MÖBIUS LEARNING: CYCLIC DEPTH FOLDING IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer-based language models organize computation along an ordered depth axis, where shallow and deep blocks often develop distinct representational roles. We introduce Möbius Learning, a training architecture that allows the same block parameters to serve multiple depth roles. Through cyclic depth folding, different data streams traverse shared block groups in cyclically shifted orders, jointly training each group in shallow and deep positions, a phenomenon we call depth-role superposition. In mixture-of-experts (MoE) language-model pretraining, Möbius Learning combines effectively with looping, which reuses parameters through repeated passes over the same block sequence. At eight loop passes, Möbius Learning achieves higher five-task zero-shot accuracy than standard looping across all four evaluated model sizes. It also supports effective data reuse: under eightfold repetition, Möbius Learning raises five-task zero-shot accuracy from 46.13% to 48.29% relative to standard looping for the largest evaluated model. These findings establish depth-role superposition as a viable training mechanism that combines with looping and improves downstream performance under repeated-data training.

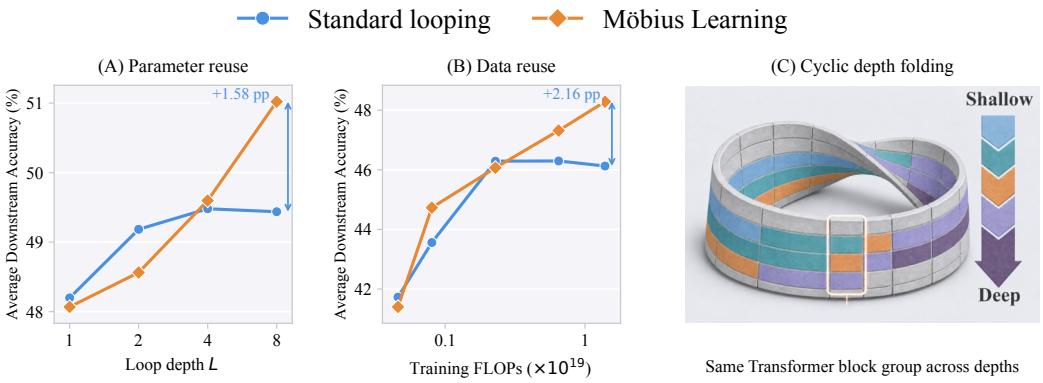


Figure 1: **Möbius Learning for parameter and data reuse.** (A–B) Five-task mean downstream accuracy for MoEs pretrained on 2.67B FineWeb token presentations. **(A) Parameter reuse:** Small MoE (approximately 294M total and 124M active parameters), with loop depth $L = 1, 2, 4, 8$ on the x-axis and no data replay; scores also appear in Figure 4 (C). **(B) Data reuse:** eight data repetitions with one pass ($L = 1$), from Tiny to Large, with estimated training FLOPs on the x-axis. Orange shows one fixed cyclic origin selected by five-task accuracy. Arrows show endpoint differences in percentage points. Protocols appear in Appendix D and Appendix C. **(C)** An illustration of cyclic depth folding: the same block group occupies different depth roles across cyclic orders.

1 INTRODUCTION

Transformer blocks encounter different representations at different depths, and prior studies report depth-dependent linguistic features, attention behavior, and feed-forward memories (Vaswani et al., 2017; Radford et al., 2019; Tenney et al., 2019; Voita et al., 2019; Clark et al., 2019; Geva et al., 2021). This depth–function association makes training one block group in both shallow and deep

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

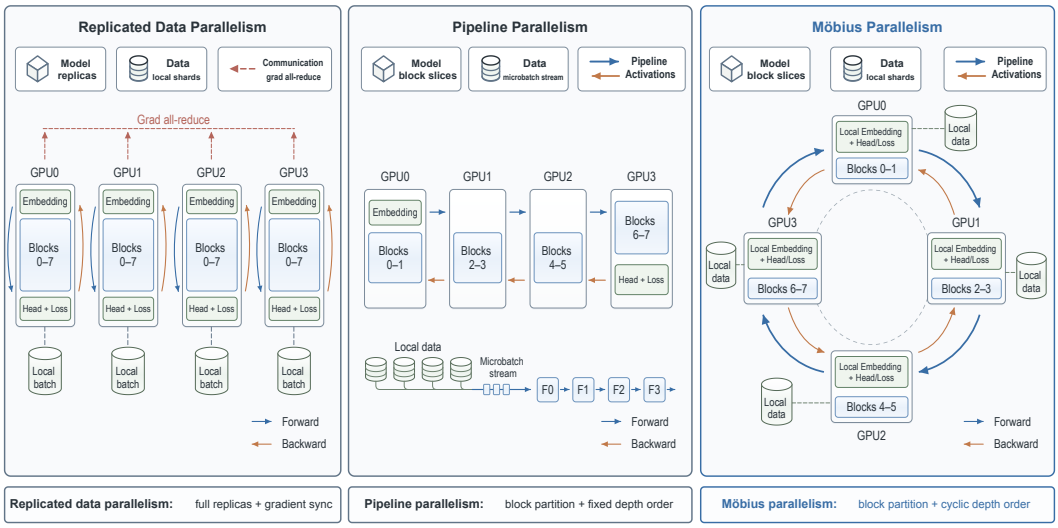


Figure 2: Conceptual comparison of three forms of parallelism. Replicated data parallelism retains a full ordered model at each worker, while pipeline parallelism partitions the model under a fixed stage order. In Möbius parallelism, each worker stores one block group and begins the cycle there, allowing the same group to serve at different depths for data from different workers.

roles nontrivial. Flexible-depth methods vary computation while generally retaining a common block order. Recurrent and cross-layer-sharing models reuse parameters across depth (Dehghani et al., 2019; Lan et al., 2020; Reid et al., 2021); depth-reduction and token-adaptive methods vary computation at the model or token level (Fan et al., 2020; Raposo et al., 2024). Fixed-order sharing schedules and recent looped language models repeatedly apply shared blocks in one prescribed order (Takase & Kiyono, 2023; Csordás et al., 2024; Geiping et al., 2025; Zhu et al., 2025; Gao et al., 2026). Here, loop depth is the number of complete block-sequence passes; it adds computation without changing block-group positions within that order.

Question: *Can the same Transformer block group be trained to serve both shallow and deep roles?*

Möbius Learning introduces cyclic depth folding: each worker introduces its local data at its own block group, so data streams traverse the same groups in cyclically shifted orders. Each group is therefore trained in both shallow and deep *source-relative depth* roles, a phenomenon we call *depth-role superposition*. A group's source-relative depth is its within-pass position measured from the stream's starting point; the construction is formalized in Section 2.

We call the distributed execution strategy induced by Möbius Learning *Möbius parallelism*. Figure 2 contrasts it with replicated data parallelism, which stores a full ordered model at each worker, and pipeline parallelism, which partitions that model across a fixed stage order (Huang et al., 2019; Narayanan et al., 2019).

The central optimization question is whether learning signals from shallow and deep roles interfere. Building on depth-role superposition, we study parameter reuse through repeated loop passes and data reuse through repeated exposure to the training corpus. We pretrain MoE variants of modded GPT-2 (Radford et al., 2019) from random initialization on FineWeb (Penedo et al., 2024).

LayerShuffle fine-tunes fixed-order pretrained Vision Transformers for robustness to shuffled execution (Freiberger et al., 2024). Möbius Learning instead applies cyclic orders throughout language-model pretraining, jointly learning shallow and deep roles; Appendix B.2 compares these methods.

Cyclic depth folding changes both how Transformer depth is represented and how Transformer blocks are distributed. Shallow and deep roles coexist within the same block parameters, removing the usual one-to-one association between parameter groups and depth roles. Compared with replicated data parallelism, each worker stores one block group rather than the complete Transformer block stack, reducing per-worker memory for Transformer block parameters and enabling training when the full stack does not fit on one device.

Our contributions are:

- **New training architecture.** We introduce Möbius Learning, a training architecture in which different data streams traverse shared Transformer block groups in cyclically shifted orders. Each group thereby learns to serve both shallow and deep roles, giving rise to depth-role superposition.
- **Learning with parameter reuse.** Depth-role superposition combines effectively with looping: at eight loop passes, both the origin-averaged and best-fixed-origin five-task accuracies exceed standard looping across all four evaluated language-model sizes (Table 3). To understand this advantage, we analyze a neural model of next-token prediction and prove that Möbius Learning with all cyclic starting points achieves lower population risk than standard looping (Theorem 1).
- **Learning with data reuse.** Averaged across five model sizes, the models selected from Möbius Learning runs achieve higher downstream accuracy than standard looping under both fourfold and eightfold data reuse. To understand how Möbius Learning benefits from reusing data, we study the same neural model with finite training data and prove that Möbius Learning with all cyclic starting points achieves lower prediction loss on unseen sequences than standard looping, with cyclic or random data reassignment yielding further gains (Theorem 2).

2 Methodology

Möbius Learning assigns one Transformer block group to each worker. The worker at which a batch originates determines that batch's cyclic starting point. The input embedding associated with that path is applied once, its hidden states circulate through the assigned block groups in worker-indexed cyclic order, and the corresponding prediction map is applied once at the end. Different originating workers therefore induce cyclic shifts of the same block groups.

2.1 Cyclic Depth Folding

Block groups and ordered computation. Let $\mathcal{X}, \mathcal{Z}, \mathcal{H}$, and $\mathcal{G}$ denote the spaces of input sequences, targets, hidden states, and vocabulary logits. We partition the Transformer into $P \geq 1$ groups F_i, indexed by $\mathbb{Z}_P = \{0, \ldots, P-1\}$. Each group contains one or more blocks and maps $\mathcal{H}$ to itself. Write $\langle a \rangle_P$ for the remainder of a modulo P, taking values in $\mathbb{Z}_P$.

An ordered Transformer consists of an input embedding $E : \mathcal{X} \to \mathcal{H}$, block-group maps $F_i : \mathcal{H} \to \mathcal{H}$, and a prediction map $R : \mathcal{H} \to \mathcal{G}$. Using $(A \circ B)(x) = A(B(x))$, one complete ordered pass is

$$\Psi = F_{P-1} \circ F_{P-2} \circ \cdots \circ F_0. \tag{1}$$

Thus, F_0 is applied first and F_{P-1} last.

Worker-indexed cyclic computation. Worker i owns F_i. For input x_s originating at worker s, the *originating-worker index* s specifies the first group in a complete cyclic pass:

$$\Psi_s = F_{\langle s+P-1 \rangle_P} \circ \cdots \circ F_{\langle s+1 \rangle_P} \circ F_s, \qquad \Psi_0 = \Psi. \tag{2}$$

The path produces logits $R_s(\Psi_s(E_s(x_s)))$, applying its embedding E_s and prediction map R_s once each. These subscripts identify the invoked maps without specifying whether endpoint parameters are shared or distinct. All paths use the same physical block groups.

Depth-role superposition. For the path from worker s, the *source-relative depth* of F_i counts preceding groups:

$$d_{s,i} = \langle i - s \rangle_P. \tag{3}$$

For $P = 4$ and $s = 2$, the order F_2, F_3, F_0, F_1 gives $d_{2,0} = 2$. As s ranges over $\mathbb{Z}_P$, each group occupies every within-pass depth exactly once: for fixed i, the map $s \mapsto d_{s,i}$ is a bijection with inverse $s = \langle i - d \rangle_P$. Joint training therefore optimizes the same group parameters across these depth roles, a phenomenon we call *depth-role superposition*.

2.2 OBJECTIVES

Let $\mathcal{Q} = (\mathcal{Q}_s)_{s=0}^{P-1}$ denote the collection of worker-local data distributions over input–target pairs (x, y). Let $\mathrm{CE} : \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$ denote next-token cross-entropy, averaged over target positions in each sequence.

Objective. For Möbius Learning, we define

$$\mathcal{L}_{\text{Möbius}}(\mathcal{Q}) = \frac{1}{P} \sum_{s=0}^{P-1} \mathbb{E}_{(x,y)\sim\mathcal{Q}_s} \left[\mathrm{CE}(R_s(\Psi_s(E_s(x))), y) \right]. \tag{4}$$

The standard objective $\mathcal{L}_{\text{standard}}(\mathcal{Q})$ uses the same expression with (E_s, Ψ_s, R_s) replaced by (E, Ψ, R), where E and R are shared across workers.

Parameter dependence is omitted for brevity. Training minimizes the corresponding objective on the training distributions $\mathcal{D} = (\mathcal{D}_s)_{s=0}^{P-1}$ with respect to the model parameters. Evaluation computes the same loss on the validation distributions $\mathcal{V} = (\mathcal{V}_s)_{s=0}^{P-1}$, with the trained parameters held fixed. Both objectives average scalar losses across workers, not logits or predictive distributions.

Each originating worker backpropagates its path-specific loss on a distinct local minibatch. Equal participation gives the source-mean weighting above; gradients from all paths using a block group contribute to its update, with each F_i participating at source-relative depth $d_{s,i}$.

Raw input–target pairs remain at their originating workers; hidden states and their gradients cross worker boundaries. Further implementation details appear in Appendix C.

Both studies evaluate completed paths and report the mean and best-fixed-origin metrics.

2.3 LEARNING UNDER PARAMETER REUSE

Recent work on looped and recurrent-depth language models explores scaling computation through repeated block applications without increasing model parameter count (Geiping et al., 2025; Zhu et al., 2025; Prairie et al., 2026).

Depth-role superposition shares block parameters across depth roles, while looping reuses the block sequence across repeated passes. We study whether these two forms of parameter reuse combine effectively by comparing standard looping and Möbius Learning across loop depths.

Repeated passes. A standard looped Transformer repeats the same block sequence in a fixed order. For an integer $L \geq 1$, standard looping composes L complete passes:

$$\Psi^L = \underbrace{\Psi \circ \cdots \circ \Psi}_{L \text{ times}}, \qquad R \circ \Psi^L \circ E : \mathcal{X} \to \mathcal{G}. \tag{5}$$

We call L the *loop depth*: the number of complete Transformer block-sequence passes. These passes comprise $K = LP$ successive block-group applications. The input embedding and prediction map are each applied once and are not counted in L or K. The case $L = 1$ recovers the standard ordered stack.

Cyclic passes. For any integer $K \geq 1$, initialize $h_{s,0} = E_s(x_s)$ and define its path by

$$h_{s,t+1} = F_{\langle s+t \rangle_P}(h_{s,t}), \qquad t = 0, \ldots, K-1. \tag{6}$$

The index t counts block-group computations along network depth, not token positions. The terminal state produces logits $R_s(h_{s,K}) \in \mathcal{G}$. Standard looping instead selects $F_{\langle t \rangle_P}$ at every origin. Equivalently,

$$\Phi_{s,K} = F_{\langle s+K-1 \rangle_P} \circ \cdots \circ F_{\langle s+1 \rangle_P} \circ F_s. \tag{7}$$

The resulting logits are $R_s(\Phi_{s,K}(E_s(x_s)))$; E_s and R_s are each applied exactly once regardless of K. Although the map is defined for arbitrary K, the subsequent analysis and all reported experiments use complete passes, $K = LP$. Periodicity gives

$$\Phi_{s,LP} = \Psi_s^L, \qquad \Psi_0 = \Psi, \qquad \Phi_{0,LP} = \Psi^L. \tag{8}$$

Indeed, $\langle s + t + P \rangle_P = \langle s + t \rangle_P$, so each successive set of P block-group applications forms the same pass Ψ_s. For $s \neq 0$, Ψ_s generally differs from Ψ, since the group maps need not commute.

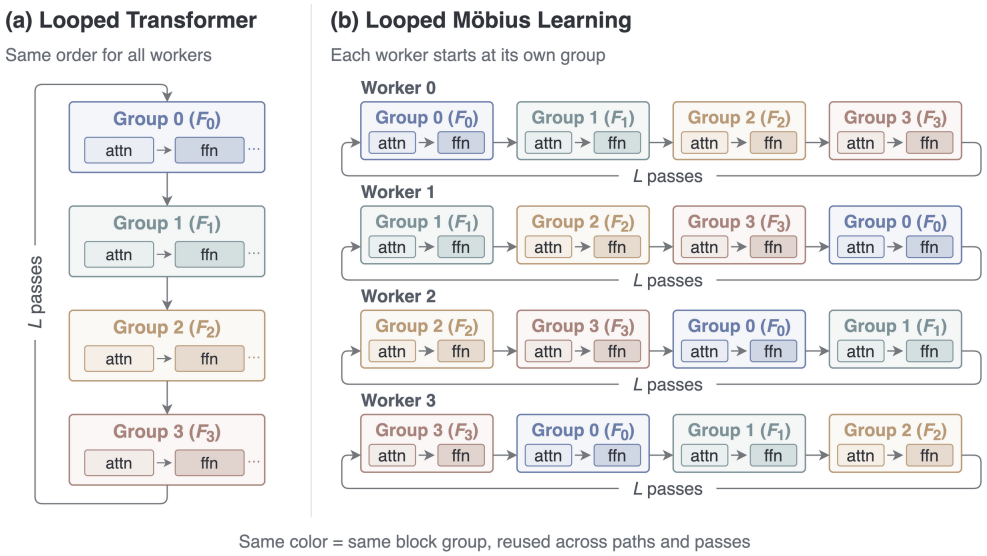


Figure 3: **Standard looping and Möbius Learning for $P = 4$. Left:** each worker-local batch repeats F_0, F_1, F_2, F_3 for L passes. **Right:** row s shows a batch originating at worker s, starting at F_s and repeating the cyclic order for L passes. Each path comprises $K = LP$ block-group applications. Rows represent logical paths through shared block groups, not separate copies of those groups.

For $s = 0$, the group application order matches standard looping, but Möbius Learning jointly trains paths from all origins. At a fixed origin, each group repeats its within-pass role across passes; across origins, it occupies all P roles. Figure 3 illustrates how Möbius Learning combines cyclic starting points with repeated loop passes, alongside standard looping. Explicit orders and block-call indices are given in Appendix C.1.

All paths reuse the same physical block groups. Each F_i is reused across passes and originating-worker paths, while distinct groups F_i and F_j need not have identical parameters. Thus depth-role superposition does not impose exact tying between groups. An abstract comparison with model-loop and layer-loop sharing schedules appears in Appendix B.1.

Objective at loop depth L. Replacing Ψ_s in Equation (4) by its L-pass counterpart gives

$$\mathcal{L}_{\text{Möbius}}(\mathcal{Q}; L) = \frac{1}{P} \sum_{s=0}^{P-1} \mathbb{E}_{(x,y)\sim\mathcal{Q}_s} \left[\mathrm{CE}\big(R_s\big(\Psi_s^L(E_s(x))\big), y\big) \right]. \tag{9}$$

The standard objective $\mathcal{L}_{\text{standard}}(\mathcal{Q}; L)$ uses the same expression with (E_s, Ψ_s^L, R_s) replaced by (E, Ψ^L, R), where E and R are shared across workers.

The case $L = 1$ recovers the objectives in Subsection 2.2. Training minimizes the corresponding objective with $\mathcal{Q} = \mathcal{D}$; evaluation computes it with $\mathcal{Q} = \mathcal{V}$.

Experimental comparison. We pretrain Tiny, Mini, and Small MoE variants of modded GPT-2 (Radford et al., 2019) on FineWeb (Penedo et al., 2024), using eight experts per block with top-2 routing. Each 12-block backbone comprises four three-block groups. We compare Möbius Learning and standard looping at $L \in \{1, 2, 4, 8\}$, matching the backbone, token budget, optimizer configuration, learning-rate schedule, and global batch at each loop depth. Both methods perform $K = LP$ block-group applications per path, so compute matching is within each loop depth. Model dimensions and parameter counts appear in Table 1; training budgets and protocols are detailed in Appendix C and Appendix D. We report FineWeb validation loss, WikiText-103 validation loss on a train-split sample (Merity et al., 2016), and mean zero-shot accuracy over ARC-Easy (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2019), SciQ (Welbl et al., 2017), and CommonsenseQA (Talmor et al., 2018). Mean averages the reported metric across cyclic starting points. Best reports the highest five-task mean accuracy or the lowest loss among these starting points,

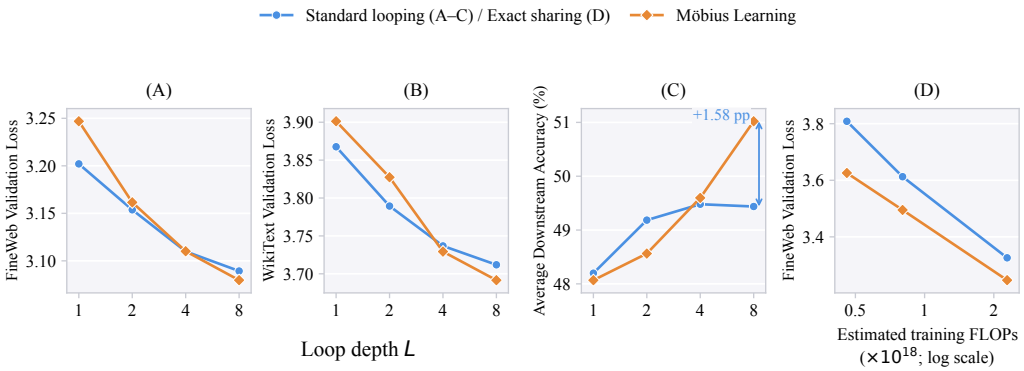


Figure 4: Parameter-reuse pretraining on FineWeb (2.67B tokens per run). (A–C) Small MoE at $L = 1, 2, 4, 8$: FineWeb validation loss, WikiText validation loss, and five-task mean accuracy, respectively. The arrow in (C) shows the best-origin gain over standard looping in percentage points. (D) FineWeb validation loss for exact sharing versus untied Möbius Learning at the $L = 1$ application depth, across Tiny, Mini, and Small (left to right); paired models have equal nominal FLOPs but different independent parameter counts. Blue denotes exact sharing in (D) only; orange denotes the best fixed cyclic origin. The full Small comparison, configurations, and evaluation details appear in Figure 6 and Appendix D.

selecting one fixed origin for each metric. Evaluation tasks and the post-hoc selection protocol are detailed in Appendix C and Appendix D.

Figure 4 shows a larger advantage at greater loop depth for the Small model: from $L = 4$ to $L = 8$, the best-origin accuracy gain over standard looping grows from 0.12 to 1.58 percentage points, while its FineWeb and WikiText losses decrease. At eight loop passes, both the origin-averaged and best-fixed-origin five-task accuracies exceed standard looping across all four evaluated model sizes, from Tiny to Medium (Table 3). Appendix D provides the remaining comparisons across model sizes and loop depths, together with per-task accuracies in Table 3.

Is Depth Folding Implicit Looping? One possible explanation for the gains above is that symmetric depth roles encourage different block groups to learn similar weights, implicitly approximating repeated applications of a shared core. We performed an exact-sharing ablation to test the strict form of this hypothesis, replacing the four distinct three-block groups with one three-block core applied four times. This ablation follows the preceding $L = 1$ training setup across Tiny, Mini, and Small, retaining twelve block applications and the same estimated training FLOPs at each scale (approximately 0.5, 0.8, and 2.3×10^{18} FLOPs, respectively; implementation details in Appendix C). In Figure 4 (D), exact sharing increases FineWeb validation loss by 0.183, 0.118, and 0.079 relative to the untied model's best cyclic origin, respectively. Thus, exact repetition of one shared core does not recover Möbius Learning's performance at these scales. This motivates our controlled-model analysis of learning dynamics induced by training across depth roles in Section 3.

2.4 Learning under Data Reuse

As language-model training scales, the limited availability of fresh training data makes effective reuse of existing corpora increasingly important. Repeated-data training yields diminishing returns (Muennighoff et al., 2025), motivating stronger regularization (Kim et al., 2026) and looped weight sharing (Chen et al., 2026); Appendix B.3 discusses these approaches.

Depth-role superposition requires the same parameters to serve multiple depth roles. Does this shared learning become more demanding when repeated exposure to a limited corpus provides a weaker learning signal? Here data reuse concerns repeated training exposure, not additional loop passes.

Using the same MoE architecture family and total token budget, we study data reuse across five model sizes, from Tiny to Large, with one complete block-stack pass per input ($L = 1$). We repeat an approximately 668M-token window four times or a 334M-token window eight times, reassigning repeated examples across cyclic starting points in Möbius Learning. We report the same five-task

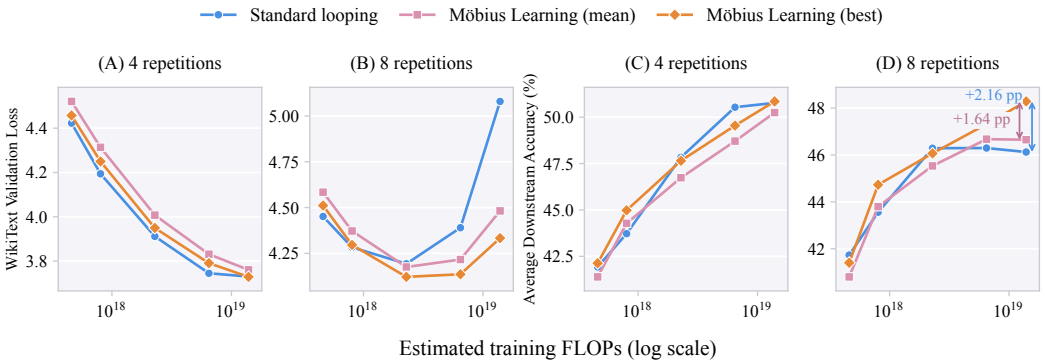


Figure 5: Data-constrained FineWeb pretraining across five MoE scales at a fixed total token budget ($L = 1$). (A,B) WikiText-103 validation loss on a train-split sample; (C,D) five-task mean accuracy. (A,C) use four repetitions of an approximately 668M-token window; (B,D), eight repetitions of an approximately 334M-token window. Blue and pink arrows show Best-minus-standard-looping and Best-minus-Mean accuracy gaps in percentage points, respectively. Best selects one fixed origin separately for each aggregate metric, post hoc; Mean averages origin-wise metrics. Model specifications and training/evaluation details appear in Appendix C.

zero-shot accuracy and WikiText validation loss. Model configurations and the repetition schedule are given in Appendix C.

Under eightfold data reuse, the selected Möbius Learning model's five-task accuracy increases with model size, whereas standard looping levels off at Small and Medium and declines at Large. At 1.39×10^{19} estimated FLOPs, a single run's post-hoc Best fixed origin attains 48.29%, versus 46.13% for standard looping and 46.65% for Mean across origins (a 1.64-percentage-point Best–Mean gap; Figure 5). The fixed origin is selected on the same evaluation data; origin-specific predictors share the trained block weights, so selection requires evaluation but no additional training. Under eightfold data reuse, Möbius Learning mitigates diminishing downstream accuracy gains as models scale.

3 Theoretical Understanding of Cyclic Depth Folding

3.1 Theoretical Setting

The preceding experiments (Subsections 2.3 and 2.4) motivate two fundamental questions.

1. **Looping.** Why can training shared blocks across depth roles improve over standard looping?
2. **Data reuse.** How can Möbius Learning use a limited training corpus more effectively?

The complementary analysis of additional loop passes and its mechanism are presented in Subsection E.3.

To study these questions, we use a trainable neural network for next-token prediction, with binary observations and local causal blocks that make the training dynamics tractable.

The data follow a stationary hidden Markov model. Hidden states S_j and observed tokens X_j take values in $\{-1, 1\}$, with

$$\mathbb{P}(S_{j+1} = s' \mid S_j = s) = \frac{1 + \rho s s'}{2}, \qquad \mathbb{P}(X_j = x \mid S_j = s) = \frac{1 + \kappa x s}{2}. \tag{10}$$

Here ρ controls how strongly the hidden state persists, and $\kappa > 0$ controls how reliably a token reveals that state. The learner observes only the prefix $X_{1:j}$ and predicts X_{j+1}. Earlier tokens help infer the hidden state, which in turn informs the next-token prediction. Each sequence supplies T prediction positions. We fix $B \geq 2$ physical blocks and $T \geq 4$, take $p = \kappa^2$ in a compact subset of $(0, 1)$, and analyze weak temporal dependence, $\rho \downarrow 0$.

The network uses a shared token embedding, causal residual MLP blocks, and a shared readout. Each block combines the current position's hidden state with the preceding position's state from the previous layer, so successive block calls carry information farther back in the prefix. We fix B physical blocks and allow $P_e \mid B$ equally spaced cyclic starting points. Standard looping uses one starting point ($P_e = 1$) while retaining all B blocks, regardless of worker count; full entrances have $P_e = B$. In the equal-sized P-group construction of Section 2, $P_e = P$; varying P_e can therefore require a different block partition. Each path makes L complete passes, giving $D_{\mathrm{blk}} = BL$ physical block applications. Layout details and a counting example are in Subsection E.1.

All parameters, including the embedding and readout, train by Euclidean gradient descent with sufficiently small step size η, starting from the exact paired reset: a symmetric initialization with zero temporal weights. Each reused block receives the sum of its occurrence-wise gradient contributions to the mean loss. Write $\mathcal{R}_{P_e,L}(N)$ for the expected next-token cross-entropy on an independent fresh sequence after N updates, averaged over the T positions and P_e entrance losses. Lower risk means better prediction of unseen tokens; this averages losses, not predictions. The first theorem and the supplementary loop-depth analysis use population gradients, computed from this expected loss; the data-reassignment result uses a finite training sample. Architecture, reset, and update details are in Subsection E.1. Below, $\asymp$ means the same asymptotic order, and $o(\rho^r)$ denotes a remainder smaller than ρ^r as $\rho \downarrow 0$.

3.2 Learning across Depth Roles

Does training across depth roles improve prediction? We compare independently trained standard looping and full entrances at the same early update, with backbone, loop depth, reset, and learning rate fixed. A positive difference means lower fresh-sequence loss for full-entrance training.

Theorem 1 (Population risk under multi-entrance training (Informal))**.** *For fixed finite L, independently train ordinary looping and full entrances from the same reset. At any common integer population-GD update N in the certified early window $\eta N \asymp \rho^{-2}\log(1/\rho)$,*

$$\mathcal{R}_{1,L}(N) - \mathcal{R}_{B,L}(N) = \frac{T-2}{2T} p^2(1-p)^4\rho^6 + o(\rho^6) > 0. \tag{11}$$

The complete version is given in Proposition E.39.

The gain comes from how learning at one depth role becomes useful at another. Full-entrance training reuses the first learned temporal rule across block positions, making a useful history coefficient available before standard looping learns its next distinct role. This mechanism provides a theoretical explanation for the advantage of Möbius Learning over standard looping observed in Subsection 2.3: training across depth roles improves how shared computation is learned, even when loop depth is unchanged.

Theoretically Optimal Cyclic Partitioning. Among legal, equally spaced cyclic layouts of our controlled model, the later-stage analysis identifies $P_e = B/3$, or three blocks per stage, as the unique population-risk minimizer under the conditions of Corollary E.30. For fixed B, T with $6 \mid B$ and $T \geq B + 2$, this holds uniformly over finite $L \geq 1$ for sufficiently small ρ, subject to the corollary's two nondegeneracy conditions. Layouts are trained independently from the same reset at the same loop depth and compared at each shared integer update in the later window $\eta N \asymp \rho^{-6}\log(1/\rho)$ specified in Equation (178).

For the twelve-block Tiny, Mini, and Small MoE backbones in our parameter-reuse experiments, this optimum gives four cyclic starting points, spaced three blocks apart. It therefore matches the four-stage partition used for these models (Appendix D), providing a controlled-model rationale for this architectural choice.

3.3 Reassigning Repeated Data across Entrances

With a fixed training corpus, later epochs revisit the same examples. We now train the same network on this finite sample rather than population gradients, and ask whether changing data–entrance assignments improves prediction on fresh sequences (Subsection 2.4).

Partition M independent complete sequences into P_e equal shards. Fixed binding (F) keeps assignments; cyclic reassignment (C) shifts them by one entrance per epoch; random reassignment (U) draws independent uniform whole-shard permutations each epoch after the shared first epoch. Partitions and schedules are independent of token values. All protocols use the same data union per update. Fresh-sequence risks $\mathcal{R}_A^{\text{test}}$, $A \in \{F, C, U\}$, average entrance losses and are denoted $\mathcal{T}_A$ in the appendix. When $P_e = B$, standard looping (O) processes the same data union at entrance zero, with single-entrance risk $\mathcal{R}_O^{\text{test}}$, denoted $\mathcal{T}_O$ in the appendix. Protocol details and a four-epoch example appear in Subsection E.14.

We analyze one-step cyclic and random reassignment as controlled counterparts to the antipodal schedule used in our experiments (Appendix C).

A positive $\mathcal{R}_F^{\text{test}} - \mathcal{R}_a^{\text{test}}$, $a \in \{C, U\}$, favors reassignment. Its leading term $\mathcal{A}_D \geq 0$ measures weighted squared differences between shard-level next-token statistics and their pooled mean (Equation (303)). The limit takes $\rho \downarrow 0$ and growing M under the stated sample and epoch conditions. Here $\Rightarrow$ denotes convergence in distribution, χ_d^2 a chi-squared variable with d degrees of freedom, and $o_{\mathbb{P}}(M^{-1})$ a remainder negligible relative to M^{-1} in probability. Probability includes the sampled sequences and, for U, its random schedule.

Theorem 2 (Test-Risk Reduction under Data Reassignment (Informal))**.** *For a fixed legal $P_e > 1$, use the same reset, step size, loop depth, data union, and preselected complete-epoch checkpoint in the common early window $\eta N \asymp \rho^{-2}\log(1/\rho)$. Under the sample and epoch-length conditions of Corollary E.41, both reassignment schedules satisfy*

$$\mathcal{R}_F^{\text{test}} - \mathcal{R}_a^{\text{test}} = \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}), \quad a \in \{C, U\}, \qquad M\mathcal{A}_D \Rightarrow \frac{\chi^2_{4(P_e-1)}}{2T}. \tag{12}$$

Both gains are positive with probability tending to one. When $P_e = B > 1$, the matched ordinary arm additionally obeys

$$\mathcal{R}_O^{\text{test}} - \mathcal{R}_F^{\text{test}} = \frac{T-2}{2T}p^2(1-p)^4\rho^6 + o_{\mathbb{P}}(\rho^6),$$
$$\mathbb{P}\{\mathcal{R}_O^{\text{test}} > \mathcal{R}_F^{\text{test}} > \max(\mathcal{R}_C^{\text{test}}, \mathcal{R}_U^{\text{test}})\} \longrightarrow 1. \tag{13}$$

The benefit comes from changing which depth roles repeatedly encounter a shard's accidental next-token patterns, without adding data. At full entrances ($P_e = B$), the theorem's high-probability ordering combines two distinct improvements at one checkpoint: fixed binding improves on standard looping at scale ρ^6, and reassignment improves on fixed binding at scale M^{-1}. These results provide a theoretical basis for the data-reuse gains in Subsection 2.4: Möbius Learning improves prediction over standard looping, while reassigning repeated data across starting points yields an additional benefit without increasing the training corpus. Cyclic and random schedules share the leading gain but are not ranked. The full result is in Corollary E.41.

4 Applications and Implications

Parameter and data reuse. Depth-role superposition combines effectively with parameter reuse through looping, supporting better performance through additional loop passes without increasing model size. The repeated-data results also make Möbius Learning particularly promising for training with repeated data, an increasingly important setting for scaling training with limited unique data.

Worker-local model ownership. Möbius Learning combines local data ownership with worker-local model capacity. Federated learning commonly keeps data local but gives each participant a full model replica (McMahan et al., 2017). Pipeline training partitions the model across workers and routes data through a designated input stage. Split learning keeps raw examples local and exchanges activations, usually through fixed split points or stage roles (Vepakomma et al., 2018; Thapa et al., 2022). In Möbius Learning, each worker forms batches from local data and stores one block group; hidden states circulate through the groups in cyclic order. This provides a path to collaborative training in which workers contribute local data and model capacity to jointly realize a deeper model.

AI USE STATEMENT

We used large language models to assist with drafting, polishing, and revising portions of the manuscript, including improving clarity, readability, and presentation. For the theoretical analysis, these tools assisted with formulating assumptions, exploring derivations, and developing, checking, and revising proofs. They also supported internal review of mathematical arguments and exposition. Generative AI tools assisted with preparing conceptual illustrations to explain the method.

ETHICS STATEMENT

Our research focuses on the design and theoretical understanding of cyclic depth folding in language-model training. Our experiments use the publicly available FineWeb corpus and established evaluation benchmarks. The findings contribute to understanding how shared parameters can learn across depth roles and how language models can reuse parameters and training data. Responsible downstream use of models trained with these methods should account for privacy, bias, and potential misuse.

REPRODUCIBILITY STATEMENT

Section 2 describes the model construction, training objectives, and experimental comparisons. Section C provides model configurations, training budgets, data-repetition schedules, and evaluation procedures, including the averaging and selection of cyclic starting points. Section 3 presents the theoretical results, with detailed assumptions, derivations, and proofs provided in Section E.

REFERENCES

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise LoRA. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.20672.

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron C. Courville, and Se-Young Yun. Mixture-of-Recursions: Learning dynamic recursive depths for adaptive token-level computation. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2507.10524.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Zixi Chen, Akshay Vegesna, Samip Dahal, and Andrew Gordon Wilson. How model growth, recursion, and boundary operators influence scaling exponents, 2026. URL https://arxiv.org/abs/2609.19107.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention, 2019. URL https://arxiv.org/abs/1906.04341.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Róbert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D. Manning. MoEUT: Mixture-of-experts universal transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0897. URL https://arxiv.org/abs/2405.16039.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers, 2019. URL https://arxiv.org/abs/1807.03819.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylO2yStDr.

Matthias Freiberger, Peter Kun, Anders Sundnes Løvlie, and Sebastian Risi. Layershuffle: Enhancing robustness in vision transformers by randomizing layer execution order, 2024. URL https://arxiv.org/abs/2407.04513.

Zitian Gao, Yilong Chen, Yihao Xiao, Xinyu Yang, Ran Tao, Joey Zhou, and Bryan Dai. Loop the loopies!, 2026. URL https://arxiv.org/abs/2607.16051.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach, 2025. URL https://arxiv.org/abs/2502.05171.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176. URL https://doi.org/10.52202/068431-2176.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. GPipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems*, volume 32, pp. 103–112, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/093f65e080a295f8076b1c5722a46aa2-Abstract.html.

Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. LoopFormer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.11451.

Atindra Jha, Margaret Li, Jure Leskovec, Percy Liang, and Luke Zettlemoyer. Data scarcity and model sparsity: Mixtures-of-experts overfit more to repeated data, 2026. URL https://arxiv.org/abs/2609.11917.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, et al. modded-nanogpt: Speedrunning the NanoGPT baseline, 2024. URL https://github.com/KellerJordan/modded-nanogpt.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79682ef54677a2cb3dfb8383b372037f-Abstract-Conference.html.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations, 2020. URL https://arxiv.org/abs/1909.11942.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2017. URL https://arxiv.org/abs/1602.05629.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Journal of Machine Learning Research*, 26(53):1–66, 2025. URL https://jmlr.org/papers/v26/24-1000.html.

Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. PipeDream: Generalized pipeline parallelism for DNN training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pp. 1–15, 2019. doi: 10.1145/3341301.3359646. URL https://doi.org/10.1145/3341301.3359646.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970. URL https://doi.org/10.52202/079017-0970. Datasets and Benchmarks Track.

Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models, 2026. URL https://arxiv.org/abs/2604.12946.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models, 2024. URL https://arxiv.org/abs/2404.02258.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. Subformer: Exploring weight sharing for parameter efficiency in generative transformers, 2021. URL https://arxiv.org/abs/2101.00234.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers, 2025. URL https://arxiv.org/abs/2502.17416.

Kristian Schwethelm, Daniel Rueckert, and Georgios Kaissis. How much is one recurrence worth? iso-depth scaling laws for looped language models, 2026. URL https://arxiv.org/abs/2604.21106.

Sho Takase and Shun Kiyono. Lessons on parameter sharing across layers in transformers. In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 78–90, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.5. URL https://aclanthology.org/2023.sustainlp-1.5/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge, 2018. URL https://arxiv.org/abs/1811.00937.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline, 2019. URL https://arxiv.org/abs/1905.05950.

Chandra Thapa, M. A. P. Chamikara, Seyit Camtepe, and Lichao Sun. SplitFed: When federated learning meets split learning, 2022. URL https://arxiv.org/abs/2004.12088.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data, 2018. URL https://arxiv.org/abs/1812.00564.

Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives, 2019. URL https://arxiv.org/abs/1909.01380.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

Zhiwei Xu, Shihao Wu, Hanseul Cho, Wei Hu, and Yixin Wang. Data-constrained language model pretraining: Improved regularization and scaling laws, 2026. URL https://arxiv.org/abs/2606.06888.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, et al. Scaling latent reasoning via looped language models, 2025. URL https://arxiv.org/abs/2510.25741.

APPENDIX CONTENTS

The appendix presents limitations and future work, additional related work, implementation and data-repetition details, parameter-reuse experiments, and the formal theory. The page numbers and section numbers below link to their destinations.

A Limitations and Future Work 15

B Additional Related Work 15

B.1 Recurrent-Depth Architectures and Sharing Schedules 15

B.2 Layer-Order Adaptation and Depth Sharing 16

B.3 Learning from Repeated Data 16

C Implementation Notes 18

C.1 Complete-Pass Orders and Block-Call Positions 22

D Parameter-Reuse Experiments 23

E Theory: Model, Results, and Proofs 28

E.1 Prediction task and physical model 29

E.2 Learning depth roles under full physical gradient descent 32

E.3 Independently trained depth and finite ladders 41

E.4 Proof of independently trained depth gains 43

E.5 Entrance count at the third training stage 51

E.6 Proofs for the common-window partition results 52

E.7 Path-selected head bias and the two-role risk coefficient 58

E.8 Proof of the path-selected two-role risk 60

E.9 Late calibration of the two-role predictor 65

E.10 Proof of late calibration and its consequences 66

E.11 Physical Data Response 70

E.11.1 Proof of the physical response 72

E.12 Ordinary Margin at a Common Checkpoint 78

E.12.1 Proof of the common-window margin 78

E.13 Cross-entropy energy bounds for empirical training 81

E.14 Data reassignment across entrances 85

E.14.1 Proof of the reassignment certificate 87

A Limitations and Future Work

Our experiments run on an eight-GPU machine and therefore use up to eight block groups, each defining a cyclic starting point. Future work should evaluate partitions with more than eight starting points and multi-node, multi-GPU deployments to measure how communication and scheduling affect efficiency with more pipeline stages. Although Möbius Learning is a training paradigm for sharing block parameters across depth roles, our empirical evaluation is limited to Transformer language models; vision experiments could examine how depth-role superposition affects learning across architectures and tasks.

