mapsto (b_g + P_{\pi_g} y, P_{\pi_g} z)$, where

P_π permutes coordinates and additive notation is used in B. The point $(0,(0,1,\dots,s-1))$ has trivial stabilizer. Every cyclic coordinate is a one-dimensional block with $U=0$, and $|A|=|B|^s s^s = |G|^s \le |G|^M$. The case $s=1$ is abelian. This covers D_m using its cyclic subgroup of index two and $A_{\boldsymbol{L}} \rtimes P$ using its abelian subgroup of index $|P|$.

All basic groups therefore have common constants $(r_*, C_*) = (0, M)$. The direct-product and permutation-wreath constructions above retain these constants under any finite number of compositions. In particular, each group has a free affine orbit in a finite abelian environment of size at most Q^M.

Character codes and comparison energy. Write the finite abelian environment as $A = \prod_j \mathbb{Z}/n_j\mathbb{Z}$, omitting trivial factors, and set $B = \sum_j \lceil \log_2 n_j \rceil \le 2\log_2 |A| = O_M(\ell)$. Let $x_g = c_g + L_g x_0$ be the free orbit. For an additive character $\chi : A \to \mathbb{R}/\mathbb{Z}$, use the triangular-wave code $\varphi(t) = (\psi(t), \psi(t+1/4))$, where $\psi(t) = \operatorname{dist}(t,\mathbb{Z})$. The separation inequality in Appendix J.2 gives $\|\varphi(t)-\varphi(s)\|_\infty \ge \frac{1}{2}\operatorname{dist}(t-s,\mathbb{Z})$. If $x_g - x_h$ has order $q \ge 2$, a uniform character evaluates it uniformly on $\{0, 1/q, \dots, (q-1)/q\}$. The mean squared distance of this set from $\mathbb{Z}$ is $(q^2-1)/(12q^2)$ for odd q and $(q^2+2)/(12q^2)$ for even q, in either case at least $1/16$.

Choose $N = \lceil 16384\ell \rceil$ independent uniform characters. Each squared code distance lies in $[0,1/2]$ and has mean at least $1/64$. Hoeffding's inequality and a union bound give failure probability at most $Q^2 e^{-N/2048}/2 < 1$. Fix a successful choice, yielding codes $y_g \in [0,1/2]^{2N}$ with $\|y_g - y_h\|_2^2 \ge N/128$ for $g \ne h$.

The pullback $\chi_r \circ L_a$ is a character, so write $\chi_r(L_a x) = \sum_j \eta_{rj}(a) x_j/n_j$ modulo one. Let $\epsilon_{ju}(b)$ be the uth binary digit of $x_{b,j}$ for $0 \le u < \lceil \log_2 n_j \rceil$, and put $v_{rju}(a) = [2^u \eta_{rj}(a)]_{n_j}/n_j \in [0,1)$, where $[\cdot]_{n_j}$ denotes the least nonnegative residue. For $s \in \{0, 1/4\}$, let $d_{rs}(a) = \{\chi_r(c_a)\} + s$. The first hidden layer computes

$$\sigma(d_{rs}(a)), \qquad \sigma(v_{rju}(a) + \epsilon_{ju}(b) - 1) = v_{rju}(a)\epsilon_{ju}(b), \qquad \sigma(1).$$

These are permitted separated-input responses. Their sum $T_{rs} = d_{rs}(a) + \sum_{j,u} v_{rju}(a)\epsilon_{ju}(b)$ lies in $[0, B+2)$ and equals $\chi_r(x_{ab}) + s$ modulo one. The second hidden layer computes $q_{rs,t} = \sigma(T_{rs} - t/2)$ for $0 \le t \le 2B+4$, using the first-layer constant for the shifts. Then

$$\psi(T_{rs}) = q_{rs,0} + 2\sum_{t=1}^{2B+4} (-1)^t q_{rs,t}.$$

Thus the output can recover y_{ab} exactly. Set $c_0 = \lceil \sqrt{2N} \rceil$, transmit this constant through the hidden layers, and use scores $F_c^0 = 2y_c^\top y_{ab} - (\|y_c\|_2^2/c_0)c_0$. Their correct-class margin is at least $N/128$. For $D=3$, the output directly combines the hinges; for $D \ge 4$, the third hidden layer forms the nonnegative code and later layers transmit it. Every bias after the first layer, including the output bias, is zero.

