

# Flat Loss, Evolving Features: Grokking in Modular Addition

Akash Kumar\*

University of California, San Diego

Karthik Sankaranarayanan

Microsoft AI

## Abstract

Can a network keep learning while its training cross-entropy stays nearly flat at a positive level? We show that it can. On modular addition, a standard testbed for grokking, ReLU networks trained with supported-Polar hidden updates, normalized output-layer updates and weight decay can keep 100% training accuracy while their hidden features change and held-out accuracy rises from zero to above 90%, at relatively flat training cross-entropy. Correct classification need not minimize cross-entropy, which still depends on prediction confidence. We prove this coexistence for one-hidden-layer networks in two finite-window theorems from constructed initializations: one gives a constant gain in the Fourier fraction of the hidden features, and the other holds at arbitrarily low positive loss, where a readout refitted on the training data generalizes before the network. Experiments from random initialization show the same coexistence on a window fixed in advance, and in deeper networks held-out accuracy also rises over post-hoc windows with a net increase in training cross-entropy. With a last-layer rate set per training fraction and head decay scaled with the prime, deeper networks at  $p = 47$  and  $61$  meet a criterion on post-hoc windows, fixed in advance, in nearly all fresh runs. A one-block transformer trained with the same updates and a slowly decaying output gain keeps every training pair correct at every logged checkpoint while its training cross-entropy rises and its held-out accuracy approaches one; with two blocks this holds through the held-out rise, after which training pairs begin to flip. A plateau in training loss need not be a plateau in learning. Code: <https://github.com/akashkumar-d/flat-loss-evolving-features>.

## 1 Introduction

The training loss is the standard measure of a neural network’s progress, and a plateau in the loss is commonly interpreted as a sign that learning has stalled. Grokking calls this interpretation into question. On small algorithmic tasks such as modular addition, networks fit their training data long before they generalize [1]. Progress measures reveal the formation of a generalizing circuit while the training loss remains nearly flat [2], and the learned features change gradually before held-out accuracy recovers [24]. Such hidden progress has been proved for online sparse-parity learning [4] and for teacher-student regression during loss plateaus, under orthogonalized updates [31] and for the population loss [21], and analyzed after fitting under the squared loss with coupled weight decay [22].

Classification networks, however, are trained with the cross-entropy (CE) loss, for which fitting the training set does not end training. A training pair is classified correctly as soon as its margin, the gap between the logit of its true class and the largest other logit, is positive. The CE also measures confidence: a network that classifies every training pair correctly can have any CE below  $\log p$ , the loss of the uniform prediction over  $p$  classes. Without regularization, the CE can keep decreasing as the logits grow [12]; with weight decay, it can level off at a positive value. A level CE then fixes only the mean log-probability of the true class on the training pairs. It does not determine whether the hidden features, or the predictions on held-out pairs, are still changing. This leads to the following question.

*Can a network that has fitted its training data continue to improve its hidden features, and recover held-out accuracy, during a window of training on which its CE stays relatively flat at a positive level?*

We study this question on a window  $I$  of the actual training trajectory. Flatness is measured by the relative variation  $v(I) = \max_{t \in I} L_t / \min_{t \in I} L_t - 1$  of the raw training CE, checked at every update, and

---

\*For correspondence: [akk002@ucsd.edu](mailto:akk002@ucsd.edu)

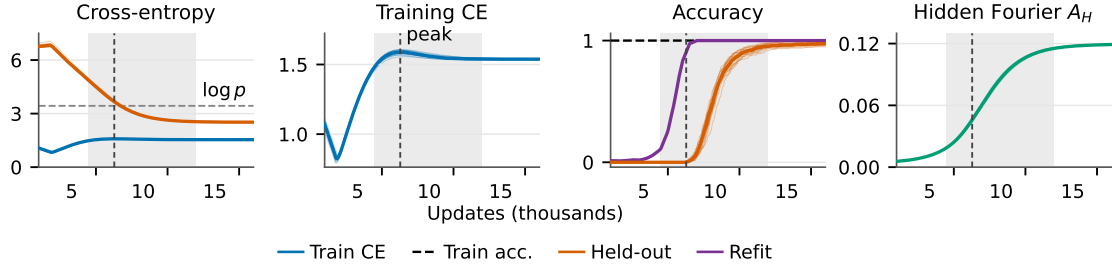


Figure 1: **Coexistence on a fixed window** (1HL ReLU network,  $p = 31$ , width 512, iid initialization, head decay  $\lambda_V = 0.2$ ). All 20 runs, each with a fresh initialization and training split (thin), and the median (thick). From the left: training and held-out CE; the training CE on its own scale; training accuracy (dashed) and the held-out accuracy of the network and of the train-only ridge refit; the Fourier fraction  $A_H$ . Gray: the locked window 4.5k–12k; dashed line: median per-run loss peak. Loss curves show cycle means. On the window, every run keeps perfect training accuracy, has raw per-update loss variation 6.33–7.81% and reaches 90% held-out accuracy. In every run, a few training pairs are misclassified again for a while after the first fit, long before the window (Appendix B.2).

generalization by the held-out accuracy of the network’s own head. Two observables describe the features. The Fourier fraction  $A_H$ , the share of the centered hidden-activation energy that lies in mean-zero functions of  $a + b$ , measures their structure. A ridge readout fitted on the training pairs alone, and never installed in the network, measures whether they already support held-out prediction (Section 3).

Figure 1 shows that the answer can be affirmative in practice. It follows a one-hidden-layer (1HL) ReLU network at  $p = 31$ , trained from fresh iid initializations and training splits, on a window fixed before the runs were drawn. On this window, every run keeps all training pairs correct, with raw training CE varying by less than 8%, while held-out accuracy climbs from zero to above 90%, the train-only refit reaches 90% more than 3,000 updates before the network, and  $A_H$  rises from about 0.02 to 0.12 (Section 5.1).

**Our Contributions.** For 1HL ReLU networks started in constructed open sets of initializations, we prove joint finite-window guarantees that place a relatively flat positive CE, a gain in  $A_H$  and held-out recovery on the same training trajectory. The hidden weights take supported-Polar steps, an exact and momentum-free form of the orthogonalized updates of Muon [27], and the head takes normalized gradient steps. Apart from weight decay, both steps have a fixed length for every nonzero gradient, so a flat loss does not imply that the weights have stopped moving.

- **High-loss coexistence (Theorem 4.1).** On a structured split of nearly half the grid, for every prime  $p \geq 101$  and sufficiently large width, the network keeps all training pairs correct with CE within 0.0201 of  $\log p$ , while  $A_H$  gains more than 0.0998 and held-out accuracy rises from zero to one.
- **Low-loss, refit-first recovery (Theorem 4.2).** For large primes and a random split, with probability at least  $1 - 2/\sqrt{p}$ , the CE stays within 5% of a prescribed positive level, which can be arbitrarily small, and  $A_H$  increases at every update; the train-only refit is accurate from the first update, and the network only at the end of the window.
- **Experiments (Section 5).** From iid initialization at practical widths, every fresh 1HL run shows this coexistence on a window fixed in advance, and networks with three to five hidden layers at  $p = 31$  meet a pre-specified joint criterion in nearly every run. At  $p = 47$  and 61, networks with three and four hidden layers meet a criterion fixed before the runs, on post-hoc windows, in nearly every run, but the joint criterion only rarely. A one-block transformer with the same updates and a slowly decaying output gain meets an analogous criterion, with a Fourier-filtered refit, in every run at  $p = 89$  and 113 while its training CE rises; with two blocks, the CE also rises with held-out accuracy, but training pairs then begin to flip.

Entry into the constructed initial regions from iid initialization is not proved; the experiments probe the behavior beyond these constructions. Section 2 reviews related work, Section 3 defines the setting, Section 4 states the theorems, Section 5 reports the experiments, and Section 6 discusses the limits of both. The appendices give the experimental details (Appendices A–D), the notation (Appendix N) and the proofs (Appendices E–G).

## 2 Related Work

**Grokking and modular arithmetic.** Proposed explanations of grokking include representation formation [10], weight-norm dynamics [5], competition between memorizing and more efficient circuits [11], transitions from lazy to rich training [6], and early and late implicit biases [7]. For shallow networks on modular addition, prior work analyzes margin maximization [8], the sample complexity of kernel and rich regimes [15] and, more recently, the training dynamics [16, 17, 18].

**Initialization scale and weight decay.** From a large initialization, a network fits its training set quickly, and weight decay then shrinks the weights slowly until the network generalizes [5]; with initialization scale  $\alpha$  and a small decay  $\lambda$ , the switch from a kernel predictor to a max-margin one comes after a time of order  $\lambda^{-1} \log \alpha$  [7]. With decoupled weight decay, the weights average the recent updates over about  $1/(\eta\lambda)$  steps, a timescale that works best when held at a fixed number of epochs [34], so in practice the weights can keep relaxing over much of training. Our networks share this ingredient: the one-hidden-layer experiments start from a large first layer ( $c = 24$ ) under weight decay, and the transformer’s output gain relaxes from twice its equilibrium (Section 5.3). These works follow the weight norm and the test loss as the training loss becomes small; we follow the training CE at a positive level together with the features, and bound both on finite windows.

**Fourier features and margin maximization.** Nanda et al. [2] identify circuit formation while the training loss remains flat, tracked by losses restricted to, and excluding, the key frequencies. Gromov [3] gives explicit Fourier solutions for shallow networks, and mechanistic studies identify several algorithms for arithmetic and group operations [13, 14]. Morwani et al. [8] prove that, on the full grid, sufficiently wide bias-free one-hidden-layer networks with quadratic activation and maximum  $L_{2,3}$ -margin consist of single-frequency neurons that together use every frequency; this describes the maximum-margin solution, not when the structure forms. Mallinar et al. [24] measure the circulant deviation of learned feature matrices in recursive feature machines and shallow networks. Our  $A_H$  is instead the fraction of hidden-activation energy in a task-aligned subspace, measured along the trajectory at positive loss. Even the maximum-margin network has  $A_H = 2/5$ , because squaring the sum of two single-frequency inputs of equal amplitude also produces terms in  $a - b$ ,  $2a$  and  $2b$ .

**Hidden progress and plateaus.** Barak et al. [4] prove hidden progress for online sparse-parity learning, and plateaus that precede fitting have also been analyzed [19, 18, 20]. In teacher-student regression, the leading eigenspace of the average gradient outer product aligns with the teacher subspace while the loss plateaus: under full-batch matrix Muon near the edge of stability, where the loss cycles with period two and its average stays flat or rises [31], and under population gradient descent from a small initialization, where a refit on the learned subspace lowers the error before the loss falls [21]. Here the task is classification with the CE, the plateau comes after the training set is fit, and its payoff is held-out accuracy. Related mechanisms include feature evolution after fitting under the squared loss with coupled weight decay [22] and late generalization with fixed features [23]. Recursive feature machines empirically grok at zero training loss [24], and probes reveal delays between representations and readouts [25]. Our constructions separate training confidence, the usefulness of the representation and the recovery of the readout. Their bounds are joint: on one finite window of the actual run, under fixed updates, they control the relative CE variation, the Fourier fraction and the predictions together, for every initialization in the constructed open sets. In these constructions, recovery need not wait for a near-zero CE, and the random-split theorem certifies that the refit recovers before the network without installing the refit as the head.

**Orthogonalized updates.** Our feature update rescales the supported Polar factor of the negative gradient to unit Frobenius norm and uses no momentum. Before rescaling, this factor maximizes the first-order decrease of the loss over the unit spectral-norm ball [28], which connects the update to spectral descent and Muon [26, 27, 9]. Orthogonalized updates, and updates that balance the singular directions, can shorten grokking delays [29, 30]. We study a different question: the change of the representation and the recovery of held-out accuracy during a window of relatively flat loss, rather than acceleration.

### 3 Problem Setup

#### 3.1 Task, model, and training objective

For an odd prime  $p$ , let  $\mathbb{F}_p = \{0, \dots, p-1\}$  with arithmetic modulo  $p$ , and let  $\Omega = \mathbb{F}_p^2$ . The task is modular addition [1]: a pair  $(a, b) \in \Omega$  is encoded as the concatenated one-hot vector  $x_{ab} = (e_a, e_b) \in \mathbb{R}^{2p}$  [3], where  $e_c$  is a standard basis vector, and its label is  $y^*(x_{ab}) = y_{ab} = a + b \pmod{p}$ . A set  $\mathcal{S} \subset \Omega$  of  $n$  training pairs is fixed, and  $\mathcal{D} = \Omega \setminus \mathcal{S}$  is the held-out set; we write  $x \in \mathcal{S}$  for  $x = x_{ab}$  with  $(a, b) \in \mathcal{S}$ , and similarly for  $\mathcal{D}$ . Each theorem specifies how  $\mathcal{S}$  is drawn.

The theory concerns a bias-free one-hidden-layer (1HL) ReLU network of width  $M$ , with hidden weights  $\mathbf{W} \in \mathbb{R}^{M \times 2p}$  and output head  $\mathbf{V} \in \mathbb{R}^{p \times M}$ , trained on the mean cross-entropy (CE):

$$\begin{aligned} h(x) &= \text{ReLU}(\mathbf{W}x), & f(x) &= \mathbf{V}h(x), \\ \text{CE}(f(x), y) &= \log \sum_{c \in \mathbb{F}_p} e^{f_c(x)} - f_y(x), \\ L(\mathbf{W}, \mathbf{V}) &= \frac{1}{n} \sum_{x \in \mathcal{S}} \text{CE}(f(x), y^*(x)). \end{aligned} \tag{1}$$

The network predicts  $\arg \max_c f_c(x)$ , with a fixed tie-breaking rule, and a subscript  $t$  denotes a quantity after  $t$  updates. The hidden table  $\mathbf{H}_t \in \mathbb{R}^{p^2 \times M}$  has the row  $h_t(x)^\top$  for each of the  $p^2$  inputs; it serves only to measure the features and does not enlarge the training set. A subscript  $B \subseteq \Omega$  restricts its rows, as in  $\mathbf{H}_{t, \mathcal{S}}$ . We write  $\mathbb{I}_N$  for the  $N \times N$  identity matrix and  $\mathbf{1}_N$  for the all-ones vector in  $\mathbb{R}^N$ . A network with  $D$  hidden layers computes  $h^{(\ell)} = \text{ReLU}(\mathbf{W}^{(\ell)} h^{(\ell-1)} + b^{(\ell)})$  for  $\ell = 1, \dots, D$ , with  $h^{(0)} = x$ , and  $f = \mathbf{V}h^{(D)}$ ; layer  $\ell$  has the table  $\mathbf{H}_t^{(\ell)}$ . The baseline networks have no biases; the dotted medians in Figures 2 and 3 come from runs with trained hidden biases (Appendix C.6).

#### 3.2 Fixed normalized updates

We train the hidden weights with a supported-Polar update and the head with normalized gradient descent. Both use the negative gradients of the training CE at the same state,  $\mathbf{F}_t = -\nabla_{\mathbf{W}} L(\mathbf{W}_t, \mathbf{V}_t)$  and  $\mathbf{G}_t = -\nabla_{\mathbf{V}} L(\mathbf{W}_t, \mathbf{V}_t)$ . For a nonzero matrix  $\mathbf{F}$  with thin singular value decomposition  $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{Z}^\top$  over its positive singular values, the supported Polar factor is  $\text{pol}(\mathbf{F}) = \mathbf{U}\mathbf{Z}^\top$ , and  $\text{pol}(0) = 0$ . The two normalizations are

$$\mathcal{N}_{\text{pol}}(\mathbf{F}) = \frac{\mathbf{U}\mathbf{Z}^\top}{\sqrt{\text{rank } \mathbf{F}}}, \quad \mathcal{N}_F(\mathbf{G}) = \frac{\mathbf{G}}{\|\mathbf{G}\|_F}, \tag{2}$$

with  $\mathcal{N}_{\text{pol}}(0) = \mathcal{N}_F(0) = 0$ ; for a nonzero argument, each returns a matrix of unit Frobenius norm. The Polar normalization moves the weights equally along every singular direction of the gradient, whereas  $\mathcal{N}_F$  preserves the ratios of the singular values. With fixed rates  $\eta_W, \eta_V > 0$  and decays  $\lambda_W, \lambda_V \geq 0$ , the updates are

$$\begin{aligned} \mathbf{W}_{t+1} &= (1 - \eta_W \lambda_W) \mathbf{W}_t + \eta_W \mathcal{N}_{\text{pol}}(\mathbf{F}_t), \\ \mathbf{V}_{t+1} &= (1 - \eta_V \lambda_V) \mathbf{V}_t + \eta_V \mathcal{N}_F(\mathbf{G}_t). \end{aligned} \tag{3}$$

Each update shrinks the weights by a fixed factor and adds a step of fixed length, with no momentum, projection or loss-dependent control. For every nonzero gradient, however small, the step length is the same; the weights can therefore keep moving while the loss is flat, although the fixed step by itself guarantees neither a flat loss nor useful motion. The reported loss  $L$  is the CE alone, without decay penalties. The theorems use  $\lambda_V = 0$ , and each experiment states its head decay. In the experiments, the positive singular values are selected with a numerical cutoff (Appendices B and C). Algorithm 1 states the training loop for  $D$  hidden layers, each with its own rate and decay; with one hidden layer it reduces to (3), with  $(\eta_1, \lambda_1) = (\eta_W, \lambda_W)$ . With trained biases,  $[\mathbf{W}^{(\ell)} \mid b^{(\ell)}]$  takes the feature step (Appendix C.6). In the transformer of Section 5.3, every weight matrix except the unembedding takes the feature step, the unembedding takes the head step, and a scalar output gain takes a sign step with decay. The Fourier fraction and the refit defined below are computed from the weights and never enter the updates.

---

**Algorithm 1** Training with supported-Polar feature steps and normalized head steps

---

**Input:** training pairs  $\mathcal{S}$ ,  $n = |\mathcal{S}|$ ; initial hidden matrices  $\mathbf{W}_0^{(1)}, \dots, \mathbf{W}_0^{(D)}$  and head  $\mathbf{V}_0$   
**Parameters:** rates  $\eta_1, \dots, \eta_D, \eta_V > 0$ ; decays  $\lambda_1, \dots, \lambda_D, \lambda_V \geq 0$ ; number of updates  $T$   
**Zero gradients:**  $\mathcal{N}_{\text{pol}}(0) = \mathcal{N}_F(0) = 0$ , as in (2), so a matrix whose gradient is zero only decays

- 1: **for**  $t = 0, 1, \dots, T-1$  **do**
- 2:    $L_t \leftarrow \frac{1}{n} \sum_{x \in \mathcal{S}} \text{CE}(f_t(x), y^*(x))$   $\triangleright$  full batch: every training pair
- 3:    $\mathbf{F}_t^{(\ell)} \leftarrow -\nabla_{\mathbf{W}^{(\ell)}} L_t$  for  $\ell = 1, \dots, D$ , and  $\mathbf{G}_t \leftarrow -\nabla_{\mathbf{V}} L_t$   $\triangleright$  all at the same state
- 4:   **for**  $\ell = 1, \dots, D$  **do**
- 5:      $\mathbf{U}\Sigma\mathbf{Z}^\top \leftarrow$  thin SVD of  $\mathbf{F}_t^{(\ell)}$  over its positive singular values
- 6:      $\mathbf{W}_{t+1}^{(\ell)} \leftarrow (1 - \eta_\ell \lambda_\ell) \mathbf{W}_t^{(\ell)} + \eta_\ell \mathbf{U}\mathbf{Z}^\top / \sqrt{\text{rank } \mathbf{F}_t^{(\ell)}}$   $\triangleright \eta_\ell \mathcal{N}_{\text{pol}}(\mathbf{F}_t^{(\ell)})$ , eq. (2)
- 7:    $\mathbf{V}_{t+1} \leftarrow (1 - \eta_V \lambda_V) \mathbf{V}_t + \eta_V \mathbf{G}_t / \|\mathbf{G}_t\|_F$   $\triangleright \eta_V \mathcal{N}_F(\mathbf{G}_t)$

---

### 3.3 Fitting, plateaus, and delayed recovery

The margin of an input  $x$  at time  $t$  is  $m_t(x) = f_{t,y^*(x)}(x) - \max_{c \neq y^*(x)} f_{t,c}(x)$ , and  $\text{Acc}_B(t)$  is the fraction of inputs  $x \in B$  with  $\arg \max_c f_{t,c}(x) = y^*(x)$ . The network has *fitted* the training set when every training margin is positive. Fitting does not require a small loss: correct predictions with small positive margins can have a loss arbitrarily close to  $\log p$ , the loss of the uniform prediction, so a fitted network can have any training CE in  $(0, \log p)$ . Flatness on an integer window  $I = \{t_0, \dots, t_1\}$  is measured by the relative variation of the raw training CE,

$$v(I) = \frac{\max_{t \in I} L_t}{\min_{t \in I} L_t} - 1, \quad (4)$$

evaluated at every update rather than on plotted or averaged losses. A 10% plateau is a window with  $v(I) \leq 0.1$ ; it allows small changes and oscillations and claims neither stationarity nor optimality. Some experiments also use the cycle mean  $\bar{L}_t = (L_t + L_{t+1})/2$ , the average over two consecutive updates. For an observable  $A$ ,  $\Delta A(I) = A(t_1) - A(t_0)$  is its change over the window. *Delayed recovery* means that held-out accuracy becomes high after a phase of fitted training with low held-out accuracy. Our coexistence claims place the change of the features and the recovery on the same flat window and assert nothing after it.

### 3.4 The Fourier fraction

Since the target depends only on  $a + b$ , we measure the features against the *addition-Fourier subspace*, the real mean-zero functions on the grid that depend only on the sum:

$$\mathcal{F}_{\text{add}} = \{g(a + b) : \sum_{y \in \mathbb{F}_p} g(y) = 0\}. \quad (5)$$

It is spanned by the centered class indicators  $\mathbf{1}_{\{a+b=c\}} - 1/p$ , and also by the real and imaginary parts of the Fourier modes  $e^{2\pi i k(a+b)/p}$ ,  $k \neq 0$ ; its dimension is  $p - 1$  (Appendix A.1). Let  $\mathbf{P}_{\text{add}}$  be the orthogonal projector onto  $\mathcal{F}_{\text{add}}$ , and let  $\mathbf{C}_N = \mathbb{I}_N - \mathbf{1}_N \mathbf{1}_N^\top / N$  and  $\mathbf{C}_{\text{grid}} = \mathbf{C}_{p^2}$ . Then  $\mathbf{P}_{\text{add}} \mathbf{H}$  averages each hidden feature over the pairs with the same sum and subtracts its mean over the grid. The *Fourier fraction* of a table  $\mathbf{H}$  is the share of its centered energy that lies in  $\mathcal{F}_{\text{add}}$ :

$$A_H(\mathbf{H}) = \frac{\|\mathbf{P}_{\text{add}} \mathbf{H}\|_F^2}{\|\mathbf{C}_{\text{grid}} \mathbf{H}\|_F^2} \in [0, 1], \quad (6)$$

defined when  $\mathbf{C}_{\text{grid}} \mathbf{H} \neq 0$ . It equals one when the centered response of every neuron is a function of  $a + b$ , and zero when no response has a component in  $\mathcal{F}_{\text{add}}$ . It is invariant to scaling  $\mathbf{H}$  by a nonzero constant, and it can rise because the energy in  $\mathcal{F}_{\text{add}}$  grows or because the remaining energy shrinks. Since it weights neurons by their amplitude, it can also change without a change in the span of the features. A gain of 0.10 in  $A_H$  means ten percentage points, not a 10% relative increase. For deep networks, we compute it for each table  $\mathbf{H}_t^{(\ell)}$ .

### 3.5 The train-only refit

A high Fourier fraction does not by itself imply that the features support correct predictions. At each snapshot  $\mathbf{H}_t$ , we therefore fit a separate ridge head on the training pairs alone:

$$\hat{\mathbf{B}}_t = \arg \min_{\mathbf{B} \in \mathbb{R}^{M \times p}} \left\{ \|\mathbf{H}_{t,\mathcal{S}} \mathbf{B} - \mathbf{T}_{\mathcal{S}}\|_F^2 + \tau_t \|\mathbf{B}\|_F^2 \right\}, \quad (7)$$

where  $\mathbf{T}$  has the rows  $T_x = e_{y^*(x)} - \mathbf{1}_p/p$  and  $\tau_t > 0$ ; the features are raw, and there is no intercept. The refitted logits are  $\hat{f}_t(x) = \hat{\mathbf{B}}_t^\top h_t(x)$ , computed from the last hidden table in deep networks. The refit is evaluated on  $\mathcal{D}$  but never installed in the network, and the training run is not modified. It measures what ridge regression extracts from the features, which can differ from what the network’s own head predicts. In the sum-loss units of (7), the penalty rule is fixed in advance: with  $R_t = \|\mathbf{H}_{t,\mathcal{S}}\|_F^2$ ,

$$\tau_t = \begin{cases} 10^{-3}n, & \text{canonical experiments,} \\ R_t/(32n), & \text{Theorem 4.1,} \\ \kappa R_t/[4p(p-1)], & \text{Theorem 4.2,} \end{cases} \quad (8)$$

where  $\kappa$  is the constant of Theorem 4.2. Appendix A.2 gives the Gram form and the experimental variants. Theorem 4.2(iii) measures the refit by its normalized held-out squared error,

$$E_{\text{ref}}(t) = \frac{\|\mathbf{H}_{t,\mathcal{D}} \hat{\mathbf{B}}_t - \mathbf{T}_{\mathcal{D}}\|_F^2}{|\mathcal{D}|(p-1)/p}, \quad (9)$$

which equals one for the zero predictor, since  $\|T_x\|_2^2 = (p-1)/p$ .

## 4 Main Theoretical Results

The two theorems below prove that coexistence of the kind seen in Figure 1 can occur under the actual CE training dynamics of a 1HL network, started from constructed initializations. In both, the two layers follow the updates (3) with  $\lambda_V = 0$ . Under each theorem’s conditions on the split, every trajectory from a specified, possibly small, open set of initializations keeps perfect training accuracy and recovers held-out accuracy during a CE plateau; entry into this set from iid initialization is not established. The first theorem holds at a loss near  $\log p$  on a structured split, and the second at a prescribed positive loss, which can be made arbitrarily small, on a random split.

### 4.1 A constant Fourier gain on a structured half-grid

The first construction isolates coexistence at high training loss: the network classifies every training pair correctly while its CE stays just below  $\log p$ .

**The split.** The training set consists of the pairs whose difference is a nonzero quadratic residue, that is, a value  $u^2 \bmod p$  with  $u \neq 0$ :

$$\mathcal{S} = \{(a, b) : b - a \in \text{QR}(\mathbb{F}_p)\}, \quad (10)$$

where  $\text{QR}(\mathbb{F}_p)$  is the set of nonzero quadratic residues. It contains  $p(p-1)/2$  pairs, just under half the grid, and each operand and each label occurs  $(p-1)/2$  times in it. This balance supports the Fourier analysis, but the split selects whole difference classes rather than independent pairs, and it holds out every diagonal pair. The theorem requires width at least  $M_{\min}(p) = p(p^2 + 3)/2$ .

**Idea of the construction.** We first analyze a reference run in which the head is frozen and the hidden weights take Polar steps with decay. Its hidden units form three groups. An *arithmetic core* classifies every training pair correctly by a small margin, and its responses and training logits stay fixed along the run. *Helper units* have zero head weights, so they affect no prediction, and they have no component in  $\mathcal{F}_{\text{add}}$ . *Query units*

are active only on held-out inputs, where their logits are initially harmful. As the helper and query units decay, the training logits do not move, so every training pair stays correct and the CE stays near  $\log p$ . On held-out inputs, the harmful query logits decay until the correctly signed contribution of the core dominates, and held-out accuracy goes from zero to one. The decay of the helper units removes energy outside  $\mathcal{F}_{\text{add}}$ , which, with bounds for the other two groups, yields a net gain in  $A_H$ . These bounds then carry over to the actual run, in which both layers train.

**Theorem 4.1** (Structured half-grid).

**Parameters.** Let  $p \geq 101$  be prime, and let the width satisfy  $M \geq p(p^2 + 3)/2$ .

**Split.** The training set is  $\mathcal{S} = \{(a, b) : b - a \text{ is a nonzero quadratic residue modulo } p\}$ .

There exist fixed rates and decay  $(\eta_W, \lambda_W, \eta_V)$ , a horizon  $J$ , sets  $\mathcal{I}_{\text{early}}, \mathcal{I}_{\text{late}} \subseteq \{1, \dots, J\}$  of size at least 16 with  $\max \mathcal{I}_{\text{early}} < \min \mathcal{I}_{\text{late}}$ , and a nonempty open set of initializations in  $\mathbb{R}^{3pM}$  such that every trajectory of (3) with  $\lambda_V = 0$  started in this set satisfies all four conclusions below.

- (i) **Training fit and CE plateau.** For every  $0 \leq t \leq J$ , training accuracy is one and

$$\log p - 0.0201 < L_t < \log p.$$

Consequently,  $v([0, J]) < 0.0045$ .

- (ii) **Early held-out failure.** For every  $t \in \mathcal{I}_{\text{early}}$ , the held-out accuracies of the network and of the refit are zero.

- (iii) **Late network recovery and Fourier gain.** For every  $t \in \mathcal{I}_{\text{late}}$ , the actual trained network's held-out accuracy is one and

$$A_H(\mathbf{H}_t) - A_H(\mathbf{H}_0) > 0.0998.$$

- (iv) **Late refit margin.** If moreover  $p \geq 1764$ , then for every  $t \in \mathcal{I}_{\text{late}}$  the refit has margin greater than  $11/50$  against every rival class at every held-out pair whose difference is a nonsquare, so its held-out accuracy is at least  $(p - 1)/(p + 1)$ .

**Interpretation.** Recovery and a net gain of more than 9.98 percentage points in  $A_H$  thus occur within one plateau that is checked at every update. The theorem asserts neither monotone growth of  $A_H$  nor persistence after  $J$ , and it does not order the recoveries of the refit and the network; clause (iv) says nothing about the refit on the diagonal pairs.

**Proof outline.** The proof is in Appendix E and is assembled in Appendix E.4. In the reference run, every training contrast  $m_{t,c}(x) = f_{t,y^*(x)}(x) - f_{t,c}(x)$  stays in  $(0, 0.0201)$ , so the identity

$$\text{CE}(f_t(x), y^*(x)) = \log \left( 1 + \sum_{c \neq y^*(x)} e^{-m_{t,c}(x)} \right) \quad (11)$$

gives perfect training classification with CE near  $\log p$ . Quantitative bounds on the three groups of units control the switch of the held-out predictions and the net gain in  $A_H$ , and the refit is analyzed separately, from the full ridge problem with the prescribed penalty. Strict signs of the training gates, a locally constant rank of the supported gradient and a nonzero head gradient give continuity near the reference trajectory. Over the finite horizon, the strict bounds therefore carry over to the actual run for a small enough positive head step and an open set of initializations.

## 4.2 Random splitting and refit-first recovery

The second construction removes both restrictions of the first, the engineered split and the loss near  $\log p$ . Each pair enters the training set independently with probability  $4/5$ , a value set by our proof estimates rather than a known threshold, and the training CE stays near a prescribed positive level  $L_*$ , set by  $a_{\text{conf}}$  below, which can be made arbitrarily small and is not the minimum of the CE. The refit becomes accurate early, and the network recovers late.

**Idea of the construction.** The initialization pairs a fixed rectangular dictionary of hidden units, which already spans every function on the grid, with a head calibrated on the training data. Along a reference feature trajectory with fixed decay and a rule-directed update, the rates balance decay against the growth of the useful component, which keeps the leading training confidence, and with it the CE, stable. On held-out inputs, however, the competing logit contributions change their balance, and the network’s held-out accuracy becomes high only at a prescribed time  $J$ . Ridge fitting on the training rows selects an accurate predictor from the first update on. Since the initial features already span every function on the grid, the refit improves because it selects a better predictor, not because the features gain new span.

**Theorem 4.2** (Random split).

**Parameters.** Fix an integer  $J \geq 30$ , constants  $\alpha, \theta, a_{\text{conf}} > 0$ ,  $\beta \in [1/2, 1)$ ,  $\kappa \in (0, 1/8]$ , and tolerances  $\epsilon_{\text{lag}}, \epsilon_{\text{test}}, \epsilon_{\text{obs}} \in (0, 1)$ . Set  $\rho = 4/5$  and  $L_* = \log(1 + \rho^2 e^{-a_{\text{conf}}})$ .

**Sampling and initialization.** There is a threshold  $p_0$  such that the following holds for every odd prime  $p \geq p_0$ . Use width  $M = p^2(p-1)^2$  and the prescribed rates and decay  $(\eta_W, \lambda_W, \eta_V)$ , chosen before sampling. Place each pair in  $\mathcal{S}$  independently with probability  $\rho$ .

With probability at least  $1 - 2/\sqrt{p}$  over the split, the split is proper and there is a nonempty open ball in  $\mathbb{R}^{3pM}$  around a prescribed initialization built from the training data such that every trajectory of (3) with  $\lambda_V = 0$  started in this ball satisfies all four conclusions below.

(i) **Training fit and CE plateau.** For every  $0 \leq t \leq J$ , training accuracy is one and

$$\frac{20}{21}L_* < L_t < \frac{22}{21}L_*.$$

Consequently,  $v([0, J]) < 0.1$ .

(ii) **Delayed network recovery.** The network’s held-out accuracy is 0 at  $t = 0$ , below  $\epsilon_{\text{lag}}$  for  $0 \leq t < J$ , and above  $1 - \epsilon_{\text{test}}$  at  $t = J$ .

(iii) **Refit accuracy and error.** The refit’s held-out accuracy is 0 at  $t = 0$  and at least  $1 - \epsilon_{\text{obs}}$  for  $1 \leq t \leq J$ . Its normalized error satisfies

$$E_{\text{ref}}(0) - E_{\text{ref}}(t) \geq \frac{401}{1568} > \frac{1}{4}, \quad t \in \{J-1, J\}.$$

(iv) **Stepwise Fourier improvement.** For every  $0 \leq t < J$ ,

$$A_H(\mathbf{H}_{t+1}) - A_H(\mathbf{H}_t) > \frac{1}{24J(3p+1)}.$$

**Parameter choices.** Here  $\alpha$  scales the initial dictionary,  $\beta$  its offsets and  $\theta$  the head step. For each sufficiently large prime, the width and the rates (119) are chosen before sampling, and the initial ball depends on the realized training data. The probability is over the split only and says nothing about entry from iid weights. Increasing  $a_{\text{conf}}$  makes  $L_*$  arbitrarily small but may change the uncalibrated threshold  $p_0$ , and changing  $J$  changes the initialization and the rates rather than extending one fixed run.

**Interpretation.** Both predictors start with zero held-out accuracy. If  $\epsilon_{\text{lag}} + \epsilon_{\text{obs}} < 1$ , clauses (ii) and (iii) give

$$\text{Acc}_{\mathcal{D}}(t) < \epsilon_{\text{lag}} < 1 - \epsilon_{\text{obs}} \leq \text{Acc}_{\text{ref}, \mathcal{D}}(t), \quad 1 \leq t < J, \quad (12)$$

where  $\text{Acc}_{\text{ref}, \mathcal{D}}$  is the refit’s held-out accuracy: the refit is accurate from the first update, and the network recovers only at  $J$ , within the same CE plateau. The error gain in (iii) compares the times  $J-1$  and  $J$  with the initial error of the same trajectory; it is neither a per-step decrease of the error nor a 25-percentage-point accuracy gain. Clause (iv) gives a Fourier gain at every update, although  $A_H$  remains  $O(1/p)$  in this construction.

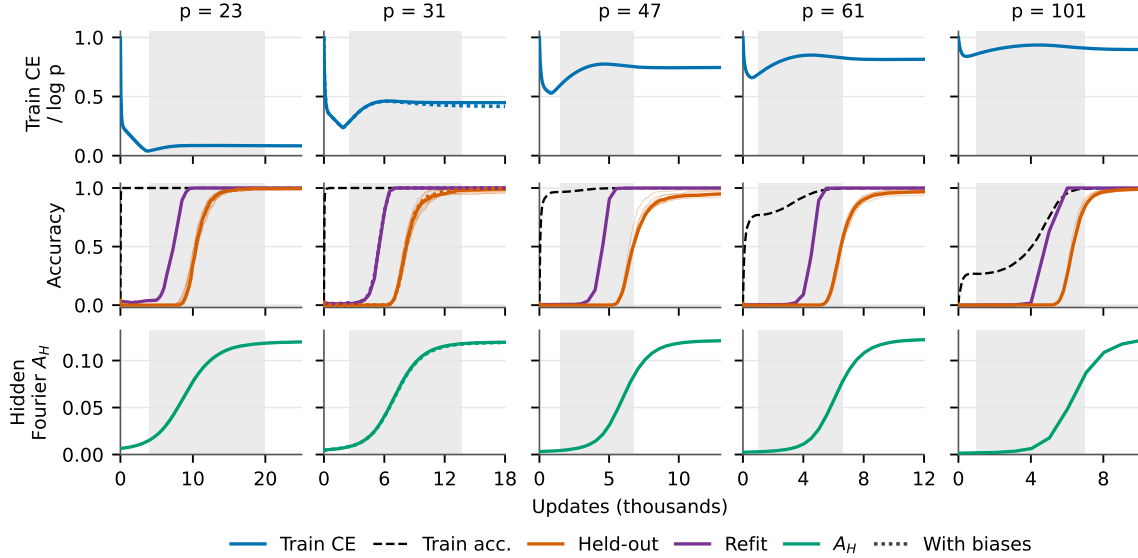


Figure 2: **One hidden layer across primes** (width 512, 10 runs per prime;  $\lambda_W = 0.01, 0.015, 0.02, 0.02, 0.02$  at  $p = 23, 31, 47, 61, 101$ , chosen on development seeds, for  $p = 101$  at  $p = 97$ ). Rows: cycle-mean training CE divided by  $\log p$ ; median training accuracy (dashed), held-out accuracy of the network (runs and median) and of the train-only refit (median);  $A_H$ . Gray: from the median start to the median end of the runs’ post-hoc windows (Appendix A.3). The primary endpoint of Section 5.2 is met by 9, 4, 0, 0 and 0 runs; at  $p \geq 47$  some training pairs stay misclassified inside the window. Runs stopped early by the stopping rule keep their last value. Dotted ( $p = 31$ ): medians with a hidden bias (Appendix C.6).

**Proof outline.** Appendix F constructs the reference feature trajectory. Its rates keep the leading training confidence stable, signed-margin estimates on held-out inputs place the network’s recovery at the prescribed time  $J$ , and an analysis of the complete feature kernel with the prescribed ridge penalty gives accurate refit predictions from time 1. Uniform error bounds transfer these estimates to the actual updates for every initialization in the prescribed ball (Appendix F.4). Both layers train, but the small head step makes the recovery mainly feature-driven.

## 5 Experiments

In this section, we discuss our experimental verification of whether qualitative counterparts of the coexistence proved in Section 4 appear outside the assumptions of its theorems. The theorems start from constructed initializations, at widths that grow with  $p$ ; the experiments start from iid initialization at practical widths, in 1HL networks (Section 5.1), in deeper ReLU networks (Section 5.2) and in transformers with one and two blocks (Section 5.3). Appendix B.6 also runs the split and the updates of Theorem 4.1 from iid initialization at width 512. Appendices A–C give the window definitions, protocols and further results, including stricter evaluations that require the first training fit to persist (Appendices B.2 and C.2); Appendix D details the transformer.

The runs of Figures 1–3 use head decay  $\lambda_V = 0.2$ ; the deeper networks across training fractions use  $\lambda_V = 0.2 \cdot 31/p$ , and the transformers a value set per prime. The theory allows head decay only under the construction-specific sufficient conditions of Remark G.4, which in both constructions keep the cumulative decay  $J\eta_V\lambda_V$  below one; the cumulative decay on the window of Figure 1,  $7500 \cdot 0.03 \cdot 0.2 = 45$ , lies outside them.

**Windows fixed in advance and post-hoc windows.** Since a statement about a window is only as reliable as the rule that chooses it, the primary counts use a window, or a band-selection rule, fixed before evaluation: the locked interval of Figure 1 for one hidden layer, and for deeper networks a band that a loss-only rule selects in each run (Section 5.2). Figures 2 and 3 instead shade looser post-hoc non-falling

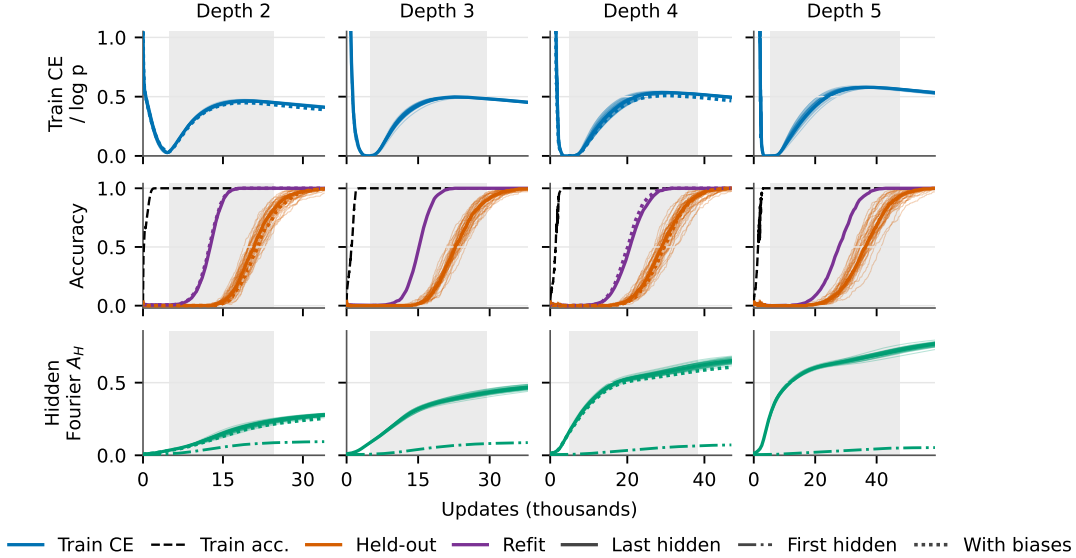


Figure 3: **Two to five hidden layers at  $p = 31$ : every run (thin) and the median (thick).** Width 256. Depth 2 uses a fixed rate set on development runs; depths 3–5 tune the last-layer rate on a preliminary run (Appendix C.1); every depth shows the same 40 evaluation seeds. The primary endpoint is met by 34, 40, 39 and 38 runs; the six depth-2 misses fail the share threshold, and one of them also ends with held-out accuracy below 0.99. Rows as in Figure 2, with the last-hidden-layer  $A_H$  (solid) and the median first-hidden-layer  $A_H$  (dash-dot). Gray: from the median start to the median end of the runs’ post-hoc windows (Appendix A.3), which contain the rise in loss after the knee. One hidden layer at  $p = 31$  is the second column of Figure 2. Dotted (depths 2, 4): medians with hidden biases.

windows. Such a window lies after the head norm has nearly reached the ceiling  $1/\lambda_V$  set by its decay, and it is the stretch over which the cycle-mean loss rises the most while never falling more than a factor 1.03 below its running maximum (Appendix A.3). These windows make no flatness claim, enter only the secondary analysis (Appendix C.5) and do not determine the primary counts. Figure 4 shades the same windows, which there also enter a criterion fixed before its runs (Appendix C.4); the transformer has no locked window and is scored on a post-hoc window of its own (Appendix D.2). Every plotted cohort shows all its runs, except in the two appendix figures on the transformer’s mirror pairs and readouts, which show medians.

## 5.1 One-hidden-layer ReLU networks

**Protocol.** Figure 1 shows 20 runs at  $p = 31$  and width  $M = 512$ , each with a fresh initialization and a fresh uniform random 80% split (768 training and 193 held-out pairs). The weights are iid Gaussian,  $W_{ij} = cZ_{ij}/\sqrt{p}$  and  $V_{kj} = \tilde{Z}_{kj}/(c\sqrt{M})$ , with  $c = 24$  and independent standard Gaussian arrays  $\mathbf{Z}, \tilde{\mathbf{Z}}$ . Training applies the updates (3), full batch in float64, for 40,000 updates with fixed  $\eta_W = \eta_V = 0.03$ ,  $\lambda_W = 0.015$  and  $\lambda_V = 0.2$ . A second development search held  $c$ , both rates and the head decay fixed and swept the prime, the training fraction and the feature decay. A predeclared rule that uses held-out accuracy selected this configuration and the window  $I = [4500, 12000]$ , and both were locked before the 20 initializations and splits were drawn (Appendix B), so the evaluated runs played no part in choosing them.

**Results.** All 20 runs classify every training pair correctly at every update of  $I$ . The training loss first dips to about 0.8 and rises to about 1.5; on  $I$ , it stays within the declared 10% plateau tolerance, with minimum 1.456–1.498 and  $v(I) = 6.33$ –7.81%. Over  $I$ ,  $A_H$  rises from 0.018–0.020 to 0.115–0.116, and held-out accuracy climbs from zero to 91.7–98.4%. The features support an accurate train-only readout before the network’s own head is accurate: the refit first records 90% at update 6,500 in every run, whereas the network first crosses 90% at updates 9,631–11,665. The window is not anchored at the first fit of the training set. In every run, a few training pairs are misclassified again for a while after the first fit, long before the window, so the stricter endpoint that requires the fit to persist from its first attainment fails (Appendix B.2). Pre-registered

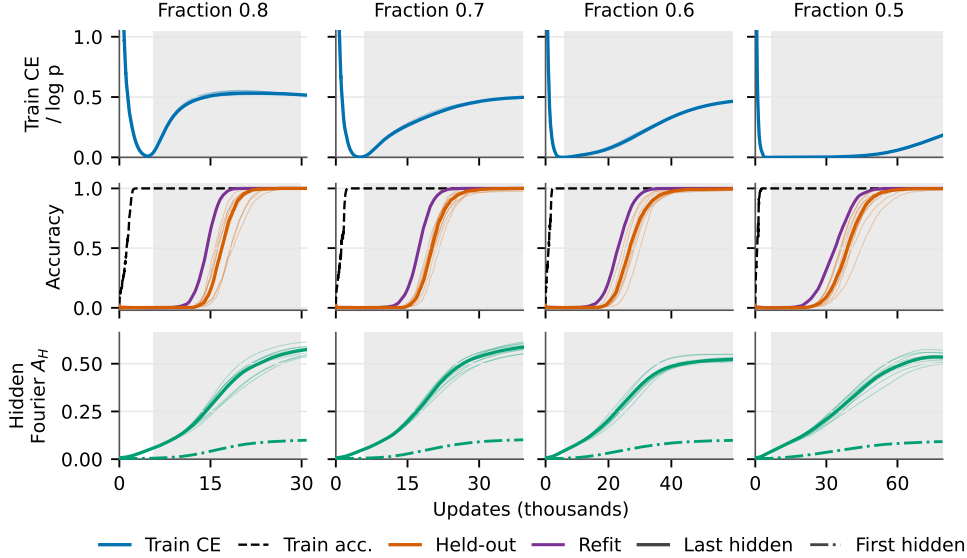


Figure 4: **Depth 3 at  $p = 47$  across training fractions: every run (thin) and the median (thick).** Width 256, 10 fresh runs per fraction, last-layer multiplier  $r = 0.45, 0.30, 0.18$  and  $0.08$  chosen per fraction on development runs, head decay  $\lambda_V = 0.2 \cdot 31/p$  (Appendix C.4). Rows and gray bands as in Figure 3. The criterion of Appendix C.4, fixed before the runs, is met by 10, 10, 10 and 8 runs; the two misses end at held-out accuracy 0.959 and 0.971 when they reach the update budget. No training pair is misclassified at any log inside any window. Each column ends no later than its shortest run.

tests of recovery without a decrease of the CE, at  $p = 23$ , are in Appendix B.3.

**Other primes.** Figure 2 shows separate 1HL cohorts of 10 runs at  $p = 23, 31, 47, 61$  and  $101$ . The feature decay was chosen on one or two development seeds per prime (for  $p = 101$ , at  $p = 97$ ), so these cohorts are not a pre-specified evaluation. The primary endpoint of Section 5.2 is met by 9, 4, 0, 0 and 0 runs; at  $p \geq 47$ , some training pairs stay misclassified inside the window.

## 5.2 Extension to deeper ReLU networks

**Protocol.** We train networks with  $D = 2$  to 5 hidden layers of width 256 at  $p = 31$ , from iid Gaussian initialization (Appendix C), with supported-Polar hidden updates and fixed decays 0.2 for the head, 0.015 for the first hidden layer and 0.07 for the later layers. For  $D = 3, 4, 5$ , only the last-layer rate is tuned, by a pre-specified rule applied to a preliminary run, after which training restarts from the same initialization and split with that rate fixed (Appendix C.1). Depth 2 uses a fixed rate chosen on development runs. The evaluation seeds are disjoint from development and shared across depths, depth 2 included, so the cohorts are dependent.

**The primary endpoint.** Deeper networks have no locked window. Instead, a rule that uses no held-out outcome selects a band  $I_C$  around the CE peak: after the knee, the first checkpoint at which the head norm reaches 99% of its ceiling  $1/\lambda_V$ , it takes the largest cycle-mean loss and the maximal interval around it on which the cycle-mean loss stays at least that peak divided by 1.03 (Appendix A.3). The pre-specified primary endpoint is a joint criterion with four parts:  $I_C$  contains at least 75% of the rise of held-out accuracy from 10% to 90%; the final held-out accuracy is at least 0.99; the Fourier fraction of the last hidden layer gains at least 0.02 across  $I_C$ ; and every training margin is positive at every 20-update log in  $I_C$ . A depth succeeds if the lower end of the two-sided 95% Wilson interval for its pass rate is at least 0.80, that is, if at least 37 of its 40 runs pass.

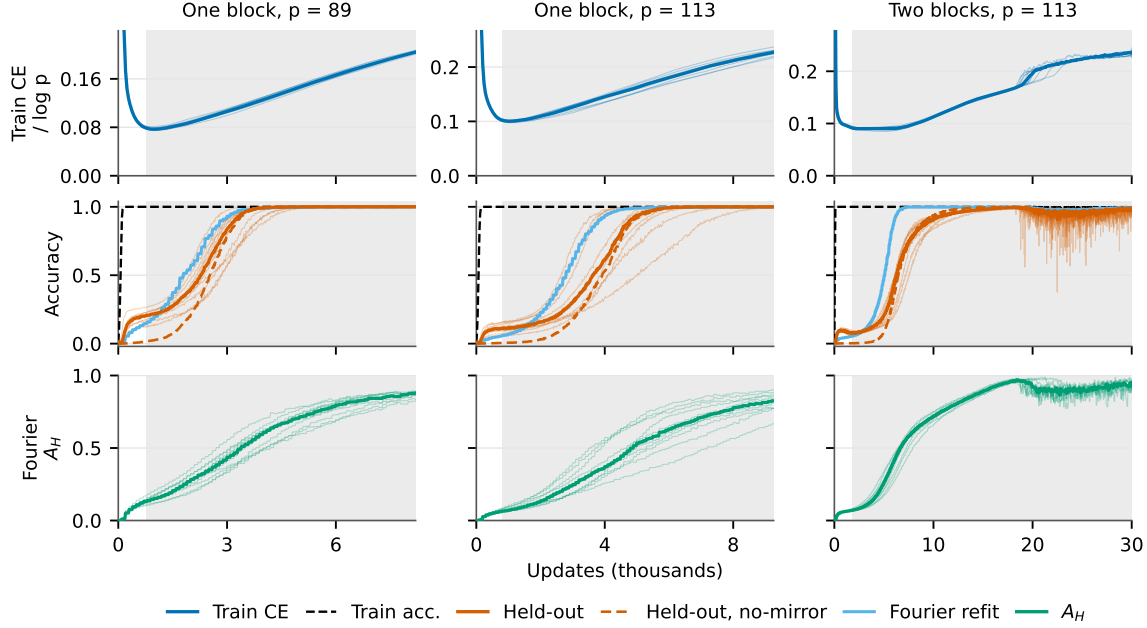


Figure 5: **Transformers with one and two blocks: every run (thin) and the median (thick).** One block at  $p = 89$  (training fraction 0.25,  $\lambda_V = 0.33$ ) and at  $p = 113$  (0.2,  $\lambda_V = 0.275$ ), with the output gain decaying as  $g_t \approx 1 + e^{-t/5000}$ , and two blocks at  $p = 113$  (0.2,  $\lambda_V = 0.275$ ,  $g_t \approx 1 + e^{-t/10000}$ ); 10 runs per column (seeds  $i = 11$ –20). Rows as in Figure 2, with the training CE averaged over 201 updates (each column on its own scale) and  $A_H$  of the normalized head input  $\hat{r}$ . The accuracy row adds the two medians used by the pass rule of Section 5.3: held-out accuracy on the no-mirror pairs (dashed) and the Fourier-filtered refit, which replaces the refit of Figure 2. The refit, not drawn, trails the network at its 50% crossing in all 30 runs, by a median of 0.7k, 1.2k and 0.5k updates in the three columns. Gray: from the median start to the median end of the post-hoc windows on all held-out pairs (Appendix D.2). Each column ends at its shortest run. One block passes in 10 of 10 runs at both primes; its runs stop at 8.2k–10.2k updates ( $p = 89$ ) and 9.3k–14k ( $p = 113$ ). Two blocks pass in none. Through the held-out rise they behave like one block: the CE rises and every training pair is correct at every 20-update log. The first training errors are logged at 17.8k–22.0k updates, 1.4k to 5.7k updates after held-out accuracy first reaches 0.99; held-out accuracy then falls back and jitters, and 6 of the 10 runs end below 0.99 (Appendix D.3, Figure 15).

**Results at  $p = 31$ .** The primary endpoint is met by 40/40, 39/40 and 38/40 runs at  $D = 3, 4, 5$ , so these depths succeed, and by 34/40 at depth 2, whose Wilson lower end, 0.71, is below 0.80 (Figure 3). The selected bands of these cohorts have cycle-mean loss variation at most 3% (Appendix A.3). All 120 runs at depths 3–5 finish with perfect held-out accuracy, and 118 have raw-loss variation at most 10% across the first logged 10%-to-90% transition.

The recipe does not transfer unchanged: with the same recipe, changing the prime to 23, the width to 512 or the feature update to a normalized gradient gives no joint successes. The normalized-gradient runs still generalize, but their loss keeps falling during the recovery (Appendix C.2). Appendix C.3 reports further fixed settings, selected on development runs, with depths 3–4 at  $p = 23, 47, 61, 97, 101$  and depth 2 at  $p = 23, 47$ , and separates primary from secondary post-hoc outcomes.

**Other primes and training fractions.** A re-tuned recipe carries the post-hoc behavior to other primes and smaller training sets. With the last-layer multiplier  $r$  set per training fraction on development runs and the head decay scaled with the prime,  $\lambda_V = 0.2 \cdot 31/p$ , depths 3 and 4 at  $p = 47$  and 61, with training fractions 0.5 to 0.8, meet a criterion on the post-hoc non-falling window, fixed before the runs, in 155 of 160 fresh runs (Appendix C.4). A run passes if the window holds at least 75% of both the network’s and the refit’s 10%-to-90% held-out rises, the refit gains at least 0.2 and the last-layer  $A_H$  at least 0.02 inside it, and held-out accuracy gains at least 0.2 from the start of the window to the end of the run and ends at 0.99 or above; the training fit is not required. For depth 3 at  $p = 47$  (Figure 4), no training pair is misclassified at any log inside any window. The windows contain the CE’s climb after the knee, which at fraction 0.5 starts

from near zero. From the refit’s 20% crossing to the network’s 90% crossing, an interval that at fraction 0.8 contains the network’s whole 10%-to-90% held-out rise, the CE changes by  $-2.9\%$  to  $+9.5\%$  at that fraction (setting medians); over the same interval, it rises at the smaller fractions. The primary endpoint holds in only 9 of the 160 runs, so these runs extend the post-hoc picture rather than the primary endpoint;  $p = 23$  was not rerun in this way.

### 5.3 Transformers

To test the behavior beyond fully connected networks, we train a one-block transformer at  $p = 113$  in the layout of Nanda et al. [2], with no LayerNorm and, here, no biases (Appendix D.1), on a random 20% of the grid (2,553 pairs); it reads  $(a, b, =)$  and predicts at the last token. The logits are  $f = g \mathbf{V} \hat{r}$ , where  $\hat{r} = r/\text{rms}(r)$  is the final residual vector  $r$  at that token divided by its root mean square [33], the unembedding  $\mathbf{V} \in \mathbb{R}^{p \times 128}$  is the head and  $g$  is a trained scalar gain. The other weight matrices take the supported-Polar step of (3) with  $\eta_W = 0.1$  and  $\lambda_W = 0.01$ , and  $\mathbf{V}$  takes the normalized step with  $\eta_V = 0.03$  and  $\lambda_V = 0.275$ , full batch in float64.

**The output gain.** Without the gain, the logit scale is set by  $\mathbf{V}$  alone, whose norm settles within a few hundred updates, and in all 10 runs the training CE falls by 7–10% during the held-out rise. The gain starts at  $g_0 = 2$  and takes a sign step with decay,  $g \leftarrow g - \eta_g \text{sign}(\partial L / \partial g) - \eta_g \lambda_g g$ , with  $\eta_g = 2 \cdot 10^{-4}$  and  $\lambda_g = 1$ . Once the training set is fitted,  $\partial L / \partial g < 0$ , so  $g$  relaxes toward 1 with time constant  $1/(\eta_g \lambda_g) = 5,000$  updates; the logit scale then shrinks slowly, and the training CE can rise while every training pair stays correct.

**Pass rule and results.** A held-out pair whose mirror  $(b, a)$  is a training pair can be answered from that pair, so we also score the *no-mirror* pairs, whose mirrors are held out too. Each run is scored on a post-hoc window that starts at or after the first full fit, on which the averaged training CE never falls more than 3% below its running maximum, with the start chosen to hold the largest share of the held-out rise (Appendix D.2). A run passes if, on both sets of held-out pairs, the window holds at least 75% of the 10%-to-90% held-out rise and, inside it, a train-only Fourier-filtered refit (below) gains at least 0.2,  $A_H$  of  $\hat{r}$  at least 0.02, and no training pair is misclassified at any 20-update log, and if the final accuracy on all held-out pairs is at least 0.99. All 10 runs pass (Figure 5, middle). After its minimum of about  $0.1 \log p$ , the averaged training CE rises by a median 33% (18–55%) across the no-mirror held-out rise; inside the window,  $A_H$  of  $\hat{r}$  gains 0.78–0.83, and the final held-out accuracy is at least 0.999. With the gain held fixed at 1.5, 2 or 3, the CE falls during the rise and no run passes (0/9), so the decay of the gain is what lifts the CE.

**Readouts.** Fitted to  $\hat{r}$ , the train-only ridge refit (7), which leads the network in Section 5.1, trails the transformer in every run, by a median of 1.2k updates at 50% accuracy on all held-out pairs: it weights all 56 frequencies of the label equally, while  $\hat{r}$  carries only 7 to 13 of them at the network’s 50% crossing. Restricted to the 8 frequencies with the most training energy, a choice that uses training data only, the same readout leads the network in all 10 runs, by a median of 0.7k updates on all held-out pairs; the network’s own logits, restricted to the same 8 frequencies, lead it by about as much, so the restriction removes memorized terms rather than adding signal (Appendix D.5).

**Other primes and two blocks.** With the head decay and the training fraction set per prime, the fraction so that the held-out rise starts after the CE minimum, the recipe passes in 10/10 runs at  $p = 89$  (Figure 5, left), 7/10 at  $p = 149$  (10/10 if up to 10 training pairs may be misclassified) and 5/10 at  $p = 61$  (10/10 on the no-mirror pairs, and 9/10 with a split that keeps every training pair with its mirror). With two blocks at the main runs’ fraction (Figure 5, right), the CE still rises during the held-out rise and no training pair is misclassified at any 20-update log, but the first training errors are logged 1.4k to 5.7k updates after held-out accuracy first reaches 0.99, and held-out accuracy then falls back. No run at the four planned two-block settings passes (0/10 at that fraction, Figure 15, and 0/15 at the other three); lowering the feature rate after the held-out rise kept the fit in three exploratory runs (Table 6). These are development runs: the gain and the rates were chosen on pilot runs, and the pass thresholds were fixed before the runs with the gain (Appendix D).

## 6 Discussion

Relatively flat training cross-entropy can coexist with evolving features and delayed generalization. Our theorems establish this for 1HL networks from constructed initial regions, and the experiments show related behavior from iid initialization and in deeper networks and a one-block transformer, outside these guarantees. The loss itself explains why this is possible: once every training pair is correct, the CE measures confidence rather than correctness. A flat training loss is therefore not by itself evidence that learning has stopped, and diagnostics of the representation, such as the Fourier fraction or a train-only refit, can reveal progress that the loss does not show.

The two diagnostics answer different questions. The Fourier fraction summarizes the structure of the representation, while signed margins underlie our accuracy guarantees. Since the fraction ignores decoder signs and class coverage, we pair it with the train-only refit, which tests predictive usefulness directly. Refit-first recovery shows earlier decodability under the specified ridge rule; it does not show that the actual head follows the refit, and it does not separate the direction of the features from their scale. In the transformer runs of Appendix D.5, the same ridge rule trails the network in every run, while its restriction to the 8 Fourier frequencies with the most training energy leads at the 50% crossing in nearly all of them. The rule weights all frequencies equally, which suits the 1HL features, where every frequency carries held-out signal, but not the transformer’s, where about 8 do.

The theory has clear limits. Theorem 4.1 gives its initialization basin, positive head step and small decay allowance only qualitatively, and its width extension uses inactive padding. Neither theorem shows how training from iid initialization reaches the constructed regions. Quantifying these possibly small neighborhoods and rates, calibrating the prime threshold of Theorem 4.2 and allowing training-active memory copies are natural next steps that require further trajectory estimates.

The deeper-network experiments show qualitative counterparts of our 1HL results, but their configurations rely on hyperparameter search and tuning (Appendix C.1); at  $p = 47$  and  $61$ , a last-layer rate set per training fraction, with head decay scaled with the prime, suffices for the post-hoc criterion of Appendix C.4 but not for the primary endpoint. Understanding how these choices support relatively flat training loss, feature evolution and delayed generalization is a step toward guarantees for deeper networks from iid initialization at practical widths and learning rates.

In the one-block transformer, the behavior depends on a slowly decaying output gain, without which the training CE falls during the held-out rise. A second block, at the main runs’ training fraction, keeps the rising CE during the held-out rise but loses the training fit later, except in a few exploratory runs in which the feature rate is lowered after the rise (Appendix D.3). We have no theory for these cases.

## AI use statement

We used large language model assistants to support the development of the initial ideas of the work, in formulating mathematical claims and hypotheses, and aspects of proof development and writing. These tools also assisted with experimental design, implementation, and interpretation of results. Additionally, we used generative AI tools to check literature and references, and to draft and edit parts of the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## References

- [1] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. 2022. arXiv:2201.02177.
- [2] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. ICLR, 2023. arXiv:2301.05217.
- [3] Andrey Gromov. Grokking modular arithmetic. 2023. arXiv:2301.02679.

- [4] Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. NeurIPS, 2022. arXiv:2207.08799.
- [5] Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking Beyond Algorithmic Data. ICLR, 2023. arXiv:2210.01117.
- [6] Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the Transition from Lazy to Rich Training Dynamics. ICLR, 2024. arXiv:2310.06110.
- [7] 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. ICLR, 2024. arXiv:2311.18817.
- [8] Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. ICLR, 2024. arXiv:2311.07568.
- [9] Jeremy Bernstein and Laker Newhouse. Modular Duality in Deep Learning. 2024. arXiv:2410.21265.
- [10] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. NeurIPS, 2022. arXiv:2205.10343.
- [11] Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. 2023. arXiv:2309.02390.
- [12] Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. ICLR, 2025. arXiv:2501.04697.
- [13] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. ICML, 2023. arXiv:2302.03025.
- [14] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. NeurIPS, 2023. arXiv:2306.17844.
- [15] Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? A theoretical analysis of grokking modular addition. ICML, 2024. arXiv:2407.12332.
- [16] Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. ICLR, 2026. arXiv:2509.21519.
- [17] Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. 2026. arXiv:2602.16849.
- [18] Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. NeurIPS, 2025. arXiv:2506.06489.
- [19] Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics. COLT, 2023. arXiv:2302.11055.
- [20] Pulkit Gopalani and Wei Hu. What happens during the loss plateau? Understanding abrupt learning in transformers. NeurIPS, 2025. arXiv:2506.13688.
- [21] Akash Kumar. Awakening of the Buddha: Subspace Learning During Population-Loss Plateaus. 2026. arXiv:2609.39408.
- [22] Lenz Pracher, Pascal de Jong, Oskar Lieshaus, Alan Jeffares, and Steffen Rulands. A spectral theory of grokking: Weight decay induces feature learning. 2026. arXiv:2609.26679.
- [23] Taeyoung Kim. Grokking on the weight-decay clock: A rate hierarchy from softly broken symmetries. 2026. arXiv:2607.23967.