B Additional Related Work

B.1 Recurrent-Depth Architectures and Sharing Schedules

Recurrent-depth architectures reuse learned transformations at multiple computational depths rather than assigning independent parameters to every depth. Universal Transformers (Dehghani et al., 2019) recurrently apply a shared self-attention and transition block, while ALBERT (Lan et al., 2020) and Subformer (Reid et al., 2021) study broader forms of cross-layer parameter sharing. Recent looped language models extend this idea to decoder-only pretraining by repeatedly applying a shared recurrent core or block stack (Geiping et al., 2025; Zhu et al., 2025; Prairie et al., 2026).

Recent architectures also vary how recurrent computation is allocated. Mixture-of-Recursions learns token-dependent recursion depths with recursion-aware key–value caching (Bae et al., 2025b), while LoopFormer uses shortcut modulation to support variable inference depth (Jeddi et al., 2026). Iso-depth scaling analyses quantify the parameter-equivalent value of recurrence (Schwethelm et al., 2026), complementing work on stable looped pretraining (Prairie et al., 2026). The parameter-equivalent value of recurrence is distinct from its benefit at matched training compute. Related theory studies the reasoning capabilities of looped Transformers (Saunshi et al., 2025).

The pattern of reuse also depends on the order in which shared blocks are applied. Takase & Kiyono (2023) compare Sequence, which assigns the same parameters to consecutive depth positions, with Cycle, which repeats one fixed multi-layer order, and Cycle(Rev), which reverses that order in the final traversal. MoEUT (Csordás et al., 2024) reports that cyclic interleaving of distinct mixture-of-experts layers performs better than consecutively repeating each layer. More recently, Loopie (Gao et al., 2026) distinguishes *model-loop*, which repeats the full block sequence, from *layer-loop*, which repeats each block before proceeding to the next; in its 6B-A0.6B comparison, layer-loop's average downstream score overtakes that of model-loop after approximately 1.2 trillion training tokens. These studies show that recurrent sharing depends not only on which parameters are reused, but also on the model-wide schedule that orders their reuse.

To make the recurrent unit explicit, we use *Layer* for Loopie and *Group* for this paper.

Application-order comparison

Layer-level recurrence in Loopie

model-loop: Layer 1 → Layer 2 → Layer 3 → Layer 1 → Layer 2 → Layer 3,
layer-loop: Layer 1 → Layer 1 → Layer 2 → Layer 2 → Layer 3 → Layer 3.

Block-group order and loop depth in this paper ($P = 4$)

Indices $0, 1, 2, 3$ denote Groups $0, 1, 2, 3$, respectively. Group i is a sequence of one or more Transformer blocks assigned to worker i; its composite map is F_i in Section 2.

Path	**Pass 1**	**Pass 2**		**Pass L**
standard looping (every origin)	$0 \to 1 \to 2 \to 3$	$0 \to 1 \to 2 \to 3$	$\cdots$	$0 \to 1 \to 2 \to 3$
Möbius Learning, $s = 0$	$0 \to 1 \to 2 \to 3$	$0 \to 1 \to 2 \to 3$	$\cdots$	$0 \to 1 \to 2 \to 3$
Möbius Learning, $s = 1$	$1 \to 2 \to 3 \to 0$	$1 \to 2 \to 3 \to 0$	$\cdots$	$1 \to 2 \to 3 \to 0$
Möbius Learning, $s = 2$	$2 \to 3 \to 0 \to 1$	$2 \to 3 \to 0 \to 1$	$\cdots$	$2 \to 3 \to 0 \to 1$
Möbius Learning, $s = 3$	$3 \to 0 \to 1 \to 2$	$3 \to 0 \to 1 \to 2$	$\cdots$	$3 \to 0 \to 1 \to 2$

Arrows denote within-forward application order. The upper comparison reproduces Loopie's three-layer, two-step example. In the lower comparison, each cell gives the within-pass block-group order; the Möbius Learning rows encode worker-indexed cyclic orders, while the columns repeat each order to loop depth L. Loopie layer-loop is included as a structural reference rather than an evaluated baseline. In particular, the $s = 0$ Möbius Learning path is order-equivalent to standard looping; the distinction lies in jointly training the complete family of worker-indexed cyclic orders.

A complementary approach retains the global schedule but relaxes strict sharing across depth. Relaxed Recursive Transformers (Bae et al., 2025a) repeat one fixed multi-layer block using Takase and Kiyono's CYCLE schedule and add layer-specific low-rank adapters at the corresponding effective-depth positions. The model thus preserves a common recurrent core while allowing depth-specific deviations.

These works vary which transformations are shared, how recurrent depth is allocated and ordered, and whether depth-specific capacity is added. Möbius Learning jointly trains the complete set of P originating-worker-indexed cyclic shifts over P distinct block groups. The standard-looping baseline repeats one fixed group order for every originating worker, whereas Möbius Learning lets origin s select shift s. The defining distinction of Möbius Learning is exact, balanced source-relative-depth coverage: across all P origins, every group occupies each within-pass source-relative depth exactly once.

B.2 LAYER-ORDER ADAPTATION AND DEPTH SHARING

LayerShuffle (Freiberger et al., 2024) fine-tunes fixed-order pretrained Vision Transformers under randomly permuted block orders to improve robustness to execution-order perturbations and layer removal at inference. Its results show that block outputs and predictive contributions can adapt to execution position, although this robustness is accompanied by lower accuracy under the original fixed order. The relation to Möbius Learning can be separated along three dimensions:

Training stage. LayerShuffle begins with blocks already optimized under one fixed order and then adapts them through fine-tuning. Möbius Learning instead imposes cyclic block orders throughout language-model pretraining, without a preceding fixed-order stage. It therefore addresses the more demanding joint problem of learning the language-modeling task while training the same block parameters in shallow and deep roles.

Effective depth. LayerShuffle applies each distinct block once within a randomly permuted forward pass. Möbius Learning repeatedly traverses the complete block sequence, reaching $K = LP$ block-group applications along each input path and evaluating the construction across a broad range of loop depths.

Order structure and coverage. LayerShuffle samples arbitrary block permutations independently across minibatches. Möbius Learning uses the P deterministic cyclic shifts selected by the originating-worker index. Across these paths, every block group occupies every within-pass source-relative depth exactly once, yielding balanced supervision across depth roles.

These differences lead to a distinct empirical question. LayerShuffle asks whether a pretrained model can remain functional under arbitrary order perturbations, whereas Möbius Learning asks whether balanced joint pretraining across depth roles can improve standard looping under matched token budgets and loop depths. The resulting evidence goes beyond order robustness: at eight loop passes, Möbius Learning achieves higher mean downstream accuracy than standard looping across all four evaluated model sizes, from Tiny to Medium (Figure 1 (A)).

B.3 LEARNING FROM REPEATED DATA

Repeated-data pretraining separates the amount of unique data from the number of token presentations. Muennighoff et al. (2025) characterize diminishing returns from repeated exposure and model their interaction with parameter count. Kim et al. (2026) study stronger regularization and parameter scaling under a fixed original-data budget. Xu et al. (2026) add an auxiliary next-token prediction loss on randomly masked inputs. For MoE models, Jha et al. (2026) report greater repetition-induced overfitting than in dense baselines and investigate dropout and output masking as mitigations.

Chen et al. (2026) report that recursive weight sharing provides effective regularization in multi-epoch training. Möbius Learning approaches parameter reuse along a different dimension: training shared blocks across depth roles. Surprisingly, this construction also mitigates diminishing returns from model scaling in downstream accuracy under data reuse (see Figure 5).

C Implementation Notes

All reported models are trained from random initialization. Möbius Learning uses cyclic block orders from the start, without a preceding fixed-order training stage.

The exact-sharing comparison in Figure 4 (D) uses MoE models with three independent blocks applied four times, versus four distinct three-block groups in the untied construction. Both perform twelve block applications. Sharing reduces the number of independent block parameters; this ablation tests exact sharing, not whether the learned untied weights are approximately similar. FineWeb validation reports the lowest cyclic-origin loss for untied Möbius Learning; exact sharing has a single prediction path. Tiny and Mini share the recorded validation index hash across constructions. For Small, the recorded evaluation settings match, but the original untied validation index hash was not recovered. The MoE configurations and evaluation procedures are specified below.

MoE Architecture. We build on the modded GPT architecture of Jordan et al. (2024), using five MoE model scales with approximately 41M to 2.19B total parameters. Table 1 lists their Transformer block counts, hidden widths, and total parameter counts.

Following Mixtral's sparse MoE design (Jiang et al., 2024), each block has eight experts and activates two per token. Each block also contains bias-free causal self-attention. Our experts use two projections of dimensions $d \times 2d$ and $2d \times d$, with squared-ReLU activation. All models use 128-dimensional attention heads, RoPE, affine RMSNorm, and query/key normalization. Input embeddings and the output projection share a $50{,}304 \times d$ weight matrix.

Table 1: Architectures and total learned parameters for the five FLOPs points in Figure 5. Parameter counts are in millions; the counting convention is given below. Möbius counts correspond to the fourfold data-reuse configuration; standard looping has the same count for both repetition schedules. FLOPs are nominal training estimates, shared by both methods at each scale.

				Total parameters (M)		
Scale	Blocks	Width	Heads	Standard	Möbius	FLOPs (10^{18})
Tiny	12	256	2	41.25	41.31	0.459
Mini	12	384	3	83.10	83.20	0.802
Small	12	768	6	293.60	294.08	2.285
Medium	24	1024	8	957.85	958.20	6.479
Large	36	1280	10	2188.40	2188.93	13.902

The counts in Table 1 are derived from the implemented parameter shapes and include all experts, routers, normalization, and auxiliary trainable parameters. Shared parameters are counted once, excluding redundant distributed storage copies.

Training budget. We set a common training budget of approximately 2.67B token presentations, using the Small model's approximately 124M active parameters as the reference. This gives about 22 tokens per active parameter, close to the Chinchilla-optimal reference of roughly 20 (Hoffmann et al., 2022), used here as an approximate budgeting rule. Each run performs 5,100 updates with global batch 512 and context length 1,024 ($D = 2{,}673{,}868{,}800$). This budget is fixed across all five scales and both repetition schedules and counts repeated presentations rather than unique data. The active-parameter estimate follows the parameter-dependent term in the FLOPs calculation below.

For the caption-level active-parameter estimates, $N_{\text{active}} \approx Vd + 12nd^2$ counts the shared vocabulary matrix, attention projections, and two active experts per block. This gives approximately 22M, 41M, 124M, 354M, and 772M for Tiny through Large; smaller routing, normalization, and depth-conditioning terms are omitted. These are nominal per-pass estimates, not the total parameters across all experts, and are not multiplied by loop depth.

The data-reuse runs use $L = 1$: increasing compute here changes model size, not the number of block-stack passes. We evaluate 20 final checkpoints: five scales, two repetition schedules, and two methods. These are single-run comparisons, with no across-seed uncertainty estimate.

Training FLOPs. We use the nominal forward–backward estimate

$$C = 6D(12nd^2 + Vd) + 12DnSd,$$

where n and d are the block count and width in Table 1, $S = 1{,}024$ is context length, and $V = 50{,}304$ is the padded vocabulary size. Attention projections contribute $4nd^2$ weights and the two active experts contribute $8nd^2$, giving the $12nd^2$ term. This proxy omits routing, auxiliary-module computation, elementwise operations, optimizer updates, activation recomputation, communication, and evaluation; its attention term is approximate. It is not a measurement of hardware work. Since D already includes repeated presentations, C is not multiplied by the repetition count again.

Optimization. The data-reuse runs use Muon with base and Muon learning rates of 0.003. We selected a learning rate of 3×10^{-3} through sweeps on the GPT-2 Small-based MoE model, with similar preferred learning rates observed across loop depths. We use this setting across model sizes to maintain a consistent optimization setup in the scaling comparison. The warmup–stable–decay schedule uses 1,024 warmup updates and a final 1,500-update linear decay toward 0.1 times the peak learning rate, within the same training budget.

Computational resources. We implement our models in PyTorch and conduct training on NVIDIA A100 GPUs with 80 GB of memory per GPU, using multiple GPUs for distributed execution.

Data reuse strategy. Unless otherwise specified, training does not deliberately repeat data windows. To construct the data-constrained settings, we retain the Small-referenced training budget above and restrict the data window available for repeated training. Dividing the 5,100-update budget into four equal segments gives 1,275 updates per segment: a window of approximately 668M token positions is replayed four times, resetting to its beginning between segments. For eightfold repetition, we use a window of approximately 334M token positions and alternate 638- and 637-update segments to preserve the same total budget. Both settings therefore use approximately 2.67B token presentations; they differ in the size of the replayed window and the number of repetitions. Window sizes count token positions, not deduplicated text.

Standard looping reuses the fixed block order. For Möbius Learning, repeated data are reassigned to cyclic starting points in antipodal order. With four block groups, the offsets are $(0, 2, 1, 3)$; with eight, they are $(0, 4, 1, 5, 2, 6, 3, 7)$. Four-repetition runs use four groups, except Small, which uses eight groups and offsets $(0, 4, 1, 5)$. All eight-repetition runs use eight groups. In each longer eight-repetition segment, the extra update uses the legacy cyclic-shift offset: $0, 6, 4, 2$ for the four such updates. Thus the four-repetition curve combines two block partitions, rather than isolating group count as a controlled variable.

WikiText Validation Loss. We construct a large validation sample from the WikiText-103 training split (Merity et al., 2016), evaluating next-token negative log-likelihood on 10,485,760 predicted tokens with context length 1,024. We use this larger evaluation sample to reduce sensitivity to finite-sample variation, with identical frozen windows across checkpoints. Here, validation denotes the evaluation role of this train-split sample, not the official WikiText validation or test partition; those partitions are not included. This validation loss assesses out-of-domain language modeling relative to FineWeb pretraining. WikiText is not used for optimization in these runs; overlap with the web pretraining corpus has not been audited.

Downstream tasks. All tasks use zero-shot likelihood scoring, without task-specific fine-tuning. ARC-Easy (Clark et al., 2018) tests grade-school science question answering (2,376 test questions). HellaSwag (Zellers et al., 2019) asks for a plausible continuation of an everyday event (10,042 validation examples). PIQA (Bisk et al., 2019) evaluates physical commonsense by choosing between two proposed solutions (1,838 validation examples). SciQ (Welbl et al., 2017) contains science multiple-choice questions; we include the provided support passage (1,000 test examples). CommonsenseQA (Talmor et al., 2018) asks five-way commonsense questions (1,221 validation examples), scored by answer-text likelihood.

For each question, we choose the candidate with the highest summed token log-probability, without length normalization; ties use the first candidate. Evaluation preserves the native sequence-length buckets with an 8,192-token scoring budget and no prompt truncation. The reported accuracy is the equally weighted mean over the five complete task splits, not an example-weighted average. Gold-answer Micro NLL was also collected for the first four tasks but is not plotted here.

Mean and best-fixed-origin metrics. For each checkpoint, we evaluate every cyclic starting point separately. We first compute each origin's mean accuracy across the five downstream tasks. Mean averages these scores across origins, whereas Best reports the score of the highest-scoring origin. The per-task results for Best all come from that same origin. For loss metrics, Mean averages the origin-wise losses and Best reports the lowest loss; the selected origin may differ between accuracy and loss. Origin selection uses the reported evaluation data and requires no additional training.

Table 2 reports the individual task accuracies underlying Figure 5, together with their five-task average.

Table 2: Per-task downstream accuracy (%; higher is better) at step 5,100. Each scale contains four- and eight-repetition comparisons. Mean averages cyclic-origin accuracies; Best uses one fixed origin selected by the five-task average, not a separate origin per task. SciQ includes support passages; CSQA denotes CommonsenseQA. Avg. weights the five tasks equally. Bold marks the highest displayed score within each scale and repetition group. All entries are single-run results.

Repeat	Method	ARC-Easy	HellaSwag	PIQA	SciQ	CSQA	Avg.
Tiny							
4	Möbius Learning (mean)	36.04	27.33	60.36	60.92	22.32	41.39
	Möbius Learning (best)	**36.41**	27.55	**61.81**	**62.20**	22.69	**42.13**
	Standard looping	36.20	**27.56**	61.43	61.70	**22.77**	41.93
8	Möbius Learning (mean)	35.17	27.13	60.60	60.00	21.09	40.80
	Möbius Learning (best)	35.06	27.16	**61.04**	**62.20**	21.54	41.40
	Standard looping	**37.37**	**27.41**	60.61	60.80	**22.44**	**41.73**
Mini							
4	Möbius Learning (mean)	39.56	27.85	62.47	66.22	25.29	44.28
	Möbius Learning (best)	**40.45**	27.92	62.73	**67.60**	**26.21**	**44.98**
	Standard looping	39.39	**28.01**	**62.79**	65.00	23.42	43.72
8	Möbius Learning (mean)	38.02	27.97	62.14	66.70	24.14	43.79
	Möbius Learning (best)	38.51	28.12	62.46	**69.50**	**25.06**	**44.73**
	Standard looping	**39.14**	**28.14**	**62.51**	63.10	24.90	43.56
Small							
4	Möbius Learning (mean)	42.88	29.46	64.35	69.46	27.53	46.74
	Möbius Learning (best)	**44.19**	29.58	64.20	71.30	**28.99**	47.65
	Standard looping	43.14	**30.14**	**65.07**	**71.90**	28.91	**47.83**
8	Möbius Learning (mean)	41.16	29.37	64.19	67.78	25.17	45.53
	Möbius Learning (best)	**42.00**	29.43	**64.36**	68.50	26.04	46.07
	Standard looping	39.81	**29.71**	64.20	**70.70**	**27.03**	**46.29**
Medium							
4	Möbius Learning (mean)	45.45	31.56	66.97	70.62	28.93	48.71
	Möbius Learning (best)	46.13	31.98	67.68	72.30	29.65	49.55
	Standard looping	**47.18**	**33.15**	**68.72**	**73.50**	**30.14**	**50.54**
8	Möbius Learning (mean)	41.06	30.95	65.04	68.83	27.50	46.67
	Möbius Learning (best)	**41.12**	31.52	**66.00**	**69.70**	**28.26**	**47.32**
	Standard looping	40.66	**32.08**	65.72	66.90	26.13	46.30
Large							
4	Möbius Learning (mean)	46.84	32.94	67.23	73.10	31.12	50.25
	Möbius Learning (best)	**47.64**	33.38	67.68	73.60	**31.94**	**50.85**
	Standard looping	45.58	**35.51**	**67.90**	**74.80**	30.06	50.77
8	Möbius Learning (mean)	41.71	31.96	64.62	66.88	28.09	46.65
	Möbius Learning (best)	**44.02**	32.69	**65.29**	**70.60**	28.83	**48.29**
	Standard looping	40.11	**33.34**	64.36	63.50	**29.32**	46.13

C.1 Complete-Pass Orders and Block-Call Positions

For $P = 4$ and $L = 2$, the forward application orders, with a semicolon separating passes, are

$$\begin{aligned} \text{standard looping, and Möbius Learning at } s = 0: &\quad (F_0, F_1, F_2, F_3)\,;\,(F_0, F_1, F_2, F_3), \\ \text{Möbius Learning at } s = 2: &\quad (F_2, F_3, F_0, F_1)\,;\,(F_2, F_3, F_0, F_1). \end{aligned} \tag{14}$$

At the level of group application order, the first row coincides with standard looping. The distinction is joint training of the full family of worker-indexed cyclic orders; the $s = 1$ and $s = 3$ paths complete that family.

Within a pass, F_i has the source-relative depth $d_{s,i}$ defined in Equation (3). Across L passes, its zero-indexed application positions are

$$d_{s,i} = \langle i - s \rangle_P, \qquad \mathrm{Pos}_{s,L}(i) = \{d_{s,i} + \ell P : \ell = 0, \ldots, L-1\}. \tag{15}$$

For example, $\mathrm{Pos}_{2,3}(0) = \{2, 6, 10\}$ when $P = 4$. A fixed origin repeats the same within-pass role, whereas across origins every group is trained at all P roles. Standard looping places F_i at $\{i + \ell P : \ell = 0, \ldots, L-1\}$ for every origin.

D PARAMETER-REUSE EXPERIMENTS

The Tiny, Mini, Small, and Medium backbone dimensions are listed in Table 1 in Appendix C; the accompanying text gives the architecture and parameter-counting convention.

Extending the training-compute estimate in Appendix C to L passes gives $C_L = 6D(12nd^2L + Vd) + 12DnSdL$: block computation scales with L, while the vocabulary projection is applied once. Figure captions round this nominal estimate to one decimal place in units of 10^{18} FLOPs.

We evaluate 24 final checkpoints: three model scales, four loop depths, and two methods. Unlike the data-reuse experiments, these runs do not deliberately replay training windows. Each backbone has twelve blocks and four block groups ($P = 4$); $L = 1, 2, 4, 8$ therefore corresponds to $K = 4, 8, 16, 32$ block-group applications. All runs use global batch 512, context length 1,024, and 5,100 updates. Both the base and Muon learning rates are 0.003, using the warmup–stable–decay schedule described in Appendix C. Increasing L reuses the full backbone, rather than tying the parameters of distinct block groups.

FineWeb validation and WikiText-103 validation on a train-split sample each score 10,485,760 predicted tokens. For Möbius Learning, all four origins score the same fixed evaluation sample before their scalar losses are averaged or the lowest-loss origin is selected. The WikiText sample and five-task scoring procedure are those described in Appendix C. Downstream Best uses one fixed origin chosen by the equally weighted five-task accuracy, not a separate origin per task. Selection is post hoc on the reported scores. All comparisons use one run per configuration; no across-seed uncertainty is estimated.

Figures 7 and 8 compares model sizes at each loop depth; Figure 9 complements the Small results in Figure 4 with the Tiny and Mini depth comparisons. Table 3 reports the individual task scores and their mean. This full depth grid covers Tiny, Mini, and Small.

We additionally evaluate a Medium pair at $L = 8$, using the same token budget, optimizer settings, and five-task evaluation protocol. The Medium backbone has 24 blocks and four block groups; its nominal training compute is 46.0×10^{18} FLOPs. Standard looping attains 53.776% five-task accuracy, compared with 54.342% averaged across the four cyclic origins and 54.381% at the fixed origin selected by the five-task mean. Per-task scores for this pair are included in Table 3. These two additional runs extend the $L = 8$ scale comparison, not the full depth grid; Large is not included in the parameter-reuse experiments.

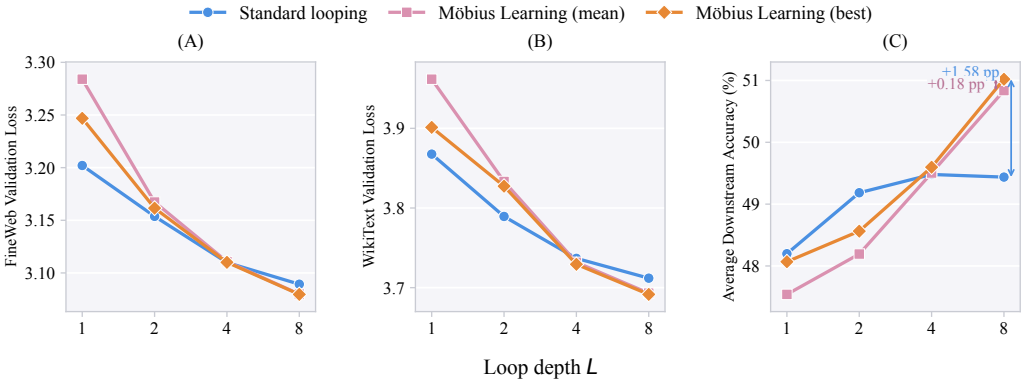


Figure 6: Parameter-reuse pretraining on FineWeb for Small, including the mean across cyclic origins. The MoE has approximately 294M total parameters and 124M active parameters per token, trained on 2.67B tokens at $L = 1, 2, 4, 8$ (estimated training FLOPs: 2.3, 4.0, 7.3, and 13.9×10^{18}, respectively). (A) FineWeb validation loss; (B) WikiText validation loss; (C) average downstream accuracy over five tasks. Endpoint arrows show Best minus standard looping (blue) and Mean (pink), in percentage points. This is the full version of Figure 4.

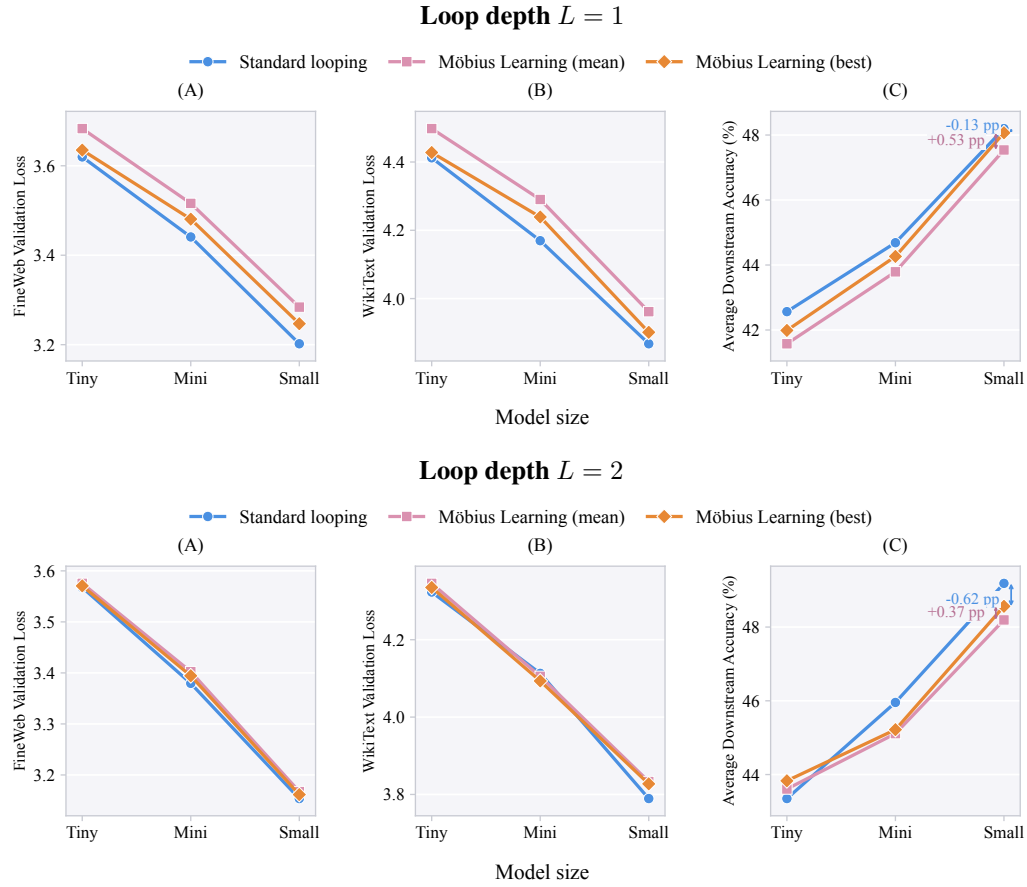


Figure 7: Parameter-reuse pretraining on FineWeb at $L = 1$ and $L = 2$. Tiny, Mini, and Small MoEs span approximately 41M–294M total parameters and 22M–124M active parameters per token, each trained on 2.67B tokens (estimated training FLOPs: $(0.5\text{–}2.3) \times 10^{18}$ at $L = 1$ and $(0.7\text{–}4.0) \times 10^{18}$ at $L = 2$). (A) FineWeb validation loss; (B) WikiText validation loss; (C) average downstream accuracy over five tasks. Endpoint arrows show Best minus standard looping (blue) and Mean (pink), in percentage points.

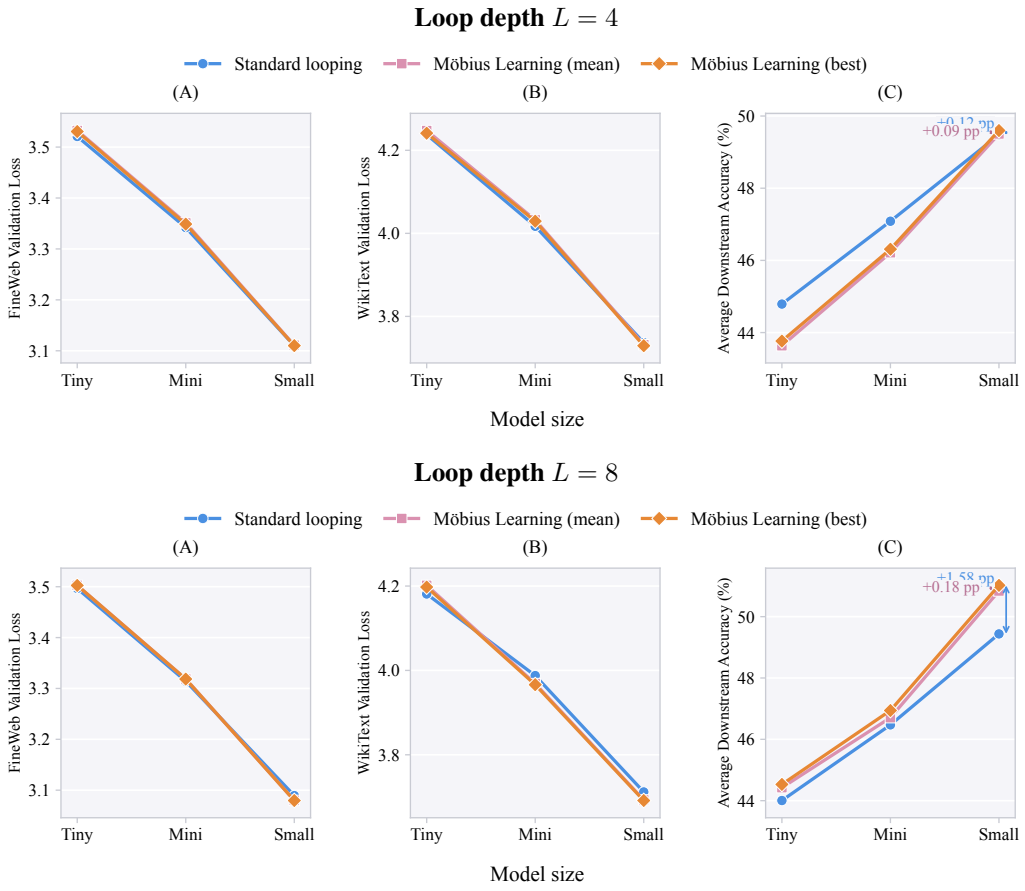


Figure 8: Parameter-reuse pretraining on FineWeb at $L = 4$ and $L = 8$. Tiny, Mini, and Small MoEs span approximately 41M–294M total parameters and 22M–124M active parameters per token, each trained on 2.67B tokens (estimated training FLOPs: $(1.2\text{–}7.3) \times 10^{18}$ at $L = 4$ and $(2.2\text{–}13.9) \times 10^{18}$ at $L = 8$). (A) FineWeb validation loss; (B) WikiText validation loss; (C) average downstream accuracy over five tasks. Endpoint arrows show Best minus standard looping (blue) and Mean (pink), in percentage points.

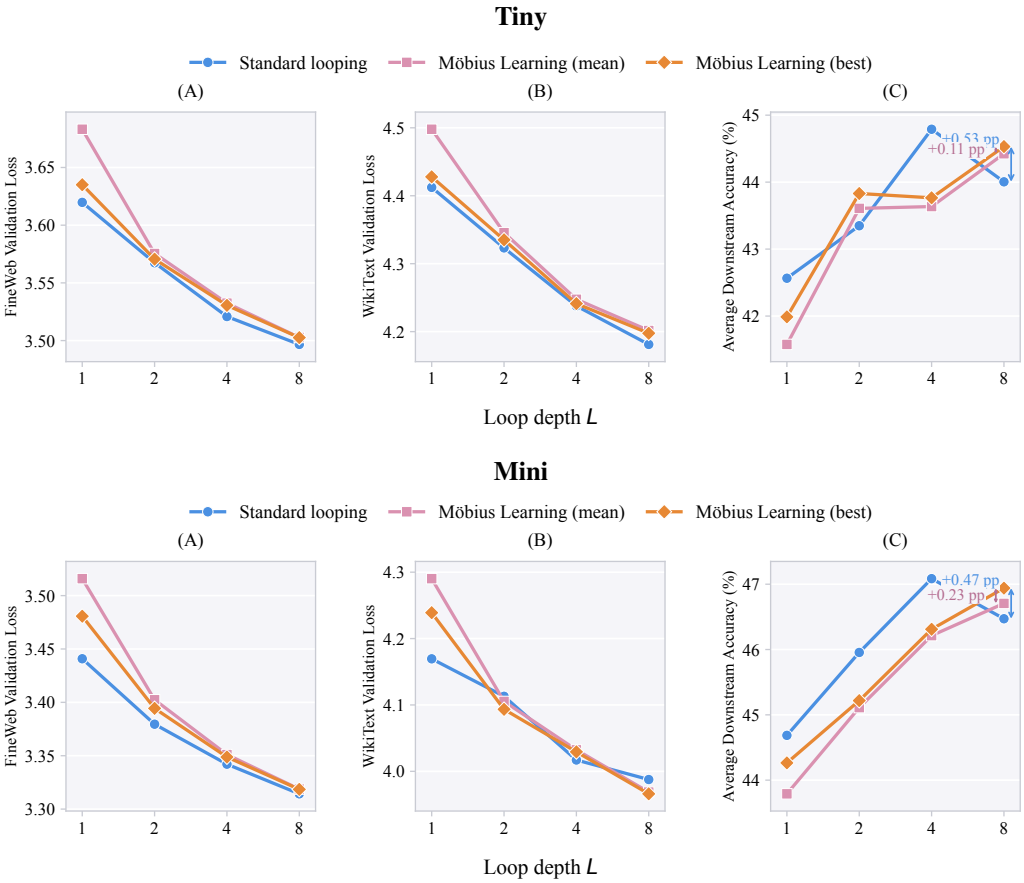


Figure 9: Parameter-reuse pretraining on FineWeb for Tiny and Mini. The two MoEs have approximately 41M and 83M total parameters (22M and 41M active per token), respectively, each trained on 2.67B tokens at $L = 1, 2, 4, 8$ (estimated training FLOPs: $(0.5\text{–}2.2) \times 10^{18}$ for Tiny and $(0.8\text{–}4.2) \times 10^{18}$ for Mini). (A) FineWeb validation loss; (B) WikiText validation loss; (C) average downstream accuracy over five tasks. Endpoint arrows show Best minus standard looping (blue) and Mean (pink), in percentage points.

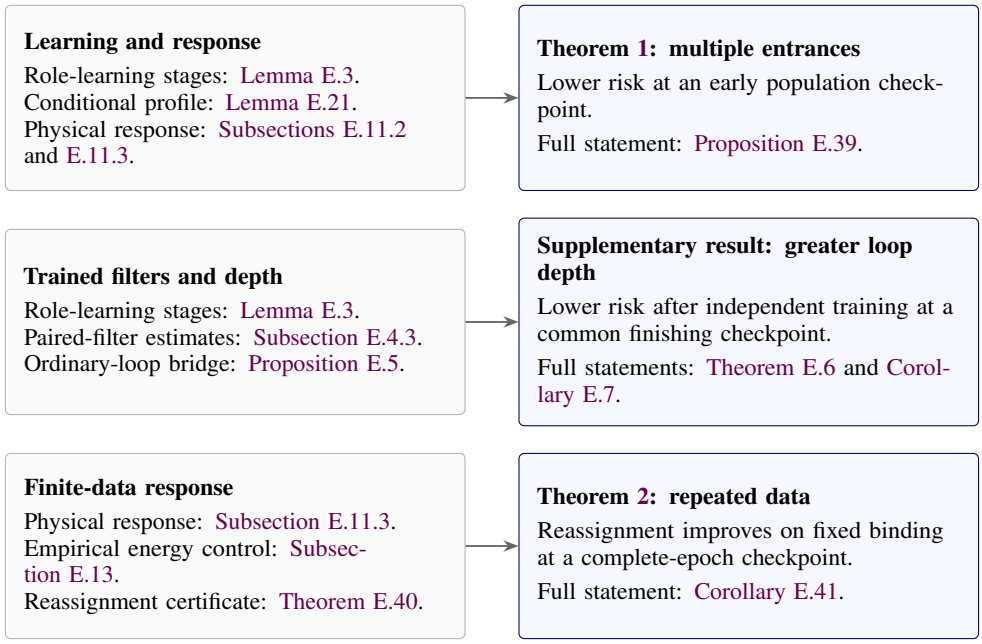
Table 3: Per-task downstream accuracy (%; higher is better) after 2.67B training tokens, for Tiny, Mini, and Small at $L = 1, 2, 4, 8$, and Medium at $L = 8$. Mean averages four cyclic-origin accuracies; Best selects one fixed origin by the five-task average, not a separate origin per task. SciQ includes support passages; CSQA denotes CommonsenseQA. Avg. weights the five tasks equally, excluding WinoGrande. Bold marks the highest displayed score within each scale and depth group. All entries are single-run results.

Depth L	Method	ARC-Easy	HellaSwag	PIQA	SciQ	CSQA	Avg.
Tiny							
1	Möbius Learning (mean)	36.38	27.19	60.70	61.42	22.17	41.58
	Möbius Learning (best)	**36.78**	**27.20**	60.94	62.50	22.52	41.99
	Standard looping	36.62	27.18	**61.10**	**63.20**	**24.73**	**42.56**
2	Möbius Learning (mean)	38.45	27.38	61.96	64.83	25.43	43.61
	Möbius Learning (best)	**38.72**	27.35	**61.97**	65.40	**25.72**	**43.83**
	Standard looping	37.37	**27.50**	61.92	**66.20**	23.75	43.35
4	Möbius Learning (mean)	38.80	**27.71**	61.03	65.75	24.88	43.64
	Möbius Learning (best)	38.97	27.59	61.10	66.10	25.06	43.77
	Standard looping	**39.39**	27.60	**62.13**	**68.20**	**26.62**	**44.79**
8	Möbius Learning (mean)	39.26	27.78	61.60	**68.05**	25.43	44.42
	Möbius Learning (best)	**39.60**	27.85	**61.64**	68.00	25.55	**44.53**
	Standard looping	38.17	**27.93**	61.10	66.20	**26.62**	44.00
Mini							
1	Möbius Learning (mean)	38.87	27.88	62.06	65.80	24.34	43.79
	Möbius Learning (best)	39.48	28.03	**62.84**	66.40	**24.57**	44.26
	Standard looping	**42.42**	**28.10**	61.92	**66.50**	24.49	**44.69**
2	Möbius Learning (mean)	39.98	28.26	63.38	67.68	26.25	45.11
	Möbius Learning (best)	40.15	28.29	63.44	68.00	26.21	45.22
	Standard looping	**41.20**	**28.60**	**63.87**	**69.40**	**26.70**	**45.96**
4	Möbius Learning (mean)	41.44	29.03	63.49	69.62	27.48	46.21
	Möbius Learning (best)	41.33	28.94	63.44	70.00	27.85	46.31
	Standard looping	**42.21**	**29.15**	**64.53**	**71.20**	**28.34**	**47.09**
8	Möbius Learning (mean)	**43.05**	29.49	63.41	70.70	26.88	46.71
	Möbius Learning (best)	42.97	**29.56**	**63.60**	**71.30**	27.27	**46.94**
	Standard looping	41.54	29.25	62.89	70.50	**28.17**	46.47
Small							
1	Möbius Learning (mean)	44.02	29.43	65.18	72.10	26.97	47.54
	Möbius Learning (best)	44.91	29.55	65.40	**73.30**	27.19	48.07
	Standard looping	**45.16**	**30.13**	**65.51**	72.10	**28.09**	**48.20**
2	Möbius Learning (mean)	44.63	**30.72**	66.49	72.30	26.82	48.19
	Möbius Learning (best)	45.37	30.67	**67.30**	72.20	27.27	48.56
	Standard looping	**45.71**	30.69	66.49	**73.80**	**29.24**	**49.18**
4	Möbius Learning (mean)	**45.81**	31.80	66.65	73.50	29.75	49.50
	Möbius Learning (best)	45.75	**32.00**	66.43	74.00	**29.81**	**49.60**
	Standard looping	45.29	31.53	**67.14**	**74.70**	28.75	49.48
8	Möbius Learning (mean)	44.96	**32.60**	67.90	77.15	31.57	50.84
	Möbius Learning (best)	44.95	32.50	**68.01**	**77.70**	**31.94**	**51.02**
	Standard looping	**44.99**	32.07	67.74	73.40	28.99	49.44
Medium							
8	Möbius Learning (mean)	**50.98**	35.91	69.78	79.40	**35.65**	54.34
	Möbius Learning (best)	50.88	**36.03**	**70.13**	79.40	35.46	**54.38**
	Standard looping	49.62	35.83	69.15	**79.80**	34.48	53.78

E THEORY: MODEL, RESULTS, AND PROOFS

PROOF ROADMAP

The two main-text theorems and the supplementary loop-depth result answer different training questions. All use the prediction task, physical network, and paired reset of Subsection E.1. The rows below connect their proof ingredients to their full statements.



Combining the early margins. The population margin in Proposition E.39, empirical control in Subsection E.13, and reassignment gaps in Corollary E.41 yield the four-arm risk differences in Equation (411). At full entrances, $P_e = B$, these imply

$$\mathcal{T}_O > \mathcal{T}_F > \max\{\mathcal{T}_C, \mathcal{T}_U\} \qquad \text{with probability tending to one.}$$

Here O, F, C, U denote ordinary looping, fixed binding, cyclic reassignment, and random reassignment. This combination uses the common early empirical window, not the later depth-finishing checkpoint.

Complementary results. The entrance-count selection in Corollary E.30 and late calibration in Subsection E.9 describe later training stages; they are not prerequisites for the three routes above.