For $A_i = \|W_i\|_F^2 + \|b_i\|_2^2$, the explicit coefficients give $A_1 \le CQN(B+1)$ and $A_2 \le CN(B+1)^3$. If $D=3$, $A_3 \le CQN(B+1)$; otherwise $A_3 \le CN(B+1)$, the intervening identity layers cost at most CN, and $A_D \le CQN$. The chosen c_0 includes the output constant in these bounds. Hence $\prod_i A_i \le C_D Q^2 N^D (B+1)^5$, using $O_M(\ell^2)$ positions per layer. Scaling the output by $128/N$ gives unit margin. Positive homogeneous balancing makes the quadratic penalty $\frac{D}{2}[(128/N)^2 \prod_i A_i]^{1/D}$. Scaling the margin further to 4ℓ yields

$$\mathcal{L}_S(F_G) \le Q^{-3}, \qquad R(F_G) \le CQ^{2/D}\ell^\rho, \qquad \rho = 1 + 5/D. \tag{J.24}$$

The same construction gives the required comparison on each nonempty rectangle $I \times J$. Every nonconstant first-layer preactivation has the form $u(a) + v(b) + b_1$, with $0 \le u \le 5/4$, $0 \le v \le 1$, and $b_1 \in \{-1, 0, 1\}$. Replace it by $\mathbf{1}_I(a)(u(a)+4) + \mathbf{1}_J(b)(v(b)+4) + b_1 - 8$, and replace the constant response by $\sigma(\mathbf{1}_I(a) + \mathbf{1}_J(b) - 1)$. All first-layer responses agree inside the rectangle and vanish outside. All later biases are zero, so the entire score vanishes outside. For $v = |I| + |J|$, the first-layer cost becomes at most $CvN(B+1)$. Balancing to unit correct margin therefore gives

$$R(G_{I,J}) \le CQ^{2/D}\ell^{\rho - 2/D}(v/Q)^{1/D}. \tag{J.25}$$

All constants and rectangle restrictions are realized by paid parameters in the original architecture whenever the stated positions fit.

Finite conditions. Set $\nu = 2-1/D$, $a_D = 2+\rho-3/D$, $p = \rho(D-1)+2$, $\eta_0 = \nu/8$, $\eta_1 = \eta_0+\omega$, and $m_0 = Q^{7/8}/\ell$. For a sufficiently large constant C depending only on the fixed parameters, require

$$\alpha Q \ge C\ell, \quad m_0 \ge C\ell^2, \quad \alpha m_0/\ell \ge 2, \quad \ell \ge C, \quad \ell^{D+7} \le Q^{3/128}.$$

In addition, require that C times the sum of the following nonnegative quantities be at most one:

$$\begin{gathered}\frac{Q^{-3/4+2\eta_1}\ell^{\rho(2D-2)+5+2a_D}}{\alpha^2}, \quad \frac{Q^{-5/8}\ell^{\rho D+7}}{\alpha^2}, \quad Q^{-1/8}\ell^{\rho D/2+5}, \quad Q^{-13/(8D)}\ell^{9/D}, \\ \alpha Q^{-\eta_0}\ell^{-(a_D+p)}, \quad \frac{\alpha}{\ell}, \quad \frac{Q^{-3/4}\ell^3}{\alpha}, \quad \frac{Q^{-5/8}\ell^{\rho D+4}}{\alpha}, \quad Q^{-3+\eta_1}\ell^{a_D-\rho}, \\ \alpha Q^{-15/8}, \qquad \alpha Q^{-1/8+5/(32D)}\ell^{-a_D}.\end{gathered} \tag{J.26}$$

These conditions involve only the task size and fixed parameters. They hold simultaneously beyond a finite threshold because $\beta < 1/8$ and $\eta_1 < 3/8 - \beta$. They also imply $n \ge (\alpha_0/2)Q^{2-\beta}$. The constants absorb the fixed ratio between $\log(2Q)$ and ℓ; all formulas here use the same $\ell = \log Q$.

Exact compression with the new energy bound. Write $\widehat{m} = m_0 + Q\sqrt{\Delta\ell^p/(\alpha\mu)}$, and first suppose $m = \widehat{m} < Q$. Then

$$\frac{\Delta Q^2\ell^p}{\alpha\mu m^2} \le 1, \qquad \Delta < \alpha\mu\ell^{-p}.$$

If the smallest original width k is at most $C_{\mathrm{rec}}m$, retain the original realization, whose hidden widths are at most $C_{\mathrm{w}}C_{\mathrm{rec}}m$ and whose unreduced affine output is constant, of rank at most one. Otherwise $k > C_{\mathrm{rec}}m$ and the architecture contains (J.24). Original-objective comparison gives $1 + J(\theta)/\lambda \le E_0 := CQ^{2/D}\ell^\rho$.