- [24] Neil 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. ICML, 2025. arXiv:2407.20199.
- [25] Chi-Ning Chou, Oscar Uzdelewicz, Neng-Chun Chiu, Yao-Yuan Yang, and SueYeon Chung. Two speeds of learning: A representation-readout decomposition of grokking and double descent. 2026. arXiv:2605.27078.
- [26] David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. NeurIPS, 2015.
- [27] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, 2024. <https://kellerjordan.github.io/posts/muon/>.
- [28] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. NeurIPS Workshop on Optimization for Machine Learning, 2024. arXiv:2409.20325.
- [29] Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. 2025. arXiv:2504.16041.
- [30] Ali Saheb Pasand and Elvis Dohmatob. Egalitarian gradient descent: A simple approach to accelerated grokking. ICLR, 2026. arXiv:2510.04930.
- [31] Akash Kumar. Can Representation Learning Decouple from Loss Minimization? Polar Updates Have an Answer. 2026. arXiv:2609.36240.
- [32] Kenneth Ireland and Michael Rosen. A Classical Introduction to Modern Number Theory. Second edition. Graduate Texts in Mathematics 84. Springer-Verlag, New York, 1990.
- [33] Biao Zhang and Rico Sennrich. Root mean square layer normalization. NeurIPS, 2019. arXiv:1910.07467.
- [34] Xi Wang and Laurence Aitchison. How to set AdamW’s weight decay as you scale model and dataset size. ICML, 2025. arXiv:2405.13698.

# Appendices

|                                                                                   |           |
|-----------------------------------------------------------------------------------|-----------|
| <b>A Representation diagnostics, windows and endpoints</b>                        | <b>18</b> |
| A.1 Fourier features, projection, and energy . . . . .                            | 18        |
| A.2 Actual prediction, train-only refitting, and signed rule margins . . . . .    | 19        |
| A.3 Windows and endpoints of the experiments . . . . .                            | 19        |
| <b>B One-hidden-layer experiments</b>                                             | <b>20</b> |
| B.1 Protocol and selection . . . . .                                              | 21        |
| B.2 The headline cohort . . . . .                                                 | 21        |
| B.3 Pre-registered tests of recovery without CE decrease . . . . .                | 22        |
| B.4 Results across primes . . . . .                                               | 22        |
| B.5 Replays and other one-hidden-layer studies . . . . .                          | 25        |
| B.6 Theorem-inspired split and update settings from iid initialization . . . . .  | 25        |
| <b>C Deeper-network experiments</b>                                               | <b>27</b> |
| C.1 Network settings and learning-rate selection . . . . .                        | 27        |
| C.2 Depths three to five . . . . .                                                | 27        |
| C.3 Results across primes . . . . .                                               | 28        |
| C.4 Depths three and four across training fractions . . . . .                     | 29        |
| C.5 Post-hoc window counts . . . . .                                              | 33        |
| C.6 Biases in the hidden layers . . . . .                                         | 33        |
| <b>D A one-block transformer</b>                                                  | <b>35</b> |
| D.1 Model, updates and readouts . . . . .                                         | 35        |
| D.2 Window and pass criterion . . . . .                                           | 36        |
| D.3 The gain and the recipe at $p = 113$ . . . . .                                | 36        |
| D.4 Other primes and mirror pairs . . . . .                                       | 37        |
| D.5 Readouts . . . . .                                                            | 38        |
| <b>N Cross-appendix notation</b>                                                  | <b>43</b> |
| <b>E Structured half-grid theorem: construction and estimates</b>                 | <b>43</b> |
| E.1 Statement and conventions . . . . .                                           | 43        |
| E.2 Claim I: construction and uniformly controlled reference dynamics . . . . .   | 43        |
| E.3 Claim II: useful-representation consequences on the same trajectory . . . . . | 56        |
| E.4 Assembly: positive head step and an initial ball at every width . . . . .     | 71        |
| <b>F A random-split low-CE theorem with refit-first recovery</b>                  | <b>73</b> |
| F.1 Statement and construction . . . . .                                          | 73        |
| F.2 Claim I: actual dynamics, plateau and delayed network recovery . . . . .      | 74        |
| F.3 Claim II: refit-first recovery and Fourier growth . . . . .                   | 92        |
| F.4 Assembly: the prescribed ball and one common threshold . . . . .              | 94        |
| <b>G Small decoupled readout decay</b>                                            | <b>95</b> |
| G.1 The structured half-grid construction . . . . .                               | 95        |
| G.2 The random-split construction . . . . .                                       | 96        |

## A Representation diagnostics, windows and endpoints

### A.1 Fourier features, projection, and energy

**The addition-Fourier features.** For the Euclidean inner product on full-grid functions, the orthonormal complex modes and the transform of hidden column  $m$  are

$$\begin{aligned}\psi_{r,s}(a,b) &= \frac{1}{p} \exp\left(\frac{2\pi i(ra+sb)}{p}\right), \quad r,s \in \mathbb{F}_p, \\ \hat{H}_m(r,s) &= \frac{1}{p} \sum_{a,b \in \mathbb{F}_p} H_{(a,b),m} \exp\left(-\frac{2\pi i(ra+sb)}{p}\right).\end{aligned}\tag{13}$$

The addition modes are  $\phi_k = \psi_{k,k}$  for  $k \neq 0$ . Each centered class indicator has the expansion

$$\mathbf{1}_{\{a+b=c\}} - \frac{1}{p} = \frac{1}{p} \sum_{k=1}^{p-1} \exp\left(\frac{2\pi i k(a+b-c)}{p}\right).\tag{14}$$

Thus these targets span  $\mathcal{F}_{\text{add}}$  of (5). For real features, let  $\mathbf{Q}_{\text{add}} \in \mathbb{R}^{p^2 \times (p-1)}$  have columns

$$\frac{\sqrt{2}}{p} \cos\left(\frac{2\pi k(a+b)}{p}\right), \quad \frac{\sqrt{2}}{p} \sin\left(\frac{2\pi k(a+b)}{p}\right)\tag{15}$$

for  $1 \leq k \leq (p-1)/2$ . Fourier orthogonality gives  $\mathbf{Q}_{\text{add}}^\top \mathbf{Q}_{\text{add}} = \mathbb{I}$ ,  $\dim \mathcal{F}_{\text{add}} = p-1$ ,  $\mathbf{P}_{\text{add}} = \mathbf{Q}_{\text{add}} \mathbf{Q}_{\text{add}}^\top$  and  $\mathbf{P}_{\text{add}} \mathbf{C}_{\text{grid}} = \mathbf{P}_{\text{add}}$ . Summing the nonzero addition modes gives the equivalent projector formula

$$(\mathbf{P}_{\text{add}} \mathbf{H})_{(a,b),:} = \frac{1}{p} \sum_{u \in \mathbb{F}_p} \mathbf{H}_{(u,a+b-u),:} - \frac{1}{p^2} \sum_{u,v \in \mathbb{F}_p} \mathbf{H}_{(u,v),:}.\tag{16}$$

**Absolute energies and their fraction.** We normalize energies per input pair, summing over all physical hidden columns. Parseval's identity gives

$$\begin{aligned}E_{\text{add}}(\mathbf{H}) &= \frac{\|\mathbf{P}_{\text{add}} \mathbf{H}\|_F^2}{p^2} = \frac{1}{p^2} \sum_m \sum_{k \neq 0} |\hat{H}_m(k,k)|^2, \\ E_{\text{ctr}}(\mathbf{H}) &= \frac{\|\mathbf{C}_{\text{grid}} \mathbf{H}\|_F^2}{p^2} = \frac{1}{p^2} \sum_m \sum_{(r,s) \neq (0,0)} |\hat{H}_m(r,s)|^2, \\ E_{\text{other}}(\mathbf{H}) &= E_{\text{ctr}}(\mathbf{H}) - E_{\text{add}}(\mathbf{H}) = \frac{\|(\mathbf{C}_{\text{grid}} - \mathbf{P}_{\text{add}}) \mathbf{H}\|_F^2}{p^2}.\end{aligned}\tag{17}$$

The zero-frequency energy is  $E_{\text{DC}}(\mathbf{H}) = \left\| p^{-2} \sum_{a,b} \mathbf{H}_{(a,b),:} \right\|_2^2$ . The uncentered energy is  $p^{-2} \|\mathbf{H}\|_F^2 = E_{\text{ctr}} + E_{\text{DC}}$ . These per-input units match the empirical observers; unnormalized Frobenius energies in the proof appendices differ by the common factor  $p^2$ . The fraction and window change are those of (6), with  $t_0 = \min I$ ,  $t_1 = \max I$  and  $E_{\text{ctr}} > 0$ . These identities distinguish absolute addition-energy growth from a fraction gain caused by decreasing other energy. The same calculation gives the difference-Fourier control by projecting onto  $\{g(a-b) : \sum_y g(y) = 0\}$ , replacing  $(k,k)$  with  $(k,-k)$  and retaining the centered-energy denominator. It applies separately at every hidden layer.

For the orthogonal projector  $\mathbf{P}_H$  onto  $\text{col}(\mathbf{C}_{\text{grid}} \mathbf{H})$ , hidden-span availability is  $A_+ = \|\mathbf{P}_H \mathbf{Q}_{\text{add}}\|_F^2 / (p-1)$ . It measures how much of the rule space the centered hidden span contains, without weighting feature amplitudes; it can already be one while  $A_H$  and readout performance change. It is an oracle full-grid diagnostic, not the train-only refit below. For an isotropic hidden-feature direction in the  $(p^2-1)$ -dimensional centered grid space, the expected  $A_H$  is  $(p-1)/(p^2-1) = 1/(p+1)$ . This is a geometric reference, not the expectation under our nonlinear network initialization. Full-grid projections use the known arithmetic task only for observation; held-out targets do not enter training-gradient or refit objectives.

## A.2 Actual prediction, train-only refitting, and signed rule margins

Section 3.5 defines the diagnostic train-only head. Its full score table is  $\hat{\mathbf{Y}}_j = \mathbf{H}_j \hat{\mathbf{B}}_j$ . The complete Gram form of its held-out scores is

$$\hat{\mathbf{Y}}_{j,\mathcal{D}} = \mathbf{H}_{j,\mathcal{D}} \mathbf{H}_{j,\mathcal{S}}^\top (\mathbf{H}_{j,\mathcal{S}} \mathbf{H}_{j,\mathcal{S}}^\top + \tau_j \mathbb{I})^{-1} \mathbf{T}_{\mathcal{S}}. \quad (18)$$

These are the scores of the unique ridge solution of (7); all hidden columns enter the fit. In the canonical experiments,  $\alpha = 10^{-3}$  is the penalty in mean-loss units. Equivalently, for the last hidden layer,

$$\hat{\mathbf{B}}_t = \arg \min_{\mathbf{B}} \left\{ \frac{1}{n} \left\| \mathbf{H}_{t,\mathcal{S}}^{(D)} \mathbf{B} - \mathbf{T}_{\mathcal{S}} \right\|_F^2 + \alpha \left\| \mathbf{B} \right\|_F^2 \right\}. \quad (19)$$

The empirical MSE is

$$\text{MSE}_{\text{ref}}(t) = |\mathcal{D}|^{-1} \left\| \mathbf{H}_{t,\mathcal{D}} \hat{\mathbf{B}}_t - \mathbf{T}_{\mathcal{D}} \right\|_F^2 = \frac{p-1}{p} E_{\text{ref}}(t). \quad (20)$$

For depth  $D$ ,  $\mathbf{H}_t$  in this equation means  $\mathbf{H}_t^{(D)}$ .

The saved empirical diagnostics use  $\text{MSE}_{\text{ref}}$  with the empirical head; the theory uses  $E_{\text{ref}}$  with its separately prescribed head. The normalization is the squared norm of a centered one-hot target. The theoretical penalties in (8) scale with training-feature energy, unlike the fixed empirical  $\alpha$ . Legacy deep-development refits instead subtract the training-row feature mean and fit training-mean-centered one-hot targets. With  $\mathbf{H}_c$  the centered training-feature matrix, their ridge penalty in sum-loss units is  $10^{-3} \text{tr}(\mathbf{H}_c^\top \mathbf{H}_c) / M_D + 10^{-12}$ . Held-out features are centered by the same training mean before applying the fitted coefficients, without adding a label-mean intercept. These refits remain separately labeled. The representation-first flags of Appendix A.3 use this legacy refit; the depth-two runs of Figure 3 log the same refit as depths three to five. A better training refit objective alone does not imply better held-out accuracy, and fixed- $\alpha$  refits are not invariant to feature scale.

**Signed rule versus residual logits.** The actual logit table and its class-centered rule/residual split are

$$\begin{aligned} \mathbf{F}_t &= \mathbf{H}_t^{(D)} \mathbf{V}_t^\top, & \mathbf{R}_t &= \mathbf{P}_{\text{add}} \mathbf{F}_t \mathbf{C}_p, \\ \mathbf{N}_t &= \mathbf{F}_t \mathbf{C}_p - \mathbf{R}_t. \end{aligned} \quad (21)$$

For  $y = y^*(x)$ , define the signed rule and adverse residual contrasts

$$\begin{aligned} \gamma_t(x) &= \min_{c \neq y} (\mathbf{R}_{t,xy} - \mathbf{R}_{t,xc}), \\ \beta_t(x) &= \max_{c \neq y} (\mathbf{N}_{t,xc} - \mathbf{N}_{t,xy}). \end{aligned} \quad (22)$$

Their names are local to this diagnostic; they are unrelated to the construction parameters  $\gamma, \beta$  in the proofs.

The residual includes input-constant, class-dependent logits; centering hidden features in  $A_H$  does not remove their effect on classification. Then  $m_t(x) \geq \gamma_t(x) - \beta_t(x)$ ;  $\beta_t(x)$  may be negative when the residual helps classification. This elementary bound explains which signed information a scalar energy discards. A gain  $\Delta A_H > 0$  alone does not guarantee a larger margin, a successful refit, or better test accuracy. The theorems require their separate signed-prediction arguments, not a universal alignment threshold.

**Observation cadence.** In the headline study, refit observations every 500 updates do not resolve crossings or dips between observations. Network accuracy is saved every update there and every 20 updates in the deeper-network experiments. A first 90% crossing need not be permanent.

## A.3 Windows and endpoints of the experiments

Plateau variation  $v(I)$ , the cycle mean  $\bar{L}_t$  and window gains such as  $\Delta A_H(I)$  are those of Section 3.3.

**Locked windows.** In the fresh one-hidden-layer cohorts (the *pre* rows of Table 1 and the cohorts of Appendix B.3), the window  $I = [t_0, t_1]$  is fixed on development runs and locked, with the recipe, before the evaluated pairs are drawn. On it we report  $v(I)$  against the declared 10% tolerance, training correctness at every update,  $\Delta A_H(I)$ , and the held-out accuracy of the network and of the train-only refit.

**Loss-selected band and primary endpoint.** The deeper-network evaluations, and the one-hidden-layer runs of Figure 2, select a CE peak after a head-norm knee, without using held-out outcomes to locate the band. The knee is the first logged head-norm checkpoint with  $\|\mathbf{V}_t\|_F \geq 0.99/\lambda_V$ . The band  $I_C = [a, b]$  is the maximal interval containing the largest post-knee  $\bar{L}$  on which  $\bar{L}$  stays at least that peak divided by 1.03. With  $\text{clip}(z) = \min(0.9, \max(0.1, z))$ , the captured share of the held-out rise is

$$q(I_C) = [\text{clip}(\text{Acc}_{\mathcal{D}}(b)) - \text{clip}(\text{Acc}_{\mathcal{D}}(a))]/0.8, \quad (23)$$

using the nearest logged accuracies. A run meets the *primary endpoint* if  $q(I_C) \geq 0.75$ , its final held-out accuracy is at least 0.99,  $\Delta A_H^{(D)}(I_C) \geq 0.02$  for the last hidden layer, and every logged minimum training margin in  $I_C$  is positive. Accuracies and margins are logged every 20 updates, and margins between logs are not certified; head-norm, Fourier and refit diagnostics are logged every 500 updates (1000 for the  $p = 97, 101$  across-prime runs). Fourier and refit values at band endpoints use the nearest observation. A depth succeeds if the lower end of the two-sided 95% Wilson interval is at least 0.80, that is, at least 37 of 40 passes. Selection guarantees a 3% cycle-mean range if the band’s maximum is the selected post-knee peak, as in the reported primary and selected across-prime cohorts. The band can extend before the knee, so this bound is not universal across development runs and is not an additional pass condition. A short flat band can occur away from recovery, which is why the share condition is needed.

**Strict first-fit-anchored endpoints.** These stronger endpoints anchor at the first fit  $T_{\text{fit}}$ : the first update at which every training margin exceeds  $10^{-10}$  (one hidden layer) or the first positive logged minimum training margin (deeper networks). For one hidden layer the strict joint endpoint requires  $T_{\text{fit}} \leq t_0$ ; the 10% plateau and  $\Delta A_H \geq 0.02$  on  $[t_0, t_1]$ ; strict fit from  $T_{\text{fit}}$  through the window end and the recovery; held-out accuracy at most 20% at  $t_0$  and during a delay of  $\max(1000, 0.02 \times \text{horizon})$  updates after  $T_{\text{fit}}$ ; and a first 90% crossing inside the window. For deeper networks it uses  $I_C$  and requires, besides the fit,  $\Delta A_H^{(D)}(I_C) \geq 0.02$ , raw band variation of at most 10%, band-start held-out accuracy of at most 20%, a first 90% crossing inside the band, no fit relapse through the band end, and held-out accuracy of at most 20% for  $\max(1000, [0.02T])$  updates after the fit. Its persistent-fit variant starts the delay after the last nonpositive margin. Two representation-first flags combine these joints with a useful legacy refit (Appendix A.2) on  $I_C$ : from at most 0.70 at the start of the band, it crosses 90% by its end, gains at least 0.20, and crosses before the network, whose accuracy is then at most 0.20.

**Post-hoc non-falling window.** After every run had been seen, we also scored runs with a looser, loss-only rule. With  $\bar{L}_t$  and the knee as above, call an interval  $[a, b]$  after the knee *non-falling* if  $\bar{L}_t \geq \max_{a \leq s \leq t} \bar{L}_s/1.03$  for every  $t \in [a, b]$ . The window is the maximal non-falling interval with the largest rise  $\max_{[a, b]} \bar{L} - \bar{L}_a$ ; ties go to the longer. A run passes if its held-out accuracy, or that of the train-only refit, rises by at least 0.2 between the first and the last log inside the window; the captured share, final accuracy, training fit and  $\Delta A_H$  are not required. For the one-hidden-layer studies of Table 1, Figures 6 and 7 and the fixed-recipe sweep of Appendix B.4, the knee is instead the first logged update, rather than a 500- or 1000-update checkpoint, with  $\|\mathbf{V}_t\|_F \geq 0.99/\lambda_V$ . Of these, the fixed-recipe, replay and quadratic studies do not log the head norm and start instead at the CE minimum before their largest relative CE climb; on the 80 headline and per-prime runs, which log both, this start lies within 230 updates of the knee and gives the same verdicts. The gray bands of Figures 2, 3, 4, 7 and 10–12 run from the median start to the median end of these windows, and Appendix C.5 gives the counts, except for Figures 4 and 10–12, whose counts are in Appendix C.4.

## B One-hidden-layer experiments

This appendix supports Section 5.1. *Fresh* pairs were drawn after a recipe and its window had been locked. *Earlier* runs predate this protocol; they are reanalyzed from saved states or replayed, and are reported

descriptively. Windows and endpoints are defined in Appendix A.3, and Table 1 summarizes the headline, per-prime, replay and quadratic studies. The experiments are evidence for coexistence, not replications of either initial-region theorem.

## B.1 Protocol and selection

The headline configuration and initialization are those of Section 5.1. The supported Polar step keeps singular values above  $10^{-10}$  times the largest. CE and accuracies are saved at all 40,001 states of every pair,  $A_H$  every 100 updates, and the refit every 500. No run is excluded; plotted loss uses two-update cycle means without additional smoothing. The refit uses the raw-feature, fixed- $\alpha = 10^{-3}$  rule (7). A uniform shrinkage of the features strengthens its penalty, unlike the training-RMS-normalized rules of the theory.

A first search of 96 runs at  $p = 17$  found no qualifying recipe. A second tested 48 runs across  $p \in \{23, 31\}$ , training fractions 0.5 and 0.8 and  $\lambda_W \in \{0.01, 0.015, 0.02\}$ , with four initialization/split pairs per recipe,  $c = 24$ ,  $\eta_W = \eta_V = 0.03$ ,  $\lambda_V = 0.2$  and 40k updates. A rule fixed before that search scanned windows on the 500-update grid, with start at least 1000 and length at least 2000. Its strict tier required, for every development pair of a recipe,  $v \leq 8\%$ , minimum CE at least 0.12,  $\Delta A_H \geq 0.07$ , perfect window training accuracy, held-out accuracy at most 20% at the start, and a first 90% crossing in  $(t_0, t_1 - 500]$ . Qualifying recipes were ordered by worst-pair Fourier gain, with fixed tie-breaks. The rule selected the headline recipe and [4500, 12000], and both were locked before 20 fresh pairs were drawn with NumPy’s `default_rng(20260930)`. Auxiliary selection and evaluation criteria, not all of which were met, are recorded with the code. Selection used held-out accuracy, so the evaluation tests this locked choice, not an untuned optimizer family.

## B.2 The headline cohort

Section 5.1 gives the window outcomes. The median of  $v(I)$  is 7.06%: these are positive-loss plateaus at the predeclared 10% tolerance, not 3% plateaus. Over  $I$ ,  $A_H$  gains 0.0955–0.0972 (median 0.0965). When the refit first records 90%, at update 6,500, the previous checkpoint is below it in every run. This is earlier availability under a fixed rule on a 500-update grid, not a proof that the head tracks the ridge solution.

**Two phases.** The CE peak separates two phases. While CE rises,  $A_H$  makes 25–39% of its window gain and the refit exceeds 70% by update 6,000, whereas the network stays at most 5.7%. The network’s own rise follows the peak, with cycle-mean CE  $\bar{L}$  falling by 1.5–3.4% between its first 10% and 90% crossings. Even so,  $\bar{L}$  ends 2.7–11.4% higher at the network’s first 90% than at the update at which the refit starts rising, in every run. Here and below, the refit’s rise starts at its last 500-update observation strictly before its first observed 10% crossing.

**Norms and energies.** Before the window, CE falls to 0.792–0.854 at updates 1,801–1,926 and then rises to 1.456–1.498 at update 4,500. During the window,  $\|\mathbf{V}_t\|_F = 4.997$ –5.000, at the cap  $1/\lambda_V = 5$ , while  $\|\mathbf{W}\|_F$

Table 1: One-hidden-layer studies (ReLU rows: width 512, 80% random split,  $\eta_W = 0.03$ ,  $\lambda_V = 0.2$ ). Window: *pre* = fixed on development runs before the fresh pairs; *per run* = persistent held-out 10%→90% crossings (descriptive). Flat:  $v \leq 10\%$ . Rec.: held-out 90% within the horizon. First-fit: the strict joint endpoint of Appendix A.3; in particular, every training margin must stay above  $10^{-10}$  from  $T_{\text{fit}}$  through recovery and the window end. –: not computed. The fixed-recipe sweep is in Appendix B.4.

| Study                  | $c/\eta_V/\lambda_W$ | Window         | $p$ | Runs | Flat | Rec. | First-fit | Med. $v$ | Med. $\Delta A_H$ | Med. min CE         |
|------------------------|----------------------|----------------|-----|------|------|------|-----------|----------|-------------------|---------------------|
| Headline (fresh)       | 24/.03/.015          | pre [4.5, 12]k | 31  | 20   | 20   | 20   | 0         | 7.06%    | .0965             | 1.48                |
| Per-prime (fresh)      | 24/.01/.005          | pre [13, 21]k  | 17  | 20   | 16   | 20   | 15        | 6.3%     | .039              | $1.8 \cdot 10^{-4}$ |
|                        | 48/.03/.005          | pre [15, 21]k  | 23  | 20   | 20   | 20   | 0         | 5.4%     | .041              | $5.1 \cdot 10^{-3}$ |
|                        | 96/.03/.005          | pre [18, 24]k  | 29  | 20   | 20   | 20   | 0         | 6.5%     | .043              | $3.9 \cdot 10^{-2}$ |
| Replays                | 24/.03/.01           | per run        | 23  | 5    | 5    | 5    | –         | 2.4%     | .038              | 0.26                |
|                        | 24/.03/.015          | per run        | 31  | 5    | 5    | 5    | –         | 2.7%     | .049              | 1.5                 |
| Quadratic <sup>†</sup> | –                    | per run        | 13  | 40   | 0    | 40   | –         | 23.6%    | .108              | –                   |

<sup>†</sup>Width 64, 135 training pairs, row-sphere features, steps  $5 \cdot 10^{-4}$  and 0.5, head decay  $10^{-3}$ , loss logged every 500 updates.

falls from 128.9–130.5 to 57.4–57.7 and the centered training-logit RMS from 0.720–0.740 to 0.464–0.480. For  $\lambda_V > 0$  and  $q = 1 - \eta_V \lambda_V \in [0, 1]$ , the normalized head update (3) gives the envelope

$$\|\mathbf{V}_t\|_F \leq q^t \|\mathbf{V}_0\|_F + (1 - q^t)/\lambda_V, \quad (24)$$

an upper bound, not a proof that norm saturation causes recovery. Over  $I$ ,  $E_{\text{ctr}}$  falls to 0.204–0.209 of its initial value while  $E_{\text{add}}$  grows by a factor 1.22–1.31, so the fraction gain is mostly suppression of energy outside the addition subspace. With fixed ridge regularization on raw features, the refit-first observation does not separate directional change from scale.

**Retrospective suffix windows.** Let  $t_\epsilon$  be the earliest update after which CE stays within a factor  $1 + \epsilon$  of its minimum until update 40,000. The rule uses training CE only but needs the whole future trace, so it is retrospective, not a stopping rule. At  $\epsilon = 10\%$ ,  $t_\epsilon$  is 4,170–4,311, and at 5% it is 4,653–6,421, both with zero held-out accuracy. At 3% it is 4,946–8,376, after improvement has begun in some runs. Training stays fitted on every suffix, held-out accuracy at 40k is 94.30–99.48%, and  $A_H$  ends in  $[0.1194, 0.1203]$ .

**The strict first-fit-anchored endpoint.** No headline run meets the strict endpoint of Appendix A.3. The first strict fit, at updates 203–796, is followed by intermittent misclassification of a few training pairs until a permanent fit at updates 239–1,085, so fit is not persistent from its first attainment. Held-out accuracy stays zero through the delay. Reanchoring at the persistent fit would change the definition post hoc.

### B.3 Pre-registered tests of recovery without CE decrease

In the headline cohort,  $\bar{L}$  decreases by 1.5–3.4% between the network’s first 10% and 90% held-out crossings. After this evaluation we screened the 12 recipes of the second development search for recovery without such a decrease. Only one passed the screen in all four of its pairs:  $p = 23$  with the headline hyperparameters. A window rule applied to these four pairs, relaxed once after its first version admitted no window, chose  $I = [6500, 16500]$ .

Before any new run we locked the recipe, the window, 20 fresh pairs, a 20,000-update horizon, the code and a decision rule. The joint endpoint requires  $v(I) \leq 10\%$ , perfect training accuracy on  $I$ , held-out accuracy at most 20% at  $t_0$ ,  $\Delta A_H \geq 0.04$ , and no CE decrease during recovery: the first 90% held-out crossing  $T_{90}$  occurs by  $t_1 - 500$ ,  $\bar{L}$  never falls more than 0.5% below its running maximum on  $[t_0, T_{90}]$ , and  $\bar{L}(T_{90}) \geq \bar{L}(t_0)$ . At least 18 of the 20 runs had to pass. It holds in 10 of the 20 runs, so this test fails (Figure 6, top). Every run meets the other four clauses; the failures concern the end of the recovery, a late or missing 90% crossing or a small loss decrease before it.

In these runs  $\bar{L}$  stays within 0.001% of its running maximum up to the first 50% crossing. This pattern motivated a second lock, written before 20 new pairs were launched, that keeps the joint endpoint with  $T_{90}$  replaced by the first 50% crossing. The joint endpoint then holds in 20 of 20 runs (95% Wilson interval  $[0.84, 1.00]$ ; Figure 6, bottom, and Figure 7a). Held-out accuracy rises from at most 1.9% to 50% at updates 8,278–9,191, while  $\bar{L}$  never falls more than 0.05% below its running maximum and ends 3.8–5.9% above  $\bar{L}(t_0)$ . As secondary endpoints, the joint criterion with the registered 0.5% drawdown tolerance holds up to the first 70% crossing in 20 runs and up to the first 90% crossing in 9, and 17 runs reach 90% within 20,000 updates. The confirmed statement covers recovery to 50%, not to 90%.

### B.4 Results across primes

**One hidden layer across primes.** The runs of Figure 2 have one hidden layer of width 512 and use the deeper-network implementation of the updates (Appendix C.1). At each prime,  $\lambda_W$  is the best of a development sweep on one or two seeds. It was fixed before any new run and then run on 8 or 9 fresh initialization/split seeds, for 10 runs per prime; at  $p = 101$  the setting chosen at  $p = 97$  was run unchanged on 10 fresh seeds. Runs stopped early by the stated rule (held-out accuracy of at least 0.99, maintained) keep their last value. The figure’s caption gives the primary-endpoint counts. These are development settings on a few seeds each, not a pre-specified evaluation.

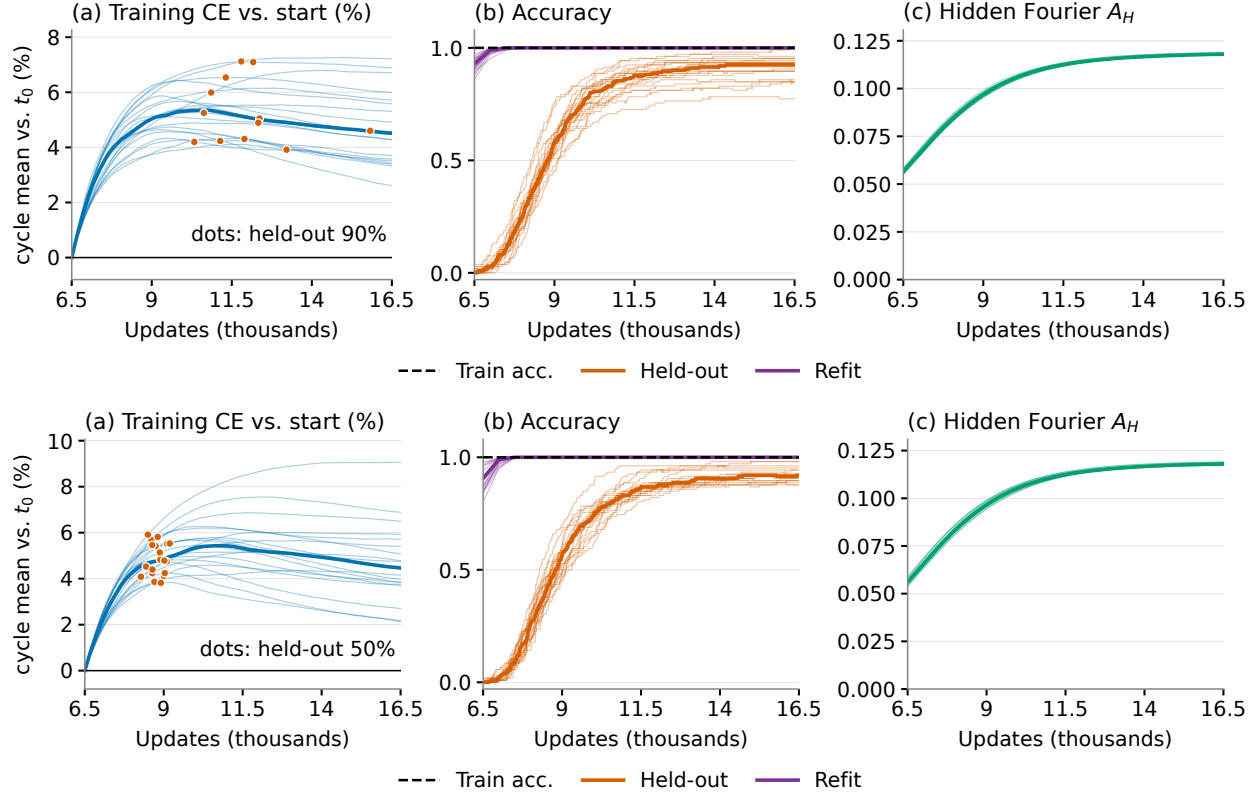


Figure 6: **Pre-registered**  $p = 23$  cohorts (20 fresh pairs each; window  $I = [6500, 16500]$  chosen on four development pairs and locked before these runs). (a) Cycle-mean training CE relative to its value at  $t_0$ ; dots mark each run's first 90% (top) or 50% (bottom) held-out crossing on  $I$ . (b) Median training accuracy (dashed) and held-out accuracy of the network and of the train-only refit (500-update grid). (c)  $A_H$ . Thin curves are runs and thick curves medians. The joint endpoint, which limits cycle-mean drawdown to 0.5% before the dot and requires the loss at the dot to be no smaller than the loss at  $t_0$ , holds in 10 of 20 runs (top) and 20 of 20 (bottom); 18 were required.

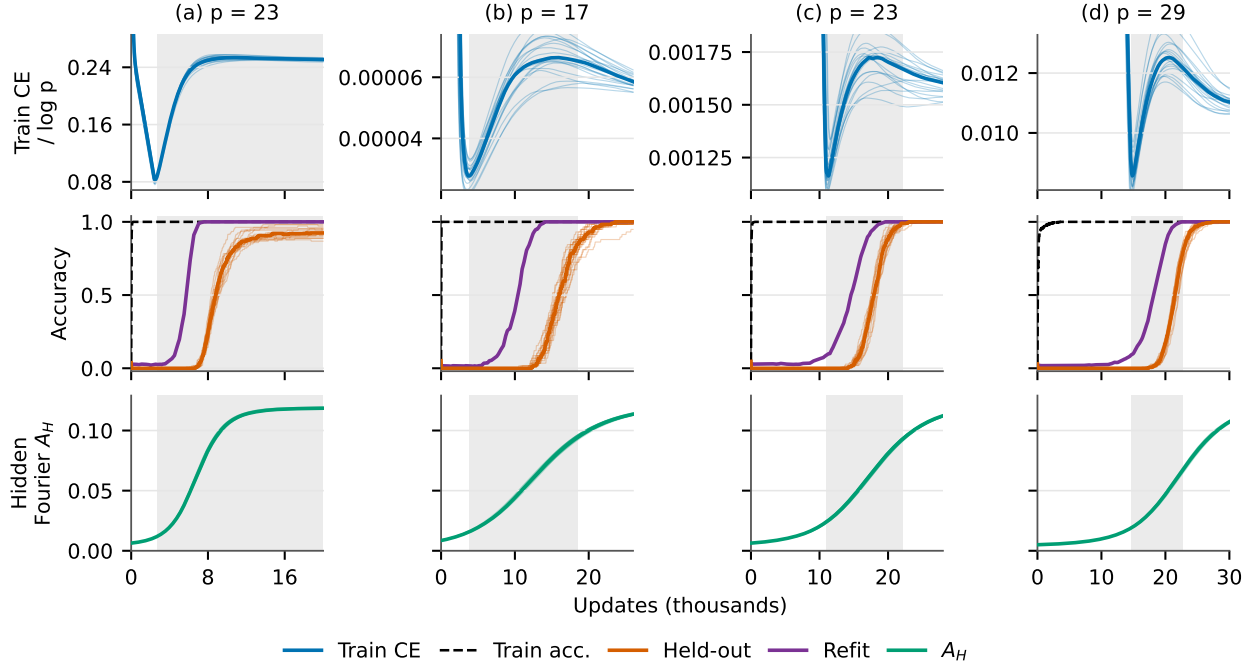


Figure 7: **Locked and pre-registered one-hidden-layer cohorts** (20 fresh pairs per column). (a)  $p = 23$  with the headline hyperparameters: the second pre-registered cohort, window  $[6500, 16500]$ . (b)–(d) The per-prime recipes at  $p = 17, 23, 29$ , with windows  $[13000, 21000]$ ,  $[15000, 21000]$  and  $[18000, 24000]$ . Rows: cycle-mean training CE divided by  $\log p$  (each column on its own scale); median training accuracy (dashed), held-out accuracy of the network (runs and median) and median train-only refit;  $A_H$ . Gray: from the median start to the median end of the runs’ post-hoc windows. The headline runs at  $p = 31$  are in Figure 1.

**Per-prime recipes.** Three recipes at  $p = 17, 23, 29$  were fixed, each with its own window, after 54 development runs and before 20 fresh pairs per prime were drawn (Figure 7b–d). Each run lasts 50k updates, with width 512, training fraction 0.8,  $\eta_W = 0.03$ ,  $\lambda_W = 0.005$ ,  $\lambda_V = 0.2$  and  $(c, \eta_V) = (24, 0.01), (48, 0.03), (96, 0.03)$ . The 10%-plateau counts are 16/20, 20/20 and 20/20, with median  $\Delta A_H$  of 0.039, 0.041 and 0.043, and all 60 runs reach 90% held-out accuracy. The median minimum CE is  $1.8 \cdot 10^{-4}$ ,  $5.1 \cdot 10^{-3}$  and  $3.9 \cdot 10^{-2}$ , far below the headline study, and the strict joint endpoint holds in 15/20, 0/20 and 0/20 runs, so the plateau and recovery counts are not joint successes. From the refit-start checkpoint defined above to the network’s first 90%,  $\bar{L}$  ends above its starting value in 20, 13 and 15 runs (medians +36%, +37% and +33%).

**One fixed recipe across primes.** Earlier runs hold one recipe fixed across  $p = 11, 13, 17, 19, 23, 29$ , with 20 pairs each:  $c = 24$ ,  $\eta_W = \eta_V = 0.03$ ,  $\lambda_W = 0.005$ ,  $\lambda_V = 0.2$ , width 512 and 40k updates. Observers and the refit were recomputed from saved states every 1,000 updates. On the common window [15000, 20000] the 10% plateau holds in most runs only at  $p = 17$ : in 0, 0, 17, 2, 0 and 0 of the 20 runs at  $p = 11, 13, 17, 19, 23, 29$ . Success of a retuned family is thus no evidence that one practical recipe works across moduli.

## B.5 Replays and other one-hidden-layer studies

**Replayed trajectories.** We replay five earlier pairs at each of  $p = 23$  ( $\lambda_W = 0.01$ ) and  $p = 31$  (the headline hyperparameters) for 30k updates, with a Gram-eigendecomposition Polar factor (relative cutoff  $10^{-7}$ ). Relative raw-CE differences from the originals stay below  $2 \cdot 10^{-14}$  through 15k updates, first exceed  $10^{-9}$  at updates 17,470–28,134, and are at most 0.069% at the 30k horizon; every scored window ends by 13.5k. The windows come from each run’s persistent 10%→90% held-out transition, snapped to the observer grid, so they are descriptive. All ten runs have 10% plateaus, with median  $v$  of 2.4% and 2.7%, and recover, with median  $\Delta A_H$  of 0.038 and 0.049. These trajectories also belong to the deeper study’s development inventory and are not counted twice.

**Quadratic-activation contrast.** A separate study uses 40 earlier pairs at  $p = 13$ , width 64, with a square activation, row-sphere feature updates and their own head dynamics (feature step  $5 \cdot 10^{-4}$ , head step 0.5, head decay  $10^{-3}$ ). All reach 90% at least once and gain at least 0.02 in Fourier fraction, but none has a 10% CE plateau (median variation 23.6%, with loss logged every 500 updates). Activation and optimizer both differ, so this is not an activation-only ablation.

**Other settings.** The second search’s recipes with a 50% training fraction did not recover within 40k updates, and a later search at  $p = 31$  with no head decay or a quadratic-residue split met the selection rule in no arm. These are limits of the tested recipes and horizons, not counterexamples to the theorems.

## B.6 Theorem-inspired split and update settings from iid initialization

We trained a bias-free 1HL ReLU network at  $p = 101$  using Theorem 4.1’s split (10) ( $n = 5,050$  training and 5,151 held-out pairs), the clock (60) with  $\lambda_W = \lambda = 1$  and  $\eta_W = s = 1 - (9/10)^{1/32}$ ,  $\lambda_V = 0$ , and the refit penalty  $\tau_t = R_t/(32n)$  of (8). This is not a theorem replication: the width is  $M = 512$  instead of at least  $M_{\min}(101) = 515,302$ , and the initialization is iid as in Section 5.1 ( $c = 24$ ) instead of the engineered ball. The head steps are not checked against the theorem’s existential cap, and the observed windows are not its certified windows. In a plan fixed after one development run and before these runs, we chose four head steps  $\eta_V$ , with  $10^{-3}$  as the primary one, and four per-run criteria for  $T = 40,000$  updates: (a)  $v([1000, T]) \leq 0.01$ ; (b)  $A_H$  rises by at least 0.02 over  $[1000, T]$ ; (c) the refit’s held-out accuracy is at most 0.1 at update 1,000 and reaches 0.9; (d) the refit reaches 0.5 at least 1,000 updates before the network, or the network does not reach 0.5 by  $T$ . Network accuracy is logged every 50 updates; the refit and Fourier fraction are observed every 250. Crossings refer to these checkpoints.

With  $\eta_V = 3 \cdot 10^{-4}$  all four criteria hold in 10/10 runs (Figure 8, top; the plan’s 6 runs and 4 more seeds run for this version). With the primary  $\eta_V = 10^{-3}$ , (b)–(d) hold in every run but (a) in none, with  $v([1000, T]) = 1.34$ –1.37% (bottom); the larger steps  $3 \cdot 10^{-3}$  and  $10^{-2}$  give  $v$  of about 4% and 15% (0/6 each). In the two displayed arms, the training loss varies by at most 0.05% and 0.17% over each run’s own interval from refit accuracy 0.1 to 0.9. Over these intervals  $A_H$  rises from about 0.009 to 0.025, and the

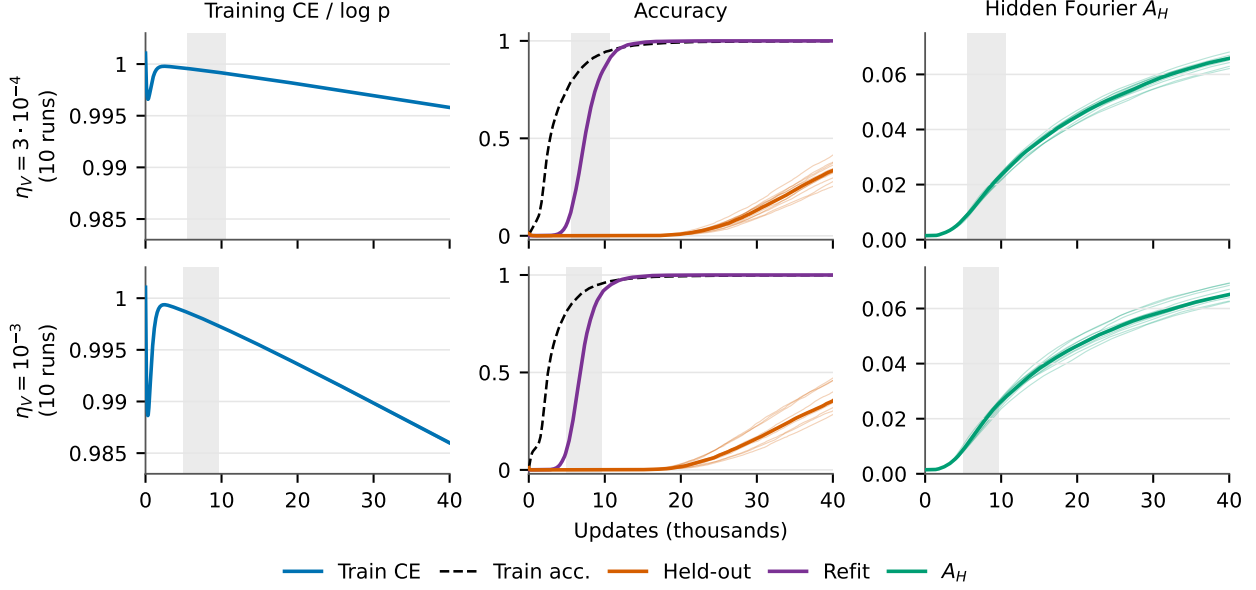


Figure 8: **Theorem-inspired settings at  $p = 101$  from iid initialization** (width 512; all runs in the two displayed arms). Rows: head steps  $\eta_V = 3 \cdot 10^{-4}$  (10 runs; all criteria of Appendix B.6 hold in each) and the primary  $\eta_V = 10^{-3}$  (10 runs; the training loss varies by 1.34–1.37% over  $[1000, 40000]$ , above the 1% criterion). Columns: training cross-entropy loss divided by  $\log p$  (the axis spans 2% of  $\log p$ ); median training accuracy (dashed), held-out accuracy of the network (runs and median) and median held-out accuracy of the theorem’s train-only refit;  $A_H$ . Gray: from the median first observed checkpoint at which the refit’s held-out accuracy reaches 0.1 to the median first at which it reaches 0.9; these medians need not define any individual run’s interval.

network’s held-out accuracy stays zero at recorded checkpoints. By update 40,000 the refit is at one in every run and the network at 0.25–0.47.

The displayed runs do not reproduce Theorem 4.1’s clauses (i) and (iii): training accuracy is 0.76–0.84 when the refit reaches 0.1, 2–14 training pairs are still misclassified at update 40,000, the network has not recovered by then, and the loss stays within about 1.5% of  $\log p$ , a plateau near uniform prediction rather than a fitted one.

The plan also fixed two controls with two runs each. With a nearly frozen head,  $\eta_V = 10^{-6}$ , the runs stall: at update 40,000 training accuracy is 0.42–0.46, the refit’s held-out accuracy is below 0.01, and  $A_H$  has risen by 0.002. With  $\eta_V = 10^{-3}$  and the experiments’ slower clock,  $\eta_W = 0.03$  and  $\lambda_W = 0.02$ , criteria (b)–(d) hold but  $v([1000, T])$  is 105–106%.

## C Deeper-network experiments

This appendix supports Section 5.2: networks with three to five hidden layers at  $p = 31$  and width 256, changes of the modulus, width and feature optimizer, settings across primes, and hidden biases. The 164 evaluated runs use 64 seed indices, separate from the 199 development runs that chose the settings using held-out accuracy and windows maximizing the captured accuracy rise under a flat cycle-mean-loss condition. The development runs, which include the ten replays of Appendix B.5, are not counted in the evaluation rates. Seed indices recur across depths and settings, so these observations are not independent. Windows and endpoints are defined in Appendix A.3.

### C.1 Network settings and learning-rate selection

The task, bias-free ReLU architecture, random 80% split and full-batch float64 CE are as in Sections 3 and 5. The primary configuration uses  $p = 31$  and width 256 in every hidden layer. With independent standard Gaussian arrays  $\mathbf{Z}_\ell, \tilde{\mathbf{Z}}$ , the initial weights are

$$\mathbf{W}^{(1)} = \frac{24}{\sqrt{p}} \mathbf{Z}_1, \quad \mathbf{W}^{(\ell)} = 8\sqrt{\frac{2}{M}} \mathbf{Z}_\ell \ (\ell \geq 2), \quad \mathbf{V} = \frac{\tilde{\mathbf{Z}}}{24 \cdot 8^{D-1} \sqrt{M}}. \quad (25)$$

Every layer trains throughout the main run. Feature updates use a Gram-eigendecomposition implementation of supported Polar normalization. It keeps eigenvalues of  $\mathbf{F}^\top \mathbf{F}$  (or  $\mathbf{F} \mathbf{F}^\top$  for a wide  $\mathbf{F}$ ) above  $10^{-14}$  times the largest, a relative singular-value cutoff of  $10^{-7}$ , and divides by the square root of the retained rank. This differs numerically from the headline SVD implementation, and near-threshold rank sensitivity remains a robustness limitation.

The head uses  $\eta_V = 0.03$  and  $\lambda_V = 0.2$ . The first layer has step 0.03 and decay 0.015, so  $K_1 = \eta_1 \lambda_1 = 4.5 \cdot 10^{-4}$ . For  $\ell \geq 2$ ,  $\lambda_\ell = 0.07$  and  $\eta_\ell = r_\ell K_1 / 0.07$ , where the shrink-rate multiplier  $r_\ell$  is 1 for intermediate layers and  $r$  for the last. To select  $r$ , we first train at multiplier  $r_0$  and record  $t_{10}^{\text{sel}}$ , the first logged update with 10% held-out accuracy. We then solve

$$r K_1 t_{10}^{\text{sel}} \exp\{0.9(r - r_0)\} = x_*, \quad (26)$$

then round  $r$  to a multiple of 0.005 and clip it to  $[0.10, 1.20]$ . For  $D = 3, 4, 5$ ,  $r_0 = 0.50, 0.35, 0.35$  and  $x_* = 3.85$ . We restart from the same initialization on the same split and hold the selected  $r$  fixed throughout training, so there is no within-run loss controller.

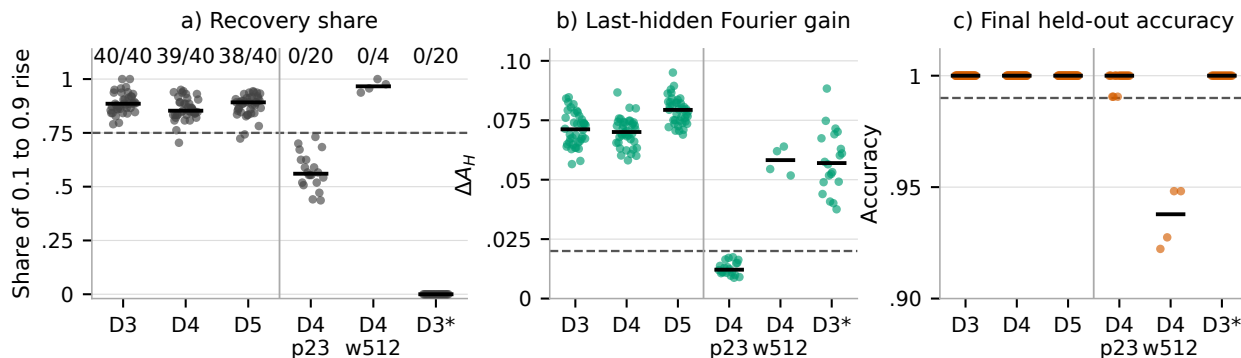
### C.2 Depths three to five

The evaluation criteria, rate-selection rule and seeds were fixed before training, and all 164 evaluated runs completed. Table 2 lists the settings and results. Initialization and split seeds are  $1000 + i$  and  $2000 + i$ : indices 1–40 recur at each primary depth, the normalized-gradient arm reuses depth-3 indices 1–20, the  $p = 23$  arm uses 41–60 and the width-512 arm 61–64.

**Primary endpoint.** The 40/40, 39/40 and 38/40 passes at depths 3, 4 and 5 have per-depth (not simultaneous) Wilson intervals  $[0.91, 1.00]$ ,  $[0.87, 1.00]$  and  $[0.83, 0.99]$ , whose lower ends are at least 0.80, so each depth meets the criterion (Figure 9). The three misses fail only the share threshold. The median band-peak CE is about 0.50, 0.54 and  $0.58 \log p$ .

**Fixed versus selected learning rates.** On 20 paired seeds per depth, runs at the fixed multiplier  $r_0$  pass 19/20, 5/20 and 19/20, against 20/20 for the recipe (exact McNemar  $p = 1, 6.1 \cdot 10^{-5}$  and 1). The recipe’s benefit is thus clear at depth 4, not at every depth.

**Changing the modulus, width or optimizer.** None of the three changed settings meets the primary endpoint (Table 2):  $p = 23$  at depth 4 (0/20), width 512 at depth 4 (0/4; final held-out accuracy 0.92–0.95 at 100k updates), and Frobenius-normalized gradient (NGD) feature updates at depth 3 (0/20, against 20/20 for the paired Polar recipe; exact McNemar  $p = 1.9 \cdot 10^{-6}$ ). The NGD runs generalize, but their CE keeps



Counts: joint passes. p23:  $p = 23$ ; w512: width 512; D3\*: normalized GD.  
Other settings: Polar. Black bars: medians. Dashed lines: primary thresholds.

Figure 9: **Results for all 164 evaluated deep-network runs.** Dots are runs and bars medians. Left of each divider: the primary cohorts ( $p = 31$ , width 256,  $D = 3, 4, 5$ ); right:  $D = 4$  with  $p = 23$ ,  $D = 4$  with width 512, and  $D = 3$  with Frobenius-normalized hidden updates (NGD). (a) Share of the 10%-to-90% held-out rise inside the loss-only band  $I_C$ , with primary-endpoint pass counts above; (b) last-hidden-layer Fourier-fraction gain  $\Delta A_H$  on  $I_C$ ; (c) held-out accuracy at the saved horizon. Dashed lines mark the primary thresholds in (a)–(c); the fourth primary criterion, positive logged training margins in  $I_C$ , holds in all 164 runs. Rates were selected by rule (26).

Table 2: Deep-network settings and outcomes, retaining all evaluated runs. Rate-selection runs stop at  $t_{10}^{\text{sel}}$ , except those also used for fixed-rate comparisons, which continue to the full horizon. Each evaluated run restarts from the same initialization and split used to select its rate. Primary: the per-run primary endpoint.  $J_1/J_{\text{pers}}$ : first-fit and persistent-fit strict joints;  $P_{10}/P_3$ : raw-CE range at most 10% or 3% from the first 10% to the first 90% held-out accuracy; RF: legacy-refit representation-first flag. These secondary tests do not alter the primary verdict.

| Setting | $(D, M, p)$  | Indices $i$  | $r_0$ | $x_*$ | Horizon | Primary | $J_1/J_{\text{pers}}$ | $P_{10}/P_3$ | RF |
|---------|--------------|--------------|-------|-------|---------|---------|-----------------------|--------------|----|
| Primary | (3, 256, 31) | 21–40 / 1–20 | .50   | 3.85  | 80k     | 40/40   | 0/17                  | 40/2         | 0  |
| Primary | (4, 256, 31) | 21–40 / 1–20 | .35   | 3.85  | 100k    | 39/40   | 0/13                  | 39/0         | 0  |
| Primary | (5, 256, 31) | 21–40 / 1–20 | .35   | 3.85  | 80k     | 38/40   | 0/19                  | 39/0         | 0  |
| Modulus | (4, 256, 23) | 41–60        | .35   | 3.85  | 120k    | 0/20    | 0/0                   | 0/0          | 0  |
| Width   | (4, 512, 31) | 61–64        | .40   | 4.50  | 100k    | 0/4     | 0/1                   | 4/1          | 0  |
| NGD     | (3, 256, 31) | 1–20         | .50   | 3.85  | 80k     | 0/20    | 0/0                   | 0/0          | 0  |

falling during recovery. They reuse the Polar-calibrated recipe, so they are not a separately tuned baseline. Other primes are not ruled out by these counts: at  $p = 47$  and 61, with  $r$  set per training fraction and  $\lambda_V = 0.2 \cdot 31/p$ , a post-hoc-window criterion holds in 155 of 160 fresh runs, though the primary endpoint holds in only 9 (Appendix C.4);  $p = 23$  was not rerun that way.

**Secondary endpoints.** Of the 120 primary runs, the pre-specified strict first-fit joint of Appendix A.3 passes in none, its pre-specified persistent-fit variant in 49, and the two representation-first flags in none. These checks do not alter the primary verdict, and no multiplicity correction is applied. The experiments do not establish a uniform refit-before-head chronology or a causal role for scalar Fourier growth.

### C.3 Results across primes

Depths 3 and 4 were also run at  $p = 23, 47, 61, 101$ , with width 256 and last-layer multiplier  $r = 0.45$  (depth 4 at  $p = 23$ : 0.32). Each setting is the best of a development sweep on one or two seeds, with a tuned last-layer multiplier  $r$  and all other layers at  $r_\ell = 1$ . It was fixed before any new run and then run on 8 or 9 fresh initialization/split seeds, for 10 runs per setting; at  $p = 101$  the settings chosen at  $p = 97$  were run unchanged on 10 fresh seeds. Depth 2 was run at  $p = 23, 31, 47$  with its development configurations; at  $p = 31$  (5 development and 5 fresh seeds) it passes the primary endpoint in 9 of 10 runs. Rerun with the same fixed rate on the 40 seed indices of the primary depths, as the first column of Figure 3, it passes in 34/40 (Wilson interval  $[0.71, 0.93]$ , whose lower end is below the 0.80 of the criterion that depths 3–5 meet); the

six misses fail the share threshold, and one also ends at held-out accuracy 0.984. Off  $p = 31$ , among depths 2 to 4 only depth 3 at  $p = 47$  passes (2/10); the other depth-3 and depth-4 settings at  $p = 23, 47, 61, 101$  pass in none of their 10 runs, and at  $p \geq 47$  some training pairs are misclassified inside the post-hoc window. The four settings chosen at  $p = 97$  (one hidden layer of width 512 or 2048, and depths 3 and 4) were also run at  $p = 97$ , 10 runs each, and none passes. These are development settings on a few seeds each, not a pre-specified evaluation like the depth-three-to-five experiment. Under the looser post-hoc rule every run of these settings passes (Appendix C.5). Appendix C.4 reruns depths 3 and 4 at  $p = 47$  and 61 with  $r$  chosen per training fraction, on fresh seeds and with a criterion fixed in advance.

## C.4 Depths three and four across training fractions

This experiment asks whether a re-tuned recipe carries the deeper-network behavior to other primes and smaller training sets. It runs depths 3 and 4 at  $p = 47$  and 61 with training fractions 0.8, 0.7, 0.6 and 0.5, in the configuration of Appendix C.1 (width 256, the same initialization scales, rates and decays), with two changes. The head decay scales with the prime,  $\lambda_V = 0.2 \cdot 31/p$  (0.132 at  $p = 47$  and 0.102 at  $p = 61$ ), and the last-layer multiplier  $r$  is fixed for each depth and fraction, without the per-run rule (26) (Table 3). Each  $r$  was chosen on 6 development runs per candidate value (3 seeds at each prime) by a fixed order: the most passes of the criterion below, then the most passes with training accuracy at least 0.97 throughout the window, then the larger median share, then the earlier median 90% held-out crossing. The criterion, the values of  $r$  and the analysis were fixed before ten fresh initialization/split seeds per setting were run, 160 runs in all; their seed indices were not used in any development run. One split seed per index nests the training sets across fractions, so each seed is a paired ladder. Runs stop at 100k, 150k, 200k and 250k updates at fractions 0.8, 0.7, 0.6 and 0.5, or once held-out accuracy has been at least 0.99 at every log for  $\max(0.5t, 10^4)$  updates since the start  $t$  of that unbroken stretch.

Table 3: Depths 3 and 4 at  $p = 47, 61$  across training fractions (10 fresh runs per setting;  $\lambda_V = 0.2 \cdot 31/p$ ). Pass: the criterion of Appendix C.4.  $\bar{L}/\log p$ : training CE (cycle mean) when held-out accuracy crosses 0.5.  $\Delta\bar{L}$ : its relative change from the refit’s 0.2 crossing to the held-out 0.9 crossing, with the number of runs in which it rose.  $\Delta A_H(W)$ : last hidden layer. Lead: held-out minus refit 0.5 crossing.  $t_{90}$ : held-out 0.9 crossing. Mis.: the most training pairs misclassified at any log inside  $W$ , over all 10 runs. Medians over the 10 runs unless stated; updates in thousands.

| $D$ | $p$ | Fraction | $r$ | Pass  | $\bar{L}/\log p$ | $\Delta\bar{L}$ (rose) | $\Delta A_H(W)$ | Lead | $t_{90}$ | Mis. |
|-----|-----|----------|-----|-------|------------------|------------------------|-----------------|------|----------|------|
| 3   | 47  | 0.8      | .45 | 10/10 | 0.52             | +9.5% (10)             | .51             | 2.4  | 19.5     | 0    |
| 3   | 47  | 0.7      | .30 | 10/10 | 0.35             | +49% (10)              | .54             | 2.8  | 24.2     | 0    |
| 3   | 47  | 0.6      | .18 | 10/10 | 0.15             | +233% (10)             | .48             | 3.7  | 34.1     | 0    |
| 3   | 47  | 0.5      | .08 | 8/10  | 0.007            | +1117% (10)            | .48             | 4.7  | 48.1     | 0    |
| 3   | 61  | 0.8      | .45 | 10/10 | 0.58             | +0.7% (8)              | .53             | 1.2  | 17.6     | 18   |
| 3   | 61  | 0.7      | .30 | 10/10 | 0.43             | +13% (10)              | .63             | 1.3  | 22.0     | 1    |
| 3   | 61  | 0.6      | .18 | 10/10 | 0.21             | +51% (10)              | .64             | 1.5  | 28.3     | 0    |
| 3   | 61  | 0.5      | .08 | 10/10 | 0.014            | +136% (10)             | .69             | 1.8  | 40.7     | 0    |
| 4   | 47  | 0.8      | .45 | 9/10  | 0.63             | −0.5% (5)              | .40             | 2.7  | 26.4     | 2    |
| 4   | 47  | 0.7      | .28 | 10/10 | 0.52             | +29% (10)              | .56             | 4.9  | 36.4     | 0    |
| 4   | 47  | 0.6      | .13 | 10/10 | 0.17             | +303% (10)             | .62             | 4.8  | 47.5     | 0    |
| 4   | 47  | 0.5      | .06 | 9/10  | 0.044            | +1312% (10)            | .59             | 9.5  | 88.3     | 0    |
| 4   | 61  | 0.8      | .45 | 9/10  | 0.67             | −2.9% (0)              | .38             | 1.5  | 23.3     | 14   |
| 4   | 61  | 0.7      | .28 | 10/10 | 0.56             | +11% (10)              | .61             | 2.1  | 31.0     | 4    |
| 4   | 61  | 0.6      | .13 | 10/10 | 0.23             | +69% (10)              | .68             | 2.0  | 41.0     | 0    |
| 4   | 61  | 0.5      | .06 | 10/10 | 0.045            | +238% (10)             | .69             | 3.3  | 67.8     | 0    |

**Criterion.** The criterion uses the post-hoc non-falling window  $W$  of Appendix A.3. A run passes if the train-only refit gains at least 0.2 inside  $W$ ; held-out accuracy gains at least 0.2 from the start of  $W$  to the end of the run and ends at least 0.99;  $W$  holds at least 75% of both the held-out and the refit 10%-to-90% rises; and  $A_H$  of the last hidden layer gains at least 0.02 across  $W$ . A setting works if at least 8 of its 10 runs pass. Training fit is not required, but adding training accuracy at least 0.97 at every logged update of  $W$  changes no count.

**Results.** All 16 settings work, and 155 of the 160 runs pass (Table 3, Figures 4 and 10–12). The runs are not independent: each seed index fixes the split and the initialization in all 16 settings, so at a given prime the training sets are nested across fractions and the same at both depths, and at a given depth and prime the initialization is the same across fractions. We therefore give counts per setting, not an interval for the pooled rate. Three of the five failures are at fraction 0.5 and  $p = 47$ : the whole held-out rise lies inside  $W$ , but the runs reach their 250k-update budget with held-out accuracy 0.959, 0.971 and 0.988. The other two are at fraction 0.8 and depth 4: the CE peaks and turns down during the held-out rise, so  $W$  closes early and holds 0.749 and 0.736 of it. In every setting the refit crosses 50% held-out accuracy before the network, by a median of 1.2k to 9.5k updates. The last hidden layer’s  $A_H$  gains 0.38 to 0.69 across  $W$  (setting medians), while the first hidden layer’s rises only to about 0.1.