E.1 PREDICTION TASK AND PHYSICAL MODEL

A partially observed next-token task. Each example is a binary token sequence $X_1, \dots, X_{T+1}$. Its latent states $(S_t)_{t\in\mathbb{Z}}$ form a stationary binary Markov chain; the learner sees only the X_t, not the S_t. Conditional on the states, token emissions are independent, with

$$\Pr(S_{t+1} = s' \mid S_t = s) = \frac{1+\rho ss'}{2}, \qquad \Pr(X_t = x \mid S_t = s) = \frac{1+\kappa xs}{2}, \tag{16}$$

for $s, s', x \in \{-1, 1\}$. The symmetric stationary law gives each state equal probability. The parameter ρ controls persistence of the hidden state, while $\kappa = \sqrt{p} > 0$ controls emission reliability: $\Pr(X_t = S_t) = (1+\kappa)/2$. In particular, $\mathbb{E}[X_t X_{t+j}] = p\rho^j$ for $j \geq 1$. The observed tokens can retain information about earlier observations through the unobserved state, so (X_t) need not itself be Markov. We fix finite B, T, take sufficiently small $\rho > 0$, and keep p in a compact interval $\mathcal{I} \Subset (0, 1)$. Lowercase $p = \kappa^2$ is unrelated to the source count P_{e} below.

At every scored position $t = 1, \dots, T$, the network receives only the prefix $X_1, \dots, X_t$ and predicts the next token X_{t+1}; the final target X_{T+1} is used for scoring, not supplied as input. A scalar logit z denotes the binary softmax logits $(z, -z)$ for labels $(+1, -1)$. Its cross-entropy is

$$\ell(z, y) = \log(2\cosh z) - yz. \tag{17}$$

The loss has unit curvature at $z = 0$, fixing the constants in the risk tables. As a reference class, let one coefficient vector $u = (u_0, \dots, u_{T-1})$ serve all prefix lengths:

$$F_{\rho,T}(u) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\ell\left(\sum_{j=0}^{t-1} u_j X_{t-j}, X_{t+1}\right), \qquad V_T = \min_{u\in\mathbb{R}^T} F_{\rho,T}(u). \tag{18}$$

Thus every position contributes to the expected loss, and position t uses only the available lags $0, \dots, t-1$. The reference V_T is the best shared *linear-history* cross-entropy, not the HMM's Bayes risk.

E.1.1 THE FINITE-WIDTH CAUSAL NETWORK

There are B distinct physical residual blocks. In every call, h_t and h_{t-1} are taken from the same preceding layer. At a given layer, block i maps a scalar hidden sequence $h = (h_t)_{t=1}^T$ to

$$[T_i(h)]_t = h_t + \sum_{a=1}^{6} v_{ia}\,\mathrm{ReLU}(w_{ia}h_t + \pi_{ia}h_{t-1} + b_{ia}) + d_i, \qquad h_0 = 0. \tag{19}$$

Each block therefore has $6\times 4 + 1 = 25$ trainable parameters. The input embedding has two trainable entries, $h_t^{(0)} = E_{X_t}$, and the shared readout is

$$H(h) = r_+\,\mathrm{ReLU}(a_+h + b_+) + r_-\,\mathrm{ReLU}(a_-h + b_-) + b_{\mathrm{out}}. \tag{20}$$

The head has seven parameters. The full model consequently has exactly $25B + 2 + 7 = 25B + 9$ physical parameters. Equation (19) is causal: the logit used to score X_{t+1} depends only on $X_1, \dots, X_t$, and the target token is never supplied to its own predictor. This finite-width model has no attention layer.

Choose a source count $P_{\mathrm{e}} \mid B$ and put $q = B/P_{\mathrm{e}}$. Sources are indexed by $s = 0, \dots, P_{\mathrm{e}} - 1$. Source s applies the physical blocks in the order

$$i_{s,k} = (sq + k) \bmod B, \qquad k = 0, \dots, BL - 1. \tag{21}$$

Thus every source executes L complete backbone passes and makes $D_{\mathrm{blk}} = BL$ physical block calls. Equivalently, it traverses $P_{\mathrm{e}}L$ groups of q blocks. Here D_{blk} counts physical calls; the methodology's $K = LP$ counts applications of its P worker-owned block groups and need not equal $P_{\mathrm{e}}L$. Hidden states are separate across source paths, while the blocks, embedding, and head are shared.

Counting blocks, groups, and entrances. For $B = 12$ blocks divided into four equal worker-owned groups, $P = 4$ and the worker-indexed construction has $P_{\mathrm{e}} = 4$. Standard looping still has $P_{\mathrm{e}} = 1$ on all twelve blocks. The theory's $P_{\mathrm{e}} = 12$ layout starts at every block and is not available as a

starting-point choice in that fixed four-group construction. In the four-group method layout, $K = 4L$ counts group applications; $D_{\mathrm{blk}} = 12L$ counts physical block calls.

Let $z_t^{(s)}(\Theta)$ be source s's logit at position t. Training uses the mean over positions and sources,

$$R_{B,P_{\mathrm{e}},L}(\Theta) = \frac{1}{P_{\mathrm{e}}T}\sum_{s=0}^{P_{\mathrm{e}}-1}\sum_{t=1}^{T}\mathbb{E}\ell(z_t^{(s)}(\Theta), X_{t+1}), \qquad \Theta_{n+1} = \Theta_n - \eta\nabla R_{B,P_{\mathrm{e}},L}(\Theta_n). \tag{22}$$

Every occurrence of a reused block and both shared endpoints is differentiated before this single physical Euclidean update. Losses, rather than logits or predictions, are averaged across sources. If an implementation instead sums source losses, the equivalent scalar rate is $\eta_{\mathrm{mean}} = P_{\mathrm{e}}\eta_{\mathrm{sum}}$ for scalar GD. This conversion does not automatically extend to nonlinear clipping or adaptive updates.

E.1.2 Reset, invariance, and the physical metric

Assumption E.1 (Paired reset)**.** For every block, the first three tuples (v, w, π, b) are

$$(1, 1, 0, \tfrac{1}{2}), \qquad (-1, 1, 0, -\tfrac{1}{2}), \qquad (-1, 1, 0, 0). \tag{23}$$

The remaining three tuples are, respectively, $(-v, -w, -\pi, b)$, and $d_i = 0$. The embedding is $(E_{+1}, E_{-1}) = (1, -1)$. The two head triples are

$$(r_+, a_+, b_+) = (\omega_0, 1, 0), \qquad (r_-, a_-, b_-) = (-\omega_0, -1, 0), \tag{24}$$

with $b_{\mathrm{out}} = 0$ and $0 \le \omega_0 \le c_h\rho$ for a fixed c_h. All of these parameters remain trainable.

The reset contains no learned temporal rule: every temporal weight is zero, and the readout may also start at zero. It is a deliberately structured initialization, not a claim about a typical random initialization.

Global token-sign reversal, paired-neuron exchange, and the corresponding endpoint sign changes are orthogonal symmetries of the population risk. Cyclically rotating block indices by q permutes the source losses and is another orthogonal symmetry. Since the reset is fixed by these symmetries, full GD preserves neuron pairing, odd embedding and readout, zero additive output biases, and q-periodicity of the blocks. Here "zero additive output biases" means the block residual biases d_i and the readout bias b_{out}. The neuron biases b_{ia} and the paired-head bias b remain trainable; in particular, some b_{ia} are initialized at $\pm 1/2$. This is an invariant trajectory of the original optimizer, not a projected or tied-parameter training rule.

To display its physical scaling, retain the first three neurons of each of the q block roles and write the independent endpoints as (e, r, a, b), where

$$E_{\pm 1} = \pm e, \qquad H(h) = r\,\mathrm{ReLU}(ah + b) - r\,\mathrm{ReLU}(-ah + b). \tag{25}$$

If ξ collects these $12q + 4$ raw coordinates and $J_{P_{\mathrm{e}}}\xi$ denotes their linear embedding into the $25B + 9$ physical parameters, then

$$(J_{P_{\mathrm{e}}}^{\mathsf{T}}J_{P_{\mathrm{e}}})^{-1} = \mathrm{diag}\big((2P_{\mathrm{e}})^{-1}I_{12q}, \tfrac{1}{2}I_4\big), \qquad \xi^+ = \xi - \eta(J_{P_{\mathrm{e}}}^{\mathsf{T}}J_{P_{\mathrm{e}}})^{-1}\nabla_\xi R_{B,P_{\mathrm{e}},L}. \tag{26}$$

In particular, $\|\nabla_\Theta R\|^2 = \nabla_\xi R^{\mathsf{T}}(J_{P_{\mathrm{e}}}^{\mathsf{T}}J_{P_{\mathrm{e}}})^{-1}\nabla_\xi R$. The factor $1/P_{\mathrm{e}}$ in a role's mobility is the reason the stage lengths below carry an explicit factor P_{e}. For arguments that compare two parameter trajectories, one may instead choose an orthonormal basis of this fixed linear subspace. In those physical isometric coordinates the update is ordinary Euclidean GD; no nonlinear coordinate connection term is then omitted.

E.1.3 Exact linear-history filter in the gate region

At reset a block maps fixed intervals around $+1$ and -1 to their centers, all six preactivation signs are separated from zero, and the state derivative vanishes. Hence there is a fixed open neighborhood in which those intervals are invariant and every block is contractive. On the paired invariant subspace, a role-j block is exactly

$$h_t^+ = \alpha_j h_t + \beta_j h_{t-1} + \gamma_j X_t, \qquad \alpha_j = 1 + \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{w}_j, \quad \beta_j = \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{\pi}_j, \quad \gamma_j = \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{b}_j. \tag{27}$$

The token term is generated by the fixed ReLU gates; it is not an additional input to the architecture.

Number roles backward from the output and work in the exact truncated polynomial ring $\mathbb{R}[z]/(z^T)$. Define

$$D_j(z) = \alpha_j + \beta_j z, \quad A(z) = \prod_{j=1}^{q} D_j(z), \quad C(z) = \sum_{j=1}^{q} \gamma_j \prod_{r<j} D_r(z). \tag{28}$$

One q-block group maps a history filter Q to $AQ + C$. Since a source executes $P_{\mathrm{e}}L$ such groups, its exact terminal filter is

$$Q_L(z) = eA(z)^{P_{\mathrm{e}}L} + C(z) \sum_{m=0}^{P_{\mathrm{e}}L-1} A(z)^m. \tag{29}$$

The first term is the propagated input embedding and must not be dropped. If $Q_{L,j} = [z^j]Q_L(z)$, then every source has the same logit on the invariant trajectory,

$$z_t = \sum_{j=0}^{t-1} u_j X_{t-j}, \qquad u_0 = r(aQ_{L,0} + b), \qquad u_j = raQ_{L,j} \quad (j \geq 1). \tag{30}$$

Thus $R_{B,P_{\mathrm{e}},L} = F_{\rho,T}(u)$ there. The forward function is a linear-history predictor in this certified gate region, although its factorized physical training dynamics remain nonlinear and all $25B + 9$ parameters are updated.

Shrinking the neighborhood if necessary gives $|\alpha_j| + 2|\beta_j| < 1/4$. Product differentiation on $|z| \leq 2$ then bounds every fixed-order derivative of (29) uniformly in finite L: differentiating k factors loses at most k contraction factors, while the number of assignments grows only polynomially in the number of groups. This depth-uniform gate and derivative control is used in both proof appendices.

Remark E.2 (Scope of the physical setup). The population results concern GD from the exact reset, fixed finite B and T, and weak ρ. The finite-data results use the same network and reset but allow empirical trajectories to leave the population-symmetric subspace. Neither setting asserts that arbitrary initializations stay in the gate region or that the network learns a nonlinear HMM posterior. Constants are not uniform when B or T grows. The separately processed population-path proxy $W = P_{\mathrm{e}}BLN$ is neither measured wall-clock time nor an optimizer-independent compute model.

E.1.4 Matched training and risk comparisons

For depth comparisons, fix $B, T, P_{\mathrm{e}}, p, \rho, \eta$, the reset, and a common number of updates; each (P_{e}, L) follows its own uninterrupted trajectory. Partition comparisons use one common update count rather than each partition's role-finishing checkpoint. For population comparisons define

$$\mathcal{R}_{P_{\mathrm{e}},L}(N) := R_{B,P_{\mathrm{e}},L}(\Theta_{P_{\mathrm{e}},L,N}), \qquad E_{P_{\mathrm{e}},L,N} := \mathcal{R}_{P_{\mathrm{e}},L}(N) - V_T, \qquad \ell_\rho := \log(1/\rho). \tag{31}$$

This excess is relative to the shared linear-history comparator in (18). The depth result uses a separate finishing checkpoint $N_{P_{\mathrm{e}}}$ for each entrance count; a common third-stage window, a later calibration window, and the empirical first-stage window are distinct.

For finite-data comparisons, draw $M = P_{\mathrm{e}}J_{\mathrm{ep}}m$ independent complete sequences of length $T + 1$. A data-independent partition forms cells $\mathcal{D}_{s,b}$ of m whole sequences, where $0 \leq s < P_{\mathrm{e}}$ and $0 \leq b < J_{\mathrm{ep}}$. At update $n = eJ_{\mathrm{ep}} + b$, fixed binding sends $\mathcal{D}_{s,b}$ to entrance s, while cyclic reassignment sends $\mathcal{D}_{s+e \bmod P_{\mathrm{e}},b}$ to entrance s. Write these assigned cells as $\mathcal{D}^{\mathrm{fix}}_{s,e,b} = \mathcal{D}_{s,b}$ and $\mathcal{D}^{\mathrm{cyc}}_{s,e,b} = \mathcal{D}_{s+e \bmod P_{\mathrm{e}},b}$. Every epoch visits each sequence once; its T scored targets are not independent samples. For either schedule a, a physical mean-GD update uses

$$\Theta^a_{n+1} = \Theta^a_n - \eta \nabla \widehat{R}^a_{e,b}(\Theta^a_n), \qquad \widehat{R}^a_{e,b}(\Theta) = \frac{1}{P_{\mathrm{e}}mT} \sum_{s=0}^{P_{\mathrm{e}}-1} \sum_{X \in \mathcal{D}^a_{s,e,b}} \sum_{t=1}^{T} \ell(z_t^{(s)}(\Theta; X), X_{t+1}). \tag{32}$$

The two schedules use the same sample union at every update. The ordinary comparator instead sends that entire union through the sole entrance, while retaining all B physical blocks and the same η and update count. For a complete-epoch checkpoint $N = E_{\mathrm{ep}}J_{\mathrm{ep}}$, all empirical arms process $\mathcal{W} = (M/J_{\mathrm{ep}})TBLN = MTBLE_{\mathrm{ep}}$ block-token calls. This accounting has no extra entrance-count factor. Their test risk is fresh-sequence source-average cross-entropy, not an average of source predictions; empirical paths need not preserve the population pairing or linear-history form.

E.2 Learning depth roles under full physical gradient descent

This subsection proves the training estimates used by Theorem E.6. A "stage" is only a time interval in the analysis: the optimizer always applies (22), without freezing, resetting, or replacing any physical parameter. Throughout we use the single-mode task (16).

Put $\ell_\rho = \log(1/\rho)$ and

$$c_j = p(1-p)^j, \qquad w_j = \frac{T-j}{T}, \qquad 0 \le j < T, \tag{33}$$

and, for $k \ge 1$, put

$$d_{k+1} = \frac{c_k c_1}{c_0} = c_{k+1}. \tag{34}$$

The coefficients c_j are the leading optimal linear-history coefficients, not initialized weights. The last equality in (34) is the geometric matching special to the single-mode task.

Lemma E.3 (Stage learning in the physical network)**.** *Fix $B \ge 2$, $T \ge 4$, $\mathcal{I} \Subset (0,1)$, and c_h. There are $\rho_0, \eta_0 > 0$, constants $A_1, \dots, A_{\min(B,T-2)}$, and finite $C, m_\star$ such that the following holds. For $P_{\mathrm{e}} \mid B$, set $q = B/P_{\mathrm{e}}$ and*

$$n_{P_{\mathrm{e}},0} = 0, \qquad n_{P_{\mathrm{e}},k} = n_{P_{\mathrm{e}},k-1} + \left\lceil \frac{A_k P_{\mathrm{e}} \ell_\rho}{\eta \rho^{2k}} \right\rceil, \qquad 1 \le k \le \min(q, T-2). \tag{35}$$

For every finite $L \ge 1$, $p \in \mathcal{I}$, $0 < \rho \le \rho_0$, and $0 < \eta \le \eta_0$, full physical Euclidean GD from Assumption E.1 preserves the paired, q-periodic fixed-gate subspace through $n_{P_{\mathrm{e}},k}$. That linear subspace admits orthonormal physical coordinates in which the update remains ordinary Euclidean GD.

Stage one starts from reset with excess $O(\rho^2)$. Each later stage $k \ge 2$ starts with excess at most $C\rho^{2k+2}$, and stage k lasts $O(P_{\mathrm{e}}\rho^{-2k}\ell_\rho/\eta)$ updates. Throughout the stage, the first $k-1$ roles stay in fixed positive multiples of ρ, the current role is $O(\rho)$, and every complementary raw-coordinate deviation from reset is $O(\rho^2 \ell_\rho^{m_\star})$. At the checkpoint the first k temporal roles obey

$$t_j = (1-p)\rho + O(\rho^2 \ell_\rho^{m_\star}), \qquad 1 \le j \le k, \tag{36}$$

If the initial head is zero, stage one includes a fixed physical-time startup, after which its active Jacobian is used. Including that startup, its total gradient path satisfies

$$\sum_{n=0}^{n_{P_{\mathrm{e}},1}-1} \eta \|\nabla R_{B,P_{\mathrm{e}},L}(\Theta_n)\| \le C\rho. \tag{37}$$

If $u_{\mathrm{lo}} = (u_0, \dots, u_k)$ and

$$\mathcal{E}_k(u) = F_{\rho,T}(u) - \min_{v \in \mathbb{R}^{k+1}} F_{\rho,T}(v, u_{k+1:T-1}), \tag{38}$$

then

$$\mathcal{E}_k(\Theta_{n_{P_{\mathrm{e}},k}}) \le C\rho^{2k+6}\ell_\rho^{m_\star}. \tag{39}$$

Let $D_{\mathrm{blk}} = BL$ and $I_{k,P_{\mathrm{e}},L} = \mathbf{1}_{\{q=k,\, D_{\mathrm{blk}} \ge k+1\}}$. The actual source-mean risk satisfies

$$R_{B,P_{\mathrm{e}},L}(\Theta_{n_{P_{\mathrm{e}},k}}) - V_T = \frac{w_{k+1}}{2}\left(c_{k+1} - I_{k,P_{\mathrm{e}},L} d_{k+1}\right)^2 \rho^{2k+4} + O(\rho^{2k+5}\ell_\rho^{m_\star}). \tag{40}$$

When $I_{k,P_{\mathrm{e}},L} = 1$, the sharper one-mode bound is

$$0 \le R_{B,P_{\mathrm{e}},L}(\Theta_{n_{P_{\mathrm{e}},k}}) - V_T \le C\rho^{2k+6}\ell_\rho^{m_\star}. \tag{41}$$

More specifically, after the final role stage $k = q$, whenever $D_{\mathrm{blk}} = BL \ge q+1$, every represented coefficient matches the one-mode target to one extra power:

$$u_i = c_i \rho^{i+1} + O(\rho^{i+2}\ell_\rho^{m_\star}), \qquad 0 \le i \le \min(D_{\mathrm{blk}}, T-1), \tag{42}$$

and (41) holds with $k = q$. For ordinary training ($P_{\mathrm{e}} = 1, q = B$), the same absolute-risk statement at $n_{1,B}$ gives, for every finite $L' \ge 2$ and $T \ge B+2$,

$$R_{B,1,1}(\Theta_{1,n_{1,B}}) - R_{B,1,L'}(\Theta_{L',n_{1,B}}) = \frac{T-B-1}{2T} c_{B+1}^2 \rho^{2B+4} + O(\rho^{2B+5}\ell_\rho^{m_\star}) > 0. \tag{43}$$

The constants are uniform over the finite divisor set $P_{\mathrm{e}} \mid B$ and finite L, with B, T fixed; no growing-B or growing-T uniformity is claimed.

The indicator in (40) has a simple meaning. If all $q = k$ roles have learned and one further physical occurrence is available, the repeated first role supplies the leading $(k+1)$st coefficient. In the single-mode model that repeated coefficient equals the target c_{k+1}. For $q > k$, the next independent role has not yet learned; for $D_{\mathrm{blk}} = k$, the next lag is not representable at all.

E.2.1 Weighted cross-entropy geometry

For the HMM in (16),

$$\mathbb{E}[X_t X_{t+d}] = p\rho^d, \qquad d \geq 1, \tag{44}$$

and every odd token moment vanishes. More generally, expanding emissions and pairing equal hidden-state indices bounds an even token moment by $C_T \rho^{d(I)}$, where $d(I)$ is the minimum total distance of a pairing of the time multiset I; repeated indices may pair at zero distance.

Write $F_i = \partial F_{\rho,T}/\partial u_i$. In the weighted region $|u_j| \leq C\rho^{j+1}$, expanding $\tanh z = z - z^3/3 + O(z^5)$ and using the pairing bound gives

$$F_i = w_i \left[u_i + \sum_{j<i} p\rho^{i-j} u_j - p\rho^{i+1} \right] + O(\rho^{i+3}). \tag{45}$$

Indeed, a degree-$r \geq 3$ term with history indices $j_1, \ldots, j_r$ contains $\rho^{j_1+\cdots+j_r+r}$ from its coefficients. The moment paired with X_{t-i} contributes enough distance that $j_1 + \cdots + j_r + d(\{i, j_1, \ldots, j_r\}) \geq i$, so its order is at least ρ^{i+r}. This proves the displayed remainder rather than assuming that nonlinear terms are uniformly negligible.

The coefficient Hessian satisfies

$$F_{ii} = w_i + O(\rho^2), \tag{46}$$

$$F_{ij} = w_{\max(i,j)} p\rho^{|i-j|} + O(\rho^{|i-j|+2}), \qquad i \neq j. \tag{47}$$

At $\rho = 0, u = 0$, this Hessian is $\mathrm{diag}(w_0, \ldots, w_{T-1})$, consistently with $\ell''(0) = 1$. Since T is fixed, it remains uniformly positive definite in a fixed small-logit neighborhood. Solving (45) triangularly gives the unique minimizer

$$u_j^* = c_j \rho^{j+1} + O(\rho^{j+3}), \qquad c_j = p(1-p)^j. \tag{48}$$

For fixed high coefficients $u_{k+1:T-1}$, the low conditional minimizer is also unique. The implicit-function formula and a Neumann expansion of the low Hessian yield

$$\left| \frac{\partial u_i^{\mathrm{cond}}}{\partial u_j} \right| \leq C\rho^{j-i}, \qquad i \leq k < j. \tag{49}$$

Every product path in an off-diagonal entry of the inverse Hessian has total index distance at least $|i-j|$, which is why profiling higher histories does not destroy the row scaling used below.

E.2.2 The active chart and row-scaled cancellation

Let $\boldsymbol{v}^0 = (1, -1, -1)$ and decompose the temporal vector of role j as

$$\boldsymbol{\pi}_j = t_j \boldsymbol{v}^0/3 + \boldsymbol{\pi}_{j,\perp}, \qquad t_j = \boldsymbol{v}^{0\mathsf{T}} \boldsymbol{\pi}_j. \tag{50}$$

To make the zero-head initialization regular, let $\Phi = Q_{L,0}$, $\Phi_j = \partial_{\alpha_j} \Phi$, and set

$$\begin{aligned} C_0 &= a\Phi + b, & s_1 &= \boldsymbol{v}_1^{\mathsf{T}} \boldsymbol{v}^0/3, \\ A_0 &= a\Phi_1 s_1/C_0, & B_0 &= \frac{a}{C_0} \left(\Phi_1 \boldsymbol{v}_1^{\mathsf{T}} \boldsymbol{\pi}_{1,\perp} + \sum_{j=2}^{q} \Phi_j \beta_j \right), \\ g &= rC_0, & y &= A_0 t_1 + B_0. \end{aligned} \tag{51}$$

At reset, $C_0 = A_0 = 1$ and $B_0 = DB_0 = 0$. The inverse $r = g/C_0$, $t_1 = (y - B_0)/A_0$ is smooth even when $r = g = 0$, and $u_0 = g$, $u_1 = gy$ exactly. Once the complementary coordinates are $O(\rho^2 \ell_\rho^{m_\star})$, this inverse also gives $t_1 = y + O(\rho^2 \ell_\rho^{m_\star})$.

Degree counting in the complete filter (29), including its propagated-input term, gives for $2 \le j \le q$

$$u_j = g\left[y\prod_{a=2}^{j} t_a + E_{j+1}\right], \qquad \|D^b E_{j+1}\| \le C\varpi^{j+1-b}, \quad b = 0,1,2, \tag{52}$$

where ϖ is the norm of all deviations from reset in this chart. A lag j requires at least j temporal factors. Its unique degree-j path uses the last j temporal occurrences; changing an injection, endpoint, or current-state factor adds a degree. If $j > D_{\mathrm{blk}}$, the coefficient and all of its parameter derivatives vanish identically.

In stage k, define active coordinates $f = (g, y, t_2, \dots, t_k)$ and let ν contain all remaining raw linear coordinates. Suppose the first $k-1$ roles lie in fixed positive multiples of ρ, while $|t_k| \le C\rho$. For $u_{\mathrm{lo}} = (u_0, \dots, u_k)$, (52) yields the row-scaled factorizations

$$J_f := D_f u_{\mathrm{lo}} = \mathrm{diag}(1, \rho, \dots, \rho^k)[\mathsf{L}_k + O(\rho\ell_\rho^{m_\star})], \tag{53}$$

where L_k is lower triangular with diagonal bounded away from zero, and

$$J_\nu := D_\nu u_{\mathrm{lo}} = \mathrm{diag}(1, \rho, \dots, \rho^k)\mathsf{E}_k, \qquad \|\mathsf{E}_k\| \le C\rho\ell_\rho^{m_\star}. \tag{54}$$

The final diagonal entry of L_k contains the already learned roles but not t_k itself. Thus the current role need not be learned before its Jacobian becomes usable. Although $\|J_f^{-1}\| \le C\rho^{-k}$ is large, the common row scaling cancels:

$$\mathsf{L} := J_\nu^{\mathsf{T}} J_f^{-\mathsf{T}} = O(\rho\ell_\rho^{m_\star}). \tag{55}$$

Using only the unweighted inverse norm would lose this conclusion.

Let G be the mobility induced by (26) in the regular chart. On the fixed neighborhood,

$$\lambda_{\min}(G) \ge \frac{1}{8P_{\mathrm{e}}}, \qquad G_{\nu g}, G_{\nu y} = O(\varpi), \qquad G_{\nu, t_j} = 0 \ (j \ge 2), \qquad G_{\nu\nu} = D_\nu, \tag{56}$$

where D_ν is a fixed positive diagonal block. The g, y cross terms are retained. The actual update has the form

$$\Delta f = -\eta(G_{ff}R_f + G_{f\nu}R_\nu) + T_f, \qquad \|T_f\| \le C\eta^2\|\nabla_\Theta R\|^2, \tag{57}$$

while the raw ν update has no chart Taylor remainder.

Write $u_{\mathrm{hi}} = u_{k+1:T-1}$ and

$$h_f = (D_f u_{\mathrm{hi}})^{\mathsf{T}} F_{\mathrm{hi}}, \qquad h_\nu = (D_\nu u_{\mathrm{hi}})^{\mathsf{T}} F_{\mathrm{hi}}, \qquad \zeta = h_\nu - \mathsf{L}h_f. \tag{58}$$

The chain rule gives the exact identity

$$R_\nu = \mathsf{L}R_f + \zeta. \tag{59}$$

Define

$$\mathsf{A} = (G_{\nu f} + D_\nu\mathsf{L})(G_{ff} + G_{f\nu}\mathsf{L})^{-1}. \tag{60}$$

Then $\mathsf{A} = O(\rho\ell_\rho^{m_\star})$, and eliminating R_f from the true coordinate updates gives

$$\Delta\nu = \mathsf{A}(\Delta f - T_f) + \eta(\mathsf{A}G_{f\nu} - D_\nu)\zeta. \tag{61}$$

This is an algebraic cancellation inside full physical GD, not a projection of the low-coefficient gradient.

E.2.3 Closing the stage induction

The first stage: startup, moving root, and physical descent. We establish the first stage directly from the physical update. In this subsection, ν denotes the remaining raw linear coordinates, centered at reset, after removing r and t_1. For the two coordinates of $\boldsymbol{\pi}_{1,\perp}$, choose an orthonormal basis of $(\boldsymbol{v}^0)^\perp$. Thus the raw coordinates (r, t_1, ν) have the constant mobility

$$\mathsf{D}_{\mathrm{raw}} = \mathrm{diag}\left(\tfrac{1}{2}, d_t, \mathsf{D}_\nu\right), \qquad d_t = \frac{3}{2P_{\mathrm{e}}}, \tag{62}$$

where D_ν is a fixed positive diagonal matrix. The factor 3 comes from $t_1 = (\boldsymbol{v}^0)^{\mathsf{T}}\boldsymbol{\pi}_1$ and $\|\boldsymbol{v}^0\|^2 = 3$. All remaining block coordinates and all remaining endpoint coordinates are included in ν.

Lemma E.4 (Uniform first-stage estimates for physical GD). *Fix $B \geq 2$, $T \geq 4$, $\mathcal{I} \Subset (0,1)$, and c_h. For a sufficiently large fixed A_1, there are $\rho_0, \eta_0, c_f, c_s, C > 0$ such that the following holds uniformly over $P_{\mathrm{e}} \mid B$, all positive integers L, $p \in \mathcal{I}$, $0 < \rho \leq \rho_0$, $0 < \eta \leq \eta_0$, and $0 \leq \omega_0 \leq c_h \rho$.*

Let

$$n_1 = n_{P_{\mathrm{e}},1} = \left\lceil \frac{A_1 P_{\mathrm{e}} \ell_\rho}{\eta \rho^2} \right\rceil, \qquad \ell_\rho = \log(1/\rho).$$

The uninterrupted physical GD trajectory remains in the paired, q-periodic fixed-gate region through n_1. Write its unchanged risk in the regular chart as $\mathcal{J}(g, y, \nu)$, let $\widehat{g}(y, \nu)$ be the local root of $\mathcal{J}_g = 0$, and put

$$e_n = g_n - \widehat{g}(y_n, \nu_n), \qquad \delta_n = y_n - (1-p)\rho.$$

For every $0 \leq n \leq n_1$,

$$|e_n| \leq C\rho e^{-c_f \eta n} + C\rho^5, \tag{63}$$

$$|\delta_n| \leq C\rho e^{-c_s \eta \rho^2 n / P_{\mathrm{e}}} + C\rho^3, \tag{64}$$

$$\|\nu_n\| \leq C\rho^2. \tag{65}$$

Moreover,

$$\sum_{n < n_1} \eta \|\nabla_\Theta R_{B,P_{\mathrm{e}},L}(\Theta_n)\| \leq C\rho. \tag{66}$$

At n_1,

$$e_{n_1} = O(\rho^5), \qquad y_{n_1} = (1-p)\rho + O(\rho^3), \qquad \|\nu_{n_1}\| = O(\rho^2). \tag{67}$$

The constants may depend on the fixed quantities above and A_1, but not on ρ, η, L. No growing-B or growing-T assertion is made.

Proof. **1. Exact chart and depth-uniform coefficient bounds.** The population symmetries preserve the stated linear physical subspace whenever the gradients are evaluated in the strict gate region. We first work on a stopped trajectory in that region and subsequently rule out exit.

Let $\Phi(\alpha, \gamma, e) = Q_{L,0}$, retaining the propagated-input term in (29). Since the entire filter is obtained by replacing α_j by $\alpha_j + \beta_j z$ in Φ, its coefficients satisfy the exact identity

$$Q_{L,j} = \frac{1}{j!} \sum_{i_1, \ldots, i_j = 1}^{q} \beta_{i_1} \cdots \beta_{i_j} \partial_{\alpha_{i_1}} \cdots \partial_{\alpha_{i_j}} \Phi. \tag{68}$$

For $j = 0$ the right side means Φ. In particular, $Q_{L,1} = \sum_j \Phi_j \beta_j$, exactly.

Every fixed-order derivative of Φ is bounded uniformly in finite L on a sufficiently small fixed raw-coordinate neighborhood. Indeed, a derivative of order k of a product involving qm role factors is bounded by a constant times $(1+qm)^k \theta^{(qm-k)_+}$, for a fixed $\theta < 1$. The geometric sum of these bounds is finite. The same bound applies to the separate propagated-input product. Only finitely many derivative orders, depending on fixed T, are needed below.

The definitions in (51) consequently give

$$C_0 = 1 + O(\|\nu\|), \quad A_0 = 1 + O(\|\nu\|), \quad B_0 = O(\|\nu\|^2), \quad D_\nu B_0 = O(\|\nu\|). \tag{69}$$

Here C_0, A_0 are independent of all temporal coordinates. The vanishing of B_0 and DB_0 at reset uses $\Phi_1 = 1$, $\Phi_j = 0$ for $j \geq 2$, and $(\boldsymbol{v}^0)^{\mathsf{T}} \boldsymbol{\pi}_{1,\perp} = 0$. These identities also hold at the boundary $BL = 2$.

Choose fixed amplitude bounds G_*, Y_*, specified below, and consider the scaled box

$$|g| \leq G_* \rho, \qquad |y| \leq Y_* \rho, \qquad \|\nu\| \leq \rho. \tag{70}$$

All estimates are established on a slightly larger box as well, so they apply to trial update segments. For small ρ, C_0, A_0 stay separated from zero, $r = g/C_0 = O(\rho)$, and $t_1 = (y - B_0)/A_0 = O(\rho)$. All temporal factors are $O(\rho)$.

Write

$$u = g\, s(y, \nu), \qquad s = (1, y, \mathcal{V}_2, \ldots, \mathcal{V}_{T-1}). \tag{71}$$

The chart relation for y eliminates the first temporal coefficient exactly:

$$\beta_1 = \frac{C_0}{a\Phi_1} y - \sum_{i=2}^{q} \frac{\Phi_i}{\Phi_1} \beta_i.$$

On the coarse box, $\Phi_1 = 1 + O(\|\nu\|)$ and $\Phi_i = O(\|\nu\|)$ for $i \geq 2$. Consequently every β_i is $O(\rho)$, with first and second (y, ν) derivatives bounded uniformly; in particular, $D_\nu \beta_1 = O(\rho)$. These bounds include the first role's orthogonal temporal coordinates, all other temporal roles, and the block and endpoint coordinates. The coefficient $Q_{L,j}$ in (68) is homogeneous of degree j in the temporal coefficients. Its remaining coefficient derivatives are uniformly bounded. The chain rule therefore gives, for $j \geq 2$ and $b = 0, 1, 2$,

$$\|D^b_{(y,\nu)} \mathcal{V}_j\| \leq C\rho^{j-b}. \tag{72}$$

A temporal derivative removes at most one temporal factor. Derivatives of the remaining coefficient functions are uniformly bounded by the preceding geometric majorant. For $j > BL$, $\mathcal{V}_j$ and all its derivatives are identically zero; no bound on the corresponding target residual is inferred from that fact.

2. Risk derivatives and the moving fast root. Let $\langle \cdot \rangle$ denote expectation averaged over the T scored prefixes, and put

$$f_t = \sum_{j=0}^{t-1} s_j X_{t-j}.$$

The exact HMM correlations and the coefficient bounds give

$$\begin{aligned} \langle X_{t+1} f_t \rangle &= p\rho + w_1 p\rho^2 y + O(\rho^5), \\ \langle f_t^2 \rangle &= 1 + w_1(2p\rho y + y^2) + O(\rho^4), \\ \langle f_t^4 \rangle &= 1 + O(\rho^2), \qquad w_1 = \frac{T-1}{T}. \end{aligned} \tag{73}$$

For example, a term containing $\mathcal{V}_j X_{t-j}$ once in the second moment is $O(\rho^{2j})$; the first such term has $j = 2$. The same HMM moment bound used in (45) controls the remaining terms.

Expanding the actual hyperbolic tangent, uniformly for $|g| \leq G_* \rho$, therefore gives

$$\mathcal{J}_g = g - p\rho + w_1(2p\rho y + y^2)g - w_1 p\rho^2 y - \frac{g^3}{3} + O(\rho^5), \qquad \mathcal{J}_{gg} = 1 + O(\rho^2). \tag{74}$$

The local root exists uniquely and obeys

$$\widehat{g} = p\rho + \frac{p^3}{3}\rho^3 - w_1 p\rho y^2 + w_1(p - 2p^2)\rho^2 y + O(\rho^5). \tag{75}$$

The coefficient chain rule controls the mixed derivatives throughout the coarse box. Here $F_j = O(\rho^{j+1})$, while (46)–(47) give

$$|(F'' s)_j| \leq C \sum_{i=0}^{T-1} \rho^{|i-j|+i} \leq C\rho^j.$$

Since s_0, s_1 are independent of ν,

$$\begin{aligned} \mathcal{J}_\nu &= g \sum_{j \geq 2} (D_\nu \mathcal{V}_j) F_j = O(\rho^5), \\ \mathcal{J}_{g\nu} &= \sum_{j \geq 2} (D_\nu \mathcal{V}_j)\{F_j + g(F'' s)_j\} = O(\rho^4). \end{aligned} \tag{76}$$

Similarly $\mathcal{J}_{gy} = O(\rho^2)$. Implicit differentiation of $\mathcal{J}_g(\widehat{g}, y, \nu) = 0$ thus yields

$$|\widehat{g}_y| \leq C\rho^2, \qquad \|D_\nu \widehat{g}\| \leq C\rho^4, \qquad \|D^2_{(y,\nu)} \widehat{g}\| \leq C\rho. \tag{77}$$

For the last bound, the root is zero at $\rho = 0$ for every nearby (y, ν): the next token is then independent and symmetric. The finite-alphabet risk, its required mixed derivatives, and the inverse $\mathcal{J}_{gg}^{-1}$ are

uniformly bounded. Differentiating the implicit equation twice and then in ρ therefore bounds the displayed second derivative by $C\rho$. This argument is uniform in finite L by (68).

The slow score can also be computed before setting $g = \widehat{g}$:

$$\mathcal{J}_y = w_1\{g^2 y + p\rho g^2 - p\rho^2 g\} + O(\rho^5). \tag{78}$$

Indeed, $F_1 = w_1\{g(y + p\rho) - p\rho^2\} + O(\rho^4)$, and $\sum_{j\geq 2} g(D_y\mathcal{V}_j)F_j = O(\rho^5)$. Consequently, with $e = g - \widehat{g}$ and $\delta = y - (1-p)\rho$,

$$\mathcal{J}_g = c_g e, \quad \tfrac{1}{2} \leq c_g \leq 2, \qquad \mathcal{J}_y = w_1 p^2\rho^2\delta + O(\rho^2|e| + \rho^5), \qquad \|\mathcal{J}_\nu\| \leq C\rho^5. \tag{79}$$

These estimates hold throughout the coarse box, including $g = 0$.

3. Exact physical updates and anisotropic coordinate errors. Put

$$\boldsymbol{k} = t_1 D_\nu A_0 + D_\nu B_0.$$

The raw updates, including every endpoint and block occurrence, are exactly

$$\begin{aligned} \Delta r &= -\frac{\eta C_0}{2}\mathcal{J}_g, \\ \Delta t_1 &= -\eta d_t A_0 \mathcal{J}_y, \\ \Delta\nu &= -\eta \mathsf{D}_\nu\{r(D_\nu C_0)\mathcal{J}_g + \boldsymbol{k}\mathcal{J}_y + \mathcal{J}_\nu\}. \end{aligned} \tag{80}$$

The complementary-coordinate update includes both the chart-induced cross terms and the direct score $\mathcal{J}_\nu$.

The pushforward mobility has

$$\begin{aligned} G_{gg} &= C_0^2/2 + r^2(D_\nu C_0)^{\mathsf{T}}\mathsf{D}_\nu D_\nu C_0, \\ G_{gy} &= r(D_\nu C_0)^{\mathsf{T}}\mathsf{D}_\nu \boldsymbol{k}, & G_{g\nu} &= r(D_\nu C_0)^{\mathsf{T}}\mathsf{D}_\nu, \\ G_{yy} &= d_t A_0^2 + \boldsymbol{k}^{\mathsf{T}}\mathsf{D}_\nu \boldsymbol{k}, & G_{y\nu} &= \boldsymbol{k}^{\mathsf{T}}\mathsf{D}_\nu. \end{aligned} \tag{81}$$

On the coarse box, $G_{gg} \geq c > 0$, $G_{gy} = O(\rho^2)$, $G_{g\nu}, G_{y\nu} = O(\rho)$, and $c/P_{\mathrm{e}} \leq G_{yy} \leq C/P_{\mathrm{e}}$. These entries are the pushforward of the physical mobility in (62).

The raw score bounds imply

$$\begin{aligned} |\Delta r| &\leq C\eta|e|, \\ |\Delta t_1| &\leq C\eta\{\rho^2(|\delta| + |e|) + \rho^5\}, \\ \|\Delta\nu\| &\leq C\eta\{\rho|e| + \rho^3|\delta| + \rho^5\}. \end{aligned} \tag{82}$$

The chart remainders are more structured than a generic $O(\eta^2\|\nabla R\|^2)$ bound. Because $g = rC_0(\nu)$ and $y = A_0(\nu)t_1 + B_0(\nu)$,

$$\begin{aligned} |T_g| &\leq C\{|\Delta r|\|\Delta\nu\| + \rho\|\Delta\nu\|^2\}, \\ |T_y| &\leq C\{|\Delta t_1|\|\Delta\nu\| + \|\Delta\nu\|^2\}. \end{aligned} \tag{83}$$

There is no $(\Delta r)^2$ term in either expression. Using $|e|, |\delta| \leq C\rho$ gives the explicit bounds

$$\begin{aligned} |T_g| &\leq C\eta^2\{\rho e^2 + \rho^4|e| + \rho^9\}, \\ |T_y| &\leq C\eta^2\{\rho^2 e^2 + \rho^4|e| + \rho^7\}. \end{aligned} \tag{84}$$

We now track the moving root rather than a fixed target for g. The exact first-order chart update is

$$\Delta g = -\eta(G_{gg}\mathcal{J}_g + G_{gy}\mathcal{J}_y + G_{g\nu}\mathcal{J}_\nu) + T_g,$$

and analogously for Δy. The root increment is

$$\Delta\widehat{g} = \widehat{g}_y\Delta y + (D_\nu\widehat{g})^{\mathsf{T}}\Delta\nu + O\big(\rho(|\Delta y|^2 + \|\Delta\nu\|^2)\big).$$

The two mobility cross terms in Δg are bounded by $C\eta(\rho^4|e| + \rho^5)$ and $C\eta\rho^6$. The first-order root motion is bounded by $C\eta(\rho^4|e| + \rho^5)$. Its quadratic remainder is bounded by $C\eta^2(\rho^3 e^2 + \rho^7)$. Finally,

since $|e| \le C\rho$, the potentially larger term $\eta^2\rho e^2$ in T_g is at most $C\eta^2\rho^2|e|$. It can be absorbed into the fast contraction by reducing ρ_0, with a fixed η_0 independent of ρ. We obtain

$$|e_{n+1}| \le (1 - c_f\eta)|e_n| + C\eta\rho^5. \tag{85}$$

For the slow coordinate define the reference coefficient $a_n = w_1 p^2 G_{yy,n}$. Equations (79)–(84) give

$$y_{n+1} = y_n - \eta a_n \rho^2 \{y_n - (1-p)\rho\} + \epsilon_n, \qquad \frac{c}{P_e} \le a_n \le \frac{C}{P_e},$$
$$|\epsilon_n| \le C\eta(\rho^2|e_n| + \rho^5). \tag{86}$$

For example, $\eta^2\rho^2 e_n^2$ is bounded by a constant times $\eta\rho^2|e_n|$, and the remaining data-independent part of T_y is smaller than $\eta\rho^5$. The nuisance estimate retains the slow displacement:

$$\|\Delta\nu_n\| \le C\eta\{\rho|e_n| + \rho^3|\delta_n| + \rho^5\}. \tag{87}$$

Retaining the slow displacement in this bound controls its accumulated contribution over the first stage.