We use the same approximation center and sample events as in Corollary G.7. The readout comparison of Appendix J.3, with the actual excess Δ, gives

$$s_F \le C\left[E_0/Q + E_0^{D-1}\ell/(\alpha\lambda^2Q^3) + \Delta/\lambda\right].$$

Thus the common row-and-column approximation, Lemma G.3, at integer budget K, has class-difference error at most $1/8$ outside at most

$$h_e \le C\left[\frac{Q\ell^{\rho(2D-2)+2}}{\alpha\mu^2K} + \frac{\Delta Q^2\ell^{\rho(D-1)+1}}{\mu K} + \frac{Q^2\ell^{\rho D+2}}{K^2}\right] \tag{J.27}$$

rows and columns. Its readout and row-and-column bounds hold simultaneously for this center. After deleting another $O(h_0)$ strips, Lemma G.4 supplies, on a rectangle of size $v = |I| + |J|$, a $1/64$ score grid of logarithmic size at most

$$C\ell\left[vK + K^2 + Q + Q^2\ell^{\rho D}/h_0\right].$$

Indeed its norm parameter is now $P_0^2 = CQ^2\ell^{\rho D}$, while the reference scores remain polynomially bounded in Q. Allocate total failure probability Q^{-3} over all integer $K, h_0 \le Q$, rectangles, and grid elements exactly as in that corollary. The row, column, and label-fiber concentration event has failure probability at most Q^{-4}. These events are independent of the later choices of λ, Δ, θ.

For the rectangle comparison, Lemma G.2 now releases $O(\ell^2)$ positions for (J.25), rather than $O(\ell)$. Its parameter-norm estimate and mean-square score estimate give the total objective perturbation

$$e_k \le C\left[\frac{\sqrt{Q}\,\ell^{\rho D/2+2}}{k} + \mu\ell^\rho(\ell^2/k)^{2/D}\right].$$

The paid constant channels are included in the released positions. Adding the unit-margin rectangle score decreases the loss by a fixed positive amount at every observed point whose top-two gap is at most $1/2$, and does not increase it at other points of the rectangle. Therefore the original objective bounds the fraction of such observed points by $C[\mu\ell^{\rho-2/D}(v/Q)^{1/D} + e_k + \Delta]$.

Apply the preceding score-grid event with the same thresholds $1/64, 1/8, 1/4, 1/2$ as in the original proof. For the center's small-gap graph outside the deleted strips, its degeneracy d_A satisfies

$$\begin{aligned}d_A \le C\Bigg[&K\ell/\alpha + Q(\mu\ell^{\rho-2/D})^{1/\nu} + Q\sqrt{e_k+\Delta} \\ &+ \sqrt{\frac{\ell}{\alpha}\left(K^2 + Q + Q^2\ell^{\rho D}/h_0\right)}\Bigg].\end{aligned} \tag{J.28}$$

This follows by applying the rectangle bound to an induced subgraph of minimum degree d: it has at least $dv/2$ edges and $v \geq 2d$, as in (G.31).

Take $K = \lceil C_s \alpha m/\ell \rceil$ and $h_0 = \lceil m \rceil$. The finite conditions ensure $1 \leq K, h_0 \leq Q$. Dividing (J.27) by m, the first and third terms are bounded by the first two quantities in (J.26); the middle term is at most a fixed constant by the definition of m. The second term in (J.28) obeys the exact identity

$$Q\left(Q^{-\eta_0}\ell^{-a_D}\ell^{\rho-2/D}\right)^{1/\nu} = m_0.$$

For the remaining terms, square and divide by m^2, using $k > m \geq m_0$. The two position-release terms give $Q^{-1/8}\ell^{\rho D/2+5}$ and $Q^{-13/(8D)}\ell^{9/D}$; the excess gives at most $\alpha\mu\ell^{-p}$. The three score-cover terms give α/ℓ, $Q^{-3/4}\ell^3/\alpha$, and $Q^{-5/8}\ell^{\rho D+4}/\alpha$. These are precisely the corresponding quantities in (J.26), so $h_e + d_A + K + h_0 + 1 \leq Cm$.

Lemma I.3 now corrects every remaining disagreement, including the original errors and tie winners, using at most $C'm$ positions per layer. Choosing $C_{\mathrm{rec}} > C'$ fits this realization in the original wide architecture; the narrow case was already retained. When $\widehat{m} \geq Q$, the row-table construction in Corollary G.7 gives the same conclusion with $m = Q$, using either a $Q+1$-position realization or the original bounded-width-ratio realization. This proves the exact EGC budget for all original widths and all objective excesses.

Training fit and independent generalization. The separate $O(\ell)$-width construction in Appendix J.2 fits every original architecture. By (J.15) and the finite condition $\ell^{D+7} \leq Q^{3/128}$, it has loss at most Q^{-2} and penalty at most $CQ^{2/D+\tau}$, where $\tau = 1/(32D)$. Hence

$$\mathcal{L}_S(F_\theta) \leq Q^{-2} + C\mu Q^{\tau} + \Delta.$$

This comparison controls training fit even when the character construction does not fit. In the nonsaturated case, $\Delta < \alpha\mu\ell^{-p}$, and division by $Q^{-1/8}/\alpha$ bounds the training loss by a fixed multiple of

$$\alpha Q^{-15/8} + \alpha Q^{1/8-\eta_0+\tau}\ell^{-a_D}, \qquad 1/8 - \eta_0 + \tau = -1/8 + 5/(32D) < 0.$$

The last two finite conditions absorb this into the claimed error rate, since $\mathrm{err}_{\mathrm{in}} \leq \mathcal{L}_S/\log 2$.