**The training loss.** The training fraction sets the CE level at which the held-out rise happens; the values below are setting medians. At fraction 0.8 the training CE climbs after the fit to  $0.52\text{--}0.67 \log p$  and stays near that level: from the refit’s 20% crossing to the network’s 90% crossing it changes by  $-2.9\%$  to  $+9.5\%$ . In every run at this fraction the refit’s 20% crossing comes before the network’s 10% crossing, so the interval contains the network’s whole 10%-to-90% held-out rise. Training accuracy stays at or near one; at  $p = 61$ , 3 to 18 of the 2,976 training pairs are misclassified at some log inside  $W$ , and at  $p = 47$  at most 2 of 1,767. At smaller fractions the rise comes earlier on the CE’s climb: at fraction 0.5 the CE is  $0.007\text{--}0.045 \log p$  when held-out accuracy crosses 50%, and from the refit’s 20% crossing to the network’s 90% crossing it grows by 136% to 1312%, large relative changes at a small level.

**Primary endpoint.** On the same runs the primary endpoint of Section 5.2, which asks for the rise inside a flat band at the CE peak, holds in only 9 of 160 (4 of 10 for each depth at  $p = 47$  and fraction 0.8, and 1 of 10 for depth 4 at  $p = 61$  and fraction 0.7). At fractions 0.5 to 0.7 the CE peaks after the held-out rise, so the band holds too little of it; at fraction 0.8 and  $p = 61$  the band holds the rise, but some training pairs are misclassified inside it. These runs thus extend the post-hoc picture of Appendix C.5 to other primes and training fractions, under a criterion fixed in advance; they do not show the fitted, flat band of the primary endpoint at these primes.

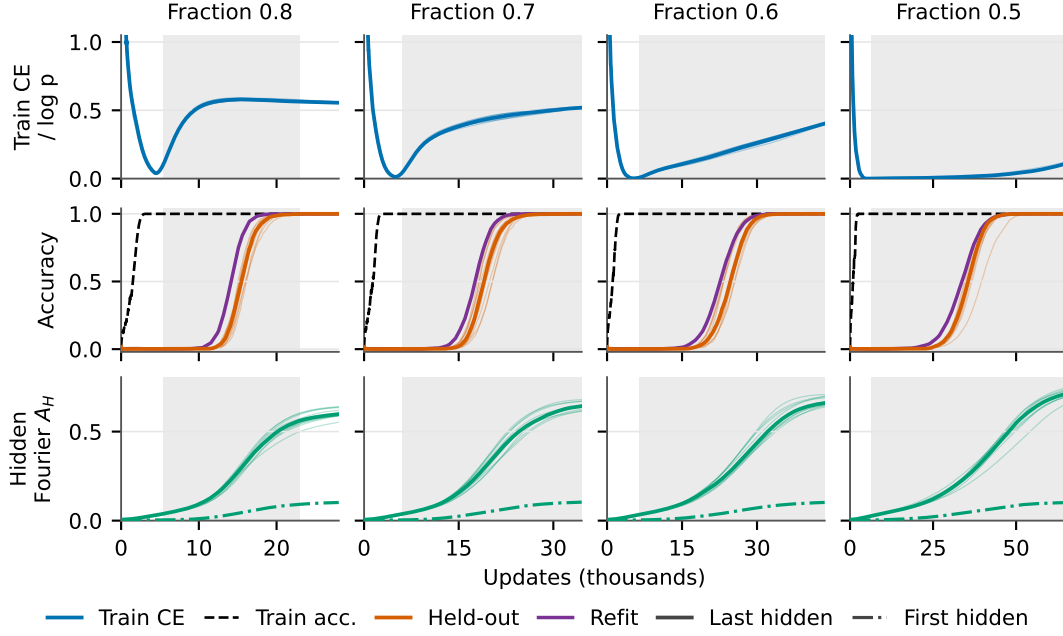


Figure 10: **Depth 3 at  $p = 61$  across training fractions: every run (thin) and the median (thick)**, in the format of Figure 4 (depth 3 at  $p = 47$ ); 10 fresh runs per fraction. Rows as in Figure 3; gray: from the median start to the median end of the post-hoc windows  $W$ . Table 3 gives  $r$ , the passes and the misclassified training pairs. Each column ends at its shortest run, so every run contributes at every plotted point; in this setting and in those of Figures 11 and 12, the shortest run ends after the last run's 90% held-out crossing.

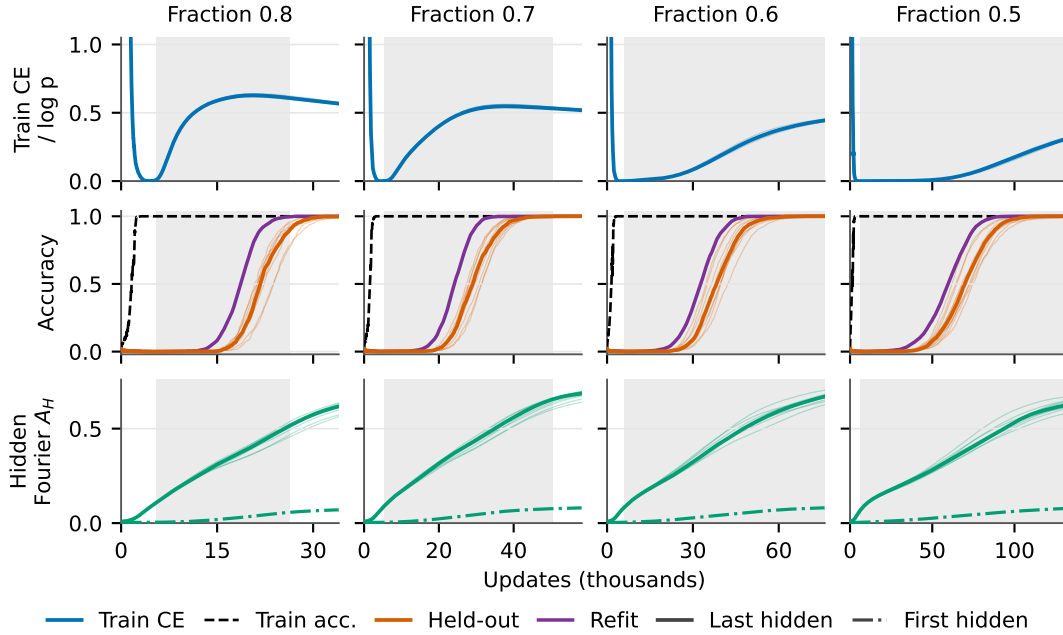


Figure 11: **Depth 4 at  $p = 47$  across training fractions: every run (thin) and the median (thick)**, as in Figure 10.

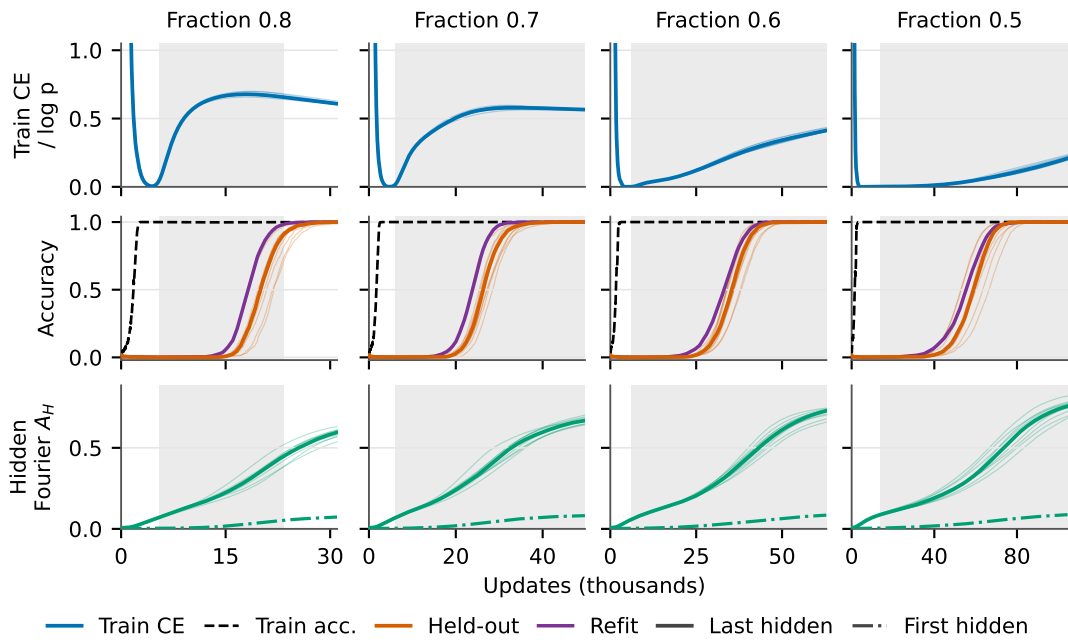


Figure 12: **Depth 4 at  $p = 61$  across training fractions: every run (thin) and the median (thick), as in Figure 10.**

## C.5 Post-hoc window counts

After every run had been seen, we also scored the runs with the post-hoc non-falling window of Appendix A.3. The following pass:

- with one hidden layer, all 20 headline runs and the 20 runs of the second  $p = 23$  cohort; all 20 per-prime runs at each of  $p = 17, 23, 29$ ; 20, 20, 20, 9, 0 and 0 fixed-recipe runs at  $p = 11, 13, 17, 19, 23, 29$ ; all five replayed trajectories at each of  $p = 23, 31$ ; and none of the 40 quadratic-activation runs;
- all 170 runs of the 17 development-selected settings with one to four hidden layers, all 30 runs at  $p = 101$  and all 120 primary runs;
- the fixed-rate arms (60/60),  $D = 4$  at  $p = 23$  (20/20) and width 512 (4/4), but none of the 20 normalized-gradient runs, whose CE keeps falling.

On the 127-run development sweep that chose 14 of the 17 settings, 115 runs pass. Outside that sweep and the one-hidden-layer cohorts of Appendix B, moving the threshold anywhere from 0.1 to 0.5 changes no count. A pass requires only a rise of 0.2 in held-out or refit accuracy, not recovery to 90%. Because the window is chosen to contain the largest CE rise, raw CE varies by orders of magnitude inside it at depths 3 to 5, and at  $p \geq 47$  nearly every run misclassifies some training pairs inside it. The rule thus shows held-out or refit accuracy rising while the training CE does not fall, not the fitted, flat window of the primary endpoint.

## C.6 Biases in the hidden layers

The baseline networks are bias-free. The dotted curves of Figures 2 and 3 come from a small first experiment with trained hidden biases. We reran the  $p = 31$  runs of Figure 2 (one hidden layer, 10 runs) and Figure 3 (depth 2, 40 runs; depth 4, 40 runs, each with its preliminary run that selects  $r$ ) with a bias in every hidden layer,  $h^{(\ell)} = \text{ReLU}(\mathbf{W}^{(\ell)}h^{(\ell-1)} + b^{(\ell)})$ . The bias starts at zero and is one more column of  $\mathbf{W}^{(\ell)}$ : the matrix  $[\mathbf{W}^{(\ell)} \mid b^{(\ell)}]$  receives the supported Polar step and the decay of (3). The head stays bias-free. Each run starts from the same network as its bias-free counterpart and keeps its split, fixed hyperparameters, rate-selection rule and stopping rule. At depth 4 the preliminary run, which now has biases, reselects  $r$ , and the selected value differs from the bias-free one in 37 of the 40 runs. Depths 3 and 5 were not rerun. With one-hot inputs a first-layer bias adds no functions, because it can be folded into the first  $p$  columns of  $\mathbf{W}^{(1)}$ ; it changes only the update. In Figures 2 and 3 the dotted medians stay close to the bias-free ones; the most visible difference is a lower training loss during and after the held-out rise. For one hidden layer, the relative decrease of  $\bar{L}$  in Figure 13 (left) has median 5.3% with the bias and 2.6% without, and is larger in all 10 pairs. With the bias, the captured share  $q(I_C)$  has median 0.62 (range 0.40–0.74), below the 0.75 that the primary endpoint requires; without it, the median is 1. At depth 2 the primary endpoint holds in 18 of the 40 runs with biases and in 34 without: 16 seeds lose it with the biases and none gains it (exact McNemar  $p = 3.1 \cdot 10^{-5}$ ), and 20 of the 22 misses with biases fail the share threshold.

With one hidden layer we also tried three other ways of updating the bias, on the same 10 initializations and splits (Figure 13): (i) the joint step until the knee, then the bias frozen (no step and no decay); (ii) the joint step until the knee, then the bias only decays, which is the bias-free optimizer from the knee on, applied to the network with the bias folded into  $\mathbf{W}$ ; (iii) from the start,  $\mathbf{W}$  takes the bias-free step of (3) and the bias a step of its own,  $b_{t+1} = (1 - \eta_W \lambda_W)b_t + \eta_W g_t / (\|g_t\|_2 \sqrt{\text{rank } \mathbf{F}_t})$  with  $g_t = -\nabla_b L$ , of about the length of the bias part of the joint step. In these runs the training loss falls about as much as without a bias in (i) and (ii), and about as much as with the joint step in (iii). These experiments are preliminary: they use one prime and one width per depth, and we do not yet draw conclusions from them.

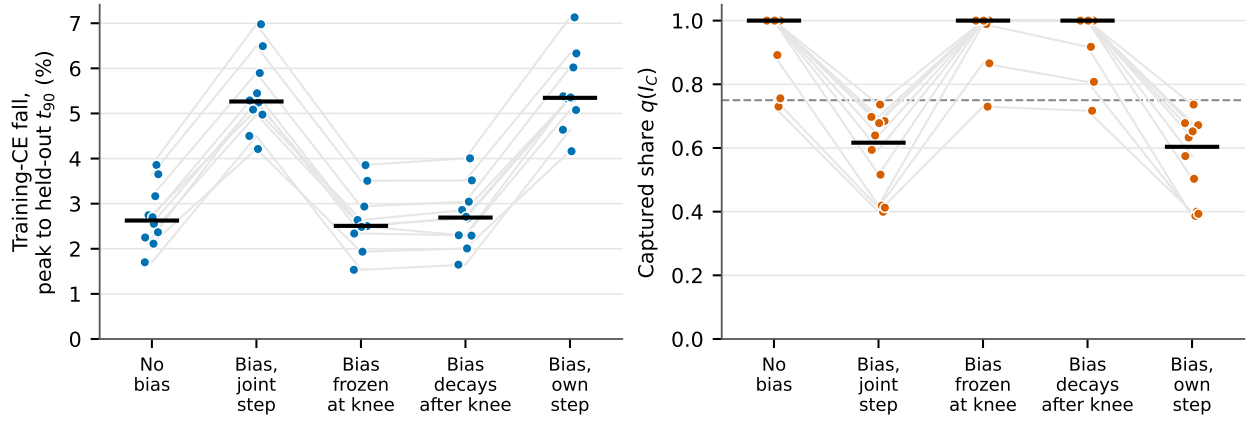


Figure 13: **Updating a hidden bias in different ways** (one hidden layer,  $p = 31$ , 10 initializations and splits; gray lines join the same run, bars are medians). Left: the relative decrease of  $\bar{L}$  from its largest post-knee value to the first update after which held-out accuracy stays at least 0.9. Right: the captured share  $q(I_C)$  of the primary endpoint, which requires at least 0.75 (dashed). Columns: no bias; the bias with the joint  $[\mathbf{W} \mid b]$  step; the joint step until the knee, then the bias frozen or only decaying; the bias updated by its own step throughout.

## D A one-block transformer

This appendix supports Section 5.3. It reports every run of the settings below. The settings were chosen on earlier development runs with other seeds, the gain on five pilot runs of one seed. Initialization and split seeds are  $7100 + i$  and  $8100 + i$ , with  $i = 11\text{--}20$  for the main runs at  $p = 113$ , the runs without the gain, the settings of Figures 14–17 and two blocks at  $\phi = 0.2$ ,  $i = 11\text{--}13$  for the recipe sweep and the two-block tests (Table 6 and the fractions 0.175 and 0.225), and  $i = 11\text{--}15$  elsewhere, so settings at the same prime and with the same number of blocks share their initial weights, whatever their gain, rates or split (one- and two-block runs share only the embeddings and the first block), and settings at the same prime share their splits when the training fraction agrees. These runs test settings found on earlier runs. The pass thresholds were fixed before all of them except the Polar runs without the gain, which had finished before and were scored afterwards (Appendix D.2). Seeds  $i = 16\text{--}20$  of the settings across primes and of two blocks were run last, after the settings had been chosen on seeds  $i = 11\text{--}15$ .

### D.1 Model, updates and readouts

**Model.** Token and position embeddings  $\mathbf{W}_E \in \mathbb{R}^{(p+1) \times d}$  and  $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{3 \times d}$ ,  $d = 128$ , feed one block: causal attention with four heads ( $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O \in \mathbb{R}^{d \times d}$ ) and a ReLU MLP ( $\mathbf{W}_{\text{in}} \in \mathbb{R}^{512 \times d}$ ,  $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d \times 512}$ ), each added to the residual stream, with no LayerNorm and no biases. The token  $p$  stands for “=”. Only the last position is computed, since the others cannot reach its logits. Its final residual  $r \in \mathbb{R}^d$  gives  $\hat{r} = r / (\frac{1}{d} \sum_i r_i^2 + 10^{-12})^{1/2}$  and the logits  $f = g \mathbf{V} \hat{r}$ , with the unembedding  $\mathbf{V} \in \mathbb{R}^{p \times d}$  ( $\mathbf{W}_U^\top$  in [2]). Initial entries are independent Gaussians with standard deviation  $1/\sqrt{d}$  for  $\mathbf{W}_E$ ,  $\mathbf{W}_{\text{pos}}$ , the attention matrices and  $\mathbf{V}$ ,  $\sqrt{2/d}$  for  $\mathbf{W}_{\text{in}}$  and  $1/\sqrt{512}$  for  $\mathbf{W}_{\text{out}}$ ; the gain starts at  $g_0$ .

**Updates.** Training uses the full-batch training CE in float64. Every matrix except  $\mathbf{V}$  takes the  $\mathbf{W}$  step of (3), with the supported Polar factor computed by an SVD that keeps singular values above  $10^{-7}$  times the largest;  $\mathbf{V}$  takes the  $\mathbf{V}$  step. The gain takes  $g \leftarrow g + \eta_g (-\text{sign}(\partial L / \partial g) - \lambda_g g)$ . Once the training set is fitted,  $\partial L / \partial g < 0$ , so with  $\lambda_g = 1$  the gain relaxes toward 1 with time constant  $\tau = 1/(\eta_g \lambda_g)$  ( $\eta_g = 2 \cdot 10^{-4}$  and  $\tau = 5,000$  unless stated). Runs stop after 30k updates, or once held-out accuracy on all held-out pairs has been at least 0.99 at every log from the start  $t$  of an unbroken stretch to  $\max(1.5t, t + 5000)$ .

**Splits and held-out sets.** A random split keeps  $\lfloor \phi p^2 \rfloor$  of the  $p^2$  pairs for training, with  $\phi$  the training fraction (2,553 at  $p = 113$  and  $\phi = 0.2$ ). About a fraction  $\phi$  of the held-out pairs have their mirror  $(b, a)$  in the training set. They can be answered by memorizing the mirror, which the shared token embedding makes easy: at the first full fit at  $p = 113$  they are 28% correct (median over runs), against 0.06% for the other held-out pairs. We therefore report every quantity on all held-out pairs and on the *no-mirror* pairs, whose mirror is also held out (8,133–8,193 of the 10,216 held-out pairs at  $p = 113$ ). A *mirror-free* split instead adds unordered pairs  $\{a, b\}$  in a random order, each with both orders, until it holds  $\lfloor \phi p^2 \rfloor$  ordered pairs, so no held-out pair has its mirror in the training set.

**Observers.** Accuracies, margins and misclassified training pairs are logged every 20 updates,  $A_H$  and the refits every 100, and the probe of Appendix D.5 every 200.  $A_H$  is (6) on the full-grid table of  $\hat{r}$ , taken before the gain. This table is the residual stream after the MLP’s output map  $\mathbf{W}_{\text{out}}$ , one linear map after its neurons, so it corresponds to the logits of the networks of Section 5.1 rather than to their hidden layer; the transformer’s MLP neurons end the main runs with  $A_H$  of only 0.03–0.04. The refit is (7) on the same table, with  $\tau_t = 10^{-3}n$ .

**Fourier-filtered refit.** With  $\omega_k = 2\pi k/p$  and  $K = (p-1)/2$ , the target  $T_x = e_y - \mathbf{1}_p/p$ ,  $y = y^*(x)$ , has entries  $\frac{2}{p} \sum_{k=1}^K [\cos(\omega_k y) \cos(\omega_k c) + \sin(\omega_k y) \sin(\omega_k c)]$  for  $c \in \mathbb{F}_p$ . The ridge solution is linear in the targets, so the refit’s logits are

$$\hat{f}_c(x) = \frac{2}{p} \sum_{k=1}^K [C_k(x) \cos(\omega_k c) + S_k(x) \sin(\omega_k c)], \quad (27)$$

Table 4: One block at  $p = 113$ ,  $\phi = 0.2$ : the gain (seeds  $i = 11$ –20; the 3-run rows use  $i = 11$ –13). Fit: first full fit (median; updates).  $\Delta\tilde{L}$ : relative change of the averaged training CE from the no-mirror held-out 0.1 crossing to the 0.9 crossing (median and range).  $q$ : share of the all-pairs rise inside the all-pairs window. Final: held-out accuracy on all pairs.

| Setting                       | Runs | Pass | Fit   | $\Delta\tilde{L}$ | $q$       | Final       | Failing condition     |
|-------------------------------|------|------|-------|-------------------|-----------|-------------|-----------------------|
| Polar, gain $2 \rightarrow 1$ | 10   | 10   | 200   | +33% (+18, +55)   | 0.93–1.00 | 0.999–1.000 | –                     |
| Polar, no gain                | 10   | 0    | 1,190 | –8% (–10, –7)     | 0.35–0.51 | 0.999–1.000 | share (10), refit (9) |
| Polar, fixed $g = 1.5$        | 3    | 0    | 240   | –17% (–18, –16)   | 0.20–0.26 | 0.999–1.000 | share, refit (3)      |
| Polar, fixed $g = 2$          | 3    | 0    | 200   | –27% (–29, –24)   | 0.13–0.17 | 1.000       | share, refit (3)      |
| Polar, fixed $g = 3$          | 3    | 0    | 160   | –44% (–49, –42)   | 0.07–0.10 | 1.000       | share, refit (3)      |

where  $C_k$  and  $S_k$  are the ridge predictions, from the same table and penalty, of  $\cos(\omega_k y)$  and  $\sin(\omega_k y)$ . The refit thus weights every frequency equally. The Fourier-filtered refit keeps in (27) only the 8 frequencies with the largest training energy  $E_k = \left\| \frac{1}{n} \sum_{x \in \mathcal{S}} (\hat{r}(x) - \bar{r}_{\mathcal{S}}) e^{-i\omega_k y^*(x)} \right\|_2^2$ , where  $\bar{r}_{\mathcal{S}}$  is the training mean of  $\hat{r}$ , so it uses training data only; with all  $K$  frequencies kept it is the refit itself. The number 8 was fixed at  $p = 113$  on development runs, before the runs reported here.

## D.2 Window and pass criterion

Let  $\tilde{L}_t$  be the training CE averaged over the 201 updates centered at  $t$  (the series is padded with its first and last values), and  $T_{\text{fit}}$  the first log at which every training pair is correctly classified. For each start  $a \in \{T_{\text{fit}}, T_{\text{fit}} + 20, \dots\}$ , let  $b(a)$  be the last update before  $\tilde{L}_t$  first falls below  $0.97 \max_{a \leq s \leq t} \tilde{L}_s$ , or the end of the run. The window is the  $[a, b(a)]$  with the largest captured share  $q$  of the held-out 10%-to-90% rise, with  $q$  as in Appendix A.3 and the earliest start on ties. Unlike the post-hoc window of Appendix A.3, it has candidate starts at or after the first fit rather than after a head-norm knee, uses a 201-update average and is placed using held-out accuracy, so it is post hoc. It is computed separately for each held-out definition. A run passes on a definition if inside its window  $q \geq 0.75$ , the Fourier-filtered refit’s held-out accuracy gains at least 0.2,  $\Delta A_H \geq 0.02$  for  $\hat{r}$ , and no training pair is misclassified at any log (*strict*; *loose* allows up to 10). A run passes if it passes on both definitions and its final accuracy on all held-out pairs is at least 0.99; unless stated, counts are strict. These thresholds were fixed before the runs at  $p = 113$  with the gain, on the no-mirror definition; the requirement on both definitions was added after those runs and before every other run reported here, except the Polar runs without the gain, which predate both and were scored afterwards.

## D.3 The gain and the recipe at $p = 113$

Table 4 compares the runs with and without the gain. Without the gain, the runs fit the training set later (median 1,190 updates) and their CE falls during the held-out rise, so the window holds only 0.35 to 0.51 of it on all held-out pairs (0.43 to 0.61 on the no-mirror pairs). A gain held fixed behaves like no gain. The relaxing gain lifts the CE during the rise: in the 10 main runs, the averaged training CE never falls more than 0.05% below its running maximum across the no-mirror held-out rise.

**The recipe.** On seeds  $i = 11$ –13 we varied the initial gain  $g_0$  and the time constant  $\tau$  (Table 5). Every run with  $g_0 \in \{2, 3\}$  and  $\tau$  from 2,500 to 10,000 passes (18/18), with a median CE rise of 24% to 104% per setting across the held-out rise; with  $\tau = 20,000$  the CE is about flat and 6 of 9 runs pass. Changing one setting at a time around the main runs ( $g_0 = 2$ ,  $\tau = 5,000$ ): head decay  $\lambda_V = 0.25$  or 0.3, and feature rate  $\eta_W = 0.05$  or 0.2, pass in 3/3 each ( $\eta_W = 0.2$  with a CE rise of only 10%); feature decay  $\lambda_W = 0.005$  passes in 1/3 (two runs end below 0.99) and  $\lambda_W = 0.02$  in 2/3; training fraction 0.3 in 0/3, because the held-out rise then comes before the CE minimum. The recipe is thus a gain of 2 to 3 that relaxes to 1 on about the time scale of the held-out rise, supported-Polar feature steps, and a training fraction that puts the rise after the CE minimum.

**Two blocks.** Two blocks pass in none of 25 runs at four settings. At  $\phi = 0.2$  ( $\lambda_V = 0.275$ ,  $\tau = 10,000$ ; seeds  $i = 11$ –20, Figures 5 and 15) they behave like one block during the held-out rise: the no-mirror held-out accuracy goes from 10% at 4.5k–6.1k updates to 90% at 8.1k–10.8k, the CE rises by 9% to 29% across

Table 5: The gain at  $p = 113$ ,  $\phi = 0.2$  (seeds  $i = 11-13$ ): passes of 3 and the median change of the averaged training CE across the no-mirror held-out rise. The  $g_0 = 2$ ,  $\tau = 5,000$  entry uses the first three main runs. With  $g_0 = 1.5$  and  $\tau = 2,500$ , all three runs pass the loose criterion.

|             | $\tau = 2,500$ | 5,000      | 10,000    | 20,000   | fixed     |
|-------------|----------------|------------|-----------|----------|-----------|
| $g_0 = 1.5$ | 1/3, +10%      | 3/3, +13%  | 2/3, +6%  | 2/3, -4% | 0/3, -17% |
| $g_0 = 2$   | 3/3, +28%      | 3/3, +38%  | 3/3, +24% | 2/3, -1% | 0/3, -27% |
| $g_0 = 3$   | 3/3, +84%      | 3/3, +104% | 3/3, +58% | 2/3, +9% | 0/3, -44% |

Table 6: Two blocks at  $\phi = 0.2$  ( $\lambda_V = 0.275$ ,  $\tau = 10,000$ ): one rate changed at update 12k (seeds  $i = 11-13$ ; no change:  $i = 11-20$ ). Window ends: end of the all-pairs window, in updates. First flip: first log after 12k with a misclassified training pair; Most: the most misclassified training pairs at a log after 12k. Final: held-out accuracy on all pairs.

| Change at 12k                                  | Pass | Window ends | First flip  | Most   | Final       |
|------------------------------------------------|------|-------------|-------------|--------|-------------|
| None                                           | 0/10 | 30k (end)   | 17.8k–22.0k | 45–438 | 0.941–0.994 |
| Gain frozen ( $\eta_g = 0$ , $g = 1.30$ )      | 2/3  | 13.5k–13.7k | 18.8k–19.7k | 57–217 | 0.983–0.997 |
| Head step $\eta_V$ : 0.03 $\rightarrow$ 0.01   | 0/3  | 30k (end)   | 21.6k–23.0k | 59–395 | 0.950–0.970 |
| Feature step $\eta_W$ : 0.1 $\rightarrow$ 0.03 | 2/3  | 12.0k       | none        | 0      | 0.986–0.997 |

that rise, no training pair is misclassified at any log, the window holds the whole no-mirror rise, and the Fourier-filtered refit gains 0.64–0.98 inside it. Held-out accuracy on all pairs reaches 0.99 at 14.9k–18.2k updates. At 17.8k–22.0k updates, 1.4k to 5.7k updates later, training pairs begin to flip: 45 to 438 are misclassified at some log inside the window, which runs to the end of the run, held-out accuracy falls back, and 6 of the 10 runs end below 0.99. At  $\phi = 0.25$  ( $\lambda_V = 0.4$ ,  $\tau = 10,000$ ) and 0.3 ( $\lambda_V = 0.275$  or 0.5,  $\tau = 5,000$ ; seeds  $i = 11-15$ ) the rise comes before the CE minimum: the window holds at most 0.45 of the no-mirror rise in 14 of the 15 runs, and the fifteenth has no window.

On seeds  $i = 11-13$  at  $\phi = 0.2$  we then changed one rate at update 12k, after the held-out rise and before the late flips (Table 6). The three changed runs of a seed share their first 12k updates; through floating-point differences between thread counts they part from the runs of the same seed above after 2k–3k updates. Freezing the gain at 1.30 ends the CE’s rise: the CE falls slowly from 12k on, so the window closes at 13.5k–13.7k updates, and pairs still flip, after the window, so the two runs that pass lose their fit as well. A smaller head step delays the flips to 21.6k–23.0k. A smaller feature step keeps every training pair correct at every log to the end in all three runs; its CE dips by about 8% just after the change and then rises again, so the window closes at the change, and held-out accuracy on all pairs reaches 0.99 later (20.1k and 23.8k updates; the third run ends at 0.986). Training fractions 0.175 and 0.225 (3 runs each, base rates) pass in none: at 0.175 the no-mirror rise ends only at 14.3k–16.2k, pairs begin to flip at 17.7k–22.4k and held-out accuracy ends at 0.82–0.91; at 0.225 the rise starts (2.6k–3.3k) before the CE minimum (4.9k–5.2k) and the CE falls by 13% to 17% across it. Lowering the feature rate after the held-out rise is thus the one change that kept the fit; it has been tried on three seeds only.

## D.4 Other primes and mirror pairs

At other primes we keep the recipe ( $g_0 = 2$ ,  $\tau = 5,000$  and the rates of the main runs) and choose the head decay  $\lambda_V$  and the training fraction  $\phi$  (Table 7). In a first set of runs ( $\phi = 0.3$  at  $p = 61$  and 0.2 at  $p = 89$  and 149), the held-out rise was too slow at  $p = 61$  and 89 (most runs end below 0.99 after 30k updates) and came before the CE minimum at  $p = 149$ . A second set of runs set  $\phi$  for each prime by a rule, fixed before those runs, for when the rise starts as a function of  $n/(p \ln p)$ ; it predicted the median start of each new setting within a factor 0.63–1.34. The random-split settings chosen on seeds  $i = 11-15$ , the most strict passes at each prime with ties broken by loose passes and then by the smaller fraction, pass in 3/5, 5/5 and 3/5 runs at  $p = 61$ , 89 and 149, and in 2/5, 5/5 and 4/5 on the five further seeds, so in 5/10, 10/10 and 7/10 (10/10 at  $p = 113$ ). At  $p = 149$  all ten pass the loose criterion (at most 7 of the 3,552 training pairs are misclassified inside the window). At  $p = 61$  all ten pass on the no-mirror pairs, and the five failures come from the mirror pairs; with the mirror-free split below, which Figure 14 shows at  $p = 61$ , 9 of 10 pass.

Table 7: One block across primes (gain  $2 \rightarrow 1$ ,  $\tau = 5,000$ ; seeds  $i = 11-15$ , and  $i = 11-20$  in the rows with 10 runs). Pass: strict, on both definitions; Loose: at most 10 misclassified training pairs; No-mirror: strict on the no-mirror definition alone, with the final all-pairs accuracy at least 0.99.  $\Delta\tilde{L}$ : median change of the averaged training CE across the no-mirror rise, over the runs with a full rise. Final: held-out accuracy on all pairs. <sup>†</sup>Mirror-free split.

| $p$ | $\phi$            | $\lambda_V$ | Training pairs | Pass  | Loose | No-mirror | $\Delta\tilde{L}$ | Final        |
|-----|-------------------|-------------|----------------|-------|-------|-----------|-------------------|--------------|
| 61  | 0.30              | 0.45        | 1,116          | 1/5   | 1/5   | 1/5       | +87%              | 0.979–0.994  |
| 61  | 0.30              | 0.60        | 1,116          | 1/5   | 1/5   | 1/5       | +58%              | 0.964–0.991  |
| 61  | 0.33              | 0.45        | 1,227          | 5/10  | 5/10  | 10/10     | +106%             | 0.994–0.999  |
| 61  | 0.35              | 0.45        | 1,302          | 0/5   | 0/5   | 5/5       | +87%              | 0.998–0.999  |
| 61  | 0.55 <sup>†</sup> | 0.45        | 2,046          | 9/10  | 9/10  | 9/10      | +161%             | 0.9899–0.997 |
| 89  | 0.20              | 0.33        | 1,584          | 1/5   | 1/5   | 1/5       | +33%              | 0.39–0.996   |
| 89  | 0.20              | 0.40        | 1,584          | 0/5   | 0/5   | 0/5       | +22%              | 0.55–0.98    |
| 89  | 0.25              | 0.33        | 1,980          | 10/10 | 10/10 | 10/10     | +35%              | 1.000        |
| 89  | 0.275             | 0.33        | 2,178          | 2/5   | 2/5   | 5/5       | +4%               | 1.000        |
| 113 | 0.20              | 0.275       | 2,553          | 10/10 | 10/10 | 10/10     | +33%              | 0.999–1.000  |
| 149 | 0.15              | 0.23        | 3,330          | 0/5   | 4/5   | 0/5       | +28%              | 0.12–0.999   |
| 149 | 0.16              | 0.23        | 3,552          | 7/10  | 10/10 | 7/10      | +19%              | 0.992–1.000  |
| 149 | 0.20              | 0.23        | 4,440          | 0/5   | 0/5   | 0/5       | −23%              | 0.999–1.000  |
| 149 | 0.20              | 0.275       | 4,440          | 0/5   | 0/5   | 0/5       | −18%              | 0.996–1.000  |

**Mirror pairs.** A window that opens when the all-pairs held-out accuracy is  $a_0$  can hold at most  $(0.9 - \max(a_0, 0.1))/0.8$  of the all-pairs rise, so the share falls below 0.75 whenever  $a_0 > 0.3$ . At  $p = 61$  the training fraction that lets the rise finish within 30k updates also puts the mirrors of about a third of the held-out pairs in the training set. These held-out pairs are learned early: at  $\phi = 0.33$ , 90% of them are correct by update 240–520 in 8 of the 10 runs (by 711 and 1,615 in the other two), and the all-pairs window opens at 560–700. At  $\phi = 0.33$  the window therefore opens at  $a_0 = 0.27-0.34$ , and the all-pairs share sits near the threshold: 0.751–0.787 in the five passing runs, 0.694–0.744 in the five failing ones (Figure 16). A mirror-free split at  $\phi = 0.55$  has about as many distinct unordered training pairs (1,038–1,043) as the random split at 0.33 (1,021–1,039). With it,  $a_0$  is 0.007–0.036, the share is 1.00, and 9 of 10 runs pass; the tenth ends at held-out accuracy 0.9899. The CE minimum (median update 724, against 741 with the random split) and the 10% crossing of the no-mirror held-out accuracy (1.9k, against 1.8k) are close to those of the random split, but the rise is slower: the 50% and 90% crossings come at 4.3k and 7.6k, against 3.5k and 5.4k. At  $\phi = 0.33$  a mirror-free split, with 622 distinct pairs, did not generalize within 15k updates (one development run). The number of distinct unordered training pairs, rather than of ordered ones, thus appears to set when the held-out rise starts, and in a random split the held-out pairs whose mirror is a training pair behave almost like training pairs.

## D.5 Readouts

Table 8 and Figure 17 compare the readouts of  $\hat{r}$  with the network on the no-mirror pairs. The refit trails the network in all 50 runs of the appendix settings, at the 50% and at the 90% crossing. The Fourier-filtered refit leads at the 50% crossing in 49 of 50 runs and at the 90% crossing in all 30 runs at  $p \geq 89$ , but at  $p = 61$  it levels off near 0.96 (median final 0.957 with the random split and 0.962 with the mirror-free split) and crosses 90% after the network. At the network’s 90% crossing, 11 to 19 of the 30 frequencies at  $p = 61$  carry held-out signal (held-out  $R^2$  of  $(C_k, S_k)$  above 0.05, a diagnostic that uses held-out labels), against 9 to 15 of 44 to 74 at the larger primes. The top 8 keep 7 or 8 of them; the top 16 keep 11 to 16 at  $p = 61$  and all of them in 29 of the 30 runs at the larger primes, and the top-16 refit leads at both crossings in 49 of the 50 runs. The choice  $K = 16$  was made after seeing  $p = 61$ , so this readout is post hoc. A train-only multinomial logistic probe of  $\hat{r}$  tracks the network (median leads of 0.02k to 0.19k updates). Its logits are  $\mathbf{B}\hat{r}$  with  $\mathbf{B} \in \mathbb{R}^{p \times d}$ , without bias or feature standardization; every 200 updates it is refitted from  $\mathbf{B} = 0$  to minimize the mean training CE plus  $5 \cdot 10^{-5} \|\mathbf{B}\|_F^2$  by at most 100 L-BFGS iterations (strong-Wolfe line search, float64), and it is scored like the network, by the held-out accuracy of its argmax, with crossings interpolated between these logs. Whether a train-only readout leads the network thus depends on which frequencies it sums: the refit weights all frequencies equally, while at the network’s 50% crossing only 6 to 13 of them carry held-out signal

Table 8: Readout leads over the network on the no-mirror pairs: the network’s crossing minus the readout’s, in thousands of updates (positive: the readout crosses first), at the network’s 0.5 and 0.9 crossings; medians, with the number of leading runs in parentheses. Settings of Figure 14, and  $p = 61$  with the random split at  $\phi = 0.33$ ;  $^\dagger$ mirror-free split,  $\phi = 0.55$ . Network’s logits, top 8: the network’s own centred logits projected, in class space, onto the 8 frequencies of the Fourier refit, without fitting.

| Readout                 | Crossing | $p = 61^\dagger$ | $p = 61$   | $p = 89$   | $p = 113$  | $p = 149$  |
|-------------------------|----------|------------------|------------|------------|------------|------------|
| Refit, all frequencies  | 0.5      | −0.70 (0)        | −0.68 (0)  | −0.78 (0)  | −1.31 (0)  | −1.51 (0)  |
|                         | 0.9      | −0.68 (0)        | −0.72 (0)  | −1.26 (0)  | −2.15 (0)  | −1.86 (0)  |
| Fourier refit, top 8    | 0.5      | +0.78 (9)        | +0.80 (10) | +0.56 (10) | +0.86 (10) | +1.18 (10) |
|                         | 0.9      | −2.46 (0)        | −1.16 (0)  | +0.45 (10) | +0.90 (10) | +1.21 (10) |
| Network’s logits, top 8 | 0.5      | +0.79 (9)        | +0.76 (10) | +0.52 (10) | +0.81 (10) | +1.06 (10) |
|                         | 0.9      | −3.99 (0)        | −1.70 (0)  | +0.34 (10) | +0.74 (10) | +1.04 (10) |
| Fourier refit, top 16   | 0.5      | +1.16 (10)       | +0.73 (10) | +0.41 (10) | +0.50 (10) | +0.65 (10) |
|                         | 0.9      | +0.87 (9)        | +0.61 (10) | +0.34 (10) | +0.47 (10) | +0.73 (10) |
| Logistic probe          | 0.5      | +0.05 (8)        | +0.02 (8)  | +0.02 (9)  | +0.03 (7)  | +0.06 (9)  |
|                         | 0.9      | +0.18 (9)        | +0.08 (9)  | +0.04 (9)  | +0.04 (10) | +0.19 (10) |
| Runs                    |          | 10               | 10         | 10         | 10         | 10         |

in any run.

**Why the same refit leads with one hidden layer and trails here.** By (27) the refit’s logits are an equal-weight sum over all  $K$  frequencies, so it tests whether the features reproduce the one-hot label at every frequency. In replays of three runs of Figure 2 ( $p = 31, 61$  and  $101$ , one seed each, with the parameters saved every 200 updates), every frequency of the hidden layer carries held-out signal by update 6.4k, 5.2k and 5.6k, while the network’s held-out accuracy is still below 0.05; the network crosses 50% at 8.2k, 6.6k and 6.3k. There the full sum is the right readout: keeping only the top 8 frequencies delays it (leads at the network’s 50% crossing of 2.4k, 1.3k and 0.7k updates, against 2.8k, 1.9k and 1.4k with all frequencies). In the transformer,  $\hat{r}$  carries 6 to 13 frequencies at the network’s 50% crossing (medians 7 to 9 per setting of Table 8), and the frequencies outside the top 8 have negative held-out  $R^2$  (medians of  $-0.27$  to  $-0.79$  per setting). The ridge fits them on the training pairs, where the features encode the labels: in replays of one run of Figure 5 and one of Figure 15, at the network’s 50% crossing, their training  $R^2$  is 0.38 and 0.53 and their held-out  $R^2$  is  $-0.42$  and  $-0.49$ . The network’s own logits, restricted to the same top 8 frequencies without any fitting, lead its held-out accuracy by about as much as the Fourier-filtered refit (Table 8), so the restriction removes memorized terms rather than adding signal that the network lacks. This matches what is known of the two kinds of solution: the sufficiently wide bias-free one-hidden-layer network with quadratic activation and maximum  $L_{2,3}$ -margin on the full grid uses every frequency [8], and the one-layer transformer of [2] uses five.

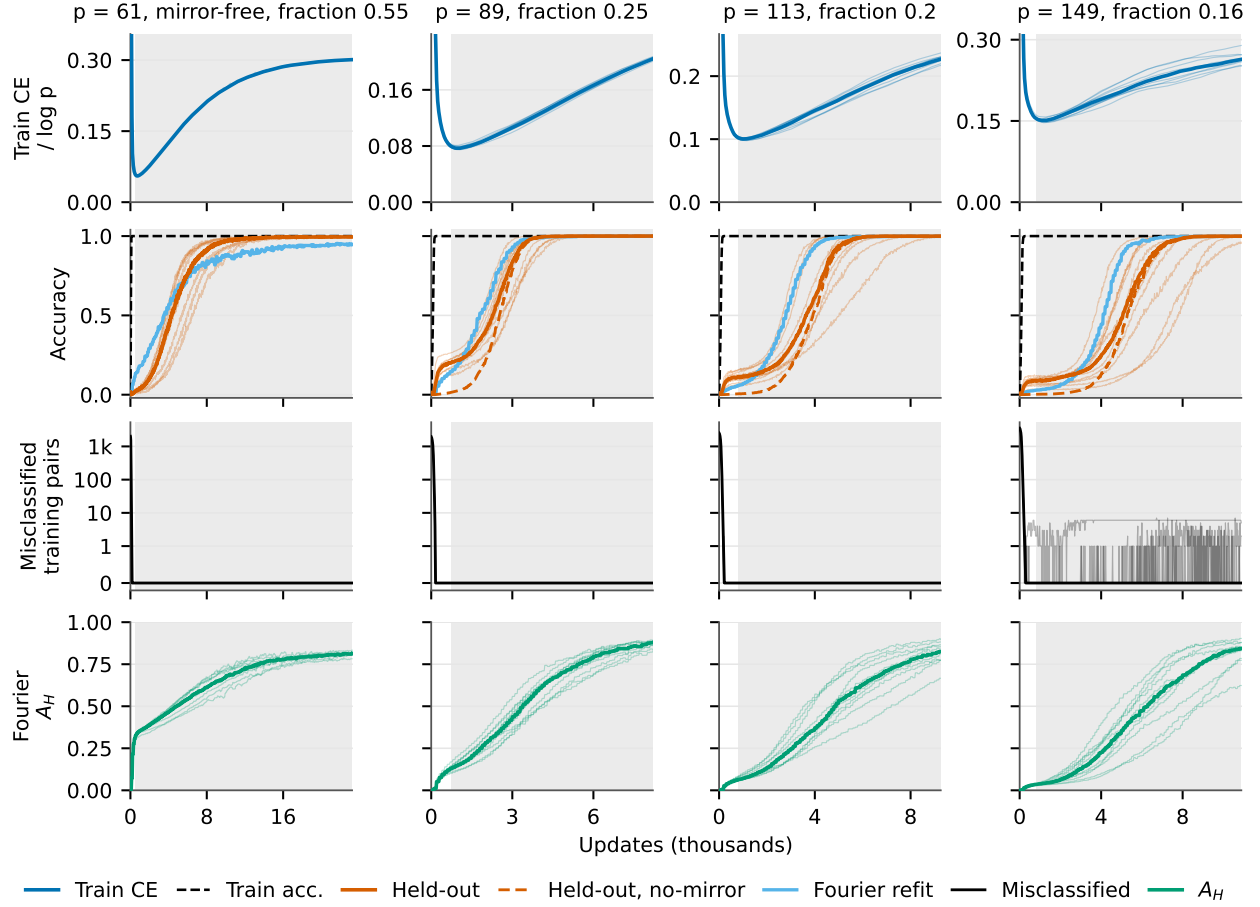


Figure 14: **One block across primes: every run (thin) and the median (thick)** (gain  $2 \rightarrow 1$ ,  $\tau = 5,000$ ;  $\lambda_V = 0.45, 0.33, 0.275$  and  $0.23$ ; 10 runs per prime, seeds  $i = 11-20$ ). At  $p = 61$  the training set is the mirror-free split of Appendix D.4 (fraction 0.55), so no held-out pair has its mirror in it and the two held-out accuracies coincide. At each  $p$  the training fraction puts the start of the held-out rise after the CE minimum. Rows as in Figure 5, plus the misclassified training pairs (symmetric log scale): the pass rule requires none inside the window, and the accuracy row cannot show a few pairs out of thousands. Gray: from the median start to the median end of the windows on all held-out pairs. Each column ends at its shortest run. Strict passes: 9/10, 10/10, 10/10 and 7/10; the run that fails at  $p = 61$  ends at held-out accuracy 0.9899, and at  $p = 149$  all 10 runs pass the loose criterion. With a random split at fraction 0.33,  $p = 61$  passes in 5/10, and all five failures come from the mirror pairs (Figure 16).

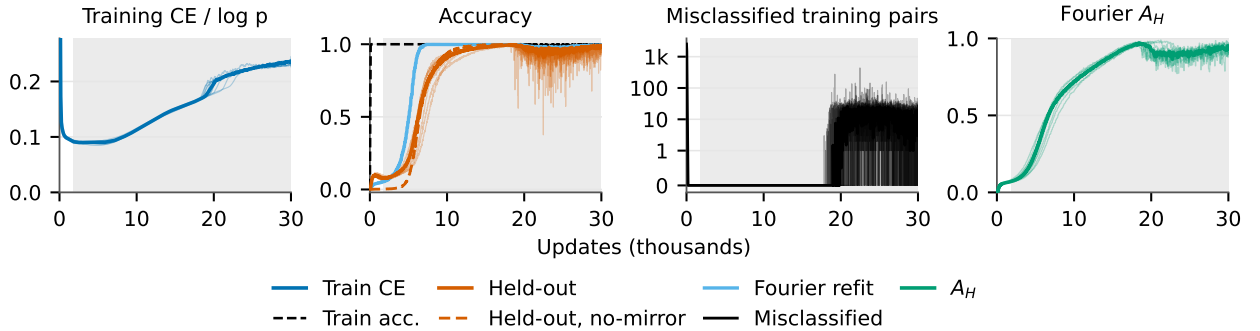


Figure 15: **Two blocks: every run (thin) and the median (thick)** (training fraction 0.2,  $\lambda_V = 0.275$ , gain  $2 \rightarrow 1$  with  $\tau = 10,000$ ; 10 runs, seeds  $i = 11-20$ : the two-block runs of Figure 5). Panels: the rows of Figure 14. No run passes: the CE rises during the held-out rise and no training pair is misclassified at any 20-update log inside the window until 17.8k–22.0k updates, when pairs begin to flip and held-out accuracy falls back. The three other two-block settings (5 runs each) and the runs of Table 6 are not shown.

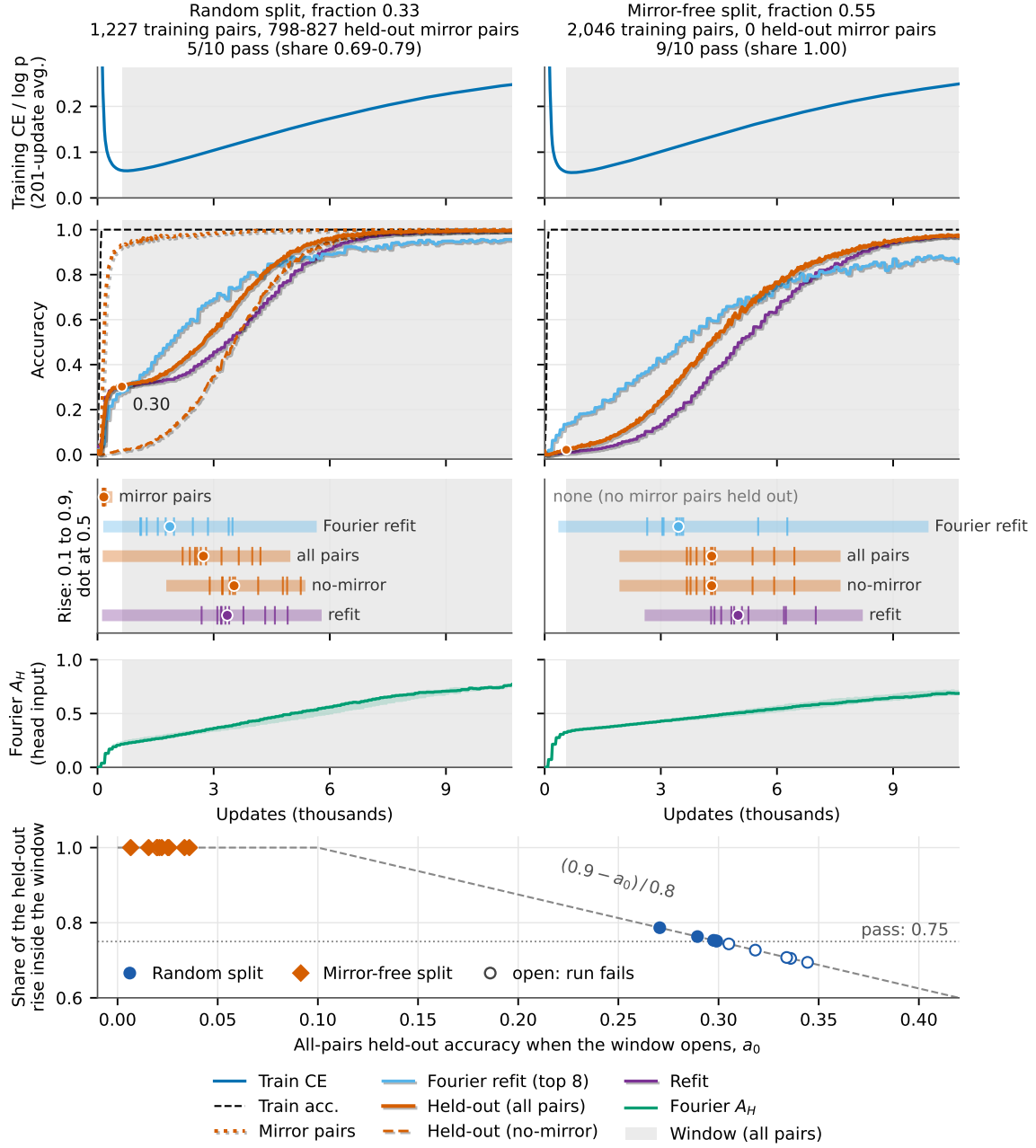


Figure 16: **Mirror pairs at  $p = 61$ :** the random split at  $\phi = 0.33$  (left; 798 to 827 of the 2,494 held-out pairs have their mirror in the training set) and a mirror-free split at  $\phi = 0.55$  (right; none), 10 runs each from the same initializations. Rows: training CE; held-out accuracy on all pairs (solid), the no-mirror pairs (dashed) and the mirror pairs (dotted), the refit and the Fourier-filtered refit on all held-out pairs, and training accuracy, medians (dot: median all-pairs accuracy when the window opens); where each accuracy rises (bar from the median 0.1 to the median 0.9 crossing, dot at the median 0.5 crossing, ticks at each run's 0.5 crossing);  $A_H$  of  $\hat{r}$ . Gray: the all-pairs window (median). Bottom: the all-pairs share of the held-out rise inside the window against  $a_0$ , the all-pairs accuracy when the window opens, one point per run. Every run lies on the bound  $(0.9 - \max(a_0, 0.1))/0.8$ ; open symbols fail (the mirror-free run that fails, on its final accuracy, is hidden among the others at  $a_0 = 0.02$ ).

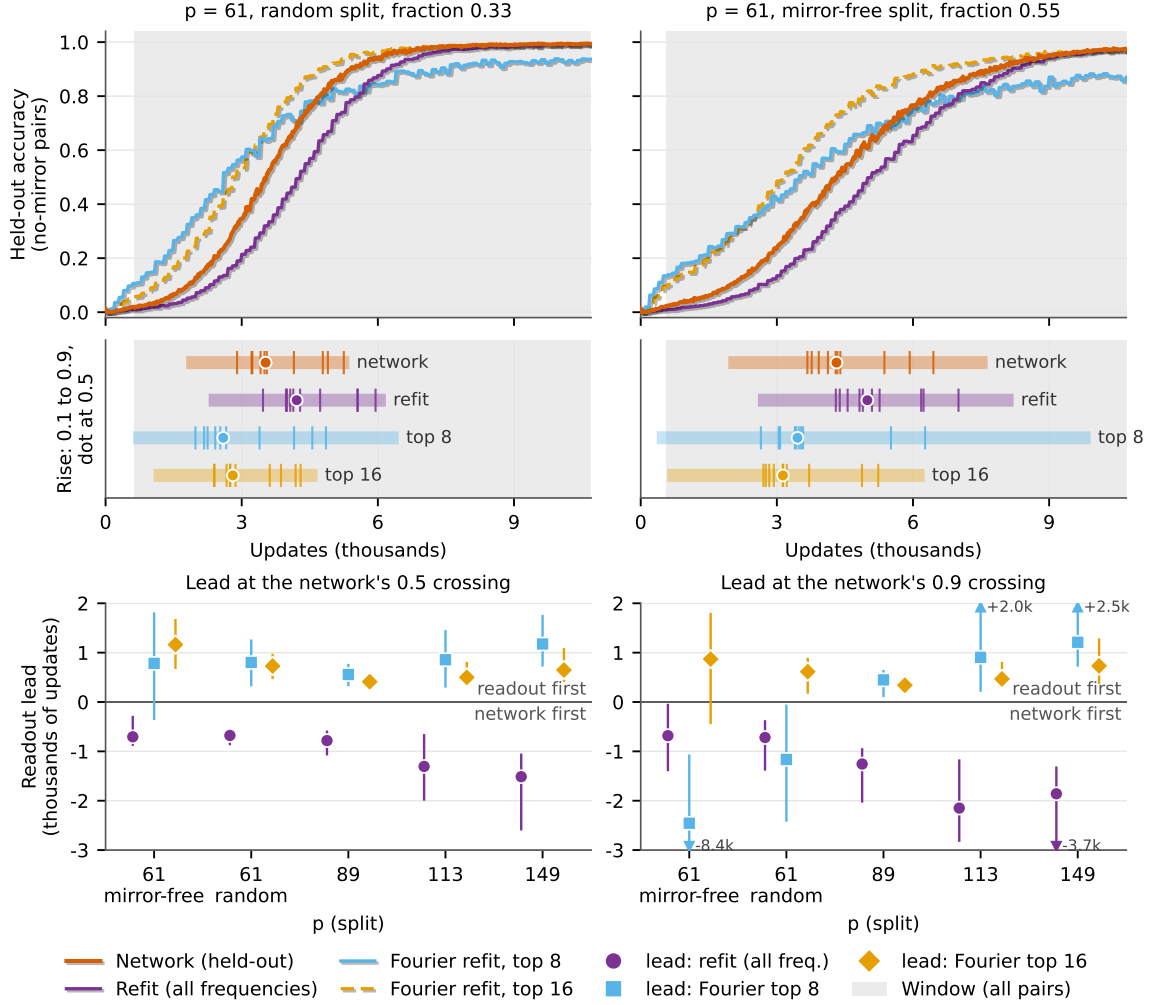


Figure 17: **Readouts against the network**, on the no-mirror pairs. Top:  $p = 61$  with the random split ( $\phi = 0.33$ ) and the mirror-free split ( $\phi = 0.55$ ): held-out accuracy of the network, of the refit (all frequencies) and of the Fourier-filtered refit with the top 8 and the top 16 frequencies, medians over 10 runs; below, where each rises (bars and dots as in Figure 16). Bottom: each readout's lead over the network at the network's 0.5 and 0.9 crossings (median over runs; bar from the smallest to the largest run) at the settings of Figure 14 and at  $p = 61$  with the random split; above zero the readout crosses first. The top-16 readout is post hoc.

## N Cross-appendix notation

Beyond Section 3, the following notation is shared across appendices; other notation is defined locally. We use  $j$  or  $t$  for update indices, with  $\mathbf{H}_j = \mathbf{H}(\mathbf{W}_j)$ ,  $L_j = L(\mathbf{W}_j, \mathbf{V}_j)$  and  $\mathbf{F}_j, \mathbf{G}_j$  the departing negative gradients of Section 3.2. For an indexed pair  $i = (a_i, b_i)$ ,  $x_i = x_{a_i b_i}$  and  $y_i = y_{a_i b_i}$ ;  $\mathbf{Y}[B]$  restricts rows and  $\mathbf{Y}(x, :)$  selects one row. For matrices indexed on the grid in both coordinates, subscripts such as  $\mathbf{Y}_{DS}$  restrict rows to  $D$  and columns to  $S$ . Matrix inner products are  $\langle \mathbf{U}, \mathbf{V} \rangle = \text{tr}(\mathbf{U}^* \mathbf{V})$ , where  $*$  is conjugate transpose ( $\top$  for real matrices);  $\mathbf{1}_B$  is the indicator of  $B$ .  $\|\cdot\|_{\text{op}}$ ,  $\odot$ ,  $\text{Var}(Z)$  and  $S^c$  denote the operator norm, entrywise product,  $\mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]$  and the complement in the stated ambient set. A positive semidefinite matrix’s supported inverse square root, written  $\mathbf{S}^{+1/2}$  or  $(\mathbf{S}^+)^{1/2}$ , acts on its positive-eigenvalue support and is zero on its kernel. For complete ridge comparison, let  $\mathbf{U}$  be a full feature table,  $S, D$  the training and query rows,  $\mathbf{T}$  the matrix of targets on  $S$  (not the full-grid target table), and  $\gamma > 0$ ; put  $q_U = \gamma \|\mathbf{U}_S\|_F^2$  and, when  $q_U > 0$ ,  $\mathbf{Y}_U = \mathbf{U}_D(\mathbf{U}_S^* \mathbf{U}_S + q_U \mathbb{I})^{-1} \mathbf{U}_S^* \mathbf{T}$ . For a same-sized comparison table  $\mathbf{V}$ , define  $q_V, \mathbf{Y}_V$  analogously and set  $e_S = \|\mathbf{V}_S - \mathbf{U}_S\|_F$ ,  $e_D = \|\mathbf{V}_D - \mathbf{U}_D\|_F$ ,  $t = \|\mathbf{T}\|_{\text{op}}$ ,  $N = \|\mathbf{T}\|_F$  and  $L = \|\mathbf{U}_D(\mathbf{U}_S^* \mathbf{U}_S + q_U \mathbb{I})^{-1/2}\|_{\text{op}}$ . All columns enter these formulas unless stated otherwise;  $t, N, L$  here are not time, dictionary size or training CE. The shared comparison is proved in Appendix E, Lemma E.38, and instantiated in Appendix F, Lemma F.34.

The two assembly results are the entry points for the decay proofs in Appendix G. In the half-grid construction,  $\mathcal{B}_M$  is the full-dimensional initial ball at width  $M$  (Definition E.53); Proposition E.54 selects this ball, the positive head-step cap  $\eta_0$  with  $\eta_V = \eta_0/2$ , and the decay cap  $d_0 \in (0, 1/2]$ . Appendices E and G, Corollary G.1, use the same ball and head step. The positive-decay corollary uses part (a) of that proposition, which preserves Theorem 4.1(i)–(iii); the late-refit part (b) assumes zero decay.

For the random split,  $h = 2\alpha(1 - \beta) > 0$  is the fixed dictionary height,  $\mathcal{E}_{\text{split}}$  the initial-data event (Definition F.10), and  $\tilde{\mathbf{V}}_0(S)$ ,  $r_{\text{new}}(S)$  the train-only center head and prescribed joint-ball radius ((135) and (141)). Appendix F constructs these objects; Appendix G, Definition G.2 and Corollary G.3, use them in the pre-sampling decay budget. Here Proposition F.36 transfers all trajectory conclusions whenever each head increment has norm at most  $\eta$ . The decay proof checks this hypothesis with data step  $\eta/2$ , keeping  $\eta$  in the prescribed radius and error budgets. Identical letters in the two constructions need not denote the same parameter. Terminating decimals are exact rationals; existence, positivity and bounds are proved where claimed, not inferred from defining formulas.

## E Structured half-grid theorem: construction and estimates

### E.1 Statement and conventions

The half-grid data and common diagnostics are defined in the main paper; Section N lists the interface reused in Appendix G; the construction and proof quantities are defined as they enter below. The default range is prime  $p \geq 97$ ; Theorem 4.1 uses  $p \geq 101$ , and the late-refit module uses  $p \geq 1764$ . The split contains whole difference classes; all diagonal pairs are held out. It is not an iid split or a guarantee for arbitrary balanced half-grids. Here  $\mathbf{P}_c = \mathbf{C}_p$  is class centering and  $T_y = e_y - \mathbf{1}_p/p$  is its centered class- $y$  target.

**Public statement and appendix supplement.** This appendix proves Theorem 4.1. It also proves the following additional bound, which is not part of that public statement: for  $p \geq 1764$ , the actual late complete train-only refit has  $E_{\text{ref}}(j) < 121/400$  at every  $j \in \mathcal{I}_{\text{late}}$ . The common theorem range is  $p \geq 101$ , although some construction estimates below hold for  $p \geq 97$ ; the public late-refit margin and accuracy guarantee is on nonsquare differences, not diagonal inputs. The scales are fixed in Definition E.3 and the clock and windows in (60)–(61); the positive head step and full-dimensional initialization-ball radius are existential, supplied by Proposition E.54.

### E.2 Claim I: construction and uniformly controlled reference dynamics

The construction has two stages. We first bound the uniform source and its Polar rows, using the mask counts and Fourier covariance blocks. Those bounds then identify the static initializer and verify the exact frozen-head trajectory in Lemma E.30.

### E.2.1 The static construction

#### E.2.1.1 Design, visible space and conventions

For the split (10), let  $Q$  be the nonzero squares in  $\mathbb{F}_p$  and  $d = (p-1)/2$  their number. The visible input space has dimension  $r = 2p-1$ , and the arithmetic dictionary has size  $N = p^2(p-1)$  and uniform mass  $\varpi = 1/N$ .

**Definition E.1** (Visible input space). Set  $E = (\mathbf{1}_{\text{left}}, -\mathbf{1}_{\text{right}})^\perp$  and let  $\mathbf{P}_E$  be its orthogonal projector.