The coefficient a_n in (86) is defined using the fast-root expansion. It is not an assertion that the actual readout already equals $p\rho$ during startup. All discrepancies caused by $g_n - \widehat{g}$ remain in ϵ_n.

4. Discrete convolutions and a noncircular first exit. Choose η_0 so that all scalar contraction factors below lie in $(0, 1)$, and write

$$a_f = 1 - c_f\eta, \qquad b_s = 1 - c_s\eta\rho^2/P_e.$$

For small ρ, $b_s - a_f \ge c_f\eta/2$. The exact convolution identity

$$\eta \sum_{j=0}^{n-1} b_s^{n-1-j} a_f^j = \eta \frac{b_s^n - a_f^n}{b_s - a_f} \le C b_s^n \tag{88}$$

is uniform in $0 < \eta \le \eta_0$. From $g_0 = \omega_0$, $y_0 = 0$, $\nu_0 = 0$, and $\widehat{g}(0,0) = \operatorname{arctanh}(p\rho)$, (85) gives $|e_n| \le C\rho a_f^n + C\rho^5$. Substituting this into (86) yields

$$|\delta_n| \le (1-p)\rho b_s^n + C\rho^3 b_s^n + C\eta(\rho^5 + \rho^7)\sum_{j=0}^{n-1} b_s^j$$
$$\le C\rho b_s^n + C\rho^3.$$

In particular, the startup forcing contributes a decaying $O(\rho^3 b_s^n)$ term, not a permanent startup error floor.

For every stopped prefix through n_1, put $H_1 = \eta n_1$. The same envelopes give

$$\sum_{n<n_1} \eta|e_n| \le C(\rho + \rho^5 H_1), \qquad \sum_{n<n_1} \eta|\delta_n| \le C(\rho^{-1} + \rho^3 H_1). \tag{89}$$

Here the fixed factor $P_e \le B$ is absorbed in C. Summing (87) and using $H_1 = O(\rho^{-2}\ell_\rho)$ gives

$$\sum_{n<n_1} \|\Delta\nu_n\| \le C(\rho^2 + \rho^3\ell_\rho) \le C\rho^2. \tag{90}$$

For clarity, the scaled box is closed without choosing a radius from a constant that itself depends on that radius. Let $p_+ = \max\mathcal{I}$ and $p_- = \min\mathcal{I}$. First choose $G_* > c_h + 2p_+ + 2$ and $Y_* > 2(1-p_-) + 2$. Obtain all derivative constants on a fixed enlargement of (70), with nuisance radius 2ρ. Then choose the fixed step ceiling and a sufficiently large A_1; finally reduce ρ_0. The fast envelope and $\widehat{g} = p\rho + O(\rho^3)$ keep $|g|$ strictly below $G_*\rho$. The slow envelope keeps $|y|$ strictly below $Y_*\rho$. Equation (90) keeps $\|\nu\| \le C\rho^2 < \rho/2$. Trial segments remain in the enlarged box because the raw increments in (82) are small. Thus no scaled-box exit occurs.

The scaled box maps, for small ρ, strictly inside a fixed physical gate neighborhood: the embedding is near ± 1, the block maps send the two token intervals into their interiors, and all preactivation

signs retain a fixed margin. The preceding stopped argument therefore also rules out gate exit. The uninterrupted physical GD trajectory therefore stays in the population invariant subspace and the strict gate region.

Having obtained $\nu = O(\rho^2)$, the exact mobility sharpens to

$$a_n = \frac{3w_1 p^2}{2P_e} + O(\rho^2) \quad \text{throughout the first stage.} \tag{91}$$

This remains the reference-rate interpretation specified above.

Choose A_1 so that $c_s A_1 > 2$. The slow transient at n_1 is then at most $O(\rho^3)$; the fast transient is smaller than $O(\rho^5)$. This proves (67) and the three uniform envelopes in the lemma.

5. Startup and total physical path. The exact raw readout update can be written

$$r_{n+1} = \left(1 - \frac{\eta C_0^2 c_g}{2}\right) r_n + \frac{\eta C_0 c_g}{2} \widehat{g}.$$

For the chosen step ceiling it preserves $r_n \geq 0$. The fast envelope implies $r_n \asymp \rho$ once $\eta n \geq t_0$, for a fixed t_0 depending only on the fixed parameter ranges. The remaining fast transient is retained in e_n at fixed t_0. If $\omega_0 = 0$, the first physical update is exactly $r_1 = \eta p \rho/2$; no active-Jacobian lower bound is used before startup.

The raw gradient identity and (62) give, in the original physical norm,

$$\|\nabla_\Theta R(\Theta_n)\| \leq C\{|e_n| + \rho^2 |\delta_n| + \rho^5\}. \tag{92}$$

Normal components outside the population invariant subspace vanish by the orthogonal symmetries; all tangent components are included in the raw metric. Using (89),

$$\sum_{n<n_1} \eta \|\nabla_\Theta R(\Theta_n)\| \leq C(\rho + \rho^3 \ell_\rho) \leq C\rho,$$

including the head startup. This proves (66) and completes the first-stage trajectory proof. □

Consequences for the stage interface. At n_1, (69) gives

$$r = p\rho + O(\rho^3), \qquad t_1 = (1-p)\rho + O(\rho^3),$$

and (92) gives the full physical terminal bound $\|\nabla_\Theta R\| \leq C\rho^5$.

The first-stage conditional deficit follows without replacing the physical optimizer. The chain rule gives

$$\mathcal{J}_y = gF_1 + \sum_{j\geq 2} g(D_y \mathcal{V}_j) F_j, \qquad \mathcal{J}_g = F_0 + yF_1 + \sum_{j\geq 2} \mathcal{V}_j F_j.$$

At the checkpoint, the score bounds and $g \asymp \rho$ imply $F_1 = O(\rho^4)$ and $F_0 = O(\rho^5)$. Uniform strong convexity of the coefficient risk in its first two rows therefore yields

$$\mathcal{E}_1 \leq C(|F_0|^2 + |F_1|^2) \leq C\rho^8.$$

If $q \geq 2$, every temporal role other than the first is part of ν. Thus the exact filter gives $u_2 = O(\rho^4)$, whereas $u_2^* = p(1-p)^2\rho^3 + O(\rho^5)$. Profiling the first two rows and charging the higher rows gives

$$R(\Theta_{n_1}) - V_T = \frac{T-2}{2T} p^2 (1-p)^4 \rho^6 + O(\rho^7).$$

If $q = 1$, write $\Phi' = \partial_\alpha \Phi$. The exact coefficient identity instead gives

$$\mathcal{V}_2 = \frac{C_0 \Phi''}{2a(\Phi')^2} y^2.$$

Its prefactor equals one at reset for every $BL \geq 2$. When $BL = 2$, $\Phi'' = 2e$ at reset: this is precisely the propagated-input contribution. Since $\nu = O(\rho^2)$, the prefactor is $1 + O(\rho^2)$, so $u_2 = p(1-p)^2\rho^3 + O(\rho^5)$ and

$$0 \leq R(\Theta_{n_1}) - V_T \leq C\rho^8.$$

Omitted rows contribute their target-risk cost separately; for $BL = 2$ the first omitted row is lag three and has squared scale ρ^8.

These are per-partition first-stage checkpoints. A common-update comparison between different entrance counts still requires the separate continuation to a common first-stage window.

Later stages. Now let $k \geq 2$ and assume the previous stage conclusion. Before learning role k, the risk excess is at most $C\rho^{2k+2}$. Stop at the first exit from a slightly enlarged gate neighborhood. A uniform physical Hessian bound and sufficiently small η imply descent up to that stopping time, hence

$$\sum_{n=n_{P_e,k-1}}^{n_{P_e,k}-1} \eta\|\nabla R_n\|^2 \leq C\rho^{2k+2}, \qquad \sum_n \|\Theta_{n+1} - \Theta_n\| \leq C\rho\sqrt{\ell_\rho}. \tag{93}$$

The fixed factor $P_e \leq B$ is absorbed into C. Strong convexity of the coefficient risk gives $\|u - u^*\| \leq C\rho^{k+1}$. Inverting the product relations in (52) then keeps roles $1, \ldots, k-1$ in positive $O(\rho)$ intervals and gives $|t_k| \leq C\rho$. Therefore (53)–(61) hold throughout the stopped stage. Taking ρ_0 smaller makes the path in (93) too short to reach the outer boundary, closing the first-exit argument.

For $j \geq k+1$, the moment and network bounds give

$$|F_j| \leq C\rho^{j+1}\ell_\rho^{m_\star}, \qquad \|D_f u_j\| + \|D_\nu u_j\| \leq C\rho^j \ell_\rho^{m_\star}, \qquad \|\zeta\| \leq C\rho^{2k+3}\ell_\rho^{m_\star}. \tag{94}$$

Summing (61), its path term is $O(\rho\ell_\rho^{m_\star})O(\rho\sqrt{\ell_\rho})$; its accumulated chart remainder is bounded by the energy in (93); and its direct high-order forcing is at most

$$O(P_e\rho^{-2k}\ell_\rho)\, O(\rho^{2k+3}\ell_\rho^{m_\star}) = O(\rho^3 \ell_\rho^{m_\star+1}), \tag{95}$$

with fixed P_e. Consequently

$$\sup_{n_{P_e,k-1} \leq n \leq n_{P_e,k}} \|\nu_n\| \leq C\rho^2 \ell_\rho^{m_\star}. \tag{96}$$

This is the nuisance bootstrap needed for later stages; future temporal roles are controlled, not frozen.

It remains to prove learning rather than only gate stability. By coefficient strong convexity, the active Jacobian, and the physical metric lower bound,

$$\|\nabla_\Theta R\|^2 \geq \frac{a_k}{P_e}\rho^{2k}\mathcal{E}_k - C\rho^{4k+6}, \qquad a_k > 0. \tag{97}$$

The derivative of the moving profile in (38) is $O(\rho^{2k+3})$, by (49) and (94). Taylor expansion along the actual GD segment, using the complete restricted Hessian rather than only its Gauss–Newton part, gives

$$\mathcal{E}_{k,n+1} \leq \left(1 - \frac{a_k\eta}{8P_e}\rho^{2k}\right)\mathcal{E}_{k,n} + C\eta\rho^{4k+6}. \tag{98}$$

Therefore

$$\mathcal{E}_{k,n} \leq C\rho^{2k+2}\exp\left[-\frac{a_k\eta\rho^{2k}}{8P_e}(n - n_{P_e,k-1})\right] + C\rho^{2k+6}. \tag{99}$$

Choosing A_k sufficiently large proves (39). The constants are selected in the noncircular order: gate and curvature neighborhoods, active-Jacobian inverse bounds, stage constants A_k, and finally a small enough ρ_0. There are only finitely many stages for fixed B, T.

E.2.4 Risk extraction and the ordinary one-loop bridge

Equation (39) calibrates the first $k+1$ coefficients. At the final role stage $k = q$, the learned factors satisfy $t_j = (1-p)\rho + O(\rho^2\ell_\rho^{m_\star})$, while the current-state, injection, and endpoint deviations that can alter a lowest-degree history path are $O(\rho^2\ell_\rho^{m_\star})$. In the exact filter (29), the degree-i contribution to any represented lag $i \leq \min(D_{\mathrm{blk}}, T-1)$ uses exactly i temporal occurrences. Its leading value is therefore $p(1-p)^i\rho^{i+1}$; replacing one temporal factor by its error, or changing a non-temporal factor, costs at least one further power of ρ. The contractive geometric sum is uniform in finite L. This proves (42) simultaneously for all represented lags.

If $q = k$ and $D_{\mathrm{blk}} \geq k+1$, the next temporal occurrence repeats role one, so

$$u_{k+1} = gy^2 t_2 \cdots t_k + O(\rho^{k+3}\ell_\rho^{m_\star}) = c_{k+1}\rho^{k+2} + O(\rho^{k+3}\ell_\rho^{m_\star}). \tag{100}$$

If $q > k$, the coefficient needs the unlearned role t_{k+1} and is $O(\rho^{k+3}\ell_\rho^{m_\star})$; if $D_{\mathrm{blk}} = k$, it is zero. Profiling the lower coefficients, using (46)–(47), now gives

$$R_{B,P_e,L} - V_T = \frac{w_{k+1}}{2}(u_{k+1} - u^*_{k+1})^2 + O(\rho^{2k+5}\ell_\rho^{m_\star}). \tag{101}$$

This proves (40). In the matched one-mode case the leading square vanishes; the deficit bound and the next missing coefficient give the sharper (41).

We finally record the interface that supplies the only adjacent-depth case not covered directly by the stability range of Theorem E.6. It is stated for the same long horizon used by Corollary E.7.

Proposition E.5 (Ordinary one-loop versus deeper loops on a common horizon)**.** *Fix $T \geq B + 2$ and take $P_e = 1$, so $q = B$. At any checkpoint consisting of $n_{1,B}$ followed by a fixed sufficiently long multiple of the final Bth-stage window, full GD from the common reset satisfies, uniformly over finite $L' \geq 2$,*

$$R_{B,1,1}(\Theta_{1,N}) - V_T = \frac{T - B - 1}{2T} c^2_{B+1}\rho^{2B+4} + O(\rho^{2B+5}\ell_\rho^{m_\star}), \tag{102}$$

$$0 \leq R_{B,1,L'}(\Theta_{L',N}) - V_T \leq C\rho^{2B+6}\ell_\rho^{m_\star}. \tag{103}$$

Consequently

$$R_{B,1,1}(\Theta_{1,N}) - R_{B,1,L'}(\Theta_{L',N}) = \frac{T - B - 1}{2T} c^2_{B+1}\rho^{2B+4} + O(\rho^{2B+5}\ell_\rho^{m_\star}) > 0. \tag{104}$$

In particular, these statements hold at the prescribed finishing checkpoint N_1 in (106).

Proof. A one-loop ordinary model makes exactly B physical calls. Throughout the fixed-gate trajectory its logit has no coefficient at lag $B + 1$ or beyond. Minimizing $F_{\rho,T}$ over all coefficients supported on lags $0, \ldots, B$ and using (49) gives a structural lower bound equal to the right-hand side of (102) up to $O(\rho^{2B+5}\ell_\rho^{m_\star})$; the first excluded coefficient is $u^*_{B+1} = c_{B+1}\rho^{B+2} + O(\rho^{B+4})$ and its profiled curvature is $w_{B+1} + O(\rho^2)$. This representational lower bound remains valid while the shallow model continues to train.

At $n_{1,B}$, (40) supplies the matching upper bound for the shallow risk. Subsequent sufficiently small-step GD is risk nonincreasing, so the shallow risk at N is no larger than that upper bound. The structural floor and this monotonic upper bound sandwich it to (102).

For $L' \geq 2$, all B roles have learned by $n_{1,B}$ and at least one further temporal occurrence exists. The single-mode repetition in (100) matches the first otherwise missing coefficient, so (41) gives the upper bound at $n_{1,B}$. Subsequent sufficiently small-step GD is risk nonincreasing. The same first-exit and nuisance estimates remain valid over any fixed multiple of the last stage window, so they include the finishing interval defining N_1. The positive leading term in (104) dominates its remainder as $\rho \to 0$. □

For the ladder horizon $T = BL_{\max} + 2$, Proposition E.5 proves the ordinary $L = 1$ to $L = 2$ step. The remaining ordinary steps, and every step for $P_e \geq 2$, satisfy the paired-training range of Theorem E.6. This use of the stage lemma is purely single-mode. In a multi-timescale HMM, repeated roles need not match c_{k+1}, and old-coefficient residuals can contribute at the same order as the new-history gain; the present signed higher-depth result does not follow for that setting.

E.3 Independently trained depth and finite ladders

We now hold the prediction task, physical backbone, partition, initialization, and gradient step size fixed. Only the number of complete loops changes. Use the independently trained risk in (31). Unlike the third-stage comparison in Theorem E.19, this result waits until all q roles finish. It therefore includes ordinary $P_e = 1$ and every legal partition, but does not compare different partitions at the same update count.

E.3.1 THE COMMON TRAINING CHECKPOINT

The $q = B/P_{\mathrm{e}}$ distinct temporal roles are learned on successively slower scales. Choose sufficiently large constants $A_1, \dots, A_q, A_{\mathrm{fin}} > 0$, and define

$$n_{P_{\mathrm{e}},0} = 0, \qquad n_{P_{\mathrm{e}},k} = n_{P_{\mathrm{e}},k-1} + \left\lceil \frac{A_k P_{\mathrm{e}} \ell_\rho}{\eta \rho^{2k}} \right\rceil, \quad 1 \le k \le q, \tag{105}$$

$$N_{P_{\mathrm{e}}} = n_{P_{\mathrm{e}},q} + \left\lceil \frac{A_{\mathrm{fin}} P_{\mathrm{e}} \ell_\rho}{\eta \rho^{2q}} \right\rceil. \tag{106}$$

The final interval uses the same full-gradient update as every preceding step. It neither freezes parameters nor selects a favorable iterate after training. Its purpose is to reduce the actual terminal gradient enough to resolve a very small difference between depths. Constants can be chosen jointly over all divisors of a fixed B. The checkpoint is a sufficient training time, not a matching lower bound on the time needed to learn.

Theorem E.6 (Strict depth improvement after independent training)**.** *Use the stationary single-mode binary HMM and source-mean loss in Subsection E.1, the physical network in Subsection E.1.1, and the exact paired reset in Assumption E.1. Fix $B \ge 2$, T, a compact interval $\mathcal{I} \subset (0,1)$ for $p = \kappa^2$, and the allowed initial-head constant. There are positive stage constants, $\eta_0, \rho_0, C, m_\star$ such that the following holds for $P_{\mathrm{e}} \mid B$, $p \in \mathcal{I}$, $0 < \eta \le \eta_0$, and $0 < \rho \le \rho_0$.*

Let $L < L'$ be positive integers, $q = B/P_{\mathrm{e}}$, and $D_{\mathrm{blk}} = BL$, with

$$2q \le D_{\mathrm{blk}} \le T - 2. \tag{107}$$

Train both depths from the same physical initialization for the common number of updates $N_{P_{\mathrm{e}}}$ in (106). Then

$$\mathcal{R}_{P_{\mathrm{e}},L}(N_{P_{\mathrm{e}}}) - \mathcal{R}_{P_{\mathrm{e}},L'}(N_{P_{\mathrm{e}}}) = \frac{T - D_{\mathrm{blk}} - 1}{2T} p^2 (1-p)^{2D_{\mathrm{blk}}+2} \rho^{2D_{\mathrm{blk}}+4} + \varepsilon_{P_{\mathrm{e}},L,L'}, \tag{108}$$

$$|\varepsilon_{P_{\mathrm{e}},L,L'}| \le C \rho^{2D_{\mathrm{blk}}+5} \ell_\rho^{m_\star} \exp\!\big(C\sqrt{\ell_\rho}\big). \tag{109}$$

In particular the risk difference is strictly positive for sufficiently small ρ. The constants depend on the fixed task and backbone, not on ρ or η; the finite-filter estimates are uniform in the deeper finite depth L'. No uniform growing-B or growing-T assertion is made.

More loop passes let the learned temporal rule use more distant tokens without adding block parameters. The coefficient in (108) quantifies this benefit. A shallow network with D_{blk} physical applications cannot use lag $D_{\mathrm{blk}} + 1$. The learned deeper network supplies its leading optimal coefficient $p(1-p)^{D_{\mathrm{blk}}+1}\rho^{D_{\mathrm{blk}}+2}$. Cross-entropy assigns a quadratic gain equal to half its square, multiplied by the fraction $(T - D_{\mathrm{blk}} - 1)/T$ of positions where this lag is available. The remainder accounts for changes in older coefficients and in the parameters learned by the two networks.

The sufficient condition $D_{\mathrm{blk}} \ge 2q$ becomes $P_{\mathrm{e}} L \ge 2$. It covers ordinary loop training from $L = 2$, and every genuine multi-entry partition from $L = 1$. It is a stability condition used by the proof, not a necessary condition for depth to help. Ordinary $L = 1$ is handled separately below.

Depth-role superposition and looping play different roles: the former changes how shared blocks learn, while the latter reuses learned computation over a longer history. Theorem E.6 establishes the latter benefit at matched updates, not matched block calls. The study in Subsection 2.3 instead compares methods across loop depths; their relative risk at each depth is not determined by this within-method result.

Corollary E.7 (A finite depth ladder on one common task)**.** *Fix a finite integer $L_{\max} \ge 2$ and set $T = BL_{\max} + 2$ once for the entire comparison. Under the assumptions of Theorem E.6, for every fixed $P_{\mathrm{e}} \mid B$,*

$$\mathcal{R}_{P_{\mathrm{e}},1}(N_{P_{\mathrm{e}}}) > \mathcal{R}_{P_{\mathrm{e}},2}(N_{P_{\mathrm{e}}}) > \cdots > \mathcal{R}_{P_{\mathrm{e}},L_{\max}+1}(N_{P_{\mathrm{e}}}). \tag{110}$$

For ordinary training, the first gap follows from Proposition E.5; all remaining gaps follow from Theorem E.6. Once both depths are at least $L_{\max}+1$, the comparison becomes an unsigned saturation bound:

$$|\mathcal{R}_{P_{\mathrm{e}},L}(N_{P_{\mathrm{e}}}) - \mathcal{R}_{P_{\mathrm{e}},L'}(N_{P_{\mathrm{e}}})| \le C \rho^{2B(L_{\max}+1)+5} \ell_\rho^{m_\star} e^{C\sqrt{\ell_\rho}}, \qquad L, L' \ge L_{\max} + 1. \tag{111}$$

This result does not compare different P_{e} at a common update count: the checkpoints $N_{P_{\mathrm{e}}}$ generally differ. Nor does it assert monotonic improvement through infinitely many loops on a fixed finite task. The horizon is long enough to contain each of the finitely many history contributions being compared, and is identical for those comparisons.

E.3.2 Why strict improvement can have little practical magnitude

Let $\Delta_L = \mathcal{R}_{P_{\mathrm{e}},L}(N_{P_{\mathrm{e}}}) - \mathcal{R}_{P_{\mathrm{e}},L+1}(N_{P_{\mathrm{e}}})$. Whenever two successive differences are covered, the leading terms satisfy

$$\frac{\Delta_{L+1}}{\Delta_L} \sim \frac{T - B(L+1) - 1}{T - BL - 1}\big[(1-p)\rho\big]^{2B}. \tag{112}$$

An extra loop reaches B additional token lags. Their predictive signal has already decayed, and risk improvement is quadratic in that signal. For weak ρ, the multiplier is very small. Consequently, a strict trained-risk ordering is compatible with nearly indistinguishable practical losses and does not by itself justify additional compute.

E.4 Proof of independently trained depth gains

This appendix proves Theorem E.6 and Corollary E.7. The point that requires more than a forward expressivity calculation is that the two depths are trained independently. We first compare their objectives at a common physical parameter, then control the difference between the two full-gradient trajectories. Throughout, constants may depend on the fixed quantities B, T, the compact set of admissible p, and the fixed gate neighborhood. They do not depend on ρ, η, or the deeper finite depth L'.

Proof roadmap. The argument compares the two training trajectories rather than subtracting coarse risk upper bounds:

1. Exact pairing and source symmetry reduce the network while retaining the physical gradient metric. Stage training learns the q temporal roles and controls the other trainable directions.
2. The finite-filter identity controls all coefficients changed by additional loops, including old lags and endpoints. The resulting full-gradient forcing is $O(\rho^{2D_{\mathrm{blk}}+2})$, up to logarithmic factors.
3. The negative part of the risk Hessian is bounded by the gradient divided by ρ, plus a small residual. Descent bounds its accumulated effect by $e^{C\sqrt{\ell_\rho}}$, instead of $e^{C\eta N_{P_{\mathrm{e}}}}$.
4. The finishing interval bounds the terminal gradient. Readout contraction improves the head's trajectory error by a further factor of ρ, placing retraining effects below the depth signal.
5. Single-mode geometric matching suppresses old-lag contributions. The first newly available lag supplies the positive term in (108).

E.4.1 Proof interface and notation

We use exactly one inherited optimization result, the stage lemma Lemma E.3. We record here the parts of its interface that enter the paired argument. For $P_{\mathrm{e}} \mid B$, put $q = B/P_{\mathrm{e}}$. Full physical gradient descent from the paired reset remains in the paired, q-periodic, fixed-gate subspace. In an orthonormal coordinate system ϑ on this invariant linear subspace, the update is still Euclidean gradient descent,

$$\vartheta_{n+1} = \vartheta_n - \eta \nabla R_{P_{\mathrm{e}},L}(\vartheta_n). \tag{113}$$

Thus no factor-space metric is omitted below. At the stage times $n_{P_{\mathrm{e}},k}$, the first k temporal roles have been learned by this same uninterrupted update. During stage k, the already learned roles are positive and of order ρ, the active role is $O(\rho)$, and every complementary raw deviation from reset is $O(\rho^2 \ell_\rho^{m_{\star,0}})$. For $k \geq 2$, the stage starts with excess at most $C\rho^{2k+2}$ and lasts $O(P_{\mathrm{e}} \rho^{-2k} \ell_\rho)$ units of physical time. The first stage is handled separately by its fast–slow recurrence, which gives

$$\sum_{n < n_{P_{\mathrm{e}},1}} \eta \|\nabla R_{P_{\mathrm{e}},L}(\vartheta_n)\| \leq C\rho. \tag{114}$$

After all q roles have been learned, the single-mode matching conclusion is

$$R_{P_{\mathrm{e}},L}(\vartheta_{n_{P_{\mathrm{e}},q}}) - V_T \leq C\rho^{2q+6} \ell_\rho^{m_{\star,0}}, \qquad BL \geq q+1. \tag{115}$$

The lemma also provides the weighted first- and second-derivative estimates used below, including all repeated occurrences, the embedding, and the head. The short initial interval needed when the head is initialized at zero is part of the first stage. These are consequences of the physical stage proof, not extra assumptions on a coefficient-space optimizer.

For two depths $L < L'$, define

$$D_{\mathrm{blk}} = BL = qP_{\mathrm{e}}L, \qquad D_{\mathrm{blk}}{}' = BL' = qP_{\mathrm{e}}L'. \tag{116}$$

Thus $P_{\mathrm{e}}L$ counts theory-group applications and D_{blk} counts physical block calls per source. The methodology's $K = PL$ retains its worker-group meaning. We write $u^{(L)}(\vartheta) = (u_0^{(L)}, \dots, u_{T-1}^{(L)})$ for the linear-history coefficients produced at depth L. Source symmetry makes every source prediction equal on the invariant trajectory, so the source-mean network risk is exactly

$$R_{P_{\mathrm{e}},L}(\vartheta) = F_{\rho,T}(u^{(L)}(\vartheta)). \tag{117}$$

The loss normalization matters for the leading constant. For $\ell(z, y) = \log(2\cosh z) - yz$,

$$\partial_z \ell(z, y) = \tanh z - y, \qquad \partial_z^2 \ell(z, y) = \operatorname{sech}^2 z = 1 + O(z^2). \tag{118}$$

In particular, the curvature at zero is one, not $1/4$.

E.4.2 Risk geometry of the single-mode task

For $u \in \mathbb{R}^T$, let

$$z_t(u) = \sum_{i=0}^{t-1} u_i X_{t-i}, \qquad F_{\rho,T}(u) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\ell(z_t(u), X_{t+1}), \tag{119}$$

and write $F_i = \partial_{u_i} F_{\rho,T}$ and $F_{ij} = \partial_{u_i}\partial_{u_j} F_{\rho,T}$. Set $w_i = (T - i)/T$. The stationary one-mode HMM has

$$\mathbb{E}[X_s X_{s+h}] = p\rho^h \quad (h \geq 1), \qquad p = \kappa^2. \tag{120}$$

Lemma E.8 (Weighted coefficient geometry)**.** *In a fixed weighted neighborhood satisfying $|u_i| \leq C\rho^{i+1}$, the following estimates hold uniformly for $0 \leq i, j < T$:*

$$|F_i(u)| \leq C\rho^{i+1}, \qquad F_{ii}(u) = w_i + O(\rho^2), \qquad |F_{ij}(u)| \leq C\rho^{|i-j|}. \tag{121}$$

The unique optimal linear-history coefficient vector obeys

$$u_i^* = c_i\rho^{i+1} + O(\rho^{i+3}), \qquad c_i = p(1-p)^i. \tag{122}$$

If a coefficient vector matches these leading terms through lag m, namely $u_i = u_i^* + O(\rho^{i+2}\ell_\rho^a)$ for $i \leq m$, and has $u_i = O(\rho^{i+1})$ afterwards, then

$$F_i(u) = O(\rho^{i+2}\ell_\rho^a), \qquad i \leq m. \tag{123}$$

Proof. Differentiating (119) and using (118) gives

$$F_i(u) = \frac{1}{T}\sum_{t=i+1}^{T} \mathbb{E}[(\tanh z_t(u) - X_{t+1})X_{t-i}], \tag{124}$$

$$F_{ij}(u) = \frac{1}{T}\sum_{t>\max\{i,j\}} \mathbb{E}[\operatorname{sech}^2 z_t(u) X_{t-i}X_{t-j}]. \tag{125}$$

The weighted coefficient bound implies $z_t = O(\rho)$ uniformly. Expanding $\operatorname{sech}^2 z_t$ and applying the HMM moment-pairing bound proves the two Hessian estimates in (121). The same moment count applied to (124) proves the gradient bound. More explicitly, at the order ρ^{i+1} the stationarity equation is

$$0 = u_i + p\sum_{j<i} \rho^{i-j}u_j - p\rho^{i+1} + O(\rho^{i+3}). \tag{126}$$

Writing $u_j = c_j\rho^{j+1} + O(\rho^{j+3})$ yields $c_i = p(1 - \sum_{j<i} c_j)$. Starting from $c_0 = p$, this recursion gives $c_i = p(1-p)^i$, proving (122).

For the last claim, use $F_i(u^*) = 0$ and integrate the Hessian on the segment from u^* to u. For $j \leq m$, the contribution is bounded by $C\rho^{|i-j|}\rho^{j+2}\ell_\rho^a$. It is at most $C\rho^{i+2}\ell_\rho^a$. For $j > m \geq i$, the first omitted contribution is of order $\rho^{j-i}\rho^{j+1}$, which is no larger than the same bound. This proves (123). □

E.4.3 THE EXACT FINITE-FILTER DIFFERENCE

Inside the invariant fixed-gate subspace, number the q distinct block roles backwards from the output and write

$$H_j(z) = \alpha_j + \beta_j z, \qquad A(z) = \prod_{j=1}^{q} H_j(z), \qquad C(z) = \sum_{j=1}^{q} \gamma_j \prod_{r<j} H_r(z). \tag{127}$$

If e denotes the initial-state endpoint, the exact filter after $P_{\mathrm{e}}L$ group applications is

$$Q_L(z) = eA(z)^{P_{\mathrm{e}}L} + C(z) \sum_{a=0}^{P_{\mathrm{e}}L-1} A(z)^a. \tag{128}$$

This identity retains the initial product and every endpoint. Subtracting the two geometric sums gives the exact relation

$$Q_{L'} - Q_L = A^{P_{\mathrm{e}}L}\big[(A-1)e + C\big] \sum_{a=0}^{P_{\mathrm{e}}(L'-L)-1} A^a. \tag{129}$$

In the stage tube supplied by Lemma E.3,

$$|\alpha_j| + |\gamma_j - e| \le C\rho^2 \ell_\rho^{m_\star,0}, \qquad |\beta_j| + |r| \le C\rho, \tag{130}$$

with bounded endpoint derivatives. The neighborhood may be chosen so that $|\alpha_j| + 2|\beta_j| < 1/4$. Therefore the last sum in (129), together with any fixed number of its physical derivatives, has a geometric majorant independent of L'. This is the source of uniformity in the deeper finite depth; it is not a growing-B or growing-T statement.

Lemma E.9 (Weighted depth perturbation). *Let $\delta u_i = u_i^{(L')} - u_i^{(L)}$ be evaluated at one common physical parameter in the stage tube. For $i \le D_{\mathrm{blk}}$,*

$$|\delta u_i| \le C\rho^{2D_{\mathrm{blk}}-i+3}\ell_\rho^{m_\star,1}, \qquad \|D_\vartheta \delta u_i\| \le C\rho^{2D_{\mathrm{blk}}-i+1}\ell_\rho^{m_\star,1}. \tag{131}$$

For $i > D_{\mathrm{blk}}$,

$$|\delta u_i| \le C\rho^{i+1}, \qquad \|D_\vartheta \delta u_i\| \le C\rho^{i}. \tag{132}$$

After all q roles have learned their single-mode ratios, the first new coefficient satisfies

$$\delta u_{D_{\mathrm{blk}}+1} = p(1-p)^{D_{\mathrm{blk}}+1}\rho^{D_{\mathrm{blk}}+2} + O(\rho^{D_{\mathrm{blk}}+3}\ell_\rho^{m_\star,1}). \tag{133}$$

The depth-L coefficient at lag $D_{\mathrm{blk}}+1$ is exactly zero.

Proof. To form z^i from $A^{P_{\mathrm{e}}L}$, one chooses i temporal factors and $D_{\mathrm{blk}} - i$ current-state factors. The corresponding scale is $\rho^i(\rho^2)^{D_{\mathrm{blk}}-i} = \rho^{2D_{\mathrm{blk}}-i}$. The constant coefficient of $[(A-1)e + C]$ is $O(\rho^2\ell_\rho^{m_\star,0})$. Its positive-degree coefficients contain the required additional temporal factors and give no larger term. Multiplication by the readout $r = O(\rho)$ proves the first estimate in (131). A physical derivative can remove either one $O(\rho)$ factor or one $O(\rho^2)$ factor. Allowing the larger two-power loss gives the second estimate. This loss is essential: differentiating a current-state or injection deviation can remove two powers, so counting only temporal derivatives would be incorrect.

For $i > D_{\mathrm{blk}}$, every contributing path contains at least i temporal factors; including the readout gives (132). Finally, after stage q, each learned temporal ratio is $(1-p)\rho + O(\rho^2\ell_\rho^{m_\star,0})$, while the current-token coefficient is $p\rho + O(\rho^2\ell_\rho^{m_\star,0})$. The first path absent from depth L uses $D_{\mathrm{blk}} + 1$ temporal transfers. This gives (133). The degree of Q_L is at most D_{blk}, so its lag-$(D_{\mathrm{blk}}+1)$ coefficient vanishes identically. □

Lemma E.10 (Full-gradient forcing). *Throughout the common enlarged stage tube,*

$$\|\nabla R_{P_{\mathrm{e}},L'}(\vartheta) - \nabla R_{P_{\mathrm{e}},L}(\vartheta)\| \le \mathfrak{f}_\rho, \qquad \mathfrak{f}_\rho = C\rho^{2D_{\mathrm{blk}}+2}\ell_\rho^{m_\star,1}. \tag{134}$$

The bound includes the embedding, head, endpoints, and all repeated physical occurrences.

Proof. With $u = u^{(L)}$ and $\delta u = u^{(L')} - u^{(L)}$, the chain rule gives

$$\nabla R_{P_e,L'} - \nabla R_{P_e,L} = (D\delta u)^\top F'(u + \delta u) + (Du)^\top \{F'(u + \delta u) - F'(u)\}. \tag{135}$$

For $i \le D_{\mathrm{blk}}$, the ith summand in the first line is at most $C\rho^{2D_{\mathrm{blk}}-i+1}\rho^{i+1}\ell_\rho^{m_\star,1} = C\rho^{2D_{\mathrm{blk}}+2}\ell_\rho^{m_\star,1}$. The terms with $i > D_{\mathrm{blk}}$ begin at order $\rho^{2D_{\mathrm{blk}}+3}$. In the second line, a typical old-coordinate term has scale

$$\rho^i\rho^{|i-j|}\rho^{2D_{\mathrm{blk}}-j+3}\ell_\rho^{m_\star,1} \le C\rho^{2D_{\mathrm{blk}}+3}\ell_\rho^{m_\star,1}. \tag{136}$$

New-coordinate terms have the same or higher order. Since T is fixed, the finite sums prove (134). □

E.4.4 Negative curvature along the actual trajectory

The objective is nonconvex in physical parameters, so a global smoothness bound would amplify (134) by the useless factor $\exp(C\eta N_{P_e})$. The next estimate controls only the negative part of the Hessian, which is what can expand the distance between two gradient steps.

Lemma E.11 (Gradient-controlled negative curvature)**.** *Outside the fixed head-startup interval and during stage k, $1 \le k \le q$,*

$$\lambda_{\min}(\nabla^2 R_{P_e,L}(\vartheta)) \ge -\frac{C}{\rho}\|\nabla R_{P_e,L}(\vartheta)\| - C\rho^{2k+2}. \tag{137}$$

The inequality holds for the Hessian on the full invariant physical subspace, not merely for its active-coordinate principal block.

Proof. Let $f = (r, t_1, \dots, t_k)$ denote the active physical coordinates and let $u_{\mathrm{lo}} = (u_0, \dots, u_k)$. The inherited weighted Jacobian factorization is

$$J_k = D_f u_{\mathrm{lo}} = \mathrm{diag}(1, \rho, \dots, \rho^k)\mathsf{K}_k, \qquad \|\mathsf{K}_k\| + \|\mathsf{K}_k^{-1}\| \le C. \tag{138}$$

The last diagonal entry does not contain the not-yet-learned t_k. Weighted second derivatives imply, for every unit physical direction v,

$$\left\|J_k^{-1} D^2 u_{\mathrm{lo}}[v, v]\right\| \le C/\rho. \tag{139}$$

Indeed the ith second derivative is $O(\rho^{i-1})$ for $i \ge 1$ and is $O(1)$ for $i = 0$; row scaling by J_k^{-1} loses at most one power of ρ.

The Hessian decomposes exactly as

$$\nabla^2 R = (Du)^\top F''(u) Du + \sum_i F_i(u) D^2 u_i. \tag{140}$$

The first term is positive semidefinite because $F_{\rho,T}$ is convex. For the low residual term, even when v has nuisance components,

$$\sum_{i \le k} F_i D^2 u_i[v, v] = (J_k^\top F_{\mathrm{lo}})^\top J_k^{-1} D^2 u_{\mathrm{lo}}[v, v]. \tag{141}$$

Here $J_k^\top F_{\mathrm{lo}}$ is the active part of the full physical gradient minus the high-coefficient contribution. The latter is bounded by

$$\sum_{i>k} \|D_f u_i\|\,|F_i| \le C \sum_{i>k} \rho^i \rho^{i+1} \le C\rho^{2k+3}. \tag{142}$$

Equations (139)–(142) therefore bound the low residual curvature by $C\|\nabla R\|/\rho + C\rho^{2k+2}$. The high residual curvature is also bounded by

$$\sum_{i>k} |F_i|\,\|D^2 u_i\| \le C \sum_{i>k} \rho^{i+1}\rho^{i-1} \le C\rho^{2k+2}. \tag{143}$$

Combining these estimates with (140) proves the claim for every full physical direction v.

If the head starts at zero, (138) is temporarily singular. The first-stage recurrence in Lemma E.3 reaches $r \asymp \rho$ in $O(1)$ physical time. On this initial interval the full Hessian has a fixed upper bound, so its total expansion is only a constant. No inverse Jacobian is used before the head has started. □

Lemma E.12 (Integrated expansion)**.** *Along either depth trajectory through the q learning stages and the finishing interval,*

$$\sum_{n<N_{P_e}} \eta a_n \le C\sqrt{\ell_\rho}, \tag{144}$$

where a_n is an admissible upper bound for the magnitude of the negative curvature in (137). The finishing interval contributes $o(1)$ to this sum.

Proof. The first-stage gradient path is (114). For $k \ge 2$, descent and the stage-start excess from Lemma E.3 give

$$\sum_{n\in\text{stage } k} \eta\|\nabla R_n\|^2 \le C\rho^{2k+2}. \tag{145}$$

Cauchy–Schwarz and the physical duration $O(P_e\rho^{-2k}\ell_\rho)$ then imply

$$\sum_{n\in\text{stage } k} \eta\|\nabla R_n\| \le C\rho\sqrt{\ell_\rho}. \tag{146}$$

After division by ρ, the first stage contributes $O(1)$ and every later stage contributes $O(\sqrt{\ell_\rho})$. There are only the fixed number q of stages. The additive residual in (137) contributes

$$O(\rho^{2k+2})\,O(P_e\rho^{-2k}\ell_\rho) = O(\rho^2\ell_\rho) \tag{147}$$

per stage. The fixed startup contributes a constant.