After exact recovery, the independent EGC count is at most $CQm\ell$. Allocating confidence $\delta/\lceil CQ \rceil$ to each integer budget, Lemma D.1 gives

$$\mathrm{err}_{\mathrm{out}}(B_\theta) \leq \min\left\{1, \frac{2-\alpha}{1-\alpha}\,\mathrm{err}_{\mathrm{in}}(B_\theta) + \frac{CQm\ell + 8\log(1/\delta)}{(1-\alpha)n}\right\}.$$

Substitute the training estimate and $\widehat{m}\ell/(\alpha Q) = Q^{-1/8}/\alpha + \sqrt{\Delta\ell^{p+2}/(\alpha^3\mu)}$ to obtain (J.23). In the saturated case the latter untruncated bound is at least $\ell/\alpha \geq 1$. Finally, $\Delta \leq \Delta_{\mathrm{tol}}(\lambda)$ gives $m \leq 2m_0$ and (J.21). The three sample events have total failure probability at most $\delta + Q^{-3} + Q^{-4}$. □

The fixed-density main theorem. Theorem 4.1 is the case $\beta = 0$ and $\lambda = \lambda_0$ of Corollary J.2. Set the corollary's α_0 equal to the fixed nominal density α of Section 4.1, and fix $\omega = 1/32$. The required finite conditions are those in Appendix G.10 for abelian groups and (J.26), together with its preceding displayed conditions, otherwise, with these substitutions. The corollary's actual sampling fraction is n/Q^2; its finite conditions give $n \geq (\alpha/2)Q^2$. Direct substitution gives

$$n\lambda_0 Q^{2/D-9/4}(\log Q)^{-[\rho(D-1)+4]} = nQ^{-5/2+1/(8D)}(\log Q)^{-(\rho D+6-3/D)} = \Delta_Q.$$

The corollary first preserves every prediction in the original architecture at budget $t_0 \leq 1$, $\max_i t_i \leq CQ^{7/8}/\log Q$, and then applies independent EGC counting. At $\beta = 0$ its error bound is $CQ^{-1/8} + C\log(1/\delta)/n$, which gives (4.4).

K Experimental Protocol and Additional Measurements

Task and paired training. We use addition modulo 127, separate one-hot inputs in $\mathbb{R}^{254}$, two width-64 ReLU hidden layers, and 127 output scores, with biases in all affine maps. Each seed

2901–2910 uniformly permutes the full domain: the first $4{,}838$ pairs train the network and the other $11{,}291$ are held out. The five strengths $\lambda \in 10^{-4}\{0, 0.5, 1, 2, 5\}$ share splits and initial parameters. All seeds and outcomes are retained. The objective is mean cross-entropy plus $\lambda\|\theta\|_2^2/2$ on all weights and biases. Full-batch Adam uses $(0.9, 0.999)$, $\epsilon = 10^{-8}$, and no optimizer weight-decay term. Weights are Xavier-uniform and biases zero. At step s of T, the learning rate is $0.003[0.2 + 0.8(1 + \cos(\pi(s-1)/T))/2]$; paired training uses $T = 24{,}000$.

Continuation and fresh initializations. Figure 1(b,c) continues each unregularized endpoint on its original training set for $24{,}000$ steps at fixed $\lambda = 5 \cdot 10^{-4}$, resetting Adam. We retain steps $0, 2000, 4000, 6000, 8000, 12000, 16000, 20000, 24000$ from each run. All ninety checkpoints fit training perfectly. Each plotted coordinate is the checkpoint's measured J, held-out accuracy, or verified EGC bound; even the starting J uses this same penalty. Red circles mark starts and filled blue circles endpoints. Comparisons along a trajectory use one training set and one objective. Panel (f) uses four fresh initialization seeds 12901–12904 per training set and strength, with $T = 48{,}000$. All 200 networks fit training perfectly. In increasing strength order, $0, 30, 37, 40, 40$ out of forty match by $k = 12$, unchanged through 60; their mean held-out accuracies are $0.105\%, 75.005\%, 92.501\%, 99.905\%, 99.995\%$. The ten training sets are independent sampling units; the four initializations are repeated measurements, not independent datasets.