The design identities needed below are:

**Lemma E.2** (Training design). *Every  $x \in \Omega$  lies in  $E$ . Every  $a \in \mathbb{F}_p$  is the left operand of exactly  $d$  training inputs and the right operand of exactly  $d$ , and every label occurs exactly  $d$  times in  $\mathcal{S}$ .*

*Proof.* Each  $x_{ab}$  has one unit entry in each half. The training inputs with left operand  $a$  are  $(a, a+q)$ ,  $q \in Q$ , and similarly on the right. For a difference  $q \in Q$  and a label  $y$  the unique pair is  $((y-q)/2, (y+q)/2)$ , so each label occurs once per  $q$ .  $\square$

#### E.2.1.2 The arithmetic core

**Fixed construction scales.** Write  $s = \eta_W$  and  $\lambda = \lambda_W$  for the parameters of (3). The half-grid learner has  $\lambda_V = 0$  except in Appendix G; its positive head step is selected in Proposition E.54.

**Definition E.3** (Construction scales). Fix a prime  $p \geq 101$ ,  $\lambda > 0$ , and then choose  $\epsilon > 0$  sufficiently small. The construction scales are

$$R = (\lambda\sqrt{r})^{-1}, \quad v = \epsilon/R, \quad L_{\text{lift}} = R\epsilon^{-3}. \quad (28)$$

**Definition E.4** (Arithmetic masks and reflection pairs). Let  $\mathbf{l} = \{0, 1, \dots, d-1\} \subset \mathbb{F}_p$ . The core dictionary is  $\Lambda = \{(\ell, s, k) : \ell \in \mathbb{F}_p^\times, s, k \in \mathbb{F}_p\}$ , of size  $N$ . For  $l = (\ell, s, k)$  define

$$A_{\ell,s} = \{a : \ell a + s \in \mathbf{l}\}, \quad B_{\ell,s,k} = k - A_{\ell,s}^c, \quad D_l(a, b) = \mathbf{1}_{A_{\ell,s}}(a)\mathbf{1}_{B_{\ell,s,k}}(b), \quad k_l = k. \quad (29)$$

Write  $A_l, B_l$ ; the mask shift  $s$  is unrelated to the step size  $s$ . Every core row has mass  $\pi_l = 1/N$ . Let  $\Lambda_c$  consist of the pairs  $\{(\ell, s, k), (-\ell, d-1-s, k)\}$  in  $\Lambda$ . Each pair  $B \in \Lambda_c$  has the same mask and class; put  $P_B = 2/N$  and denote these shared objects by  $D_B, k_B$  (Lemma E.6). Let  $\mathbf{C}_C \in \mathbb{R}^{\Lambda \times \Lambda_c}$  have entry  $1/\sqrt{2}$  on the two rows of each pair and zero elsewhere. Thus  $\mathbf{C}_C^\top \mathbf{C}_C = \mathbb{I}$ .

**Definition E.5** (Inactive indicators and core heads). For  $l = (\ell, s, k) \in \Lambda$ , let  $\mathbf{j}_l \in \{0, 1\}^{2p}$  indicate the inactive left and right operand coordinates, and define the analytic head matrix by

$$\bar{\mathbf{V}}_{C,l} = -v\sqrt{\pi_l} T_{k_l}. \quad (30)$$

Write  $\iota(\ell, s, k) = (-\ell, d-1-s, k)$  for the reflection defining the pairs in  $\Lambda_c$ . Thus the physical heads satisfy  $\bar{\mathbf{V}}_{C,c} = \bar{\mathbf{V}}_C \mathbf{C}_C$ , with  $\bar{\mathbf{V}}_{C,c,B} = -v\sqrt{P_B} T_{k_B}$ .

**Lemma E.6** (The rectangles). *For every  $l \in \Lambda$ :*

- (i)  $|A_l| = d$  and  $|B_l| = (p+1)/2$ ;  $\mathbf{j}_l$  has exactly  $p$  ones, and for  $x \in \Omega$ ,  $\mathbf{j}_l^\top x = 0$  if  $D_l(x) = 1$  while  $\mathbf{j}_l^\top x \geq 1$  otherwise.
- (ii) (Label exclusion)  $D_l(x) = 1$  implies  $y^*(x) \neq k_l$ .
- (iii) (Coverage) For  $x \in \Omega$  and  $k \neq y^*(x)$ , exactly  $d(p+1)/2$  rows  $l$  with  $k_l = k$  have  $D_l(x) = 1$ .
- (iv) (Class mass and head frame)  $\sum_{l:k_l=k} \pi_l = 1/p$  for every class  $k$ , and  $\bar{\mathbf{V}}_C \bar{\mathbf{V}}_C^\top = (v^2/p) \mathbf{P}_c$ .
- (v) (Reflection pairs)  $\iota$  is a fixed-point-free involution of  $\Lambda$  with  $A_{\iota(l)} = A_l$ ,  $B_{\iota(l)} = B_l$ ,  $k_{\iota(l)} = k_l$  and  $\mathbf{j}_{\iota(l)} = \mathbf{j}_l$ . Thus  $\Lambda$  is the disjoint union of  $N/2$  reflection pairs of rows with identical masks, classes, inactive indicators and masses.

*Proof.* (i)  $z \mapsto \ell z + s$  is a bijection of  $\mathbb{F}_p$ , so  $|A_l| = |\mathbb{I}| = d$  and  $|B_l| = |A_l^c| = p - d$ ;  $\mathbf{j}_l$  has  $(p - d) + d = p$  ones, and  $\mathbf{j}_l^\top x_{ab} = \mathbf{1}\{a \notin A_l\} + \mathbf{1}\{b \notin B_l\}$ .

(ii) If  $a \in A_l$  and  $a + b = k_l$ , then  $b \in k_l - A_l$ , which is disjoint from  $B_l$ .

(iii) Let  $x = (a, b)$ ; then  $a \neq k - b$ . By the definition (29) of the masks,  $D_l(x) = 1$  with  $k_l = k$  means  $\ell a + s \in \mathbb{I}$  and  $\ell(k - b) + s \notin \mathbb{I}$ . An affine bijection  $z \mapsto \ell z + s$  is determined by the images of the two distinct points  $a, k - b$ , and every pair of distinct images arises, so there are exactly  $|\mathbb{I}| |\mathbb{I}^c| = d(p + 1)/2$  such rows.

(iv) Each class has  $p(p - 1)$  rows of mass  $1/N$ . Hence

$$\bar{\mathbf{V}}_C \bar{\mathbf{V}}_C^\top = v^2 \sum_l \pi_l T_{k_l} T_{k_l}^\top = (v^2/p) \sum_k T_k T_k^\top = (v^2/p) \mathbf{P}_c.$$

(v)  $z \in \mathbb{I}$  if and only if  $d - 1 - z \in \mathbb{I}$ , since  $z \mapsto d - 1 - z$  maps  $\{0, \dots, d - 1\}$  onto itself and  $\{d, \dots, p - 1\}$  onto  $\{d - p, \dots, -1\} \equiv \{d, \dots, p - 1\}$ . As  $-\ell a + d - 1 - s = d - 1 - (\ell a + s)$ , the rows  $l$  and  $\iota(l)$  have the same  $A$ , hence the same  $B$  and  $\mathbf{j}$ . Clearly  $\iota \circ \iota$  is the identity, and  $\iota(l) = l$  would require  $\ell = -\ell$ , impossible for odd  $p$ .  $\square$

**Definition E.7** (Uniform source and mask tables). Let  $\mathbf{X}_S$  and  $\mathbf{X}_{\text{grid}}$  have rows  $x^\top$  on the indicated input sets. Write  $g_l = D_l \in \mathbb{R}^\Omega$  and let  $\mathbf{D}_{\text{grid}} \in \{0, 1\}^{\Omega \times \Lambda}$  have columns  $g_l$ ;  $\mathbf{D}_S$  is its training-row restriction. The uniform source and its fixed physical compression are

$$b_l = \frac{1}{np} \sum_{x \in S} D_l(x)x, \quad \mathbf{F}_0 \in \mathbb{R}^{\Lambda \times 2p}, \quad \mathbf{F}_{0,l} = \sqrt{\pi_l} b_l^\top, \quad \Sigma_0 = \mathbf{F}_0^\top \mathbf{F}_0, \quad \mathbf{F}_{0,c} = \mathbf{C}_C^\top \mathbf{F}_0. \quad (31)$$

By (31), put  $\mathbf{F}_{\text{unif}} := \mathbf{F}_0 = (\sqrt{\omega}/(np)) \mathbf{D}_S^\top \mathbf{X}_S$  and  $\mathbf{Q}_{\text{QR}} = \text{pol}(\mathbf{F}_{\text{unif}})$ . The positive-pair source has the same nonzero singular values as  $\mathbf{F}_0$ .

**Definition E.8** (Uniform anchor). With  $\mathbf{Q}_{\text{QR}}$  just defined, let  $\mathbf{H} = \mathbf{H}^{\text{QR}} = \mathbf{D}_{\text{grid}} \odot (\mathbf{X}_{\text{grid}} \mathbf{Q}_{\text{QR}}^\top)$ .

By (64), at uniform probabilities the physical core source is  $v\mathbf{F}_{0,c}$  and its analytic lift is  $v\mathbf{F}_0$ . The table  $\mathbf{H}^{\text{QR}}$  will supply the Fourier anchor in Section E.3.1.

### E.2.1.3 Counts, covariance blocks and inverse roots

For the interval calculations, write  $b = d$  and  $c = (p + 1)/2$  for the left- and right-operand mask sizes, so the rectangle size is  $u = bc = (p^2 - 1)/4$ . The base interval is  $\mathbb{I}$  and its reflected complement is  $\mathbb{I}' = -\mathbb{I}^c = \{1, \dots, c\}$ .

**Definition E.9** (Character operators and Fourier modes). Let  $\chi_2$  be the quadratic character, with  $\chi_2(0) = 0$ . Let  $(\mathbf{C}_-)_{a,z} = \chi_2(z - a)$  and  $\mathbf{C}_+ = \mathbf{C}_-^\top$ . Put  $e(y) = e^{2\pi i y/p}$  and  $\hat{f}(t) = \sum_a f(a)e(-ta)$ . Define the Hermitian multiplier  $\mathbf{H}_\chi$  to have eigenvalue  $\chi_2(t)$  on the Fourier vector  $(p^{-1/2}e(ta))_{a \in \mathbb{F}_p}$ , and put  $\tau = \sum_a \chi_2(a)e(-a)$ . Write  $e_t^L = ((p^{-1/2}e(ta))_a, 0)$  and  $e_t^R = (0, (p^{-1/2}e(ta))_a)$  in  $\mathbb{C}^{2p}$ , and  $e_c = (\mathbf{1}, \mathbf{1})/\sqrt{2p}$ . Put  $\mathbf{T} = \begin{bmatrix} 0 & \bar{\tau} \\ \tau & 0 \end{bmatrix}$ .

**Lemma E.10** (Character operators). *For every odd prime  $p$ ,  $|\tau|^2 = p$  and  $\bar{\tau} = \chi_2(-1)\tau$  (see also [32]). Moreover  $\mathbf{C}_- = \bar{\tau}\mathbf{H}_\chi$ ,  $\mathbf{C}_+ = \tau\mathbf{H}_\chi$ ,  $\mathbf{H}_\chi^2 = \mathbf{P}_c$ ,  $\mathbf{C}_\pm \mathbf{1} = 0$ ,  $\mathbf{C}_+ = \chi_2(-1)\mathbf{C}_-$  and  $\mathbf{C}_- \mathbf{C}_+ = \mathbf{C}_+ \mathbf{C}_- = p\mathbf{P}_c$ . In particular  $\chi_2(y) = (\tau/p) \sum_{m \neq 0} \chi_2(m)e(my)$  for all  $y$ .*

*Proof.* Write  $e_t(a) = p^{-1/2}e(ta)$  for the normalized Fourier vector. For  $t \neq 0$ , substitution  $y = z - a$  gives  $\mathbf{C}_- e_t = \chi_2(t)\bar{\tau}e_t$ ;  $\mathbf{C}_-$  annihilates constants. Also  $\bar{\tau} = \chi_2(-1)\tau$  by  $a \mapsto -a$ , and

$$|\tau|^2 = \sum_u \chi_2(u) \sum_{b \neq 0} e(b(1 - u)) = (p - 1) - \sum_{u \neq 1} \chi_2(u) = p.$$

Here  $\sum \chi_2 = 0$  by nonsquare dilation. Transpose and Fourier inversion give all remaining identities. In particular,  $\mathbf{T}$  is Hermitian,  $\mathbf{T}^2 = p\mathbb{I}$  and  $\|\mathbf{T}\| = \sqrt{p}$ .  $\square$

Thus  $\tau$  is real for  $p \equiv 1 \pmod{4}$  and imaginary for  $p \equiv 3 \pmod{4}$ ; neither its sign nor a reality assumption is used.

For a fixed mask, write  $\alpha = \mathbf{1}_{A_l}$  and  $\beta = \mathbf{1}_{B_l}$  for the left and right indicators, with intersection indicator  $L = \alpha \odot \beta$  and size  $h = |A_l \cap B_l|$ . The count calculation uses the signed character count  $\Xi = \alpha^\top \mathbf{C}_- \beta$  and the number  $n_l = E_l$  of active training inputs.

**Definition E.11** (Training counts and interval powers). The QR training-count vector  $\mathbf{n} = npb_l$  has operand blocks

$$\begin{aligned}\mathbf{n}^L(a) &= \alpha(a) \# \{z \in B_l : (a, z) \in \mathcal{S}\}, \\ \mathbf{n}^R(z) &= \beta(z) \# \{a \in A_l : (a, z) \in \mathcal{S}\}.\end{aligned}$$

Expectations  $\mathbb{E}$  average over  $l$  uniform on  $\Lambda$ . Put  $\mathbf{B} = \sqrt{\varpi} \mathbf{D}_{\text{grid}}$  and

$$\omega_v = |\hat{\mathbf{1}}_I(v)|^2, \quad \omega'_v = |\hat{\mathbf{1}}_{I'}(v)|^2, \quad \Sigma_2 = \sum_{v \neq 0} \omega_v^2.$$

**Lemma E.12** (Exact counts and mask moments). *Let  $p \geq 5$  be prime.*

(i) *The QR training counts are*

$$\begin{aligned}\mathbf{n}^L &= \frac{1}{2}(c\alpha - L + \alpha \odot \mathbf{C}_-\beta), \\ \mathbf{n}^R &= \frac{1}{2}(b\beta - L + \beta \odot \mathbf{C}_+\alpha).\end{aligned}$$

*The source row is  $\mathbf{F}_{\text{unif},l} = (\sqrt{\varpi}/(np))\mathbf{n}^\top$ .*

(ii) *Both count blocks sum to the number of active training inputs:  $\mathbf{1}^\top \mathbf{n}^L = \mathbf{1}^\top \mathbf{n}^R = n_l = (u - h + \Xi)/2$ , with  $|\Xi| \leq u/\sqrt{p}$ .*

(iii) *Write  $l \leftrightarrow (\mu, t_1, t_2)$  with  $\mu = \ell^{-1}$ ,  $t_1 = -s$  and  $t_2 = \ell k + s$ . This is a bijection  $\Lambda \rightarrow \mathbb{F}_p^\times \times \mathbb{F}_p \times \mathbb{F}_p$ , and  $A_l = \mu(I + t_1)$ ,  $B_l = \mu(I' + t_2)$  with  $I' = \{1, \dots, c\} = -I^c$ . For fixed  $\mu$ ,  $t_1, t_2$  are independent and uniform.*

(iv) *With  $\text{Var}(h)$  the variance under  $\mathbb{E}$ :  $\mathbb{E}[h] = u/p$ ,  $\mathbb{E}[h^2] = u/3$ ,  $\text{Var}(h) = u(p^2 + 3)/(12p^2)$ ,  $\mathbb{E}[\Xi] = \mathbb{E}[h\Xi] = 0$  and  $\mathbb{E}[\Xi^2] = p \text{Var}(h)$ . Hence  $\mathbb{E}[n_l] = u(p-1)/(2p)$  and  $\text{Var}(n_l) = (p+1) \text{Var}(h)/4$ .*

*Proof.* Parts (i)–(iii) give the count formulas and a convenient parameterization of the masks. For part (iv), we use that parameterization to diagonalize the mask kernel and read off the moments of  $h$  and  $\Xi$ .

(i) Expand  $1_Q(t) = [1 - 1_{t=0} + \chi_2(t)]/2$  in each QR count; the source factor is (31).

(ii) Sum the two count formulas, using  $\mathbf{1}^\top(\alpha \odot \mathbf{C}_-\beta) = \Xi = \mathbf{1}^\top(\beta \odot \mathbf{C}_+\alpha)$ . By Lemma E.10,  $\Xi = (\mathbf{P}_c\alpha)^\top \mathbf{C}_-(\mathbf{P}_c\beta)$  with  $\|\mathbf{C}_-\|_{\text{op}} = \sqrt{p}$  and  $\|\mathbf{P}_c\alpha\|^2 = \|\mathbf{P}_c\beta\|^2 = u/p$ .

(iii) This is direct:  $A_l = \ell^{-1}(I - s)$  and  $B_l = k - \ell^{-1}(I^c - s) = \ell^{-1}(I' + \ell k + s)$ .

(iv) Put  $\mathbf{B} = \sqrt{\varpi} \mathbf{D}_{\text{grid}}$  and  $\mathbf{K} = \mathbf{B}\mathbf{B}^\top$ . In the orthonormal grid basis  $\phi_{r,s}(a, z) = p^{-1}e(ra + sz)$ , its eigenvalues are

$$\lambda(r, s) = \frac{1}{p^2(p-1)} \sum_{\mu \neq 0} \omega_{\mu r} \omega'_{\mu s}, \quad \omega_v = |\hat{\mathbf{1}}_I(v)|^2, \quad \omega'_v = |\hat{\mathbf{1}}_{I'}(v)|^2. \quad (32)$$

Indeed, the coefficient of column  $l$  is  $\sqrt{\varpi} p^{-1} \hat{\mathbf{1}}_{A_l}(r) \hat{\mathbf{1}}_{B_l}(s)$ . Conditional on  $\mu$ , the independent translations of  $A_l, B_l$  from (iii) annihilate all unequal frequency pairs. The remaining diagonal average is (32). Here  $\omega_0 = b^2$ ,  $\omega'_0 = c^2$  and  $\omega'_v = \omega_v$  for  $v \neq 0$ . Here  $|y|$  is the absolute value of the representative of  $y \in \mathbb{F}_p$  in  $\{-(p-1)/2, \dots, (p-1)/2\}$ . The cyclic autocorrelation  $|I \cap (I + y)| = b - |y|$  gives, by Parseval,

$$\Sigma_2 = p \sum_y (b - |y|)^2 - b^4 = (p^2 - 1)(p^2 + 3)/48. \quad (33)$$

For the two grid functions  $d_0(a, z) = \mathbf{1}[a = z]$  and  $c_\chi(a, z) = \chi_2(z - a)$ , the Fourier coefficients at  $(r, s)$  are  $\mathbf{1}[r + s = 0]$  and  $\mathbf{1}[r + s = 0]\tau\chi_2(s)$ . Since  $h = \langle g_l, d_0 \rangle$  and  $\Xi = \langle g_l, c_\chi \rangle$ , their second moments are the corresponding quadratic forms of  $\mathbf{K}$ . Using  $\lambda(0, 0) = u^2/p^2$  and  $\lambda(-s, s) = \Sigma_2/[p^2(p-1)]$  for  $s \neq 0$  gives

$$\mathbb{E}[h^2] = \frac{u^2 + \Sigma_2}{p^2} = \frac{u}{3}, \quad \mathbb{E}[\Xi^2] = \frac{\Sigma_2}{p}, \quad \mathbb{E}[h\Xi] = 0. \quad (34)$$

The first moments are  $\mathbb{E}[h] = u/p$ ,  $\mathbb{E}[\Xi] = 0$ , since  $\mathbb{E}[g_l] = u/p^2$ . Combining these first moments with (34) gives  $\text{Var}(h) = \Sigma_2/p^2$ , with  $\Sigma_2$  evaluated in (33). Substituting into  $n_l = (u - h + \Xi)/2$  gives the stated mean and variance of  $n_l$ .  $\square$

### Line-sum evaluations.

**Definition E.13** (Line-sum quantities). For  $z \neq 0, 1$ , define  $W(z) = \sum_{t \neq 0} \omega_{tz} \omega_{t(1-z)}$  and  $J_\chi = \sum_{v \neq 0} \chi_2(v) \omega_v$ . In the following line sums, every  $z$  ranges over  $\mathbb{F}_p \setminus \{0, 1\}$ :

$$J = \frac{\sum_z W(z)}{p^2(p-1)}, \quad R = \frac{W(1/2)}{p^2(p-1)}, \quad U = \frac{\sum_z \chi_2(z) W(z)}{p^2(p-1)}, \quad V = \frac{\chi_2(-1) \sum_z \chi_2(z(1-z)) W(z)}{p^2(p-1)}.$$

Put  $\Sigma_1 = \sum_{v \neq 0} \omega_v$ ,  $g_0 = 4p/\sqrt{(p+1)(p^2+1)}$  and  $\vartheta = 2p/(p^2+1)$ .

Write  $j = g_0^2 J$ ,  $r_0 = g_0^2 R$ ,  $u = g_0^2 U$  and  $w = g_0^2 V$  for the normalized total, central, one-character and two-character line sums, respectively. Directly,

$$\begin{aligned} J &= \frac{(p+1)(p^2-3)}{24p^2}, \quad R = \frac{W(1/2)}{p^2(p-1)} = \frac{\Sigma_2}{p^2(p-1)} = \frac{(p+1)(p^2+3)}{48p^2}, \\ V &= \frac{\chi_2(-1)J_\chi^2 - \Sigma_2}{p^2(p-1)}. \end{aligned} \tag{35}$$

The evaluations follow from  $\sum_z W(z) = \Sigma_1^2 - \Sigma_2$  and  $\sum_z \chi_2(z(1-z)) W(z) = J_\chi^2 - \chi_2(-1)\Sigma_2$ . The first is obtained by writing the double sum  $\sum_{t \neq 0} \sum_{m \neq 0, t} \omega_m \omega_{t-m} = \sum_{m \neq 0} \omega_m (\Sigma_1 - \omega_m)$  with  $m = tz$ . The second is the same double sum with the weights  $\chi_2(m)\chi_2(t-m)$ ; the full convolution square of  $\chi_2\omega$  is  $J_\chi^2$ , and its  $t=0$  term is  $\chi_2(-1)\Sigma_2$ .

Consequently  $j = 2(p^2-3)/(3(p^2+1))$  and  $r_0 = (p^2+3)/(3(p^2+1))$ .

**Definition E.14** (Covariance blocks). Set

$$\begin{aligned} e_0 &= \frac{(p+1)(5p^2-3)}{3(p-1)^2(p^2+1)}, \quad x = \vartheta \left(1 - \frac{p+1}{(p-1)^2}\right), \quad e_1 = -\frac{2}{p-1} + \frac{j+pw}{(p-1)^2}, \\ f &= \frac{2}{p-1} - \frac{2u}{(p-1)^2}, \quad k_0 = \frac{2(1+u)}{(p-1)^2}, \quad k_1 = \frac{2\vartheta}{(p-1)^2}. \end{aligned} \tag{36}$$

$$\mathbf{E} = \begin{bmatrix} e_0 + x & e_1 \\ e_1 & e_0 - x \end{bmatrix}, \quad \mathbf{D} = \begin{cases} f \mathbf{T}, & \chi_2(-1) = -1, \\ \tau \begin{bmatrix} -k_0 + k_1 & f \\ f & -k_0 - k_1 \end{bmatrix}, & \chi_2(-1) = 1. \end{cases} \tag{37}$$

Define  $\mathbf{M}_\pm = g_0^{-2}(\mathbb{I} + \mathbf{E} \pm \mathbf{D})$  and  $c_0 = \mathbb{E}[n_t^2]/b^2$ . For the covariance computation,  $\mathcal{P}_Q$  denotes multiplication by  $\mathbf{1}_Q(z-a)$ .

**Lemma E.15** (Exact source covariance). *Let  $p \geq 5$  be prime. For  $t \neq 0$  with  $\varsigma = \chi_2(t)$ , the matrix of  $p^5 \mathbf{F}_{\text{unif}}^\top \mathbf{F}_{\text{unif}}$  on the ordered pair  $(e_t^L, e_t^R)$  is*

$$\mathbf{M}_\varsigma = g_0^{-2}(\mathbb{I} + \mathbf{E} + \varsigma \mathbf{D}). \tag{38}$$

On constants,  $p^5 \mathbf{F}_{\text{unif}}^\top \mathbf{F}_{\text{unif}} e_c = 2c_0 e_c$  and  $\mathbf{F}_{\text{unif}}(\mathbf{1}, -\mathbf{1}) = 0$ , where

$$c_0 = \frac{\mathbb{E}[n_t^2]}{b^2} = \frac{u^2}{p^2} + \frac{(p+1)^2(p^2+3)}{48p^2(p-1)} \geq \frac{u^2}{p^2}. \tag{39}$$

Equivalently,  $\mathbf{F}_{\text{unif}}^\top \mathbf{F}_{\text{unif}} = p^{-5}[\mathbf{M}_+ \otimes \mathbf{\Pi}_+ + \mathbf{M}_- \otimes \mathbf{\Pi}_- + 2c_0 e_c e_c^\top]$  with  $\mathbf{\Pi}_\pm = (\mathbf{P}_c \pm \mathbf{H}_\chi)/2$ . Here  $\mathbf{M} \otimes \mathbf{\Pi}_\pm$  acts by  $\mathbf{M}$  on the operand index and by  $\mathbf{\Pi}_\pm$  on the input coordinate.

*Proof.* We compute the covariance one Fourier plane at a time. The mask-kernel eigenvalues in (32) supply the weights; the two exceptional terms  $s=0, t$  and the remaining line sums then give the block in (38).

Write  $q = \mathbf{1}_Q$  and  $\mathbf{F}_{\text{unif}} = (np)^{-1} \mathbf{B}^\top \mathcal{P}_Q \mathbf{X}_{\text{grid}}$ . The indicator has  $\hat{q}(0) = b$  and  $\hat{q}(s) = (-1 + \tau \chi_2(s))/2$  for  $s \neq 0$ . The two masked operand modes have grid Fourier coefficients  $p^{-1/2} \mathbf{1}[r+s=t] \hat{q}(s)$  and  $p^{-1/2} \mathbf{1}[r+s=t] \hat{q}(s-t)$ . By (32), different  $t$  do not mix, and their Gram is

$$\mathbf{M}_t = \frac{1}{b^2} \sum_s \lambda(t-s, s) \begin{bmatrix} |\hat{q}(s)|^2 & \overline{\hat{q}(s)} \hat{q}(s-t) \\ \hat{q}(s-t) \hat{q}(s) & |\hat{q}(s-t)|^2 \end{bmatrix}. \tag{40}$$

Here  $p^5/(n^2p^3) = b^{-2}$  accounts for all normalizations.

Put  $L = bc^3/[p^2(p-1)]$  and  $R' = b^3c/[p^2(p-1)]$ . For  $t \neq 0$ , the terms  $s = 0, t$  use  $\lambda(t, 0) = L$ ,  $\lambda(0, t) = R'$ , with  $L + R' = 2g_0^{-2}$  and  $L - R' = 2g_0^{-2}\vartheta$ . The remaining terms, indexed by  $s = tz$ , have weight  $W(z)/[p^2(p-1)]$ . Their character sums are precisely  $J, U, V$  of Definition E.13. For example, using  $W(z) = W(1-z)$  and  $\tau\chi_2(-1) = \bar{\tau}$  gives

$$(\mathbf{M}_t)_{LR} = \frac{L + R'}{2b}(\chi_2(t)\bar{\tau} - 1) + \frac{J + pV - 2\chi_2(t)\bar{\tau}U}{4b^2}.$$

For the diagonal entries use  $|\hat{q}(s)|^2 = (p+1 - 2\operatorname{Re} \tau \chi_2(s))/4$  for  $s \neq 0$ . Substitution of (36)–(37) now gives  $\mathbf{M}_t = g_0^{-2}(\mathbb{I} + \mathbf{E} + \chi_2(t)\mathbf{D})$  in both Gauss phases. At  $t = 0$  the two columns agree, each Gram entry is  $\mathbb{E}[n_l^2]/b^2 = c_0$  by Lemma E.12(iv), and the constant block is  $c_0\mathbf{1}\mathbf{1}^\top$ . Its eigenvalues are  $2c_0$  and zero.  $\square$

**Lemma E.16** (Character-sum envelope). *For every prime  $p \geq 5$ ,  $|u| \leq j < 2/3$  and  $-r_0 \leq w \leq j$ . If  $\chi_2(-1) = -1$ , then  $J_\chi = 0$  and  $V = -R$ , so  $w = -r_0$ .*

*Proof.* The nonnegative weights  $W(z)$  give  $|U|, |V| \leq J$ . In the imaginary phase, pairing  $x, -x$  in  $J_\chi$  gives zero since  $\omega_{-x} = \omega_x$ ; now use (35). In the real phase the same identity gives  $V = (J_\chi^2 - \Sigma_2)/[p^2(p-1)] \geq -R$ .  $\square$

**Definition E.17** (Supported inverse-root quantities). The inverse roots  $\mathbf{M}_\pm^{-1/2}$  exist for  $p \geq 97$  by Lemma E.18(i). Define

$$\mathbf{G}_{\text{avg}} = \frac{1}{2}(\mathbf{M}_+^{-1/2} + \mathbf{M}_-^{-1/2}), \quad \mathbf{G}_{\text{diff}} = \frac{1}{2}(\mathbf{M}_+^{-1/2} - \mathbf{M}_-^{-1/2}), \quad \mathbf{G}_{\text{const}} = \frac{1}{2\sqrt{2c_0}} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Put  $\mathbf{A}_e = \mathbf{G}_{\text{avg}}/g_0$ ,  $\mathbf{A}_o = \mathbf{G}_{\text{diff}}/g_0$ , and  $f_p = \sqrt{p}g_0/b$ .

For  $p \geq 97$  both blocks  $\mathbf{M}_\pm$  are positive definite (Lemma E.18(i) below). The inverse root of  $\mathbf{F}_{\text{unif}}^\top \mathbf{F}_{\text{unif}}$  on its support  $E$ , with  $\mathbf{P}_0 = \mathbf{1}\mathbf{1}^\top/p$ , is therefore

$$(\mathbf{F}_{\text{unif}}^\top \mathbf{F}_{\text{unif}})^{+1/2} = p^{5/2}[\mathbf{G}_{\text{avg}} \otimes \mathbf{P}_c + \mathbf{G}_{\text{diff}} \otimes \mathbf{H}_\chi + \mathbf{G}_{\text{const}} \otimes \mathbf{P}_0]. \quad (41)$$

The quantities  $\mathbf{G}_{\text{avg}}, \mathbf{G}_{\text{diff}}, \mathbf{G}_{\text{const}}, \mathbf{A}_e, \mathbf{A}_o, f_p$  are defined in Definition E.17; only the structural constant-minus direction is null. Multiplying the QR source row in Lemma E.12(i) by the supported inverse root (41) gives

$$\begin{aligned} \mathbf{Q}_{\text{QR},l}^\top &= \sqrt{\varpi} \frac{\sqrt{p}}{b} [\mathbf{G}_{\text{avg}} \otimes \mathbf{P}_c + \mathbf{G}_{\text{diff}} \otimes \mathbf{H}_\chi \\ &\quad + \mathbf{G}_{\text{const}} \otimes \mathbf{P}_0] \mathbf{n}. \end{aligned} \quad (42)$$

These operators are real. If  $\chi_2(-1) = -1$  then  $\mathbf{M}_- = \overline{\mathbf{M}_+}$ , so  $\mathbf{G}_{\text{avg}}$  is real symmetric and  $\mathbf{G}_{\text{diff}}$  is imaginary Hermitian. Also  $\mathbf{T} \otimes \mathbf{H}_\chi$  acts as  $(w, v) \mapsto (\mathbf{C}_-v, \mathbf{C}_+w)$ .

**Lemma E.18** (Inverse-root bounds). *Let  $p \geq 97$  be prime.*

- (i)  $\frac{11}{15}g_0^{-2}\mathbb{I} \prec \mathbf{M}_\pm \prec \frac{19}{15}g_0^{-2}\mathbb{I}$ .
- (ii)  $\|\mathbf{A}_e - \mathbb{I}\| < 3/p$ . The matrix  $\mathbf{A}_e = [a_{ij}]$  is real symmetric, with  $1 - 3/p < a_{11}, a_{22} < 1 + 3/p$  and  $|a_{12}| = |a_{21}| < 3/p$ .
- (iii)  $\|\mathbf{A}_o + \mathbf{T}/p\| < 15/(2p^{3/2})$ ; hence  $\|\mathbf{A}_o\| < (1 + 15/(2p))/\sqrt{p}$ .
- (iv)  $f_p < 8/(p-1)$  and  $c_0 \geq u^2/p^2$ .

*Proof.* First we bound the two covariance perturbations to prove (i) and justify the inverse-root series. We then separate terms containing even or odd numbers of  $\mathbf{D}$  factors: the even part gives (ii), while a cancellation in the odd part gives the sharper approximation in (iii). Part (iv) follows from the explicit scalar formulas.

*Scalar inputs and positivity, part (i).* Write  $a = e_0$ ,  $h = \operatorname{tr} \mathbf{D}/2$ ,  $\mathbf{E} = \mathbf{E} - a\mathbb{I}$ ,  $\mathbf{D} = \mathbf{D} - h\mathbb{I}$ ,  $e^2 = \operatorname{tr}(\mathbf{E}^2)/2$ ,  $d^2 = \operatorname{tr}(\mathbf{D}^2)/2$  with  $e, d \geq 0$ , and  $c = \operatorname{tr}(\mathbf{E}\mathbf{D})/2$ . Set  $s = \|\mathbf{E}\|$ ,  $t = \|\mathbf{D}\|$ ,  $\ell = 39/8$  and  $m = 21/10$ .

These letters are local to this proof. The traceless Hermitian matrices satisfy  $\mathbf{E}^2 = e^2 \mathbb{I}$ ,  $\mathbf{D}^2 = d^2 \mathbb{I}$  and  $\mathbf{ED} + \mathbf{DE} = 2c \mathbb{I}$ . Substituting the envelope of Lemma E.16 in (36) gives

$$\begin{aligned} \frac{5}{3p} < a < \frac{7}{4p}, \quad e < \frac{25}{8p}, \quad \frac{4}{p} < d^2 < \frac{1089}{256p}, \quad d < \frac{33}{16\sqrt{p}}, \\ |h| < \frac{7}{2p^{3/2}}, \quad |c| < \frac{24}{5p^{3/2}}. \end{aligned} \quad (43)$$

For explicit checks,  $p\mathbf{x} < 2$ ,  $2 < p\mathbf{f} < 2 + 7/(2p) \leq 41/20$ , while  $p^2\mathbf{k}_0 < 7/2$  and  $p^3\mathbf{k}_1 < 25/6$  in the real phase. The lower  $a$  bound reduces to  $15p^3 - 13p^2 + 7p - 5 > 0$ . The envelope also gives  $\mathbf{e}_1 < 0$  and

$$p|\mathbf{e}_1| \leq \frac{2p}{p-1} + \frac{1}{3} + \frac{1+8/p}{3(p-1)^2} < \frac{12}{5},$$

so  $p^2e^2 < 4 + (12/5)^2 < (25/8)^2$ . Also  $pd^2 \leq (41/20)^2 + 625/(36p^4) < (33/16)^2$ . In the imaginary phase  $h = c = 0$ ; otherwise  $h = -\tau\mathbf{k}_0$  and  $c = \tau(\mathbf{x}\mathbf{k}_1 + \mathbf{e}_1\mathbf{f})$ , whence

$$p^{3/2}|c| < \frac{25}{3p^2} + \left[ \frac{2p}{p-1} + \frac{1}{3} + \frac{1+8/p}{3(p-1)^2} \right] \left( 2 + \frac{7}{2p} \right) < \frac{24}{5}.$$

The last bounds hold at 97 and improve thereafter. Thus  $\|\mathbf{E}\| + \|\mathbf{D}\| \leq a + e + |h| + d < 39/(8p) + (33/16 + 7/(2p))/\sqrt{p} < 4/15$ , proving (i); the last inequality holds at 97 using  $\sqrt{97} > 49/5$  and improves thereafter.

*Separating the inverse-root series.* The bound  $s + t < 4/15$  just proved ensures convergence for both signs in  $(\mathbb{I} + \mathbf{E} \pm \mathbf{D})^{-1/2}$ . Put  $z_n = \binom{2n}{n}/4^n$  and  $\mathbf{C} = 3(\mathbf{ED} + \mathbf{DE})/8 - 5\mathbf{D}^3/16$ . Define the convergent parity remainders

$$\begin{aligned} \mathbf{R}_e &= \frac{1}{2} \sum_{n \geq 3} (-1)^n z_n [(\mathbf{E} + \mathbf{D})^n + (\mathbf{E} - \mathbf{D})^n], \\ \mathbf{R}_o &= \frac{1}{2} \sum_{n \geq 4} (-1)^n z_n [(\mathbf{E} + \mathbf{D})^n - (\mathbf{E} - \mathbf{D})^n]. \end{aligned}$$

The convergent binomial series for  $(\mathbb{I} + \mathbf{E} \pm \mathbf{D})^{-1/2}$  gives

$$\begin{aligned} \mathbf{A}_e - \mathbb{I} &= -\frac{\mathbf{E}}{2} + \frac{3}{8}(\mathbf{E}^2 + \mathbf{D}^2) + \mathbf{R}_e, \\ \mathbf{A}_o &= -\frac{\mathbf{D}}{2} + \mathbf{C} - \frac{5}{16}(\mathbf{E}^2\mathbf{D} + \mathbf{EDE} + \mathbf{DE}^2) + \mathbf{R}_o. \end{aligned} \quad (44)$$

The even and odd combinations retain only words containing respectively an even and an odd number of  $\mathbf{D}$  factors; no commutation is assumed. The coefficients  $z_n$  decrease, so geometric summation gives

$$\begin{aligned} \|\mathbf{R}_e\| &\leq \frac{5}{16} \frac{s^3 + 3st^2 - s^4 + t^4}{(1-s)^2 - t^2}, \\ \|\mathbf{R}_o\| &\leq \frac{35}{128} \frac{4s^3t + 4st^3 - 3s^4t + 2s^2t^3 + t^5}{(1-s)^2 - t^2}. \end{aligned}$$

Here  $s < \ell/p$ ,  $t < m/\sqrt{p}$ , and  $(1-s)^2 - t^2 > 1 - (2\ell + m^2)/p > 17/20$ . Dropping the negative numerator terms and evaluating at 97 yields

$$\begin{aligned} p\|\mathbf{R}_e\| &< \frac{25}{68} \left[ \frac{\ell^3}{97^2} + \frac{3\ell m^2 + m^4}{97} \right] < \frac{7}{20}, \\ p^{3/2}\|\mathbf{R}_o\| &< \frac{175}{544} \left[ \frac{4\ell^3 m + 2\ell^2 m^3}{97^2} + \frac{4\ell m^3 + m^5}{97} \right] < \frac{4}{5}. \end{aligned} \quad (45)$$

These bounds improve thereafter.

The even part, part (ii). We combine the first line of (44) with its remainder budget in (45). Since  $-a/2 + 3d^2/8 > 0$ , the quadratic even term satisfies

$$p \left\| -\frac{\mathbf{E}}{2} + \frac{3}{8}(\mathbf{E}^2 + \mathbf{D}^2) \right\| < \frac{4681}{6144} + \frac{25}{16} + \frac{3\ell^2/8 + 693/128}{p} + \frac{147}{32p^2} < \frac{53}{20}.$$

Adding  $p\|\mathbf{R}_e\| < 7/20$  from (45) proves  $\|\mathbf{A}_e - \mathbb{I}\| < 3/p$ . The real-phase roots are real symmetric and the imaginary-phase roots are conjugates, proving the reality and entry bounds in (ii).

The odd part, part (iii). In the second line of (44), the terms collected in  $\mathbf{C}$  must be estimated together. The traceless identities give the cancellation directly:

$$\begin{aligned} \mathbf{C} = & \left( \frac{3ah}{4} + \frac{3c}{4} - \frac{5h^3}{16} - \frac{15hd^2}{16} \right) \mathbb{I} \\ & + \left( \frac{3a}{4} - \frac{5d^2}{16} - \frac{15h^2}{16} \right) \mathbf{D} + \frac{3h}{4} \mathbf{E}. \end{aligned} \quad (46)$$

The scalar bounds (43) imply  $|3a/4 - 5d^2/16| < 15/(128p)$ . Applying them to the three coefficients in (46) consequently gives

$$p^{3/2}\|\mathbf{C}\| < \frac{18}{5} + \frac{495}{2048} + \frac{151977/8192 + 525/64}{p} + \frac{24255}{1024p^2} + \frac{1715}{128p^3} < \frac{33}{8}.$$

The mixed cubic terms contribute at most  $15s^2t/16 < 1/(2p^{3/2})$ . Finally the explicit form of  $\mathbf{D}$  gives  $p^{3/2}\|\mathbf{D}/2 + \mathbf{T}/p\| \leq p^2(2 + 1/p)/(p-1)^2 < 33/16$ : in the real phase use  $1 + u + \sqrt{u^2 + (u - (p-1)/p)^2} \leq 2 + 1/p$ ; the imaginary-phase expression is smaller. The second line of (44) now has four contributions: the leading-term error,  $\mathbf{C}$ , the mixed cubic terms, and the remainder bounded in (45). Adding them gives

$$p^{3/2}\|\mathbf{A}_o + \mathbf{T}/p\| < \frac{33}{16} + \frac{33}{8} + \frac{1}{2} + \frac{4}{5} = \frac{599}{80} < \frac{15}{2}. \quad (47)$$

Adding  $\|\mathbf{T}\|/p = p^{-1/2}$  to (47) proves the remaining bound in (iii).

(iv)  $g_0^{-2} > p/16$  gives  $g_0 < 4/\sqrt{p}$ , so  $f_p < 8/(p-1)$ . The bound on  $c_0$  is (39).  $\square$

For a source supported on  $E$ , write  $\sigma_{\min,E}(\mathbf{F}) = \min_{z \in E, \|z\|=1} \|\mathbf{F}z\|$ .

**Lemma E.19** (Uniform source). (i)  $\sigma_{\min,E}(\mathbf{F}_0) > 1/(5p^2)$ .

(ii) For every  $l \in \Lambda$ :  $\text{pol}(\mathbf{F}_0)_l x > \frac{13}{60}\sqrt{\pi_l}$  whenever  $D_l(x) = 1$ , and  $\text{pol}(\mathbf{F}_0)_l x < \frac{17}{2}\sqrt{\pi_l}$  for every  $x \in \Omega$ .

*Proof.* The two bounds use the notation and formulas of Appendix E.2.1.3, where  $\mathbf{F}_{\text{unif}} = \mathbf{F}_0$  and  $\mathbf{Q}_{\text{QR}} = \text{pol}(\mathbf{F}_0)$ . The singular-value bound in (i) follows from the covariance blocks. For the row bounds in (ii), we first control the training counts, then insert them into (42). The resulting row formula separates a signed count contribution from the inverse-root errors.

(i) On the nonconstant Fourier planes,  $p^5\mathbf{\Sigma}_0$  has blocks  $\mathbf{M}_{\pm}$ ; on the constant line in  $E$  it has eigenvalue  $2c_0$ . The root bounds give  $\mathbf{M}_{\pm} \succ (11/15)g_0^{-2}\mathbb{I} \succ (11p/240)\mathbb{I}$ , while  $2c_0 \geq 2u^2/p^2 > g_0^{-2} > p/16$ . Thus  $\sigma_{\min,E}(\mathbf{F}_0) > \sqrt{11/240}p^{-2} > 1/(5p^2)$ .

(ii) Fix a row and write  $\alpha = \mathbf{1}_A$ ,  $\beta = \mathbf{1}_B$ ,  $L = \alpha \odot \beta$ ,  $h = |A \cap B|$  and  $\mathbf{n} = (\mathbf{n}^L, \mathbf{n}^R) = npb_l$ . Set  $E_- = \frac{1}{2}(u - b - u/\sqrt{p})$  and  $E_+ = \frac{1}{2}(u + u/\sqrt{p})$ . Lemma E.12(i)–(ii), together with  $0 \leq h \leq b$ , gives

$$E_- \leq E_l \leq E_+.$$

We first give the two count budgets used below. Put  $M_p = \sqrt{p}[\log(p+1) + 3 - 2\log 2]/(2\sqrt{2})$ . For every affine image  $D$  of  $l$  or  $l' = -l^c$  and every  $y$ ,

$$\left| \sum_{w \in D} \chi_2(w - y) \right| \leq M_p < \frac{9p}{40}. \quad (48)$$

Indeed, Fourier inversion and  $|\tau| = \sqrt{p}$  bound this sum by  $p^{-1/2} \sum_{m \neq 0} |\widehat{\mathbf{1}}_1(m)|$ . Put  $\theta_m = \pi m/(2p)$  and  $K = \lceil b/2 \rceil$ . The summand is  $(2 \sin \theta_m)^{-1}$  for odd  $m$  and  $(2 \cos \theta_m)^{-1}$  for even  $m$ . Pair  $m$  with  $p - m$ , use  $\sin \theta \geq 2\sqrt{2}\theta/\pi$  and  $\cos \theta \geq 1/\sqrt{2}$  for  $\theta \leq \pi/4$ , and use  $K \leq (p+1)/4$ . Then the sum is at most

$$\frac{p}{\sqrt{2}} \sum_{j=1}^K \frac{1}{2j-1} + \sqrt{2} \lfloor b/2 \rfloor < \frac{p}{\sqrt{2}} (1 + \frac{1}{2} \log K) + \frac{p}{2\sqrt{2}} \leq \sqrt{p} \mathbf{M}_p.$$

The harmonic bound follows by integrating  $1/w$  on  $[2j-2, 2j]$  for  $j \geq 2$ . The ratio  $\mathbf{M}_p/p$  decreases for  $p \geq 97$ , and  $\mathbf{M}_{97} < 21.7 < 9 \cdot 97/40$ . Because  $A, B$  have the required affine forms (Lemma E.12), this bounds every entry of  $\mathbf{C}_-\beta$  and  $\mathbf{C}_+\alpha$ .

The character-sum cap (48) gives two count budgets: a coordinate bound for the even inverse-root error and a norm bound for the odd error. Set  $\nu_\infty = \max(\|\mathbf{P}_c \mathbf{n}^L\|_\infty, \|\mathbf{P}_c \mathbf{n}^R\|_\infty)$  and  $\nu_2^2 = \|\mathbf{P}_c \mathbf{n}^L\|^2 + \|\mathbf{P}_c \mathbf{n}^R\|^2$ . Each centered count lies between  $-E_+/p$  and  $(c + \mathbf{M}_p)/2 - E_-/p$ ; the latter is below  $19p/80 + \sqrt{p}/8 + 1/2 + 1/(8p) < p/3$ , so  $\nu_\infty < p/3$ . The two uncentered blocks have the same sum  $E_l$  and respective entry caps  $(c + \mathbf{M}_p)/2, (b + \mathbf{M}_p)/2$ . Therefore

$$\nu_2^2 \leq (p/2 + \mathbf{M}_p)E_l - 2E_l^2/p < (29p/40)E_l - 2E_l^2/p < p^3/16. \quad (49)$$

For the final inequality put  $t = p^{-1/2} < 1/9$ . Since  $E_l \leq E_+ < p^2(1+t)/8 < 5p^2/36 < 29p^2/160$ , the quadratic is increasing through this interval, and its upper bound divided by  $p^3$  is  $(19+9t-10t^2)/320 < 1/16$ .

*Decomposing a Polar row.* We now apply  $\nu_\infty < p/3$  and (49) to the inverse-root row. Put  $\gamma_l = 2E_l/(pg_0\sqrt{2c_0})$  and, for  $x = x_{az}$ , define

$$\begin{aligned} S_l(x) &= (\mathbf{P}_c \mathbf{n}^L)(a) + (\mathbf{P}_c \mathbf{n}^R)(z) - [(\mathbf{C}_- \mathbf{n}^R)(a) + (\mathbf{C}_+ \mathbf{n}^L)(z)]/p, \\ \Delta_e(x) &= x^\top ((\mathbf{A}_e - \mathbb{I}) \otimes \mathbf{P}_c) \mathbf{n}, \\ \Delta_o(x) &= x^\top ((\mathbf{A}_o + \mathbf{T}/p) \otimes \mathbf{H}_\chi) \mathbf{n}. \end{aligned}$$

The inverse-root row formula (42) now reads

$$\frac{\text{pol}(\mathbf{F}_0)_l x}{\sqrt{\pi_l}} = f_p [S_l(x) + \Delta_e(x) + \Delta_o(x) + \gamma_l]. \quad (50)$$

Lemma E.18(ii)–(iii) and  $\mathbf{H}_\chi^2 = \mathbf{P}_c$  give

$$\begin{aligned} |\Delta_e| &\leq (3/p)2\sqrt{2}\nu_\infty < 2\sqrt{2}, & |\Delta_o| &\leq \frac{15\sqrt{2}}{2p^{3/2}}\nu_2 < \frac{15\sqrt{2}}{8}, \\ & & |\Delta_e| + |\Delta_o| &< \frac{11}{2}. \end{aligned} \quad (51)$$

The formulas for  $g_0, c_0$  give

$$\begin{aligned} \frac{8}{p+2} < f_p < \frac{8}{p-1}, & (u/p)^2 \leq c_0 < (u/p)^2(1 + 1/(2p)), \\ \frac{\sqrt{p}}{8} < \gamma_l < \frac{\sqrt{p}}{4}. \end{aligned} \quad (52)$$

The two sides of the  $\gamma_l$  estimate follow from the count and covariance bounds as follows. Indeed  $2E_l \geq u(1 - 1/c - p^{-1/2}) > 3u/4$  and  $pg_0\sqrt{2c_0} < 6u/\sqrt{p}$  prove the lower bound on  $\gamma_l$ . For the upper bound, substitution of  $E_+$  gives  $\gamma_l/\sqrt{p} \leq (1 + p^{-1/2})\sqrt{1 + 2/p}/(4\sqrt{2}) < 1/4$  for  $p \geq 97$ .

*Lower bound on active inputs.* With the errors in (50) controlled, it remains to bound  $S_l(x)$ . On an active mask,  $a \in A, z \in B$ . Put  $\epsilon_x = \beta(a) + \alpha(z)$ . Substitution of the count formulas yields

$$\begin{aligned} S_l(x) &= (p - \epsilon_x)/2 - 2E_l/p + [c(\mathbf{C}_-\beta)(a) + b(\mathbf{C}_+\alpha)(z)]/(2p) \\ &\quad + [(\mathbf{C}_-L)(a) + (\mathbf{C}_+L)(z)]/(2p) \\ &\quad - [(\mathbf{C}_- \text{diag}(\beta)\mathbf{C}_+\alpha)(a) + (\mathbf{C}_+ \text{diag}(\alpha)\mathbf{C}_-\beta)(z)]/(2p). \end{aligned} \quad (53)$$

In (53), we bound the middle-line intersection term first, then apply a projection estimate to each quadratic term in the last line. The middle line is at least  $(\epsilon_x - 2h)/(2p)$ . For an orthogonal projection  $\mathbf{P}$ ,  $|v^\top \mathbf{P} w| \leq (|v^\top w| + \|v\| \|w\|)/2$ , since  $2\mathbf{P} - \mathbb{I}$  is orthogonal. Apply this with  $(v, w, \mathbf{P}) = (\mathbf{C}_+ e_a, \mathbf{C}_+ \alpha, \text{diag } \beta)$  and then with  $(\mathbf{C}_- e_z, \mathbf{C}_- \beta, \text{diag } \alpha)$ . The two quadratic terms are at most  $(c + \sqrt{(p-1)u})/2$  and  $(b + \sqrt{(p-1)u})/2$ . The  $h$  terms cancel, giving

$$\begin{aligned} S_l(x) &\geq p/2 - u/p - 1 + 1/p - u/p^{3/2} \\ &\quad - M_p/2 - 1/4 - \sqrt{(p-1)u}/(2p) \\ &> 11p/80 - \sqrt{p}/2 - 5/4. \end{aligned} \tag{54}$$

*Upper bound on all inputs.* Here the count budgets suffice, without assuming  $D_l(x) = 1$ . For arbitrary  $x$ , Cauchy-Schwarz on the character rows and  $\|\mathbf{P}_c \mathbf{n}^L\| + \|\mathbf{P}_c \mathbf{n}^R\| \leq \sqrt{2}\nu_2 < 3p^{3/2}/8$  give

$$S_l(x) \leq p/2 + M_p - 2E_-/p + \frac{\sqrt{p-1}}{p} \sqrt{2}\nu_2 < 17p/20 + \sqrt{p}/4 + 1/2. \tag{55}$$

*Assembling the row margins.* Insert (54) and (55) into (50), using the error budget (51) and the scales (52). The bracket in the row formula is consequently above  $11p/80 - 3\sqrt{p}/8 - 27/4 > 0$  on the mask and below  $17p/20 + \sqrt{p}/2 + 6$  everywhere. Thus

$$\begin{aligned} \frac{\text{pol}(\mathbf{F}_0)_l x}{\sqrt{\pi_l}} &> \frac{11p - 30\sqrt{p} - 540}{10(p+2)} > \frac{13}{60} \quad (D_l(x) = 1), \\ \frac{\text{pol}(\mathbf{F}_0)_l x}{\sqrt{\pi_l}} &< \frac{34p/5 + 4\sqrt{p} + 48}{p-1} < \frac{17}{2}. \end{aligned}$$

For the upper bound a nonpositive bracket is immediate. The final scalar inequalities are  $53p - 180\sqrt{p} - 3266 > 0$  and  $17p - 40\sqrt{p} - 565 > 0$ ; both hold at 97 using  $\sqrt{97} < 10$ , and increase thereafter.  $\square$

#### E.2.1.4 Helpers and training-invisible queries

**Definition E.20** (Fixed active bank and width padding). The active physical bank consists of  $N/2$  core pair rows  $B \in \Lambda_c$ ,  $p$  ordinary helpers  $u \in \mathbb{F}_p$ , and  $p$  training-invisible queries  $u \in \mathbb{F}_p$ , with the roles described in Section 4.1. At every admitted width  $M \geq M_{\min}$ , append  $M - (N/2 + 2p)$  dormant rows; this is at least  $n$  and equals  $n$  at  $M_{\min}$ . The canonical incoming row of each dormant unit is  $-\mathbf{1}_{2p}^\top$  and its head column is zero. Its grid preactivation is  $-2$ . No coordinate is tied in the full-dimensional initialization ball.

**Definition E.21** (Row helpers). Define

$$\mathbf{J}_{\text{row}} \in \{0, 1\}^{p^2 \times p}, \quad (\mathbf{J}_{\text{row}})_{(a,b),u} = \mathbf{1}[a = u]. \tag{56}$$

The incoming rows are  $\mathbf{W}_{N,u}^{\text{nom}} = \mathbf{W}_{N,u}^0 = L_{\text{lift}}(2e_u - \mathbf{1}; 0)$ , and their heads are

$$\mathbf{V}_{N,u}^0 = 0. \tag{57}$$

Their nominal feature table is  $\mathbf{H}_N^{\text{nom}} = L_{\text{lift}} \mathbf{J}_{\text{row}}$ . At the nominal helper rows the helper features are  $\text{ReLU}(\mathbf{W}_{N,u}^0 x_{ab}) = L_{\text{lift}} \mathbf{1}\{a = u\}$ , so exactly one helper is active at each input.

**Definition E.22** (Training-invisible query rows). For  $u \in \mathbb{F}_p$ , let  $\mathcal{D}_u = \{(u, w) \in \mathcal{D}\}$  and  $c(u) = 2u + 1$ . Put  $A_* = 21R/10$  and  $v = \epsilon$ , and set

$$\mathbf{W}_{Q,u}^{\text{wit}} = \mathbf{W}_{Q,u}^0 = A_*(2e_u - \mathbf{1}; -2\mathbf{1}_{u+Q}).$$

The query heads are

$$\mathbf{V}_{Q,u}^{\text{wit}} = \mathbf{V}_{Q,u}^0 = \frac{24\epsilon}{p(21R/10)} T_{c(u)} = \frac{24v}{pA_*} T_{c(u)}. \tag{58}$$

Let  $\mathbf{H}_Q^{\text{wit}}$  denote the query ReLU table.

**Lemma E.23** (Query witness). (i)  $c(u) \neq y^*(x)$  for every  $x \in \mathcal{D}_u$ , and  $c(u) = y_i$  for the training input  $i = (u, u + 1)$ .

(ii) The witness just defined has preactivations equal to  $A_*$  on  $\mathcal{D}_u$  and in  $\{-A_*, -3A_*\}$  elsewhere; in particular they are negative on  $\mathcal{S}$ . Its feature table  $\mathbf{H}_Q^{\text{wit}}$  has column  $u$  equal to  $A_* \mathbf{1}_{\mathcal{D}_u}$ , which is nonnegative and supported in one first-operand row, and  $\|\mathbf{H}_Q^{\text{wit}}\|_F^2 = p(p+1)A_*^2/2 < A_*^2 p^2$ . Its logits vanish on  $\mathcal{S}$  and equal  $(24\epsilon/p)T_{c(u)}$  on  $\mathcal{D}_u$ .

*Proof.* (i)  $c(u) = u + w$  means  $w - u = 1 \in Q$ , so  $(u, w) \in \mathcal{S}$ ; and  $(u, u + 1) \in \mathcal{S}$  has label  $2u + 1$ . (ii) The left part of the preactivation at  $(a, w)$  is  $\pm A_*$  according to  $a = u$  or  $a \neq u$ , and the right part is  $-2A_*$  if  $w - u \in Q$  and 0 otherwise. On row  $u$  this gives  $A_*$  at held-out and  $-A_*$  at training inputs, and on the other rows  $-A_*$  or  $-3A_*$ . The energy is  $p \cdot \frac{p+1}{2} A_*^2$ . On  $\mathcal{D}_u$  only row  $u$  is active.  $\square$

**Actual source identities.** Write  $P(f)_c = e^{f_c} / \sum_k e^{f_k}$  for softmax and  $r_x = e_{y^*(x)} - P(f(x))$  for the negative derivative of cross-entropy with respect to the example logits. The sources  $\mathbf{F} = -\nabla_{\mathbf{W}} L$  and  $\mathbf{G} = -\nabla_{\mathbf{V}} L$  have, for a hidden row with incoming row  $w$  and head column  $c$ , the row and column

$$\mathbf{F}_{\text{row}} = \frac{1}{n} \sum_{x \in \mathcal{S}} \mathbf{1}\{wx > 0\} (c^\top r_x) x^\top, \quad \mathbf{G}_{\text{col}} = \frac{1}{n} \sum_{x \in \mathcal{S}} r_x \text{ReLU}(wx), \quad (59)$$

since  $\text{ReLU}'(0) = 0$ . Rows of  $\mathbf{F}$  lie in  $E$  (Lemma E.2), so  $\mathbf{F}\mathbf{P}_E = \mathbf{F}$  and  $\text{rank } \mathbf{F} \leq r$  at every state.

### E.2.1.5 Physical realization and positive compression

**Definition E.24** (Fixed positive pair isometry). Put  $M_{\text{act}} = N/2 + 2p$ . In core/helper/query order define

$$\mathbf{C}_{\text{act}} = \text{diag}(\mathbf{C}_C, \mathbb{I}_p, \mathbb{I}_p) \in \mathbb{R}^{(N+2p) \times M_{\text{act}}}, \quad \mathbf{C}_{\text{act}}^\top \mathbf{C}_{\text{act}} = \mathbb{I}.$$

It lifts an active physical state by  $\mathcal{E}_{\text{act}}(\mathbf{W}, \mathbf{V}) = (\mathbf{C}_{\text{act}} \mathbf{W}, \mathbf{V} \mathbf{C}_{\text{act}}^\top)$ . The paired rows have identical masks, so positive homogeneity gives the analytic observer table  $\mathbf{H}^{\text{ana}} = \mathbf{H}^{\text{act}} \mathbf{C}_{\text{act}}^\top$  and preserves logits and the example-row Gram matrix  $\mathbf{H} \mathbf{H}^\top$ . Dormant rows are ordinary independent physical coordinates, not part of this isometry.

**Lemma E.25** (Positive lifts and full-column observers). Let  $\mathbf{C} \in \mathbb{R}^{m \times r}$  satisfy  $\mathbf{C}^\top \mathbf{C} = \mathbb{I}$ , with each row a positive multiple of a coordinate row. For  $(\mathbf{W}, \mathbf{V}) = (\mathbf{C} \mathbf{W}_c, \mathbf{V}_c \mathbf{C}^\top)$ , the feature table is  $\mathbf{H} = \mathbf{H}_c \mathbf{C}^\top$  and the logits agree. For every example-row map  $\mathbf{P}$ ,  $\|\mathbf{P} \mathbf{H}\|_F = \|\mathbf{P} \mathbf{H}_c\|_F$ , also on each feature block preserved by  $\mathbf{C}$ . The example-row Gram matrix and prescribed ridge penalty agree. When that penalty is positive, the complete train-only refit predictions, errors and accuracies agree, and its coefficient matrix is  $\mathbf{C} \mathbf{B}_c$ . Appending zero feature columns preserves these same quantities, with the refit assertions again restricted to a positive penalty. For every compatible matrix  $\mathbf{F}_c$ ,

$$(\mathbf{C} \mathbf{F}_c)^\top (\mathbf{C} \mathbf{F}_c) = \mathbf{F}_c^\top \mathbf{F}_c, \quad \text{pol}(\mathbf{C} \mathbf{F}_c) = \mathbf{C} \text{pol}(\mathbf{F}_c),$$

including the zero-source case.

*Proof.* Positive homogeneity of ReLU gives  $\mathbf{H} = \mathbf{H}_c \mathbf{C}^\top$  rowwise, including at zero preactivations. The identity  $\mathbf{C}^\top \mathbf{C} = \mathbb{I}$  then gives equality of logits and example-row Gram matrices, and

$$\|\mathbf{P} \mathbf{H}\|_F^2 = \text{tr}(\mathbf{P} \mathbf{H}_c \mathbf{C}^\top \mathbf{C} \mathbf{H}_c^\top \mathbf{P}^\top) = \|\mathbf{P} \mathbf{H}_c\|_F^2.$$

The same calculation applies blockwise when the lift preserves that partition. Lemma E.38(i) gives the complete refit assertions and the zero-column assertion.

If  $\mathbf{F}_c = \mathbf{U} \mathbf{\Sigma} \mathbf{Z}^\top$  is a thin positive SVD, then  $\mathbf{C} \mathbf{F}_c = (\mathbf{C} \mathbf{U}) \mathbf{\Sigma} \mathbf{Z}^\top$  is one because  $(\mathbf{C} \mathbf{U})^\top \mathbf{C} \mathbf{U} = \mathbb{I}$ . This proves both source identities; if  $\mathbf{F}_c = 0$ , both Polar factors vanish.  $\square$

### E.2.2 Explicit initialization and uniformly controlled reference dynamics

Use the construction scales of Definition E.3. Put  $\rho_* = \epsilon^4$  and  $\rho_g = \rho_*/8$ . The fixed geometric clock and the certified windows are

$$\begin{aligned} q_{\text{clk}} &= (9/10)^{1/32}, & s &= (1 - q_{\text{clk}})/\lambda, \\ J &= \left\lceil \frac{\log(4/\rho_*)}{-\log q_{\text{clk}}} \right\rceil, & \rho_j &= q_{\text{clk}}^j \quad (0 \leq j \leq J). \end{aligned} \quad (60)$$

The certified windows are

$$\mathcal{I}_{\text{early}} = \{1 \leq j \leq J : \rho_j \geq 9/10\}, \quad \mathcal{I}_{\text{late}} = \{1 \leq j \leq J : \rho_*/2 \leq \rho_j \leq \rho_*\}. \quad (61)$$

**Lemma E.26** (Clock and windows). *For  $0 < \epsilon < 1/200$  and the prescribed windows (61),  $0 < \lambda s < 1/8$ ,  $\rho_j \geq \rho_J > \rho_g$  for  $0 \leq j \leq J$ , and  $\rho_J < \rho_*/2$ . Moreover  $\mathcal{I}_{\text{early}} = \{1, \dots, 32\}$ ,  $\mathcal{I}_{\text{late}}$  contains at least sixteen integers, and  $\max \mathcal{I}_{\text{early}} < \min \mathcal{I}_{\text{late}}$ .*

*Proof.* We first locate the last arrival relative to  $\rho_*$ , then count late arrivals starting at the first crossing of  $\rho_*$ . Let  $a = -\log q_{\text{clk}} = \log(10/9)/32$ . Since  $q_{\text{clk}} > 7/8$ ,  $\lambda s = 1 - q_{\text{clk}} < 1/8$ . With  $J = \lceil \log(4/\rho_*)/a \rceil$ ,

$$q_{\text{clk}}\rho_*/4 < \rho_J \leq \rho_*/4, \quad (62)$$

so  $\rho_J > \rho_g$  and  $\rho_J < \rho_*/2$ . Also  $q_{\text{clk}}^{32} = 9/10$ . Since  $\rho_J \leq \rho_*/4 < 9/10 = \rho_{32}$ , we have  $J > 32$  and hence  $\mathcal{I}_{\text{early}} = \{1, \dots, 32\}$ . Let  $j_0$  be the first index with  $\rho_{j_0} \leq \rho_*$ ; it exists by (62), and  $j_0 \geq 1$  since  $\rho_0 = 1 > \rho_*$ . Minimality gives, for  $0 \leq k \leq 15$ ,

$$\rho_* \geq q_{\text{clk}}^{j_0+k} > q_{\text{clk}}^{k+1} \rho_* \geq q_{\text{clk}}^{16} \rho_* > \rho_*/2,$$

since  $q_{\text{clk}}^{16} = \sqrt{9/10} > 1/2$ . As  $\rho_J < \rho_*/2$ , these indices are strictly below  $J$ . Thus  $j_0, \dots, j_0 + 15$  belong to  $\mathcal{I}_{\text{late}}$ . Finally  $\epsilon < 1/200$  gives  $\rho_* < 9/10$ , so every late index follows 32.  $\square$

Here  $\eta = \eta_V$  is the head data step and  $d_V = \eta_V \lambda_V$  is its per-step decoupled shrinkage; the continuation below allows both parameters to vary about zero.