At $n_{P_e,q}$, (115) and descent give a finishing path of size

$$\sum_{n=n_{P_e,q}}^{N_{P_e}-1} \eta\|\nabla R_n\| \le C\rho^3\ell_\rho^{m_\star,2}. \tag{148}$$

Its gradient-controlled negative curvature is $O(\rho^2\ell_\rho^{m_\star,2})$, and its additive residual is also $O(\rho^2\ell_\rho)$. Both vanish as $\rho\to 0$. This proves (144). □

E.4.5 The nonlinear paired-distance bootstrap

Let

$$\theta_n = \vartheta_{L,n}, \qquad \theta'_n = \vartheta_{L',n}, \qquad d_n = \|\theta'_n - \theta_n\|. \tag{149}$$

Both trajectories begin from the same physical reset, so $d_0 = 0$.

Proposition E.13 (Distance between independently trained depths)**.** *If $D_{\text{blk}} \ge 2q$, then*

$$\sup_{n\le N_{P_e}} d_n \le \overline{d}_\rho, \qquad \overline{d}_\rho = C\rho^{2D_{\text{blk}}+2-2q}\ell_\rho^{m_\star,2}\exp(C\sqrt{\ell_\rho}). \tag{150}$$

In particular $\overline{d}_\rho = o(\rho^2)$.

Proof. Subtracting the two updates and applying the integral mean-value formula to $R_{P_e,L}$ gives

$$\theta'_{n+1} - \theta_{n+1} = (I - \eta\overline{H}_n)(\theta'_n - \theta_n) - \eta\{\nabla R_{P_e,L'}(\theta'_n) - \nabla R_{P_e,L}(\theta'_n)\}, \tag{151}$$

where $\overline{H}_n$ is the average Hessian of $R_{P_e,L}$ on the segment from θ_n to θ'_n. The gradient is Lipschitz in the fixed tube. Thus the negative-curvature estimate on this segment differs from the estimate at θ_n by at most Cd_n/ρ. Taking η below the fixed positive-curvature step ceiling, Lemmas E.10 and E.11 yield

$$d_{n+1} \le (1+\eta a_n)d_n + \frac{C\eta}{\rho}d_n^2 + \eta\mathfrak{f}_\rho, \qquad \sum_n \eta a_n \le C\sqrt{\ell_\rho}. \tag{152}$$

The total physical time is $O(P_e\rho^{-2q}\ell_\rho)$. Ignoring the quadratic term momentarily and applying discrete Gronwall to the forcing in (152) gives the scale in (150). To close the argument, stop at the

first iterate where d_n would exceed that scale. Before this exit, the cumulative relative effect of the quadratic term is bounded by

$$\frac{C\eta N_{P_e}}{\rho}\overline{d}_\rho \le C\rho^{2D_{\mathrm{blk}}+1-4q}\ell_\rho^{m_{\star,2}+1}\exp(C\sqrt{\ell_\rho}). \tag{153}$$

When $D_{\mathrm{blk}} \ge 2q$, the power of ρ is at least one. Since $\exp(C\sqrt{\ell_\rho}) = \rho^{-o(1)}$, the right-hand side tends to zero. Discrete Gronwall then returns a strict fraction of the proposed exit radius, which rules out the exit. At the boundary $D_{\mathrm{blk}} = 2q$, the exponent in (150) is $2q+2 \ge 4$, so the entire segment between the trajectories remains in the enlarged weighted tube. This also verifies that the segment estimates used above are not circular. □

E.4.6 Terminal stationarity and readout damping

The distance bound alone is too coarse: multiplying $\overline{d}_\rho$ by a fixed risk Lipschitz constant would be much larger than the depth signal. Two endpoint refinements recover the missing powers.

Lemma E.14 (Gradient at the prescribed terminal iterate)**.** *For every depth covered by Theorem E.6,*

$$\|\nabla R_{P_e,L}(\theta_{N_{P_e}})\| \le C\rho^{2q+3}\ell_\rho^{m_{\star,3}}. \tag{154}$$

The bound holds at the actual endpoint N_{P_e}; no iterate is selected retrospectively.

Proof. Let $g_n = \nabla R_{P_e,L}(\theta_n)$. Along one GD step,

$$g_{n+1} = (I - \eta\widetilde{H}_n)g_n, \tag{155}$$

where $\widetilde{H}_n$ is the Hessian averaged over that step. The positive part of the Hessian is below the step ceiling, and Lemma E.12 shows that the accumulated negative curvature over the finishing interval is $o(1)$. Hence $\|g_{N_{P_e}}\| \le C\|g_n\|$ for every iterate n in that interval. Descent and (115) give

$$\sum_{n=n_{P_e,q}}^{N_{P_e}-1}\eta\|g_n\|^2 \le C\rho^{2q+6}\ell_\rho^{m_{\star,0}}. \tag{156}$$

The finishing duration is $\Theta(P_e\rho^{-2q}\ell_\rho)$. Averaging (156) and using the preceding propagation inequality proves (154). □

Separate the physical readout amplitude r from all other physical coordinates ζ. Pairing makes every logit exactly

$$z = rf_\zeta(X). \tag{157}$$

In the stage tube, $r = O(\rho)$, $f_\zeta = X_t + o(1)$, and the first two derivatives of f_ζ are bounded. Moreover $|\mathbb{E}[X_{t+1} \mid X_{\le t}]| \le \kappa\rho$. Applying (118) to (157) gives the full Hessian block bounds

$$R_{rr} \in [c, C], \qquad \|R_{r\zeta}\| \le C\rho, \qquad \|R_{\zeta\zeta}\| \le C\rho^2. \tag{158}$$

For completeness, the residual part of the last block is $r\mathbb{E}[(\tanh z - X_{t+1})D^2_{\zeta\zeta}f_\zeta]$. Conditioning on the observed history makes the expectation $O(\rho)$, and the outer r supplies the second power. Thus (158) does not silently replace the physical Hessian by its Gauss–Newton part.

Lemma E.15 (Readout difference and endpoint risk)**.** *Let $\Delta r_n = r'_n - r_n$. Under $D_{\mathrm{blk}} \ge 2q$,*

$$|\Delta r_{N_{P_e}}| \le C\rho\overline{d}_\rho, \tag{159}$$

$$|R_{P_e,L'}(\theta'_{N_{P_e}}) - R_{P_e,L'}(\theta_{N_{P_e}})| \le C\rho^{2D_{\mathrm{blk}}+5}\ell_\rho^{m_\star}\exp(C\sqrt{\ell_\rho}). \tag{160}$$

Proof. Use the first row of the Hessian in the mean-value formula for the two head updates, and use Lemma E.10 for the objective mismatch. The result is

$$|\Delta r_{n+1}| \le (1 - c\eta)|\Delta r_n| + C\eta\rho d_n + C\eta\mathfrak{f}_\rho. \tag{161}$$

The initial head difference is zero. Geometric summation, the distance bound, and $\mathfrak{f}_\rho \le C\rho\overline{d}_\rho$ give (159). This last inequality follows directly from $D_{\mathrm{blk}} \ge 2q$.

Expand $R_{P_e,L'}$ at its own terminal point $\theta'_{N_{P_e}}$. The linear term is bounded by Lemma E.14 as

$$C\rho^{2q+3}\ell_\rho^{m_\star,3}\overline{d}_\rho \le C\rho^{2D_{\mathrm{blk}}+5}\ell_\rho^{m_\star}e^{C\sqrt{\ell_\rho}}. \tag{162}$$

Writing the displacement as $(\Delta r, \Delta\zeta)$ and using (158)–(159), the quadratic term is at most

$$C(\Delta r)^2 + C\rho|\Delta r|\|\Delta\zeta\| + C\rho^2\|\Delta\zeta\|^2 \le C\rho^2\overline{d}_\rho^2. \tag{163}$$

Its power of ρ is $4D_{\mathrm{blk}} + 6 - 4q$, which is at least $2D_{\mathrm{blk}} + 6$ when $D_{\mathrm{blk}} \ge 2q$. Absorbing logarithmic and subpolynomial factors proves (160). A crude $C\overline{d}_\rho^2$ estimate would have order $\rho^{2D_{\mathrm{blk}}+4}$ at $D_{\mathrm{blk}} = 2q$ and would fail to separate the desired signal; the readout estimate is therefore necessary. □

E.4.7 The same-parameter risk signal

At the terminal shallow parameter $\theta_{N_{P_e}}$, the single-mode stage result and the small finishing path give

$$u_i^{(L)} = u_i^* + O(\rho^{i+2}\ell_\rho^{m_\star,0}), \qquad 0 \le i \le D_{\mathrm{blk}}. \tag{164}$$

The conclusion extends beyond the q independently learned roles because the single learned transfer ratio is geometric at leading order. This is the step that would fail for a general multi-timescale target.

Lemma E.16 (Risk gain at a common physical parameter)**.** *If $D_{\mathrm{blk}} \le T - 2$, then at $\theta_{N_{P_e}}$,*

$$\begin{aligned} &R_{P_e,L}(\theta_{N_{P_e}}) - R_{P_e,L'}(\theta_{N_{P_e}}) \\ &\quad = \frac{T - D_{\mathrm{blk}} - 1}{2T}p^2(1-p)^{2D_{\mathrm{blk}}+2}\rho^{2D_{\mathrm{blk}}+4} + O(\rho^{2D_{\mathrm{blk}}+5}\ell_\rho^{m_\star,4}). \end{aligned} \tag{165}$$

Proof. By Lemma E.8 and (164), all old residuals satisfy

$$F_i(u^{(L)}) = O(\rho^{i+2}\ell_\rho^{m_\star,0}), \qquad i \le D_{\mathrm{blk}}. \tag{166}$$

Combining this with (131) gives

$$\sum_{i \le D_{\mathrm{blk}}} F_i(u^{(L)})\delta u_i = O(\rho^{2D_{\mathrm{blk}}+5}\ell_\rho^{m_\star,4}). \tag{167}$$

This is the old-coefficient cancellation: the old residuals are one power smaller because the one-mode leading coefficients match. Without that extra power, this term would be of the same order as the new-history gain.

Put

$$a_{D_{\mathrm{blk}}} = p(1-p)^{D_{\mathrm{blk}}+1}, \qquad w_{D_{\mathrm{blk}}+1} = \frac{T - D_{\mathrm{blk}} - 1}{T}. \tag{168}$$

Since the shallow model has $u^{(L)}_{D_{\mathrm{blk}}+1} = 0$ whereas $u^*_{D_{\mathrm{blk}}+1} = a_{D_{\mathrm{blk}}}\rho^{D_{\mathrm{blk}}+2} + O(\rho^{D_{\mathrm{blk}}+4})$, the triangular risk geometry gives

$$F_{D_{\mathrm{blk}}+1}(u^{(L)}) = -w_{D_{\mathrm{blk}}+1}a_{D_{\mathrm{blk}}}\rho^{D_{\mathrm{blk}}+2} + O(\rho^{D_{\mathrm{blk}}+3}\ell_\rho^{m_\star,0}). \tag{169}$$

Taylor-expand $F_{\rho,T}(u^{(L)} + \delta u)$ at $u^{(L)}$. Equations (133) and (169) show that the negative linear term in $F(u^{(L)}) - F(u^{(L)} + \delta u)$ is $w_{D_{\mathrm{blk}}+1}a^2_{D_{\mathrm{blk}}}\rho^{2D_{\mathrm{blk}}+4}$ to leading order. By (121), the corresponding negative quadratic term is $-\frac{1}{2}w_{D_{\mathrm{blk}}+1}a^2_{D_{\mathrm{blk}}}\rho^{2D_{\mathrm{blk}}+4}$. Their sum is therefore

$$\frac{w_{D_{\mathrm{blk}}+1}}{2}a^2_{D_{\mathrm{blk}}}\rho^{2D_{\mathrm{blk}}+4}. \tag{170}$$

The loss normalization in (118) is exactly what produces $w_{D_{\mathrm{blk}}+1}$ here. The old linear terms are controlled by (167); old–old and old–new quadratic terms start at order $\rho^{2D_{\mathrm{blk}}+6}$; and lags at least $D_{\mathrm{blk}} + 2$ start at order $\rho^{2D_{\mathrm{blk}}+6}$. Bounded third derivatives give a still higher-order Taylor remainder. Substituting (168) proves (165). □

Proof of Theorem E.6. Let $\theta_{N_{P_e}}$ and $\theta'_{N_{P_e}}$ be the two independently trained terminal parameters. Insert and subtract the deeper objective at the shallow terminal parameter:

$$R_{P_e,L}(\theta_{N_{P_e}}) - R_{P_e,L'}(\theta'_{N_{P_e}}) = [R_{P_e,L}(\theta_{N_{P_e}}) - R_{P_e,L'}(\theta_{N_{P_e}})] \tag{171}$$

$$+ [R_{P_e,L'}(\theta_{N_{P_e}}) - R_{P_e,L'}(\theta'_{N_{P_e}})]. \tag{172}$$

The first bracket is Lemma E.16; the second is bounded by Lemma E.15. This proves (108)–(109). Finally,

$$\rho \ell_\rho^{m_\star} \exp(C\sqrt{\ell_\rho}) \longrightarrow 0 \qquad (\rho \downarrow 0), \tag{173}$$

so the explicit leading coefficient is strictly positive for sufficiently small ρ. □

E.4.8 Finite ladders and saturation

The upper restriction $D_{\mathrm{blk}} \le T-2$ was used only to ensure that lag $D_{\mathrm{blk}}+1$ is scored and hence supplies a signed leading term. The forcing, curvature, trajectory, and head arguments remain valid when every scored lag is already available. This gives the following unsigned endpoint estimate.

Proposition E.17 (Finite-horizon saturation)**.** *Fix a depth L_0 with $D_{\mathrm{blk}0} = BL_0 \ge \max\{2q, T-1\}$. For every finite $L' > L_0$,*

$$|\mathcal{R}_{P_e,L_0}(N_{P_e}) - \mathcal{R}_{P_e,L'}(N_{P_e})| \le C\rho^{2D_{\mathrm{blk}0}+5}\ell_\rho^{m_\star} \exp(C\sqrt{\ell_\rho}). \tag{174}$$

Proof. Use $D_{\mathrm{blk}0}$ in Lemmas E.9 and E.10. Since $i \le T-1 \le D_{\mathrm{blk}0}$ for every scored coefficient, there is no new-lag term. The learned one-mode matching gives $F_i = O(\rho^{i+2}\ell_\rho^{m_\star,0})$ for every scored i. Hence the common- parameter risk difference is $O(\rho^{2D_{\mathrm{blk}0}+5}\ell_\rho^{m_\star})$ by exactly the sum in (167). The paired-distance and readout arguments use only $D_{\mathrm{blk}0} \ge 2q$ and add the subpolynomial factor shown in (174). □

Proof of Corollary E.7. Set $T = BL_{\max}+2$. For $P_e \ge 2$, the shallow side of every adjacent comparison $L \to L+1$, $1 \le L \le L_{\max}$, satisfies

$$D_{\mathrm{blk}} = BL \ge B \ge 2B/P_e = 2q, \qquad D_{\mathrm{blk}} \le BL_{\max} = T-2. \tag{175}$$

Thus Theorem E.6 proves every strict inequality. For $P_e = 1$, the same argument covers $L = 2, \dots, L_{\max}$. It remains to justify the first ordinary comparison without applying the sufficient condition $D_{\mathrm{blk}} \ge 2q$ outside its range.

At $k = q = B$, the absolute-risk interface in Lemma E.3 gives the single-mode depth-$L \ge 2$ model excess $O(\rho^{2B+6}\ell_\rho^{m_\star,0})$ at $n_{1,B}$; descent preserves this upper bound through the finishing interval. A one-loop filter has degree at most B, so its lag-$(B+1)$ coefficient is identically zero. Minimizing $F_{\rho,T}$ over all lower coefficients while imposing this missing coefficient gives, by the Schur complement of the Hessian in Lemma E.8,

$$\inf_{u_{B+1}=\cdots=u_{T-1}=0}[F_{\rho,T}(u) - V_T] = \frac{T-B-1}{2T}p^2(1-p)^{2B+2}\rho^{2B+4} + O(\rho^{2B+5}). \tag{176}$$

The finishing path for the one-loop trajectory is $O(\rho^2\sqrt{\ell_\rho})$, so it remains in the same fixed-gate linear-filter region. Consequently (176) is a lower bound on its actual terminal risk and proves $\mathcal{R}_{1,1}(N_1) > \mathcal{R}_{1,2}(N_1)$ for small ρ.

Finally take $L_0 = L_{\max}+1$. Then $D_{\mathrm{blk}0} = B(L_{\max}+1) \ge T-1$ and $D_{\mathrm{blk}0} \ge 2q$. Applying Proposition E.17 to L_0 and any $L' > L_0$, and using the triangle inequality when both displayed depths exceed L_0, proves (111). The construction uses one fixed finite horizon and does not pass to an infinite-depth limit. □

Remark E.18 (Why the gains rapidly diminish). For two consecutive comparisons covered by Theorem E.6, division of their leading coefficients gives

$$\frac{T-B(L+1)-1}{T-BL-1}\big[(1-p)\rho\big]^{2B}. \tag{177}$$

Thus strict positivity and rapid saturation are compatible: another loop reaches B more lags, while both temporal correlation and the quadratic risk signal decay over those lags. This is a finite-horizon asymptotic statement, not a claim of practical finite-ρ improvement or an experiment.

E.5 Entrance count at the third training stage

The model fixes a single HMM and scores the same shared history coefficients at every available prefix. We now compare every legal partition of one physical backbone, at one update count. This is a comparison of independently trained networks, not an ensemble or a choice of entrance during training. Put $q = B/P_{\mathrm{e}}$ and $\ell_\rho = \log(1/\rho)$. Throughout this section $B \geq 6$ and $T \geq B + 2$ are fixed, $P_{\mathrm{e}} \mid B$, $L \geq 1$ is finite, and all trajectories use the exact reset and source-mean population GD of Subsection E.1 and Assumption E.1. The conclusions are uniform over finite L.

Theorem E.19 (All-divisor third-stage risk table)**.** *Fix $B \geq 6$, $T \geq B + 2$, $\Gamma > 1$, a compact $\mathcal{I} \Subset (0, 1)$ for $p = \kappa^2$, and the allowed initial-head constant c_h. There exist $A_*, \eta_0, \rho_0, C > 0$ and finite $m_\star$, depending only on these fixed quantities, such that, for $p \in \mathcal{I}$, $0 < \eta \leq \eta_0$, $0 < \rho \leq \rho_0$, every $P_{\mathrm{e}} \mid B$, every finite $L \geq 1$, and every integer*

$$N_- \leq N \leq N_+ := \lfloor \Gamma N_- \rfloor, \qquad N_- := \left\lceil \frac{A_* \ell_\rho}{\eta \rho^6} \right\rceil, \tag{178}$$

the actual trained excess $E_{P_{\mathrm{e}},L,N} := \mathcal{R}_{P_{\mathrm{e}},L}(N) - V_T$ obeys

q	*excess risk after full physical GD*
$q \geq 4$	$E_{P_{\mathrm{e}},L,N} = A_T(p)\rho^{10} + O(\rho^{11}\ell_\rho^{m_\star})$
$q = 3$	$0 \leq E_{P_{\mathrm{e}},L,N} \leq C\rho^{14}\ell_\rho^{m_\star}$
$q = 2$	$0 \leq E_{P_{\mathrm{e}},L,N} \leq C\rho^{12}\ell_\rho^{m_\star}$
$q = 1$	$0 \leq E_{P_{\mathrm{e}},L,N} \leq C\rho^{12}\ell_\rho^{m_\star}$.

(179)

Here the explicit positive coefficient is

$$A_T(p) = \frac{T-4}{2T} p^2 (1-p)^8 > 0. \tag{180}$$

If, in addition, $\inf_{p \in \mathcal{I}} |2 - (T-1)p| > 0$, then the last row has the lower bound

$$E_{B,L,N} \geq C_T(p)\rho^{12} - C\rho^{13}\ell_\rho^{m_\star},$$
$$C_T(p) = \frac{p^4(1-p)^6[2-(T-1)p]^2}{2T(T-1)^2(T-2)^2(T-3)} > 0. \tag{181}$$

All constants and remainders are uniform over the stated $P_{\mathrm{e}}, L, N, p, \eta$. The minimizer set $\mathcal{P}^*_{B,L,N} := \arg\min_{P_{\mathrm{e}}|B} \mathcal{R}_{P_{\mathrm{e}},L}(N)$ satisfies

$$\begin{aligned} &6 \mid B: && \mathcal{P}^*_{B,L,N} \subseteq \{B/3, B/2\}, \\ &3 \mid B,\ 2 \nmid B: && \mathcal{P}^*_{B,L,N} = \{B/3\}, \\ &2 \mid B,\ 3 \nmid B: && \mathcal{P}^*_{B,L,N} \subseteq \{B/2, B\}, \\ &\gcd(B,6) = 1: && \mathcal{P}^*_{B,L,N} = \{B\}. \end{aligned} \tag{182}$$

Only the first two lines require the extra nondegeneracy condition.

This baseline table does not yet use the path-selected head bias. Its $q = 2$ upper bound alone cannot rank two roles against three roles. For example, it leaves $P_{\mathrm{e}} \in \{2, 3\}$ when $B = 6$, whereas $B = 9$ has the unique optimum $P_{\mathrm{e}} = 3$ and $B = 7$ the unique optimum $P_{\mathrm{e}} = 7$. Theorem E.29 and Corollary E.30 below sharpen the two-role row and resolve the remaining pair away from an additional, explicit degeneracy. The positive $q \geq 4$ coefficient excludes every such divisor because the always-legal $q = 1$ upper bound has smaller order. When $3 \mid B$, the $q = 3$ upper bound also excludes $q = 1$ by (181).

The one-role obstruction is specific to shared, finite-prefix linear-history cross-entropy. At the optimum, the first non-geometric minor is

$$u_1^* u_3^* - (u_2^*)^2 = \frac{p^3(1-p)^4[2-(T-1)p]}{(T-1)(T-2)(T-3)} \rho^8 + O(\rho^{10}). \tag{183}$$

The one-role filter has a geometric lag tail up to a negligible finite-depth correction. Its minor therefore cannot match (183); the necessary lag-three coefficient error is order ρ^6, giving risk order ρ^{12}. The sign in (183) is immaterial after squaring. At $p = 2/(T-1)$ this leading obstruction vanishes, so we do not exclude full entry there. The appendix proves the statistical expansion, the finite-L control, and the training bounds in Theorem E.19.

Corollary E.20 (Depth improvement after optimizing the partition)**.** *Under the nondegenerate assumptions of Theorem E.19, suppose $3 \mid B$. Fix $L < L'$ with $BL \le T-2$ and use the same integer N from (178) for every partition and both depths. Define $\mathcal{V}_{B,L,N} := \min_{P_e|B} \mathcal{R}_{P_e,L}(N)$. Then*

$$\mathcal{V}_{B,L,N} - \mathcal{V}_{B,L',N} = G_{B,L,T}(p)\rho^{2BL+4} + O\Big(\rho^{2BL+5}\ell_\rho^{m_\star} e^{C\sqrt{\ell_\rho}}\Big) > 0, \tag{184}$$

$$G_{B,L,T}(p) = \frac{T-BL-1}{2T}p^2(1-p)^{2BL+2} > 0. \tag{185}$$

The minimizing partition may differ between the two depths. This is a finite same-update comparison; it does not assert an infinite ladder at fixed T.

For a fixed finite depth range $1 \le L \le L_{\max}$, the table also admits one specially chosen work comparison. Choose $\Gamma > BL_{\max} + 1$ before setting (178), put $W = B^2 L_{\max} N_-$, and stop configuration (P_e, L) at $\lfloor W/(P_e BL)\rfloor$ updates. Every stop lies in the same certified window, so (179)–(182) continue to hold at these stops. This is a block-call proxy, not a general fixed-work depth ranking: the depths now receive different update counts.

E.6 Proofs for the common-window partition results

All estimates below concern the original physical update (22). We use $\ell = \log(1/\rho)$ and write $O_\ell(\rho^a)$ for $O(\rho^a \ell^{m_\star})$ with a finite exponent $m_\star$ that may increase between displays. Constants are uniform over the finite divisor set of fixed B, compact p-set, $0 < \eta \le \eta_0$, all finite depths, and the certified update interval. No constant is claimed uniform in growing B or T. The exact symmetry, physical mobility, gate region, and role-learning estimates are proved in Subsection E.1 and Lemma E.3; the finite-depth derivative estimates and paired-trajectory machinery are proved in Subsection E.4. We give the additional arguments needed to extend those interfaces to one common third-stage window.

E.6.1 A conditional-risk descent identity for the actual update

Fix a cutoff $m < T-1$ and split the exact filter coefficients into $a = (u_0, \dots, u_m)$ and $b = (u_{m+1}, \dots, u_{T-1})$. In the small-logit region, $F = F_{\rho,T}$ has a uniformly positive coefficient Hessian; Lemma E.8 gives diagonal entries $w_i = (T-i)/T + O(\rho^2)$ and off-diagonal lag decay. Set

$$a_*(b) = \arg\min_v F(v,b), \quad H_m(\Theta) = F(a_*(b(\Theta)), b(\Theta)), \quad \mathcal{E}_m(\Theta) = F(a(\Theta), b(\Theta)) - H_m(\Theta). \tag{186}$$

The minimizer $a_*(b)$ is smooth by the implicit function theorem; its first two derivatives are bounded by the inverse *coefficient* Hessian, not by an inverse network Jacobian. On a closed stopped tube strictly inside the gate neighborhood, (29) and its first two physical derivatives are uniformly bounded: on $|z| \le 2$ each role factor has norm below $1/4$, and $\sup_{m\ge 0} m^2(1/4)^m < \infty$ controls two product derivatives. The same bounds hold on line segments in a sufficiently small convex sub-tube. Chain differentiation of (186) therefore gives a uniform physical Hessian bound on every actual GD step segment, including nuisance directions and both endpoints.

Lemma E.21 (Physical conditional-profile descent)**.** *Let $J_a = D_\Theta a$ and $J_b = D_\Theta b$ in orthonormal coordinates of the invariant physical subspace. Along the segment $(a_*(b) + t(a - a_*(b)), b)$, put*

$$A = \int_0^1 F_{aa}\, dt, \quad C = \int_0^1 F_{ba}\, dt, \quad \mathsf{Q} = CA^{-1}. \tag{187}$$

Then, exactly,

$$\nabla\mathcal{E}_m = (J_a + \mathsf{Q}^\mathsf{T} J_b)^\mathsf{T} F_a, \qquad \nabla H_m = J_b^\mathsf{T} F_b(a_*(b), b). \tag{188}$$

For sufficiently small fixed η, one full physical GD step satisfies

$$\mathcal{E}_m(\Theta^+) \le \mathcal{E}_m(\Theta) - \frac{\eta}{2}\|\nabla\mathcal{E}_m\|^2 + C\eta\|\nabla H_m\|^2. \tag{189}$$

Proof. Writing $v = a - a_*(b)$ and integrating the fundamental theorem of calculus gives $F_a = Av$ and $F_b(a,b) - F_b(a_*,b) = Cv = \mathsf{Q}F_a$. The envelope theorem gives $D_b F(a_*(b), b) = F_b(a_*, b)$; subtracting its chain rule from the chain rule for $F(a,b)$ proves (188). Taylor expansion of $\mathcal{E}_m$ at $\Theta^+ = \Theta - \eta(\nabla\mathcal{E}_m + \nabla H_m)$, the bounded physical Hessian, and Young's inequality absorb the cross term and the $O(\eta^2)$ term into half of the negative square, proving (189). The update is the unrestricted, full-parameter GD step; H_m is not separately optimized. □

For later use, coefficient strong convexity implies $\|F_a\|^2 \geq c\mathcal{E}_m$. Moreover, on the weighted role tube, $\|\mathsf{Q}\| = O(\rho)$: at $\rho = 0$ distinct history lags are orthogonal, and the finite-prefix off-diagonal Hessian has the lag decay of Lemma E.8. If selected physical columns f satisfy

$$\sigma_{\min}(J_{a,f}) \geq c\rho^s/\sqrt{P_{\mathrm{e}}}, \qquad \|J_{b,f}J_{a,f}^{-1}\| = O_\ell(\rho), \tag{190}$$

then $(J_a+\mathsf{Q}^{\mathsf{T}}J_b)_f = (I+\mathsf{Q}^{\mathsf{T}}J_{b,f}J_{a,f}^{-1})J_{a,f}$ has the same least-singular-value order. Thus $\|\nabla\mathcal{E}_m\|^2 \geq (c/P_{\mathrm{e}})\rho^{2s}\mathcal{E}_m$. The $1/P_{\mathrm{e}}$ is the physical role mobility in (26).

E.6.2 Training precision through the third-stage window

The preceding stage lemma gives, after role q is learned, the actual excess $O_\ell(\rho^{2q+6})$ and conditional deficit $\mathcal{E}_q = O_\ell(\rho^{2q+6})$. Its weighted filter expansion, obtained from (29) by product differentiation in the stopped tube, is

$$u_j = g\prod_{a=1}^{j} t_{1+(a-1) \bmod q} + O_\ell(\rho^{j+3}), \qquad j \leq \min(D_{\mathrm{blk}}, T-1), \tag{191}$$

with $g = p\rho + O_\ell(\rho^3)$ after the first stage and $t_a = O(\rho)$. The error has two extra powers because $\alpha_a, \gamma_a - e = O_\ell(\rho^2)$ and each temporal deviation from its effective t_a contributes $O_\ell(\rho^3)$. The initial-product term $eA^{P_{\mathrm{e}}L}$ is retained; at the boundary $j = D_{\mathrm{blk}}$ the same order follows from the cancellation in (29). The coefficient optimum in Subsection E.2.1 satisfies $u_j^* = c_j\rho^{j+1} + O(\rho^{j+3})$, $c_j = p(1-p)^j$. Strong convexity and the conditional deficit, followed by successive ratios of (191), sharpen the learned roles to

$$t_a = (1-p)\rho + O_\ell(\rho^3), \qquad 1 \leq a \leq q. \tag{192}$$

Indeed the low-coefficient error is $O_\ell(\rho^{q+3})$; its effect on the ath ratio is $O_\ell(\rho^{q+3-a})$, which is at least two powers beyond ρ for $a \leq q$. It follows that every represented high coefficient differs from its optimum by $O_\ell(\rho^{j+3})$. The coefficient Hessian's lag decay gives the same bound for the conditional high residual $F_j(a_*(b), b)$ for represented rows $j \leq D_{\mathrm{blk}}$. In detail, a low-row correction has size $O_\ell(\rho^{q+3})$ in the last low row, and its contribution to high row j is multiplied by the Hessian factor $O(\rho^{j-q})$; earlier rows are smaller after their own weighted scaling. Thus the sum is $O_\ell(\rho^{j+3})$, including the finite-filter boundary term retained in (29). For omitted rows $j > D_{\mathrm{blk}}$, $u_j \equiv 0$ and $Du_j \equiv 0$; their residual may be only $O(\rho^{j+1})$, but it contributes to risk only at order $\rho^{2D_{\mathrm{blk}}+4}$ and contributes nothing to the physical profile gradient.

Lemma E.22 (Finishing the learned roles)**.** *For $q \leq 3$ and $D_{\mathrm{blk}} = BL \geq q+2$, after the qth role checkpoint and an additional $AP_{\mathrm{e}}\ell/(\eta\rho^{2q})$ updates, with sufficiently large fixed A, the actual risk obeys*

$$R_{B,P_{\mathrm{e}},L}(\Theta) - V_T \leq C\rho^{2q+8}\ell^{m_\star}. \tag{193}$$

The bound persists through any fixed multiple of this training time that remains below the next independent-role scale.

Proof. Use $m = q$ and active raw coordinates $f = (r, t_1, \ldots, t_q)$. Triangular differentiation of (191) gives $J_{a,f} = \mathrm{diag}(1, \rho, \ldots, \rho^q)\mathsf{K}$ with $\|\mathsf{K}\| + \|\mathsf{K}^{-1}\| \leq C$, after the bounded physical mobility conversion. A high row $j \geq q+1$ has active derivative $O_\ell(\rho^j)$; weighted triangular inversion gives $J_{b,f}J_{a,f}^{-1} = O_\ell(\rho)$. Hence (190) holds with $s = q$. By the refined high residual and the derivative bound $\|Du_j\| = O_\ell(\rho^j)$, obtained by differentiating the contractive finite product in every physical coordinate,

$$\|\nabla H_q\| \leq C\sum_{j\geq q+1} \rho^j\rho^{j+3}\ell^{m_\star} \leq C\rho^{2q+5}\ell^{m_\star}. \tag{194}$$

Combining (189) with strong convexity gives

$$\mathcal{E}_q^+ \leq (1 - c\eta\rho^{2q}/P_{\mathrm{e}})\mathcal{E}_q + C\eta\rho^{4q+10}\ell^{m_\star}. \tag{195}$$

Geometric summation over the stated interval reduces the first term to $O(\rho^{2q+10})$ and leaves a floor $O_\ell(\rho^{2q+10})$. The first remaining represented high-coefficient error is $O_\ell(\rho^{q+4})$, contributing $O_\ell(\rho^{2q+8})$ to risk; later represented lags are smaller. Since $D_{\mathrm{blk}} \geq q+2$, an omitted lag's leading risk is $O(\rho^{2D_{\mathrm{blk}}+4})$, no earlier than the displayed order. This proves (193).

For closure of the stopped argument, the starting total excess is $O_\ell(\rho^{2q+6})$. Ordinary risk descent and Cauchy–Schwarz bound the path over $O(P_{\mathrm{e}}\rho^{-2q}\ell/\eta)$ steps by $O_\ell(\rho^3)$. This is smaller than the fixed gate margin and preserves the $O_\ell(\rho^2)$ nuisance tube and (192). The first-exit time therefore cannot precede the stated endpoint. □

The case $q = 1$ needs one more trainable coefficient. The shared head's hidden bias b_h is a physical parameter; it is not an added feature.

Lemma E.23 (Head-bias calibration under full GD)**.** *For $q = 1$, $B \geq 6$, and every finite $L \geq 1$, after $AP_{\mathrm{e}}\ell/(\eta\rho^6)$ additional updates beyond the role-finishing interval, with sufficiently large fixed A,*

$$R_{B,B,L}(\Theta) - V_T \leq C\rho^{12}\ell^{m_\star}. \tag{196}$$

This estimate holds on a fixed-multiple continuation of the same third-stage window.

Proof. Take $m = q+1 = 2$ and select the raw physical columns $f = (r, \beta, b_h)$, holding the nuisance coordinates (α, γ, e, a_h) fixed for the Jacobian calculation. This is a coordinate selection inside the full update, not a frozen-parameter optimizer. Put

$$s = \frac{a_h\gamma}{1-\alpha}, \qquad \tau = \frac{\beta}{1-\alpha}, \qquad D_{\mathrm{blk}} = BL. \tag{197}$$

The *exact finite* filter is $Q_{D_{\mathrm{blk}}} = e(\alpha+\beta z)^{D_{\mathrm{blk}}} + \gamma[1-(\alpha+\beta z)^{D_{\mathrm{blk}}}]/(1-\alpha-\beta z)$. Its infinite part has

$$u_{\infty,0} = r(s+b_h), \qquad u_{\infty,j} = rs\tau^j \quad (j \geq 1). \tag{198}$$

The finite correction $ra_h(\alpha+\beta z)^{D_{\mathrm{blk}}}(e - Q_\infty)$ is retained. On the stopped tube,

$$r = p\rho + O_\ell(\rho^3), \quad s = 1 + O_\ell(\rho^2), \quad b_h = O_\ell(\rho^2), \quad \tau = (1-p)\rho + O_\ell(\rho^3). \tag{199}$$

For each represented row $0 \leq j \leq \min(D_{\mathrm{blk}}, T-1)$, set $D_0 = \max\{B, j\}$. Omitted rows $j > D_{\mathrm{blk}}$ have $u_{D_{\mathrm{blk}},j} = 0$ and are treated separately below. Coefficient extraction gives a fixed-D_{blk} bound $C_{D_{\mathrm{blk}}}\rho^{2D_{\mathrm{blk}}-j+3}\ell^{m_{\star,D_{\mathrm{blk}}}}$, whose logarithmic exponent need not be uniform in D_{blk}. Factor the first D_0 powers and use $\sup_{D_{\mathrm{blk}}\geq D_0} {D_{\mathrm{blk}}}^T|\alpha|^{D_{\mathrm{blk}}-D_0} < \infty$ for fixed T and $|\alpha| \leq 1/4$. This gives the uniform estimates

$$|u_{D_{\mathrm{blk}},j} - u_{\infty,j}| \leq C\rho^{2D_0-j+3}\ell^{m_\star}, \qquad \left|\partial_{r,\beta}(u_{D_{\mathrm{blk}},j} - u_{\infty,j})\right| \leq C\rho^{2D_0-j+2}\ell^{m_\star}, \tag{200}$$

with one fixed $m_\star$ independent of D_{blk}. For $j = 0, 1, 2, 3$, $D_0 = B$; these corrections and derivatives are smaller than the displayed coefficient and Jacobian scales by several powers because $B \geq 6$.

For the infinite part, solve $du_0 = du_1 = 0$ under $db_h = 1$. The two equations give $dr = -r/(s+b_h)$ and $d\tau = \tau/(s+b_h)$. Therefore the compensated Schur derivative is

$$S_\infty := \left.\frac{\partial u_{\infty,2}}{\partial b_h}\right|_{u_{\infty,0},u_{\infty,1}} = \frac{rs\tau^2}{s+b_h} = p(1-p)^2\rho^3 + O_\ell(\rho^5). \tag{201}$$

The first-two-row inverse has entries at worst $O_\ell(\rho^{-1})$. Using (200) for the first three rows in its Schur formula gives $|S_{D_{\mathrm{blk}}} - S_\infty| \leq C\rho^{2B-5}\ell^{m_\star}$, hence $S_{D_{\mathrm{blk}}} = p(1-p)^2\rho^3 + O_\ell(\rho^5)$ uniformly for $D_{\mathrm{blk}} \geq B \geq 6$. This is a relative $o(1)$ bound on the actual Schur pivot, not an inference from raw $O(\rho^2)$ perturbations alone. The remaining two pivots have orders 1 and ρ, so the full 3×3 inverse has norm $O_\ell(\rho^{-3})$. Physical mobility changes these estimates only by fixed P_{e} factors.

For each represented high row $j \geq 3$, eliminate the selected columns through the first three rows. In the infinite filter, $u_{\infty,j} = u_{\infty,1}^{2-j}u_{\infty,2}^{j-1}$, so its derivatives with respect to (u_0, u_1, u_2) are

$$\left(0, (2-j)\frac{u_j}{u_1}, (j-1)\frac{u_j}{u_2}\right) = (0, O(\rho^{j-1}), O(\rho^{j-2})). \tag{202}$$

For a represented high row $3 \leq j \leq D_{\mathrm{blk}}$, (200) has $D_0 \geq j$, so its derivative error is $O_\ell(\rho^{j+2})$. Multiplication by the low-row inverse $O_\ell(\rho^{-3})$ changes the ratio by at most $O_\ell(\rho^{j-1})$, hence at most $O_\ell(\rho^2)$; the correction to that inverse is smaller by the first-three-row estimate. For omitted $j > D_{\mathrm{blk}}$, the row derivative is zero. Thus $\|J_{b,f}J_{a,f}^{-1}\| = O_\ell(\rho)$, and (190) holds with $s = 3$ throughout the stopped tube.

The first represented high row is now $j = 3$. For each represented $j \geq 3$, the conditional coefficient residual is $O_\ell(\rho^{j+3})$ and $\|Du_j\| = O_\ell(\rho^j)$; for $j > D_{\mathrm{blk}}$, $Du_j = 0$. Therefore the full physical high-profile gradient satisfies

$$\|\nabla H_2\| \leq C \sum_{j=3}^{\min(D_{\mathrm{blk}},T-1)} \rho^j \rho^{j+3} \ell^{m_\star} \leq C\rho^9 \ell^{m_\star}. \tag{203}$$

Lemma E.21 yields

$$\mathcal{E}_2^+ \leq (1 - c\eta\rho^6/P_{\mathrm{e}})\mathcal{E}_2 + C\eta\rho^{18}\ell^{m_\star}. \tag{204}$$

The starting excess from Lemma E.22 is $O_\ell(\rho^{10})$. Geometric summation makes the conditional deficit $O_\ell(\rho^{12})$. After the moving low profile is calibrated, the first represented high residual at lag three is $O_\ell(\rho^6)$, contributing $O_\ell(\rho^{12})$. Omitted rows start at $j = D_{\mathrm{blk}} + 1 \geq 7$ and contribute $O(\rho^{2D_{\mathrm{blk}}+4})$. More explicitly, coefficient smoothness and the conditional optimum give

$$H_2 - V_T \leq C \left[\sum_{j=3}^{\min(D_{\mathrm{blk}},T-1)} |u_j - u_j^*|^2 + \sum_{j=D_{\mathrm{blk}}+1}^{T-1} |u_j^*|^2 \right] = O_\ell(\rho^{12}), \tag{205}$$

where an empty sum is zero; the represented-row bound uses $t_1 = (1-p)\rho + O_\ell(\rho^3)$ and optimized u_0, u_1, u_2. This proves (196).

For the first-exit check, define the stopped tube by a fixed gate margin, $|\alpha|+|\gamma-e|+|b_h| \leq C_0\rho^2\ell^{m_{\star,0}}$, $|r - p\rho| \leq C_0\rho^3\ell^{m_{\star,0}}$, and $|\tau - h\rho| \leq C_0\rho^3\ell^{m_{\star,0}}$, with C_0 chosen after the established checkpoint bounds. On every prefix before first exit, ordinary full-GD descent and the starting $O_\ell(\rho^{10})$ risk sublevel give

$$\sum \eta\|\nabla R\| \leq \sqrt{\left(\sum \eta\right)\left(\sum \eta\|\nabla R\|^2\right)} \leq C\rho^2\ell^{m_{\star,1}}. \tag{206}$$

This controls each nuisance displacement, including b_h, before using the desired final risk. The same starting sublevel forces the first three coefficients within $O_\ell(\rho^5)$ of their conditional optima; triangular inversion of u_1, u_2 preserves $\tau = h\rho + O_\ell(\rho^3)$. Likewise $u_0 = p\rho + O_\ell(\rho^3)$, and $u_0 = r(s + b_h)$ plus the finite-depth correction gives $r = p\rho + O_\ell(\rho^3)$. These conclusions use the starting risk sublevel and the controlled nuisance path, not the final calibrated bound. Choosing C_0 larger than these derived constants, then shrinking ρ_0, contradicts a first exit. Hence the Schur pivot, row ratio, and high-gradient estimates hold on every actual update segment until the claimed endpoint, without assuming (196) in advance. □

Lemma E.24 (Uniform third-stage stops)**.** *For a suitable A_* in (178), uniformly for all integer $N \in [N_-, N_+]$ and all finite $L \geq 1$,*

$$q \geq 4: \quad E_{P_{\mathrm{e}},L,N} = \frac{w_4}{2} c_4^2 \rho^{10} + O_\ell(\rho^{11}), \tag{207}$$

$$q = 3: \quad E_{P_{\mathrm{e}},L,N} = O_\ell(\rho^{14}),$$

$$q = 2: \quad E_{P_{\mathrm{e}},L,N} = O_\ell(\rho^{12}),$$

$$q = 1: \quad E_{P_{\mathrm{e}},L,N} = O_\ell(\rho^{12}). \tag{208}$$

Proof. For $q = 3$, choose A_* so that role three and the finishing interval in Lemma E.22 end before $N_-/2$. For $q = 2$ the same bound is finished on the earlier $\rho^{-4}\ell$ scale; subsequent full GD decreases the objective. For $q = 1$, choose A_* also to finish Lemma E.23 before $N_-/2$. There are finitely many divisors, so one A_* works. Continuing through N_+ remains in the gate tube: Cauchy–Schwarz with the already attained excess gives post-completion paths $O_\ell(\rho^4)$, $O_\ell(\rho^3)$, and $O_\ell(\rho^3)$ for $q = 3, 2, 1$, respectively. Monotone descent preserves their upper bounds.