Searching for exact realizations. For $k \in \{8, 12, 16, \ldots, 60\}$, search uses a width-$(k, k)$ ReLU network and all $16{,}129$ teacher predictions, without reading true held-out labels. Zero padding embeds it in the original architecture with budget $(1, k, k)$: affine rank at most one, followed by the two layerwise nonlinear budgets. Each unresolved teacher–loss–budget pair receives two Xavier-uniform starts, paired across WD conditions and checkpoints, with the same Adam moment parameters. The search learning rate follows the same cosine rule from 0.003 to 0.0006 over $4{,}000$ steps, then from 0.0006 to 0.00012 over $12{,}000$ steps. Complete matching is checked every 100 steps and the first verified match is retained. A verified realization carries to larger budgets by zero padding and rechecking, only for the same teacher and loss. The 64^* point denotes the original realization, not a new search. Panel (d) uses cross-entropy on teacher labels. Panel (e) adds 0.1 times a mean-squared centered-logit discrepancy: teacher scores are centered across classes and divided by their global root-mean-square value, floored at 0.1; student scores are centered across classes. All other search settings are shared. Across the five strengths, the label-only searches match $0, 7, 8, 9, 10$ of ten by $k = 12$, unchanged through 60; the second loss reaches the same counts by 56. Neither matches an unregularized network.

Numerical verification. Stored floating-point weights are treated as exact dyadic values. For a row-vector affine map $y = hW + b$ of fan-in m, let $|h - h_{\mathrm{exact}}| \le e$, $\epsilon = 2^{-52}$, and $g = 8(m+2)\epsilon < 0.01$. With nextup denoting the next larger binary64 value componentwise and ν the smallest positive normal binary64 value, the verifier propagates

$$e' = \mathrm{nextup}(2e|W| + 2g(|h||W| + |b|) + (m+2)\nu)\,.$$

The factors two dominate rounding in the nonnegative sums; the last term allows underflow. ReLU preserves the enclosure by its Lipschitz bound. For the required class c, round $h_c - e_c$ down, each rival $h_d + e_d$ up, and the differences down. Acceptance requires strictly positive lower bounds on all margins at all $16{,}129$ inputs. Nonfinite computations are rejected; a zero lower bound is inconclusive. Teachers and proposed realizations undergo the same check. No structural rank is inferred from small singular values.

Complete records and interpretation. The accompanying records contain all 340 plotted networks and checkpoints: 50 paired endpoints, 90 continuation checkpoints, and 200 fresh endpoints. They retain training indices, complete predictions, measured metrics, original and realizing weights, search initializations, successes and failures, and standalone verifiers. The comparisons on the uniform budget grid use $3{,}394$ search starts; the continuation measurements retain 222 search starts. Every plotted original and reported realizing network was reopened and checked on the full domain. There are 160 distinct complete prediction maps. Identical maps share their best verified same-architecture bound in panel (c); the 162 group-law records therefore share $(1, 8, 8)$. Panels (d–f) count each network's own searches and transfer no successes across networks. Successful search gives an EGC upper bound; failure gives no lower bound. Measured objective values do not determine the gap to $J^*_{\mathcal{A}}$.

L FITTING ENERGY FOR RANDOM LABELS

The group-law constructions admit low-loss realizations with controlled quadratic energy. We compare their fitting cost with that of independently random labels, using the same inputs, architecture, loss, and energy R from (4.1). Requiring the same cross-entropy tolerance is essential: rescaling all logits preserves their predicted labels but changes both energy and confidence.

The network has $h_0 = (e_a, e_b)$, $h_l = \sigma(W_l h_{l-1} + b_l)$ for $1 \le l < D$, and $F_\theta = W_D h_{D-1} + b_D$, where $\sigma(t) = \max\{t, 0\}$ acts coordinatewise and $R(\theta) = \frac{1}{2}\sum_{l=1}^{D}(\|W_l\|_F^2 + \|b_l\|_2^2)$.

Proposition L.1 (Random-label fitting energy)**.** *Fix an affine depth $D \in \mathbb{N}$, $D \ge 2$, and $\alpha \in (0,1)$. Let $S \subseteq [Q]^2$ be any fixed set of $n = \lfloor \alpha Q^2 \rfloor$ distinct inputs, and assign independent uniform labels $Y_x \in [Q]$ to $x \in S$. There are constants $c, C, Q_0 > 0$, depending only on D, α, such that, for every integer $Q \ge Q_0$, with probability at least $1 - CQ^{-10}$, simultaneously for every finite-width depth-D ReLU network,*

$$R(\theta) + \frac{D-1}{2} \ge c\left[\frac{(\log Q - 2\mathcal{L}_S(F_\theta; Y))_+\sqrt{nQ}}{(\log Q)^{3/2}}\right]^{2/D}. \tag{L.1}$$

Here $z_+ = \max\{z, 0\}$, and $\mathcal{L}_S(F; Y) = n^{-1}\sum_{x\in S}[\log\sum_c e^{F_c(x)} - F_{Y_x}(x)]$. In particular, every network satisfying $\mathcal{L}_S(F_\theta; Y) \le \frac{1}{4}\log Q$ obeys

$$R(\theta) \ge c\left(\frac{Q^3}{\log Q}\right)^{1/D} - \frac{D-1}{2}. \tag{L.2}$$

All hidden biases are included in R; the output bias is either included and penalized or fixed to zero. The event covers parameters chosen after observing the random labels and does not require optimality of a training objective.