**Definition E.27** (Actual update map). For head parameters  $\eta, d_V \in \mathbb{R}$ , write  $\mathcal{U}_{\eta, d_V}(\mathbf{W}, \mathbf{V})$  for

$$(q_{\text{clk}} \mathbf{W} + s \mathcal{N}_{\text{pol}}(\mathbf{F}), (1 - d_V) \mathbf{V} + \eta \mathcal{N}_F(\mathbf{G})), \quad (63)$$

with  $\mathbf{F} = \mathbf{F}(\mathbf{W}, \mathbf{V})$ ,  $\mathbf{G} = \mathbf{G}(\mathbf{W}, \mathbf{V})$  and the zero conventions of (3). Use the departure conventions of Section 3.2; the frozen-head reference sets  $\eta = d_V = 0$ .

For each pair  $B$ , let  $\mathbf{j}_B$  be its shared inactive-coordinate vector. Put  $\mathbf{D}_C(x) = \text{diag}((D_B(x))_{B \in \Lambda_c})$ ,  $\mathbf{J}_C = (\mathbf{j}_B^\top)_{B \in \Lambda_c}$ ,  $g = 4\sqrt{2}R/\rho_g$  and  $g_c = \sqrt{2}g$ . The physical anchor is

$$\mathbf{Q}_{\text{QR},c} = \text{pol}(\mathbf{F}_{0,c}) = \mathbf{C}_C^\top \mathbf{Q}_{\text{QR}}, \quad \mathbf{H}_C^{\text{QR},c}(x) = \mathbf{D}_C(x) \mathbf{Q}_{\text{QR},c} x = (\mathbf{H}_{x,:}^{\text{QR}} \mathbf{C}_C)^\top.$$

Here  $\mathbf{H}_C^{\text{QR},c}(x)$  is a feature column; its full table is  $\mathbf{H}_C^{\text{QR},c} = \mathbf{H}^{\text{QR}} \mathbf{C}_C$ . The uniform-source row bounds in Lemma E.19(ii) apply to the physical pair rows with  $\sqrt{P_B} = \sqrt{2}\sqrt{\pi_l}$ .

**Definition E.28** (Static convex initialization). Before initializing the network, solve the following fixed-mask program:

$$\begin{aligned} \mathcal{A} &= \{\mathbf{A} \in \mathbb{R}^{(N/2) \times 2p} : \mathbf{A} \mathbf{P}_E = \mathbf{A}, \|\mathbf{A}\|_{\text{op}} \leq 1\}, \\ \Phi_\epsilon(\mathbf{A}) &= \frac{1}{n} \sum_{x \in \mathcal{S}} \text{CE}(R \bar{\mathbf{V}}_{C,c} \mathbf{D}_C(x) \mathbf{A} x, y^*(x)), \\ \mathbf{Q}_\epsilon &= \arg \min_{\mathbf{A} \in \mathcal{A}} \Phi_\epsilon(\mathbf{A}). \end{aligned}$$

Lemma E.30 proves existence and uniqueness for small  $\epsilon$ . This is a one-time training-data-dependent initializer, not a projection or an additional operation in the trained optimizer. The analytic lift is  $\mathbf{Q}_\epsilon^{\text{ana}} = \mathbf{C}_C \mathbf{Q}_\epsilon$ . For this static program only,  $P_A(\cdot | x)$  denotes the softmax of  $R\bar{\mathbf{V}}_{C,c} \mathbf{D}_C(x) \mathbf{A}x$ , and define

$$\hat{\mathbf{F}}(\mathbf{A})_B = \frac{\sqrt{P_B}}{n} \sum_{x \in \mathcal{S}} D_B(x) P_A(k_B | x) x^\top. \quad (64)$$

The formal masks are used only to define this convex program; its minimizer will be shown to have precisely these masks in the actual ReLU network.

The analytic core rows (used only as the positive lift of the physical pair bank) are

$$\mathbf{W}_{C,l}^{0,\text{ana}} = R\mathbf{Q}_{\epsilon,l}^{\text{ana}} - g\mathbf{j}^\top, \quad \mathbf{V}_{C,l}^{0,\text{ana}} = -v\sqrt{\pi_l} T_{k_l}. \quad (65)$$

The physical incoming rows are  $\mathbf{W}_{C,B}^0 = R\mathbf{Q}_{\epsilon,B} - g_c\mathbf{j}_B^\top$ , with heads

$$\mathbf{V}_{C,B}^0 = -v\sqrt{P_B} T_{k_B}. \quad (66)$$

They equal  $\mathbf{C}_C^\top \mathbf{W}_C^{0,\text{ana}}$  and  $\mathbf{V}_C^{0,\text{ana}} \mathbf{C}_C$ , respectively. Starting from these rows, the helper/query rows and the dormant rows of Definition E.20, keep the heads frozen and update by  $\mathcal{U}_{0,0}$ . Let  $\mathbf{W}_{C,j}^\circ$  be the physical core incoming block at arrival  $j$ , so  $\mathbf{H}_{C,j}^{\text{phys}} = \text{ReLU}(\mathbf{X}_{\text{grid}}(\mathbf{W}_{C,j}^\circ)^\top)$ .

**Definition E.29** (Discrete reference feature tables). The reference trajectory is constructed in Lemma E.30; only its discrete arrivals are used. The physical active feature table and its padded width- $M$  table are

$$\mathbf{H}_j^{\text{act}} = [\mathbf{H}_{C,j}^{\text{phys}} | \mathbf{H}_{N,j}^{\text{phys}} | \mathbf{H}_{Q,j}^{\text{phys}}], \quad \mathbf{H}_j^M = [\mathbf{H}_j^{\text{act}} | 0_{p^2 \times (M - (N/2 + 2p))}]. \quad (67)$$

Each block is the ReLU feature table of the corresponding physical rows at arrival  $j$ . The analytic observer, a lift of this same active trajectory, is

$$\mathbf{H}_j^{\text{ana}} = [\mathbf{H}_{C,j}^{\text{phys}} \mathbf{C}_C^\top | \mathbf{H}_{N,j}^{\text{phys}} | \mathbf{H}_{Q,j}^{\text{phys}}]. \quad (68)$$

At every reference arrival the helper and query blocks are respectively  $\rho_j L_{\text{lift}} \mathbf{J}_{\text{row}}$  and  $\rho_j \mathbf{H}_Q^{\text{wit}}$ ; the dormant feature columns vanish. The complete refit with its prescribed penalty is computed from any of  $\mathbf{H}_j^M, \mathbf{H}_j^{\text{act}}, \mathbf{H}_j^{\text{ana}}$ ; these tables have the same example-row Gram matrix  $\mathbf{H}\mathbf{H}^\top$  and Frobenius norms after the fixed isometric lift and zero padding. The point feature  $\mathbf{H}_{C,j}^{\text{phys}}(x) = h_{C,j}^{\text{phys}}(x)$  is a column vector, and  $H_{B,j}(x)$  is its component indexed by  $B$ .

**Feature-table order.** Analytic observer tables use the fixed order core/helper/query. Physical width- $M$  tables append dormant zero columns at the right. The analytic lift in (68) preserves the example-row Gram matrix; zero padding preserves it as well.

**Lemma E.30** (Static convex core and exact reference dynamics). *Fix a prime  $p \geq 101$  and  $\lambda > 0$ . For all sufficiently small  $\epsilon > 0$ , the static program just defined has a unique minimizer  $\mathbf{Q}_\epsilon$ . It is a supported partial isometry,  $\mathbf{Q}_\epsilon^\top \mathbf{Q}_\epsilon = \mathbf{P}_E$ , and  $\mathbf{Q}_\epsilon \rightarrow \mathbf{Q}_{\text{QR},c}$  as  $\epsilon \downarrow 0$ .*

*The actual frozen-head reference initialized above has, for every  $0 \leq j \leq J$ , the exact core weights and feature vector*

$$\mathbf{W}_{C,j}^\circ = R\mathbf{Q}_\epsilon - \rho_j g_c \mathbf{J}_C, \quad \mathbf{H}_{C,j}^{\text{phys}}(x) = R\mathbf{D}_C(x) \mathbf{Q}_\epsilon x. \quad (69)$$

*Every departure source has rank  $r$ . The grid gates are the prescribed ones: core row  $B$  is active exactly on  $\mathcal{D}_B$ , helper row  $u$  exactly when the first operand is  $u$ , query row  $u$  exactly on  $\mathcal{D}_u$ , and dormant rows never. Uniformly in  $j \leq J$ , the physical core table divided by  $R$  converges to  $\mathbf{H}_C^{\text{QR},c}$ ; on each active mask its values satisfy*

$$\frac{1}{5}R\sqrt{P_B} < H_{B,j}(x) < 9R\sqrt{P_B}.$$

*The reference training logits, and hence its training CE, are constant throughout this horizon.*

*Proof.* The proof has three steps. Uniform closeness to the source  $\mathbf{F}_{0,c}$  gives full supported rank throughout the static program. First-order optimality then identifies its minimizer as a Polar factor and proves uniqueness. Finally, strict gate signs let us substitute that same factor into the actual update and verify (69) by induction.

**The static program.** The objective  $\Phi_\epsilon$  is continuous and convex on the compact convex set  $\mathcal{A}$ , so it has a minimizer. Lemma E.6(iv) and the fixed pair lift give  $\bar{\mathbf{V}}_{C,c} \bar{\mathbf{V}}_{C,c}^\top = (v^2/p) \mathbf{P}_c$ . For every  $\mathbf{A} \in \mathcal{A}$ , therefore,

$$\|R \bar{\mathbf{V}}_{C,c} \mathbf{D}_C(x) \mathbf{A} x\|_2 \leq \epsilon \sqrt{2/p}. \quad (70)$$

Label exclusion gives  $-\nabla_{\mathbf{A}} \Phi_\epsilon(\mathbf{A}) = \epsilon \hat{\mathbf{F}}(\mathbf{A})$ , with the fixed-mask source (64). Each softmax coordinate has gradient  $\ell_1$  norm at most  $1/2$ , so (70) controls its deviation from  $1/p$ . Using  $\sum_B P_B D_B(x) \leq 1$  and  $\|x\|_2 = \sqrt{2}$  gives

$$\|\hat{\mathbf{F}}(\mathbf{A}) - \mathbf{F}_{0,c}\|_F \leq \sqrt{2} \sup_{x,k} |P_A(k | x) - 1/p| \leq \epsilon/\sqrt{p}. \quad (71)$$

All sources are supported on  $E$ . The uniform bound (71), together with the singular-value floor in Lemma E.19(i) and the pair isometry, therefore gives full supported rank for every  $\mathbf{A} \in \mathcal{A}$  when  $\epsilon$  is small.

**Optimality identifies the Polar factor.** At a minimizer, the first-order inequality says that  $\mathbf{A}$  maximizes  $\langle \hat{\mathbf{F}}(\mathbf{A}), \mathbf{B} \rangle_F$  over  $\mathbf{B} \in \mathcal{A}$ . For a source of rank  $r$  on  $E$ , the unique maximizer is its supported Polar factor: in an SVD, each positive singular value forces  $\mathbf{B} v_i = u_i$  in the inequality  $\langle u_i, \mathbf{B} v_i \rangle \leq 1$ . Thus every minimizer satisfies the fixed-point and isometry identities

$$\mathbf{A} = \text{pol}(\hat{\mathbf{F}}(\mathbf{A})), \quad \mathbf{A}^\top \mathbf{A} = \mathbf{P}_E. \quad (72)$$

If  $\mathbf{A}, \mathbf{B}$  are minimizers, convexity makes  $\Phi_\epsilon$  constant on their segment. Thus  $\langle \hat{\mathbf{F}}(\mathbf{A}), \mathbf{B} - \mathbf{A} \rangle_F = 0$ , so  $\mathbf{B}$  also maximizes  $\langle \hat{\mathbf{F}}(\mathbf{A}), \cdot \rangle_F$  on  $\mathcal{A}$ . The unique-maximizer statement above gives  $\mathbf{B} = \mathbf{A}$ . Apply (72) at this unique minimizer. Continuity of the supported Polar map at  $\mathbf{F}_{0,c}$  and (71) give  $\mathbf{Q}_\epsilon \rightarrow \mathbf{Q}_{\text{QR},c}$ .

**Verification of the reference trajectory.** The anchor row bounds of Lemma E.19(ii) transfer to  $\mathbf{Q}_\epsilon$ , giving the stated active values for small  $\epsilon$ . For an inactive mask,  $\mathbf{J}_C x \geq 1$  in that row while  $|(\mathbf{Q}_\epsilon x)_B| \leq \sqrt{2}$ . Lemma E.26 and the shield give  $\rho_j g_c \geq \rho_j g_c > 16 q_{\text{clk}} R > \sqrt{2} R$ . Thus every candidate core preactivation in (69) has the prescribed strict sign.

Helpers have zero heads; queries are strictly inactive on training; dormant rows are negative everywhere. Their sources are zero and they decay exactly. A zero source row has zero Polar row, and appending such rows preserves source rank and normalization. Noncore training logits therefore vanish. The core training logits are exactly the fixed-mask logits at  $\mathbf{Q}_\epsilon$ , with source  $v \hat{\mathbf{F}}(\mathbf{Q}_\epsilon)$ . By (72) and rank  $r$ , its normalized Polar factor is  $\mathbf{Q}_\epsilon / \sqrt{r}$ .

Finally  $s/\sqrt{r} = (1 - q_{\text{clk}})R$ , so substituting the proposed incoming weights into the actual feature update gives

$$q_{\text{clk}}(R \mathbf{Q}_\epsilon - \rho_j g_c \mathbf{J}_C) + s \mathbf{Q}_\epsilon / \sqrt{r} = R \mathbf{Q}_\epsilon - \rho_{j+1} g_c \mathbf{J}_C. \quad (73)$$

The initialization gives the case  $j = 0$ , and (73) supplies the induction step. The gate and source checks above therefore prove the exact trajectory, gates and rank through  $J$ . The feature convergence now follows directly from  $\mathbf{Q}_\epsilon \rightarrow \mathbf{Q}_{\text{QR},c}$ , uniformly in all arrivals, without a drift accumulation. The constant core logits prove the final CE assertion.  $\square$

## E.3 Claim II: useful-representation consequences on the same trajectory

### E.3.1 The Fourier-fraction bound

We derive the supporting source and energy estimates for  $p \geq 97$ , then prove the  $p \geq 101$  anchor bound and net gain needed by Theorem 4.1. The transfer uses the discrete actual-arrival limits of Lemma E.35. All terminating decimals below are exact rationals. The argument first isolates the anchor's common component in (75)–(76), then controls its remainder by comparing the two Polar factors. Proposition E.36 combines these estimates and transfers the resulting Fourier bound to the arrivals.

#### E.3.1.1 The anchor and common-component orthogonality

**Definition E.31** (Anchor energy and masked incoming map). Put  $K_0 = \sqrt{u(1-u/p^2)}$ ,  $a_p = (p^2 + 3)/[3(3p^2 + 1)]$  and  $\mathcal{J}_l = \sqrt{\varpi}g_l$ . Let  $\bar{\mathbf{F}} = \sqrt{\varpi}\mathbf{D}_{\text{grid}}^\top \mathbf{X}_{\text{grid}}$ ,  $\bar{\mathbf{Q}} = \text{pol}(\bar{\mathbf{F}})$ ,  $\mathbf{G} = p^3 \mathbf{F}_{\text{unif}}$ ,  $\bar{\alpha} = (p+1)/(4p)$  and  $m = \langle \mathbf{Q}_{\text{QR}}, \bar{\mathbf{F}} \rangle_F / u$ . The masked incoming map is  $\mathcal{L}(\mathbf{T}) = \mathbf{D}_{\text{grid}} \odot (\mathbf{X}_{\text{grid}} \mathbf{T}^\top)$  for an incoming matrix  $\mathbf{T}$ . The mask indicators and the full-bank average  $\mathbb{E}$  are as above, equivalently evaluated on the positive pair lift with the original weights.

Here  $\mathcal{J}$  is a table, distinct from the scalar line sum  $J$ . For row  $l$  write  $\mathbf{Q}_{\text{QR},l} = (w_l^\top, v_l^\top)$  in the two operand blocks. Then its masked column is

$$H_l(a, z) = 1_{A_l}(a) 1_{B_l}(z) [w_l(a) + v_l(z)]. \quad (74)$$

**Lemma E.32** (Full-mask common-component identity). *For the uniform-source anchor  $\mathbf{H} = \mathbf{H}^{\text{QR}}$ ,*

$$\|\mathbf{C}_{\text{grid}} \mathbf{H}\|_F^2 = m^2 K_0^2 + \|\mathbf{C}_{\text{grid}}(\mathbf{H} - m\mathcal{J})\|_F^2, \quad \|\mathbf{C}_{\text{grid}} \mathcal{J}\|_F^2 = K_0^2. \quad (75)$$

Moreover,

$$\|\mathbf{P}_{\text{add}} \mathbf{H}\|_F^2 = a_p K_0^2 m^2 + \|\mathbf{P}_{\text{add}}(\mathbf{H} - m\mathcal{J})\|_F^2, \quad \|\mathbf{P}_{\text{add}} \mathcal{J}\|_F^2 = a_p K_0^2. \quad (76)$$

For every incoming matrix  $\mathbf{T}$ , the centered masked map satisfies

$$\|\mathbf{C}_{\text{grid}} \mathcal{L}(\mathbf{T})\|_F^2 \leq (p - u/p) \|\mathbf{T}\|_F^2.$$

*Proof.* The rectangle calculation gives the indicator energies and the masked-map bound. The remaining task is to identify the two cross moments that make the energy decompositions orthogonal.

*Rectangle geometry.* Every rectangle has size  $u$ , so its centered indicator has squared norm  $K_0^2$ . Common dilation and independent translations reduce addition counts to  $t(y) = y + 1$  for  $0 \leq y < b$ ,  $t(y) = 2b - y$  for  $b \leq y < 2b$ , and  $t(2b) = 0$ . Thus  $\sum_y t(y)^2 = up/3$  and  $\mathbf{P}_{\text{add}} g(a, z) = t(a+z)/p - u/p^2$ , proving the two indicator norms.

For one incoming row, restrict its operand profiles to  $A, B$ , and decompose them into active means  $w_0, v_0$  and mean-zero parts  $w^\circ, v^\circ$ . The centered masked squared norm is

$$c\|w^\circ\|^2 + b\|v^\circ\|^2 + K_0^2(w_0 + v_0)^2.$$

Its input squared norm is at least  $\|w^\circ\|^2 + \|v^\circ\|^2 + bw_0^2 + cv_0^2$ . Weighted Cauchy-Schwarz gives the row-map norm squared  $\max\{b, c, K_0^2(1/b + 1/c)\} = p - u/p$ . Summing rows proves the map bound.

*The full-mask cross moment.* Let  $\mathcal{L}^*$  be the Frobenius adjoint of  $\mathcal{L}$ , and put  $\mathbf{Z} = \mathcal{L}^* \mathbf{P}_{\text{add}} \mathcal{J} - a_p(1 - u/p^2) \bar{\mathbf{F}}$ . For each mask write  $\alpha = \mathbf{1}_A$ ,  $\beta = \mathbf{1}_B$  and let  $\psi_A$  be the centered restriction to  $A$  of  $a \mapsto \sum_{z \in B} (\mathbf{P}_{\text{add}} g_l)(a, z)$ , extended by zero outside  $A$ ;  $\psi_B$  reverses operands. Set  $\mathbf{M}_A = \mathbb{E}[\alpha \psi_A^\top]$  and  $\mathbf{M}_B = \mathbb{E}[\beta \psi_B^\top]$ . The adjoint map gives

$$\mathcal{L}^* \mathbf{C}_{\text{grid}} \mathcal{J} = (1 - u/p^2) \bar{\mathbf{F}}, \quad \mathbf{Z}_l = \sqrt{\varpi}(\psi_A^\top, \psi_B^\top). \quad (77)$$

Indeed each active restriction has mean  $a_p K_0^2/b$  or  $a_p K_0^2/c$ , because its total is  $\|\mathbf{P}_{\text{add}} g_l\|^2 = a_p K_0^2$ . For fixed slope and  $A$ ,

$$\sum_{z \in B} (\mathbf{P}_{\text{add}} g_l)(a, z) = \frac{1}{p} \sum_{a' \in A} |B \cap (B + a' - a)| - \frac{cu}{p^2},$$

so  $\psi_A$  is independent of the translation of  $B$ . The analogous statement holds for  $\psi_B$ .

Full-mask averaging is invariant under every common affine map of the field. Under  $(a, z) \mapsto (\sigma a + t, \sigma z + t)$ , the addition coordinate becomes  $\sigma(a + z) + 2t$ . Thus  $\mathbf{P}_{\text{add}}$  commutes with the action and the two restricted profiles transform covariantly. Consequently  $\mathbf{M}_A$  and  $\mathbf{M}_B$  commute with its permutation matrices. The affine group is two-transitive, so each matrix has the form  $r\mathbb{I} + s\mathbf{1}\mathbf{1}^\top$ . For  $\mathbf{M}_A$ , both  $\mathbf{M}_A \mathbf{1} = 0$  and  $\text{tr } \mathbf{M}_A = \mathbb{E}[\langle \alpha, \psi_A \rangle] = 0$ . These equations force  $r = s = 0$ . The same argument gives  $\mathbf{M}_B = 0$ . Since  $\mathbf{1}^\top \mathbf{M}_A = b\mathbb{E}[\psi_A^\top]$  and  $\mathbf{1}^\top \mathbf{M}_B = c\mathbb{E}[\psi_B^\top]$ , both profile expectations also vanish.

The exact QR count formulas, averaged over the translation of  $B$ , give

$$\mathbb{E}_B[\mathbf{n}^L] = \frac{c}{2}(1 - p^{-1})\alpha, \quad \mathbb{E}_B[\mathbf{n}^R] = \frac{c}{2p}(b\mathbf{1} - \alpha + \mathbf{C}_+ \alpha).$$

For the opposite conditioning they give

$$\mathbb{E}_A[\mathbf{n}^R] = \frac{b}{2}(1 - p^{-1})\beta, \quad \mathbb{E}_A[\mathbf{n}^L] = \frac{b}{2p}(c\mathbf{1} - \beta + \mathbf{C}_-\beta).$$

The source definition gives  $\mathbf{G}_l = (p/b)\sqrt{\varpi}(\mathbf{n}^{L\top}, \mathbf{n}^{R\top})$ . Since  $\mathbf{M}_A = \mathbf{M}_B = 0$  and  $\mathbf{C}_\pm$  are fixed linear operators, the four averaged count identities therefore imply  $\mathbf{G}^*\mathbf{Z} = 0$ . Writing the supported Polar factor as  $\mathbf{Q}_{\text{QR}} = \mathbf{G}(\mathbf{G}^*\mathbf{G})^{\dagger/2}$  therefore gives  $\langle \mathbf{Q}_{\text{QR}}, \mathbf{Z} \rangle_F = 0$ .

Using (77) and  $\langle \mathbf{Q}_{\text{QR}}, \mathbf{Z} \rangle_F = 0$  in the definition of  $\mathbf{Z}$  gives the required cross moments:

$$\langle \mathbf{C}_{\text{grid}}\mathbf{H}, \mathbf{C}_{\text{grid}}\mathcal{J} \rangle = mK_0^2, \quad \langle \mathbf{P}_{\text{add}}\mathbf{H}, \mathbf{P}_{\text{add}}\mathcal{J} \rangle = a_pmK_0^2. \quad (78)$$

Together with the indicator norms, (78) identifies  $m\mathbf{C}_{\text{grid}}\mathcal{J}$  and  $m\mathbf{P}_{\text{add}}\mathcal{J}$  as the respective orthogonal projections. Pythagoras gives (75) and (76).  $\square$

### E.3.1.2 Same-source energy estimates

**Definition E.33** (Binary source comparison). For  $t \neq 0$ , subscripts  $t$  indicate the ordered input pair  $(e_t^L, e_t^R)$ . Put  $\varepsilon_p = (p^2 + 3)/[3(p - 1)^3]$  and  $\delta_Q^2 = \|\mathbf{Q}_{\text{QR}} - \bar{\mathbf{Q}}\|_F^2$ . The common active value used below is  $h_o = 2\sqrt{p/(p+1)} + \sqrt{2/p}$ . For  $t \neq 0$ , use the ordered Fourier pair  $(e_t^L, e_t^R)$  and put

$$\eta_t = \frac{-1 + \chi_2(t)\tau}{p-1}, \quad \mathbf{S}_t = \sqrt{\bar{\alpha}} \text{diag}(c, b), \quad \mathbf{K}_t = \bar{\mathbf{Q}}_t^* \mathbf{G}_t.$$

For the constant mode put  $\nu_c = \sqrt{2}u/\sqrt{p}$ ,  $\bar{\mathbf{Q}}_c = \mathbf{1}/\sqrt{N}$  and  $\ell_c = \mathbf{G}e_c - \nu_c\bar{\mathbf{Q}}_c$ .

**Lemma E.34** (Polar distance from the binary comparison). *For every prime  $p \geq 97$ :*

- (i)  $\bar{\mathbf{F}}_t = \bar{\mathbf{Q}}_t \mathbf{S}_t$  and  $\mathbf{K}_t = \sqrt{\bar{\alpha}} \begin{bmatrix} c & c\bar{\eta}_t \\ b\eta_t & b \end{bmatrix}$ . On constants,  $\bar{\mathbf{F}}e_c = \nu_c\bar{\mathbf{Q}}_c$  and  $\bar{\mathbf{F}}(\mathbf{1}, -\mathbf{1}) = 0$ . On every active rectangle the corresponding row of  $\bar{\mathbf{Q}}$  has the same value  $\sqrt{\varpi}[\bar{\alpha}^{-1/2} + \sqrt{2/p}]$ .
- (ii)  $\mathbf{G}e_c = \nu_c\bar{\mathbf{Q}}_c + \ell_c$ , with  $\ell_c \perp \bar{\mathbf{Q}}_c$  and  $\|\ell_c\|^2 = \varepsilon_p\nu_c^2$ .
- (iii)  $\delta_Q^2 < 2$ .

*Proof.* (i) Apply (32) with  $\bar{\mathbf{F}} = \mathbf{B}^\top \mathbf{X}_{\text{grid}}$  and  $\mathbf{G} = (p/b)\mathbf{B}^\top \mathcal{P}_Q \mathbf{X}_{\text{grid}}$ . The unmasked input modes have sole Fourier coefficients  $\sqrt{p}$  at  $(t, 0)$  and  $(0, t)$ ; the masked coefficients in (40) give

$$\bar{\mathbf{F}}_t^* \bar{\mathbf{F}}_t = \bar{\alpha} \text{diag}(c^2, b^2), \quad \bar{\mathbf{F}}_t^* \mathbf{G}_t = \frac{1}{2p} \begin{bmatrix} c^3 & c^3\bar{\eta}_t \\ b^2c\eta_t & b^2c \end{bmatrix} = \mathbf{S}_t \mathbf{K}_t.$$

Every row of  $\bar{\mathbf{F}}$  is  $\sqrt{\varpi}(c\mathbf{1}_A, b\mathbf{1}_B)$ , and its constant coordinate is  $\sqrt{\varpi}\sqrt{2/p}u$ . Consequently

$$\bar{\mathbf{Q}} = \bar{\mathbf{F}}[\bar{\alpha}^{-1/2} \text{diag}(c^{-1}, b^{-1}) \otimes \mathbf{P}_c + \nu_c^{-1}e_c e_c^\top], \quad \bar{\mathbf{Q}}_l x_{az} = \sqrt{\varpi}[\bar{\alpha}^{-1/2}(\mathbf{1}_A(a) + \mathbf{1}_B(z) - 1) + \sqrt{2/p}].$$

This proves all claims in (i), including the common active value.

(ii) The constant coordinate of row  $l$  of  $\mathbf{G}$  is  $\sqrt{\varpi}(p/b)\sqrt{2/p}n_l$ . Lemma E.12(iv) gives its mean and variance, so  $\|\ell_c\|^2 = (2p/b^2) \text{Var}(n_l) = \varepsilon_p\nu_c^2$ .

(iii) We bound the nonconstant Polar blocks by a source-error estimate and lower singular-value bounds, then add the constant block from (ii). Write  $\text{Herm}(\mathbf{A}) = (\mathbf{A} + \mathbf{A}^*)/2$  and set

$$\mathbf{S}'_t = \sqrt{\bar{\alpha}} \begin{bmatrix} c & p\bar{\eta}_t/2 \\ p\eta_t/2 & b \end{bmatrix}.$$

By covariance and (i),  $\mathbf{G}_t^* \mathbf{G}_t = p\mathbf{M}_{\chi_2(t)}$ ,  $\bar{\mathbf{Q}}_t^* \mathbf{G}_t = \mathbf{K}_t$ , and  $\mathbf{S}'_t = \text{Herm}(\mathbf{K}_t)$ . Set  $\mathbf{E}_t = \mathbf{G}_t - \bar{\mathbf{Q}}_t \mathbf{S}'_t$  and  $\Gamma = p\sqrt{\bar{\alpha}}/2$ . Thus  $\|\mathbf{E}_t\|_F^2 = \text{tr}(\mathbf{G}_t^* \mathbf{G}_t) - \text{tr}((\mathbf{S}'_t)^2)$ . Equal phase counts,  $\mathbf{M}_+ + \mathbf{M}_- = 2g_0^{-2}(\mathbb{I} + \mathbf{E})$ , and  $\sum_{t \neq 0} |\eta_t|^2 = (p+1)/(p-1)$  give the exact trace sums

$$\sum_{t \neq 0} \text{tr}(\mathbf{G}_t^* \mathbf{G}_t) = 2p(p-1)g_0^{-2}(1 + \mathbf{e}_0), \quad \sum_{t \neq 0} \text{tr}((\mathbf{S}'_t)^2) = \bar{\alpha} \left[ (p-1)(b^2 + c^2) + \frac{p^2(p+1)}{2(p-1)} \right].$$

The Gauss-sum identity  $|\tau|^2 = p$  and  $\sum_{t \neq 0} \chi_2(t) = 0$  give the stated  $\eta$  sum, so

$$\begin{aligned} \sum_{t \neq 0} \frac{\|\mathbf{E}_t\|_F^2}{\Gamma^2} &= 2(p-1) \left[ (1+p^{-2})\mathbf{e}_0 - \frac{p+1}{(p-1)^2} \right] \\ &= \frac{2(p+1)(2p^2-3)}{3p^2(p-1)} = \frac{4}{3} \frac{p+1}{p-1} \left( 1 - \frac{3}{2p^2} \right) \\ &< \frac{4}{3} \cdot \frac{49}{48} = \frac{49}{36}. \end{aligned} \quad (79)$$

For full-column-rank  $\mathbf{X} = \mathbf{Q}\mathbf{T}$ ,  $\mathbf{Y} = \mathbf{U}\mathbf{S}$  with  $\mathbf{T}, \mathbf{S} \succ 0$  and  $\mathbf{Q}, \mathbf{U}$  isometries,

$$2 \operatorname{Re} \langle \mathbf{X} - \mathbf{Y}, \mathbf{Q} - \mathbf{U} \rangle_F = 2 \operatorname{tr} [\mathbf{T}(\mathbb{I} - \operatorname{Herm}(\mathbf{Q}^*\mathbf{U})) + \mathbf{S}(\mathbb{I} - \operatorname{Herm}(\mathbf{U}^*\mathbf{Q}))] \geq (\sigma_{\min}(\mathbf{T}) + \sigma_{\min}(\mathbf{S})) \|\mathbf{Q} - \mathbf{U}\|_F^2,$$

because each  $\mathbb{I} - \operatorname{Herm}(\mathbf{Q}^*\mathbf{U}) \succeq 0$  and its trace is  $\|\mathbf{Q} - \mathbf{U}\|_F^2/2$ . Cauchy–Schwarz therefore yields  $\|\mathbf{Q} - \mathbf{U}\|_F \leq 2\|\mathbf{X} - \mathbf{Y}\|_F/(\sigma_{\min}(\mathbf{X}) + \sigma_{\min}(\mathbf{Y}))$ . Lemma E.18(i) and  $g_0^2\bar{\alpha} = 4p/(p^2+1)$  give

$$\frac{\sigma_{\min}(\mathbf{G}_t)}{\Gamma} > \sqrt{\frac{11}{15}(1+p^{-2})} > \frac{17}{20}. \quad (80)$$

Also  $\mathbf{S}'_t = \sqrt{\bar{\alpha}} \begin{bmatrix} c & p\bar{\eta}_t/2 \\ p\eta_t/2 & b \end{bmatrix}$  and  $|\eta_t| \leq (\sqrt{p}+1)/(p-1)$  imply

$$\frac{\lambda_{\min}(\mathbf{S}'_t)}{\Gamma} \geq 1 - \frac{1}{p} - |\eta_t| \geq 1 - \frac{1}{p} - \frac{1}{\sqrt{p}-1} > \frac{7}{8} \quad (p \geq 97). \quad (81)$$

The last inequality follows from  $1/p+1/(\sqrt{p}-1) < 1/97+5/44 < 1/8$ . In particular  $\mathbf{S}'_t \succ 0$ . Apply the Polar inequality to  $\mathbf{X} = \mathbf{G}_t$ ,  $\mathbf{Y} = \mathbf{Q}_t\mathbf{S}'_t$ , whose polar factors are  $\mathbf{Q}_{\text{QR},t}$  and  $\bar{\mathbf{Q}}_t$ . The two lower bounds (80)–(81) give the denominator  $69\Gamma/40$ ; summing and using (79) bounds the nonconstant squared Polar distance by

$$\frac{4 \sum_{t \neq 0} \|\mathbf{E}_t\|_F^2}{(69\Gamma/40)^2} < \frac{78400}{42849} < \frac{19}{10}.$$

By (ii), the constant block contributes exactly  $2[1 - (1 + \varepsilon_p)^{-1/2}] \leq \varepsilon_p$ , where  $\varepsilon_p = (p^2+3)/(3(p-1)^3) < 1/[2(p-1)] \leq 1/192$ . Thus  $\delta_Q^2 < 19/10 + 1/192 < 2$ .  $\square$

### E.3.2 Actual-arrival table and Fourier transfer

Use the scales in Definition E.3 and clock (60), and write  $M_{\text{act}} = N/2 + 2p$  for the active physical width. The analytic observer table is the fixed active-bank lift (68); it is not a second trajectory. The physical table at width  $M$  is  $[\mathbf{H}_j^{\text{act}} \mid \mathbf{0}_{M-M_{\text{act}}}]$ ; these zero columns are appended to the physical active table, not to its analytic lift.

**Lemma E.35** (Actual-arrival table limits). *Fix a prime  $p \geq 101$  and  $\lambda > 0$ . As  $\epsilon \downarrow 0$ , the actual arrivals of the frozen-head reference path satisfy*

$$\sup_{\substack{0 \leq j \leq J \\ \rho_j \geq 9/10}} \left\| \frac{\mathbf{H}_j^{\text{ana}}}{\rho_j L_{\text{lift}}} - [0 \mid \mathbf{J}_{\text{row}} \mid 0] \right\|_F \longrightarrow 0, \quad (82)$$

$$\sup_{\substack{0 \leq j \leq J \\ \rho_j \leq \epsilon^4}} \left\| \frac{\mathbf{H}_j^{\text{ana}}}{R} - [\mathbf{H}^{\text{QR}} \mid 0 \mid 0] \right\|_F \longrightarrow 0. \quad (83)$$

Both limits are uniform over the indicated discrete arrival sets.

*Proof.* Both limits follow from the same three-block table. The early normalization keeps the helper block at unit scale; the late normalization keeps the core at unit scale. The exact trajectory from Lemma E.30 and  $\mathbf{H}^{\text{QR}} \mathbf{C}_C \mathbf{C}_C^\top = \mathbf{H}^{\text{QR}}$  (reflection-paired feature columns coincide) give

$$\begin{aligned} \mathbf{H}_j^{\text{ana}} &= [\mathbf{H}_{C,0}^{\text{phys}} \mathbf{C}_C^\top \mid \rho_j R \epsilon^{-3} \mathbf{J}_{\text{row}} \mid \rho_j \mathbf{H}_Q^{\text{wit}}], \\ R^{-1} \mathbf{H}_{C,0}^{\text{phys}} \mathbf{C}_C^\top &\longrightarrow \mathbf{H}^{\text{QR}}. \end{aligned} \quad (84)$$

For (82), divide (84) by  $\rho_j R \epsilon^{-3}$ : the core block is at most  $(10/9)\epsilon^3 \|\mathbf{H}_{C,0}^{\text{phys}}/R\|_F$  and the query block is less than  $(21p/10)\epsilon^3$ . Both vanish uniformly on the early arrival set; the helper block is exactly  $\mathbf{J}_{\text{row}}$ .

For (83), divide (84) by  $R$ . Since  $\rho_j \leq \epsilon^4$  and  $\|\mathbf{J}_{\text{row}}\|_F = p$ , the total error is bounded by

$$\left\| \mathbf{H}_{C,0}^{\text{phys}} \mathbf{C}_C^\top / R - \mathbf{H}^{\text{QR}} \right\|_F + p\epsilon + (21p/10)\epsilon^4 \longrightarrow 0.$$

This bound is uniform on the late arrival set. Positive lift and zero padding preserve all row Gram matrices and Fourier energies, so the same limits serve the physical observers at every admitted width.  $\square$

**Proposition E.36** (Fourier gain on the frozen-head reference path). *For every prime  $p \geq 101$ , all sufficiently small positive  $\epsilon$ , and every actual arrival  $0 \leq j \leq J$ , the centered table has positive energy. The reference path satisfies*

$$A_H(\mathbf{H}^{\text{QR}}) > \frac{1275}{11879} > 0.107, \quad (85)$$

and, for sufficiently small  $\epsilon$ ,

$$A_H(\mathbf{H}_0^{\text{ana}}) < 0.004, \quad A_H(\mathbf{H}_j^{\text{ana}}) > 0.104 \quad (j \in \mathcal{I}_{\text{late}}). \quad (86)$$

Consequently

$$A_H(\mathbf{H}_j^{\text{ana}}) - A_H(\mathbf{H}_0^{\text{ana}}) > 0.100 > 0.0998 \quad (j \in \mathcal{I}_{\text{late}}).$$

*Proof.* First we lower-bound the anchor's Fourier fraction by its angle with the common component. We then use the two table limits to obtain the early and late bounds on the reference path. Lemma E.34(i) gives  $\mathcal{L}(\bar{\mathbf{Q}}) = h_\circ \mathbf{J}$ , where  $h_\circ > 2$ : use  $2\sqrt{p/(p+1)} > 2 - 2/(p+1)$  and  $\sqrt{2/p} > 2/(p+1)$ . Put  $x = \mathbf{C}_{\text{grid}} \mathbf{H}^{\text{QR}}$  and  $y = h_\circ \mathbf{C}_{\text{grid}} \mathbf{J}$ . The map bound in Lemma E.32, the bound  $\delta_Q^2 < 2$  from Lemma E.34(iii), and  $(p - u/p)/K_0^2 = 4p/(p^2 - 1)$  give

$$\frac{\|x - y\|_F^2}{\|y\|_F^2} < \frac{8p}{(p^2 - 1)h_\circ^2} < \frac{2p}{p^2 - 1} < 1. \quad (87)$$

Thus  $x \neq 0$ . Let  $\theta$  be the angle between  $x$  and  $y$ . The cross moments (78) and the numerator decomposition (76) give  $\langle x, y \rangle = h_\circ m K_0^2$  and  $\|\mathbf{P}_{\text{add}} \mathbf{H}^{\text{QR}}\|_F^2 \geq a_p m^2 K_0^2$ , so  $A_H(\mathbf{H}^{\text{QR}}) \geq a_p \cos^2 \theta$ . The distance from  $y$  to  $\mathbb{R}x$  is at most  $\|x - y\|_F$ , hence  $\sin^2 \theta \leq \|x - y\|_F^2 / \|y\|_F^2$ . Using (87) and  $a_p > 1/9$  therefore gives

$$A_H(\mathbf{H}^{\text{QR}}) > \frac{1}{9} \left( 1 - \frac{2p}{p^2 - 1} \right) \geq \frac{4999}{45900} > \frac{1275}{11879} > 0.107.$$

Here  $p/(p^2 - 1)$  decreases for  $p \geq 101$ .

View the tables as vectors in their Frobenius Hilbert space, with  $\mathbf{C}_{\text{grid}}$  and  $\mathbf{P}_{\text{add}}$  acting columnwise. Thus  $A_H$  is the Rayleigh quotient  $\|\mathbf{P}_{\text{add}} \mathbf{H}\|_F^2 / \|\mathbf{C}_{\text{grid}} \mathbf{H}\|_F^2$  wherever its denominator is positive. For the early limit,  $\mathbf{P}_{\text{add}} \mathbf{J}_{\text{row}} = 0$  and  $\|\mathbf{C}_{\text{grid}} \mathbf{J}_{\text{row}}\|_F^2 = p(p - 1) > 0$ . Thus (82) and continuity give  $A_H(\mathbf{H}_0^{\text{ana}}) \rightarrow 0$ . For late arrivals, (83) gives uniform convergence to the anchor, so  $A_H(\mathbf{H}_j^{\text{ana}})$  converges uniformly to  $A_H(\mathbf{H}^{\text{QR}}) > 0.107$ . This proves both strict bounds in (86).

Finally  $\mathbf{C}_{\text{grid}} \mathbf{H}^{\text{QR}} \neq 0$  was proved above. The uniform core limit in Lemma E.35 therefore keeps the centered core energy positive at every arrival for small enough  $\epsilon$ . The full table contains that core block, so its centered energy is positive as well. Zero padding does not change it.  $\square$

### E.3.3 Margins, cross-entropy, and prediction chronology

Write  $m_{j,k}(x) = f_{j,y^*(x)}(x) - f_{j,k}(x)$  for the individual true-versus-rival contrast, so  $m_j(x) = \min_{k \neq y^*(x)} m_{j,k}(x)$ . For a reference arrival and a rival  $k \neq y^*(x)$ , put

$$m_{j,k}^C(x) = (e_{y^*(x)} - e_k)^\top \bar{\mathbf{V}}_{C,c} h_{C,j}^{\text{phys}}(x).$$

For each rival the uniform active mass is

$$\sum_{B \in \Lambda_c} P_B D_B(x) \mathbf{1}[k_B = k] = m_0, \quad m_0 = \frac{p+1}{4p^2}.$$

**Proposition E.37** (Margins and CE on the frozen-head reference path). *Fix  $p \geq 101$ . For all sufficiently small  $\epsilon > 0$ , every actual arrival  $0 \leq j \leq J$ , grid input  $x$ , and rival  $k \neq y^*(x)$  has positive core margin  $m_{j,k}^C(x)$ , with*

$$\frac{\epsilon m_0}{5} < m_{j,k}^C(x) < 9\epsilon m_0 < \frac{3\epsilon}{p}, \quad m_0 = \frac{p+1}{4p^2}. \quad (88)$$

On training inputs, the reference path has training accuracy one and

$$\log p - 0.0201 < L_j < \log p, \quad \frac{\max_{0 \leq j \leq J} L_j}{\min_{0 \leq j \leq J} L_j} - 1 < 0.0045. \quad (89)$$

On every held-out input, its actual accuracy is zero when  $\rho_j \geq 9/10$  and one when  $\rho_j \leq \epsilon^4$ . In particular these conclusions hold on the prescribed early and late clock windows, including diagonal held-out inputs.

*Proof.* We first bound each core contrast. On training inputs this also bounds the full contrast and hence CE; on held-out inputs we compare it with the decaying query contribution. For each rival class, exactly  $u/2 = (p^2 - 1)/8$  active reflection-pair core blocks contribute. Their masses sum to  $P_B u/2 = (2/N)(u/2) = u/N = m_0$ . By the active-row bounds in Lemma E.30, their normalized active values lie in  $(1/5, 9)$ . The frozen negative-class heads and  $vR = \epsilon$  give the exact contrast formula

$$m_{j,k}^C(x) = v \sum_{B: k_B=k, D_B(x)=1} \sqrt{P_B} H_{B,j}(x). \quad (90)$$

Label exclusion means no active core unit has excluded class  $y^*(x)$ ; the common centered offset in each  $T_k$  head cancels from a true-versus-rival contrast. Inserting the active-row bounds and the mass sum into (90) gives (88). Helpers have zero heads, and the query rows are inactive on training inputs, so these are also the training margins of the full network.

If  $\epsilon < 1/200$ , then  $9\epsilon m_0 < 3\epsilon/p < 0.0201$ . The CE contrast identity (11) and  $0 < m_{j,k} < 9\epsilon m_0$  yield

$$\log p - 9\epsilon m_0 < L_j < \log p.$$

This implies the CE interval in (89). Its relative variation is at most  $0.0201/(\log p - 0.0201) < 0.0045$  for  $p \geq 101$ .

For  $x = (a, b) \in \mathcal{D}$ , only the query indexed by  $a$  is active, and its named class  $c(a) = 2a + 1$  is a rival. In the early window its contribution to the  $c(a)$ -versus-true score difference is at least  $24(9/10)\epsilon/p = 21.6\epsilon/p$ . The core can oppose this by less than  $3\epsilon/p$ , so the named rival wins by more than  $18\epsilon/p$ . Hence every held-out point is misclassified.

In the late window, the adverse query contribution to any margin is at most  $24\epsilon^5/p$ . If  $\epsilon^4 < 1/480$ , then for  $p \geq 101$

$$24\epsilon^5/p < \epsilon m_0/5,$$

because  $m_0 p/120 = (p+1)/(480p) > 1/480$ . The core lower bound (88) therefore leaves every held-out margin strictly positive. The same core-plus-query formulas apply to diagonal inputs.  $\square$

Proposition E.39 uses (82) to retain zero held-out refit accuracy and normalized error greater than one. For  $p \geq 1764$ , Proposition E.52 uses (83) and the QR-anchor refit bound; it gives held-out accuracy at least  $(p-1)/(p+1)$ , nonsquare-difference margins greater than  $11/50$ , and full held-out normalized error below  $121/400$ . The latter accuracy statement deliberately makes no diagonal-correctness claim; the actual network's late accuracy-one conclusion above does.

### E.3.4 The complete refit at early arrivals

The diagnostic uses every current hidden column and the prescribed ridge penalty (8). We first record the normalization and perturbation facts also used by the random-split proof. All matrices below retain their full column sets.

In this generic matrix lemma, the matrix letters, their SVD factors and eigenvalues, and differences such as  $\mathbf{V} - \mathbf{U}$  are freshly quantified objects of compatible dimensions, unrelated to construction objects denoted by the same letter. For the following lemma, write  $\mathbf{B}_U = \mathbf{U}_S^* \mathbf{U}_S + q_U \mathbb{I}$  and  $\mathbf{w}_U = \mathbf{B}_U^{-1} \mathbf{U}_S^* \mathbf{T}$ , with analogous definitions for  $\mathbf{B}_V, \mathbf{w}_V$ ; put  $\mathbf{r}_V = \mathbf{T} - \mathbf{V}_S \mathbf{w}_V$ . For a real scalar  $a$ ,  $(a)_+ = \max(a, 0)$ .

**Lemma E.38** (Full ridge: normalization and perturbation). *Use the full-ridge notation of Appendix N with real matrices,  $\gamma > 0$  and  $q_U > 0$ .*

- (i) *If  $\mathbf{V} = c\mathbf{U}\mathbf{C}^\top$ ,  $c > 0$  and  $\mathbf{C}^\top \mathbf{C} = \mathbb{I}$ , then  $q_V = c^2 q_U$ ,  $\mathbf{V}_A \mathbf{V}_B^\top = c^2 \mathbf{U}_A \mathbf{U}_B^\top$  for any row sets  $A, B$ ,  $\mathbf{w}_V = c^{-1} \mathbf{C} \mathbf{w}_U$ , and  $\mathbf{Y}_V = \mathbf{Y}_U$ . Appending columns which vanish on  $S$  changes neither the penalty nor any prediction; their fitted coefficients are zero. If  $\mathbf{T} \mathbf{1} = 0$ , every predicted row is centered.*

*The complete prediction  $\mathbf{U} \mapsto \mathbf{Y}_U$  is continuous on  $q_U > 0$ . Any finite collection of strict inequalities in its scores and squared errors to fixed targets, with fixed positive error denominators, therefore persists near  $\mathbf{U}$ .*

- (ii) *For any table  $\mathbf{V}$  of the same size with  $q_V > 0$ , the comparison quantities of Appendix N satisfy*

$$\|\mathbf{Y}_V - \mathbf{Y}_U\|_F \leq \frac{te_D}{2\sqrt{q_V}} + L \left[ te_S \left( \frac{1}{\sqrt{q_U}} + \frac{1}{2\sqrt{q_V}} \right) + \frac{|q_V - q_U|N}{2\sqrt{q_U q_V}} \right]. \quad (91)$$

*For the prescribed penalties,*

$$q_V \geq \gamma(\|\mathbf{U}_S\|_F - e_S)_+^2, \quad |q_V - q_U| \leq \gamma e_S(2\|\mathbf{U}_S\|_F + e_S). \quad (92)$$

*Proof.* Part (i) supplies the invariances and continuity needed to transfer both limiting refits. Part (ii) records a quantitative comparison: subtract the normal equations, control the resulting coefficient error, and then pass to query predictions. The dual formula  $\mathbf{w}_U = \mathbf{U}_S^\top (\mathbf{U}_S \mathbf{U}_S^\top + q_U \mathbb{I})^{-1} \mathbf{T}$  proves the normalization, zero-column and centering assertions. On  $q_U > 0$ , continuity of the inverse gives continuity of the scores and squared errors. Finitely many strict inequalities have a common neighborhood.

For (ii), write  $\Delta = \mathbf{V} - \mathbf{U}$  and  $\mathbf{r}_V = \mathbf{T} - \mathbf{V}_S \mathbf{w}_V$ . The normal equations give the exact identity

$$\mathbf{B}_U(\mathbf{w}_V - \mathbf{w}_U) = \Delta_S^\top \mathbf{r}_V - \mathbf{U}_S^\top \Delta_S \mathbf{w}_V - (q_V - q_U) \mathbf{w}_V. \quad (93)$$

The singular-value filters give  $\|\mathbf{r}_V\|_{\text{op}} \leq t$ ,  $\|\mathbf{w}_V\|_{\text{op}} \leq t/(2\sqrt{q_V})$ ,  $\|\mathbf{w}_V\|_F \leq N/(2\sqrt{q_V})$ ,  $\|\mathbf{B}_U^{-1/2}\|_{\text{op}} \leq q_U^{-1/2}$  and  $\|\mathbf{B}_U^{-1/2} \mathbf{U}_S^\top\|_{\text{op}} \leq 1$ . Multiply (93) by  $\mathbf{B}_U^{-1/2}$  and use  $\mathbf{Y}_V - \mathbf{Y}_U = \Delta_D \mathbf{w}_V + \mathbf{U}_D(\mathbf{w}_V - \mathbf{w}_U)$  to obtain (91). The reverse triangle inequality and  $q_V - q_U = \gamma(2\langle \mathbf{U}_S, \Delta_S \rangle + \|\Delta_S\|_F^2)$  prove (92).  $\square$

**Proposition E.39** (Early complete-refit failure for the frozen-head reference). *For all sufficiently small  $\epsilon > 0$ , every prescribed arrival  $0 \leq j \leq J$  with  $\rho_j \geq 9/10$ , and every held-out input  $x = (a, b)$ , let  $\hat{\mathbf{Y}}_j$  be the prediction of the complete train-only refit. Its true-class score is negative, the named rival  $2a + 1$  exceeds it by more than*

$$\frac{1}{2(d + 1/32)},$$

*and its normalized held-out error is greater than one. In particular, its accuracy is zero, including on diagonal inputs.*

*Proof.* We compute the complete fit of the limiting helper table explicitly, then use its strict wrong-class margins and error reserve to transfer the conclusion to the early arrivals. Put  $\mathbf{H}^\circ = [0 \mid \mathbf{J}_{\text{row}} \mid 0]$  and

write  $\mathbf{J}_S = \mathbf{J}_{\text{row},S} = \mathbf{J}_{\text{row}}|_S$ . Each first operand occurs in exactly  $d$  training rows, so  $\mathbf{J}_S^\top \mathbf{J}_S = d\mathbb{I}_p$  and  $\|\mathbf{J}_S\|_F^2 = n = pd$ . Thus  $\mathbf{H}^\circ$  has training penalty  $1/32$ , and its complete refit prediction is

$$\mathbf{Y}^\circ(a, b) = \frac{1}{d + 1/32} \sum_{q \in Q} T_{2a+q}. \quad (94)$$

For  $x = (a, b) \in \mathcal{D}$ , the difference  $b - a$  is either zero or a nonsquare, so the true label  $y^*(x) = 2a + (b - a)$  is absent from  $2a + Q$ . Hence

$$\mathbf{Y}^\circ(x)_{y^*(x)} = -\frac{d}{p(d + 1/32)} < 0.$$

Since  $1 \in Q$ , the class  $2a + 1$  occurs once in the sum and its score exceeds the true score by  $1/(d + 1/32)$ .

Every ridge prediction is centered because the training targets are centered. Expanding the squared error and using the negative true-class score from (94) gives

$$\frac{\|\mathbf{Y}^\circ(x) - T_{y^*(x)}\|_2^2}{(p-1)/p} = 1 + \frac{\|\mathbf{Y}^\circ(x)\|_2^2 - 2\mathbf{Y}^\circ(x)_{y^*(x)}}{(p-1)/p} > 1. \quad (95)$$

By (82), the full-column tables  $\mathbf{H}_j^{\text{ana}}/(\rho_j L_{\text{lift}})$  converge uniformly to  $\mathbf{H}^\circ$  over the indicated actual arrivals. The limiting penalty is positive, so Lemma E.38(i) retains the strict score inequalities and the conclusion from (95) that normalized error exceeds one, with the rival reserve halved. Positive lift and zero-column invariance transfer the complete refit to the physical network at every admitted width.  $\square$

### E.3.5 The complete refit at late arrivals

This subsection proves the static QR-anchor estimate and transfers it to every late arrival for each prime  $p \geq 1764$ . We use the uniform normalized-table limit of Lemma E.35; no quantitative transport estimate is needed. The Fourier lemmas below serve only this subsection. The complete fit always uses every feature column and the prescribed ridge penalty. The proof has three stages. Lemma E.42 reduces the anchor error to two representative Fourier problems through (98). The binary comparison and sparse teacher then control the complete fit in Proposition E.51. Finally, Proposition E.52 converts this error bound into margins and uses (83) to reach every late arrival.

**The late-refit anchor and its symmetries.** The actual-arrival table limits are the discrete pair (82)–(83); no continuous- $\rho$  table is defined.

**Definition E.40** (Normalized late-refit tables). Let  $Q^c = \mathbb{F}_p \setminus Q$  and fix a nonsquare  $\nu$ . Write  $x = (a, a + t)$  and  $\mathbf{F}_{s,t} = p^{-1/2}e(-st)$ . With  $h_\circ$  from Definition E.33, put

$$\widetilde{\mathbf{H}} = \frac{\mathbf{H}^{\text{QR}}}{h_\circ}, \quad \widetilde{\mathbf{B}} = \sqrt{\varpi} \mathbf{D}_{\text{grid}}, \quad \widetilde{\mathbf{E}} = \widetilde{\mathbf{H}} - \widetilde{\mathbf{B}}, \quad \ell = \frac{\|\widetilde{\mathbf{H}}_S\|_F^2}{32n}. \quad (96)$$

**Definition E.41** (Translation blocks and complete ridge fit). For common translations,  $r \cdot (l_1, l_2, l_3) = (l_1, l_2 - l_1 r, l_3 + 2r)$  and  $\Lambda_0 = \{l : l_3 = 0\}$ . A table is equivariant if  $\mathbf{Y}(x + (r, r), r \cdot l) = \mathbf{Y}(x, l)$ . The target-aligned Fourier blocks are

$$\mathbf{Y}_k(t, l) = \sum_{a \in \mathbb{F}_p} e(-k(2a + t)) \mathbf{Y}((a, a + t), l) \quad (k, t \in \mathbb{F}_p, l \in \Lambda_0). \quad (97)$$

Let  $\mathbf{H}_k, \mathbf{B}_k, \mathbf{E}_k = \mathbf{H}_k - \mathbf{B}_k$  be the blocks of  $\widetilde{\mathbf{H}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{E}}$ . For  $k \neq 0$  put  $\mathbf{A}_k = \mathbf{H}_k[Q]$ ,  $\mathbf{K}_k = \mathbf{A}_k^* \mathbf{A}_k + \ell \mathbb{I}$ ,  $\hat{c}_k = \mathbf{K}_k^{-1} \mathbf{A}_k^* \mathbf{1}_Q$  and  $g_k = \mathbf{H}_k \hat{c}_k \in \mathbb{C}^p$ . Let  $\mathbf{Y}^{\text{QR}}$  be the complete-ridge prediction of  $\mathbf{H}^{\text{QR}}$  and  $E^{\text{QR}}$  its held-out error (9); common rescaling means  $\mathbf{Y}^{\text{QR}}$  is also the prediction of  $\widetilde{\mathbf{H}}$ .

Common feature rescaling leaves complete-ridge predictions unchanged, so we work with  $\widetilde{\mathbf{H}}$  and its actual penalty  $\ell > 1/129$ . Common translation maps each mask to  $r \cdot l$ ; its action is free and  $\Lambda_0$  meets each orbit

once. Since the split is translation invariant, the source and its Polar factor are equivariant, hence so are  $\tilde{\mathbf{H}}, \tilde{\mathbf{B}}, \tilde{\mathbf{E}}$ . Positive ridge penalties give the unique minimizers specified by their normal equations.

For the alias reduction, let  $Q_-$  be the nonsquares and define

$$\mathcal{U}f(k, t) = p^{-1/2} \sum_a e(-k(2a+t))f(a, a+t), \quad \mathfrak{F}_{y,k} = p^{-1/2}e(ky).$$

For  $\sigma \in Q$  put  $d_\sigma(l_1, l_2, l_3) = (l_1/\sigma, l_2, \sigma l_3)$ .

**Lemma E.42** (Reduction to two alias problems). *Let  $p \geq 1764$  be prime. For the complete ridge fit of  $\mathbf{H}^{\text{QR}}$ , let  $g_k$  be its prediction blocks from (97).*

(i) *The map  $\mathcal{U}$  is unitary and sends training and test supports onto  $\bigoplus_k \mathbb{C}^Q$  and  $\bigoplus_k \mathbb{C}^{Q^c}$ , respectively. For every equivariant  $\mathbf{Y}$ ,*

$$\mathcal{U}\mathbf{Y}\mathbf{Y}^*\mathcal{U}^* = \bigoplus_k \mathbf{Y}_k \mathbf{Y}_k^*, \quad \|\mathbf{Y}\|_F^2 = \sum_k \|\mathbf{Y}_k\|_F^2.$$

(ii) *For every input  $x = (a, a+t)$ ,*

$$\left\| \mathbf{Y}^{\text{QR}}(x) - T_{y^*(x)} \right\|^2 = p^{-1} \sum_{k \in \mathbb{F}_p^\times} |1 - g_k(t)|^2.$$

(iii) *For every  $\sigma \in Q$  and  $k \in \mathbb{F}_p^\times$ :  $g_{\sigma k}(t) = g_k(\sigma t)$  and  $\|\mathbf{E}_{\sigma k}\|_F = \|\mathbf{E}_k\|_F$ .*

(iv) *The anchor error is*

$$(p+1)E^{\text{QR}} = \sum_{k \in \{1, \nu\}} \|\mathbf{1}_{Q^c} - g_k[Q^c]\|^2, \quad (98)$$

*and every held-out input  $x$  whose difference is a nonsquare satisfies  $\left\| \mathbf{Y}^{\text{QR}}(x) - T_{y^*(x)} \right\|^2 \leq (p+1)E^{\text{QR}}/p$ .*

*Proof.* Common translations first separate the complete ridge problem into Fourier blocks. Square dilations then identify all nonzero frequencies with the two representatives 1 and  $\nu$ , which yields both the total error formula and the pointwise nonsquare bound. Since 2 is invertible,  $\mathcal{U}$  is a Fourier transform followed by unit phases. Equivariance gives, for  $l \in \Lambda_0$ ,

$$(\mathcal{U}\mathbf{Y})((k, t), r \cdot l) = p^{-1/2}e(-2kr)\mathbf{Y}_k(t, l).$$

The characters  $p^{-1/2}e(2kr)$  are orthonormal on each free column orbit. This proves (i) and supplies a unitary full-column decomposition of ridge, with the same penalty  $\mathbb{I}$  in every block.

The centered class- $k$  target has Fourier coefficient  $p^{-1/2}e(k(2a+t))$  for  $k \neq 0$  and zero for  $k = 0$ . Under  $\mathcal{U}$  it is  $\mathbf{1}$  in block  $k$  and zero in the others. The block normal equations are therefore precisely those for  $\hat{c}_k$ . Inverting  $\mathcal{U}$  gives

$$(\mathbf{Y}^{\text{QR}}(x)^* \mathfrak{F})_k = p^{-1/2}e(k(2a+t))g_k(t),$$

so class Parseval proves (ii).

For  $\sigma \in Q$ , dilation permutes training inputs, masks by  $d_\sigma$ , and target classes. It intertwines the source and Polar factor and preserves the actual penalty  $\ell$ . Uniqueness of full ridge thus gives the stated identity for  $g_{\sigma k}$ . The block definition also gives  $\mathbf{E}_{\sigma k}(t, l) = \mathbf{E}_k(\sigma t, d_\sigma l)$  by the substitution  $a \mapsto \sigma a$ . These row and column permutations preserve Frobenius norm, proving (iii).

Writing each nonzero frequency uniquely as  $\sigma k$  gives

$$p \left\| \mathbf{Y}^{\text{QR}}(x) - T_{y^*(x)} \right\|^2 = \sum_{k \in \{1, \nu\}} \sum_{\sigma \in Q} |1 - g_k(\sigma t)|^2. \quad (99)$$

Every  $\sigma$  permutes  $Q^c$ , including 0. Summing over all held-out inputs and using  $|\mathcal{D}|(p-1)/p = (p+1)|Q|$  gives (98). For a nonsquare  $t$ ,  $\sigma t$  runs once through  $Q_-$ . Thus (99) is bounded by the two full  $Q^c$  norms in (98), proving the pointwise assertion in (iv).  $\square$

### The binary alias geometry.

**Definition E.43** (Binary alias spectrum). Write  $a_v = |\hat{\mathbf{1}}_I(v)|^2$  (so  $a_0 = b^2$ ),  $Q_4 = \sum_{v \neq 0} a_v^2$ , and, for  $q \neq 0$ ,  $A_q(v) = \hat{\mathbf{1}}_I(v/q)$  and  $B_q(v) = p\mathbf{1}[v = 0] - \overline{A_q(v)}$ . Set  $a'_0 = c^2$  and  $a'_v = a_v$  for  $v \neq 0$ . Define  $\lambda_k(s) = \varpi \sum_{q \neq 0} a_{(k-s)/q} a'_{(k+s)/q}$ ,  $\lambda_+ = (p-1)^2(p+1)/(16p^2)$  and  $\lambda_- = (p+1)^3/(16p^2)$ . These are the Gram eigenvalues of Lemma E.44. Put  $\lambda_0 = (p+1)(p^2+3)/(48p^2)$ .