For $q \geq 4$, apply the stage lemma Lemma E.3 at $k = 3$. The next role is not learned at this time: its scale is $\rho^{-8}\ell$, versus the present $\rho^{-6}\ell$. Repeating its conditional-deficit recurrence for the fixed extra multiple $[N_-, N_+]$ reduces the transient below $O_\ell(\rho^{11})$ but changes the not-yet-learned fourth coefficient only within that remainder. More explicitly, the fourth role begins at $O_\ell(\rho^2)$; its physical gradient is $O_\ell(\rho^9)$ because the lag-four residual is $O(\rho^5)$ and the derivative of that lag through the fourth role is $O(\rho^4)$. A time $O(\rho^{-6}\ell)$ moves it by at most $O_\ell(\rho^3)$, so its contribution to u_4 is at most $O_\ell(\rho^6)$ and its effect on the leading missing-lag risk is $O_\ell(\rho^{11})$. Later roles have higher powers.

The already learned low coefficients remain controlled by their conditional contraction; the same physical path bound prevents gate exit. The risk extraction (40), with $I_{3,P_e,L} = 0$, then gives (207). The first-exit path estimate of Lemma E.3 remains $o(1)$ over this fixed-multiple interval. All choices precede shrinking ρ_0, so the quantifiers include the entire closed integer window. □

E.6.3 True cross-entropy statistics and the one-role obstruction

Write $h = 1 - p$ and $w_i = (T - i)/T$. At $\rho = 0$, the coefficient Hessian of $F_{\rho,T}$ at zero is $\mathrm{diag}(w_0, \dots, w_{T-1})$. It is positive definite, and $F_{\rho,T}$ is globally strictly convex: its Hessian is an average of positive-weight outer products of observed history vectors with strictly positive sech^2 weights, and each lag has a positive-probability independent innovation. The stationary point near zero is thus its unique global minimizer.

Lemma E.25 (Finite-prefix CE minor)**.** *For fixed T and $p \in \mathcal{I}$, let u^* minimize $F_{\rho,T}$. For $1 \le j \le T - 3$,*

$$u_j^* u_{j+2}^* - (u_{j+1}^*)^2 = \frac{p^3 h^{2j+2}[2 - (T-j)p]}{(T-j)(T-j-1)(T-j-2)} \rho^{2j+6} + O(\rho^{2j+8}). \tag{209}$$

Proof. For lag indices $i_1, \dots, i_m$, binary products reduce repeated indices modulo two. If the remaining distinct indices are $a_1 < \cdots < a_{2r}$, the exact observed-token moment is $p^r \rho^{(a_2 - a_1) + \cdots + (a_{2r} - a_{2r-1})}$; odd products vanish. This follows by conditioning successively on the hidden Markov chain and using $\mathbb{E}[X_t \mid S_t] = \kappa S_t$. The target-history pair has moment $p\rho^{i+1}$. Because a monomial is present only in prefixes containing all its lags, its weight is $w_{\max(i_1,\dots,i_m)}$.

Analytic implicit solving and the alternating-token symmetry give

$$u_i^* = c_i \rho^{i+1} + s_i \rho^{i+3} + O(\rho^{i+5}), \qquad c_i = p h^i. \tag{210}$$

The leading stationary equation is $c_i + p \sum_{a<i} c_a = p$. At order ρ^{i+3}, the linear part receives the s coefficients and only one higher-lag term, from $a = i + 1$, weighted by $r_i = w_{i+1}/w_i = (T - i - 1)/(T - i)$ (set $r_{T-1} = 0$). In $\tanh z = z - z^3/3 + O(z^5)$, the only cubic monomials at this order have at least two zero lags. Their coefficient is $K_0 = p^3$ for $i = 0$ and, for $i \ge 1$,

$$K_i = p c_0^3 + 3 c_0^2 \left(c_i + p \sum_{a=1}^{i-1} c_a \right) = p^3 (3 - 2p). \tag{211}$$

The quintic and higher terms have at least two additional powers. Consequently

$$s_i + p \sum_{a<i} s_a = K_i/3 - p r_i c_{i+1}. \tag{212}$$

For $i \ge 2$, subtract adjacent equations and use $K_i = K_{i-1}$: $s_i - h s_{i-1} = p c_i (r_{i-1} - h r_i)$. With $\zeta_i = s_i/c_i$, this is $\zeta_i - \zeta_{i-1} = p(r_{i-1} - h r_i)$. The leading ρ^{2j+4} term of the minor cancels since $c_j c_{j+2} = c_{j+1}^2$. Its next term is $c_j c_{j+2}(\zeta_j + \zeta_{j+2} - 2\zeta_{j+1})\rho^{2j+6}$. Substituting the two adjacent r differences and simplifying the rational weights yields (209). The remainder is uniform on the compact p-set by analyticity. □

Lemma E.26 (Finite-depth one-role tail)**.** *For $P_e = B$, $D_{\mathrm{blk}} = BL \ge B \ge 6$, and a parameter in the certified weighted tube of Lemma E.24,*

$$|u_1 u_3 - u_2^2| \le C \rho^{2B+2} \ell^{m_\star}, \tag{213}$$

uniformly over all finite $L \ge 1$.

Proof. All roles are identical by the $q = 1$ symmetry. With $t(z) = \alpha + \beta z$, the exact recurrence gives

$$Q_{D_{\mathrm{blk}}} = e t^{D_{\mathrm{blk}}} + \gamma \sum_{m=0}^{D_{\mathrm{blk}}-1} t^m, \quad Q_\infty = \frac{\gamma}{1 - t}, \quad Q_{D_{\mathrm{blk}}} - Q_\infty = t^{D_{\mathrm{blk}}}(e - Q_\infty). \tag{214}$$

For $j \ge 1$, $u_{\infty,j} = r a_h \gamma \beta^j/(1 - \alpha)^{j+1}$, whose adjacent minor is exactly zero. The weighted tube gives $r, \beta = O(\rho)$ and $\alpha, \gamma - e = O_\ell(\rho^2)$; the constant term of $e - Q_\infty$ is $O_\ell(\rho^2)$ and its jth positive coefficient is $O(\rho^j)$. Expanding $t^{D_{\mathrm{blk}}}$ binomially therefore gives, for fixed $1 \le j \le 4$,

$|u_{D_{\mathrm{blk}},j} - u_{\infty,j}| \le C_{D_{\mathrm{blk}}} \rho^{2D_{\mathrm{blk}}-j+3} \ell^{m_\star, D_{\mathrm{blk}}}$. The fixed-$D_{\mathrm{blk}}$ logarithmic exponent is not asserted uniform in D_{blk}. Factoring the first B powers of t and using $\sup_{D_{\mathrm{blk}} \ge B} {D_{\mathrm{blk}}}^T |\alpha|^{D_{\mathrm{blk}}-B} < \infty$ for fixed T and $|\alpha| \le 1/4$ gives $|u_{D_{\mathrm{blk}},j} - u_{\infty,j}| \le C\rho^{2B-j+3}\ell^{m_\star}$ with one $m_\star$ independent of D_{blk}, as in (200). Multiplication by $u_{\infty,j} = O(\rho^{j+1})$ gives (213); the quadratic correction is smaller. The head hidden bias affects only u_0 and cannot change the minor. □

Lemma E.27 (Risk cost of the minor mismatch)**.** *Assume* $\inf_{p\in\mathcal{I}} |2-(T-1)p| > 0$. *For the actual* $q = 1$ *trajectory throughout the common window,*

$$F_{\rho,T}(u) - V_T \ge C_T(p)\rho^{12} - C\rho^{13}\ell^{m_\star}, \tag{215}$$

where $C_T(p)$ *is given in* (181)*.*

Proof. The actual training upper bound (208) and local strong convexity imply $\varepsilon_i := u_i - u_i^* = O_\ell(\rho^6)$ for every scored lag. By Lemma E.25 at $j = 1$,

$$\Delta^* := u_1^* u_3^* - (u_2^*)^2 = \delta_T(p)\rho^8 + O(\rho^{10}), \quad \delta_T(p) = \frac{p^3h^4[2-(T-1)p]}{(T-1)(T-2)(T-3)}. \tag{216}$$

Expand the network minor around u^*:

$$u_1u_3 - u_2^2 = \Delta^* + u_3^*\varepsilon_1 - 2u_2^*\varepsilon_2 + u_1^*\varepsilon_3 + \varepsilon_1\varepsilon_3 - \varepsilon_2^2. \tag{217}$$

Here $u_1^* = ph\rho^2 + O(\rho^4)$, $u_2^* = ph^2\rho^3 + O(\rho^5)$, and $u_3^* = ph^3\rho^4 + O(\rho^6)$. The first two linear error terms are $O_\ell(\rho^9)$ or smaller, and the quadratic errors are $O_\ell(\rho^{12})$; (213) is $O_\ell(\rho^{14})$ because $B \ge 6$. Hence

$$\varepsilon_3 = -\frac{\delta_T(p)}{ph}\rho^6 + O_\ell(\rho^7). \tag{218}$$

To turn this coefficient error into an objective lower bound, profile all coordinates except u_3: let $G_3(v) = \min_{u_i : i\ne 3} F_{\rho,T}(u_0, \dots, u_2, v, u_4, \dots)$. The Schur complement of the coefficient Hessian is $G_3''(u_3^*) = w_3 + O(\rho^2)$, $w_3 = (T-3)/T$; its off-diagonal correction starts at order ρ^2. Taylor's theorem and $F(u) \ge G_3(u_3)$ give $F(u) - V_T \ge (w_3/2)\varepsilon_3^2 - O(\rho^2\varepsilon_3^2) - O(|\varepsilon_3|^3)$. Inserting (218) yields $C_T = (w_3/2)[\delta_T/(ph)]^2$, exactly the coefficient in (181). The uniform distance from degeneracy makes it strictly positive. □

The lower-bound order is attainable by a *coefficient-space construction*, but this fact is not used as a training claim. Taking $\alpha = 0$, $\gamma = e = a_h = 1$, $\beta = u_2^*/u_1^*$, $r = (u_1^*)^2/u_2^*$, and $b_h = u_0^*/r - 1$ makes u_0, u_1, u_2 exact and $u_3 = (u_2^*)^2/u_1^*$. With $D_{\mathrm{blk}} \ge B \ge 6$, the remaining lag errors contribute only $O(\rho^{14})$, so the construction has $F - V_T = C_T\rho^{12} + O(\rho^{14})$. This identifies the obstruction's scale without asserting that GD reaches this particular parameter vector.

E.6.4 Partition, depth, and work comparisons

Proof of Theorem E.19. Lemma E.24 gives (179), since $w_4c_4^2/2 = A_T(p) > 0$. Lemma E.27 gives (181). For sufficiently small ρ, the always available $q = 1$ upper bound is strictly below the positive $q \ge 4$ expansion. Thus every minimizer has $q \le 3$. When $3 \mid B$ and the nondegeneracy condition holds, the $q = 3$ upper bound is strictly below the $q = 1$ lower bound, leaving $q \in \{2, 3\}$. The allowed values are divisors of B: $q = 3$ exists exactly when $3 \mid B$, and $q = 2$ exactly when $2 \mid B$. Enumerating those cases gives (182). When $3 \nmid B$, no comparison between the $q = 2$ and $q = 1$ rows is made; if B is coprime to six, $q = 1$ is the only remaining candidate. Uniformity over the finite divisor set follows by taking maxima of the finitely many constants. □

Lemma E.28 (Same-window paired-depth estimate for two or three roles)**.** *Let* $k = 3$, $q \in \{k, k-1\}$, $P_{\mathrm{e}} = B/q$, $L < L'$, *and* $D_{\mathrm{blk}} = BL \le T - 2$. *For every* N *in* (178)*,*

$$\mathcal{R}_{P_{\mathrm{e}},L}(N) - \mathcal{R}_{P_{\mathrm{e}},L'}(N) = \frac{w_{D_{\mathrm{blk}}+1}}{2} c_{D_{\mathrm{blk}}+1}^2 \rho^{2D_{\mathrm{blk}}+4} + O\Big(\rho^{2D_{\mathrm{blk}}+5}\ell^{m_\star} e^{C\sqrt{\ell}}\Big). \tag{219}$$

Proof. The common-parameter filter calculation in Lemmas E.9, E.10 and E.16 uses only the weighted tube and $D_{\mathrm{blk}} \ge 2k$, which holds because $B \ge 6$ and $L \ge 1$. In particular, the depth gradient forcing remains $O_\ell(\rho^{2D_{\mathrm{blk}}+2})$. The paired-depth proof controls the initial role-learning path.

For the longer continuation, stop both trajectories when either leaves a convex weighted tube chosen with a smaller fixed gate margin. Their joining parameter segments stay in that tube. The finite-filter coefficient and derivative bounds therefore apply on every such segment, uniformly in finite L'; no pointwise estimate is used outside its established region. After Lemma E.22, the represented high residual improves to $F_j = O_\ell(\rho^{j+3})$. In the exact Hessian decomposition (140), the low residual term is bounded using the active weighted inverse as in (139), and the high term uses $D^2 u_j = O_\ell(\rho^{j-1})$ for represented rows; omitted rows have $Du_j = D^2 u_j = 0$. The same bounds hold on a joining segment by continuity and the weighted tube estimates. This strengthens the negative-curvature bound on the continuation to

$$\lambda_{\min}(\nabla^2 R) \geq -C\|\nabla R\|/\rho - C\rho^{2q+4}\ell^{m_\star}. \tag{220}$$

For $q = 2$, the remaining time is $O(\rho^{-6}\ell)$ and its descent path is $O_\ell(\rho^3)$; therefore the integrated gradient-controlled term is $O_\ell(\rho^2)$ after division by ρ, while the residual term is $O_\ell(\rho^{2q+4-6}) = O_\ell(\rho^2)$. The $q = 3$ case is no larger. On joining segments, the gradient differs from its endpoint value by at most a bounded Hessian times the paired distance. Under the distance bootstrap $d \leq C\rho^{2D_{\mathrm{blk}}+2-2k}\ell^{m_\star}e^{C\sqrt{\ell}}$, its extra contribution to the integrated $\|\nabla R\|/\rho$ term is bounded by $C\rho^{-2k-1}\ell\, d = O\Big(\rho^{2D_{\mathrm{blk}}+1-4k}\ell^{m_\star+1}e^{C\sqrt{\ell}}\Big) = o(1)$, since $D_{\mathrm{blk}} \geq 2k$. Thus the segment estimate closes the distance bootstrap at a smaller fixed tube radius. Along with the initial bound in Lemma E.12, the total expansion factor is at most $e^{C\sqrt{\ell}}$. The nonlinear distance recurrence (149) with the same forcing gives

$$\sup_{n\leq N}\|\Theta_{L',n} - \Theta_{L,n}\| \leq C\rho^{2D_{\mathrm{blk}}+2-2k}\ell^{m_\star}e^{C\sqrt{\ell}}. \tag{221}$$

Its quadratic-feedback ratio is $\rho^{2D_{\mathrm{blk}}+1-4k}\ell^{m_\star}e^{C\sqrt{\ell}} = o(1)$ because $D_{\mathrm{blk}} \geq 2k$, closing the same bootstrap as Equation (153).

For every terminal N in the window, a preceding interval of physical duration at least $c\rho^{-6}\ell$ beginning after $N_-/2$ is available. On it, the $q = 2$ starting excess is $O_\ell(\rho^{12})$ and the $q = 3$ starting excess is $O_\ell(\rho^{14})$. Full-GD descent thus bounds $\sum \eta\|g_n\|^2$ by those respective orders. Along each actual GD step, $g_{n+1} = (I - \eta\widetilde{H}_n)g_n$ with $\widetilde{H}_n$ the segment-averaged Hessian. The positive Hessian part is below the fixed step ceiling; (220) and the just-established integrated negative part bound the amplification from any earlier gradient to g_N by a fixed factor. Averaging the energy over the preceding interval gives $\|g_N\| = O_\ell(\rho^9)$ for $q = 2$ and $O_\ell(\rho^{10})$ for $q = 3$; both are at most $C\rho^{2k+3}\ell^{m_\star}$ for $k = 3$. The same segment control ensures the Hessian block estimates (158) apply between the two terminal parameters. The fast readout direction gives $|r_{L',N} - r_{L,N}| \leq C\rho$ times (221). Since the non-head risk Hessian block is $O(\rho^2)$ and the head cross block is $O(\rho)$, Lemma E.15 places the retraining contribution at $O(\rho^{2D_{\mathrm{blk}}+5}\ell^{m_\star}e^{C\sqrt{\ell}})$. At a common parameter, the old-lag changes have this same smaller order and the newly accessible lag $D_{\mathrm{blk}}+1$ contributes $(w_{D_{\mathrm{blk}}+1}/2)c^2_{D_{\mathrm{blk}}+1}\rho^{2D_{\mathrm{blk}}+4}$ by Lemma E.16. Adding the two pieces proves (219). This argument compares two independently trained physical trajectories, not one parameter evaluated at two depths. □

Proof of Corollary E.20. At both depths Theorem E.19 confines the minimizing partitions to $\mathcal{P}_B = \{B/3\} \cup (\{B/2\} \text{ if } 2 \mid B)$. Every P_{e} in this finite set has $q \in \{3, 2\}$, so Lemma E.28 gives the same leading gap $G_{B,L,T}(p)\rho^{2BL+4}$ and a uniform remainder. If real arrays satisfy $x_{P_{\mathrm{e}}} = y_{P_{\mathrm{e}}} + a + e_{P_{\mathrm{e}}}$ with $|e_{P_{\mathrm{e}}}| \leq e$, then $a - e \leq \min_{P_{\mathrm{e}}} x_{P_{\mathrm{e}}} - \min_{P_{\mathrm{e}}} y_{P_{\mathrm{e}}} \leq a + e$; applying this identity to the two depths proves (184). Positivity follows from $BL \leq T - 2$, $p \in \mathcal{I}$, and $\rho\ell^{m_\star}e^{C\sqrt{\ell}} \to 0$. No minimizer identity across depths is assumed. □

Finally, for the work statement after Corollary E.20, $1 \leq P_{\mathrm{e}} \leq B$ and $1 \leq L \leq L_{\max}$ give $W/(P_{\mathrm{e}}BL) = BL_{\max}N_-/(P_{\mathrm{e}}L) \geq N_-$ and at most $BL_{\max}N_-$. Taking the floor leaves the lower endpoint unchanged because it is an integer; $\Gamma > BL_{\max} + 1$ places the upper endpoint below N_+. Thus every work stop is covered by Theorem E.19.

E.7 Path-selected head bias and the two-role risk coefficient

We now resolve the $q = 2$ candidate in the common third-stage window by tracking one raw shared-head coordinate along the actual optimizer trajectory. This is not a new optimizer or a coefficient-space gradient flow. Retain the HMM, physical network, exact paired reset, population source-mean

Euclidean GD, and the shared linear-history comparator V_T from Subsection E.1 and Theorem E.19. Fix even $B \geq 6$, set $P_{\mathrm{e}} = B/2$ and $q = 2$, take finite $T \geq B + 2$, p in a compact subset of $(0, 1)$, $0 < \eta \leq \eta_0 < 1$, and $\omega_0 = \omega\rho$ with $0 \leq \omega \leq c_h$ for the fixed reset constant c_h. For fixed $\Gamma > 1$, use the same third-stage window as Theorem E.19,

$$N_- = \left\lceil \frac{A_* \ell_\rho}{\eta \rho^6} \right\rceil, \qquad N_- \leq N \leq \lfloor \Gamma N_- \rfloor, \qquad \ell_\rho = \log(1/\rho), \tag{222}$$

with A_* sufficiently large. Constants below may depend on fixed B, T, the compact parameter sets, and Γ, but are uniform in finite L, N in this window, and the stated step sizes.

Define the path and target corrections

$$\sigma_{P_{\mathrm{e}}}(p, \eta, \omega) = \frac{p^2 - \omega^2}{2} - \frac{P_{\mathrm{e}}(1-p)^2}{6} - \frac{\eta(p-\omega)^2}{2(4-\eta)}, \tag{223}$$

$$\tau_T(p) = -\frac{2p^2}{3} + 2p\left[\frac{1-p}{(T-1)(T-3)} - \frac{1}{T(T-2)}\right], \tag{224}$$

$$\Xi_{P_{\mathrm{e}},T}(p, \eta, \omega) = \sigma_{P_{\mathrm{e}}}(p, \eta, \omega) - \tau_T(p). \tag{225}$$

Theorem E.29 (Two-role path selection and risk expansion). *There exists $\rho_0 > 0$, depending only on the fixed B, T, compact parameter sets, c_h, η_0, and Γ, such that for every $0 < \rho < \rho_0$, the full physical GD trajectory satisfies, uniformly for every N in Equation (222), every finite $L \geq 1$, and $0 < \eta \leq \eta_0$,*

$$b_{h,N} = \sigma_{P_{\mathrm{e}}}(p, \eta, \omega)\rho^2 + O(\rho^3 \ell_\rho^{m_\star}), \tag{226}$$

$$R_{B,P_{\mathrm{e}},L}(\Theta_N) - V_T = \frac{T-3}{2T} p^2(1-p)^6 \Xi_{P_{\mathrm{e}},T}(p, \eta, \omega)^2 \rho^{12} + O\big(|\Xi_{P_{\mathrm{e}},T}|\rho^{13}\ell_\rho^{m_\star} + \rho^{14}\ell_\rho^{m_\star}\big). \tag{227}$$

Here b_h is the trainable hidden bias in the paired shared head. In particular, if $|\Xi_{P_{\mathrm{e}},T}| \geq \xi_0 > 0$ on the parameter set, the two-role excess is bounded above and below by positive constant multiples of ρ^{12}. If $\Xi_{P_{\mathrm{e}},T} = 0$, the theorem gives only the upper bound $O(\rho^{14}\ell_\rho^{m_\star})$; it does not rank $q = 2$ against $q = 3$ on that degeneracy set. The finite exponent $m_\star$ is uniform in L, N, and η.

Corollary E.30 (Unique entrance count on a nondegenerate divisor family). *Suppose additionally that $6 \mid B$ and, uniformly on a compact parameter set,*

$$|2 - (T-1)p| \geq \delta_0 > 0, \qquad |\Xi_{B/2,T}(p, \eta, \omega)| \geq \xi_0 > 0. \tag{228}$$

Then, for all sufficiently small ρ, every finite $L \geq 1$, and every common-window update N,

$$\arg\min_{P_{\mathrm{e}}' \mid B} R_{B,P_{\mathrm{e}}',L}(\Theta_{P_{\mathrm{e}}',L,N}) = \{B/3\}. \tag{229}$$

An explicit nonempty region satisfying both exclusions, for each fixed $6 \mid B$ and $T \geq B + 2$, is $p \in [0.49, 0.51]$, $0 \leq \omega \leq \min\{c_h, 1/8\}$, and $0 < \eta \leq \min\{\eta_0, 1/16\}$. The constants are not asserted uniform as B or T grows. This is a common-update, fixed-B, T classification, not a permanent optimum over training time. If the $q = 2$ nondegeneracy fails but the full-entry condition $|2-(T-1)p| \geq \delta_0$ still holds, the risk table then yields only $\mathcal{P}^*_{B,L,N} \subseteq \{B/3, B/2\}$; if the full-entry condition also fails, this argument does not exclude $P_{\mathrm{e}} = B$.

Corollary E.31 (Depth envelope after entrance selection). *Under the assumptions of Corollary E.30, the minima over legal entrance counts in the original common third-stage window are attained at $P_{\mathrm{e}} = B/3$. Consequently the already established same-N depth comparison in Corollary E.20 applies to the selected $q = 3$ trajectory whenever its stated horizon and depth conditions hold. This corollary alone does not extend that result to the later ρ^{-8} window or to an arbitrary fixed-work comparison.*

The coefficient in Equation (227) is a training-selection effect, not a local expressivity obstruction. The next proposition records the relevant capacity statement separately.

Proposition E.32 (Nearby four-coefficient realization). *For $q = 2$ and any finite physical depth $D_{\mathrm{blk}} = BL \geq 6$, let $u_0^*, \ldots, u_3^*$ be the first four coefficients of the minimizer of the shared-history*

risk $F_{\rho,T}$. For sufficiently small ρ, a parameter point in the fixed-gate family has $u_j(\Theta) = u_j^*$ for $j = 0, 1, 2, 3$, with

$$r = O(\rho), \qquad \beta_1 = (1-p)\rho + O(\rho^3), \quad \beta_2 = (1-p)\rho + O(\rho^3), \quad b_h = \tau_T(p)\rho^2 + O(\rho^4). \tag{230}$$

The resulting coefficient-space excess is $O(\rho^{14})$ (with constants for fixed B, T). This is an existence statement for a nearby parameter point; it does not assert that GD reaches it, nor that the full network risk equals a Bayes or nonlinear-posterior risk.

E.8 Proof of the path-selected two-role risk

We prove Theorem E.29, Corollary E.30, and Proposition E.32 for the same physical network, source-mean population update, reset, and linear-history comparator as in Subsections E.1 and E.5. Throughout this section $q = 2$, $P_{\mathrm{e}} = B/2$, $B \geq 6$ is even, $D_{\mathrm{blk}} = BL$, and $\ell = \log(1/\rho)$. Constants can depend on fixed B, T, the compact p-set, the bounded initial-head ratio $\omega = \omega_0/\rho$, and the window multiplier, but not on finite L, the terminal update N, or $0 < \eta \leq \eta_0 < 1$. We use $O_\ell(\rho^a)$ for $O(\rho^a \ell^{m_\star})$ with a finite, depth-independent exponent $m_\star$. All assertions about parameters are on the invariant physical trajectory; no parameter is held fixed during GD. The preceding stage and profile interfaces are Lemmas E.3, E.21, E.22 and E.24. We give the additional path calculation in full.

E.8.1 Discrete startup from a possibly zero head

The first $O(\ell/\eta)$ steps determine a non-negligible term in the head-bias coefficient. A continuous-time approximation would omit it. Let $\boldsymbol{v}^0 = (1, -1, -1)$ and $t_1 = (\boldsymbol{v}^0)^{\mathsf{T}} \boldsymbol{\pi}_1$, as in (50). Let ν collect all independent raw coordinates other than r and the two temporal vectors, with ν^0 their reset values.

Lemma E.33 (Head startup and its squared increments)**.** *Fix $C_b > 4$ and put $n_b = \lceil C_b \ell/\eta \rceil$. For sufficiently small ρ, the full physical update satisfies, uniformly for $0 \leq n \leq n_b$,*

$$\begin{aligned} |r_n| &\leq C\rho, \\ \|\boldsymbol{\pi}_{j,n}\| &\leq C\rho^3 \ell^{m_\star} \quad (j = 1, 2), \\ \|\nu_n - \nu^0\| &\leq C\rho^2 \ell^{m_\star}, \end{aligned} \tag{231}$$

$$r_{n+1} = r_n - \frac{\eta}{2}(r_n - p\rho) + O(\eta\rho^3 \ell^{m_\star}), \tag{232}$$

$$r_n = p\rho + (\omega - p)\rho(1 - \eta/2)^n + O(\rho^3 \ell^{m_\star}). \tag{233}$$

At n_b, $R(\Theta_{n_b}) - V_T = O_\ell(\rho^4)$, and

$$\sum_{n<n_b} (\Delta r_n)^2 = \frac{\eta}{4-\eta}(p-\omega)^2\rho^2 + O_\ell(\rho^4). \tag{234}$$

The estimates include $\omega = 0$.

Proof. Stop at the first exit from the fixed startup tube $|r| \leq C_r\rho$, $\|\boldsymbol{\pi}_j\| \leq C_\pi \rho^3 \ell^2$, and $\|\nu - \nu^0\| \leq C_\nu \rho^2 \ell^2$ inside the fixed gate neighborhood. Choose the constants once, larger than their reset values; the powers ℓ^2 do not change during the bootstrap. The exact polynomial (29) has $Q_0 = 1$ and $Q_{j\geq 1} = 0$ at reset. Within the stopped set, degree counting in its current-state and temporal factors gives

$$Q_0 = 1 + O(\rho^2\ell^2), \quad Q_1 = O(\rho^3\ell^2), \quad u_0 = r + O(\rho^3\ell^2), \quad u_1 = O(\rho^4\ell^2). \tag{235}$$

For fixed T, all first derivatives of the finite filter are uniformly $O(1)$ in finite L on this fixed gate neighborhood, by contractivity of the remaining product as in Subsection E.1.3. In particular, Q_0 is independent of the temporal vectors. The CE/HMM expansion of Subsection E.2.1 then yields, throughout the stopped interval,

$$F_0 = r - p\rho + O(\rho^3\ell^{m_\star}), \qquad F_1 = O(\rho^2), \qquad \sum_{j\geq 2} |F_j| = O(\rho^3). \tag{236}$$

The last bound uses fixed T: any logarithm in a higher-lag filter coefficient is absorbed by its extra power of ρ. These are bounds at the current iterate, not just at reset. The uniform first-derivative bound, $r = O(\rho)$, and the raw mobility (26) now give the *coarse bounds without logarithms*

$$R_r = r - p\rho + O(\rho^3\ell^{m_\star}), \qquad \|R_{\boldsymbol{\pi}_j}\| = O(\rho^3), \qquad \|R_\nu\| = O(\rho^2). \tag{237}$$

The temporal estimate uses $\partial_{\boldsymbol{\pi}_j} Q_0 = 0$ and the factor $r = O(\rho)$ multiplying $|F_1| + \sum_{j\geq 2} |F_j| = O(\rho^2)$. The non-temporal estimate includes the trainable embedding, head input weight and hidden bias, neuron biases, and current-state block parameters. Their F_0 contribution has factor $r = O(\rho)$, while higher-lag contributions are no larger. Since B, T are fixed, the physical mobility changes these orders only by fixed constants.

The r equation is a stable scalar recurrence. Starting at $r_0 = \omega\rho$, it bounds $|r_n| \leq C\rho$ and gives (232)–(233). Summing the other two gradient bounds over $n_b\eta = O(\ell)$ gives $\|\boldsymbol{\pi}_j\| \leq C\rho^3\ell$ and $\|\nu - \nu^0\| \leq C\rho^2\ell$. These are strictly inside the chosen ℓ^2 boundaries when ρ is sufficiently small. The scalar recurrence similarly keeps $|r|$ strictly inside its fixed $C_r\rho$ boundary. Hence first exit cannot occur before n_b; (231) follows without assuming a positive initial readout or escalating any logarithmic exponent.

Write $\lambda = 1 - \eta/2$. Since $\lambda^{n_b} \leq \exp(-\eta n_b/2) \leq \rho^{C_b/2}$, (233) gives $u_0 - p\rho = O_\ell(\rho^3)$ at n_b. The target has $u_1^* = p(1-p)\rho^2 + O(\rho^4)$, whereas (235) gives $u_1 = O_\ell(\rho^4)$; all later target lags are smaller. Coefficient smoothness near u^* therefore gives $R(\Theta_{n_b}) - V_T = O_\ell(\rho^4)$. Finally, $\Delta r_n = \frac{\eta}{2}(p-\omega)\rho\lambda^n + O(\eta\rho^3\ell^{m_\star})$. The infinite leading geometric sum is

$$\frac{\eta^2(p-\omega)^2\rho^2}{4(1-\lambda^2)} = \frac{\eta(p-\omega)^2\rho^2}{4-\eta}.$$

Its tail after n_b is smaller than ρ^4. For the cross term use $\sum_{n<n_b} \eta\lambda^n \leq 2$; for squared errors use $n_b\eta = O(\ell)$ and $\eta \leq \eta_0$. Both are $O_\ell(\rho^4)$ uniformly even as $\eta \downarrow 0$. This proves (234). □

E.8.2 The physical head-bias balance

Write a_h, b_h for the paired head input weight and hidden bias, and put $C_0 = a_h Q_0 + b_h$. For $q = 2$, Q_0 has no temporal factor, so it is independent of t_1 at fixed other raw coordinates. The raw chain rule gives the following exact identities for the *finite* filter:

$$R_r = C_0 F_0 + a_h \sum_{j\geq 1} Q_j F_j, \qquad R_{b_h} = rF_0, \qquad R_{t_1} = ra_h \sum_{j\geq 1} Q_{j,t_1} F_j. \tag{238}$$

The raw head mobility is $1/2$. Since $t_1 = (\boldsymbol{v}^0)^\mathsf{T}\boldsymbol{\pi}_1$ and $\|\boldsymbol{v}^0\|^2 = 3$, its mobility is $3/(2P_\mathrm{e})$. Consequently $\Delta r = -\eta R_r/2$, $\Delta b_h = -\eta R_{b_h}/2$, and $\Delta t_1 = -3\eta R_{t_1}/(2P_\mathrm{e})$. Eliminating F_0 from (238) yields the exact discrete balance

$$\Delta b_h = \frac{r}{C_0}\Delta r - \frac{P_\mathrm{e}}{3C_0} t_1 \Delta t_1 + \frac{\eta r a_h}{2C_0} \sum_{j\geq 1} (Q_j - t_1 Q_{j,t_1}) F_j. \tag{239}$$

No endpoint, occurrence, or source-sum learning-rate replacement is implicit in this identity.

Lemma E.34 (Integrated two-role balance)**.** *Let $\bar{N}_2$ be the completed q=2 role-finishing checkpoint from Lemma E.22. Then*

$$b_{h,\bar{N}_2} = \left\{ \frac{p^2-\omega^2}{2} - \frac{P_\mathrm{e}(1-p)^2}{6} - \frac{\eta(p-\omega)^2}{2(4-\eta)} \right\} \rho^2 + O_\ell(\rho^3). \tag{240}$$

The same expansion holds at every N in (178).

Proof. Set $\mathcal{U}_j = Q_j - t_1 Q_{j,t_1}$. On the established two-role weighted tube, $\alpha_i = O_\ell(\rho^2)$, $\beta_i = O(\rho)$, $\gamma_i = 1 + O_\ell(\rho^2)$, and $r = O(\rho)$. Also $\beta_1 = s_1 t_1 + O_\ell(\rho^4)$ with $s_1 = (\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}^0)/3 = 1 + O_\ell(\rho^2)$; the orthogonal temporal component is controlled by the nuisance estimate in Lemma E.3. The lowest weighted-order term of Q_1 without a current-state factor is proportional to t_1; the corresponding term of Q_2 is proportional to $t_1\beta_2$. The Euler operator $t_1\partial_{t_1}$ cancels these leading terms. Other monomials of the same temporal degree can contain an α_i; they are retained in the remainder, along with non-temporal deviations and higher weighted-order terms. In the finite polynomial (29) this gives, uniformly over finite L,

$$|\mathcal{U}_1| \leq C\rho^3\ell^{m_\star}, \quad |\mathcal{U}_2| \leq C\rho^4\ell^{m_\star}, \quad |\mathcal{U}_j| \leq C\rho^j\ell^{m_\star} \quad (3 \leq j < T). \tag{241}$$

For uniformity, extract the fixed B physical factors first; each additional complete group is contractive, and differentiating at most T factors costs a bounded multiple of $\sup_{m\geq m_0} m^T\vartheta^{m-m_0}$ for some

fixed $\vartheta < 1$. Thus one logarithmic exponent suffices for all L. Rows $j > D_{\mathrm{blk}}$ vanish identically and are absent from the sum.

The cumulative remainder in (239) has the following orders; each entry includes the outside factor ηr and the duration.

	duration	residual bound	$\sum \eta r\, \mathcal{U}_j F_j$
stage 1, $j = 1$	$O(\rho^{-2}\ell/\eta)$	$\lvert F_1\rvert = O(\rho^2)$	$O_\ell(\rho^4)$
stage 1, $j \geq 2$	$O(\rho^{-2}\ell/\eta)$	$\lvert F_j\rvert = O_\ell(\rho^{j+1})$	$O_\ell(\rho^6)$
stage 2, $j = 1$	$O(P_{\mathrm{e}}\rho^{-4}\ell/\eta)$	$\lvert F_1\rvert = O_\ell(\rho^3)$	$O_\ell(\rho^3)$
stage 2, $j = 2$	$O(P_{\mathrm{e}}\rho^{-4}\ell/\eta)$	$\lvert F_2\rvert = O_\ell(\rho^3)$	$O_\ell(\rho^4)$
stage 2, $j \geq 3$	$O(P_{\mathrm{e}}\rho^{-4}\ell/\eta)$	$\lvert F_j\rvert = O_\ell(\rho^{j+1})$	$O_\ell(\rho^4)$.

(242)

The stage-one estimates follow from Subsection E.2.3; after its checkpoint the low conditional deficit and the stage-two sublevel give the displayed, deliberately coarse F_1, F_2 bounds. Thus the Euler contribution is $O_\ell(\rho^3)$, not a hidden ρ^2 term.

Throughout these two stages $C_0 = 1 + O_\ell(\rho^2)$. To replace C_0^{-1} by 1 in the first two terms of (239), use $r, t_1 = O(\rho)$ and the total parameter variation. Stage one has variation $O(\rho)$ by (37); stage two has $O_\ell(\rho\sqrt{\ell})$ by (93). The resulting error is $O_\ell(\rho^4)$. Summing (239) from reset, where $b_h = t_1 = 0$ and $r_0 = \omega\rho$, and retaining the exact discrete integration by parts gives

$$b_{h,\bar{N}_2} = \frac{r_{\bar{N}_2}^2 - r_0^2}{2} - \frac{P_{\mathrm{e}}}{6} t_{1,\bar{N}_2}^2 - \frac{1}{2}\sum_{n<\bar{N}_2} (\Delta r_n)^2 + \frac{P_{\mathrm{e}}}{6}\sum_{n<\bar{N}_2} (\Delta t_{1,n})^2 + O_\ell(\rho^3). \tag{243}$$

In particular, the sign of the r-square term is negative.

The startup lemma gives the leading r-square sum. After n_b, descent from its $O_\ell(\rho^4)$ excess implies $\sum_{n\geq n_b}(\Delta r_n)^2 = O(\eta\rho^4\ell^{m_\star})$ up to $\bar{N}_2$. For the temporal square sum, total stage-one descent energy alone would be too coarse. The coordinate estimate $\lvert R_{t_1}\rvert \leq C\rho^3\ell^{m_\star}$ on the first-stage tube, over $O(\rho^{-2}\ell/\eta)$ steps, gives $\sum_{\text{stage 1}}(\Delta t_1)^2 = O(\eta\rho^4\ell^{m_\star})$. Stage two begins with $O_\ell(\rho^6)$ excess by Lemma E.3, so its descent energy bounds the remaining temporal squares by $O(\eta\rho^6\ell^{m_\star})$. Finally, the refined role and head estimates of Subsection E.2.3 and Equation (192) give $r_{\bar{N}_2} = p\rho + O_\ell(\rho^3)$ and $t_{1,\bar{N}_2} = (1-p)\rho + O_\ell(\rho^3)$. Substitution into (243) proves (240).

The completed q=2 finishing excess is $O_\ell(\rho^{12})$. On the interval from $\bar{N}_2$ to any N in (178), the physical duration is $O(\rho^{-6}\ell)$; descent and Cauchy–Schwarz give total parameter path $O_\ell(\rho^3)$. The first-exit argument in Lemma E.24 keeps this path inside the weighted gate tube. Therefore b_h changes by only $O_\ell(\rho^3)$, proving the uniform-window assertion. □

E.8.3 The finite-filter ratio and the CE target

The path-selected bias matters because it changes one ratio among the first four history coefficients. Define the stable auxiliary filter $Q_\infty = C/(1-A)$ only for algebra; the trained network always uses the finite Q_L in (29). For $q = 2$ put

$$d_0 = 1 - \alpha_1\alpha_2, \qquad A_{\mathrm{net}} = \frac{\alpha_1\beta_2 + \alpha_2\beta_1}{d_0}, \qquad B_{\mathrm{net}} = \frac{\beta_1\beta_2}{d_0}. \tag{244}$$

Coefficient extraction from $Q_\infty = C/(1-A)$ gives

$$\begin{aligned} Q_{\infty,0} &= \frac{\gamma_1 + \gamma_2\alpha_1}{d_0}, & Q_{\infty,1} &= \frac{\gamma_2\beta_1}{d_0} + A_{\mathrm{net}}Q_{\infty,0}, \\ Q_{\infty,2} &= A_{\mathrm{net}}Q_{\infty,1} + B_{\mathrm{net}}Q_{\infty,0}, & Q_{\infty,3} &= A_{\mathrm{net}}Q_{\infty,2} + B_{\mathrm{net}}Q_{\infty,1}. \end{aligned} \tag{245}$$

Consequently

$$Q_{\infty,0}Q_{\infty,3} - Q_{\infty,1}Q_{\infty,2} = A_{\mathrm{net}}(Q_{\infty,0}Q_{\infty,2} - Q_{\infty,1}^2). \tag{246}$$

On the completed two-role tube, $\alpha_i = O_\ell(\rho^2)$, $\beta_i = (1-p)\rho + O_\ell(\rho^3)$, and $\gamma_i = 1 + O_\ell(\rho^2)$. Hence $A_{\mathrm{net}} = O_\ell(\rho^3)$, $Q_{\infty,0}Q_{\infty,2} - Q_{\infty,1}^2 = O_\ell(\rho^4)$, and $Q_{\infty,1}Q_{\infty,2} \asymp \rho^3$. The last $O_\ell(\rho^4)$ cancellation uses $\beta_2 - \beta_1 = O_\ell(\rho^3)$; treating the two roles as merely $O(\rho)$ would not suffice.