Proof. Write $S = \{(a_i, b_i) : 1 \le i \le n\}$, $d = n/Q$, and $\ell = \log Q$. For sufficiently large Q, every input row and column has at most κd observed entries, where $\kappa = 2/\alpha$. Constants below may depend on D, α.

We first establish a uniform centered-score estimate. Set

$$\widetilde{x}_{ab} = (e_a, e_b, 1), \qquad \widetilde{x}_0 = (Q^{-1}\mathbf{1}, Q^{-1}\mathbf{1}, 1).$$

Let $\mathcal{H}_D$ be the class of scalar ReLU networks with a linear output and at most D matrices, no separate biases, arbitrary finite widths, and each matrix having Frobenius norm at most one. For $h \in \mathcal{H}_D$, put $g_h(a,b) = h(\widetilde{x}_{ab}) - h(\widetilde{x}_0)$. There are nonnegative vectors u, v such that

$$|g_h(a,b)| \le u_a + v_b, \qquad \|u\|_2^2 + \|v\|_2^2 \le 1. \tag{L.3}$$

Indeed, if U, V are the two input blocks of the first matrix, take $u_a = \|U(e_a - Q^{-1}\mathbf{1})\|_2$ and define v_b analogously; the remaining layers are 1-Lipschitz. For independent Rademacher signs ϵ_i and arbitrary real q_i,

$$\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}_D}\left|\sum_i \epsilon_i q_i g_h(a_i, b_i)\right| \le C_D\left(\sum_i q_i^2\right)^{1/2}. \tag{L.4}$$

To verify this bound, first consider a scalar network with exactly $s \ge 2$ matrices. If a_j is its last-layer coefficient and r_j is the norm of the corresponding incoming row in the preceding layer, positive homogeneity writes its output as $\sum_j a_j r_j \sigma(h'_j)$, where $h'_j \in \mathcal{H}_{s-1}$ and

$$\sum_j |a_j| r_j \le \Big(\sum_j a_j^2\Big)^{1/2}\Big(\sum_j r_j^2\Big)^{1/2} \le 1.$$

Zero rows contribute nothing. Thus its responses belong to the absolute convex hull of $\sigma(\mathcal{H}_{s-1})$. A linear supremum is unchanged by taking this hull, and Rademacher contraction for ReLU costs at most a factor two. Iterating and taking the finite union over depths at most D reduces the estimate to the linear class, for which $\mathbb{E}\|\sum_i \epsilon_i q_i \widetilde{x}_i\|_2 \le \sqrt{3}\|q\|_2$. For the reference term, $|h(\widetilde{x}_0)| \le \sqrt{3}$ gives $\mathbb{E}\sup_h |h(\widetilde{x}_0)\sum_i \epsilon_i q_i| \le \sqrt{3}\|q\|_2$.

Let $r_a = |\{i : a_i = a\}|$ and define the random count matrices

$$K_{ac} = |\{i : a_i = a, Y_i = c\}|, \qquad M_{ac} = K_{ac} - r_a/Q.$$

The independent centered summands $M = \sum_i e_{a_i}(e_{Y_i} - Q^{-1}\mathbf{1})^\top$ have norm at most one and matrix variance at most κd. The rectangular matrix Bernstein inequality [67, Theorem 1.6] therefore gives $\|M\|_{\mathrm{op}} \le C\sqrt{d\ell}$ with failure probability at most CQ^{-30}. The same estimate holds for the corresponding matrix formed from the second operand.

We need a bound that is uniform in the large coordinates of a network. For any real $Q \times Q$ matrix M, nonnegative unit-ball vectors u, w, and $t > 0$,

$$\|M \odot \mathbf{1}_{\{u_a w_c > t\}}\|_{\mathrm{op}} \le (1 + \lceil \log_2 Q \rceil)\|M\|_{\mathrm{op}}. \tag{L.5}$$

Here $\odot$ denotes entrywise multiplication. Sort w in decreasing order and pad with zeros to a power of two. Each selected row is a prefix of the columns, which decomposes into at most one dyadic interval of each length. At a fixed length, distinct intervals have disjoint column sets and disjoint assigned row sets. The resulting block-diagonal submatrix has norm at most $\|M\|_{\mathrm{op}}$; summing over lengths proves (L.5). Consequently,

$$\sum_{a,c}(K_{ac} + r_a/Q)u_a w_c \mathbf{1}_{\{u_a w_c > t\}} \le (1 + \lceil \log_2 Q \rceil)\|M\|_{\mathrm{op}} + \frac{2\kappa d}{Qt}. \tag{L.6}$$

For the mean term, use $\sum u_a w_c \mathbf{1}_{\{u_a w_c > t\}} \le t^{-1} \sum u_a^2 w_c^2 \le t^{-1}$.