Thus  $A_q(0) = b$ ,  $B_q(0) = c$  and  $|A_q(v)|^2 = |B_q(v)|^2 = a_{v/q}$  for  $v \neq 0$ . For  $l = (q, l_2, l_3)$ , translation of the masks gives  $\hat{\mathbf{1}}_{A_l}(v) = e(vl_2/q)A_q(v)$  and  $\hat{\mathbf{1}}_{B_l}(v) = e(-v(l_3 + l_2/q))B_q(v)$ .

**Lemma E.44** (Binary alias geometry). *Let  $p \geq 1764$  be prime and  $k \neq 0$ . Then*

$$\mathbf{B}_k \mathbf{B}_k^* = \mathbf{F}^* \text{diag}(\lambda_k) \mathbf{F}, \quad \lambda_k(s) > 0. \quad (100)$$

Moreover  $\lambda_k(s) = \lambda_1(s/k)$ ,

$$\lambda_k(0) = \varpi Q_4 = \lambda_0, \quad \lambda_k(k) = \lambda_+, \quad \lambda_k(-k) = \lambda_-, \quad \sum_s \lambda_k(s) = \frac{p+1}{6}.$$

*Proof.* The block definition and the mask transforms give, for  $l = (q, l_2, 0)$ ,

$$(\mathbf{F} \mathbf{B}_k)(s, l) = \sqrt{\varpi/p} A_q(k-s) B_q(k+s) e(-2sl_2/q). \quad (101)$$

The phases in (101) make distinct Fourier rows orthogonal after summing over  $l_2$ ; their squared norms are exactly  $\lambda_k(s)$ , proving (100). The geometric-series formula for  $\hat{\mathbf{1}}_I(v)$  is nonzero at every  $v$ : for  $v \neq 0$  both  $1 - e(-bv)$  and  $1 - e(-v)$  are nonzero. This proves positivity. Substitute  $q = kr$  for scaling. Parseval and Lemma E.12(iv) give  $\sum_{v \neq 0} a_v = u$  and  $Q_4 = \Sigma_2 = p^2 \text{Var}(h) = u(u+1)/3$ , hence all three special eigenvalues. To obtain the trace, substitute  $n = k/q \neq 0$ ,  $w = s/q$  and then  $(i, j) = (n - w, n + w)$ . This is a bijection onto  $i + j \neq 0$ , so

$$\sum_s \lambda_k(s) = \varpi \left[ (pb)(pc) - \sum_i a_i a'_{-i} \right] = \varpi [p^2 u - u^2 - Q_4] = \frac{p+1}{6}.$$

□

### Source-defect and penalty bounds.

**Definition E.45** (Defect and penalty envelopes). Put

$$\begin{aligned} \kappa_p &= \sqrt{\frac{p}{p-1}}, \quad \Delta_p^2 = \frac{4p}{(p-1)h_\circ^2}, \\ \ell_{\text{bin}} &= \frac{p^2-1}{128p^2}, \quad \ell_+ = \left(1 + \frac{2p}{(p-1)\sqrt{p+1}}\right)^2 \ell_{\text{bin}}. \end{aligned} \quad (102)$$

**Lemma E.46** (Defect and penalty). *Let  $p \geq 1764$  be prime.*

- (i)  $\mathbf{D}_{\text{grid}} \odot (\mathbf{X}_{\text{grid}} \bar{\mathbf{Q}}^\top) = h_\circ \tilde{\mathbf{B}}$ ; that is, every row of  $\bar{\mathbf{Q}}$  has the constant value  $h_\circ \sqrt{\varpi}$  on its active rectangle.
- (ii)  $h_\circ > 2$  and  $\|\tilde{\mathbf{E}}\|_F^2 < 2p/h_\circ^2 < p/2$ .
- (iii)  $\|\mathbf{E}_1\|_F^2 + \|\mathbf{E}_\nu\|_F^2 < \Delta_p^2 < \kappa_p^2$ . In particular  $\|\mathbf{E}_k\|_{\text{op}} < \kappa_p$  for every  $k \neq 0$ .
- (iv)  $\ell_{\text{bin}} \leq \ell \leq \ell_+$ .

*Proof.* Parts (i)–(iii) turn the earlier Polar comparison into a bound on the two alias defects. Part (iv) controls the penalty computed from the actual QR table, so the subsequent ridge estimates use that table’s prescribed normalization. (i) Since  $h_o = 2\sqrt{p/(p+1)} + \sqrt{2/p} = \bar{\alpha}^{-1/2} + \sqrt{2/p}$ , this is exactly the active-rectangle value proved in Lemma E.34(i).

(ii) The inequalities  $2\sqrt{p/(p+1)} > 2 - 2/(p+1)$  and  $\sqrt{2/p} > 2/(p+1)$  give  $h_o > 2$ . By (i),  $h_o \tilde{\mathbf{E}} = \mathbf{D}_{\text{grid}} \odot (\mathbf{X}_{\text{grid}}(\mathbf{Q}_{\text{QR}} - \tilde{\mathbf{Q}})^\top)$ . For each incoming row  $(w, v)$ , weighted Cauchy–Schwarz gives

$$|w(a) + v(z)|^2 \leq \frac{p}{c}|w(a)|^2 + \frac{p}{b}|v(z)|^2, \quad \sum_{a \in A_l, z \in B_l} |w(a) + v(z)|^2 \leq p \|(w, v)\|^2.$$

Summing over rows and using Lemma E.34(iii) yields  $\|\tilde{\mathbf{E}}\|_F^2 \leq p\delta_Q^2/h_o^2 < 2p/h_o^2 < p/2$ .

(iii) Lemma E.42(i),(iii) gives

$$d(\|\mathbf{E}_1\|_F^2 + \|\mathbf{E}_\nu\|_F^2) \leq \sum_{k \in \mathbb{F}_p} \|\mathbf{E}_k\|_F^2 = \|\tilde{\mathbf{E}}\|_F^2 < 2p/h_o^2.$$

This is the first claim; the second follows from  $h_o > 2$  and  $\|\mathbf{E}_k\|_{\text{op}} \leq \|\mathbf{E}_k\|_F$ .

(iv) Write  $\|\mathbf{M}\|_*$  for the sum of the singular values of  $\mathbf{M}$ . Let  $\mathbf{M} = \tilde{\mathbf{B}}_S^\top \mathbf{X}_S$ , a positive multiple of  $\mathbf{F}_{\text{unif}}$ , so that  $\mathbf{Q}_{\text{QR}} = \text{pol}(\mathbf{M})$  and  $\langle \mathbf{Q}_{\text{QR}}, \mathbf{M} \rangle_F = \|\mathbf{M}\|_*$ . Since  $\mathbf{D}_{\text{grid}}$  is binary,

$$\langle \tilde{\mathbf{H}}_S, \tilde{\mathbf{B}}_S \rangle_F = h_o^{-1} \langle \mathbf{Q}_{\text{QR}}, \mathbf{M} \rangle_F = h_o^{-1} \|\mathbf{M}\|_* \geq h_o^{-1} \langle \tilde{\mathbf{Q}}, \mathbf{M} \rangle_F = \|\tilde{\mathbf{B}}_S\|_F^2.$$

The inequality holds because  $\|\tilde{\mathbf{Q}}\|_{\text{op}} \leq 1$ ; the last equality is (i). Expanding  $\tilde{\mathbf{H}}_S = \tilde{\mathbf{B}}_S + \tilde{\mathbf{E}}_S$  gives  $\|\tilde{\mathbf{H}}_S\|_F^2 \geq \|\tilde{\mathbf{B}}_S\|_F^2 + \|\tilde{\mathbf{E}}_S\|_F^2$ . For fixed  $(l_1, l_2)$ , a pair  $(a, a') \in A_l \times A_l^c$  yields the input  $(a, l_3 - a') \in A_l \times B_l$ , and this input is a training input for exactly  $d$  values of  $l_3$ . Hence  $\|\tilde{\mathbf{B}}_S\|_F^2 = \varpi p(p-1)d^2(d+1) = d^2(d+1)/p = 32n\ell_{\text{bin}}$ , and  $\ell \geq \ell_{\text{bin}}$ . For the upper bound,  $\|\tilde{\mathbf{H}}_S\|_F \leq \|\tilde{\mathbf{B}}_S\|_F + \|\tilde{\mathbf{E}}\|_F < \|\tilde{\mathbf{B}}_S\|_F + \sqrt{p/2}$  by (ii). Dividing by  $\|\tilde{\mathbf{B}}_S\|_F = \sqrt{d^2(d+1)/p}$  gives  $\sqrt{\ell/\ell_{\text{bin}}} < 1 + 2p/((p-1)\sqrt{p+1})$ , which is precisely the stated  $\ell \leq \ell_+$ .  $\square$

**Definition E.47** (Adjacent interval correlation). Set  $C_1 = \sum_{v \notin \{0, -1\}} a_v a_{v+1}$  and  $t_I = C_1/a_1$ .

**Lemma E.48** (Adjacent interval bound). *For every odd prime  $p \geq 1764$ ,*

$$a_1 > p^2/\pi^2, \quad a_2 < 1, \quad C_1/a_1 < 5/2.$$

*Proof.* For canonical representatives  $|v| \leq b$ , the geometric sum gives

$$a_v = \frac{1}{4 \cos^2(\pi v/(2p))} \leq \frac{1}{2} \quad (v \neq 0 \text{ even}), \quad a_v = \frac{1}{4 \sin^2(\pi v/(2p))} \quad (v \text{ odd}).$$

Thus  $a_1 > p^2/\pi^2$  and  $a_2 < 1$ . Every non-wrap neighbor pair in  $C_1$  has opposite parity. Each odd coefficient meets at most two even neighbors, whose total is at most one. The sole wrap pair has product  $a_b a_{-b} = a_b^2 < 1$ : if  $b$  is even use the displayed bound, while if  $b$  is odd then  $\pi b/(2p) > \pi/6$ . Parseval gives  $\sum_{v \neq 0} a_v = u$ , so  $C_1 \leq u + 1$ . Consequently

$$\frac{C_1}{a_1} < \frac{\pi^2}{4}(1 + 3/p^2) < \frac{121}{49}(1 + 3/1764^2) < \frac{5}{2},$$

using  $\pi < 22/7$ .  $\square$

**A two-slope comparison teacher.** The complete fit still uses every column. Only its comparison coefficient is supported on the two slopes  $q = \pm k$ , so the adjacent interval correlation controls its target error.

**Definition E.49** (Sparse comparison teacher and correction). For  $k \neq 0$  put  $\mathbf{G}_k = \mathbf{B}_k \mathbf{B}_k^*$  and  $w_k = \mathbf{B}_k^* \mathbf{1} / \lambda_0$ . The sparse comparison coefficient is

$$w'_k(q, l_2, 0) = -\frac{\mathbf{1}[q \in \{k, -k\}]}{2\sqrt{\varpi}a_1}.$$

Set  $\zeta_k = w'_k - w_k$  and  $f'_k = \mathbf{H}_k w'_k$ . For its regularized binary correction put  $z'_k = \mathbf{B}_k^* (\mathbf{G}_k + \ell \mathbb{I})^{-1} (f'_k - \mathbf{1})$  and  $r'_k = \ell (\mathbf{G}_k + \ell \mathbb{I})^{-1} (f'_k - \mathbf{1})$ . Only the comparison coefficient is sparse;  $\hat{c}_k$  still minimizes over every column.

**Lemma E.50** (Sparse regularized target energy). *For every prime  $p \geq 1764$  and  $k \neq 0$ , the sparse comparison coefficient satisfies*

$$\mathbf{B}_k w'_k = \mathbf{1}, \quad \zeta_k = w'_k - w_k \in \ker \mathbf{B}_k \perp \text{range}(\mathbf{B}_k^*), \quad \|\zeta_k\|^2 < \frac{4}{5}.$$

Its actual QR teacher  $f'_k = \mathbf{H}_k w'_k$  obeys

$$(f'_k - \mathbf{1})^* (\mathbf{G}_k + \frac{1}{129} \mathbb{I})^{-1} (f'_k - \mathbf{1}) < \frac{200}{p}.$$

*Proof.* First we check the binary interpolation and coefficient-norm claims. We then contract the QR source against this coefficient and bound its nonzero and zero Fourier modes separately. Their weighted Parseval sum gives the regularized target-energy estimate. For a slope-indexed expression write  $\langle F_q \rangle_k = [F_k + F_{-k}] / (2a_1)$ . Summing the Fourier rows (101) against  $w'_k$  over the  $p$  shifts  $l_2$  gives

$$(\mathbf{F} \mathbf{B}_k w'_k)(s) = -\sqrt{p} \mathbf{1}[s = 0] \langle A_q(k) B_q(k) \rangle_k = \sqrt{p} \mathbf{1}[s = 0].$$

Here  $A_q(k) B_q(k) = -a_1$  for  $q = \pm k$ . Thus  $\mathbf{B}_k w'_k = \mathbf{1}$ . Since  $w_k = \mathbf{B}_k^* \mathbf{1} / \lambda_0$  is the minimum-norm binary solution, the asserted nullspace orthogonality follows. The  $2p$  nonzero entries of  $w'_k$ , Lemma E.48, and  $\pi < 22/7$  give

$$\|\zeta_k\|^2 = \frac{p}{2\varpi a_1^2} - \frac{p}{\lambda_0} < \frac{1880}{2401} (1 - p^{-1}) + \frac{96}{p^2} < \frac{4}{5}.$$

Here the exact difference  $48(1 - p^{-1}) - p/\lambda_0 = 48(2p^2 - 3)/[p(p+1)(p^2+3)]$  is less than  $96/p^2$ . Uniformly,  $1880/2401 + 96/1764^2 < 4/5$ .

*Source contraction.* Lift the comparison coefficient by  $W'_k(q, l_2, l_3) = e(kl_3) w'_k(q, l_2, 0)$ . For this proof put  $x = \sqrt{p}$ ,  $\alpha = (x - 1)^{-1}$ ,  $\beta_* = f_p/h_0$  and  $\gamma = (g_0 \sqrt{2c_0})^{-1}$ . For  $h \neq 0$  put  $\mathbf{R}_h = \mathbf{G}_{\text{avg}} + \chi_2(h) \mathbf{G}_{\text{diff}}$  and  $\Delta_h = \mathbf{R}_h/g_0 - \mathbb{I}$ , with  $\Delta_0 = 0$ . For this contraction, write  $\mathbf{R}_0 = \mathbf{G}_{\text{const}}$  for the constant inverse-root block and  $\mathbf{n} = npb_l$  for the QR count row, with operand blocks  $\mathbf{n}^L, \mathbf{n}^R$ . Set

$$\eta_h = \frac{-1 + \tau \chi_2(h) + p \mathbf{1}[h = 0]}{p - 1}.$$

The lifted coefficient satisfies  $W'_k((-a) \cdot l) = e(-2ka) W'_k(l)$ , so equivariance gives

$$f'_k(t) = e(-kt) \sum_l \widetilde{\mathbf{H}}((0, t), l) W'_k(l).$$

Insert the source rows (42) and the count convolutions

$$\hat{\mathbf{n}}^L(h) = \frac{b}{p} \sum_i \eta_{-i} \hat{\mathbf{1}}_{A_i}(h - i) \hat{\mathbf{1}}_{B_i}(i), \quad \hat{\mathbf{n}}^R(h) = \frac{b}{p} \sum_i \eta_i \hat{\mathbf{1}}_{B_i}(h - i) \hat{\mathbf{1}}_{A_i}(i).$$

For fixed slope  $q$ , a product with total left and right frequencies  $I_L, I_R$  has shift phase  $e(l_2(I_L - I_R)/q) e((k - I_R)l_3)$ . Summation over the two shifts forces  $I_L = I_R = k$ . Define the two inner convolutions by

$$\begin{aligned} \Sigma_q^L(s) &= \sum_i A_q(k - i) A_q(i) [\eta_s(\mathbf{R}_{i-s}/g_0)_{11} + \eta_i(\mathbf{R}_{i-s}/g_0)_{12}], \\ \Sigma_q^R(s) &= \sum_i B_q(k - i) B_q(i) [\eta_{-i}(\mathbf{R}_{i+s}/g_0)_{21} + \eta_s(\mathbf{R}_{i+s}/g_0)_{22}]. \end{aligned} \tag{103}$$

Then, with  $\widehat{f}'_k(s) = \sum_t e(-st) f'_k(t)$ , the contraction is

$$\widehat{f}'_k(s) = -\frac{\beta_* b}{p} \langle B_q(k+s) B_q(-s) \Sigma_q^L(s) + A_q(k-s) A_q(s) \Sigma_q^R(s) \rangle_k. \quad (104)$$

Indeed the source row and Fourier inversions contribute  $g_0/(h_\circ \sqrt{p}) = \beta_* b/p$  after the two shift sums. We estimate these contractions in norm without evaluating each teacher coefficient.

*Nonzero modes: a whole-convolution operator bound.* For  $s \neq 0$  and  $q = \pm k$ , first bound the inner convolutions (103); then apply the outer factors in (104) in the weighted Fourier norm. Put  $d_0 = 3/p + 15/(2x^3)$ ,  $d_1 = 1/x + d_0$ ,  $D = \alpha(d_0 + d_1)$  and

$$j_{\text{kin}} = p\alpha + c(d_1 + \alpha d_0 + D + \alpha) + D\sqrt{(p-2)t_I}.$$

Lemma E.18(ii)–(iii) gives  $|(\Delta_h)_{jj}| \leq d_0$  and  $|(\Delta_h)_{12}|, |(\Delta_h)_{21}| \leq d_1$  for  $h \neq 0$ . Since  $|\eta_h| \leq \alpha$  for  $h \neq 0$ , the root-error response is at most  $d_1 + \alpha d_0$  at  $i = 0$ , and at most  $D$  otherwise. The identity part uses  $\sum_i A_q(k-i) A_q(i) = p A_q(k)$  and  $\sum_i B_q(k-i) B_q(i) = p B_q(k)$ . The products at  $i = 0, k$  have moduli  $b\sqrt{a_1}$  on the left and  $c\sqrt{a_1}$  on the right; the remaining squared sum is  $C_1$ . Cauchy–Schwarz therefore bounds each inner sum, except its source-frequency-zero jump, by

$$[p\alpha + c(d_1 + \alpha d_0 + D) + D\sqrt{(p-2)t_I}] \sqrt{a_1}.$$

The actual constant block is  $\mathbf{R}_0/g_0 = (\gamma/2)\mathbf{1}\mathbf{1}^\top$ . At  $i = s$  on the left or  $i = -s$  on the right, its response differs from the identity response by  $(\gamma - 1)\eta_s$ . Here  $0 < \gamma < 1$ : Lemma E.15 and the formula for  $g_0$  give  $2c_0 g_0^2 \geq 2(p-1)^2(p+1)/(p^2+1) > 1$ . The multiplying mask product is at most  $c\sqrt{a_1}$ : at an endpoint this is immediate, and otherwise its squared modulus is at most  $C_1 = t_I a_1 < c^2 a_1$ . Thus both whole inner sums have modulus at most  $j_{\text{kin}} \sqrt{a_1}$ . For  $x \geq 42$  the decreasing scalar expressions satisfy

$$pD = \frac{x+6+15/x}{x-1} < \frac{6}{5}, \quad (x+2)d_0 < \frac{1}{10}, \quad \frac{x}{x-1} < \frac{21}{20}, \quad \frac{c\alpha}{x} < \frac{11}{20}.$$

Also  $pD < 6/5$  and  $t_I < 5/2$  give  $D\sqrt{(p-2)t_I}/x < (6/(5p))\sqrt{5/2} < 48/(25p) < 1/100$ . Hence

$$\begin{aligned} \frac{j_{\text{kin}}}{x} &= \frac{x}{x-1} + \frac{c\alpha}{x} [2 + (x+2)d_0] + \frac{D\sqrt{(p-2)t_I}}{x} \\ &< \frac{21}{20} + \frac{11}{20} \frac{21}{10} + \frac{1}{100} = \frac{443}{200} < \frac{9}{4}. \end{aligned} \quad (105)$$

Now apply the outer multiplication operator in the weighted norm  $\|v\|_\omega^2 = \sum_{s \neq 0} |v(s)|^2 / (\lambda_k(s) + 1/129)$ . At  $s = k, -k$ , its direct factors have moduli  $b\sqrt{a_1}, c\sqrt{a_1}$  and its cross factors are at most  $\sqrt{a_1 a_2} < \sqrt{a_1}$ . Off these axes each outer factor has squared sum at most  $C_1$ . Minkowski for the two slopes and the two operand blocks, followed by (105), gives

$$\begin{aligned} \|\widehat{f}'_k\|_\omega &\leq \frac{\beta_* b}{p} j_{\text{kin}} \sqrt{\frac{(b+1)^2}{\lambda_+} + \frac{(c+1)^2}{\lambda_-} + 516t_I} \\ &< \frac{6j_{\text{kin}}}{x} < \frac{27}{2} < 14. \end{aligned} \quad (106)$$

For the strict middle bound, the axis terms divided by  $p$  sum to

$$\begin{aligned} \frac{4p(p+1)}{(p-1)^2} + \frac{4p(p+3)^2}{(p+1)^3} &= 8 + \frac{12}{p-1} + \frac{8}{(p-1)^2} + \frac{12}{p+1} - \frac{16}{(p+1)^3} \\ &< 8 + \frac{24}{1763} + \frac{8}{1763^2} < \frac{81}{10}, \end{aligned}$$

while  $516t_I/p < 1290/1764 < 4/5$ , and  $\beta_* < 4/(p-1)$ .

*Zero frequency.* Put  $a_* = [g_0(b^2 + c^2) + 2u/\sqrt{2c_0}]/(h_0 p^{3/2})$ . For  $h \neq 0$  the two normalized row factors  $(\mathbf{R}_h)_{11}/g_0 + \eta_h(\mathbf{R}_h)_{12}/g_0$  and  $(\mathbf{R}_h)_{22}/g_0 + \eta_{-h}(\mathbf{R}_h)_{21}/g_0$  differ from one by at most  $d_0 + \alpha d_1 < 5/p$ , since

$$p(d_0 + \alpha d_1) = 4 + \frac{15}{2x} + \frac{4 + 15/(2x)}{x - 1} \leq 4 + \frac{5}{28} + \frac{4 + 5/28}{41} < 5.$$

The absolute sparse left and right convolutions are bounded by  $b + \sqrt{(p-2)t_I}$  and  $c + \sqrt{(p-2)t_I}$ , both below  $p$  because  $(p-1)^2 > 10(p-2)$ . Applying the contraction (104) at  $s = 0$  with these bounds gives

$$|p^{-1}\widehat{f'_k}(0) - a_*| < \frac{g_0}{h_0 p^{3/2}} \frac{5}{p} (cp + bp) < \frac{10}{p}. \quad (107)$$

The exact source covariance expresses  $a_*$  as

$$a_* = \frac{2\sqrt{p/(p+1)}}{h_0} \sqrt{1 + p^{-2}} + \frac{\sqrt{2/p}}{h_0} (1 + \varepsilon_p)^{-1/2}.$$

The coefficients sum to one, and  $0 < \varepsilon_p < 1/[2(p-1)]$ . Both endpoints differ from one by less than  $1/p$ , so  $|p^{-1}\widehat{f'_k}(0) - 1| < 11/p$  by (107). Since  $\lambda_0 > p/48$ , this mode contributes less than  $5808/p^2$ . By (100), the weighted Parseval sum has weights  $(\lambda_k(s) + 1/129)^{-1}$  and an overall factor  $1/p$  for the unnormalized transform  $\widehat{f'_k}$ . Combining the zero mode with (106) proves the asserted energy bound:

$$(f'_k - \mathbf{1})^*(\mathbf{G}_k + \frac{1}{129}\mathbb{I})^{-1}(f'_k - \mathbf{1}) < \frac{14^2}{p} + \frac{5808}{p^2} < \frac{200}{p}.$$

Since  $\ell > 1/129$ , the regularized correction in Definition E.49 consequently has  $\mathbf{B}_k z'_k + r'_k = f'_k - \mathbf{1}$  and  $\|z'_k\|^2 + \|r'_k\|^2/\ell < 200/p$ .  $\square$

#### A target-sensitive normal-residual frame.

**Proposition E.51** (Complete-error frame). *For every prime  $p \geq 1764$ , with  $x = \sqrt{p}$ ,*

$$\frac{p+1}{p} E^{\text{QR}} < \frac{2}{5x^3} \left[ 129(153/10)^2 + \left( \sqrt{400 + 8x^2/5} + 101 \right)^2 \right] < \frac{3}{10}. \quad (108)$$

*Proof.* We compare the full ridge minimizer with a trial obtained by correcting the sparse comparison coefficient. The proof bounds the trial's residual, derives a sampling estimate for arbitrary coefficient vectors, and controls the difference between the trial and the minimizer through the normal equations. The alias identity (98) then converts these bounds into held-out error. Here  $x = \sqrt{p} \geq 42$ . The scalar definitions and Lemma E.46(iv) give

$$\kappa_p < 1001/1000, \quad 1/129 < \ell \leq \ell_+ < (21/20)^2/128 < 1/116,$$

using  $2p/((p-1)\sqrt{p+1}) < 84/1763 < 1/20$ .

*The sparse target trial.* Set  $z_k = z'_k - \zeta_k$ ,  $c_k = w_k - z_k$  and  $e_k = \mathbf{1} - \mathbf{H}_k c_k$ . Lemma E.50 and  $\ell > 1/129$  give

$$\mathbf{B}_k \zeta_k = 0, \quad z'_k \perp \zeta_k, \quad \|\zeta_k\|^2 < 4/5, \quad \mathbf{B}_k z'_k + r'_k = f'_k - \mathbf{1}, \quad \|z'_k\|^2 + \|r'_k\|^2/\ell < 200/p.$$

Since  $c_k = w'_k - z'_k = w_k - z_k$ , its full-grid residual is  $e_k = \mathbf{E}_k z'_k - r'_k$ . The norm of  $[\mathbf{E}_k, -\sqrt{\ell}\mathbb{I}]$ , the nullspace orthogonality and Lemma E.46(iii) therefore give

$$\mathcal{R} := \sum_k \|e_k\|^2 < \frac{200}{p} (\Delta_p^2 + 2\ell), \quad \mathcal{Z} := \sum_k \|z_k\|^2 < 400/p + 8/5. \quad (109)$$

All paired sums run over  $k \in \{1, \nu\}$ .

*Sampling all Fourier modes.* For fixed  $k$  write  $\mathbf{B} = \mathbf{B}_k$ ,  $\mathbf{H} = \mathbf{H}_k$ ,  $\mathbf{A} = \mathbf{H}[Q]$  and  $\mathbf{K} = \mathbf{A}^* \mathbf{A} + \ell \mathbb{I}$ . Put  $\mathbf{M} = \mathbf{F} \text{diag}(\mathbf{1}_Q) \mathbf{F}^*$ ,  $\alpha = (p-1)/(2p)$ ,  $m = (\sqrt{p} + 1)/(2p)$  and  $\theta = \alpha + m$ . Set  $T = (p+1)/6$  and  $L = \ell + (\sqrt{T/x} + \sqrt{2\kappa_p})^2$ . The matrix  $\mathbf{M}$  has diagonal  $\alpha$  and off-diagonal moduli at most  $m$ . Thus

$\psi^* \mathbf{M} \psi \geq \theta \|\psi\|^2 - m \|\psi\|_1^2$ , where  $\theta = (x+1)/(2x) > 1/2$ . For  $\psi = \mathbf{F} \mathbf{B} \xi$ , Fourier-row orthogonality gives  $\|\psi\|_1^2 \leq T \|\xi\|^2$ . Since  $m/\theta = 1/x$ ,  $\|\mathbf{B}[Q^c] \xi\|^2 \leq \|\mathbf{B}[Q] \xi\|^2 + (T/x) \|\xi\|^2$ . Using  $\mathbf{H} - \mathbf{B} = \mathbf{E}_k$  and its orthogonal training/held-out restrictions gives

$$\begin{aligned} \|\mathbf{H}[Q^c] \xi\| &\leq \|\mathbf{A} \xi\| + (\sqrt{T/x} + \sqrt{2\kappa_p}) \|\xi\| \\ &\leq \sqrt{L/\ell} \|\mathbf{K}^{1/2} \xi\|. \end{aligned} \quad (110)$$

For  $x \geq 42$ ,  $2\ell/x < 1/1000$ ,  $\sqrt{T/p} < 41/100$  and  $\sqrt{2\kappa_p}/\sqrt{x} < 11/50$ . The last inequality follows by squaring and using  $2x/(x^2 - 1) \leq 84/1763 < (11/50)^2$ . Hence

$$(L + \ell)/x < 1/1000 + (41/100 + 11/50)^2 < 2/5. \quad (111)$$

A target-matched normal residual. Put  $s = p/(\lambda_0 d)$  and

$$u_k = \mathbf{B}_k[Q]^* \mathbf{1}_Q - w_k/s, \quad v_k = \mathbf{E}_k[Q]^* \mathbf{1}_Q, \quad \rho_k = \ell s \mathbf{1}_Q, \quad \delta_k = \hat{c}_k - c_k.$$

Over  $k \in \{1, \nu\}$ , put  $U = \sum_k \|u_k\|^2$ ,  $\mathcal{K} = \sum_k \delta_k^* \mathbf{K}_k \delta_k / \ell$  and  $e_{\text{ho}}^2 = \sum_k \|e_k[Q^c]\|^2$ . The subtraction in  $u_k$  removes exactly the constant Fourier component. Since  $|(\mathbf{F} \mathbf{1}_Q)(a)| \leq (1 + 1/x)/2 < 21/40$  for  $a \neq 0$  and  $T < p/5$ ,

$$U = \sum_k \sum_{a \neq 0} \lambda_k(a) |(\mathbf{F} \mathbf{1}_Q)(a)|^2 < 2(21/40)^2 p/5 < p/9.$$

Also  $\sum_k \|v_k\|^2 < d \Delta_p^2$ . The normal equations give

$$\begin{aligned} \mathbf{K}_k \delta_k &= \mathbf{A}_k^* (e_k[Q] - \rho_k) - \ell (c_k - \mathbf{A}_k^* \rho_k / \ell), \\ c_k - \mathbf{A}_k^* \rho_k / \ell &= -(z_k + s(u_k + v_k)). \end{aligned} \quad (112)$$

The operator  $\mathbf{K}_k^{-1/2} [\mathbf{A}_k^*, -\sqrt{\ell} \mathbb{I}]$  is a coisometry. Apply it to (112) jointly over the two aliases. Adding the held-out residual pairs  $e_k[Q]$  with  $e_k[Q^c]$ , so (109) controls their full-grid norm. The triangle inequality and  $1/\ell < 129$  then give

$$\begin{aligned} \mathcal{K} + e_{\text{ho}}^2 / \ell &< 129 \left( \sqrt{\frac{200}{p} (\Delta_p^2 + 2\ell) + \ell s \sqrt{2d}} \right)^2 \\ &\quad + \left( \sqrt{400/p + 8/5} + s\sqrt{U} + s\sqrt{d\Delta_p} \right)^2. \end{aligned} \quad (113)$$

Here  $sp = 96p^4 / [(p^2 - 1)(p^2 + 3)] < 96 < 961/10$ , since  $(p^2 - 1)(p^2 + 3) = p^4 + 2p^2 - 3 > p^4$ . Also  $h_o > 2$  and  $\ell < 1/116$  imply  $\Delta_p^2 + 2\ell < 1 + 1/1763 + 1/58 < (101/100)^2$ , and  $\sqrt{d\Delta_p} < \sqrt{p/2} < (71/100)\sqrt{p}$ . After multiplication by  $x$ , the first parenthesis is below

$$(101/100)(71/5) + (961/10)/116 < 153/10.$$

The second is below  $\sqrt{400 + 8x^2/5} + (961/10)(1/3 + 71/100) < \sqrt{400 + 8x^2/5} + 101$ .

Finally,  $\mathbf{1} - g_k = e_k - \mathbf{H}_k \delta_k$ . Applying (110) to  $\delta_k$  and using the alias identity (98) gives

$$\sqrt{(p+1)E^{\text{QR}}} \leq e_{\text{ho}} + \sqrt{L\mathcal{K}} \leq \sqrt{L + \ell} \sqrt{\mathcal{K} + e_{\text{ho}}^2 / \ell}. \quad (114)$$

Substitute the parenthesis bounds into (113), square (114), and use (111). Dividing by  $p = x^2$  gives the first inequality in (108). The quotient  $[\sqrt{400 + 8x^2/5} + 101]/x = \sqrt{400/x^2 + 8/5} + 101/x$  decreases for  $x > 0$ , and  $400 + 8 \cdot 42^2/5 = 16112/5 < 57^2$ . Thus  $\sqrt{400 + 8x^2/5} + 101 < 158x/42$  for  $x \geq 42$ . After this bound, both terms are decreasing in  $x$ , so the right side is at most

$$\frac{2[129(153/10)^2 + 158^2]}{5 \cdot 42^3} < \frac{3}{10}.$$

□

**Proposition E.52** (Late refit for every prime  $p \geq 1764$ ). *For each fixed prime  $p \geq 1764$  and  $\lambda > 0$ , all sufficiently small  $\epsilon > 0$  have the following property. At every prescribed late arrival  $j \in \mathcal{I}_{\text{late}}$ , the complete train-only refit has margin greater than  $11/50$  against every rival at each held-out input with a nonsquare difference, normalized held-out error less than  $121/400$ , and accuracy at least  $(p-1)/(p+1)$ . The same statements hold for the compressed physical table.*

*Proof.* At the QR anchor, the pointwise bound of Lemma E.42(iv) and the complete-error estimate (108) in Proposition E.51 give

$$\|\mathbf{Y}^{\text{QR}}(x) - T_y\|^2 < 3/10$$

on every nonsquare-difference input. Thus every rival  $c \neq y$  has margin

$$\mathbf{Y}^{\text{QR}}(x)_y - \mathbf{Y}^{\text{QR}}(x)_c > 1 - \sqrt{3/5} > 11/50. \quad (115)$$

The same frame gives  $E^{\text{QR}} < 3/10 < 121/400$ , with diagonals included.

The full-column anchor  $[\mathbf{H}^{\text{QR}} \mid 0 \mid 0]$  has positive prescribed penalty because  $\ell > 1/129$  for  $\mathbf{H}^{\text{QR}}/h_\circ$ . The late table limit (83) of Lemma E.35 makes  $\mathbf{H}_j^{\text{ana}}/R$  converge uniformly to this full-column anchor over the prescribed late arrivals. Lemma E.38(i) therefore retains margins greater than  $11/50$ , using the reserve in (115), and the strict error bound for all sufficiently small  $\epsilon > 0$ . The error uses the full held-out denominator, including diagonal inputs.

Every nonsquare-difference held-out input is thus correctly classified, giving accuracy at least  $pd/[p(d+1)] = (p-1)/(p+1)$ ; no diagonal correctness is asserted. Positive lift and zero-column invariance transfer the complete refit predictions and errors to the physical table.  $\square$

## E.4 Assembly: positive head step and an initial ball at every width

### E.4.1 Finite-horizon assembly at a fixed width

We now transfer the reference conclusions to a neighborhood of actual training runs. The proof checks continuity of the update at each reference departure, then uses the finite horizon to preserve all strict observer and prediction bounds on one initial ball.

**Definition E.53** (Canonical state and discrete reference). At width  $M$ , let  $\theta_M^{\text{can}}$  consist of the physical core rows constructed before Lemma E.30, the helper and query rows in Section E.2.1.4, their heads in (57), (58) and (66), and the dormant rows of Definition E.20. Define the frozen-head arrivals by

$$\theta_{M,0}^\circ = \theta_M^{\text{can}}, \quad \theta_{M,j+1}^\circ = \mathcal{U}_{0,0}(\theta_{M,j}^\circ), \quad 0 \leq j < J. \quad (116)$$

For a radius  $\delta > 0$ , write  $\mathcal{B}_M = \{\theta_0 : \|\theta_0 - \theta_M^{\text{can}}\|_F < \delta\}$ .

**Proposition E.54** (Finite-horizon assembly at every width). *Fix a prime  $p \geq 101$ , a width  $M \geq M_{\min}(p)$  and  $\lambda > 0$ . For every sufficiently small  $\epsilon > 0$  as in Definition E.3, there are  $\delta \in (0, 1/2]$ ,  $\eta_0 > 0$  and  $d_0 \in (0, 1/2]$  such that every  $\theta_0 \in \mathcal{B}_M$  and all  $0 \leq \eta < \eta_0$ ,  $0 \leq d_V < d_0$  satisfy:*

- (a) *The width- $M$  trajectory satisfies Theorem 4.1(i)–(iii) on the prescribed windows. If  $d_V = 0$  and  $\eta > 0$ , its actual head moves at every departure  $j < J$ .*
- (b) *If  $d_V = 0$  and  $p \geq 1764$ , it satisfies part (iv) and  $E_{\text{ref}}(j) < 121/400$  at every  $j \in \mathcal{I}_{\text{late}}$ .*

*The constants may depend on the fixed width  $M$ . In particular,  $\eta = \eta_0/2$ ,  $d_V = 0$  gives the public theorem and its late-refit error supplement, while the same ball and head step serve part (a) with any fixed  $0 < d_V < d_0$ .*

*Proof.* Fix  $p \geq 101$ ,  $\lambda > 0$  and an admitted width  $M$ . Choose and fix a sufficiently small  $\epsilon > 0$  satisfying the component certificates above, including  $\epsilon < 1/200$ . This fixes the rates, windows and finite horizon  $J$  before the initialization is chosen. Use the width- $M$  canonical state and frozen-head reference  $\theta_{M,j}^\circ$  of Definition E.53.

**Reference gates, rank and observers.** Lemma E.30 supplies the strict core signs and full supported source rank at every reference departure. The noncore rows have zero incoming sources: helpers have zero frozen heads, queries are inactive on training, and dormant rows are negative. They decay by  $q_{\text{clk}}$  and retain their strict training signs. Their zero source rows append zeros to the core source, preserving its rank and supported Polar factor. Thus the same active frozen-reference trajectory serves this fixed width.

The fixed positive pair lift and the appended zero feature columns preserve the complete refit and Fourier observers by Lemma E.25. Hence the reference conclusions apply at this width: Proposition E.37 supplies strict training margins, the CE interval and variation bound, and the early and late actual prediction signs; Proposition E.36 supplies the late Fourier gain and positive centered-table energy; Proposition E.39 supplies strict wrong-class margins for the complete early refit. If  $p \geq 1764$ , Proposition E.52 supplies the strict late error bound and 11/50 nonsquare refit margins. The latter margins count  $pd$  correct rows among  $p(d+1)$  held-out rows; the full error denominator still includes the diagonal rows.

**Joint continuity of the actual update.** For  $\theta = (\mathbf{W}, \mathbf{V})$  write  $\zeta = (\theta_0, \eta, d_V)$ ,  $\zeta^0 = (\theta_M^{\text{can}}, 0, 0)$  and  $\Theta_j(\zeta) = \mathcal{U}_{\eta, d_V}^{\circ j}(\theta_0)$ . Let  $\mathbf{K}_E$  have orthonormal columns spanning  $E$ , and put  $\mathbf{\Gamma}(\theta) = \mathbf{K}_E^\top \mathbf{F}(\theta)^\top \mathbf{F}(\theta) \mathbf{K}_E$ . At a frozen reference departure write  $\mathbf{G} = -\nabla_{\mathbf{V}} L$ , and let  $P(c \mid x_i)$  be its softmax probabilities. Strict training margins give

$$\langle \mathbf{V}, \mathbf{G} \rangle_F = \frac{1}{n} \sum_{i \in \mathcal{S}} \sum_{c \neq y_i} P(c \mid x_i) (f_{y_i}(x_i) - f_c(x_i)) > 0. \quad (117)$$

Equation (117) ensures that the head normalization has a nonzero denominator. On a neighborhood with the same training gates,  $\mathbf{F}, \mathbf{G}$  are continuous and  $\mathbf{F} = \mathbf{F} \mathbf{P}_E$  by Lemma E.2. The supported source rank supplied by the bootstrap makes  $\mathbf{\Gamma}(\theta)$  positive definite, so it stays positive definite nearby. There

$$\mathcal{N}_{\text{pol}}(\mathbf{F}) = \frac{\mathbf{F} \mathbf{K}_E \mathbf{\Gamma}(\theta)^{-1/2} \mathbf{K}_E^\top}{\sqrt{r}}, \quad \mathcal{N}_F(\mathbf{G}) = \mathbf{G} / \|\mathbf{G}\|_F \quad (118)$$

are continuous. Substituting (118) into the update (63) proves joint continuity of  $\mathcal{U}_{\eta, d_V}$  in state,  $\eta$  and  $d_V$  at  $(\theta_{M,j}^0, 0, 0)$ . Finite composition makes every arrival  $\Theta_j(\zeta)$ ,  $0 \leq j \leq J$ , continuous at  $\zeta^0 = (\theta_M^{\text{can}}, 0, 0)$ .

**One full-coordinate ball and the two decay scopes.** At every reference arrival the training feature table is nonzero, because vanishing features would give zero logits and zero training margins. Its prescribed refit penalty is therefore positive. The ridge formula makes the full refit scores and normalized errors continuous at that table. The positive centered-table energy likewise makes the Fourier fraction continuous. Held-out ReLU gates need not stay fixed for these continuity statements.

The strict reference conclusions include every training contrast in  $(0, 0.0201)$ , not only the aggregate CE interval. They involve finitely many arrivals, inputs and classes. Joint continuity therefore supplies a full-coordinate  $\delta$ -ball about  $\theta_M^{\text{can}}$  and positive parameter caps  $\eta_0, d_0$  preserving part (a); shrink them so  $\delta \leq 1/2$  and  $d_0 \leq 1/2$ . For  $p \geq 1764$ , continuity on the slice  $d_V = 0$  also preserves the late refit's strict error and nonsquare margins. Shrink  $\delta$  and  $\eta_0$ , keeping  $d_0$  fixed, so that  $\mathcal{B}_M \times [0, \eta_0) \times \{0\}$  lies in this slice neighborhood. Shrinking preserves part (a), and proves part (b) on the same final ball and head-step interval, without asserting a late-refit conclusion for positive decay.

Accuracy is preserved through the stated strict correct or wrong score margins, not through continuity of the accuracy function. The same positive training margins give  $\mathbf{G}_j \neq 0$  at every actual departure in the selected neighborhood, so the head increment has norm  $\eta$  when  $d_V = 0$  and  $\eta > 0$ . All  $3pM$  initial coordinates may vary in this ball.  $\square$

*Proof of Theorem 4.1.* Fix  $p \geq 101$ ,  $M \geq M_{\min}(p)$  and  $\lambda > 0$ . Choose  $\epsilon > 0$  sufficiently small for all component certificates in Proposition E.54, including the late-refit certificate when  $p \geq 1764$ . This fixes  $s, J$  and the two windows in (60)–(61) before the initialization is chosen. Lemma E.26 gives  $\mathcal{I}_{\text{early}} = \{1, \dots, 32\}$ ,  $|\mathcal{I}_{\text{late}}| \geq 16$ , and  $\max \mathcal{I}_{\text{early}} < \min \mathcal{I}_{\text{late}}$ , with both windows contained in  $\{1, \dots, J\}$ . Take the final ball and constants of Proposition E.54 and set  $(\eta_W, \lambda_W, \eta_V) = (s, \lambda, \eta_0/2)$ ,  $\lambda_V = 0$ . The ball  $\mathcal{B}_M$  is nonempty and open in  $\mathbb{R}^{3pM}$ . Part (a) of that proposition gives (i)–(iii), and part (b) gives (iv) for  $p \geq 1764$ , with all guarantees on this same ball and trajectory.  $\square$

**Scope.** The positive constants  $\delta, \eta_0, d_0$  may depend on  $p, M, \lambda$ , and the chosen witness  $\epsilon$  and its finite  $J$ . No quantitative lower bound or uniformity over widths is asserted. For each fixed width, the same selected ball, head step, horizon and windows serve the undecayed theorem and the half-grid decay conclusion.

## F A random-split low-CE theorem with refit-first recovery

The task, network and Fourier observable are those of (1) and (6). This is a different construction from Theorem 4.1: its width, readout penalty and asymptotic regime are specified in Theorem 4.2 and Section 3.5. Its dictionary, split event, calibrated head, reference path and initial ball are defined in the corresponding subsections below. Symbols in this appendix are local to it except for the shared interface listed in Section N.

### F.1 Statement and construction

The elementary construction formulas below apply to odd primes  $p \geq 5$ ; the final theorem uses the larger common threshold. Throughout this appendix,  $m = M = p^2(p-1)^2$  denotes the hidden width. With  $h$  from Section N and  $k, a$  from Definition F.17, prescribe

$$\lambda = k/h, \quad s = h(1-a)/k, \quad \eta = \theta/(hp).$$

These are fixed before sampling. The actual learner is

$$(\eta_W, \lambda_W, \eta_V, \lambda_V) = (s, \lambda, \eta, 0) \tag{119}$$

in (3), with mean CE over the realized split. For the refit, write  $\gamma = \kappa/[4p(p-1)]$  and  $q_{H_j} = \tau_j = \gamma \|\mathbf{H}_{j,S}\|_F^2$ . Its target-normalized error is also denoted  $E_{\text{obs}}(\mathbf{H}_j) = E_{\text{ref}}(j)$ ; Definition F.28 gives the same own-penalty error for any nonzero training-feature table. We write  $\text{Acc}_{\text{ref}}(\mathbf{U})$  for the held-out accuracy of this refit of a feature table  $\mathbf{U}$ , with the fixed tie rule.

**Theorem F.1** (Quantitative random-split realization). *Theorem 4.2 holds on the event  $\mathcal{E}_{\text{split}}$  of Definition F.10 with the fixed updates (119) and the training-data-built center (120)–(135). Its initial region  $\mathcal{R}_{p,S}^{\text{random}}$  is the nonempty ordinary open ball of radius (141) in all 3pm weight coordinates about that center. The refit is (18), and  $E_{\text{ref}}(j) = E_{\text{obs}}(\mathbf{H}_j)$ . In addition, the training margins in clause (i) are strictly positive, and the Fourier fraction in clause (iv) remains  $O(1/p)$  throughout  $0, \dots, J$ . The error gain compares each trajectory with its own initial error. No accuracy after  $J$  is asserted.*

The rates are fixed before sampling, and both weight matrices remain trainable and untied. The label-blind rectangular dictionary contains no added memories. The head (135) uses two one-time centered train-only calibrations:  $b_{\text{base}}$  for the canonical base, whose normalized residual is  $\hat{\mathbf{R}}$ , and  $b_*$  for the mixed head. Proposition F.22 proves their existence, uniqueness and bounds. This engineered warm start is neither a zero head nor a proved outcome of earlier training. Neither the actual initial head nor the separate complete diagnostic correctly predicts any held-out input: Proposition F.22 and Lemma F.35 establish the center claims, and Lemma F.30 and Proposition F.36 extend them to every initial state in the prescribed ball. Refitted heads are never installed in the learner.

**Proof strategy.** The construction first fixes the split event and clock, then calibrates the initial head in Proposition F.22. The reference kernels in Section F.2.4 describe how this center evolves along the fixed-direction comparison path. Lemma F.30 then controls actual states on the prescribed  $r_6$  ball, uniformly over the event and every head sequence with increments at most  $\eta$ . Its Frobenius estimate gives mean loss and held-out accuracy; its rank and operator estimates give strict training margins. These are the outputs of Claim I.

Claim II, in Section F.3, proves the full-refit and Fourier conclusions separately on those same states. Initial actual-head failure comes from Lemma F.30, while initial refit failure uses the signed estimate in Lemma F.35. Proposition F.36 assembles the conclusions with one pre-sampling threshold. The actual trajectory is not required to retain its initial gates.

## F.2 Claim I: actual dynamics, plateau and delayed network recovery

### F.2.1 Split model, dictionary and the initial-data event

**Grid notation.** Use the task and centered targets of Sections 3.1 and 3.5, and the shared height scale  $h$  of Section N. For  $x = (a, b)$ , write  $\mathbf{x}_x = x_{ab}$ ,  $y_x = y^*(x)$  and  $t_x = T_x$ . Here  $\mathbf{P} = \mathbf{C}_p$  is class centering, so  $\mathbf{T}_S = \mathbf{Y}_S \mathbf{P}$ , and  $n_D = |\mathcal{D}| = p^2 - n$ .

**Definition F.2** (Grid tables and construction scale). Let  $\mathbf{X} \in \mathbb{R}^{\Omega \times 2p}$  be the grid input table with rows  $\mathbf{x}_x^\top$ , and let  $\mathbf{Y}_S$  be the one-hot training-label table. The visible space  $E_{\text{vis}} = \{(v, v') \in \mathbb{R}^p \times \mathbb{R}^p : \mathbf{1}^\top v = \mathbf{1}^\top v'\}$  contains every  $\mathbf{x}_x$  and has dimension  $r_v = \dim E_{\text{vis}} = 2p - 1$ ; write  $A = \log(p - 1) + a_{\text{conf}}$  for the confidence scale. The fixed theorem parameters and Bernoulli split are those of Theorem 4.2.

#### Split model.

**Definition F.3** (Sampled masks and counts). The mask  $\mathbf{M} \in \{0, 1\}^{p \times p}$  has independent Bernoulli( $\rho$ ) entries, and  $\mathcal{S} = \{(a, b) : M_{ab} = 1\}$ . We write  $M_x = M_{ab}$  for  $x = (a, b)$ ,  $n_a^{\text{row}} = \sum_b M_{ab}$ ,  $n_b^{\text{col}} = \sum_a M_{ab}$  and  $n_c = \sum_a M_{a, c-a}$  for the row, column and class counts, and

$$\hat{n}_{a,c}^{\text{row}} = M_{a, c-a} - \frac{n_a^{\text{row}}}{p}, \quad \hat{n}_{b,c}^{\text{col}} = M_{c-b, b} - \frac{n_b^{\text{col}}}{p}, \quad \hat{n}_c = n_c - \frac{n}{p} = \sum_a \hat{n}_{a,c}^{\text{row}} = \sum_b \hat{n}_{b,c}^{\text{col}}$$

for their class-centered versions. The centered masks are  $\bar{\mathbf{M}} = \mathbf{M} - \rho \mathbf{1} \mathbf{1}^\top$ ,  $\bar{M}_{a,c}^{\text{row}} = M_{a, c-a} - \rho$  and  $\bar{M}_{b,c}^{\text{col}} = M_{c-b, b} - \rho$ . For  $x = (a, b)$  set  $N_{xc} = M_{a, c-a} + M_{c-b, b}$ .

Write  $\hat{n} = \mathbf{T}_S^\top \mathbf{1}$ , with coordinates  $\hat{n}_c = n_c - n/p$ .

For fixed  $x$ , rival arm pairs use distinct Bernoulli entries, so the  $N_{xc}$ ,  $c \neq y_x$ , are independent Binomial(2,  $\rho$ ) variables, independent of  $M_x$ . The three permuted centered mask matrices have independent entries individually; independence between those matrices is never used.

#### Dictionary and forces.

**Definition F.4** (Dictionary and full-grid force). Let  $b_0 = (p + 1)/2$ ,  $I_0 = \{0, \dots, b_0 - 1\}$ ,  $\mu = b_0/p$  and  $q = \mu/2 = (p + 1)/(4p)$ . The neurons of (120) are indexed by  $(u, v, u', v') \in (\mathbb{F}_p^\times \times \mathbb{F}_p)^2$ , with  $l$  and  $r$  the indicator vectors of  $uI_0 + v$  and  $u'I_0 + v'$ . Write  $\mathbb{E}_{l,r}$  for the uniform average over these  $m$  neurons and  $g_x = l_a r_b$  for  $x = (a, b)$ . Let  $\mathbf{O}_{\text{row}}$  and  $\mathbf{O}_{\text{col}}$  be the  $\Omega \times \Omega$  matrices with entries  $\mathbf{1}\{a = a'\}$  and  $\mathbf{1}\{b = b'\}$ , and  $\mathbf{B}_0 = \mathbb{I} + \mathbf{O}_{\text{row}} + \mathbf{O}_{\text{col}} + \mathbf{1} \mathbf{1}^\top$ .

Denote the prescribed center by  $(\bar{\mathbf{W}}_0, \bar{\mathbf{V}}_0)$  and its feature table by  $\bar{\mathbf{H}}_0 = \mathbf{H}(\bar{\mathbf{W}}_0)$ ; unbarred initial states may be any point in the certified ball. For the two indicator masks  $l, r \in \{0, 1\}^p$ , the dictionary row is

$$(\bar{\mathbf{W}}_0)_{(l,r)} = \frac{\alpha}{\sqrt{m}}(2l - \mathbf{1} - \beta \mathbf{1}, 2r - \mathbf{1} - \beta \mathbf{1}). \quad (120)$$

The full-grid rule force is  $c_F = \mathbf{X}^\top \text{diag}(g) \mathbf{T} \mathbf{T}^\top g$ . Forces use the neuron norm  $\|c\|_{L^2} = (\mathbb{E}_{l,r} [\|c(l, r)\|^2])^{1/2}$ . The interaction source is  $f_x = (l_a - \mu)(r_b - \mu)$ .

For the center dictionary put  $d = h^2 q^2$ ,  $\bar{\mathbf{H}} = \bar{\mathbf{H}}_{0,S}$ ,  $\mathbf{B} = \mathbf{H} \mathbf{H}^\top / d$  and  $w = \mathbf{1}_m / \sqrt{m}$ .

**Lemma F.5** (Dictionary identities). *For the dictionary and kernels in Definition F.4:*

- (i)  $\Pr(l_a = 1) = \mu$  and  $\Pr(l_a = l_{a'} = 1) = q$  for  $a \neq a'$ . Hence  $\mathbb{E}[ll^\top] = q(\mathbb{I} + \mathbf{1} \mathbf{1}^\top)$  and the covariance of  $l$  is  $q\mathbf{P}$ ; the same holds for  $r$ , which is independent of  $l$ .
- (ii)  $\mathbb{E}_{l,r}[gg^\top] = q^2 \mathbf{B}_0$ , and the interaction source satisfies  $\mathbb{E}_{l,r}[f_x f_{x'}] = q^2 P_{aa'} P_{bb'}$ .
- (iii)  $\bar{H}_{0,x,(l,r)} = h g_x / \sqrt{m}$ . Hence  $\bar{\mathbf{H}}_0 \bar{\mathbf{H}}_0^\top = d \mathbf{B}_0$  with  $\bar{\mathbf{H}}_0 w = h \mu^2 \mathbf{1}_\Omega$ , and the normalized training Gram matrix  $\mathbf{B}$  of (135) equals  $\mathbf{B}_{0,SS}$ .

$$(iv) \mathbf{B}_0 \mathbf{T} = \mathbf{T}, \mathbf{T}^\top \mathbf{T} = p\mathbf{P}, \|\mathbf{T}\|_{\text{op}} = \sqrt{p}, \|\mathbf{X}\|_{\text{op}} = \sqrt{2p}, \text{ and } \mathbb{E}_{l,r} \left[ \left\| \mathbf{T}^\top g \right\|^2 \right] = q^2 p(p-1).$$

*Proof.* (i) Fix  $a$ . For each  $k \in I_0$  and  $u \neq 0$  exactly one  $v$  gives  $uk + v = a$ , so  $a \in uI_0 + v$  for  $b_0(p-1)$  of the  $p(p-1)$  pairs. For  $a \neq a'$ , the map  $(k, k') \mapsto (u, v) = ((a - a')/(k - k'), a - uk)$  is a bijection from ordered pairs of distinct elements of  $I_0$  onto the pairs with  $a, a' \in uI_0 + v$ , so the probability is  $b_0(b_0 - 1)/(p(p-1)) = q$ . As  $\mu = 2q$  and  $\mu^2 = q(1 + 1/p)$ , the covariance is  $q(\mathbb{I} + \mathbf{1}\mathbf{1}^\top) - \mu^2 \mathbf{1}\mathbf{1}^\top = q\mathbf{P}$ . (ii) follows from (i) and the independence of  $l$  and  $r$ . (iii) At  $x$ , the preactivation of neuron  $(l, r)$  is  $\alpha(2l_a + 2r_b - 2 - 2\beta)/\sqrt{m}$ , which equals  $h/\sqrt{m}$  if  $l_a = r_b = 1$  and is at most  $-2\alpha\beta/\sqrt{m} < 0$  otherwise. Thus  $\bar{\mathbf{H}}_0 \bar{\mathbf{H}}_0^\top = h^2 \mathbb{E}_{l,r} [gg^\top]$  and  $(\bar{\mathbf{H}}_0 w)_x = h \mathbb{E}_{l,r} [g_x] = h\mu^2$ . (iv) Every operand row and column of  $\Omega$  meets each class exactly once, so  $\mathbf{O}_{\text{row}} \mathbf{T} = \mathbf{O}_{\text{col}} \mathbf{T} = 0$  and  $\mathbf{1}^\top \mathbf{T} = 0$ ;  $\mathbf{T}^\top \mathbf{T} = p\mathbf{P}$ , and  $\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} p\mathbb{I} & \mathbf{1}\mathbf{1}^\top \\ \mathbf{1}\mathbf{1}^\top & p\mathbb{I} \end{pmatrix}$  has largest eigenvalue  $2p$ . Finally  $\mathbb{E} \left[ \left\| \mathbf{T}^\top g \right\|^2 \right] = q^2 \text{tr}(\mathbf{T}^\top \mathbf{B}_0 \mathbf{T}) = q^2 \text{tr}(\mathbf{T}^\top \mathbf{T})$ .  $\square$

For a neuron  $(l, r)$ , a source  $\varphi \in \mathbb{R}^\Omega$  (which may depend on the neuron) and  $\mathcal{S}' \subseteq \Omega$ , define the gated force

$$c_\varphi(\mathcal{S}') = \mathbf{X}_{\mathcal{S}'}^\top \text{diag}(g_{\mathcal{S}'}) \mathbf{T}_{\mathcal{S}'} \mathbf{T}_{\mathcal{S}'}^\top \varphi_{\mathcal{S}'} = \sum_{i,j \in \mathcal{S}'} g_i \langle t_i, t_j \rangle \varphi_j \mathbf{x}_i \in \mathbb{R}^{2p}. \quad (121)$$

In particular  $c_F = c_g(\Omega)$ . Since  $g - f = \mu l_a + \mu r_b - \mu^2$  is additive,  $\mathbf{T}^\top (g - f) = 0$ ; hence  $c_F = c_f(\Omega)$  and  $c_1(\Omega) = 0$ .

### Count, design and arm events.

**Definition F.6** (Count thresholds and design/arm events). Use the pre-sampling mixing coefficient  $\xi$  of Definition F.17. Before sampling define

$$r_\xi = e^{-A\xi}, \quad Q_\xi = (p-1)(\rho + (1-\rho)r_\xi)^2, \quad \varsigma = \rho(1-\rho). \quad (122)$$

For  $\delta \in (0, 1)$  put  $\ell_\delta = \log(32p^2/\delta)$  and

$$D_* = \sqrt{p\ell_\delta/2}, \quad t_* = \sqrt{4p \log 9 + 2 \log(24/\delta)}, \quad Q_* = \sqrt{2p\ell_\delta}, \quad K_* = 4D_* + t_*. \quad (123)$$

Put  $n_{\max} = \rho p + D_*$  and  $n_- = p(\rho p - D_*)$ . Set

$$D_H = \sqrt{\frac{3p}{2} \log \frac{2p(p-1)}{\delta}}, \quad D_L = \sqrt{\frac{p-1}{2} \log \frac{4p^2}{\delta}}, \quad D_W = \sqrt{\frac{p(p-1)}{2} \log \frac{2p}{\delta}}.$$

Define  $\mathcal{E}_{\text{des}}(\delta)$  by

$$\max_a |n_a^{\text{row}} - \rho p|, \max_b |n_b^{\text{col}} - \rho p|, \max_c |n_c - \rho p| \leq D_*; \quad (124)$$

$$\|\bar{\mathbf{M}}\|_{\text{op}}, \|\bar{\mathbf{M}}^{\text{row}}\|_{\text{op}}, \|\bar{\mathbf{M}}^{\text{col}}\|_{\text{op}} \leq t_*; \quad (125)$$

$$\max_{a,c} \left| \sum_b \bar{M}_{ab} \hat{n}_{b,c}^{\text{col}} \right| \leq Q_*, \quad \max_{b,c} \left| \sum_a \bar{M}_{ab} \hat{n}_{a,c}^{\text{row}} \right| \leq Q_*. \quad (126)$$

For  $y \neq c$  put  $\mathbf{H}_{yc} = \sum_a M_{a,y-a} M_{a,c-a} M_{a+c-y,y-a}$ . Let  $N_x^{(2)} = \#\{c \neq y_x : N_{xc} = 2\}$  and put  $\mathbf{W}_c = \sum_{i \in \mathcal{S}, y_i \neq c} r_\xi^{2-N_{ic}^{(2)}}$ . The other three events are

$$\mathcal{E}_H(\delta) = \{|\mathbf{H}_{yc} - \rho^3 p| \leq D_H \forall y \neq c\},$$

$$\mathcal{E}_L(\delta) = \{|N_x^{(2)} - (p-1)\rho^2|,$$

$$|\#\{c \neq y_x : N_{xc} = 0\} - (p-1)(1-\rho)^2| \leq D_L \forall x\},$$

$$\mathcal{E}_W(\delta) = \{|\mathbf{W}_c - \rho[n_c(n_c - 1) + 2r_\xi n_c(p - n_c) + r_\xi^2(p - n_c)(p - n_c - 1)]| \leq D_W \forall c\}.$$

The count certificates used in the bounds are  $\mathbf{H}_{\min} = \rho^3 p - D_H$ ,  $\zeta_x^0 = \sum_{c \neq y_x} r_\xi^{2-N_{xc}^{(2)}}$  and  $\Delta_W = D_W + (3p+1)D_* + p$ .

For the count calculation, put

$$u_\xi = \rho + (1 - \rho)r_\xi, \quad N_x^{(0)} = \#\{c \neq y_x : N_{xc} = 0\},$$

and

$$F(k) = k(k-1) + 2r_\xi k(p-k) + r_\xi^2(p-k)(p-k-1).$$

**Lemma F.7** (Count and arm events). *For  $\delta \in (0, 1)$ , use the four count/arm events in Definition F.6. Each has probability at least  $1 - \delta$ . On  $\mathcal{E}_{\text{des}}$ ,  $|\hat{n}_c| \leq 2D_*$ ; on  $\mathcal{E}_H$ ,  $H_{yc} \geq H_{\min}$ ; on  $\mathcal{E}_L$ ,  $|\zeta_x^0 - Q_\xi| \leq (1 - r_\xi^2)D_L \leq D_L$ ; and on  $\mathcal{E}_{\text{des}} \cap \mathcal{E}_W$ ,*

$$|W_c - n_c Q_\xi| \leq \Delta_W.$$

*Proof. Tail bound and scalar-event families.*

We use one bounded-range estimate. For a variable of range length  $\ell$ , the second derivative of its centered log moment-generating function is a tilted variance, at most  $\ell^2/4$  by comparison with the range midpoint. Integrating twice bounds it by  $t^2\ell^2/8$ . For sums with  $r$  colors and independence within each color, Hölder with exponent  $r$  across colors, followed by exponential Markov and optimization, gives

$$\Pr\{|S - \mathbb{E}[S]| \geq u\} \leq 2 \exp[-2u^2/(r \sum_i \ell_i^2)].$$

The argument also holds conditionally. The five scalar applications have the following numbers of tests, effective range-square sums, and union costs:

| family             | tests    | $r \sum_i \ell_i^2$ | cost            |
|--------------------|----------|---------------------|-----------------|
| counts             | $3p$     | $p$                 | $3\delta/(16p)$ |
| cross sums         | $2p^2$   | $4p$                | $\delta/8$      |
| H                  | $p(p-1)$ | $3p$                | $\delta$        |
| $N^{(0)}, N^{(2)}$ | $2p^2$   | $p-1$               | $\delta$        |
| W                  | $p$      | $p(p-1)$            | $\delta$        |

Here the counts are sums of Bernoulli variables. The cross-sum terms  $\bar{M}_{ab}\hat{n}_{b,c}^{\text{col}}$  lie in  $[-1, 1]$ , use independent columns, and have means  $\varsigma(\mathbf{1}\{a = c - b\} - 1/p)$  summing to zero; transpose for the other family. For  $\Delta = c - y$ , write  $H_{yc} = \sum_a Y_a C_a C_{a+\Delta}$  with  $Y_a = M_{a,y-a}$ ,  $C_a = M_{a,c-a}$ . These mean- $\rho^3$  indicators have the three-colorable  $p$ -cycle as dependency graph. Rival arm pairs at fixed  $x$  are disjoint, giving independent indicators of means  $\rho^2, (1 - \rho)^2$ . Finally, condition on the class- $c$  line, with  $k = n_c$  sampled entries. Write each remaining input as  $(a, c - a')$ ,  $a \neq a'$ . These  $p(p-1)$  mask bits are independent; their weights are  $k(k-1)$  copies of 1,  $2k(p-k)$  copies of  $r_\xi$ , and  $(p-k)(p-k-1)$  copies of  $r_\xi^2$ , giving conditional mean  $\rho F(k)$ . Substituting  $D_*, Q_*, D_H, D_L, D_W$  gives the displayed costs.