For finite depth,

$$Q_L - Q_\infty = A^{P_e L}(e - Q_\infty), \qquad A = (\alpha_1 + \beta_1 z)(\alpha_2 + \beta_2 z). \tag{247}$$

We record the required uniform estimate rather than using a fixed-D_{blk} bound with a depth-dependent logarithmic exponent. Write $A^{P_e L} = A^{B/2} A^{(P_e L - B/2)}$. In $A^{B/2}$, extraction of lag $j \le 3$ has order $O_\ell(\rho^{2B-j})$: among its B linear factors, each constant factor is $O_\ell(\rho^2)$ and each temporal factor is $O(\rho)$. In the remaining product, a fixed lag-$k \le 3$ coefficient is bounded by $C\rho^k m^k \vartheta^{m-k}$ for $m = P_e L - B/2$ and some $\vartheta < 1$; the bound is also valid at $m = 0$ with the usual empty-product convention. The supremum over $m \ge 0$ is finite. The constant coefficient of $e - Q_\infty$ is $O_\ell(\rho^2)$, and its positive lag-k coefficients are $O(\rho^k)$. Convolution therefore gives, with one fixed $m_\star$,

$$|Q_{L,j} - Q_{\infty,j}| \le C\rho^{2B-j+2}\ell^{m_\star}, \qquad 0 \le j \le 3, \quad L \ge 1. \tag{248}$$

This fixed-B extraction is essential: the apparent fixed-D_{blk} $O(\rho^{2D_{\mathrm{blk}}-j+2}\ell^{m_{\star,D_{\mathrm{blk}}}})$ does not itself give an L-uniform logarithmic exponent.

Combining (246) and (248), and using $a_h Q_{L,0} = 1 + O_\ell(\rho^2)$, yields for the physical history coefficients (30)

$$\frac{u_0 u_3}{u_1 u_2} = \left(1 + \frac{b_h}{a_h Q_{L,0}}\right) \frac{Q_{L,0} Q_{L,3}}{Q_{L,1} Q_{L,2}} = 1 + b_h + O_\ell(\rho^4). \tag{249}$$

All denominators are nonzero in the certified tube since $p, 1-p$ are bounded away from zero. This relation is local to the trained tube; it is not an expression lower bound for all two-role parameters.

For comparison, the unique shared-linear-history CE optimum has $u_i^* = c_i\rho^{i+1} + s_i\rho^{i+3} + O(\rho^{i+5})$, $c_i = p(1-p)^i$, by (210). Put $r_i^{\mathrm{pref}} = w_{i+1}/w_i = (T-i-1)/(T-i)$ and $\zeta_i = s_i/c_i$. The actual CE stationary equations (212), including its cubic $\tanh$ term, give

$$\begin{aligned} \zeta_0 &= \frac{p^2}{3} - p(1-p) r_0^{\mathrm{pref}}, \\ \zeta_1 &= p^2(1 + r_0^{\mathrm{pref}}) - p(1-p) r_1^{\mathrm{pref}}, \\ \zeta_i - \zeta_{i-1} &= p\big(r_{i-1}^{\mathrm{pref}} - (1-p) r_i^{\mathrm{pref}}\big), \qquad i \ge 2. \end{aligned} \tag{250}$$

Thus $\zeta_0 + \zeta_3 - \zeta_1 - \zeta_2$ equals

$$\tau_T(p) := -\frac{2p^2}{3} + 2p\left[\frac{1-p}{(T-1)(T-3)} - \frac{1}{T(T-2)}\right], \tag{251}$$

and the optimal ratio is

$$\frac{u_0^* u_3^*}{u_1^* u_2^*} = 1 + \tau_T(p)\rho^2 + O(\rho^4). \tag{252}$$

This calculation uses the actual finite-prefix CE weights; replacing them by equal lag weights or a square loss changes τ_T.

E.8.4 Conditional fit and the risk coefficient

For $m = 2$, write $\mathcal{E}_2 = F(u) - H_2(\Theta)$ as in (186). The q=2 finishing recurrence (195) has contraction scale $\eta\rho^4/P_e$ and forcing $O_\ell(\eta\rho^{18})$. Its floor is therefore $O_\ell(\rho^{14})$. The completed finishing checkpoint reaches that floor, and the same recurrence preserves it through the common window. Indeed the represented high residual begins at lag three with $F_3 = O_\ell(\rho^6)$, its physical Jacobian is $O_\ell(\rho^3)$, and later rows are smaller; omitted rows have identically zero physical derivative. Hence $\|\nabla H_2\| = O_\ell(\rho^9)$ as required by Lemma E.21. The weighted q=2 role tube and first-exit closure over the window are supplied by Lemma E.24. We obtain

$$\mathcal{E}_{2,N} = O_\ell(\rho^{14}), \qquad \|u_{0:2} - a_*(u_{3:T-1})\| = O_\ell(\rho^7). \tag{253}$$

The second conclusion follows from coefficient strong convexity, not from total-risk control alone.

The q=2 weighted jet (191), refined roles (192), and the CE coefficients imply $u_3 - u_3^* = O_\ell(\rho^6)$ and $u_j - u_j^* = O_\ell(\rho^{j+3})$ for every represented $j \ge 4$. Omitted rows $j > D_{\mathrm{blk}}$ have $u_j = 0$ and first occur after $D_{\mathrm{blk}} \ge B \ge 6$. The conditional-response bound (49) makes the high-row displacement

of the conditional optimum at low lag i at most $O_\ell(\rho^{9-i})$ for $i = 0, 1, 2$: the leading contribution comes from lag three with distance factor ρ^{3-i}. Together with (253),

$$|u_i - u_i^*| = O_\ell(\rho^7), \qquad i = 0, 1, 2. \tag{254}$$

In (249), solve for u_3 and insert (254), (240), and (252). Since $u_1^* u_2^*/u_0^* = c_3\rho^4 + O(\rho^6)$, where $c_3 = p(1-p)^3$, perturbations of the first three coefficients change this quantity by $O_\ell(\rho^7)$. Thus

$$u_3 - u_3^* = c_3\big(\sigma_{P_e}(p, \eta, \omega) - \tau_T(p)\big)\rho^6 + O_\ell(\rho^7), \tag{255}$$

where

$$\sigma_{P_e}(p, \eta, \omega) = \frac{p^2 - \omega^2}{2} - \frac{P_e(1-p)^2}{6} - \frac{\eta(p-\omega)^2}{2(4-\eta)}. \tag{256}$$

Proof of Theorem E.29. Let $\Xi_{P_e,T} = \sigma_{P_e} - \tau_T$. Taylor expansion of F about u^* is valid uniformly in the certified small-logit neighborhood. The lag-three Hessian entry is $w_3 + O(\rho^2)$, with $w_3 = (T-3)/T$; off-diagonal entries decay with lag distance by (47). Equation (254) makes the low-coordinate squares $O_\ell(\rho^{14})$. Represented $j \geq 4$ errors are $O_\ell(\rho^{j+3})$, while omitted $j > D_{\text{blk}}$ target coefficients contribute $O(\rho^{2D_{\text{blk}}+4})$. Cross terms between lag three and an adjacent $O_\ell(\rho^7)$ error acquire another factor $O(\rho)$ from the Hessian and are $O_\ell(\rho^{14})$. The cubic Taylor remainder is smaller. Inserting (255) gives

$$R_{B,B/2,L}(\Theta_N) - V_T = \frac{T-3}{2T}p^2(1-p)^6\Xi_{B/2,T}^2\rho^{12} + O\big(|\Xi_{B/2,T}|\rho^{13}\ell^{m_\star} + \rho^{14}\ell^{m_\star}\big). \tag{257}$$

The bound is uniform in the common window and all finite L. When $\Xi = 0$, it gives only the $O_\ell(\rho^{14})$ upper bound; it does not order q=2 against q=3 there. □

E.8.5 Unique partition and nearby physical capacity

Proof of Corollary E.30. Suppose $6 \mid B$. Every legal $P_e \mid B$ corresponds to a divisor $q = B/P_e$; in particular q=3 is legal. The common-window analysis gives its trained excess $O_\ell(\rho^{14})$ by (208). Every legal $q \geq 4$ has positive leading excess $A_T(p)\rho^{10} + O_\ell(\rho^{11})$ by (207). The q=1 excess has the positive $C_T(p)\rho^{12} - O_\ell(\rho^{13})$ lower bound (181) when $|2 - (T-1)p| \geq \delta_0 > 0$. Finally, (257) gives a positive ρ^{12} lower bound for q=2 when $|\Xi_{B/2,T}| \geq \xi_0 > 0$. Thus q=3 is strictly better than every other legal divisor for sufficiently small ρ, uniformly in N and finite L. This proves $\mathcal{P}^*_{B,L,N} = \{B/3\}$; it does not compare q=1 directly with q=2.

For completeness, the displayed nondegenerate region in the theorem has a uniform algebraic separation for each fixed B, T. Rearrange (223)–(224) as

$$\Xi_{P_e,T} = \frac{7p^2 - P_e(1-p)^2}{6} - \frac{\omega^2}{2} - \frac{\eta(p-\omega)^2}{2(4-\eta)} + \epsilon_T(p), \tag{258}$$
$$\epsilon_T(p) = 2p\left[\frac{1}{T(T-2)} - \frac{1-p}{(T-1)(T-3)}\right].$$

For $p \in [49/100, 51/100]$ and $T \geq 8$, $\epsilon_T(p) > 0$: after multiplying positive denominators, its bracket numerator is $pT^2 - (2+2p)T + 3 > 0$, increasing in $T \geq 8$ and positive already at $T = 8$. Also $\epsilon_T(p) \leq 2p/[T(T-2)]$. Here $P_e = B/2$ is a multiple of three. If $P_e = 3$ or 6, then

$$\Xi_{P_e,T} \geq \frac{7(0.49)^2 - 6(0.51)^2}{6} - \frac{1}{128} - \frac{(0.51)^2}{126} = \frac{3407}{336000} > \frac{1}{100}. \tag{259}$$

If $P_e \geq 9$, then $B \geq 18$, $T \geq 20$, and

$$\Xi_{P_e,T} \leq \frac{7(0.51)^2 - 9(0.49)^2}{6} + \frac{1.02}{20 \cdot 18} = -\frac{101}{1875} < -\frac{1}{100}. \tag{260}$$

The negative ω^2 and finite-step terms can only strengthen this upper bound. Moreover, $2/(T-1) \leq 2/7 < 0.49$, separating this region from the q=1 degeneracy. The interval proof uses no numerical search and asserts no uniformity as B, T grow. □

Proof of Corollary E.31. Corollary E.30 fixes the minimizing divisor to $P_e = B/3$ at each depth and at the same window update N. Substituting this configuration into the q=3 paired-depth estimate (219) gives the stated envelope comparison under its horizon and depth conditions. No parameter trajectory is carried from one depth to another. □

Proof of Proposition E.32. This is a local realization statement, not a statement about the GD trajectory. Let u^* be the coefficient optimum of the same $F_{\rho,T}$, and set

$$b_h^{\mathrm{req}} = \frac{u_0^* u_3^*}{u_1^* u_2^*} - 1, \qquad r = \frac{u_0^*}{1 + b_h^{\mathrm{req}}}, \qquad \beta_1 = \frac{u_1^*}{r}, \qquad \beta_2 = \frac{u_2^*}{u_1^*}. \tag{261}$$

Keep the original paired reset values of every raw block coordinate except its temporal vector, and put $\boldsymbol{\pi}_j = \beta_j \boldsymbol{v}^0/3$ for $j = 1, 2$; repeat these two roles according to the q-periodic physical embedding. Keep $e = a_h = 1$ and set the original paired-head coordinates to $(r, a_h, b_h^{\mathrm{req}})$. The reset block coordinates give $\alpha_1 = \alpha_2 = 0$ and $\gamma_1 = \gamma_2 = 1$ exactly, while $\boldsymbol{v}_j^{\mathsf{T}} \boldsymbol{\pi}_j = \beta_j$. As $D_{\mathrm{blk}} = BL = 2P_e L \geq 6$, the initial-product term in (29) first appears at lag D_{blk}. Its first four coefficients are therefore

$$Q_0 = 1, \qquad Q_1 = \beta_1, \qquad Q_2 = \beta_1\beta_2, \qquad Q_3 = \beta_1^2\beta_2. \tag{262}$$

Equations (30) and (261) now give $u_i = u_i^*$ exactly for $0 \leq i \leq 3$. By (252), $b_h^{\mathrm{req}} = \tau_T(p)\rho^2 + O(\rho^4)$, while $r = p\rho + O(\rho^3)$ and $\beta_j = (1-p)\rho + O(\rho^3)$. The chosen physical parameters lie in the same fixed gate neighborhood for small ρ. For represented $j \geq 4$, their leading temporal product is $p(1-p)^j\rho^{j+1}$, so $u_j - u_j^* = O(\rho^{j+3})$; omitted $j > D_{\mathrm{blk}}$ contribute only their $O(\rho^{j+1})$ target coefficients. Coefficient smoothness and lag decay give $F(u) - V_T = O(\rho^{14})$. Thus the q=2 ρ^{12} trained term in (257) is not forced by a nearby local capacity obstruction. Nothing here asserts that GD reaches this constructed parameter in the common window. □

E.9 Late calibration of the two-role predictor

The third-stage comparison in Subsection E.7 is a comparison at a specified training time, not a permanent ranking. At that time the two-role predictor has learned the leading temporal coefficients, but its shared head has selected a slightly wrong ratio among the first four history coefficients. We now follow the original full-parameter GD after that checkpoint. The target and metric remain the same: population next-token cross-entropy averaged over prefixes and sources, measured against the best shared linear-history predictor V_T, not Bayes risk.

Fix even $B \geq 6$, $P_e = B/2$, $q = 2$, finite $T \geq B + 2$, and any finite loop depth $L \geq 1$. Use the paired reset, all $25B + 9$ trainable physical coordinates, and source-mean Euclidean GD of Subsection E.1 and Assumption E.1. Let $p = \kappa^2$ range in a compact subset of $(0, 1)$, $\omega_0 = \omega\rho$ with $0 \leq \omega \leq c_h$, and $0 < \eta \leq \eta_0$ for a sufficiently small fixed η_0. Choose in advance an update N_0 in the certified common third-stage window (222); no parameters are reset at N_0. Retain $\sigma_{P_e}, \tau_T, \Xi_{P_e,T}$ from (223)–(225), and set

$$\lambda_{T,p} = \frac{T-3}{2T} p^2 (1-p)^6 > 0, \qquad s_n = \eta\rho^8 n, \qquad \ell_\rho = \log(1/\rho). \tag{263}$$

Theorem E.35 (Late full-GD calibration curve)**.** *For every fixed $S > 0$, sufficiently small ρ and every integer $n \geq 0$ with $s_n \leq S\ell_\rho$, the actual physical trajectory remains in the certified gate region and obeys*

$$b_{h,N_0+n} = \rho^2 \left[\tau_T + \Xi_{P_e,T} e^{-\lambda_{T,p} s_n}\right] + O(\rho^3 \ell_\rho^{m_\star}), \tag{264}$$

$$R_{B,P_e,L}(\Theta_{N_0+n}) - V_T = \lambda_{T,p} \Xi_{P_e,T}^2 \rho^{12} e^{-2\lambda_{T,p} s_n} + O\big(|\Xi_{P_e,T}| \rho^{13} \ell_\rho^{m_\star} e^{-\lambda_{T,p} s_n} + \rho^{14} \ell_\rho^{m_\star}\big). \tag{265}$$

Here $m_\star$ is finite and independent of L, n, η; constants may depend on fixed B, T, S and the compact parameter sets. The leading rate has no explicit $1/P_e$: this late direction is the shared-head bias, whose raw mobility is $1/2$. Entrance count still enters the starting amplitude $\Xi_{P_e,T}$.

Corollary E.36 (Calibration time)**.** *If $|\Xi_{P_e,T}| \geq \xi_0 > 0$ uniformly, then for fixed $0 < \theta < 1$, the first update n_θ with excess risk at most θ times its value at N_0 satisfies*

$$\eta\rho^8 n_\theta = \frac{\log(1/\theta)}{2\lambda_{T,p}} + O(\rho\ell_\rho^{m_\star} + \eta\rho^8). \tag{266}$$

For any parameters, an additional scaled time $s_n \geq \ell_\rho/\lambda_{T,p} + O(1)$ within the theorem window is sufficient for $O(\rho^{14}\ell_\rho^{m_\star})$ excess. When $|\Xi_{P_e,T}|$ is separated from zero, reaching that order also requires $s_n \geq \ell_\rho/\lambda_{T,p} - O(\log \ell_\rho)$, with the logarithmic correction depending on the allowed polylogarithmic target. There is no such necessity on the degeneracy set $\Xi_{P_e,T} = 0$.

Corollary E.37 (Persistence of the selected entrance count)**.** *Suppose $6 \mid B$ and the two margins in (228) hold uniformly. Take $N_0 = N_-$ in (222). There is a fixed $s_* > 0$ such that, for all sufficiently small ρ, every finite L, and every common update*

$$N_- \leq N \leq N_- + \left\lfloor \frac{s_*}{\eta\rho^8} \right\rfloor, \qquad \arg\min_{P_e' \mid B} R_{B,P_e',L}(\Theta_{P_e',L,N}) = \{B/3\}. \tag{267}$$

This extends the classification for a bounded late interval, not for all future training times.

Corollary E.38 (A specified equal-work depth reversal)**.** *Keep $P_e = B/2$, $|\Xi_{P_e,T}| \geq \xi_0$, a finite $L_{\max}$, and fixed $w > 0$. Count all training, including burn-in, with the block-call proxy $W = P_eBLN$. Give each separately trained depth the same budget $W = P_eBw/(\eta\rho^8)$ and stop it at $N_L = \lfloor W/(P_eBL) \rfloor$. Uniformly for $1 \leq L \leq L_{\max}$,*

$$R_{B,P_e,L}(\Theta_{L,N_L}) - V_T = \lambda_{T,p}\Xi_{P_e,T}^2\rho^{12}e^{-2\lambda_{T,p}w/L} + O(\rho^{13}\ell_\rho^{m_\star}). \tag{268}$$

Consequently, for any fixed $L < L' \leq L_{\max}$, the deeper run has strictly larger risk for sufficiently small ρ. This fixed-budget comparison does not reverse the earlier same-update depth theorem.

The mechanism is a separation of learning rates within one physical network. Independent temporal roles carry the replica-dependent mobility $1/P_e$; after they have been learned, a residual fourth-coefficient error is repaired predominantly through a shared head coordinate. The curve proves later reachability of the $O(\rho^{14}\ell_\rho^{m_\star})$ scale, not convergence exactly to V_T. It neither orders different depths at the same late update nor extends to arbitrary P_e, finite data, noisy optimizers, or nonlinear posterior learning. The proof, including the full-gradient compensation, is in Subsection E.10.

E.10 Proof of late calibration and its consequences

Throughout, $q = 2$, $P_e = B/2$, $D_{\mathrm{blk}} = BL$, $h = 1 - p$, $c_j = ph^j$, and $w_j = (T - j)/T$. We write $O_\ell(\rho^a)$ for $O(\rho^a\ell_\rho^{m_\star})$, allowing a larger but fixed exponent between displays. Constants are uniform in finite L, the permitted time and $0 < \eta \leq \eta_0$, but not in growing B, T. All calculations use the invariant trajectory of the original $25B + 9$-parameter update. In particular, the raw mobilities below follow from (26); they are not an alternative optimizer.

E.10.1 A stopped trajectory with refined role scales

At N_0, Theorem E.29 and Lemma E.24 gives

$$E_0 := R(\Theta_{N_0}) - V_T \leq C\rho^{12}, \quad b_h = \sigma_{P_e}\rho^2 + O_\ell(\rho^3), \quad r = p\rho + O_\ell(\rho^3), \quad t_i = h\rho + O_\ell(\rho^3). \tag{269}$$

Here $t_i = (\boldsymbol{v}^0)^{\mathsf{T}}\boldsymbol{\pi}_i$ and $\boldsymbol{v}^0 = (1, -1, -1)$; the other raw coordinates differ from reset by $O_\ell(\rho^2)$. The first bound has no leading logarithm because the q=2 risk expansion is uniform and $\Xi_{P_e,T}$ is bounded.

Choose a fixed closed, convex gate neighborhood strictly inside that of Subsection E.1.3. Stop at its first exit or at $s_n = S\ell_\rho$, whichever comes first. The exact contractive filter has uniformly bounded first and second physical derivatives in this neighborhood, including on update segments. Thus, for a common small η_0, ordinary descent along every stopped step gives

$$E_{j+1} \leq E_j - \frac{\eta}{2}\|\nabla_\Theta R_j\|^2, \qquad \sum_{j<n} \eta\|\nabla_\Theta R_j\|^2 \leq 2E_0. \tag{270}$$

As $n\eta = s_n\rho^{-8}$, Cauchy–Schwarz in physical Euclidean coordinates yields

$$\sum_{j<n} \|\Theta_{j+1} - \Theta_j\| \leq (n\eta)^{1/2}\left(\sum_{j<n} \eta\|\nabla_\Theta R_j\|^2\right)^{1/2} \leq C\rho^2\sqrt{s_n}. \tag{271}$$

This only maintains the broad $O_\ell(\rho^2)$ nuisance bounds; it does not alone maintain the $O_\ell(\rho^3)$ effective-role errors.

To recover them at every stopped prefix, coefficient strong convexity near the optimum u^* gives $\|u - u^*\| \le C\rho^6$ from (270). The exact first three filter rows, using the broad nuisance bounds, are

$$C_0 := a_h Q_0 + b_h = 1 + O_\ell(\rho^2), \quad Q_1 = \beta_1 + O_\ell(\rho^3), \quad Q_2 = \beta_1\beta_2 + O_\ell(\rho^4), \qquad (272)$$
$$u_0^* = p\rho + O(\rho^3), \quad u_1^* = ph\rho^2 + O(\rho^4), \quad u_2^* = ph^2\rho^3 + O(\rho^5).$$

The first row gives $r = p\rho + O_\ell(\rho^3)$; division by r in the second gives $\beta_1 = h\rho + O_\ell(\rho^3)$, and division by $r\beta_1$ in the third gives $\beta_2 = h\rho + O_\ell(\rho^3)$. Both divisors are separated from zero by compactness of p. To return to raw temporal coordinates, write $\boldsymbol{\pi}_i = t_i \boldsymbol{v}^0/3 + \boldsymbol{\pi}_{i,\perp}$. On the broad tube, $\boldsymbol{v}_i - \boldsymbol{v}^0 = O_\ell(\rho^2)$ and $\boldsymbol{\pi}_{i,\perp} = O_\ell(\rho^2)$, so

$$\beta_i = s_i t_i + \boldsymbol{v}_i^\mathsf{T} \boldsymbol{\pi}_{i,\perp}, \qquad s_i = 1 + O_\ell(\rho^2), \qquad \boldsymbol{v}_i^\mathsf{T} \boldsymbol{\pi}_{i,\perp} = O_\ell(\rho^4). \qquad (273)$$

It follows that $t_i = h\rho + O_\ell(\rho^3)$ on every stopped prefix. First fix the outer gate margin and S, then the broad-tube constants, and finally ρ_0. Equations (271)–(273) place every stopped iterate strictly inside those boundaries. First exit is therefore impossible before $s_n = S\ell_\rho$. This order of argument avoids assuming the fine role estimates that the long-time proof needs to establish.

E.10.2 The finite-depth fourth direction

Fix nuisance coordinates temporarily only to calculate a partial Jacobian; they continue to update in GD. Put $f = (r, t_1, t_2)$, $a_{0:2} = (u_0, u_1, u_2)$, and

$$A_{\mathrm{lo}} = \partial_f a_{0:2}, \qquad v_b = \partial_{b_h} a_{0:2} = (r, 0, 0)^\mathsf{T}. \qquad (274)$$

Since Q_0 has no temporal factor and $\partial_{t_i}\beta_i = s_i$, the actual finite-$L$ matrix is

$$A_{\mathrm{lo}} = \begin{pmatrix} C_0 & 0 & 0 \\ a_h Q_1 & r a_h s_1 Q_{1,\beta_1} & r a_h s_2 Q_{1,\beta_2} \\ a_h Q_2 & r a_h s_1 Q_{2,\beta_1} & r a_h s_2 Q_{2,\beta_2} \end{pmatrix}. \qquad (275)$$

Its determinant is nonzero on the refined tube. Holding the first three outputs fixed while differentiating in b_h gives $r_b = -r/C_0$ and

$$(\beta_{1,b}, \beta_{2,b})^\mathsf{T} = C_0^{-1} J_\beta^{-1} (Q_1, Q_2)^\mathsf{T}, \qquad J_\beta = \partial_\beta (Q_1, Q_2). \qquad (276)$$

Consequently the exact fourth pivot for the finite filter is

$$S_{D_{\mathrm{blk}}} := \partial_{b_h} u_3 \big|_{u_{0:2}} = \frac{r a_h}{C_0} \left[\nabla_\beta Q_3 J_\beta^{-1} (Q_1, Q_2)^\mathsf{T} - Q_3 \right]. \qquad (277)$$

Here is the relative, rather than merely absolute, pivot estimate. Set $\beta_i = \rho b_i$. Uniformly on the stopped tube, $b_i = h + O_\ell(\rho^2)$, and the first four finite-filter rows obey

$$Q_0 = 1 + O_\ell(\rho^2), \quad \rho^{-1} Q_1 = b_1 + O_\ell(\rho^2), \quad \rho^{-2} Q_2 = b_1 b_2 + O_\ell(\rho^2), \quad \rho^{-3} Q_3 = b_1^2 b_2 + O_\ell(\rho^2), \qquad (278)$$

with the same bounds after one or two derivatives in b_i. In the leading rows, $\partial_{(b_1,b_2)}(b_1, b_1 b_2)$ equals $\left(\begin{smallmatrix} 1 & 0 \\ b_2 & b_1 \end{smallmatrix}\right)$. Its inverse is uniformly bounded, and direct substitution in (277) gives

$$\nabla (b_1^2 b_2) \begin{pmatrix} 1 & 0 \\ b_2 & b_1 \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_1 b_2 \end{pmatrix} - b_1^2 b_2 = b_1^2 b_2. \qquad (279)$$

As $r = p\rho + O_\ell(\rho^3)$,

$$S_{D_{\mathrm{blk}}} = c_3 \rho^4 [1 + O_\ell(\rho^2)], \qquad c_3 = p(1-p)^3 > 0. \qquad (280)$$

For completeness, the finite-depth control in (278) is not a fixed-depth estimate with an uncontrolled logarithmic exponent. Use the exact identity $Q_L - Q_\infty = A^{P_e L}(e - Q_\infty)$ from (247). For row j extract the first $D_j^0 = \max\{B, j\}$ physical factors. Among their lag-k terms, each temporal factor costs $O(\rho)$ and each constant factor costs $O_\ell(\rho^2)$. For the remaining m contractive factors, coefficient

extraction at fixed lag $k \le T-1$ is bounded by $C_T m^T \rho^k \vartheta^{m-k}$ for a fixed $\vartheta < 1$. Its supremum over $m \ge 0$ is finite. A raw temporal derivative may lose one power of ρ; a derivative in the scaled b_i does not. Convolution with $e - Q_\infty$, whose constant row is $O_\ell(\rho^2)$, yields for represented j the uniform bounds

$$|u_{L,j} - u_{\infty,j}| \le C\rho^{2D_j^0 - j + 3}\ell_\rho^{m_\star}, \qquad |\partial_{r,t_1,t_2}(u_{L,j} - u_{\infty,j})| \le C\rho^{2D_j^0 - j + 2}\ell_\rho^{m_\star}. \tag{281}$$

For $j \le 3$ and $B \ge 6$, these errors and their scaled first two derivatives are smaller than the relative $O_\ell(\rho^2)$ errors in (278). In particular they cannot cancel the ρ^4 pivot.

For represented $j \ge 4$ set $h_j = \partial_f u_j$ and $S_j = -h_j A_{\mathrm{lo}}^{-1} v_b$; for $j = 3$ this definition agrees with (277). Weighted differentiation of the finite filter and triangular inversion of (275) give

$$\|h_j\| = O_\ell(\rho^j), \quad \|A_{\mathrm{lo}}^{-1} v_b\| = O_\ell(\rho), \quad |S_j| = O_\ell(\rho^{j+1}), \quad \|h_j A_{\mathrm{lo}}^{-1}\| = O_\ell(\rho^{j-2}). \tag{282}$$

In the third entry, the $j = 3$ case has the sharper signed value (280). With J_4 the four-row Jacobian in columns (f, b_h), block inversion gives

$$[h_j, 0]J_4^{-1} = \left[h_j A_{\mathrm{lo}}^{-1} - \frac{S_j}{S_{D_{\mathrm{blk}}}} h_3 A_{\mathrm{lo}}^{-1}, \frac{S_j}{S_{D_{\mathrm{blk}}}}\right]. \tag{283}$$

Thus $\|J_4^{-1}\| \le C\rho^{-4}$ and the high-row inverse ratio is $O_\ell(\rho^{j-3})$, beginning at $O_\ell(\rho)$ for $j = 4$. This is the fourth-direction rank estimate for the actual finite-depth network.

E.10.3 Profile accuracy under the full physical update

Use the conditional-risk notation of (186) with cutoff $m = 3$. The stopped refined tube and (191) imply, for represented $4 \le j \le \min(D_{\mathrm{blk}}, T-1)$,

$$u_j - u_j^* = O_\ell(\rho^{j+3}), \qquad F_j(a_*(b), b) = O_\ell(\rho^{j+3}). \tag{284}$$

Indeed the low-row conditional displacement contributes at most the same order by the lag-decaying coefficient Hessian (47). For omitted rows $j > D_{\mathrm{blk}}$, $u_j \equiv 0$ and $D_\Theta u_j \equiv 0$. Their residuals can be larger, but their coupling back to a represented row $j \le D_{\mathrm{blk}}$ is at most $O(\rho^{2D_{\mathrm{blk}}+3-j})$, which is no larger than the right side of (284). Differentiating all represented rows therefore gives

$$\|\nabla H_3\| \le C \sum_{j=4}^{\min(D_{\mathrm{blk}}, T-1)} \rho^j \rho^{j+3} \ell_\rho^{m_\star} = O_\ell(\rho^{11}). \tag{285}$$

The effective fourth pivot (280), high-row inverse (283), coefficient cross-Hessian $O(\rho)$, and the true raw mobility $(1/2, 3/(2P_{\mathrm{e}}), 3/(2P_{\mathrm{e}}), 1/2)$ meet (190) with $s = 4$. The physical profile identity and descent lemma Lemma E.21 consequently give

$$\mathcal{E}_3^+ \le (1 - c\eta\rho^8/P_{\mathrm{e}})\mathcal{E}_3 + C\eta\rho^{22}\ell_\rho^{m_\star}. \tag{286}$$

The initial $\mathcal{E}_3 \le E_0 = O(\rho^{12})$ contracts to its $O_\ell(\rho^{14})$ floor in $O(P_{\mathrm{e}}\ell_\rho/(\eta\rho^8))$ steps. This supports late reachability, but the conservative profile rate is not the leading calibration rate below. Separately, taking $u_{0:3}^*$ as a feasible profile comparator and retaining omitted rows yields

$$H_3 - V_T \le C \sum_{j=4}^{\min(D_{\mathrm{blk}}, T-1)} \rho^{2j+6}\ell_\rho^{m_\star} + C \sum_{j=D_{\mathrm{blk}}+1}^{T-1} \rho^{2j+2} = O_\ell(\rho^{14}), \tag{287}$$

where an empty sum is zero and $D_{\mathrm{blk}} \ge B \ge 6$. Thus a small profile gradient has not been mistaken for a small total risk.

We also need continued precision of the first three outputs. The cutoff-two profile recurrence (195) remains valid on this stopped tube: the leading high residual is $F_3 = O_\ell(\rho^6)$, its physical derivative is $O_\ell(\rho^3)$, and hence

$$\mathcal{E}_2^+ \le (1 - c\eta\rho^4/P_{\mathrm{e}})\mathcal{E}_2 + C\eta\rho^{18}\ell_\rho^{m_\star}. \tag{288}$$

The established starting bound is $\mathcal{E}_2 = O_\ell(\rho^{14})$, so the same bound holds at every late step. Strong convexity gives distance $O_\ell(\rho^7)$ from the conditional low-three optimum. The high-to-low conditional response (49) adds at most $O_\ell(\rho^{9-i})$ at lag $i = 0, 1, 2$ from the lag-three $O_\ell(\rho^6)$ displacement; later rows are smaller. Hence

$$|u_i - u_i^*| = O_\ell(\rho^7), \qquad i = 0, 1, 2. \tag{289}$$

E.10.4 EXACT ELIMINATION AND THE SCALAR LATE MODE

Let $F_j = \partial_{u_j} F_{\rho,T}$ and $D_f = \mathrm{diag}(1/2, 3/(2P_{\mathrm{e}}), 3/(2P_{\mathrm{e}}))$. The exact chain rule, with every high represented row retained, is

$$R_{b_h} = v_b^{\mathsf{T}} A_{\mathrm{lo}}^{-\mathsf{T}} R_f + \sum_{j=3}^{\min(D_{\mathrm{blk}},T-1)} S_j F_j. \tag{290}$$

Since $\Delta f = -\eta D_f R_f$ and $\Delta b_h = -\eta R_{b_h}/2$, the actual updates satisfy

$$\Delta b_h = B_n \Delta f - \frac{\eta}{2} \sum_{j=3}^{\min(D_{\mathrm{blk}},T-1)} S_j F_j, \qquad B_n = \tfrac{1}{2} v_b^{\mathsf{T}} A_{\mathrm{lo}}^{-\mathsf{T}} D_f^{-1}. \tag{291}$$

There is no instantaneous conditional fitting in this identity. Triangular inversion gives $\|B_n\| \le C_{B,T}\rho$; in the leading two-role filter, $A_{\mathrm{lo}}^{-1} v_b$ in (r, β_1, β_2) coordinates is $(r/C_0, -\beta_1/C_0, 0)^{\mathsf{T}}$. Using the full path bound (271),

$$\sum_{j<n} |B_j \Delta f_j| \le C\rho \sum_{j<n} \|\Delta\Theta_j\| \le C\rho^3 \sqrt{s_n} = O_\ell(\rho^3). \tag{292}$$

The low-coordinate compensation is retained exactly and then bounded in total, not discarded step by step.

The finite two-role recurrence (246), its finite-depth correction (281), and the target ratio (252) give on the refined tube

$$\frac{u_0 u_3}{u_1 u_2} = 1 + b_h + O_\ell(\rho^4), \qquad \frac{u_0^* u_3^*}{u_1^* u_2^*} = 1 + \tau_T \rho^2 + O(\rho^4). \tag{293}$$

Equation (289) makes the propagated error in $u_1 u_2/u_0$ at most $O_\ell(\rho^8)$: the largest term is an $O_\ell(\rho^7)$ error in u_1 multiplied by $u_2^*/u_0^* = O(\rho^2)$, while the u_2 error is multiplied by $u_1^*/u_0^* = O(\rho)$. Thus

$$u_3 - u_3^* = c_3 \rho^4 (b_h - \tau_T \rho^2) + O_\ell(\rho^8). \tag{294}$$

The Hessian expansion about u^*, lag decay, and (289) then give

$$F_3 = w_3 c_3 \rho^4 (b_h - \tau_T \rho^2) + O_\ell(\rho^8), \qquad w_3 = (T-3)/T. \tag{295}$$

For $j \ge 4$, (282) and the represented-row residual bound imply a total $O_\ell(\rho^{12})$ contribution to the sum in (291); omitted rows have zero physical derivative. Combining (280) and (295) yields

$$\sum_{j=3}^{\min(D_{\mathrm{blk}},T-1)} S_j F_j = w_3 c_3^2 \rho^8 (b_h - \tau_T \rho^2) + O_\ell(\rho^{12}). \tag{296}$$

Since $w_3 c_3^2/2 = \lambda_{T,p}$, this is the actual scalar recursion

$$b_{h,n+1} - \tau_T \rho^2 = (1 - \eta \lambda_{T,p} \rho^8)(b_{h,n} - \tau_T \rho^2) + B_n \Delta f_n + O_\ell(\eta \rho^{12}). \tag{297}$$

The role mobility $1/P_{\mathrm{e}}$ occurs in D_f and the controlled compensation, while the leading coefficient is the shared-head mobility $1/2$.

Sum (297) from N_0. By (292), the total compensated forcing is $O_\ell(\rho^3)$, and the other forcing totals $O_\ell(\eta n \rho^{12}) = O_\ell(\rho^4)$ on $s_n \le S\ell_\rho$. The initial value is $\Xi_{P_{\mathrm{e}},T}\rho^2 + O_\ell(\rho^3)$ by (269). Therefore

$$b_{h,N_0+n} - \tau_T \rho^2 = \Xi_{P_{\mathrm{e}},T} \rho^2 (1 - \eta \lambda_{T,p} \rho^8)^n + O_\ell(\rho^3). \tag{298}$$

Uniformly on this window, the geometric factor equals $e^{-\lambda_{T,p} s_n}[1 + O(\eta \rho^8 \ell_\rho)]$, proving (264). The original finite-step startup correction remains inside $\Xi_{P_{\mathrm{e}},T}$.

For risk, Taylor-expand F around u^* using (294), (289), represented high-row errors (284), and the omitted tail. Off-diagonal Hessian entries contribute one extra power of ρ per lag distance. The result is

$$R(\Theta_{N_0+n}) - V_T = \lambda_{T,p} \rho^8 (b_{h,N_0+n} - \tau_T \rho^2)^2 + O_\ell(\rho^{14}). \tag{299}$$

Substitution of (298) gives exactly (265), including its exponentially decaying cross error. This completes the proof of Theorem E.35.

E.10.5 Timing, partition persistence, and work

Proof of Corollary E.36. When $|\Xi_{P_e,T}| \geq \xi_0$, divide (265) by its $n = 0$ value. On each bounded scaled-time interval, $E_n/E_0 = e^{-2\lambda_{T,p}s_n} + O(\rho\ell_\rho^{m_\star})$. The limiting curve crosses a fixed $\theta \in (0,1)$ with nonzero slope. Actual risk is nonincreasing by (270); bracketing the first crossing at neighboring grid values of spacing $\eta\rho^8$ proves (266). At $s_n \geq \ell_\rho/\lambda_{T,p} + O(1)$, the main term and cross error in (265) are at most $O_\ell(\rho^{14})$. Conversely, if $E_n \leq C\rho^{14}\ell_\rho^{m_0}$ with fixed C, m_0 and separated Ξ, the leading term must be at most this order up to the lower-order cross and floor. The corresponding bound on $e^{-2\lambda_{T,p}s_n}$ gives $s_n \geq \ell_\rho/\lambda_{T,p} - O(\log \ell_\rho)$. This last implication is conditional on the separated initial mismatch. □

Proof of Corollary E.37. There are only finitely many legal divisors of fixed B. At N_-, q=3 has $O_\ell(\rho^{14})$ excess. Its late path is $O_\ell(\rho^3)$ over any fixed scaled interval by descent and Cauchy–Schwarz, so the gate tube remains valid and risk descent preserves that upper bound. For q=2, (265) and $|\Xi| \geq \xi_0$ give a positive $c\rho^{12}$ lower bound on every bounded interval $0 \leq s \leq s_*$. For q=1, its starting $O_\ell(\rho^{12})$ sublevel gives an $O_\ell(\rho^2)$ path on a fixed scaled interval. Coefficient strong convexity and the first two triangular filter rows recover the fine readout and effective temporal scales on each stopped prefix, as in (272); the broad path controls the remaining raw coordinates. Thus the weighted one-role neighborhood persists. The finite-depth geometric-tail minor of (181), with the separate margin $|2 - (T-1)p| \geq \delta_0$, still gives a positive $c\rho^{12}$ lower bound there.

For q≥ 4, the fourth role has not yet been fitted at N_-: $u_4 = O_\ell(\rho^6)$ whereas $u_4^* = c_4\rho^5 + O(\rho^7)$. In a sufficiently small fixed radius-$a\rho$ physical ball about that iterate, gate signs remain fixed and all temporal factors are $O(\rho)$; current-state factors may be $O(\rho)$, not necessarily $O_\ell(\rho^2)$. The readout amplitude r is $O(\rho)$, while a_h stays near one. Temporal-factor counting in the exact filter, or a Cauchy bound on $|z| = c/\rho$ with fixed small c, gives $\|D_\Theta u_4\| \leq C\rho^4$ throughout this ball. Choosing a small preserves a gap $|u_4 - u_4^*| \geq c\rho^5$, and coefficient strong convexity gives a $c\rho^{10}$ risk lower bound. The initial excess is $O(\rho^{10})$; descent bounds the additional path before a possible ball exit by $C\rho\sqrt{s_*}$. Choose $s_* > 0$ so this is less than $a\rho/2$. First exit is impossible, and the lower bound persists. The q=3 upper bound is strictly smaller than all three competing orders for small ρ, proving (267). □

Proof of Corollary E.38. For each fixed depth, the specified total work gives $N_L = \lfloor w/(\eta\rho^8 L)\rfloor$. The common burn-in satisfies $\eta\rho^8 N_0 = O(\rho^2\ell_\rho)$, uniformly for the finite depth set. Thus $\eta\rho^8(N_L - N_0) = w/L + O(\rho^2\ell_\rho + \eta\rho^8)$. Apply (265) at that bounded scaled time to obtain (268). For fixed $L < L'$, the difference of leading factors is $e^{-2\lambda_{T,p}w/L'} - e^{-2\lambda_{T,p}w/L} > 0$. The separated Ξ makes its coefficient uniformly positive, whereas the remainder is $o(\rho^{12})$. All updates from reset were counted. This comparison uses the declared block-call proxy, not measured FLOPs or wall time. □

E.11 Physical Data Response

The preceding results concern population updates. We now keep the same binary hidden Markov task, $25B + 9$-parameter causal network, paired reset, and full physical gradient update, but train on M independent complete sequences. A sequence supplies all T causal next-token losses; its tokens are not treated as independent observations. This comparison asks whether changing which cyclic entrance receives a data shard can reduce fresh-data cross-entropy after training. The comparison is between two schedules at the same B, P_e, L, rather than between different source counts.