Fix a class direction $z \in \mathbb{R}^Q$, $\|z\|_2 \le 1$, and set $\tau = \sqrt{d}/Q$. Let χ_τ clip a real number to $[-\tau, \tau]$, and define

$$Z_z = \sup_{h \in \mathcal{H}_D} \left| \sum_i \left[\chi_\tau(z_{Y_i} g_h(a_i, b_i)) - \frac{1}{Q} \sum_c \chi_\tau(z_c g_h(a_i, b_i)) \right] \right|.$$

Symmetrization, contraction, and (L.4) give $\mathbb{E}Z_z \le C_D\sqrt{d}$. The centered summands have envelope $b = 2\tau$, and (L.3) gives total variance at most $v = 2\kappa d/Q$ for every fixed function. The concentration inequality of Klein & Rio [32, Theorem 1.1(c)], after rescaling by b, gives

$$\mathbb{P}\{Z_z \ge \mathbb{E}Z_z + t\} \le \exp\left[-\frac{t^2}{2(v + 2b\mathbb{E}Z_z) + 3bt} \right].$$

Thus $Z_z \le C\sqrt{d}$ except on an event of probability $\exp[-(20 + \log 9)Q]$, by increasing C.

Removing the clipping changes the centered sum by at most

$$\sum_{i,c}(\mathbf{1}_{\{Y_i = c\}} + Q^{-1})|z_c g_h(a_i, b_i)|\mathbf{1}_{\{|z_c g_h(a_i, b_i)| > \tau\}}.$$

Using (L.3), split this expression into the two operand envelopes with threshold $\tau/2$ and apply (L.6). Its supremum over z, h is at most $C\sqrt{d}\,\ell^{3/2}$ on the matrix event. A $1/4$-net of the class-direction unit ball has at most 9^Q points. Take the union bound on this net and then use linear approximation only for the *unclipped* expression. With probability at least $1 - CQ^{-20}$,

$$\sup_{\|z\|_2 \le 1,\, h \in \mathcal{H}_D} \left| \sum_i (z_{Y_i} - \bar{z}) g_h(a_i, b_i) \right| \le C\sqrt{d}\,\ell^{3/2}, \qquad \bar{z} = Q^{-1} \sum_c z_c. \tag{L.7}$$

These concentration arguments apply first to finite scalar classes, then to countable dense classes by monotone limits. Continuity on the finite input set extends the event to all finite widths and parameters.

To apply this estimate to a biased multi-output network, put

$$A_l = \begin{pmatrix} W_l & b_l \\ 0 & 1 \end{pmatrix} \quad (l < D), \qquad A_D = (W_D, b_D), \qquad P_b = \prod_{l=1}^{D} \|A_l\|_F.$$

This represents the original function exactly; the constant coordinates are used only for the estimate. In the remainder of this proof, evaluate this augmented representation at $x_i = \widetilde{x}_{a_i b_i}$ and $x_0 = \widetilde{x}_0$. Positive homogeneity and the arithmetic–geometric mean inequality yield

$$P_b \le \left(\frac{2R + D - 1}{D} \right)^{D/2}. \tag{L.8}$$

If $P_b = 0$, the claimed bound is immediate. Otherwise normalize every A_l to unit Frobenius norm and denote the normalized output columns by w_j. Write the last hidden responses as $r_j h_j$, where r_j is the corresponding incoming row norm. Then $\sum_j r_j^2 \le 1$ and $\sum_j \|w_j\|_2^2 \le 1$. Each h_j belongs to $\mathcal{H}_D$: a scalar identity readout after its final ReLU uses at most D matrices. By (L.7), its centered score vector has Euclidean norm at most $C\sqrt{d}\,\ell^{3/2}$. Hence the absolute centered score of the normalized vector output is at most

$$C\sqrt{d}\,\ell^{3/2} \sum_j r_j \|w_j\|_2 \le C\sqrt{d}\,\ell^{3/2}.$$

Restoring the norm product gives

$$\left| \sum_{i,c} (\mathbf{1}_{\{Y_i=c\}} - Q^{-1})[F_c(x_i) - F_c(x_0)] \right| \le C P_b \sqrt{d}\,\ell^{3/2}. \tag{L.9}$$