*Matrix-norm events and union accounting.*

For each of the three matrices in (125), a unit bilinear form has range-square sum 1. A maximal  $1/4$ -separated sphere net has at most  $9^p$  points, by comparing disjoint radius- $1/8$  balls with a radius- $9/8$  ball; approximation in both arguments bounds the matrix norm by twice the largest net bilinear form. Thus these tests cost  $6 \cdot 9^{2p} e^{-t_*^2/2} = \delta/4$ . The total design cost is  $3\delta/(16p) + 3\delta/8 < \delta$ , proving all four probability claims.

*Consequences for counts, arms and weights.*

The centered-count bound follows because  $n/p$  is the mean class count. For  $r \in [0, 1]$ ,  $\sum_{c \neq y_x} r^{2-N_{xc}} = r(p-1) + (1-r)N_x^{(2)} - r(1-r)N_x^{(0)}$ , whose mean is  $(p-1)(\rho + (1-\rho)r)^2$ ; its event error is at most  $(1-r^2)D_L$ . For  $r = r_\xi$ ,  $F'(k) = 2k(1-r)^2 + 2r(1-r)p - (1-r^2) \in [-1, 2p]$  and  $\rho F(\rho p) - \rho p Q_\xi = -\rho^2(1-\rho)(1-r)^2 p$ . Together with  $Q_\xi \leq p-1$  and  $|n_c - \rho p| \leq D_*$  these prove the stated bound on  $W_c$ .  $\square$

**Definition F.8** (Sampled-force events). Let  $c_f = c_f(\mathcal{S})$ ,  $c_g = c_g(\mathcal{S})$  and  $c_1 = c_1(\mathcal{S})$  be the forces (121) with the interaction, gate and constant sources. For  $\delta \in (0, 1)$  define the events  $\mathcal{E}_f(\delta) = \{\|c_f - \rho^2 c_F\|_{L^2} \leq t_f\}$  and  $\mathcal{E}_g(\delta) = \{\|c_g - a_g c_1 - \rho^2 c_F\|_{L^2} \leq t_g\}$ , with  $t_f = p^{3/2}(2\delta)^{-1/2}$ ,  $t_g = (2/\delta)^{1/2} p^{3/2}$  and the neuron-dependent coefficient  $a_g = \mathbf{1}_S^\top g_S / n \in [0, 1]$  when  $n > 0$  (and  $a_g = 0$  when  $n = 0$ ).

**Lemma F.9** (Sampled initial forces). *For  $p \geq 5$  and  $\delta \in (0, 1)$ , each sampled-force event  $\mathcal{E}_f(\delta), \mathcal{E}_g(\delta)$  of Definition F.8 has probability greater than  $1 - 4\delta$ .*

*Proof.* For  $\varphi \in \{f, g, \mathbf{1}\}$ , define the local finite-difference quantities

$$\begin{aligned}\partial_i C &= C|_{M_i=1} - C|_{M_i=0}, & u &= \mathbf{T}^\top \text{diag}((M_x)_{x \in \Omega})\varphi, \\ v_c &= \sum_j M_j g_j (\mathbf{1}\{y_j = c\} - 1/p) \mathbf{x}_j, & d_i &= (1 - 1/p) g_i \varphi_i \mathbf{x}_i.\end{aligned}$$

For the finite difference  $\partial_i C$ , orthogonality of products of  $M_i - \rho$  gives  $\text{Var}(C) \leq \varsigma \sum_i \mathbb{E} [\|\partial_i C\|^2]$ : each nonconstant monomial's variance is counted by its degree on the right. Apply this to  $C = c_\varphi$ ,  $\varphi \in \{f, g, \mathbf{1}\}$ , in the neuron  $L^2$  space. The quantities  $u, v_c, d_i$  satisfy, by direct differentiation,

$$\partial_i c_\varphi = g_i u_{y_i} \mathbf{x}_i + \varphi_i v_{y_i} + (1 - 2M_i) d_i;$$

the  $M_i$  terms cancel. Here  $|\varphi_i| \leq 1$  and  $\mathbb{E}_{l,r} [\|\mathbf{T}^\top \varphi\|^2] \leq p^2$ , so averaging over masks and neurons gives  $\mathbb{E} [\|u\|^2] \leq (\rho^2 + \varsigma) p^2 = \rho p^2$ . At the full mask, each coordinate of  $v_c$  has magnitude at most one (on the left it is  $l_a(r_{c-a} - \mu)$ ); also  $\sum_{c,j} \|g_j (\mathbf{1}\{y_j = c\} - 1/p) \mathbf{x}_j\|^2 \leq 2p^2$ . Thus  $\mathbb{E} [\sum_c \|v_c\|^2] \leq 2\rho p^2$ . Each class has  $p$  inputs and  $\sum_i \|d_i\|^2 \leq 2p^2$ , so the three-term squared triangle inequality bounds the variance by  $3\varsigma(4\rho p^3 + 2p^2)$ . Moreover  $\mathbb{E}_M[c_\varphi] - \rho^2 c_\varphi(\Omega) = \varsigma \sum_i d_i$ , whose squared norm is at most  $2\varsigma^2 p^3$ . Hence

$$\mathbb{E} [\|c_\varphi - \rho^2 c_\varphi(\Omega)\|_{L^2}^2] \leq 3\varsigma(4\rho p^3 + 2p^2) + 2\varsigma^2 p^3 \leq \frac{1112}{625} p^3 < 2p^3.$$

Markov gives  $\Pr(\mathcal{E}_f^c) < 4\delta$  and  $\Pr\{\|c_g - \rho^2 c_F\|_{L^2}^2 + \|c_1\|_{L^2}^2 > p^3/\delta\} < 4\delta$ . For every neuron-dependent  $a \in [0, 1]$ ,  $\|c_g - a c_1 - \rho^2 c_F\|_{L^2}^2 \leq 2(\|c_g - \rho^2 c_F\|_{L^2}^2 + \|c_1\|_{L^2}^2)$ , proving  $\mathcal{E}_g$ .  $\square$

**Definition F.10** (Complete initial-data split event). Let

$$\begin{aligned}\mathcal{E}_{\text{split}} &= \mathcal{E}_{\text{des}}(p^{-1/2}/7) \cap \mathcal{E}_H(p^{-1/2}/7) \cap \mathcal{E}_L(p^{-1/2}/7) \\ &\quad \cap \mathcal{E}_W(p^{-1/2}/7) \cap \mathcal{E}_f(p^{-1/2}/7) \cap \mathcal{E}_g(p^{-1/2}/7).\end{aligned}$$

Unless its argument is displayed, a split threshold uses  $\delta = p^{-1/2}/7$ . The nonempty proper-split and denominator conditions are proved in Proposition F.11 and the assembly of Appendix F. The event  $\mathcal{E}_H$  and its certificate  $\mathbf{H}_{\min}$  are retained in this event; the calibration argument below uses a direct bound from  $\mathcal{E}_{\text{des}}$  instead.

For the following failure-budget calculation, write  $\delta_1 = p^{-1/2}/7$ .

**Proposition F.11** (Initial-data split event). *The event  $\mathcal{E}_{\text{split}}$  of Definition F.10 satisfies:*

- (i)  $\Pr(\mathcal{E}_{\text{split}}) \geq 1 - 2/\sqrt{p}$ .
- (ii) *The thresholds are  $t_f = (7/2)^{1/2} p^{7/4}$  and  $t_g = 14^{1/2} p^{7/4}$ , and, as  $p \rightarrow \infty$ ,  $t_* = O(\sqrt{p})$ ,  $D_W = O(p\sqrt{\log p})$  and  $D_*, Q_*, K_*, D_H, D_L = O(\sqrt{p \log p})$ .*
- (iii) *For all sufficiently large  $p$ ,  $K_* < \rho p$  and  $D_* < \min(\rho, 1 - \rho)p = p/5$ . Then, on  $\mathcal{E}_{\text{split}}$ , every operand row, operand column and class contains both training and held-out inputs; in particular the split is proper and  $n_c \geq 1$  for every class.*

*Proof.* The four count/arm failures cost at most  $4\delta_1$  and the two force failures less than  $8\delta_1$ , so their sum is less than  $12/(7\sqrt{p}) < 2/\sqrt{p}$ . At  $\delta = \delta_1$  every logarithm in the count thresholds is  $O(\log p)$ , which gives the orders in (ii), and  $t_f, t_g$  follow by substitution. Hence (iii) holds for large  $p$ , and then all row, column and class counts lie strictly between 0 and  $p$ .  $\square$

**Standing assumptions and the sampled design.** From here on the theorem parameters, including  $J$ , are fixed, the split lies in  $\mathcal{E}_{\text{split}}$ , and  $p$  is a sufficiently large odd prime. All estimates are uniform over this event. One common threshold  $p_0 \geq 2048^2$  is chosen in Proposition F.36; the initial-data bounds at the beginning of the proof of Lemma F.30 establish every denominator and sign condition used below, independently of any trajectory.

**Definition F.12** (Sampled design). Let  $\Upsilon = \text{diag}(\mathbb{I}, \mathbb{I}, (p+2)^{-1})$  on  $\mathbf{1}^\perp \oplus \mathbf{1}^\perp \oplus \mathbb{R}$ ,  $z_x = (\mathbf{P}e_a, \mathbf{P}e_b, p^{-1/2})$  for  $x = (a, b)$ , and let  $\mathcal{Z}$  have rows  $z_x$ . Put  $\mathcal{K} = \mathcal{Z}_S^\top \mathcal{Z}_S$ ,  $\Psi = \mathcal{Z}_S^\top \mathbf{T}_S$ ,  $s_0 = \rho p - K_*$ ,  $\kappa_* = K_*/(\rho p)$ ,  $N_S = \|\mathbf{T}_S\|_F$ ,  $N_{\mathcal{D}} = \|\mathbf{T}_{\mathcal{D}}\|_F$  and

$$N_{\mathcal{D}}^- = \sqrt{p((1-\rho)p - D_*)(1-1/p)}, \quad N_+ = \sqrt{p(p-1)}. \quad (127)$$

Use the sampled-design quantities (127), and put

$$E_o = \mathbf{1}^\perp \oplus \mathbf{1}^\perp \oplus \mathbb{R}, \quad n_{\mathcal{D}}^- = p((1-\rho)p - D_*).$$

**Lemma F.13** (Sampled design).  $\mathcal{Z}^\top \mathcal{Z} = p\mathbb{I}_{E_o}$  and  $\mathcal{Z}\Upsilon^{-1}\mathcal{Z}^\top = \mathbf{O}_{\text{row}} + \mathbf{O}_{\text{col}} + \mathbf{1}\mathbf{1}^\top$ , so  $\mathbf{B}_0 = \mathbb{I} + \mathcal{Z}\Upsilon^{-1}\mathcal{Z}^\top$ . Moreover  $n \geq n_- \geq \rho p^2/2$ ,  $n_{\mathcal{D}} \geq n_{\mathcal{D}}^- > 0$ ,  $N_{\mathcal{D}}^- \leq N_{\mathcal{D}} \leq N_+$ ,  $N_S \leq N_+$  and  $\|\mathcal{K} - \rho p\mathbb{I}\|_{\text{op}} \leq K_*$ ; hence  $s_0\mathbb{I} \preceq \mathcal{K} \preceq (\rho p + K_*)\mathbb{I}$  and  $\kappa_* \leq 1/2$ . Also  $\|\mathbf{X}_S\|_{\text{op}} \leq \sqrt{2n_{\text{max}}}$ ,  $\|\mathbf{T}_S\|_{\text{op}} \leq \sqrt{n_{\text{max}}} \leq \sqrt{p}$ ,  $\|\mathcal{Z}_{\mathcal{D}}\|_{\text{op}} \leq \sqrt{p}$ ,  $\|\Psi\|_{\text{op}} \leq \sqrt{3}t_*$ ,  $\|\Psi\|_F \leq \sqrt{3(p-1)}t_*$ ,  $\Psi\mathbf{P} = \Psi$ ,  $\mathbf{T}_S^\top \mathbf{T}_S = \mathbf{P} \text{diag}(n_c)\mathbf{P}$ , and  $\|\mathcal{Z}_{Sc}\|^2 \geq (s_0/p)\|\mathcal{Z}c\|^2$  for  $c \in E_o$ .

*Proof.* The identities for  $\mathcal{Z}$  follow from  $z_x \Upsilon^{-1} z_{x'}^\top = (\mathbf{1}\{a = a'\} - \frac{1}{p}) + (\mathbf{1}\{b = b'\} - \frac{1}{p}) + \frac{p+2}{p}$  and from summing  $z_x z_x^\top$  over  $\Omega$ . Sum the split count bounds for  $n, n_{\mathcal{D}}$ ; each target row has squared norm  $1 - 1/p$  and each class has at most  $p$  points. In block form, with  $\delta n^{\text{row}} = n^{\text{row}} - \rho p\mathbf{1}$  and  $\delta n^{\text{col}} = n^{\text{col}} - \rho p\mathbf{1}$ ,

$$\mathcal{K} - \rho p\mathbb{I} = \begin{pmatrix} \mathbf{P} \text{diag}(\delta n^{\text{row}})\mathbf{P} & \mathbf{P}\bar{\mathbf{M}}\mathbf{P} & \mathbf{P}\delta n^{\text{row}}/\sqrt{p} \\ \mathbf{P}\bar{\mathbf{M}}^\top\mathbf{P} & \mathbf{P} \text{diag}(\delta n^{\text{col}})\mathbf{P} & \mathbf{P}\delta n^{\text{col}}/\sqrt{p} \\ \delta n^{\text{row}\top}\mathbf{P}/\sqrt{p} & \delta n^{\text{col}\top}\mathbf{P}/\sqrt{p} & (n - \rho p^2)/p \end{pmatrix}. \quad (128)$$

On  $\mathcal{E}_{\text{des}}$  the upper-left  $2 \times 2$  block has norm at most  $D_* + t_*$ , the border vector has norm at most  $\sqrt{2}D_*$ , and  $|n - \rho p^2|/p \leq D_*$ ; so  $\|\mathcal{K} - \rho p\mathbb{I}\|_{\text{op}} \leq K_*$ , and  $K_* \leq \rho p/2$  gives  $\kappa_* \leq 1/2$ . The nonnegative matrix  $\mathbf{X}_S^\top \mathbf{X}_S$  has row sums  $2n_a^{\text{row}}$  and  $2n_b^{\text{col}}$ ;  $\mathbf{T}_S = \mathbf{Y}_S\mathbf{P}$  with  $\mathbf{Y}_S^\top \mathbf{Y}_S = \text{diag}(n_c)$ ; and  $n_{\text{max}} < p$  as  $D_* < p/5$ . The three blocks of  $\Psi$  are  $\mathbf{P}\bar{\mathbf{M}}^{\text{row}}\mathbf{P}$ ,  $\mathbf{P}\bar{\mathbf{M}}^{\text{col}}\mathbf{P}$  and  $\mathbf{1}^\top \bar{\mathbf{M}}^{\text{row}}\mathbf{P}/\sqrt{p}$ , each of norm at most  $t_*$ ; since  $\Psi\mathbf{1} = 0$ , its rank is at most  $p - 1$ . Finally  $\|\mathcal{Z}\|_{\text{op}} = \sqrt{p}$  and  $\|\mathcal{Z}_{Sc}\|^2 = c^\top \mathcal{K} c \geq s_0 \|c\|^2$ .  $\square$

## F.2.2 The dictionary coefficient of the rates

**Definition F.14** (Force covariance and pre-sampling coefficient). Write  $c_F = (c_F^{\text{L}}, c_F^{\text{R}})$  and define

$$\tau_c^2 = \frac{\mathbb{E}_{l,r}[\|\mathbf{P}c_F^{\text{L}}\|^2]}{p-1}, \quad \tau_0^2 = \frac{2\mathbb{E}_{l,r}[(\mathbf{1}^\top c_F^{\text{L}})^2]}{p}, \quad \mathbf{u}_0 = \frac{(\mathbf{1}, \mathbf{1})}{\sqrt{2p}}.$$

Let  $\Pi_o$  be the orthogonal projector onto  $\mathbf{1}^\perp \oplus \mathbf{1}^\perp$ ;  $\tau_c, \tau_0$  denote the nonnegative square roots.

Let  $\Sigma_F = \mathbb{E}_{l,r}[c_F c_F^\top]$ . The coefficient in (130) is

$$c_R(p) = \frac{\text{tr} \Sigma_F^{1/2}}{q^2 p(p-1)\sqrt{2p-1}}, \quad q = \frac{p+1}{4p}, \quad (129)$$

The comparison direction  $\mathbf{U}_F$  is the normalized supported Polar factor of the matrix whose row for neuron  $(l, r)$  is  $c_F(l, r)^\top/\sqrt{m}$ ; its corresponding row is  $\mathbf{U}_F(l, r)$ . Both  $c_R(p)$  and  $\mathbf{U}_F$  are fixed before the training split is sampled.

**Definition F.15** (Coherent-force quantities). For an interval mask  $I$ ,  $\text{ac}_I(t) = |I \cap (I+t)|$ ;  $I_l = \{a : l_a = 1\}$  and  $I_r = \{b : r_b = 1\}$ . The coherent sum is

$$E_{\text{coh}} = \sum_{a=0}^{(p-1)/2} \left[ a \left( \frac{p-1}{2} - a \right) + \frac{b_0(p-1)}{4p} \right].$$

The pre-sampling coefficient is (129). Let  $\Phi_F \in \mathbb{R}^{m \times 2p}$  have rows  $c_F(l, r)^\top / \sqrt{m}$ ; then  $\Phi_F^\top \Phi_F = \Sigma_F$  and  $U_F = \mathcal{N}_{\text{pol}}(\Phi_F)$ .

**Lemma F.16** (Own covariance and bounds on  $c_R(p)$ ). *For the force-covariance quantities of Definition F.14:*

(i)  $\Sigma_F = \tau_c^2 \Pi_o + \tau_0^2 \mathbf{u}_0 \mathbf{u}_0^\top$ ; hence  $\text{tr } \Sigma_F^{1/2} = 2(p-1)\tau_c + \tau_0$  and

$$c_R(p) = \frac{2(p-1)\tau_c + \tau_0}{q^2 p(p-1) \sqrt{2p-1}}.$$

(ii) For  $p \geq 64$ ,  $\tau_c > p^{3/2}/32$  and  $\tau_0 \geq p^2/24$ . Thus  $\Sigma_F$  has rank  $2p-1$  and range  $E_{\text{vis}}$ , its smallest positive eigenvalue is  $\min(\tau_c, \tau_0)^2 > p^3/1024$ , and  $U_F(l, r) = c_F(l, r)^\top (\Sigma_F^+)^{1/2} / \sqrt{m(2p-1)}$ .

(iii) For  $p \geq 64$ ,  $2/3 < c_R(p) \leq 4p/\sqrt{p^2-1} < 7$ .

*Proof.* Write  $(l * r)_c = \sum_a l_a r_{c-a}$ ,  $n_0 = (p-1)/2$ , and  $f_{\text{coh}}(a) = a(n_0 - a) + b_0 n_0 / (2p)$  on  $I_0$ , zero off  $I_0$ ; then  $E_{\text{coh}} = \mathbf{1}^\top f_{\text{coh}}$ . Since  $(\mathbf{T}^\top g)_c = (l * r)_c - b_0^2/p$ ,

$$c_F^L(a) = l_a \sum_{b \in I_r} [(l * r)_{a+b} - b_0^2/p] = l_a \left[ \sum_{v \in I_l} a c_{I_r}(a-v) - b_0^3/p \right].$$

Summing over  $a$  gives  $\mathbf{1}^\top c_F^L = \|\mathbf{P}(l * r)\|^2$ ; the right block is symmetric.

(i) Left translations translate  $\mathbf{P} c_F^L$  and fix  $c_F^R$  and the common coordinate sum. Averaging these translations makes their cross second moments zero. Simultaneous dilations act transitively on nonzero coordinate differences, so the left centered second moment is  $\tau_c^2 \mathbf{P}$ ; exchanging the masks gives the identical right block. Together with  $c_F \in E_{\text{vis}}$  and  $\mathbf{u}_0^\top c_F = \sqrt{2/p} \mathbf{1}^\top c_F^L$  this gives (i).

(ii) The slope event  $u'/u = \pm 1$  has probability  $2/(p-1)$ . Translations and simultaneous dilation reduce to  $I_l = I_r = I_0$ . Summing  $b_0 - |a-v|$  on  $I_0$  identifies the coherent block with the function  $f_{\text{coh}}$  defined above and gives

$$E_{\text{coh}} = \frac{(p^2-1)(p^2+3)}{48p} \in [p^3/48, p^3/48 + p/24].$$

Using  $\sum_{a=0}^{n_0} a^2(n_0-a)^2 = (n_0^5 - n_0)/30$  and  $(1-1/p)^5 \geq 1-5/p$ , for  $p \geq 64$  we obtain

$$\frac{\|\mathbf{P} f_{\text{coh}}\|^2}{p^5} \geq \frac{1-5/64}{960} - \frac{1}{60 \cdot 64^4} - \frac{(1+2/64^2)^2}{2304} > \frac{1}{2048}.$$

Restricting the two nonnegative second moments to this slope event gives  $\tau_c^2 > p^3/1024$  and  $\tau_0^2 \geq 4E_{\text{coh}}^2/[p(p-1)] \geq p^4/576$ . The rank, range and supported Polar formula follow from (i).

(iii) The lower bound follows from (ii), discarding  $\tau_0$  and using  $q \leq 65/256$  and  $q^2 < 21/320$ . For each neuron,  $\text{diag}(g)\mathbf{X}$  is the incidence matrix of the complete bipartite graph on its two masks, each of size  $b_0$ , so its squared operator norm is  $2b_0 = p+1$ . Since  $(\mathbf{T}\mathbf{T}^\top)^2 = p\mathbf{T}\mathbf{T}^\top$ , Lemma F.5(iv) gives

$$\text{tr } \Sigma_F \leq (p+1)p\mathbb{E} \left[ \left\| \mathbf{T}^\top g \right\|^2 \right] = q^2 p^2 (p^2 - 1).$$

Scalar Cauchy–Schwarz on the  $2p-1$  positive eigenvalues and division by (129) give  $c_R \leq \sqrt{p+1}/(q\sqrt{p-1}) = 4p/\sqrt{p^2-1} < 7$ .  $\square$

**Definition F.17** (Geometric clock and pre-sampling choices). Set  $Z = 9/50$ . The geometric rate parameters are

$$\begin{aligned} k &= \rho c_R(p), & a &= (1+kZ)^{-1/J}, & \nu_j &= a^j, \\ z_j &= (a^{-j} - 1)/k, & \xi &= k(z_{J-1} + z_J)/4. \end{aligned} \tag{130}$$

These are pre-sampling choices. The associated clock aliases are

$$\begin{aligned} \vartheta_j &= 1 - 4z_j, & g_- &= (1-a)/(2a), & g_+ &= (1-a)/2, \\ \mathbf{W}_j^{\text{ref}} &= \nu_j(\bar{\mathbf{W}}_0 + h z_j U_F), & \mathbf{H}_j^{\text{ref}} &= \mathbf{H}(\mathbf{W}_j^{\text{ref}}). \end{aligned}$$

The clock identities and signs are proved in Lemma F.18.

These construction parameters are separate from those of Definition E.3.

For the clock comparison put  $g_Z = 1 + kZ$  and  $a_J = (137/125)^{-1/J}$ .

**Lemma F.18** (Geometric clock and reference recursion). *Let  $J \geq 30$  and  $p \geq 64$ , with the pre-sampling rates and clock of Definition F.17 and (119).*

- (i)  $0 < a < a_J$ ,  $1 - s\lambda = a$  and  $s/h = (1 - a)/k$ . For  $0 \leq j \leq J$ ,  $\frac{125}{251} < \nu_j \leq 1$ ,  $\nu_j(1 + kz_j) = 1$ ,  $0 = z_0 < z_1 < \dots < z_J = Z = \frac{9}{50}$ , and  $\vartheta_j \geq \frac{7}{25}$ .
- (ii)  $z_{J-1} > \frac{1}{6}$ , and  $z_j - z_{j-1} > \frac{45}{502J} > \frac{1}{12J}$  for  $1 \leq j \leq J$ .
- (iii)  $0 < \xi < kZ/2$  and  $\nu_j \xi > \frac{11}{540}$  for  $0 \leq j \leq J$ . The quantity  $1 - \nu_j(1 + 2\xi)$  increases strictly with  $j$  and equals  $-g_-$  at  $J - 1$  and  $g_+$  at  $J$ , where both positive gaps exceed  $(1 - a_J)/2$ .
- (iv) The reference path has  $\mathbf{W}_0^{\text{ref}} = \bar{\mathbf{W}}_0$  and  $\mathbf{W}_{j+1}^{\text{ref}} = a\mathbf{W}_j^{\text{ref}} + s\mathbf{U}_F$  for  $0 \leq j < J$ .

*Proof.* **Clock identities and recursion.**

Lemma F.16(iii) gives  $8/15 < k < 28/5$ , hence  $137/125 < g_Z < 251/125$  and  $0 < a < a_J < 1$ . The prescribed rates imply  $1 - s\lambda = a$ ,  $s/h = (1 - a)/k$ , and

$$\nu_j = a^j, \quad \nu_j k z_j = 1 - a^j, \quad 1 + 2\xi = g_Z(1 + a)/2.$$

The first two identities give the increasing clock from 0 to  $Z$ ,  $\nu_j \geq \nu_J = g_Z^{-1} > 125/251$ , and  $\vartheta_j \geq 1 - 4Z = 7/25$ . They also rewrite the reference path as

$$\mathbf{W}_j^{\text{ref}} = a^j \bar{\mathbf{W}}_0 + (h/k)(1 - a^j)\mathbf{U}_F,$$

which proves its initial state and recursion because  $s = h(1 - a)/k$ .

**Positive geometric increments.**

The positive geometric increments increase with  $j$ , sum to  $Z$ , and have maximum-to-minimum ratio

$$\frac{z_J - z_{J-1}}{z_1 - z_0} = a^{1-J} = ag_Z < g_Z.$$

Comparison with their arithmetic mean  $Z/J$  therefore gives

$$\frac{Z}{Jg_Z} < z_j - z_{j-1} < \frac{g_Z Z}{J}.$$

The lower bound exceeds  $45/(502J) > 1/(12J)$ . Since  $J \geq 30$ , the upper bound gives

$$z_{J-1} > Z(1 - g_Z/J) > \frac{9}{50} \left(1 - \frac{251}{3750}\right) > \frac{1}{6}.$$

**Midpoint mixing and signs.**

The midpoint identity and  $a > g_Z^{-1}$ , since  $J > 1$ , give directly

$$\nu_j \xi \geq \nu_J \xi = \frac{1 + a - 2/g_Z}{4} > \frac{1 - g_Z^{-1}}{4} > \frac{3}{137} > \frac{11}{540}.$$

Also  $\xi < kZ/2$  because  $z_{J-1} < Z$ . Since  $\nu_j$  decreases, the signed quantity increases strictly. The same midpoint identity gives its endpoint values  $-(1 - a)/(2a) = -g_-$  and  $(1 - a)/2 = g_+$ . Both gaps exceed  $(1 - a_J)/2 > 0$  because  $a < a_J < 1$ .  $\square$

Thus the reference path is the decayed fixed-direction recursion of the learner with  $\mathcal{N}_{\text{pol}}(\mathbf{F}_j)$  replaced by  $\mathbf{U}_F$ . It is used only in the proof.

### F.2.3 Train-only calibration of the initial head

The clock is now fixed by Definition F.17. This subsection builds the head that will be compared along that clock. Lemma F.20 first controls held-out kernel predictions; Proposition F.22 then uses the two train-only calibrations to obtain the center-margin identity (136) with uniformly small error.

**Local scales and oscillation.** Write  $\omega = \sqrt{\log p/p}$  for the asymptotic split scale and  $\text{osc}(f) = \max_k f_k - \min_k f_k$  for  $f \in \mathbb{R}^p$ .

**Definition F.19** (Shifted inverses). Fix  $c_o \geq 1$  and put  $\mathcal{A}_{c_o} = c_o \Upsilon + \mathcal{K}$ ,  $\mathbf{C}^{(c_o)} = \mathcal{A}_{c_o}^{-1} \Psi$ ,  $\Lambda_1 = c_o + \rho p$  and  $\iota = \frac{1}{p(\rho p + c_o/(p+2))} - \frac{2}{p\Lambda_1}$ , and set

$$\begin{aligned} \epsilon_{\text{pred}} = & \frac{4(D_* + Q_*)}{\Lambda_1^2} + \frac{8|\iota|D_*^2}{\Lambda_1} \\ & + \frac{K_*\sqrt{2p+2\ell_\delta}}{\rho p\sqrt{p}(\rho p + c_o/(p+2))} + \frac{\sqrt{2}\kappa_*^2\sqrt{2p+2\ell_\delta}}{(1-\kappa_*)\rho p}. \end{aligned} \quad (131)$$

These depend on  $c_o$ ; without an argument they refer to  $c_o = 1$ , and  $\mathcal{A} = \mathcal{A}_1$ .

For the constant function use the coordinate vector  $d_1 = (0, 0, \sqrt{p})$ ,  $h_x = \mathbf{B}_{0,x\mathcal{S}}\mathbf{B}^{-1}\mathbf{1}_\mathcal{S}$  and  $\epsilon_h = \sqrt{2p}/[(p+2)s_0]$ . Thus  $\mathcal{Z}d_1 = \mathbf{1}_\Omega$ ; the scalar  $d_j$  below instead denotes an actual-state distance.

The held-out behavior of the canonical head and of the center refit is governed by the realized inverses of shifted kernels, in the coordinates of Lemma F.13.

For the shifted-kernel estimate, put

$$\Lambda_g = \rho p + \frac{c_o}{p+2}, \quad B_* = \sqrt{2p+2\ell_\delta}.$$

**Lemma F.20** (Kernel predictions). *For the shifted inverse quantities in Definition F.19, for every  $c_o \geq 1$  and  $x \in \mathcal{D}$ ,*

$$\mathbf{B}_{0,x\mathcal{S}}[\mathbf{B} + (c_o - 1)\mathbb{I}]^{-1} = z_x \mathcal{A}_{c_o}^{-1} \mathcal{Z}_\mathcal{S}^\top, \quad h_x = 1 - z_x \mathcal{A}^{-1} \Upsilon d_1. \quad (132)$$

Also  $|h_x - 1| \leq \epsilon_h$  and  $\mathcal{A}_{c_o} \succeq \mathcal{K} \succeq s_0\mathbb{I}$ . For  $x = (a, b) \in \mathcal{D}$ ,  $y = a + b$  and  $c \neq y$ ,

$$\left| z_x \mathbf{C}^{(c_o)}(e_y - e_c) + \frac{N_{xc}}{\Lambda_1} \right| \leq 2|\iota|D_* + 2\epsilon_{\text{pred}}. \quad (133)$$

*Proof.* For  $x \notin \mathcal{S}$  the identity part of  $\mathbf{B}_0$  does not contribute, so  $\mathbf{B}_{0,x\mathcal{S}} = z_x \Upsilon^{-1} \mathcal{Z}_\mathcal{S}^\top$ . Since

$$\mathcal{A}_{c_o} \Upsilon^{-1} \mathcal{Z}_\mathcal{S}^\top = \mathcal{Z}_\mathcal{S}^\top [\mathbf{B} + (c_o - 1)\mathbb{I}],$$

multiplication by both inverses gives the first formula of (132). At  $c_o = 1$ , use  $\mathbf{1}_\mathcal{S} = \mathcal{Z}_\mathcal{S} d_1$ ,  $\mathcal{K} = \mathcal{A} - \Upsilon$  and  $z_x d_1 = 1$  for the second. The bound on  $h_x$  follows from  $\mathcal{A} \succeq s_0\mathbb{I}$ ,  $\|z_x\| \leq \sqrt{2}$  and  $\|\Upsilon d_1\| = \sqrt{p}/(p+2)$ .

For  $x = (a, b) \in \mathcal{D}$ ,  $y = a + b$  and rival  $c$ , put

$$\begin{aligned} v &= e_y - e_c, & r_d &= M_{d,y-d} - M_{d,c-d}, & s_d &= M_{y-d,d} - M_{c-d,d}, & \delta_c &= n_y - n_c, \\ \psi &= \Psi v, & \bar{\mathcal{A}} &= c_o \Upsilon + \rho p \mathbb{I}, & \theta_0 &= \bar{\mathcal{A}}^{-1} \psi, \\ \mathbf{E} &= \mathcal{K} - \rho p \mathbb{I}, & \mathbf{B} &= \bar{\mathcal{A}}^{-1} \mathbf{E}. \end{aligned}$$

For these contrast vectors,  $|r_d|, |s_d| \leq 1$ ,  $|\delta_c| \leq 2D_*$  and  $\psi = \Psi v = (\mathbf{P}r, \mathbf{P}s, \delta_c/\sqrt{p})$  has norm at most  $B_*$ . For these  $\bar{\mathcal{A}}, \theta_0$ ,  $z_x \theta_0 = -N_{xc}/\Lambda_1 + \iota \delta_c$ , since  $x \notin \mathcal{S}$ . The perturbation satisfies  $\|\mathbf{B}\|_{\text{op}} \leq \kappa_* < 1$  and

$$\mathbf{C}^{(c_o)} v = \theta_0 - \mathbf{B} \theta_0 + \mathbf{B}^2 (\mathbb{I} + \mathbf{B})^{-1} \theta_0.$$

Since  $\|\theta_0\| \leq B_*/(\rho p)$ , the remainder and global coordinate of the first-order term contribute at most the fourth and third terms of (131), respectively. By (128), its row block before applying  $\bar{\mathcal{A}}^{-1}$  is  $\mathbf{P}[(\delta n^{\text{row}} \odot r + \bar{\mathbf{M}}s)/\Lambda_1 + \iota \delta n^{\text{row}} \delta_c]$ . Here  $\|\bar{\mathbf{M}}s\|_\infty \leq 2Q_*$  by subtracting two instances of (126). Applying  $\mathbf{P}$  at most doubles a coordinate bound; the row and column blocks therefore contribute at most  $4(D_* + 2Q_*)/\Lambda_1^2 + 8|\iota|D_*^2/\Lambda_1$ . This is bounded by twice the first two terms of (131). Adding the global and remainder bounds, then  $|\iota \delta_c| \leq 2|\iota|D_*$ , proves (133).  $\square$

**Definition F.21** (Train-only calibrations and center). For  $A > 0$  and  $n_c \geq 1$  for every  $c$ , choose the unique balance scalar  $s_{\text{bal}} > 0$  for which

$$\sum_c \frac{\sqrt{1 + 4(e^A - 1)n_c/s_{\text{bal}}} - 1}{2(e^A - 1)} = 1,$$

and put  $q_{\text{base},c} = (\sqrt{1 + 4(e^A - 1)n_c/s_{\text{bal}}} - 1)/(2(e^A - 1))$ . Existence and uniqueness are proved in Proposition F.22. Set

$$b_{\text{base},c} = \log q_{\text{base},c} - \frac{1}{p} \sum_d \log q_{\text{base},d}, \quad \alpha_c = \frac{1 + (e^A - 1)/p}{1 + (e^A - 1)q_{\text{base},c}}, \quad \hat{\mathbf{R}}_i = \alpha_{y_i}(e_{y_i} - q_{\text{base}}) \quad (i \in \mathcal{S}). \quad (134)$$

Let  $b_*$  be the unique centered training-CE offset for the premixed training logits

$$A[(1 - \xi)\mathbf{T}_S + \xi\mathbf{B}\hat{\mathbf{R}}] + (1 - \xi)\mathbf{1}b_{\text{base}}^\top;$$

Proposition F.22 establishes this train-only calibration. The center head is

$$\begin{aligned} \bar{\mathbf{V}}_0 &= \frac{A}{d}[(1 - \xi)\mathbf{B}^{-1}\mathbf{T}_S + \xi\hat{\mathbf{R}}]^\top \mathbf{H} \\ &\quad + \frac{1 - \xi}{d} b_{\text{base}}^\top \mathbf{1}_S^\top \mathbf{B}^{-1} \mathbf{H} + \frac{b_* w^\top}{h\mu^2}. \end{aligned} \quad (135)$$

Existence and the claimed properties of both calibrations are proved in Proposition F.22.

Define  $u_{0,xc}$  by the center-margin identity

$$(\bar{\mathbf{H}}_0 \bar{\mathbf{V}}_0^\top)_{xy_x} - (\bar{\mathbf{H}}_0 \bar{\mathbf{V}}_0^\top)_{xc} = A[M_x + \xi(2M_x - N_{xc})] + Au_{0,xc}, \quad (136)$$

and set  $B_{\text{cal}} := \text{osc}(b_*)$ ,  $e_0 := \max_{x \in \Omega, c \neq y_x} |u_{0,xc}|$  and  $E_{\text{cal}} := Ae_0$ . These are proof certificates of the already prescribed center, not inputs to either train-only minimizer, the rates or the radius.

**Local train-only calibration notation.** For a training-logit matrix  $\mathbf{F} \in \mathbb{R}^{n \times p}$  with labels  $y_i$  and  $n_c \geq 1$  for every class, set  $\mathcal{C} = \{b \in \mathbb{R}^p : \mathbf{1}^\top b = 0\}$  and

$$\mathcal{L}_F(b) = \frac{1}{n} \sum_i \left[ \log \sum_c e^{F_{ic} + b_c} - F_{iy_i} - b_{y_i} \right].$$

Write  $b^F = \arg \min_{b \in \mathcal{C}} \mathcal{L}_F(b)$  and  $\mathbf{R} = \mathbf{Y}_S - \text{softmax}(\mathbf{F} + \mathbf{1}(b^F)^\top)$ . Existence and uniqueness of the centered minimizer are proved below. The softmax Jacobian at a probability vector  $\pi$  is  $\text{diag}(\pi) - \pi\pi^\top$ .

For  $A > 0$ , put  $k_A = e^A - 1$  and  $q_c(t) = (\sqrt{1 + 4k_A n_c/t} - 1)/(2k_A)$  for  $t > 0$ . The balance scalar satisfies  $\sum_c q_c(s_{\text{bal}}) = 1$ , and  $q_{\text{base}} = q(s_{\text{bal}})$ . Set  $\varpi_c = (1 + k_A q_{\text{base},c})^{-1}$ . For the premixed head define

$$\begin{aligned} \mathbf{V}_{\text{can}} &= \frac{1}{d} (A\mathbf{T}_S + \mathbf{1}b_{\text{base}}^\top)^\top \mathbf{B}^{-1} \mathbf{H}, \\ \mathbf{V}_{\text{raw}} &= \frac{A}{d} \hat{\mathbf{R}}^\top \mathbf{H}, \quad \mathbf{V}_\xi = (1 - \xi)\mathbf{V}_{\text{can}} + \xi\mathbf{V}_{\text{raw}}. \end{aligned}$$

Put  $\mathbf{F}_x^\xi = \bar{\mathbf{H}}_{0,x} \mathbf{V}_\xi^\top$ , and use  $\mathbf{F} = \mathbf{F}_S^\xi$  and  $b_* = b^F$  for its calibration. The actual center residual is  $\mathbf{R}_{\text{cal}} = \mathbf{Y}_S - \text{softmax}(\mathbf{F}_S^\xi + \mathbf{1}b_*^\top)$ .

**Proposition F.22** (Train-only calibrated initial head). *For any training-logit table  $\mathbf{F}$  with  $n_c > 0$ , the centered minimizer  $b^F$  of  $\mathcal{L}_F$  exists uniquely and is characterized by*

$$\sum_{i \in \mathcal{S}} \text{softmax}(\mathbf{F}_i + b^F) = (n_c)_c.$$

*Its calibrated residual  $\mathbf{R}$  has  $\mathbf{R}^\top \mathbf{1}_n = 0$  and  $\mathbf{R}\mathbf{1}_p = 0$ .*

For  $\mathbf{F} = \mathbf{A}\mathbf{T}_S$ , the prescribed  $s_{\text{bal}}$  exists uniquely,  $q_{\text{base}} = q(s_{\text{bal}})$ , and  $b_{\text{base}} = b^{AT_S}$ . The normalized canonical residual satisfies  $\widehat{\mathbf{R}}^\top \mathbf{1}_n = 0$ ,  $\widehat{\mathbf{R}}\mathbf{1}_p = 0$ . Uniformly on  $\mathcal{E}_{\text{split}}$ ,

$$q_{\text{base},c} = 1/p + O(\omega/p), \quad \alpha_c = 1 + O(\omega),$$

$$\text{osc}(b_{\text{base}}), \quad \|b_{\text{base}}\|_\infty = O(\omega), \quad \left\| \widehat{\mathbf{R}} - \mathbf{T}_S \right\|_{\text{op}} = O(\sqrt{p}\omega).$$

Every canonical training margin is  $A + O(\omega)$ .

On  $\mathcal{E}_{\text{split}}$ , the prescribed head (135) has logits  $\mathbf{F}_x^\xi + b_*$  on every input, and  $\mathbf{R}_{\text{cal}}$  has both zero sums. The center certificates from Definition F.21 satisfy

$$B_{\text{cal}} = O(A\omega), \quad e_0 = O(\omega), \quad E_{\text{cal}} = O(A\omega).$$

Consequently every center training margin is at least  $A(1 - e_0) > 0$ , and every held-out input has a rival margin at most  $-A(\xi - e_0) < 0$  for all sufficiently large  $p$ .

*Proof.* We first establish the canonical calibration and its residual bounds. We then estimate the premixed margins and their class imbalance, and use the balance Jacobian to bound the final offset  $b_*$ .

(1) *Convexity and the canonical root.*

For  $\pi_i(b) = \text{softmax}(\mathbf{F}_i + b)$ , the Hessian quadratic form is  $n^{-1} \sum_i \sum_{c < d} \pi_{ic}(b) \pi_{id}(b) (v_c - v_d)^2$ ; it is positive on  $\mathcal{C}$ . Also  $\mathcal{L}_F(b) \geq \min_c (n_c/n) \text{osc}(b) + n^{-1} \sum_i (\min_c F_{ic} - F_{iy_i})$ , and  $\text{osc}(b) \geq \|b\|_2 / \sqrt{p}$  on  $\mathcal{C}$ , so the objective is coercive. Its gradient lies in  $\mathcal{C}$  and vanishes exactly at class-count balance, proving the first assertion.

For canonical logits, write  $q = q_{\text{base}}$ . The factor  $\varpi$  gives a class- $y$  row residual  $\varpi_y(e_y - q)$ , so balance is equivalent to  $n_c \varpi_c = s_{\text{bal}} q_c$  for  $s_{\text{bal}} = \sum_y n_y \varpi_y$ , or  $n_c = s_{\text{bal}} q_c (1 + k_A q_c)$ . Solving gives the stated  $q_c(t)$ ; each coordinate decreases continuously from  $+\infty$  to 0 as  $t$  increases from 0 to  $+\infty$ , so  $\sum_c q_c(t) = 1$  has exactly one root. Centering  $\log q$  therefore gives the unique calibrated offset and the residual formulas.

On  $\mathcal{E}_{\text{split}}$ , the canonical balance identity gives, for  $q_i \geq q_j$ ,

$$\frac{q_i}{q_j} \leq \frac{n_i}{n_j} = 1 + O(\omega).$$

Together with  $\sum_c q_c = 1$ , this gives  $q_c = 1/p + O(\omega/p)$  uniformly. Centering  $\log q$  gives  $\text{osc}(b_{\text{base}}), \|b_{\text{base}}\|_\infty = O(\omega)$ . Also

$$|\alpha_c - 1| = \frac{k_A |1/p - q_c|}{1 + k_A q_c} \leq \frac{|1/p - q_c|}{q_c} = O(\omega).$$

Finally, with  $\mathbf{P} = \mathbb{I} - \mathbf{1}\mathbf{1}^\top/p$ ,

$$\widehat{\mathbf{R}} - \mathbf{T}_S = \mathbf{Y}_S \{ (\text{diag}(\alpha) - \mathbb{I}) \mathbf{P} - \alpha(q - \mathbf{1}/p)^\top \}.$$

The bracket has operator norm  $O(\omega)$ , since  $\|\alpha\| = O(\sqrt{p})$  and  $\|q - \mathbf{1}/p\| = O(\omega/\sqrt{p})$ . Using  $\|\mathbf{Y}_S\|_{\text{op}} \leq \sqrt{n_{\text{max}}} = O(\sqrt{p})$  proves the residual bound; the margin claim follows from the offset oscillation.

(2) *Premixed pointwise margins and initial imbalance.* Put

$$D_x^\alpha = \sum_{b'} M_{ab'} \alpha_{a+b'} + \sum_{a'} M_{a'b} \alpha_{a'+b} \quad (x = (a, b)).$$

For the mixed logits  $\mathbf{F} = \mathbf{F}_S^\xi$ , write  $\pi_i = \pi_i(0)$  with  $\pi_i(b) = \text{softmax}(\mathbf{F}_i + b)$ , and set

$$\tau = \frac{e^{-A}}{1 + e^{-A} Q_\xi}, \quad t_{ic} = r_\xi^{2-N_{ic}} \quad (c \neq y_i), \quad \mathbf{g}_c = \sum_i \pi_{ic} - n_c,$$

$$w_{cd}(b) = \sum_i \pi_{ic}(b) \pi_{id}(b), \quad w_{cd} = w_{cd}(0), \quad w_0 = \min_{c \neq d} w_{cd}.$$

The event gives  $s_0 \asymp p$ . The canonical training margins are as above; on  $\mathcal{D}$ , Lemma F.20 gives the canonical margin estimate

$$O(A/p) + A[2|\iota|D_* + 2\epsilon_{\text{pred}}] + (1 + \epsilon_h)O(\omega) = O(A\omega).$$

For  $q = q_{\text{base}}$ , the constant block in  $\mathbf{B}_{0,x\mathcal{S}}\widehat{\mathbf{R}}$  vanishes by  $\widehat{\mathbf{R}}^\top \mathbf{1} = 0$ , and its remaining blocks give

$$(\mathbf{B}_{0,x\mathcal{S}}\widehat{\mathbf{R}})_c = M_x \alpha_y (\mathbf{1}\{c = y\} - q_c) + \alpha_c N_{xc} - q_c D_x^\alpha. \quad (137)$$

Hence its normalized  $y$ -versus- $c$  margin is  $3M_x \alpha_y - N_{xc} \alpha_c - (q_y - q_c)(M_x \alpha_y + D_x^\alpha) = 3M_x - N_{xc} + O(\omega)$ : here  $D_x^\alpha = O(p)$ ,  $|\alpha_c - 1| = O(\omega)$  and  $|q_y - q_c| = O(\omega/p)$ . Mixing the two components therefore gives, uniformly on the entire grid,

$$F_{xyx}^\xi - F_{xc}^\xi = A[M_x + \xi(2M_x - N_{xc})] + O(A\omega). \quad (138)$$

The quantities  $Q_\xi, \tau, t_{ic}$  have  $Q_\xi \asymp p$  and  $\tau \asymp p^{-1}$ . Equation (138), the arm-count event and  $\sum_{c \neq y_i} t_{ic} = Q_\xi + O(D_L)$  give  $\pi_{ic} = \tau t_{ic}[1 + O(A\omega)]$  for  $c \neq y_i$  and  $1 - \pi_{iy_i} = \tau Q_\xi[1 + O(A\omega)]$ . Thus the weighted-count event yields

$$\mathbf{g}_c = \tau(W_c - n_c Q_\xi) + O(\tau A\omega(W_c + n_c Q_\xi)) = O(pA\omega),$$

since  $|W_c - n_c Q_\xi| \leq \Delta_W = O(p^2\omega)$  and  $W_c + n_c Q_\xi = O(p^2)$ .

For  $c \neq d$ , the three mask factors defining  $\mathbf{H}_{cd}$  have sums  $n_c, n_d, n_d$ . The binary inequality  $uvw \geq u + v + w - 2$  therefore gives

$$\mathbf{H}_{cd} \geq n_c + 2n_d - 2p \geq (3\rho - 2)p - 3D_* \asymp p.$$

On these two-arm rows, the true probability is bounded below and the rival probability is  $\gtrsim e^{-A} \asymp p^{-1}$ , proving  $w_0 \gtrsim 1$ .

(3) *A uniform balance-Jacobian bound.* Apply part (1) to  $\mathbf{F} = \mathbf{F}_\mathcal{S}^\xi$ . Recall  $B_{\text{cal}} = \text{osc}(b_*)$ , and let  $i, k$  maximize/minimize  $b_*$ . For  $0 \leq t \leq 1$ ,  $\pi_{jc}(tb_*) \geq e^{-tB_{\text{cal}}} \pi_{jc}$ , hence  $w_{cd}(tb_*) \geq e^{-2tB_{\text{cal}}} w_{cd}$ . Integrating the balance Jacobian from 0 to  $b_*$  gives

$$\begin{aligned} \mathbf{g}_k - \mathbf{g}_i &= \int_0^1 \sum_d [w_{id}(tb_*)(b_{*,i} - b_{*,d}) \\ &\quad + w_{kd}(tb_*)(b_{*,d} - b_{*,k})] dt \\ &\geq pw_0 B_{\text{cal}} \int_0^1 e^{-2tB_{\text{cal}}} dt = \frac{pw_0}{2} (1 - e^{-2B_{\text{cal}}}). \end{aligned}$$

For each  $d$  the coefficient differences are nonnegative and sum to  $B_{\text{cal}}$ ; a diagonal edge has zero coefficient. The integral formula also holds at  $B_{\text{cal}} = 0$ . Since  $\|\mathbf{g}\|_\infty = O(pA\omega)$  and  $w_0 \gtrsim 1$ ,  $1 - e^{-2B_{\text{cal}}} = O(A\omega) = o(1)$ , so  $B_{\text{cal}} = O(A\omega)$ .

The exact identity  $\bar{\mathbf{H}}_0 w = h\mu^2 \mathbf{1}$  realizes this same offset at every input. The balance equations give both zero sums for  $\mathbf{R}_{\text{cal}}$ , and adding  $b_*$  to (138) changes any margin by at most  $\text{osc}(b_*) = O(A\omega)$ . Thus  $e_0 = O(\omega)$  and  $E_{\text{cal}} = O(A\omega)$ . On training rows,  $M_x = 1$  and  $N_{xc} \leq 2$ , so the ideal margin in (138) is at least  $A$ . For every held-out  $x = (a, b)$ , its operand row contains a training input on the split event, giving some  $c \neq y_x$  with  $N_{xc} \geq 1$ ; then its ideal margin is at most  $-A\xi$ . Since  $e_0 = o(1) < \xi$ , the two strict signs follow.  $\square$

## F.2.4 The reference direction and kernels

Proposition F.22 supplies the starting margins. We next quantify the direction that moves the features: Lemma F.24 bounds the reference coordinates and supplies a singular-value floor for the force anchor, Lemma F.25 controls perturbations of that direction, and Lemma F.27 identifies the reference kernel's rule component and leakage. These estimates feed both the actual-state comparison and the later refit calculation.

The prescribed center, rates and radius depend only on the fixed problem, fixed parameters and initial training data; full-grid auxiliary errors are proof-only. The comparison path  $\mathbf{W}_j^{\text{ref}} = \nu_j(\bar{\mathbf{W}}_0 + h z_j \mathbf{U}_F)$  is not an optimizer state. The proof allows any head sequence with  $\|\mathbf{V}_{j+1} - \mathbf{V}_j\|_F \leq \eta$ , always using the actual departing-state CE gradient in the feature update.

**Definition F.23** (Reference force normalization). Write  $\bar{c}_F = \left\| \mathbf{T}^\top g \right\|^2 / p$ ,  $w_F = \sqrt{mr_v} \mathbf{U}_F(l, r)^\top$ , and let  $\mathbf{J}_{\text{vis}}$  be an orthonormal basis matrix for  $E_{\text{vis}}$ . Let  $\mathbf{B}_F$  have rows  $\rho^2 c_F(l, r)^\top / \sqrt{m}$ . Set  $G_F = \max_{l, r, x} |\mathbf{x}_x^\top w_F| / \sqrt{r_v}$  and  $\sigma_p = \rho^2 p^{3/2} / 32$ .

**Lemma F.24** (Reference direction). *For  $p \geq 64$ , every neuron and coordinate satisfy*

$$(w_F)_i = ((c_F)_i - \bar{c}_F)/\tau_c + \bar{c}_F/\tau_0, \quad |(w_F)_i| < \frac{8}{3}\sqrt{p} + 1, \\ G_F < 4, \quad \sigma_{\min}(\mathbf{B}_F \mathbf{J}_{\text{vis}}) > \sigma_p.$$

Moreover  $\mathbf{B}_F = \mathbf{B}_F \mathbf{J}_{\text{vis}} \mathbf{J}_{\text{vis}}^\top$  and  $\mathcal{N}_{\text{pol}}(\mathbf{B}_F) = \mathbf{U}_F$ .

*Proof.* Lemma F.16 gives  $(\Sigma_F^+)^{1/2} = \mathbf{\Pi}_o/\tau_c + \mathbf{u}_o \mathbf{u}_o^\top/\tau_0$ ; the two blocks of  $c_F$  have mean  $\bar{c}_F$ , proving the formula. For an active left coordinate,  $(c_F)_a = \sum_{v \in I_l} \text{ac}_{I_r}(a - v) - b_0^3/p$ . The nonzero-shift autocorrelations of an affine half interval are the multiset  $\{1, 1, \dots, b_0 - 1, b_0 - 1\}$ . Summing its  $b_0 - 1$  smallest or largest elements gives

$$-p^2/16 + p/4 - 1/16 - 1/(8p) \leq (c_F)_a \leq p^2/16 + 1/16 - 1/(8p).$$

Inactive coordinates vanish. Cauchy–Schwarz applied to the two autocorrelation vectors gives  $0 \leq p\bar{c}_F \leq \sum_t \text{ac}_{I_o}(t)^2 - b_0^4/p = E_{\text{coh}}$ , hence  $0 \leq \bar{c}_F \leq p^2/48 + 1/24$ . Thus  $|(c_F)_i - \bar{c}_F| \leq p^2/12$ ; the bounds  $\tau_c \geq p^{3/2}/32$ ,  $\tau_0 \geq p^2/24$  prove the coordinate cap. Consequently  $G_F < (16/3 + 2/\sqrt{p})/\sqrt{2 - 1/p} \leq 134/(3\sqrt{127}) < 4$ . Finally every force row lies in  $E_{\text{vis}}$ , and  $\mathbf{B}_F^\top \mathbf{B}_F = \rho^4 \Sigma_F$  has restricted eigenvalues  $\rho^4 \tau_c^2, \rho^4 \tau_0^2$ . Its thin singular decomposition proves both the floor and the asserted Polar identity.  $\square$

In the following matrix lemma, all matrix names, SVD factors and eigenvalues are freshly quantified objects of compatible dimensions. They are unrelated to construction quantities denoted by the same letters.

**Lemma F.25** (Anchored Polar estimate). *If  $\mathbf{B}^\top \mathbf{B} \succeq \sigma^2 \mathbb{I}_r$ ,  $\sigma > 0$ , for  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times r}$  and  $\epsilon = \|\mathbf{A} - \mathbf{B}\|_F$ , then*

$$\|\mathcal{N}_{\text{pol}}(\mathbf{A}) - \mathcal{N}_{\text{pol}}(\mathbf{B})\|_F \leq \frac{2\epsilon}{\sqrt{r}\sigma}, \quad r - \text{rank } \mathbf{A} \leq \frac{\epsilon^2}{\sigma^2}.$$

For matrices with rows in  $E_{\text{vis}}$ , replace  $r$  by  $r_v$  and  $\mathbf{B}$  by  $\mathbf{B} \mathbf{J}_{\text{vis}}$  in the floor. In particular  $\mathbf{A} \neq 0$  when  $\epsilon < \sqrt{r}\sigma$ .

*Proof.* Let  $\mathbf{Q}_A, \mathbf{Q}_B$  be the unnormalized supported polar factors and  $j = \text{rank } \mathbf{A}$ . Projection onto  $\ker \mathbf{A}$  gives  $(r - j)\sigma^2 \leq \epsilon^2$ . If  $j = 0$ , then  $\epsilon \geq \sqrt{r}\sigma$  and the direction gap is one. Otherwise put  $a = \sqrt{r}$ ,  $b = \sqrt{j}$ ,  $D = \|\mathbf{Q}_B - \mathbf{Q}_A\|_F$  and  $d = \|\mathbf{Q}_B/a - \mathbf{Q}_A/b\|_F \leq 2$ . The SVD gives  $\langle \mathbf{A}, \mathbf{Q}_A - \mathbf{Q}_B \rangle \geq 0$ . Writing  $\mathbf{B} = \mathbf{Q}_B \mathbf{H}$ , use  $\mathbf{H} \succeq \sigma \mathbb{I}$  and  $\mathbb{I} - \text{sym}(\mathbf{Q}_B^\top \mathbf{Q}_A) \succeq 0$  to obtain

$$\epsilon D \geq \langle \mathbf{B} - \mathbf{A}, \mathbf{Q}_B - \mathbf{Q}_A \rangle \geq \frac{\sigma}{2}(D^2 + a^2 - b^2) \geq \frac{\sigma a D^2}{a + b},$$

where the last step uses  $D \leq a + b$  and  $a \geq b$ . If  $D = 0$  the claim is immediate; otherwise combine this with  $4D^2 - (a + b)^2 d^2 = (a - b)^2(4 - d^2) \geq 0$  to get  $d \leq 2D/(a + b) \leq 2\epsilon/(\sigma a)$ . Right multiplication by  $\mathbf{J}_{\text{vis}}$  preserves rank, nonzero singular values and Frobenius norms, proving the visible-space assertion.  $\square$

The Fourier characters are  $\chi_\omega(x) = p^{-1} e^{2\pi i \omega \cdot x/p} = \psi_{\omega_1, \omega_2}(x)$  from (13). Write  $\mathbf{L}_R = \{(k, k) : k \neq 0\}$  and  $\mathbf{P}_R = \mathbf{P}_{\mathbf{L}_R} = \mathbf{P}_{\text{add}} = \mathbf{T} \mathbf{T}^\top / p$  for the rule projector.

**Definition F.26** (Reference kernels). Set  $\mathbf{F}_0 = \bar{\mathbf{H}}_0/(hq)$  and  $(\mathbf{P}_F)_{x, (l, r)} = g_{(l, r)}(x) \mathbf{x}_x^\top \mathbf{U}_F(l, r)^\top / q$ ; write  $\mathbf{F}(z) = \mathbf{F}_0 + z \mathbf{P}_F$ ,  $\mathbf{C} = \mathbf{P}_F \mathbf{F}_0^\top$  and  $\mathbf{B}_2 = \mathbf{P}_F \mathbf{P}_F^\top$ . Put  $\mathbf{P}_0 = \mathbf{1} \mathbf{1}^\top / p^2$ , and let  $\mathbf{P}_{\text{ax}}$  be the sum of the orthogonal projectors onto the two punctured Fourier axes. Set  $\mathbf{P}_{\text{mix}} = \mathbb{I} - \mathbf{P}_0 - \mathbf{P}_{\text{ax}} - \mathbf{P}_R$  and  $\mathbf{P}_{\text{int}} = \mathbf{P}_R + \mathbf{P}_{\text{mix}}$ .

The scalar coefficients and trace definitions are

$$c_{\text{ax}} = \frac{4q^2 p(p-1)}{\sqrt{r_v}} \left[ \frac{1-\mu}{\tau_c} + \frac{\mu}{\tau_0} \right], \quad c_0 = (p+1)c_{\text{ax}}, \quad k_w = \frac{c_{\text{ax}}}{p+1}, \quad \mathbf{C}_{\text{mix}} = \mathbf{C} \mathbf{P}_{\text{mix}}, \\ R_R = \frac{\text{tr}(\mathbf{P}_R \mathbf{B}_2)}{p-1}, \quad \lambda_R(z) = 1 + 2c_R z + R_R z^2.$$

The identification of these coefficients with the kernel eigenvalues is proved in Lemma F.27.

Use the reference tables of Definition F.26 and put  $\mathbf{H}_U = q\mathbf{P}_F$ . For a punctured frequency line  $\mathbf{L} = \{k(u, v) : k \neq 0\}$  with  $(u, v) \neq (0, 0)$ , let  $\mathbf{P}_{\mathbf{L}}$  be the real orthogonal projector onto the corresponding character span. For a diagonal subset mask  $\mathbf{M}$ , put

$$\widehat{\mathbf{M}}(w) = p^{-1} \sum_x \mathbf{M}_{xx} e^{2\pi i w \cdot x/p}.$$

**Lemma F.27** (Cross kernel and leakage). *For  $p \geq 64$ , the symmetric matrices  $\mathbf{C}, \mathbf{B}_2$  act as scalars on each invariant subspace projected onto by  $\mathbf{P}_{\mathbf{L}}$  or  $\mathbf{P}_0$ , and*

$$\begin{aligned} \mathbf{C} &= c_R \mathbf{P}_R + c_{\text{ax}} \mathbf{P}_{\text{ax}} + c_0 \mathbf{P}_0 + \mathbf{C}_{\text{mix}}, \quad c_{\text{ax}} \in (0, 4), \quad \mathbf{H}_U w = \mu^2 k_w \mathbf{1}, \\ \|\mathbf{P}_F\|_F, \|\mathbf{C}_{\text{mix}}\|_F &\leq \sqrt{2p}/q, \quad \|\mathbf{C}_{\text{mix}}\|_{\text{op}} < 8. \end{aligned}$$

On the rule line,  $\mathbf{B}_2$  has eigenvalue  $R_R \geq c_R^2$ , and  $\mathbf{F}(z)\mathbf{F}(z)^\top$  has eigenvalue  $\lambda_R(z) \geq (1 + c_R z)^2$ . For a subset mask  $\mathbf{M}$  with  $n'$  selected points and any  $\mathbf{E} = \sum_{\mathbf{L} \neq \mathbf{L}_R} e_{\mathbf{L}} \mathbf{P}_{\mathbf{L}}$ ,

$$\|\mathbf{E} \mathbf{M} \mathbf{T}\|_F^2 \leq \frac{n'}{p(p-1)} \|\mathbf{E}\|_F^2.$$

*Proof.* The entries of  $\mathbf{C}$  and  $\mathbf{B}_2$  are respectively  $\mathbb{E}[gg' \mathbf{x}_x^\top w_F] / (q^2 \sqrt{r_v})$  and  $\mathbb{E}[gg' \mathbf{x}_x^\top w_F \mathbf{x}_x^\top w_F] / (q^2 r_v)$ . Translation of either mask translates its force block; simultaneous dilation and exchange of the masks dilate and exchange the blocks. The dictionary law is invariant under these operations. Thus both kernels are translation invariant, symmetric (use dilation by  $-1$ ), and constant on frequency lines, with equal eigenvalues on the axes. The rule eigenvalue follows from

$$\text{tr}(\mathbf{T}^\top \mathbf{C} \mathbf{T}) = \frac{\mathbb{E}[c_F^\top (\boldsymbol{\Sigma}_F^+)^{1/2} c_F]}{q^2 \sqrt{r_v}} = c_R p(p-1).$$

On active entries  $\mathbf{P}_{\text{ax}} g = 2\mu(1 - \mu)$  and  $(l, r)^\top w_F = 2p\bar{c}_F[(1 - \mu)/\tau_c + \mu/\tau_0]$ . Using  $\mathbb{E}[p\bar{c}_F] = q^2 p(p-1)$  in  $\text{tr}(\mathbf{P}_{\text{ax}} \mathbf{C}) = 2(p-1)c_{\text{ax}}$  gives the formula. Its upper bound follows from  $c_{\text{ax}} < 4q^2(16 + 24/\sqrt{p})/\sqrt{2-1/p} < 3.5$  at  $p \geq 64$ . Averaging  $\mathbf{H}_U w$  and  $\mathbf{C} \mathbf{1}$  over the grid gives the two constant-mode identities. Also  $\|\mathbf{P}_F\|_F \leq \|\mathbf{X} \mathbf{U}_F^\top\|_F / q \leq \sqrt{2p}/q$ . Since  $\mathbf{F}_0 \mathbf{F}_0^\top = \mathbb{I}$  on  $\mathbf{P}_{\text{mix}}$ ,  $\|\mathbf{C}_{\text{mix}}\|_F \leq \|\mathbf{P}_F\|_F$ ; each line has dimension  $p-1$ , so its operator norm is at most  $\sqrt{2p/(q^2(p-1))} < 8$ . For a unit rule vector  $v$ ,  $\|\mathbf{F}_0^\top v\| = 1$  and  $c_R = \langle \mathbf{P}_F^\top v, \mathbf{F}_0^\top v \rangle \leq \sqrt{R_R}$ .