Center the output across classes; this preserves loss and does not increase P_b. Write $n_c = |\{i : Y_i = c\}|$, $a_c = n_c - d$, and $\beta_c = d/n_c$. Binomial concentration and a union bound give $d/2 \le n_c \le 2d$ and $\max_c |a_c| \le C\sqrt{d\ell}$, with failure probability at most CQ^{-20}. For $f_c(x) = F_c(x) - F_c(x_0)$ and $\bar{f}_c = n^{-1}\sum_i f_c(x_i)$, the class-balanced correct score satisfies the exact identity

$$T_{\mathrm{bal}} := \frac{1}{Q}\sum_c \frac{1}{n_c} \sum_{i:Y_i=c} F_c(x_i) = \frac{1}{n}\left[\sum_{i,c} (\mathbf{1}_{\{Y_i=c\}} - Q^{-1})\beta_c f_c(x_i) - \sum_c a_c \beta_c \bar{f}_c \right]. \tag{L.10}$$

The reference values cancel because $\sum_c F_c(x_0) = 0$. The first sum is bounded by (L.9) applied to $\mathrm{diag}(\beta)F$, whose norm product is at most $2P_b$. For the second sum, the input-block envelope and the row/column degree bound give $n^{-1}\sum_i \|f(x_i)\|_2^2 \le 2\kappa P_b^2/Q$. Consequently,

$$\sum_c |\bar{f}_c| \le \sqrt{Q}\,\|\bar{f}\|_2 \le \sqrt{\frac{Q}{n}\sum_i \|f(x_i)\|_2^2} \le \sqrt{2\kappa}\,P_b.$$

Hence $|T_{\mathrm{bal}}| \le C P_b \ell^{3/2}/\sqrt{nQ}$. The corresponding class-balanced cross-entropy is at most $2\mathcal{L}_S$, while $\sum_c F_c(x) = 0$ implies $\log \sum_c e^{F_c(x)} \ge \log Q$. Therefore $T_{\mathrm{bal}} \ge \log Q - 2\mathcal{L}_S$. Combining these inequalities with (L.8) proves (L.1); $n \ge \alpha Q^2/2$ then gives (L.2). Increasing Q_0, C absorbs the stated failure probabilities. □

Comparison with group labels. The construction in Appendix F.1 also bounds the fitting energy for every finite group. Here are the required norm calculations. In its notation, $d = O(\log Q)$ samples have representation dimensions $m_j \le \sqrt{Q}$ and $\sum_c |\zeta_j(c)|^2 = Q/\gamma_{m_j}^4 = O(Q)$ by Schur orthogonality. The $16m_j$ first-layer sign-product responses of sample j have total squared input norm $32Qm_j$. For $D = 2$, their squared output norm is $2Q/(d^2\gamma_{m_j}^4)$. For $D \ge 3$, the second layer aggregates each sample into the positive and negative parts of its real and imaginary features, at squared norm cost 8; each subsequent hidden layer copies these four nonnegative coordinates at cost 4. The final squared output norm is again $2Q/(d^2\gamma_{m_j}^4)$. All biases are zero. Summing over samples and multiplying layer norms gives $\prod_l \|W_l\|_F^2 \le C_D Q^{5/2}(\log Q)^{D-2}$. The complete-table margin is at least $3/4$. Scaling the scores by $4\log(2Q)$ and balancing the layers therefore gives

$$\max_{x\in[Q]^2} \mathcal{L}_{\{x\}}(F;G) \le Q^{-2}, \qquad R(\theta_G) \le C_D Q^{5/(2D)} \log(2Q). \tag{L.11}$$

This realization uses $O(\sqrt{Q}\log Q)$ positions in the first hidden layer and $O(\log Q)$ in each remaining hidden layer. It is used in a fixed architecture only when these positions are available. The random-label lower bound holds for that same architecture.

Thus, at the common loss tolerance Q^{-2}, the random-label lower bound exceeds the group-law upper bound by a factor of at least $cQ^{1/(2D)}/(\log Q)^{1+1/D}$ for sufficiently large Q. This is a separation in fitting energy, not merely in representability. For example, in a common architecture with every hidden width $Q+1$, both tasks are representable for sufficiently large Q. Index the labels by $\{0,\dots,Q-1\}$ for this construction. For an arbitrary label table y_{ab}, the first-layer responses

$h_i(a,b) = \sigma(Q\mathbf{1}_{\{a=i\}} + y_{ib} - Q)$ and $h_*(a,b) = 1$ give $\sum_i h_i = y_{ab}$. Transmit this nonnegative sum and the constant, and use scores $\gamma(2cy_{ab} - c^2)$. Their winning margin is at least γ, and all weights and biases have finite paid energy. Taking $\gamma = 3\log(2Q)$ reaches the common loss tolerance. The energy separation therefore quantifies a benefit of the target's group structure even when both label assignments can be fitted in the same original architecture.