For the mask Fourier coefficients defined above, Parseval gives  $\sum_w |\widehat{\mathbf{M}}(w)|^2 = n'$ . On a nonrule line with generator  $u$ , the map  $(k, l) \mapsto (l, l) - ku$  is injective for  $k, l \neq 0$ . Thus  $\|\mathbf{P}_{\mathbf{L}} \mathbf{M} \mathbf{P}_R\|_F^2 \leq n'/p^2$ . Orthogonality of the line ranges and  $\mathbf{T} \mathbf{T}^\top = p \mathbf{P}_R$  give the claim.  $\square$

### F.2.5 Uniform finite-horizon comparison

All theorem parameters, including  $J$ , are fixed before  $p$  and the split. The constants used for the exact initial ball and the complete refit are fixed next, before any actual-state estimates. The radius (141) is chosen using the initial refit's signed reserve (140) and the perturbation budget (139). Its positivity is checked at the start of Lemma F.30; its protection of initial refit failure is verified in Section F.4.

**Definition F.28** (Complete refit and reference-kernel budgets). For any feature table  $\mathbf{U}$  with  $\|\mathbf{U}_S\|_F > 0$ , set  $q_U = \gamma \|\mathbf{U}_S\|_F^2$ ,  $\mathbf{Y}_U = \mathbf{U}_{\mathcal{D}} \mathbf{U}_S^\top (\mathbf{U}_S \mathbf{U}_S^\top + q_U \mathbb{I})^{-1} \mathbf{T}_S$  and  $E_{\text{obs}}(\mathbf{U}) = \|\mathbf{Y}_U - \mathbf{T}_{\mathcal{D}}\|_F^2 / N_{\mathcal{D}}^2$ .

For  $0 \leq z \leq 1/5$ , put  $\mathbf{B}(z) = \mathbf{F}(z)\mathbf{F}(z)^\top$  and  $\theta_R(z) = \lambda_R(z) - 1$ . Define

$$\begin{aligned} \mathbf{E}_z &= \mathbf{P}_{\text{mix}}(\mathbf{B}(z) - \mathbb{I})\mathbf{P}_{\text{mix}}, \\ a_{\min} &= 1/5, \quad q_{\min} = \kappa \rho / 200, \\ C_{\max} &= \max \left\{ 4/\rho, (1 + 4/\rho) \frac{169}{25q_{\min}} \right\}. \end{aligned}$$

Put  $\Gamma(z) = \rho\theta_R(z)/(1 + q_{F(z)} + \rho\theta_R(z))$ .

$$\mathbf{P}_p(e) = \mathbf{A}_p e + \mathbf{B}_p e^2, \quad \mathbf{A}_p = \frac{\sqrt{p}(1 + 3\sqrt{C_{\max}})}{2\sqrt{q_{\min}N_{\mathcal{D}}}} + \frac{2\sqrt{C_{\max}}\gamma(9/5)pN_+}{q_{\min}N_{\mathcal{D}}}, \quad \mathbf{B}_p = \frac{\sqrt{C_{\max}}\gamma N_+}{2q_{\min}N_{\mathcal{D}}}. \quad (139)$$

Put  $\lambda_{\text{ob}} = 1 + 4\gamma n$ , let  $\Lambda_1, \iota, \epsilon_{\text{pred}}$  be those of Definition F.19 at  $c_o = \lambda_{\text{ob}}$ , and set

$$\mu_{\text{obs0}} = \frac{2}{\Lambda_1} - 2|\iota|D_* - 2\epsilon_{\text{pred}}. \quad (140)$$

For the Fourier transfer, put  $A_0 = 1/(3p + 1)$ .

**Definition F.29** (Prescribed initial ball). Define, from the initial data and fixed parameters,

$$r_{\text{new}} = r_6 := \frac{\mu_{\text{obs0}}}{2\sqrt{2}N_+ \mathbf{P}_p(\sqrt{2p}/(hq))}. \quad (141)$$

The corresponding initial region is  $\mathcal{R}_{p,\mathcal{S}}^{\text{random}} = \{(\mathbf{W}, \mathbf{V}) : \|(\mathbf{W} - \bar{\mathbf{W}}_0, \mathbf{V} - \bar{\mathbf{V}}_0)\|_F < r_{\text{new}}\}$ . Its positivity and eventual validity are proved in Proposition F.36.

**Actual-state comparison notation.** Write  $d_j = \|\mathbf{W}_j - \mathbf{W}_j^{\text{ref}}\|_F / h$  for the normalized feature distance and  $r = \|(\mathbf{W}_0 - \bar{\mathbf{W}}_0, \mathbf{V}_0 - \bar{\mathbf{V}}_0)\|_F$  for the joint initial distance. For this comparison put  $\phi = p^{-1/4} + A\omega$  and

$$\begin{aligned} \mathbf{U} &= (1 - \xi)\mathbf{B}^{-1}(\mathbf{T}_S + \mathbf{1}b_{\text{base}}^\top/A) + \xi\hat{\mathbf{R}}, \\ s(g) &= \mathbf{U}^\top g_S + b_*/(4A), \quad \mathbf{C}_T(g) = \mathbf{X}_S^\top \text{diag}(g_S)\mathbf{T}_S. \end{aligned}$$

For each time  $j$ , set

$$\begin{aligned} t_j &= e^{-A\nu_j\xi}, \quad Q_j = (p-1)(\rho + (1-\rho)t_j)^2, \\ \mathbf{w}_j &= \frac{e^{-A}Q_j}{1 + e^{-A}Q_j}, \quad \mathbf{k}_j = \frac{A\mathbf{w}_j}{hq^2n}, \quad \delta_j^F = \frac{\|\mathbf{F}_j - \mathbf{k}_j\mathbf{B}_F\|_F}{\mathbf{k}_j\sigma_p\sqrt{r_v}}. \end{aligned}$$

**Lemma F.30** (Two-output finite-horizon comparison). *For all sufficiently large  $p$ , the train-only center is defined on  $\mathcal{E}_{\text{split}}$  and its prescribed radius satisfies  $0 < r_6 < 1$ . Initialize at joint distance  $r < r_6$  from the center and use the feature update in (119) with any head sequence whose increments have norm at most the pre-sampling budget  $\eta = \theta/(hp)$ . Uniformly over these masks, initial states and head sequences, for every fixed  $\ell \geq 0$ ,*

$$A^\ell \max_{j \leq J} d_j \rightarrow 0, \quad A^\ell \max_{j < J} \delta_j^F \rightarrow 0.$$

Moreover  $L_j = L_* + o(1)$ ,  $\text{Acc}_{\mathcal{D}}(j) \rightarrow 0$  for  $j < J$ ,  $\text{Acc}_{\mathcal{D}}(J) \rightarrow 1$ , and  $\text{Acc}_{\mathcal{D}}(0) = 0$  exactly for any fixed tie rule. Every training margin satisfies, eventually,

$$(\mathbf{H}_j \mathbf{V}_j^\top)_{i,y_i} - (\mathbf{H}_j \mathbf{V}_j^\top)_{i,c} \geq \nu_j A/25 \quad (i \in \mathcal{S}, c \neq y_i, 0 \leq j \leq J). \quad (142)$$

*Proof.* All orders and limits below are uniform over the family in the statement. Their constants may depend only on the fixed theorem parameters. The proof has three stages. The static and source estimates control the calibrated head and the reference margins. The departing-state estimate then closes a finite induction for  $d_j$  and  $\delta_j^F$ . Finally, the resulting logit error gives mean loss and held-out accuracy, while the rank estimate gives the separate pointwise training bound (142).

**Static event, center and exact radius.**

The event bounds give  $t_* = O(\sqrt{p})$ ,  $D_*, Q_*, K_*, D_H, D_L = O(p\omega)$ ,  $D_W = O(p\sqrt{\log p})$  and  $t_f, t_g = O(p^{7/4})$ . Thus  $n, n_{\mathcal{D}} \asymp p^2$ ,  $s_0, \mathbf{H}_{\min} \asymp p$  and  $Q_\xi - D_L, Q_j - \mathbf{w}_j D_L \gtrsim p$ . In particular  $K_* \leq \rho p/2$ ,  $D_* < p/5$  and  $D_L < \rho^2(p-1)$  eventually, and every class has positive training count. These facts precede any use of a calibrated inverse. The count bounds give  $\Delta_W = O(p^2\omega)$ . Proposition F.22, using only these initial-data

bounds and the kernel identities, gives  $\|b_{\text{base}}\|_\infty, \|\alpha - \mathbf{1}\|_\infty = O(\omega)$ ,  $B_{\text{cal}}, E_{\text{cal}} = O(A\omega)$  and  $e_0 = O(\omega) < \xi$ . These certificates establish the prescribed center's signs independently of any trajectory; take also  $p \geq 2048^2$ .

Uniformly for  $1 \leq c_o \leq 2$ , the kernel formulas give  $\Lambda_1 = \rho p + O(1)$ ,  $|\iota| = O(p^{-2})$  and  $\epsilon_{\text{pred}} = O(p^{-3/2} \log p)$ . Since  $1 < \lambda_{\text{ob}} = 1 + \kappa n / [p(p-1)] < 2$ ,

$$p\mu_{\text{obs}0} \longrightarrow 2/\rho, \quad \mathbf{A}_p = O(p^{-1/2}), \quad \mathbf{B}_p = O(p^{-2}).$$

Thus  $r_6 > 0$ . The first positive summand of  $\mathbf{A}_p$ ,  $\mu_{\text{obs}0} \leq 2/(\rho p)$ , and  $N_{\mathcal{D}}^-/N_+ = \sqrt{1 - \rho - D_*/p}$  give

$$r_6 \leq \frac{hq\sqrt{q_{\min}(1-\rho)}}{\rho(1+3\sqrt{C_{\max}})p^2} = O(p^{-2}) < 1.$$

### Canonical and centered source bounds.

Put

$$\begin{aligned} d_{\text{atom}} &= (\mu(l - \mu\mathbf{1}), \mu(r - \mu\mathbf{1}), \sqrt{p}\mu^2), \\ e_{\text{can}} &= \mathbf{T}_S^\top \mathbf{B}^{-1} g_S - \mathbf{T}_S^\top f_S. \end{aligned}$$

The atom coordinates give  $g = f + \mathcal{Z}d_{\text{atom}}$ . The additive inverse identity gives

$$e_{\text{can}} = \Psi^\top \mathcal{A}^{-1}(\Upsilon d_{\text{atom}} - \mathcal{Z}_S^\top f_S).$$

Here  $\|\Psi\|_{\text{op}} = O(\sqrt{p})$ ,  $\|\Upsilon d_{\text{atom}}\| = O(\sqrt{p})$  and  $\|\mathcal{Z}_S^\top f_S\|^2 \leq t_*^2 p [2\mu(1-\mu) + \mu^2(1-\mu)^2] = O(p^2)$ . Thus  $\sup_g \|e_{\text{can}}\| = O(\sqrt{p})$ . Also  $\mathbf{1}^\top \mathbf{B}^{-1} g_S = d_1^\top \Upsilon \mathcal{A}^{-1} \mathcal{Z}_S^\top g_S = O(1)$ , using  $\|\Upsilon d_1\| = \sqrt{p}/(p+2)$  and  $\|g_S\| \leq b_0$ .

### Calibrated coefficients and force comparison.

Set

$$v_g = \mathbf{T}_S^\top g_S - a_g \hat{n}, \quad \mathcal{J}_\alpha = (\mathbb{I} - q_{\text{base}} \mathbf{1}^\top) \text{diag}(\alpha), \quad \mathbf{E} = \hat{\mathbf{R}} - \mathbf{T}_S.$$

The centered coefficients give  $\hat{\mathbf{R}}^\top g_S = \mathcal{J}_\alpha v_g$  and  $\mathbf{C}_T v_g = c_g - a_g c_1$ . Since  $\mathbf{1}^\top v_g = 0$ ,  $\|\mathcal{J}_\alpha - \mathbf{P}\|_{\text{op}} = O(\omega)$  implies  $\|(\mathcal{J}_\alpha - \mathbb{I})v_g\| \leq O(\omega) \|v_g\|$ . Dictionary covariance and  $\|\hat{n}\| \leq 2D_*\sqrt{p}$  give

$$\|\mathbf{T}_S^\top g_S\|_{L^2}^2 \leq q^2(3n + 4pD_*^2), \quad \|v_g\|_{L^2} = O(p\sqrt{\log p}).$$

Expanding  $\mathbf{C}_T s - \rho^2 c_F$ , the canonical and raw force errors are bounded by  $t_f + t_g = O(p^{7/4})$ . The remaining terms are  $\mathbf{C}_T$  applied to  $(1-\xi)e_{\text{can}}$ ,  $\xi(\mathcal{J}_\alpha - \mathbb{I})v_g$  and the two offsets. Their  $L^2$  norms are respectively  $O(\sqrt{p})$ ,  $O(\sqrt{p} \log p)$  and  $O(\sqrt{\log p})$ . Since  $\|\mathbf{C}_T\|_{\text{op}} \leq 2^{1/2} n_{\text{max}} = O(p)$ , their total is  $O(p^{3/2} \log p) = O(p^{7/4})$ .

### One head Gram budget.

Put  $\mathbf{S} = (1-\xi)\mathbf{B}^{-1}\mathbf{T}_S + \xi\hat{\mathbf{R}}$  and, in this paragraph,  $v = e_y - e_c$  for  $y \neq c$ . The maps  $\mathbf{I}_{\text{row}}$  and  $\mathbf{I}_{\text{col}}$  send a train vector to its sums over each first and second operand, respectively;  $\mathbf{I}_{\text{op}}$  denotes either map, with  $\text{op} \in \{\text{row}, \text{col}\}$ . Define the strict-training contrast budget

$$K_p = \max_{y \neq c} \frac{hq}{A} \|\bar{\mathbf{V}}_{0,y} - \bar{\mathbf{V}}_{0,c}\|_2. \quad (143)$$

The coefficient formulas give  $\|s(g)\|_\infty = O(p)$ , since  $|(\mathbf{T}_S^\top f)_c| \leq n_{\text{max}} + n/p$  and  $|\hat{\mathbf{R}}_c^\top g_S| \leq (\max_d \alpha_d) n_{\text{max}} = O(p)$ . Indeed  $g \in \{0, 1\}$ , and the zero-sum residual column has positive mass  $n_c \alpha_c (1 - q_{\text{base},c}) \leq n_c \alpha_c$ . The canonical factorization gives  $\|\mathcal{J}_\alpha^\top - \mathbf{P}\|_{\text{op}} = O(\omega)$ . The exact factorization  $\hat{\mathbf{R}} = \mathbf{Y}_S \mathcal{J}_\alpha^\top$ , balance and the centered-mask event imply, for either operand map,

$$\mathbf{I}_{\text{op}} \hat{\mathbf{R}} = \mathbf{P} \bar{\mathbf{M}}^{\text{op}} \mathcal{J}_\alpha^\top, \quad \|\mathbf{I}_{\text{op}} \hat{\mathbf{R}}\|_{\text{op}} = O(\sqrt{p}).$$

For the mixed coefficient  $\mathbf{S}$ , the dictionary Gram and  $\hat{\mathbf{R}}^\top \mathbf{1} = 0$  give the exact identity

$$\begin{aligned} \mathbf{S}^\top \mathbf{B} \mathbf{S} &= (\mathbf{T}_\mathcal{S} + \xi \mathbf{E})^\top (\mathbf{T}_\mathcal{S} + \xi \mathbf{E}) + \xi^2 \sum_{\text{op} \in \{\text{row}, \text{col}\}} (\mathbf{I}_{\text{op}} \hat{\mathbf{R}})^\top (\mathbf{I}_{\text{op}} \hat{\mathbf{R}}) \\ &\quad - (1 - \xi)^2 \mathbf{T}_\mathcal{S}^\top (\mathbb{I} - \mathbf{B}^{-1}) \mathbf{T}_\mathcal{S}. \end{aligned} \quad (144)$$

The last term is negative semidefinite; the others have operator norm  $O(p)$ , using  $\|\mathbf{E}\|_{\text{op}} = O(\sqrt{p\omega})$ . For  $v = e_y - e_c$ , every coordinate of  $\mathbf{I}_{\text{op}} \mathbf{T}_\mathcal{S} v$  is in  $\{-1, 0, 1\}$ . The preceding operand identity and count bounds therefore give

$$\|\mathbf{I}_{\text{op}} \hat{\mathbf{R}} v\| \leq \sqrt{p} + O(\sqrt{\log p}), \quad \|(\mathbf{T}_\mathcal{S} + \xi \mathbf{E})v\|^2 = 2\rho p + O(\sqrt{p \log p}).$$

Thus  $v^\top \mathbf{S}^\top \mathbf{B} \mathbf{S} v \leq 2p(\rho + \xi^2) + O(\sqrt{p \log p})$ .

Since  $\mathbf{B} \succeq \mathbb{I} + \mathbf{1}\mathbf{1}^\top$ , the two head offsets satisfy

$$\left\| \frac{hq}{A} \bar{\mathbf{V}}_0 - \frac{\mathbf{S}^\top \mathbf{H}}{hq} \right\|_{\text{op}} \leq \frac{(1 - \xi) \|b_{\text{base}}\|}{A} + \frac{q \|b_*\|}{A\mu^2} = O(\sqrt{p\omega}).$$

For a contrast  $v$ , the same bound uses the two oscillations instead of the vector norms and is  $O(\omega)$ . Since  $\mathbf{H}\mathbf{H}^\top = h^2 q^2 \mathbf{B}$ , these give

$$\|\bar{\mathbf{V}}_0\|_{\text{op}} = O(A\sqrt{p}), \quad K_p^2 \leq 2p(\rho + \xi^2) + O(\sqrt{p \log p}). \quad (145)$$

The Gram identity (144) therefore supplies two bounds: the head operator norm for the departing-state comparison, and the contrast budget  $K_p$  for the final strict-training argument. The center head column is  $As(g)/(hq^2\sqrt{m})$ , so its entrywise bound is  $O(Ap/\sqrt{m})$ ; this is separate from its operator norm.

#### Reference features and their table error.

In this paragraph set  $v = \mathcal{A}^{-1} \Upsilon d_1$ . The initial preactivation is  $h/\sqrt{m}$  on an active entry and at most  $-h/\sqrt{m}$  otherwise. Lemma F.24 bounds the increment by  $4hz/\sqrt{m}$ . Thus for  $0 \leq z \leq 1/5$  and  $\nu > 0$ ,  $\nu(\mathbf{W}_0 + hz\mathbf{U}_F)$  retains its initial signs with modulus at least  $\nu h(1 - 4z)/\sqrt{m}$ , and its features are  $\nu hq\mathbf{F}(z)$ . Since  $\|\mathbf{F}_0\|_{\text{op}} = p + 1$  and  $\|\mathbf{P}_F\|_F = O(\sqrt{p})$ ,  $\|\mathbf{H}_j^{\text{ref}}\|_{\text{op}} = O(p)$ . The coefficient errors satisfy the exact identities

$$\mathbf{U} - \mathbf{T}_\mathcal{S} = (1 - \xi) \mathbf{Z}_\mathcal{S} (v b_{\text{base}}^\top / A - \mathcal{A}^{-1} \Psi) + \xi \mathbf{E}.$$

The constant-mode identity is

$$\mathbf{1}^\top \mathbf{U} = (1 - \xi) [v^\top \Psi + (\mathbf{1}^\top \mathbf{Z}_\mathcal{S} v) b_{\text{base}}^\top / A].$$

Together these give  $\|\mathbf{U} - \mathbf{T}_\mathcal{S}\|_F = O(p\omega)$ ,  $\|\mathbf{T}_\mathcal{S}^\top \mathbf{U} / p - \rho \mathbf{P}\|_{\text{op}} = O(\omega)$  and  $\|\mathbf{1}^\top \mathbf{U}\| = O(p^{-1} + \sqrt{\log p}/A)$ .

Here use  $\|v\| = O(p^{-3/2})$ ,  $0 \leq \mathbf{1}^\top \mathbf{Z}_\mathcal{S} v < 1$ , and the calibration's  $\|\mathbf{E}\|_{\text{op}} = O(\sqrt{p\omega})$ .

#### Reference margins.

Let  $\tilde{\mathbf{U}}$  be  $\mathbf{U}$  extended by zero off  $\mathcal{S}$ , and put

$$\Delta_j = z_j (C\tilde{\mathbf{U}} - k\mathbf{T}), \quad u_{j,xc} = u_{0,xc} + z_j k_w (b_{*,y_x} - b_{*,c}) / A.$$

For  $\Delta_j$ , the rule part is  $O(p\omega)$  in Frobenius norm. The axes and constant-mode bounds are  $c_{\text{ax}}[\sqrt{2n/p} + \|\mathbf{U} - \mathbf{T}_\mathcal{S}\|_F]$  and  $c_0 \|\mathbf{1}^\top \mathbf{U}\| / p$ . The mixed part is at most  $\sqrt{2n/(q^2(p-1))} + 8 \|\mathbf{U} - \mathbf{T}_\mathcal{S}\|_F$  by Lemma F.27. Thus  $\|\Delta_j\|_F = O(p\omega)$ . The reference logits are exactly

$$\nu_j [\bar{\mathbf{H}}_0 \bar{\mathbf{V}}_0^\top + Az_j k \mathbf{T} + A \Delta_j + z_j k_w \mathbf{1} b_*^\top].$$

Use the center margins (136),  $k_w = c_{\text{ax}}/(p+1)$  and  $\text{osc}(b_*) \leq B_{\text{cal}}$ . Every reference margin is therefore

$$\nu_j A \{M_x + \xi(2M_x - N_{xc}) + kz_j + u_{j,xc} + \Delta_{j,xy_x} - \Delta_{j,xc}\}, \quad (146)$$

with  $\sup |u_{j,xc}| = O(\omega)$ ,  $\Delta_0 = 0$  and  $|u_{0,xc}| \leq e_0$ . Since  $\nu_j(1 + kz_j) = 1$ , the ideal training margin is  $A[1 + \nu_j\xi(2 - N_{xc})] \geq A$ .

### Departing-state comparison.

We now compare one actual departing state with the reference margin formula (146). The resulting force error will be the input to the finite induction below.

At a departing state  $(\mathbf{W}, \mathbf{V})$  set  $d = \|\mathbf{W} - \mathbf{W}_j^{\text{ref}}\|_F / h$  and  $v = \|\mathbf{V} - \bar{\mathbf{V}}_0\|_F$ . The ideal rival vector  $a_i$  has true entry zero and rival entries  $t_j^{2-N_{ic}}$ ; put  $\zeta_i = \mathbf{1}^\top a_i - Q_j$ . Let  $\mathbf{R}_i^*$  be the CE residual of the ideal margins  $A[1 + \nu_j\xi(2 - N_{ic})]$ . Write  $\mathbf{D}, \mathbf{D}^0$  for the actual and reference training gates and define

$$\mathbf{F}^{\text{anc}} = \mathbf{w}_j n^{-1} ((\mathbf{T}_S \bar{\mathbf{V}}_0) \odot \mathbf{D}^0)^\top \mathbf{X}_S, \quad b = \nu_j h \vartheta_j / \sqrt{m}.$$

The clock and sampled design give  $\mathbf{w}_j \asymp 1$ ,  $\mathbf{k}_j \asymp A/p^2$  and  $\mathbf{k}_j \sigma_p \sqrt{r_v} \gtrsim A$ . At any departing state with  $d \leq 1$ , the head distance obeys  $v \leq r + J\eta = O(p^{-1})$ . ReLU contraction and  $\|\mathbf{X}_E\|_{\text{op}} \leq \sqrt{2p}$  give, without any actual-gate assumption, on every subset  $E$ ,

$$\|\mathbf{H}_E \mathbf{V}^\top - \mathbf{H}_{j,E}^{\text{ref}} \bar{\mathbf{V}}_0^\top\|_F \leq O(pAd) + (O(p) + O(\sqrt{p}d))v. \quad (147)$$

For the ideal-rival vector  $a_i$ , the arm identity gives  $\mathbf{1}^\top a_i = Q_j + \zeta_i$ ,  $|\zeta_i| \leq D_L$ , and the exact residual is

$$\mathbf{R}_i^* / \mathbf{w}_j - \mathbf{T}_i = \frac{(1 - \mathbf{w}_j)\zeta_i \mathbf{T}_i - \mathbf{P}a_i}{Q_j + \mathbf{w}_j \zeta_i}.$$

Since  $Q_j - \mathbf{w}_j D_L \gtrsim p$  and  $\|\mathbf{P}a_i\| \leq \sqrt{p-1}$ , its norm is  $O(\omega)$ . Softmax is 1/2-Lipschitz in Euclidean norm. Margin errors bounded by  $b$  change the probabilities by at most  $b$  in  $\ell^1$ : after matching true logits their range is at most  $2b$ , and the Jacobian's  $\ell^1$  action is their mean absolute deviation. The reference bounds and (147) now give

$$\|\mathbf{R} - \mathbf{w}_j \mathbf{T}_S\|_F / \sqrt{n} = O(A\omega + Ad + v + dv/\sqrt{p}).$$

The reference gate gap is  $b \gtrsim m^{-1/2}$ , so entrywise

$$|\mathbf{D} - \mathbf{D}^0| \leq |\mathbf{X}_S (\mathbf{W} - \mathbf{W}_j^{\text{ref}})^\top| / b.$$

Expand  $\mathbf{F} = n^{-1} ((\mathbf{R}\mathbf{V}) \odot \mathbf{D})^\top \mathbf{X}_S$  about the anchor  $\mathbf{F}^{\text{anc}}$ . Use  $\|\mathbf{R}\|_F \leq \sqrt{2n}$ ,  $\|\mathbf{X}_S\|_{\text{op}} = O(\sqrt{p})$ , the center head operator bound in (145) and the separate entrywise bound  $|(\mathbf{T}_S \bar{\mathbf{V}}_0)_{ig}| = O(Ap/\sqrt{m})$  to obtain

$$\|\mathbf{F} - \mathbf{F}^{\text{anc}}\|_F \leq O(v/\sqrt{p}) + O(A) \|\mathbf{R} - \mathbf{w}_j \mathbf{T}_S\|_F / \sqrt{n} + O(Ad).$$

The source comparison adds  $\mathbf{k}_j O(p^{7/4})$ . Dividing by  $\mathbf{k}_j \sigma_p \sqrt{r_v} \gtrsim A$  gives

$$\delta_j^F \leq C(\phi + Ad). \quad (148)$$

### Fixed-horizon closure.

Every force row lies in  $E_{\text{vis}}$ . Lemma F.25 therefore gives the following bound without an actual-rank premise:

$$\|\mathcal{N}_{\text{pol}}(\mathbf{F}_j) - \mathbf{U}_F\|_F \leq 2\delta_j^F. \quad (149)$$

Hence, whenever  $d_j \leq 1$ ,

$$d_{j+1} \leq ad_j + (2s/h)\delta_j^F. \quad (150)$$

Together, (148), (149) and (150) propagate the feature distance and force error from one departing state to the next. The quantity  $\phi$  obeys  $A^\ell \phi \rightarrow 0$  for every fixed  $\ell$ . Induct on  $j$  with the assertion  $A^\ell d_j \rightarrow 0$  for every fixed  $\ell \geq 0$ . At  $j = 0$  it follows from  $d_0 \leq r_6/h = O(p^{-2})$ . At the inductive step,  $d_j \rightarrow 0$  licenses the departing bound  $d_j \leq 1$  eventually, and

$$A^\ell \delta_j^F \leq C(A^\ell \phi + A^{\ell+1} d_j) \rightarrow 0.$$

The distance recurrence (150) proves the assertion at  $j + 1$ . All estimates are event-, ball- and head-uniform. For each fixed power there are finitely many times, so every threshold needed for the theorem can be chosen before the split. This establishes the claimed limits without a future tube or rank premise.

### Mean loss and held-out accuracy.

Put  $u'_{j,x_c} = \nu_j A u_{j,x_c}$  and

$$\mathbf{G}_j = \nu_j A \mathbf{\Delta}_j + \mathbf{H}_j \mathbf{V}_j^\top - \mathbf{H}_j^{\text{ref}} \bar{\mathbf{V}}_0^\top.$$

This paragraph's  $\mathbf{G}_j$  is a logit-error table, distinct from the departing-state head source in Appendix G. The actual margins consequently have the form

$$\nu_j A [M_x + \xi(2M_x - N_{x_c}) + k z_j] + u'_{j,x_c} + G_{j,xy_x} - G_{j,x_c}, \quad (151)$$

where  $|u'_{j,x_c}| = O(A\omega)$  and  $\|\mathbf{G}_j\|_F / p = O(A\omega + Ad_j + p^{-1}) = o(1)$ . At zero,  $|u'_{0,x_c}| \leq Ae_0$  and  $\|\mathbf{G}_0\|_F = O(pAr) = o(A)$ , because  $\mathbf{\Delta}_0 = 0$ . The ideal training CE differs from  $L_*$  by at most  $e^{-a_{\text{conf}}}[2t_j + D_L/(p-1)] = o(1)$ , using  $\nu_j \xi > 11/540$ . CE is 1-Lipschitz in the maximum margin error and  $\sqrt{2}$ -Lipschitz in the logit vector. Averaging the errors in (151) and using  $\sum_{i \in \mathcal{S}} \|\mathbf{G}_{j,i}\| \leq \sqrt{n} \|\mathbf{G}_j\|_F$  proves  $L_j = L_* + o(1)$ .

For held-out accuracy, apply the same margin decomposition (151) to the clock signs in Lemma F.18(iii). Every held-out row has an  $N_{x_c} = 2$  rival since  $\rho^2(p-1) > D_L$ . Its pre- $J$  ideal margin is at most  $-Ag_-$ ; at  $J$  all ideal margins are at least  $Ag_+$ . The clock gaps are uniformly bounded below by  $(1 - a_J)/2 > 0$ . Adverse classification therefore requires a row norm of  $\mathbf{G}_j$  at least a fixed positive multiple of  $A$ ; counting such rows by  $\|\mathbf{G}_j\|_F^2$  proves both accuracy limits. At zero the chosen rival has ideal margin  $-2A\xi$ ;  $e_0 < \xi$  and  $\|\mathbf{G}_0\|_F = o(A)$  make it strictly negative in every row, proving the exact initial claim independently of the refit.

### Operator motion and every training margin.

Lemma F.16(iii) and  $\xi < \rho Z c_R / 2$  give  $\limsup \xi \leq 36/125$  for the prescribed rates. The anchored Polar rank estimate gives  $\text{rank } \mathbf{F}_j \geq r_v [1 - (\delta_j^F)^2]$ . Thus the forces are nonzero eventually, and their normalized Polar directions have operator norm at most  $[r_v(1 - \max_{t < J} (\delta_t^F)^2)]^{-1/2}$ . Exact unrolling of the feature update, with  $s \sum_{t < j} a^{-t-1} = h z_j$ , gives

$$\frac{\|\mathbf{W}_j - \nu_j \bar{\mathbf{W}}_0\|_{\text{op}}}{\nu_j} \leq r + \frac{h z_j}{\sqrt{r_v(1 - \max_{t < J} (\delta_t^F)^2)}}. \quad (152)$$

Positive homogeneity and ReLU contraction multiply this bound by at most  $\sqrt{2}$  for a row-feature error. Using the center training margin  $A(1 - e_0)$ , the head distance  $r + J\eta$ , and the contrast budget (145), each actual margin divided by  $\nu_j A$  is at least

$$1 - e_0 - \frac{\sqrt{2}K_p}{hq} \left( r + \frac{h z_j}{\sqrt{r_v}} (1 + o(1)) \right) - \frac{\sqrt{2}(r + J\eta)}{A} \left[ 2hq + \sqrt{2} \left( r + \frac{h z_j}{\sqrt{r_v}} (1 + o(1)) \right) \right].$$

All errors except feature motion vanish uniformly. Since  $z_j \leq 9/50$ ,  $q \rightarrow 1/4$  and  $r_v = 2p - 1$ , the squared limsup of that motion term is

$$\limsup \left( \frac{\sqrt{2}K_p(9/50)}{q\sqrt{r_v}} \right)^2 \leq 32 \left( \frac{9}{50} \right)^2 \left[ \frac{4}{5} + \left( \frac{36}{125} \right)^2 \right] = \frac{8939808}{9765625} < \left( \frac{24}{25} \right)^2.$$

Thus the normalized margins have uniform liminf greater than  $1/25$ , which proves (142). This operator argument is separate from the Frobenius comparison used for mean loss and held-out accuracy.  $\square$

**Completion of Claim I.** Lemma F.30 now gives both required forms of control: the logit decomposition (151) gives mean loss and delayed held-out recovery, while the operator-motion bound (152) gives every training margin. These dynamics conclusions are established before the refit and Fourier estimates of Claim II.

### F.3 Claim II: refit-first recovery and Fourier growth

#### F.3.1 Full refit and Fourier fraction

We first analyze the reference table  $\mathbf{F}(z)$ : Lemma F.31 bounds its kernel, Lemma F.32 identifies its refit predictions, and Lemma F.33 proves its Fourier growth. Lemma F.34 controls changes in the complete refit, including its feature-dependent penalty. Section F.4 then applies these estimates to the actual tables using Lemma F.30.

The complete rule (18), written  $\mathbf{Y}_U$  for a table  $U$ , is invariant under positive scaling by Lemma E.38(i);  $A_H(U)$  has the same invariance directly from its ratio definition (6).

**Lemma F.31** (Reference-kernel bounds). *For all sufficiently large  $p$ , uniformly over the split event and  $0 \leq z \leq 1/5$ ,  $\mathbf{B}(z)$  has a constant diagonal and*

$$\begin{aligned} \|\mathbf{P}_{\text{int}}\mathbf{F}_0\|_{\text{op}} &= 1, \quad \|\mathbf{P}_{\text{int}}\mathbf{P}_F\|_{\text{op}} < 8, \quad \|\mathbf{P}_{\text{int}}\mathbf{F}(z)\|_{\text{op}}^2 \leq 169/25, \\ (1 + 2z/3)^2 - 1 &\leq \theta_R(z), \quad \lambda'_R(z) > 4/3, \\ \mathbf{B}(z) &= (\mathbf{P}_0 + \mathbf{P}_{\text{ax}})\mathbf{B}(z)(\mathbf{P}_0 + \mathbf{P}_{\text{ax}}) + \mathbf{P}_{\text{int}} + \theta_R(z)\mathbf{P}_R + \mathbf{E}_z, \\ \|\mathbf{E}_z\|_{\text{op}} &= O(1), \quad \|\mathbf{E}_z\|_F = O(\sqrt{p}), \quad \mathbf{E}_z\mathbf{P}_R = \mathbf{E}_z\mathbf{Z} = 0. \end{aligned}$$

Moreover  $\|\mathbf{F}(z)_S\|_F / \sqrt{n} = 2 + O(p^{-1/2})$  and

$$\begin{aligned} q_{F(z)} &= \kappa\rho + o(1), \quad q_{F(z)} \geq q_{\min}, \\ \|\mathbf{F}(z)_{\mathcal{D}}v\|^2 &\leq C_{\max}[\|\mathbf{F}(z)_Sv\|^2 + q_{F(z)}\|v\|^2]. \end{aligned}$$

*Proof.* Lemma F.27 gives  $\mathbf{B}(z) = \mathbf{B}_0 + 2z\mathbf{C} + z^2\mathbf{B}_2$  and translation/dilation invariance. Thus the diagonal is constant, the additive, rule and mixed spaces reduce  $\mathbf{B}(z)$ , and  $\|\mathbf{F}(z)_S\|_F^2 = (n/p^2)\|\mathbf{F}(z)\|_F^2$ . Each eigenvalue of  $\mathbf{B}_2$  on a line is at most  $\|\mathbf{P}_F\|_F^2 / (p-1) \leq 2p/(q^2(p-1)) < 64$ . Since  $\mathbf{B}_0 = \mathbb{I}$  on the interaction space, this proves the operator bounds and the displayed decomposition. The rule lower bound and derivative follow from  $c_R > 2/3$  and  $R_R \geq c_R^2$ . On the mixed space,  $\mathbf{E}_z = 2z\mathbf{C}_{\text{mix}} + z^2\mathbf{P}_{\text{mix}}\mathbf{B}_2\mathbf{P}_{\text{mix}}$ , so its operator norm is bounded and

$$\|\mathbf{E}_z\|_F \leq (2z + 8z^2)\|\mathbf{P}_F\|_F = O(\sqrt{p}).$$

The constant diagonal,  $\|\mathbf{F}_0\|_F = 2p$  and  $\|\mathbf{P}_F\|_F \leq \sqrt{2p}/q$  give  $\|\mathbf{F}(z)_S\|_F / \sqrt{n} = 2 + O(p^{-1/2})$ . Hence  $q_{F(z)} = 4\gamma n[1 + O(p^{-1/2})] = \kappa\rho + o(1)$  uniformly, so  $q_{F(z)} \geq q_{\min}$  eventually. For a comparison vector  $v$ , put  $f = \mathbf{F}(z)v$  and decompose  $f = f_{\text{add}} + f_{\text{int}}$ , with  $f_{\text{add}} = (\mathbf{P}_0 + \mathbf{P}_{\text{ax}})f$  and  $f_{\text{int}} = \mathbf{P}_{\text{int}}f$ . The sampled design gives  $\|f_{\text{add}}\|^2 \leq (2/\rho)\|f_{\text{add},S}\|^2$  and  $\|f_{\text{int}}\|^2 \leq 169\|v\|^2/25$ . Using  $\|f_{\text{add},S}\|^2 \leq 2\|f_S\|^2 + 2\|f_{\text{int}}\|^2$  proves the leverage bound.  $\square$

For the reference-refit proof put  $c_z = 1 + q_{F(z)}$  and

$$\begin{aligned} \mathbf{U}_{\perp} &= \mathbf{T}_S - \mathbf{Z}_S\mathbf{K}^{-1}\mathbf{\Psi}, \quad \mathbf{\Delta}_{\perp} = \mathbf{T}_S^{\top}\mathbf{U}_{\perp} - \rho p\mathbf{P}, \\ \mathbf{\alpha}_z &= \mathbf{U}_{\perp}/(c_z + \rho\theta_R(z)), \quad \mathbf{Q} = (\mathbf{B}(z)_{SS} + q_{F(z)}\mathbb{I})^{-1}. \end{aligned}$$

**Lemma F.32** (Full reference predictions). *Uniformly over the split event and  $z \in [0, 1/5]$ , for all sufficiently large  $p$ ,*

$$\|\mathbf{Y}_{F(z)} - \Gamma(z)\mathbf{T}_{\mathcal{D}}\|_F / N_{\mathcal{D}} = O(\omega). \quad (153)$$

At  $z = 0$ ,

$$\mathbf{Y}_{F_0} = \mathbf{Z}_{\mathcal{D}}(\lambda_{\text{ob}}\mathbf{\Upsilon} + \mathbf{K})^{-1}\mathbf{\Psi}. \quad (154)$$

*Proof.* The trial coefficient  $\mathbf{\alpha}_z$  removes the additive part of the training targets. We estimate its normal-equation residual and its held-out prediction, then use the bound in Lemma F.31 to control the correction to the exact ridge solution.

The sampled design gives  $\mathbf{K} \succeq s_0\mathbb{I}$ ,  $s_0 \asymp p$ ,  $\|\mathbf{Z}_S\|_{\text{op}} = O(\sqrt{p})$ ,  $\|\mathbf{\Psi}\|_F = O(p)$  and  $N_{\mathcal{D}} \asymp p$ . The orthogonal residual satisfies

$$\begin{aligned} \mathbf{Z}_S^{\top}\mathbf{U}_{\perp} &= 0, \quad \|\mathbf{T}_S - \mathbf{U}_{\perp}\|_F = O(\sqrt{p}), \\ \mathbf{\Delta}_{\perp} &= \mathbf{T}_S^{\top}\mathbf{T}_S - \rho p\mathbf{P} - \mathbf{\Psi}^{\top}\mathbf{K}^{-1}\mathbf{\Psi}, \quad \|\mathbf{\Delta}_{\perp}\|_{\text{op}} \leq D_* + 3t_*^2/s_0 = O(p\omega). \end{aligned}$$

Extend  $\mathbf{U}_\perp$  by zero off  $\mathcal{S}$ . Its orthogonality annihilates the entire additive block of  $\mathbf{B}(z)$ , irrespective of that block's eigenvalues. The nonrule-line leakage estimate of Lemma F.27 and  $\|\mathbf{E}_z\|_F = O(\sqrt{p})$  bound  $(\mathbf{E}_z)_{\Omega\mathcal{S}}\mathbf{T}_\mathcal{S}$  by  $O(\sqrt{p})$ . The correction  $\mathbf{T}_\mathcal{S} - \mathbf{U}_\perp$  has the same bound, and  $\|\mathbf{E}_z\|_{\text{op}} = O(1)$ ; thus  $\|(\mathbf{E}_z)_{\mathcal{S}\mathcal{S}}\mathbf{U}_\perp\|_F + \|(\mathbf{E}_z)_{\mathcal{D}\mathcal{S}}\mathbf{U}_\perp\|_F = O(\sqrt{p})$ .

Since  $\mathbf{T}_\mathcal{S}\mathbf{P} = \mathbf{T}_\mathcal{S}$ , the trial normal residual and test prediction are

$$(\mathbf{B}(z)_{\mathcal{S}\mathcal{S}} + q_{F(z)}\mathbb{I})\boldsymbol{\alpha}_z - \mathbf{T}_\mathcal{S} = \frac{c_z(\mathbf{U}_\perp - \mathbf{T}_\mathcal{S}) + \theta_R(z)\mathbf{T}_\mathcal{S}\boldsymbol{\Delta}_\perp/p + (\mathbf{E}_z)_{\mathcal{S}\mathcal{S}}\mathbf{U}_\perp}{c_z + \rho\theta_R(z)}, \quad (155)$$

$$\mathbf{B}(z)_{\mathcal{D}\mathcal{S}}\boldsymbol{\alpha}_z = \Gamma(z)\mathbf{T}_\mathcal{D} + \frac{\theta_R(z)\mathbf{T}_\mathcal{D}\boldsymbol{\Delta}_\perp/p + (\mathbf{E}_z)_{\mathcal{D}\mathcal{S}}\mathbf{U}_\perp}{c_z + \rho\theta_R(z)}. \quad (156)$$

Both errors have Frobenius norm  $O(p\omega)$ . Applying the reference bound of Lemma F.31 to the range of  $\mathbf{F}(z)_\mathcal{S}^\top$  gives

$$\mathbf{Q}\mathbf{B}(z)_{\mathcal{S}\mathcal{D}}\mathbf{B}(z)_{\mathcal{D}\mathcal{S}}\mathbf{Q} \preceq C_{\max}\mathbf{Q}\mathbf{B}(z)_{\mathcal{S}\mathcal{S}} \preceq C_{\max}\mathbb{I}.$$

Here  $\mathbf{Q}\mathbf{B}(z)_{\mathcal{S}\mathcal{S}} = \mathbb{I} - q_{F(z)}\mathbf{Q} \preceq \mathbb{I}$ . The correction to (156) is the residual (155) acted on by  $-\mathbf{B}(z)_{\mathcal{D}\mathcal{S}}\mathbf{Q}$ , whose operator norm is at most  $\sqrt{C_{\max}}$ . Dividing the resulting prediction error by  $N_\mathcal{D} \asymp p$  proves (153).

At  $z = 0$ , the constant diagonal gives  $q_{F_0} = 4\gamma n$ . Equation (132) at  $c_o = \lambda_{\text{ob}} = 1 + q_{F_0}$ , multiplied by  $\mathbf{T}_\mathcal{S}$ , gives (154).  $\square$

For the centered reference trace, put

$$B_c = \|\mathbf{C}_{\text{grid}}\mathbf{P}_F\|_F / \sqrt{(p-1)(3p+1)}, \quad d_c(z) = \|\mathbf{C}_{\text{grid}}\mathbf{F}(z)\|_F^2 / [(p-1)(3p+1)].$$

**Lemma F.33** (Fourier growth and transfer). *For all sufficiently large  $p$ , uniformly on  $0 \leq z \leq 1/5$ ,*

$$A_H(\mathbf{F}(z)) = A_0\lambda_R(z)/d_c(z), \quad \frac{d}{dz}A_H(\mathbf{F}(z)) > A_0, \quad A_H(\mathbf{F}(z)) \leq (34/5)A_0.$$

*For any family with  $z \in [0, 1/5]$ ,  $e = \|\mathbf{V} - \mathbf{F}(z)\|_F$  and  $\sup e = o(\sqrt{p})$ , uniformly*

$$|A_H(\mathbf{V}) - A_H(\mathbf{F}(z))| = O(e/p^{3/2}) = o(A_0).$$

*Proof.* The initial identity  $4p^2 - (p+1)^2 = (p-1)(3p+1)$  gives  $d_c(0) = 1$ . Triangle/Cauchy and  $B_c = O(p^{-1/2})$  give  $d_c = 1 + O(p^{-1/2})$  and  $d'_c = O(p^{-1/2})$ . The rule-line energy is  $(p-1)\lambda_R$ , giving the fraction. Since  $\lambda'_R > 4/3$  and  $\lambda_R \leq \|\mathbf{P}_{\text{int}}\mathbf{F}(z)\|_{\text{op}}^2 \leq 169/25$ ,

$$\frac{1}{A_0} \frac{d}{dz} A_H(\mathbf{F}(z)) = \frac{\lambda'_R}{d_c} - \frac{\lambda_R d'_c}{d_c^2} \geq \frac{4}{3} - O(p^{-1/2}) > 1, \quad \frac{A_H(\mathbf{F}(z))}{A_0} \leq 169/25 + O(p^{-1/2}) < 34/5.$$

For  $N_X = \|\mathbf{P}_{\text{add}}\mathbf{C}_{\text{grid}}\mathbf{X}\|_F$ ,  $D_X = \|\mathbf{C}_{\text{grid}}\mathbf{X}\|_F$ , the reference has  $N_F = O(\sqrt{p})$  and  $D_F = \Theta(p)$ . Contraction gives  $|N_V - N_F|, |D_V - D_F| \leq e$  and hence  $N_V = O(\sqrt{p})$ ,  $D_V = \Theta(p)$ .

$$|A_H(\mathbf{V}) - A_H(\mathbf{F}(z))| \leq \frac{(N_V + N_F)e}{D_V^2} + \frac{A_H(\mathbf{F}(z))(D_V + D_F)e}{D_V^2} = O(e/p^{3/2}) = o(A_0).$$

$\square$

For the following nearby-table estimate use  $b_{\max} = 9/5$ .

**Lemma F.34** (Nearby full feature tables). *For all sufficiently large  $p$ , let  $0 \leq z \leq 1/5$ ,  $\mathbf{U} = \mathbf{F}(z)$  and  $e = \|\mathbf{V} - \mathbf{U}\|_F \leq a_{\min}\sqrt{n_-}$ , where  $a_{\min}$  is fixed in Definition F.28 and  $b_{\max} = 9/5$ . Then  $q_U, q_V \geq q_{\min}$ ,  $|q_V - q_U| \leq \gamma(4b_{\max}pe + e^2)$  and  $\|\mathbf{Y}_V - \mathbf{Y}_U\|_F \leq \mathbf{P}_p(e)N_\mathcal{D}$ , with  $\mathbf{P}_p$  as in (139).*

*Proof.* The reference training-norm bounds in Lemma F.31 give  $2a_{\min}\sqrt{n} \leq \|\mathbf{U}_\mathcal{S}\|_F \leq 2b_{\max}\sqrt{n}$ ; thus  $\|\mathbf{V}_\mathcal{S}\|_F \geq a_{\min}\sqrt{n}$  and  $q_V \geq \gamma a_{\min}^2 n \geq q_{\min}$ . Apply Lemma E.38(ii) with  $e_S, e_D \leq e$ ,  $t \leq \sqrt{p}$ ,  $N \leq N_+$  and  $L \leq \sqrt{C_{\max}}$ . Its penalty-change estimate gives the displayed penalty change, and substitution gives exactly (139).  $\square$

**Lemma F.35** (Initial refit failure). *For the signed margin reserve (140), every held-out input has a rival whose center-refit margin is at most  $-\mu_{\text{obs}0}$ . If a feature table  $\mathbf{V}$  has  $e = \|\mathbf{V} - \mathbf{F}_0\|_F \leq a_{\min}\sqrt{n_-}$  and  $\sqrt{2}N_+P_p(e) < \mu_{\text{obs}0}$ , then every held-out input has a strictly negative refit margin for  $\mathbf{V}$ , and  $\text{Acc}_{\text{ref}}(\mathbf{V}) = 0$ .*

*Proof.* Equation (154) identifies the center-refit row as  $z_x \mathbf{C}^{(\lambda_{\text{ob}})}$ . Every held-out  $x$  has a rival  $c$  with  $N_{xc} = 2$ , since  $N_x^{(2)} \geq \rho^2(p-1) - D_L > 0$ . Equation (133) bounds this margin by  $-2/\Lambda_1 + 2|\iota|D_* + 2\epsilon_{\text{pred}} = -\mu_{\text{obs}0}$ . Lemma F.34 with  $z = 0$  moves each prediction row by at most  $P_p(e)N_{\mathcal{D}} \leq P_p(e)N_+$  and each margin by at most  $\sqrt{2}$  times this quantity, proving the claim.  $\square$

## F.4 Assembly: the prescribed ball and one common threshold

The dynamics conclusions are already supplied by Lemma F.30. The remaining task is to transfer the reference refit and Fourier estimates to those trajectories, use the prescribed ball for exact initial refit failure, and choose a common threshold for all conclusions.

Use the exact prescribed radius (141) from Definition F.29. This section proves a sufficiently-large-prime statement with all theorem parameters, including  $J$ , fixed first; no conclusion is asserted for smaller primes merely because they pass finite tests.

For observer transfer put

$$\widetilde{\mathbf{H}}_j = \mathbf{H}_j / (\nu_j h q), \quad e_j^{\text{obs}} = \|\widetilde{\mathbf{H}}_j - \mathbf{F}(z_j)\|_F, \quad d_p = \frac{\sqrt{2p}}{h q}. \quad (157)$$

**Proposition F.36** (Uniform eventual window). *For the fixed parameters of Theorem F.1, there is a finite  $p_0$ , chosen before the split, such that for every odd prime  $p \geq p_0$  and every split in  $\mathcal{E}_{\text{split}}$ , the center and all static quantities used above are defined and  $0 < r_{\text{new}} < 1$ . Let  $\|(\mathbf{W}_0 - \bar{\mathbf{W}}_0, \mathbf{V}_0 - \bar{\mathbf{V}}_0)\|_F < r_{\text{new}}$ , and use*

$$\mathbf{W}_{j+1} = a\mathbf{W}_j + s\mathcal{N}_{\text{pol}}(-\nabla_{\mathbf{W}} L(\mathbf{W}_j, \mathbf{V}_j)), \quad \mathbf{V}_{j+1} = \mathbf{V}_j + \mathbf{D}_j, \quad \|\mathbf{D}_j\|_F \leq \eta \quad (j < J).$$

*Then all trajectory conclusions of Theorem F.1 hold on  $0, \dots, J$ , including strict training margins, the comparison with this trajectory's own initial refit error, and  $A_H(\mathbf{H}_j) < 7/(3p+1)$ . The head budget is  $\eta = \theta/(hp)$ , prescribed before sampling in (119).*

*Proof.* All orders below are uniform over the event, exact prescribed ball and head sequences in the statement. Lemma F.30 proves the static certificates,  $0 < r_{\text{new}} = r_6 < 1$ ,  $A \max_j d_j \rightarrow 0$  and strict training. Its loss and actual-head accuracy limits, with  $L_* > 0$  and the fixed tolerances, give all required CE and accuracy conclusions, including exactly zero initial held-out accuracy for any fixed tie rule.

*Positive-time refit accuracy.*

For the normalized observer table (157), ReLU contraction and Lemma F.30 give

$$e_j^{\text{obs}} \leq \frac{\sqrt{2p} d_j}{\nu_j q} = o(\sqrt{p}), \quad P_p(e_j^{\text{obs}}) = o(1). \quad (158)$$

Both the nearby-refit hypothesis  $e_j^{\text{obs}} \leq a_{\min}\sqrt{n_-}$  and the uniform Fourier-transfer hypothesis  $e_j^{\text{obs}} = o(\sqrt{p})$  hold eventually. The uniform reference error in Lemma F.32 is  $O(\omega)$ . Combining its prediction estimate (153) with Lemma F.34 and (158) gives

$$\frac{\|\mathbf{Y}_{\widetilde{\mathbf{H}}_j} - \Gamma(z_j)\mathbf{T}_{\mathcal{D}}\|_F}{N_{\mathcal{D}}} = o(1). \quad (159)$$

For  $j \geq 1$ , Lemma F.18(ii) gives  $z_j > 1/(12J)$ , so  $\Gamma(z_j)$  is uniformly bounded below by a positive constant. A row without a strict correct maximum costs at least  $\Gamma(z_j)^2/2$  in squared prediction error. As  $N_{\mathcal{D}}^2 = (1-1/p)n_{\mathcal{D}}$ , the erroneous fraction tends to zero, giving the required refit accuracy. At  $j = 0$ ,  $\Gamma(0) = 0$ , so this trajectory's own  $E_{\text{obs}}(\mathbf{H}_0) \rightarrow 1$ .

*Same-trajectory error reduction.*

Equation (159) also determines the normalized error along this trajectory. The following lower bound on  $\Gamma$  makes the late-time gain quantitative.

For  $j = J - 1, J$ ,  $z_j > 1/6$  and the uniform limit  $q_{F(z)} \rightarrow \kappa\rho$  in Lemma F.31 give

$$\liminf_{p \rightarrow \infty} \frac{\rho\theta_R(z_j)}{1 + q_{F(z_j)}} \geq \frac{19\rho}{81(1 + \kappa\rho)} \geq \frac{152}{891} > \frac{1}{6}.$$

Hence  $\Gamma(z_j) \geq 1/7$  eventually and

$$E_{\text{obs}}(\mathbf{H}_0) - E_{\text{obs}}(\mathbf{H}_j) \geq 1 - (6/7)^2 - o(1) = 13/49 - o(1) > 401/1568.$$

This comparison is with the same initialization, not the center.

*Exact initial refit failure.*

The asymptotic error limit above does not establish zero accuracy at  $j = 0$ ; that conclusion uses the exact prescribed radius. Here  $e_0^{\text{obs}} \leq d_p r$  and nonnegative coefficients and  $r < r_{\text{new}} < 1$  give

$$\sqrt{2}N_+ \mathbf{P}_p(e_0^{\text{obs}}) \leq \sqrt{2}N_+ \mathbf{P}_p(d_p r) \leq r\sqrt{2}N_+ \mathbf{P}_p(d_p) < \mu_{\text{obs}0}/2. \quad (160)$$

The radius choice (141) yields (160), leaving the signed margin reserve required by Lemma F.35, which proves zero initial refit accuracy.

*Fourier fraction and stepwise gain.*

Finally, with  $A_0 = 1/(3p + 1)$ , apply Lemma F.33 using the table error in (158). This gives  $|A_H(\mathbf{H}_j) - A_H(\mathbf{F}(z_j))| = o(A_0)$ . Lemma F.18(ii) and the reference derivative in Lemma F.33 give increments  $> A_0/(12J)$ , so actual increments exceed  $A_0/(24J)$  eventually; the same transfer gives  $A_H(\mathbf{H}_j) \leq (34/5)A_0 + o(A_0) < 7A_0$ . Positive scaling preserves the full refit by Lemma E.38(i), and preserves  $A_H$  by (6). There are finitely many states and every bound is uniform, so one finite  $p_0$  establishes all assertions simultaneously.  $\square$

*Proof of Theorem F.1.* Proposition F.11 gives the proper split event with probability at least  $1 - 2/\sqrt{p}$ . Proposition F.36 supplies, uniformly on that event, a finite pre-sampling threshold and a positive radius, hence the prescribed nonempty open ball in all  $3pm$  coordinates. The actual update has  $1 - s\lambda = a$  and head increments of norm  $\eta$  or zero, so the same proposition combines Claims I and II on its trajectory. No accuracy after  $J$  is asserted.  $\square$

## G Small decoupled readout decay

This appendix gives two positive readout-decay certificates, each using the assembly result for its own construction. For the half-grid, Proposition E.54(a) gives joint finite-horizon continuity: the same ball, head step and windows preserve Theorem 4.1(i)–(iii). Its late-refit clause (iv) is outside this positive-decay conclusion. For the random split, Proposition F.36 allows any head sequence whose increments have norm at most  $\eta$ . We verify that bound using data step  $\eta/2$  and a decay cap chosen before sampling. This preserves the specified event, ball and  $\eta$ -based budgets.

In both constructions, the raw training-CE sources are evaluated at the departing state, decay is outside normalization, and  $\mathcal{N}_F(0) = 0$ . The decayed trajectories need not equal the undecayed ones. Throughout,  $d_V = \eta_V \lambda_V$  denotes per-step readout shrinkage;  $\eta_V$  is the actual data step of the particular extension, not necessarily its undecayed motion budget  $\eta$ .

### G.1 The structured half-grid construction

Fix a prime  $p \geq 101$  and a width  $M \geq p(p^2 + 3)/2$ . Let  $s, \lambda, J$ , the windows (61), the head step  $\eta_V = \eta_0/2$  and the ball  $\mathcal{B}_M$  be those of Proposition E.54, and let  $d_0$  be its decay bound; all are fixed before the initialization is chosen.

**Corollary G.1** (Half-grid readout decay). *For every fixed  $d_V$  with  $0 < d_V < \min\{d_0, 1/(20J)\}$ , every initialization in  $\mathcal{B}_M$ , trained by the feature update of (3) and the head update  $\mathbf{V}_{j+1} = (1 - d_V)\mathbf{V}_j + \eta_V \mathcal{N}_F(\mathbf{G}_j)$ , satisfies Theorem 4.1(i)–(iii), with the same arrival windows (61). Moreover  $Jd_V < 1/20$ .*

*Proof.* Apply Proposition E.54(a) on its final ball with

$$\eta = \eta_V = \eta_0/2 < \eta_0, \quad 0 < d_V < d_0, \quad Jd_V < J/(20J) = 1/20. \quad (161)$$

It gives clauses (i)–(iii) on the same arrival windows (61). Part (b), which supplies the late-refit conclusion, assumes  $d_V = 0$  and is not invoked here.  $\square$

The cap  $d_0$  is a fixed positive number from finite-horizon continuity; no formula is given for it. The same ball and the same head step serve the undecayed theorem.

## G.2 The random-split construction

Use Appendix F’s event, width  $m = p^2(p-1)^2$ , step  $\eta = \theta/(hp)$ , common threshold  $p_0$  of Proposition F.36, and specified centers and radii  $\bar{\mathbf{V}}_0(\mathcal{S}), r_{\text{new}}(\mathcal{S})$ . Definition F.10 supplies the event and Definition F.29 supplies the prescribed ball; their definitions are retained when decay is added.

**Definition G.2** (Pre-sampling readout-decay budgets). For the random-split decay extension, define before sampling

$$B_p = \max_{\mathcal{S} \in \mathcal{E}_{\text{split}}} \{\|\bar{\mathbf{V}}_0(\mathcal{S})\|_F + r_{\text{new}}(\mathcal{S})\}, \quad Q_p = B_p + J\eta/2.$$

The maximum is over all event masks, not the realized split. All Appendix F radius and error formulas retain the undecayed step budget  $\eta$ .

Using these pre-sampling bounds, choose

$$0 < d_V \leq \min\{1/2, \eta/(2Q_p)\} \quad (162)$$

before sampling, keep the feature update (119), and use

$$\mathbf{V}_{j+1} = (1 - d_V)\mathbf{V}_j + \frac{\eta}{2}\mathcal{N}_F(\mathbf{G}_j). \quad (163)$$

**Corollary G.3** (Random-split readout decay). *Fix the parameters of Theorem F.1 and an odd prime  $p \geq p_0$ . Choose the pre-sampling decay (162) and use the update (163). On  $\mathcal{E}_{\text{split}}$ , of probability at least  $1 - 2/\sqrt{p}$ , every initialization in the ball (141) satisfies all trajectory conclusions of Theorem F.1 at states  $0, \dots, J$ . The width, split law, event, common threshold and prescribed radii are those of Appendix F; every radius and error formula retains  $\eta$ , although the actual data step is  $\eta/2$ .*

*Proof. Step 1: obtain a bound valid before sampling.* The finite nonempty event makes  $B_p, Q_p$  in Definition G.2 finite before sampling. After a split in that event is realized, the initial ball (141) gives

$$\|\mathbf{V}_0\|_F < \|\bar{\mathbf{V}}_0(\mathcal{S})\|_F + r_{\text{new}}(\mathcal{S}) \leq B_p. \quad (164)$$

**Step 2: bound the head along the decayed trajectory.** The normalized data source has norm at most one, including when  $\mathbf{G}_j = 0$ . Since  $0 < d_V \leq 1/2$ , the update (163) therefore gives

$$\|\mathbf{V}_{j+1}\|_F \leq (1 - d_V)\|\mathbf{V}_j\|_F + \eta/2 \leq \|\mathbf{V}_j\|_F + \eta/2. \quad (165)$$

Induction from (164) yields

$$\|\mathbf{V}_j\|_F < B_p + j\eta/2 \leq Q_p, \quad 0 \leq j \leq J. \quad (166)$$

**Step 3: verify the original motion budget.** Combine the head bound (166) with the pre-sampling cap (162). At every departure  $j < J$ ,

$$\|\mathbf{V}_{j+1} - \mathbf{V}_j\|_F \leq d_V\|\mathbf{V}_j\|_F + \eta/2 < d_V Q_p + \eta/2 \leq \eta. \quad (167)$$

This inequality also covers  $\mathbf{G}_j = 0$ . Proposition F.36 now applies with the undecayed step budget  $\eta$  and gives all the stated trajectory conclusions.  $\square$

Any deterministic finite upper bound for  $B_p$  suffices. Using only the realized split would instead give post-sampling quantifiers.

*Remark G.4 (Readout-decay extensions).* For the random split the data step decreases from  $\eta$  to  $\eta/2$ , reserving motion budget for decay; the half-grid keeps its head step and takes its decay cap from joint continuity. The resulting trajectories need not coincide with the undecayed ones. No practical decay magnitude, equal two-layer shrinkage, or positive decay lower bound uniform in  $p$  is proved. In the main-text convention  $1 - \eta_V \lambda_V$ , the half-grid has  $\eta_V \lambda_V = d_V$ , so  $\lambda_V = d_V / \eta_V$  and  $J \eta_V \lambda_V < 1/20$ ; the random split has  $\eta_V = \eta/2$ ,  $\eta_V \lambda_V = d_V$  and  $J \eta_V \lambda_V \leq J \eta_V / (B_p + J \eta_V) < 1$ . These are consequences, not substitutes for the prescriptions. Software's  $\omega_V$  replaces  $\lambda_V$  in these equalities; normalizing  $\mathbf{G} - \omega_V \mathbf{V}$  as one source is a different, uncovered optimizer.