A matched multi-update epoch. Fix $P_e \mid B$, $q = B/P_e$, and $M = P_e J_{ep} m$. A data-independent partition produces P_e shards $D_s = \bigsqcup_{b=0}^{J_{ep}-1} D_{s,b}$, each cell containing m complete sequences. At update $n = eJ_{ep} + b$, the fixed schedule sends $D_{s,b}$ to source s; the rotating schedule sends $D_{s+e \bmod P_e,b}$ to source s. For either schedule $a \in \{\mathrm{fix}, \mathrm{rot}\}$, all trainable block occurrences, embedding coordinates, and readout coordinates are differentiated at the same current parameter before one update:

$$\Theta_{n+1}^a = \Theta_n^a - \eta\nabla\widehat{R}_{e,b}^a(\Theta_n^a), \qquad \widehat{R}_{e,b}^a(\Theta) = \frac{1}{P_e}\sum_{s=0}^{P_e-1}\frac{1}{m}\sum_{X\in D_{s,e,b}^a}\frac{1}{T}\sum_{t=1}^{T}\ell(z_t^{(s)}(\Theta;X), X_{t+1}). \tag{300}$$

At every update both arms use the identical sample union $\bigcup_s D_{s,b}$; only its assignment to entrances changes. An epoch has $J_{\rm ep}$ actual updates and visits each sequence once. The checkpoint $N = E_{\rm ep}J_{\rm ep}$ is measured in updates, whereas L is the number of complete backbone passes in a forward path. Both arms use the same η, N and the same fresh-sequence test risk

$$R_{P_{\rm e}}(\Theta) = \frac{1}{P_{\rm e}} \sum_s \mathbb{E}_Q \left[\frac{1}{T} \sum_{t=1}^{T} \ell(z_t^{(s)}(\Theta; X), X_{t+1}) \right]. \tag{301}$$

This averages source losses, not source logits or predictions. The block-token work of either arm is $(M/J_{\rm ep})TBLN = MTBLE_{\rm ep}$; multiplying by $P_{\rm e}$ again would count partitioned sequences twice.

The contrast statistic. With $X_0 = 0$, define the four-dimensional statistic of one complete sequence

$$Z(X) = \frac{1}{T} \sum_{t=1}^{T} X_{t+1}(1, X_t, X_{t-1}, X_t X_{t-1})^{\mathsf{T}}, \qquad W_T = \operatorname{diag}\left(1, 1, \frac{T-1}{T}, \frac{T-1}{T}\right). \tag{302}$$

Let $\widehat{m}_{s,b}$ and $\widehat{m}_s = {J_{\rm ep}}^{-1} \sum_b \widehat{m}_{s,b}$ be its cell and whole-shard means, and put $\widehat{m} = {P_{\rm e}}^{-1} \sum_s \widehat{m}_s$. The first-stage fast-epoch limit is governed by the nonnegative shard-entry contrast

$$\mathcal{A}_D = \frac{1}{2P_{\rm e}} \sum_s (\widehat{m}_s - \widehat{m})^{\mathsf{T}} W_T^{-1} (\widehat{m}_s - \widehat{m}). \tag{303}$$

It measures an association between finite-sample fluctuations and terminal block roles. Common fluctuations of all shards are not removed.

To retain the number of updates per epoch, let $\xi_k = \exp(2\pi \mathrm{i} k/P_{\rm e})$ and $\widehat{d}_{k,b} = {P_{\rm e}}^{-1} \sum_s \xi_k^{-s} \widehat{m}_{s,b}$ for $1 \le k < P_{\rm e}$. Write

$$\begin{aligned} &v = (23/4, 19/4, 3/2, 3/2), \quad w = \operatorname{diag}(W_T), && \mu_\nu = v_\nu w_\nu, \\ &a_\nu = 1 - \eta p^2 \rho^2 \mu_\nu / P_{\rm e}, && r_\nu = a_\nu^{J_{\rm ep}}, && \varpi_{\nu,b} = \frac{(1-a_\nu) a_\nu^{J_{\rm ep}-1-b}}{1 - a_\nu^{J_{\rm ep}}}, \end{aligned} \tag{304}$$

$$\overline{d}_{k,\nu}^{J_{\rm ep}} = \sum_{b=0}^{J_{\rm ep}-1} \varpi_{\nu,b} \widehat{d}_{k,b,\nu}, \qquad \gamma_{\nu k} = \frac{\left| \sum_{e=0}^{E_{\rm ep}-1} r_\nu^{E_{\rm ep}-1-e} \xi_k^e \right|^2}{\left(\sum_{e=0}^{E_{\rm ep}-1} r_\nu^{E_{\rm ep}-1-e} \right)^2}. \tag{305}$$

Here $0 < a_\nu < 1$, $\sum_b \varpi_{\nu,b} = 1$, and $0 \le \gamma_{\nu k} \le 1$. The four rates come from the actual physical terminal-block derivative, not a separate feature-training surrogate.

For a complete-epoch checkpoint in the late first-stage window

$$A_- \rho^{-2} \ell_\rho \le \eta N \le A_+ \rho^{-2} \ell_\rho, \qquad N = E_{\rm ep} J_{\rm ep}, \tag{306}$$

choose fixed $A_+ > A_- > 0$ uniformly over the finite divisor set $P_{\rm e} \mid B$, with A_- large enough that $a_\nu^N \le \rho^2$. The physical linear response yields the explicit nonnegative quadratic

$$\mathcal{B}_{D,J_{\rm ep},N} = \frac{1}{2} \sum_{\nu=1}^{4} \frac{(1 - a_\nu^N)^2}{w_\nu} \sum_{k=1}^{P_{\rm e}-1} (1 - \gamma_{\nu k}) |\overline{d}_{k,\nu}^{J_{\rm ep}}|^2 \ge 0. \tag{307}$$

Its conversion into a finite-data trained-risk difference uses the output-energy bounds in Subsection E.13.4; nonnegativity of this quadratic alone is not a sign theorem for the full risk.

The factor $\gamma_{\nu k}$ records how much source contrast survives the sequence of changing assignments. For a complete source cycle $P_{\rm e} \mid E_{\rm ep}$,

$$\gamma_{\nu k} = \frac{(1 - r_\nu)^2}{1 - 2r_\nu \cos(2\pi k/P_{\rm e}) + r_\nu^2}. \tag{308}$$

Short epochs relative to the contrast fitting time $P_{\rm e}/(\eta p^2 \rho^2 \mu_\nu)$ have $r_\nu \approx 1$ and suppress repeated shard-role adaptation. If an epoch is long enough to refit its current assignment, $r_\nu \approx 0$ and a completed rotation cycle alone gives little terminal suppression. For $P_{\rm e} = 4$ and $4 \mid E_{\rm ep}$, the removed fraction is $2r_\nu/(1 + r_\nu^2)$ in modes $k = 1, 3$ and $4r_\nu/(1 + r_\nu)^2$ in mode $k = 2$. These fractions apply to the leading contrast-risk term, not to the entire test cross-entropy.

E.11.1 PROOF OF THE PHYSICAL RESPONSE

All norms in this section are physical Euclidean norms, except for the explicit ℓ_1 distance between laws on the finite sequence alphabet. Constants may depend on fixed $B, T, \mathcal{I}, c_h$; they are uniform over $P_{\mathrm{e}} \mid B$, the integer epoch parameters, and finite L. We write $t_N = \eta N$ and use $m_\star$ for a finite exponent of $\ell_\rho = \log(1/\rho)$, enlarging it between estimates when needed. The finite-data certificate uses a single upper envelope m_0. The proof uses the population stage-one theorem Lemma E.3; its extension to the later checkpoint in (306) is established first. No empirical trajectory is assumed to remain paired or block-periodic.

E.11.2 THE POPULATION REFERENCE THROUGH THE FULL WINDOW

Let $\bar{\Theta}_n$ be population GD with the same $B, P_{\mathrm{e}}, L, \eta$ and reset. By Lemma E.3, at its first-stage checkpoint $n_{P_{\mathrm{e}},1}$ the regular chart of (51) has

$$\begin{aligned} u_0 = g, \qquad u_1 = gy, \qquad e_g := g - \widehat{g}(y,\nu) = O(\rho^5 \ell_\rho^{m_\star}), \\ y = (1-p)\rho + O(\rho^3 \ell_\rho^{m_\star}), \qquad \|\nu\| = O(\rho^2 \ell_\rho^{m_\star}). \end{aligned} \tag{309}$$

The readout product $r_n a_{h,n}$ is $p\rho + O(\rho e^{-c\eta n}) + O(\rho^3 \ell_\rho^{m_\star})$ after the fixed-duration head startup. We need these estimates at every later legal N, not only at $n_{P_{\mathrm{e}},1}$.

For $q > 1$, (40) at $k = 1$ gives the coarse but sufficient energy bound

$$0 \le R_{P_{\mathrm{e}}}(\bar{\Theta}_{n_{P_{\mathrm{e}},1}}) - V_T \le C\rho^6 \ell_\rho^{m_\star}. \tag{310}$$

For $q = 1$ the bound is no worse; we do not use its sharper $O(\rho^8)$ matched-risk estimate. Stop the population trajectory at first exit from a slightly enlarged fixed-gate box. Within that box the physical Hessian is bounded uniformly in L. The descent lemma, for $\eta \le \eta_0$, yields up to this stopping time

$$\sum_{n=n_{P_{\mathrm{e}},1}}^{N-1} \eta \|\nabla R_{P_{\mathrm{e}}}(\bar{\Theta}_n)\|^2 \le 2\{R_{P_{\mathrm{e}}}(\bar{\Theta}_{n_{P_{\mathrm{e}},1}}) - V_T\} \le C\rho^6 \ell_\rho^{m_\star}. \tag{311}$$

Here $R_{P_{\mathrm{e}}} \ge V_T$ because, on the paired gate trajectory, every source is the same shared linear-history predictor. Cauchy–Schwarz over any $O(\rho^{-2}\ell_\rho/\eta)$ continuation gives

$$\sum_{n=n_{P_{\mathrm{e}},1}}^{N-1} \eta \|\nabla R_{P_{\mathrm{e}}}(\bar{\Theta}_n)\| \le C\rho^2 \ell_\rho^{m_\star}. \tag{312}$$

This is smaller than the fixed gate margin and the learned first temporal role's order-ρ interval. It closes the stopped first-exit argument. Since the regular chart and its inverse have bounded derivatives, the same path estimate also keeps every complementary coordinate within $O(\rho^2 \ell_\rho^{m_\star})$ of its checkpoint value, on every stopped prefix. The small chart-metric cross terms therefore retain their inherited orders. This re-establishes the tube hypotheses of the following recurrences, rather than inferring fine role accuracy from path length alone. The fast, slow, and nuisance recurrences (85)–(87) therefore remain valid in the continuation. Their discrete convolutions preserve (309) and the $O(\rho^2 \ell_\rho^{m_\star})$ raw complementary tube throughout (306). In particular, the $O(\eta\rho^4)$ nuisance forcing accumulates over only $O(\rho^{-2}\ell_\rho/\eta)$ updates. Choosing A_- sufficiently large makes the slow transient smaller than its $O(\rho^3 \ell_\rho^{m_\star})$ floor for each of the finitely many divisors $P_{\mathrm{e}} \mid B$.

For later Hessian propagation, let $\mu_t = \mathbb{E}[X_{t+1} \mid X_{1:t}]$ and set

$$\epsilon_n = \max_{s,t,X_{1:t}} |\tanh z_{s,t}(\bar{\Theta}_n; X_{1:t}) - \mu_t(X_{1:t})|. \tag{313}$$

The HMM filter gives $\mu_t = p\rho X_t + p(1-p)\rho^2 X_{t-1} + O(\rho^3)$ uniformly over the finite prefix alphabet. The preceding fast–slow bounds and the exact filter (29) give

$$\epsilon_n \le C\{\rho e^{-c\eta n} + \rho^2 e^{-c\eta\rho^2 n/P_{\mathrm{e}}} + \rho^3 \ell_\rho^{m_\star}\}, \qquad \sum_{n<N} \eta\epsilon_n \le C. \tag{314}$$

The first term includes head startup. The last term integrates to $O(\rho \ell_\rho^{m_\star+1})$ on the window, while the two exponentials have bounded integrals.

The full terminal physical gradient needs a stronger argument than the raw nuisance-update estimate. In the regular chart, $u_0 = g, u_1 = gy$ exactly, so $D_\nu u_0 = D_\nu u_1 = 0$ when g, y are held fixed. Equations (45) and (48) and (309) give the late scores

$$F_1 = O(\rho^4 \ell_\rho^{m_\star}), \qquad F_j = O(\rho^3 \ell_\rho^{m_\star}) \ (j \geq 2). \tag{315}$$

For each $j \geq 2$, every term of u_j in the exact finite-history filter contains the readout and enough temporal factors to reach lag j. Differentiating once in any tangent chart coordinate, including embedding and head coordinates, removes at most one factor. For all represented lags, not only the range $j \leq q$ displayed in (52), expand the complete filter (29). A lag-j term contains j temporal factors and one readout factor; a derivative leaves order ρ^j up to fixed logarithmic factors. The number of assignments in a path of length k grows at most polynomially in k at fixed T. After extracting at most T lag factors and the differentiated factor, the remaining current-state factors form a geometric tail. One uniform majorant is a constant times $\rho^j \ell_\rho^{c_T} \sum_{k \geq 0} (k+1)^{T+2} \varepsilon^{(k-T-2)_+}$ with $\varepsilon \leq 1/2$. This sum is bounded independently of L and includes the propagated-input term and repeated occurrences of the same role. Thus $\|Du_j\| \leq C\rho^2 \ell_\rho^{m_\star}$ for every $j \geq 2$; coefficients with $j > BL$ vanish. There are only $T-2$ high rows. The exact chain rule now reads

$$\begin{aligned} R_y &= gF_1 + \sum_{j \geq 2} F_j D_y u_j = O(\rho^5 \ell_\rho^{m_\star}), \\ R_\nu &= \sum_{j \geq 2} F_j D_\nu u_j = O(\rho^5 \ell_\rho^{m_\star}). \end{aligned} \tag{316}$$

For R_g, the fast-root relation and its tracking error $e_g = O(\rho^5 \ell_\rho^{m_\star})$ give the same order; the high-row chain terms remain $O(\rho^5 \ell_\rho^{m_\star})$. This distinction matters because $r_n a_{h,n} - p\rho = O(\rho^3)$ is not itself the fast-root error. Population sign, neuron-pair, and cyclic block symmetries set every normal component outside their invariant tangent space to zero exactly. The chart Jacobian and physical metric are bounded on the fixed box, so the complete $25B + 9$-coordinate statement is

$$\|\nabla R_{P_e}(\bar{\Theta}_N)\| \leq C\rho^5 \ell_\rho^{m_\star}. \tag{317}$$

E.11.3 Full-space differentiability, stability, and data response

The reset maps separated intervals around the two token states strictly into their interiors, with zero incoming-state Jacobian. Continuity gives a fixed physical parameter box in which all ReLU gate signs remain fixed, these intervals remain invariant, and the state Jacobian of each block has norm at most $\gamma < 1$. This box permits broken neuron pairing, nonperiodic blocks, additive biases, and independently moving endpoints. If $U_j^{(r)}$ is an rth parameter derivative of the hidden sequence after call j, then direct differentiation of (19) gives

$$\|U_{j+1}^{(1)}\| \leq \gamma \|U_j^{(1)}\| + C, \qquad \|U_{j+1}^{(r)}\| \leq \gamma \|U_j^{(r)}\| + C_r \mathcal{Q}_r(\|U_j^{(1)}\|, \ldots, \|U_j^{(r-1)}\|), \quad 2 \leq r \leq 4. \tag{318}$$

The polynomials $\mathcal{Q}_r$ have fixed coefficients. Geometric induction bounds all four derivative orders of logits and single-sequence losses uniformly over finite L; the number of physical parameter coordinates is fixed with B.

Along the population path the physical Hessian decomposes exactly as

$$H_n = \frac{1}{PT} \sum_{s,t} \mathbb{E}\left[\operatorname{sech}^2(z_{s,t}) D z_{s,t} D z_{s,t}^{\mathsf{T}} + (\tanh z_{s,t} - \mu_t) D^2 z_{s,t}\right]. \tag{319}$$

The first term is positive semidefinite; causality allows the target to be replaced by μ_t in the second. Thus $H_n \succeq -C\epsilon_n I$ and $\|H_n\| \leq C$. For sufficiently small η_0, each factor $I - \eta H_n$ has norm at most $1 + C\eta\epsilon_n$ (positive eigenvalues remain inside the unit disk). By (314), every ordered product obeys

$$\left\| \prod_{j=a}^{b} (I - \eta H_j) \right\| \leq C. \tag{320}$$

No positive-curvature assumption is imposed on all physical directions.

For a cell law let $\Delta_{s,b} = \widehat{Q}_{s,b} - Q$. The gradient field at update (e, b) is affine in these law perturbations and differs from its population field by at most $C\delta_D$, with the same derivative bound. For either schedule, the mean-value update around $\bar{\Theta}_n$, (320), and bounded third derivatives imply, for the maximum displacement d_* through time N,

$$d_* \le Ct_N\delta_D + Ct_N(d_*^2 + \delta_D d_*). \tag{321}$$

Stop at the boundary of a smaller full-space gate box. Under $\delta_D(1+t_N)^2 \le c$, continuity in the law interpolation and (321) close the first exit, yielding

$$\max_{n\le N} \|\Theta_n^a - \bar{\Theta}_n\| \le Ct_N\delta_D. \tag{322}$$

The same statement holds on each update segment and for every scalar interpolation $Q + h\Delta$, $0 \le h \le 1$. It does not claim descent of a single empirical objective under changing minibatches.

The propagator bound needed below holds on the entire interpolation, not only at $h = 0$. Let $\Theta_n^{a,h}$ be either interpolated trajectory and $\widehat{H}_n^{a,h}$ the Hessian of its interpolated update loss at that point. The full-space derivative bounds and (322) give, uniformly in h and the two schedules,

$$\|\widehat{H}_n^{a,h} - H_n\| \le C(\delta_D + \|\Theta_n^{a,h} - \bar{\Theta}_n\|) \le C\delta_D(1+t_N), \quad \sum_{n<N} \eta\|\widehat{H}_n^{a,h} - H_n\| \le C\delta_D(1+t_N)^2. \tag{323}$$

Together with $H_n \succeq -C\epsilon_n I$ and (314), small η_0 gives every interpolated ordered propagator the uniform bound

$$\left\| \prod_{j=u}^{v} (I - \eta\widehat{H}_j^{a,h}) \right\| \le C\exp\big(C\delta_D(1+t_N)^2\big) \le Ce^{Cc}. \tag{324}$$

Here $u \le v$ are arbitrary update indices and a denotes the schedule. Thus the subsequent data derivatives and the integral Taylor remainder are controlled at every intermediate law, including their moving empirical Hessians.

Write the first and second derivatives at $h = 0$ of this interpolation as U_n^a and V_n^a. If h_n^a denotes the linear data-gradient field, the exact recurrences are

$$U_{n+1}^a = (I - \eta H_n)U_n^a - \eta h_n^a, \tag{325}$$

$$V_{n+1}^a = (I - \eta H_n)V_n^a - \eta\{D^3 R_{P_e}(\bar{\Theta}_n)[U_n^a, U_n^a] + 2Dh_n^a U_n^a\}. \tag{326}$$

Taylor expansion of the terminal risk gives

$$R_{P_e}(\Theta_N^a) = R_{P_e}(\bar{\Theta}_N) + g_N^\mathsf{T} U_N^a + \frac{1}{2}(U_N^a)^\mathsf{T} H_N U_N^a + \frac{1}{2} g_N^\mathsf{T} V_N^a + \mathcal{R}_3^a, \qquad g_N = \nabla R_{P_e}(\bar{\Theta}_N). \tag{327}$$

The output-energy argument in Subsections E.13.3 and E.13.4 bounds the third-order remainder at the scale needed for the finite-data theorem.

All bounds use the maximum cell-law perturbation, so they do not grow with the number $P_e J_{ep}$ of law coordinates.

E.11.4 Cyclic projection and the shared-endpoint term

Let U shift physical block labels by q while fixing the embedding and head, and let $\Pi = {P_e}^{-1}\sum_{j=0}^{P_e-1} \mathsf{U}^j$. It is an orthogonal projector onto the common sector. Source equivariance implies $\ell_s(\mathsf{U}\Theta; X) = \ell_{s-1}(\Theta; X)$, $\Pi g_n = g_n$, and $[H_n, \Pi] = 0$ at the paired reference. For every slot set

$$\bar{\Delta}_b = {P_e}^{-1}\sum_s \Delta_{s,b}, \qquad \widetilde{\Delta}_{s,b} = \Delta_{s,b} - \bar{\Delta}_b. \tag{328}$$

The $\bar{\Delta}_b$ field is exactly the same in both schedules at every update. Since $\sum_s \widetilde{\Delta}_{s,b} = 0$, its scheduled source-specific field has zero group average at the reference: $\Pi h_{e,b}^- = 0$. Consequently the first variations split as $U_n^a = U_n^+ + U_n^{-,a}$, where U_n^+ is identical in both arms and $\Pi U_n^{-,a} = 0$. In particular, the shared endpoints have no first-order contrast forcing, although they respond to the common field.

The invariant gradient kills $g_N^\mathsf{T} U_N^{-,a}$. The invariant Hessian kills cross terms between U_N^+ and $U_N^{-,a}$. In (326), the purely common second response is identical in the two arms; a common–contrast product transforms in the contrast sector and is again orthogonal to g_N. A product of two contrast modes, however, can have a common component. Denote that second response by $V_N^{--,a}$, retaining its head and embedding coordinates. Since contrast forcing on blocks is $O(\rho\delta_D)$, (325) and (320) give $\|U_n^{-,a}\| \le C\rho t_N \delta_D$. The two forcing terms in (326) then give

$$\|V_N^{--,a}\| \le C(\rho^2 t_N^3 + \rho t_N^2)\delta_D^2. \tag{329}$$

The ρt_N^2 term allows a mixed head–block derivative without a block–block readout factor. At $t_N = O(\rho^{-2}\ell_\rho)$, (317) yields

$$|g_N^\mathsf{T} V_N^{--,a}| \le C\rho \ell_\rho^{m_\star} \delta_D^2. \tag{330}$$

Thus the shared endpoints are neither frozen nor assumed harmless; their second-order effect is below the leading contrast risk.

E.11.5 The complete shared-occurrence block derivative

For source s number its $D_{\mathrm{blk}} = BL$ block calls from 0 to $D_{\mathrm{blk}} - 1$, let $i_{s,j}$ be the physical block at call j, and define $A_j = D_h T_{i_{s,j}}$ and $G_j = D_{\theta_{i_{s,j}}} T_{i_{s,j}}$ at its incoming state. Shared use of block b gives the exact chain rule

$$D_{\theta_b} h_{D_{\mathrm{blk}}}^{(s)} = \sum_{j : i_{s,j} = b} A_{D_{\mathrm{blk}}-1} \cdots A_{j+1} G_j. \tag{331}$$

At the reference, $\|A_j\| \le \varepsilon = C\rho\ell_\rho^{m_\star} < 1/2$, $\|G_j\| \le C$, and each incoming state differs from its reset token state by $O(\varepsilon)$. The terminal call is the unique summand with empty suffix and uses block $b(s) = sq - 1 \pmod B$. Every other summand has at least one downstream A_j. Even if a physical block occurs L times, the sum over its distinct suffix lengths is bounded by the full geometric series $C\varepsilon/(1-\varepsilon)$, uniformly in L.

Here is an explicit description of the reset terminal feature in all 25 raw coordinates. For an input token pair (x, y), with $y = 0$ at $t = 1$, put $g_a = w_a x + \pi_a y + b_a$ at reset and define the vector

$$\mathcal{F}(x,y) = \big((\mathrm{ReLU}(g_a),\ v_a \mathbf{1}_{g_a>0} x,\ v_a \mathbf{1}_{g_a>0} y,\ v_a \mathbf{1}_{g_a>0})_{a=1}^6, 1\big)^\mathsf{T}. \tag{332}$$

The tuple order is (v_a, w_a, π_a, b_a) followed by d. Its Walsh expansion on $x, y \in \{-1, 1\}$ is $\mathcal{F}(x,y) = f_1 + x f_2 + y f_3 + xy f_4$; the same formula at $y = 0$ has its last two terms zero. Substituting the reset values (23) into (332) gives

$$f_\nu^\mathsf{T} f_\tau = 0 \quad (\nu \ne \tau), \qquad (\|f_1\|^2, \|f_2\|^2, \|f_3\|^2, \|f_4\|^2) = (23/4, 19/4, 3/2, 3/2). \tag{333}$$

This calculation includes the output bias in f_1 and a genuinely pairing-breaking fourth feature; it is not confined to the paired subspace. Since the head is locally affine in its input, the terminal summand in (331) gives, uniformly in s, b, t,

$$D_{\theta_b} z_{s,t}(\bar\Theta_n) = r_n a_{h,n} \mathbf{1}_{\{b=b(s)\}}(f_1 + X_t f_2 + X_{t-1} f_3 + X_t X_{t-1} f_4) + O(\rho^2 \ell_\rho^{m_\star}). \tag{334}$$

During head startup, replacing $r_n a_{h,n}$ by $p\rho$ adds $O(\rho e^{-c\eta n} + \rho^3 \ell_\rho^{m_\star})$. Every one of the $B - P_{\mathrm{e}}$ nonterminal blocks is still trained; its leading contrast derivative is zero only at this order. The geometric sum accounts for all its later and earlier repeated occurrences.

Let $\mathsf{F} = (f_1, \dots, f_4)$ and $\mathsf{K} = \mathsf{F} W_T \mathsf{F}^\mathsf{T}$. On the nontrivial cyclic sector $\Pi^\perp$, the leading physical Hessian acts on the terminal orbit by

$$(H_0^- v)_{b(s)} = \frac{p^2\rho^2}{P_{\mathrm{e}}} \mathsf{K} v_{b(s)}, \qquad \sum_s v_{b(s)} = 0, \tag{335}$$

and vanishes on other block orbits. Formula (334), including off-diagonal terminal–nonterminal blocks, proves this as an operator-norm expansion on the entire contrast space. The four positive eigenvalues are $p^2\rho^2\mu_\nu/P_{\mathrm{e}}$ per nonzero source Fourier mode, where $\mu_\nu = v_\nu w_\nu$; there are $4(P_{\mathrm{e}} - 1)$ positive modes. No positive curvature is claimed on the other modes.

For two block arguments, but not for mixed head–block arguments, the paired-head identity is exact:

$$D^2_{BB} z_{s,t} = r_n a_{h,n} D^2_{BB} h_{s,t}. \tag{336}$$

The differentiated recurrence (318) bounds the hidden second derivative uniformly in L. The residual part of the contrast Hessian in (319) is therefore at most $C\rho\epsilon_n$. The Gauss–Newton difference from (335) has three contributions: readout startup $O(\rho^2 e^{-c\eta n})$, the $O(\rho)$ weak-HMM feature covariance change, and the $O(\rho \ell_\rho^{m_\star})$ terminal-derivative change. The latter two cost $O(\rho^3 \ell_\rho^{m_\star})$ after multiplying two readout factors. Together with (314), they give

$$\|H_n^- - H_0^-\| \le C\rho^2 e^{-c\eta n} + C\rho^3 \ell_\rho^{m_\star}, \qquad \sum_{n<N} \eta \|H_n^- - H_0^-\| \le C\rho \ell_\rho^{m_\star}. \tag{337}$$

Using a generic $O(1)$ bound for $D^2 z$ here would lose the essential readout factor and would give only an $O(1)$ integrated error.

The same full occurrence derivative in the centered empirical gradient gives, on the terminal orbit,

$$(h^-_{e,b})_{b(s)} = -\frac{p\rho}{P_{\mathrm{e}}} \mathsf{F}(\widehat{m}_{s+e,b} - P_{\mathrm{e}}{}^{-1} \textstyle\sum_j \widehat{m}_{j,b}) + \mathrm{err}_{e,b,s}, \tag{338}$$

where all terminal and nonterminal components of the error have norm at most $C[\rho e^{-c\eta n} + \rho^2 \ell_\rho^{m_\star}]\delta_D$. Endpoint components cancel exactly in the centered source sum. Summing the error with the stable propagator costs at most $C\ell_\rho^{m_\star}\delta_D$.

E.11.6 Solving the epoch response and transferring it to risk

Project the terminal block response onto $f_\nu/\sqrt{v_\nu}$ and use the normalized source transform $\widehat{x}_k = P_{\mathrm{e}}{}^{-1} \sum_s \xi_k^{-s} x_s$. In a nonzero mode, the constant-coefficient part of (325) becomes

$$\widehat{U}_{\nu k,n+1} = a_\nu \widehat{U}_{\nu k,n} + \eta \frac{p\rho\sqrt{v_\nu}}{P_{\mathrm{e}}} \xi_k^e \widehat{d}_{k,b,\nu} \tag{339}$$

for rotation; fixed binding replaces ξ_k^e by 1. The sign follows from the negative gradient in (325) and (338). The exact epoch sum is

$$\sum_{b=0}^{J_{\mathrm{ep}}-1} a_\nu^{J_{\mathrm{ep}}-1-b} \widehat{d}_{k,b,\nu} = \frac{1-r_\nu}{1-a_\nu} \overline{d}^{J_{\mathrm{ep}}}_{k,\nu}. \tag{340}$$

Let $H_{\nu,k} = \sum_{e=0}^{E_{\mathrm{ep}}-1} r_\nu^{E_{\mathrm{ep}}-1-e} \xi_k^e$. Since $(1-r_\nu)H_{\nu,0} = 1 - a_\nu^N$, solving (339) gives

$$\widehat{U}^{\mathrm{fix}}_{\nu k,N} = \frac{1-a_\nu^N}{p\rho\sqrt{v_\nu} w_\nu} \overline{d}^{J_{\mathrm{ep}}}_{k,\nu},$$
$$\widehat{U}^{\mathrm{rot}}_{\nu k,N} = \frac{H_{\nu,k}}{H_{\nu,0}} \widehat{U}^{\mathrm{fix}}_{\nu k,N}. \tag{341}$$

The triangle inequality gives $|H_{\nu,k}| \le H_{\nu,0}$, hence $0 \le \gamma_{\nu k} \le 1$. Summing the finite geometric series gives (305); if $P_{\mathrm{e}} \mid E_{\mathrm{ep}}$, cancellation of $1 - r_\nu^{E_{\mathrm{ep}}}$ gives (308). For $E_{\mathrm{ep}} = 1$, $\gamma_{\nu k} = 1$, consistently with identical schedules.

The leading response has norm at most $C\delta_D/\rho$. Discrete Duhamel with (337) changes it by at most $C\rho\ell_\rho^{m_\star}(\delta_D/\rho) = C\ell_\rho^{m_\star}\delta_D$. The forcing error after (338) has the same bound. This includes leading-null block modes, whose direct leading forcing is zero. The actual first contrast response thus differs from (341) by $C\ell_\rho^{m_\star}\delta_D$ in the full physical space. At the endpoint, replacing the actual contrast Hessian by H_0^- costs $C\rho\ell_\rho^{m_\star}\delta_D^2$: both first responses are $O(\delta_D/\rho)$ and the Hessian difference is $O(\rho^3\ell_\rho^{m_\star})$.

Parseval for the normalized source transform gives $\sum_s |x_s|^2 = P_{\mathrm{e}} \sum_k |\widehat{x}_k|^2$. Inserting (341) into $\frac{1}{2}(U_N^-)^{\mathsf{T}} H_0^- U_N^-$, the P_{e} from Parseval cancels the $1/P_{\mathrm{e}}$ in (335); the remaining scalar is $(1-a_\nu^N)^2/(2w_\nu)$. The fixed-minus-rotated quadratic difference is exactly $\mathcal{B}_{D,J_{\mathrm{ep}},N}$ in (307). All linear, common, and mixed terms of (327) cancel by Subsection E.11.4; the pure-contrast second response is bounded by (330). Finally the third-order Taylor remainder is the one in (327). This identifies the nonnegative physical-response quadratic. The sharp remainder and the resulting trained-risk comparison are proved in Subsections E.13.4 and E.14.1.

E.11.7 Fast epochs, complete-sequence fluctuations, and quantifiers

Suppose $\eta\rho^2 J_{\mathrm{ep}} \to 0$. Normalized geometric weights satisfy

$$\sum_b |\varpi_{\nu,b} - {J_{\mathrm{ep}}}^{-1}| \le C\eta\rho^2 J_{\mathrm{ep}}. \tag{342}$$

The lower window constant gives $a_\nu^N \le \rho^2$. For each fixed nonzero k, the finite-series formula and $|\xi_k - 1| \ge c_{P_{\mathrm{e}}}$ show $\gamma_{\nu k} \le C_{P_{\mathrm{e}}}(\eta\rho^2 J_{\mathrm{ep}})^2$; the denominator stays away from zero because $r_\nu^{E_{\mathrm{ep}}} = a_\nu^N \le \rho^2$. Parseval applied to whole-shard means then gives

$$|\mathcal{B}_{D,J_{\mathrm{ep}},N} - \mathcal{A}_D| \le C\{\rho^2 + \eta\rho^2 J_{\mathrm{ep}} + (\eta\rho^2 J_{\mathrm{ep}})^2\}\delta_D^2. \tag{343}$$

At $\rho = 0$ the observations are iid symmetric signs. With $V_t = X_{t+1}(1, X_t, X_{t-1}, X_t X_{t-1})$, $X_0 = 0$, each V_t is a martingale difference against the observed past. Distinct t terms are orthogonal even though their predictor coordinates overlap. The first term has second-moment diagonal $(1, 1, 0, 0)$; each later term has I_4. Hence the covariance of the *whole-sequence* statistic is

$$\Sigma_0 := \mathrm{Cov}_{\rho=0} Z = W_T/T, \qquad \Sigma_\rho = W_T/T + O(\rho), \tag{344}$$

where the second statement follows from smoothness of the finite HMM sequence law. No token-level independence is used at positive ρ.

Each shard contains M/P_{e} independent sequences. Let $Y_s = \sqrt{M/P_{\mathrm{e}}}(\widehat{m}_s - \mathbb{E}Z)$. In any joint limit with $M/P_{\mathrm{e}} \to \infty$ and $\rho \downarrow 0$, the bounded triangular-array CLT gives independent $Y_s \Rightarrow N(0, W_T/T)$. Centering their common mean gives the exact identity

$$M\mathcal{A}_D = \frac{1}{2}\sum_s (Y_s - \bar{Y})^{\mathsf{T}} W_T^{-1}(Y_s - \bar{Y}) \Rightarrow \frac{\chi^2_{4(P_{\mathrm{e}}-1)}}{2T}. \tag{345}$$

There are $P_{\mathrm{e}} - 1$ real source contrasts in each of four features, not $2(P_{\mathrm{e}} - 1)$ independent complex modes. At every M the exact mean is

$$M\mathbb{E}_D\mathcal{A}_D = \frac{P_{\mathrm{e}} - 1}{2}\,\mathrm{tr}(W_T^{-1}\Sigma_\rho) = \frac{2(P_{\mathrm{e}} - 1)}{T} + O(\rho). \tag{346}$$

For fixed B, T the sequence alphabet is finite. A cell-frequency concentration union bound over $P_{\mathrm{e}}J_{\mathrm{ep}}$ cells, each of size $M/(P_{\mathrm{e}}J_{\mathrm{ep}})$, gives

$$\delta_D = O_{\mathbb{P}}\left(\sqrt{J_{\mathrm{ep}}\log(2P_{\mathrm{e}}J_{\mathrm{ep}})/M}\right). \tag{347}$$

The sharp output-energy estimate and the reassignment argument use (347) together with the response derived above. For fixed J_{ep}, L, η and a preselected complete-epoch checkpoint $N(\rho)$ in (306), the sequential expectation has the distinct quantifier order

$$\lim_{\rho\downarrow 0}\lim_{\substack{M\to\infty \\ P_{\mathrm{e}}J_{\mathrm{ep}}|M}} M\mathbb{E}_D\{R_{P_{\mathrm{e}}}(\Theta^{\mathrm{fix}}_{N(\rho)}) - R_{P_{\mathrm{e}}}(\Theta^{\mathrm{rot}}_{N(\rho)})\} = \frac{2(P_{\mathrm{e}} - 1)}{T}. \tag{348}$$

Indeed, at fixed $\rho, N, J_{\mathrm{ep}}, L$, interpolation in Subsection E.11.3 has a quadratic Taylor coefficient in finitely many cell-law fluctuations. The third-order remainder is $O(M^{-3/2})$ in expectation on the gate event; its complement has exponentially small probability by finite-alphabet concentration. For fixed ρ, N, L, the finite-iterate network has bounded test risk over the compact empirical-law simplex. Thus the inner limit exists. The sharp risk transfer in Subsection E.13.4 and the fast-epoch replacement (343) reduce that coefficient to $M\mathbb{E}_D\mathcal{A}_D + O(\rho\ell_\rho^{m\star})$ at fixed J_{ep}. Equation (346) gives the outer limit. This does not assert a joint expectation limit for growing J_{ep}.

One can retain the epoch-length dependence at the level of the explicit quadratic form. Independence of the fixed cells gives, for $k \neq 0$, $\mathbb{E}|\widehat{d}_{k,b,\nu}|^2 = J_{\mathrm{ep}}(\Sigma_\rho)_{\nu\nu}/M$; hence

$$\begin{aligned}
\mathbb{E}_D\mathcal{B}_{D,J_{\mathrm{ep}},N} &= \frac{1}{2M}\sum_{\nu=1}^{4}\frac{(\Sigma_\rho)_{\nu\nu}}{w_\nu}V_{\nu,J_{\mathrm{ep}}}(1 - a_\nu^N)^2\sum_{k=1}^{P_{\mathrm{e}}-1}(1 - \gamma_{\nu k}), \\
V_{\nu,J_{\mathrm{ep}}} &= J_{\mathrm{ep}}\sum_b \varpi_{\nu,b}^2 \\
&= J_{\mathrm{ep}}\frac{1 - a_\nu}{1 + a_\nu}\frac{1 + r_\nu}{1 - r_\nu} \ge 1.
\end{aligned} \tag{349}$$

This is the expectation of the leading quadratic, not an exact formula for finite-M trained risk. It separates weighted reuse of later minibatches from the source-rotation factor. Changing J_{ep} also changes cell size, so (349) alone is not a pure rotation-rate experiment.

E.12 Ordinary Margin at a Common Checkpoint

We compare the ordinary Loop with two full-entrance schedules on the same finite-width model, data, and stopping time. Fix $B \geq 2$, $T \geq 4$, and let Q be the law of one complete HMM sequence $(X_1, \ldots, X_{T+1})$. For $M = BJ_{\mathrm{ep}}m$ independent sequences, an exogenous partition forms cells $D_{s,b}$, $0 \leq s < B$, $0 \leq b < J_{\mathrm{ep}}$, each containing m whole sequences. Write $\widehat{Q}_{s,b}$ for its empirical law and $\widehat{Q}_{\mathrm{pool},b} = B^{-1}\sum_s \widehat{Q}_{s,b}$. The three methods, denoted O (ordinary), F (fixed binding), and R (rotation), start from the same reset and use the same B, L, η and number of updates. At update $n = eJ_{\mathrm{ep}} + b$, their empirical objectives are

$$\widehat{R}^{O}_{e,b}(\Theta) = \int \ell_0(\Theta; x)\, d\widehat{Q}_{\mathrm{pool},b}(x),$$

$$\widehat{R}^{F}_{e,b}(\Theta) = \frac{1}{B}\sum_{s=0}^{B-1} \int \ell_s(\Theta; x)\, d\widehat{Q}_{s,b}(x),$$

$$\widehat{R}^{R}_{e,b}(\Theta) = \frac{1}{B}\sum_{s=0}^{B-1} \int \ell_s(\Theta; x)\, d\widehat{Q}_{s+e \bmod B,b}(x), \qquad \Theta^a_{n+1} = \Theta^a_n - \eta\nabla\widehat{R}^a_{e,b}(\Theta^a_n), \tag{350}$$

where $\ell_s(\Theta; x) = T^{-1}\sum_{t=1}^{T} \ell(z_t^{(s)}(\Theta; x), x_{t+1})$. Each update in all three methods uses the same union $\bigcup_s D_{s,b}$, hence the same M/J_{ep} sequences. Every physical block occurrence, the embedding, and the shared head contribute to the single full-parameter Euclidean update. The test risks are fresh-sequence CE: $\mathcal{T}_O = R_{B,1,L}(\Theta^O_N)$ at entrance zero, and $\mathcal{T}_F = R_{B,B,L}(\Theta^F_N)$, $\mathcal{T}_R = R_{B,B,L}(\Theta^R_N)$ with source losses averaged, not predictions ensembled. All use the same $25B+9$ physical parameters. Their common block-token work proxy is $(M/J_{\mathrm{ep}})TBLN$; no additional factor B is charged for partitioning the source batch.

Let $\ell_\rho = \log(1/\rho)$,

$$a_T(p) = \frac{T-2}{2T}p^2(1-p)^4, \qquad \delta_D = \max_{s,b}\|\widehat{Q}_{s,b} - Q\|_1, \qquad t_N = \eta N. \tag{351}$$

Proposition E.39 (Common-window population margin)**.** *Fix finite $B \geq 2, T \geq 4$, $p \in \mathcal{I} \Subset (0,1)$, the exact paired reset of Assumption E.1, and a preselected finite depth set $\mathcal{L}$. For sufficiently small fixed η and ρ, take $L \in \mathcal{L}$ and a preselected integer checkpoint N in*

$$A\rho^{-2}\ell_\rho \leq \eta N \leq \Gamma A\rho^{-2}\ell_\rho, \qquad \Gamma > 1, \tag{352}$$

with A sufficiently large for both population trajectories. Then, for a finite logarithmic exponent m,

$$R_{B,1,L}(\bar{\Theta}^1_N) - R_{B,B,L}(\bar{\Theta}^B_N) = a_T(p)\rho^6 + O(\rho^7\ell_\rho^m), \qquad a_T(p) = \frac{T-2}{2T}p^2(1-p)^4 > 0. \tag{353}$$

The references have the same physical block count, initialization, step size, update count, and loop depth, but are trained separately.

The transfer of this margin to the three empirical arms uses the conditional cross-entropy identity below and the sharp output-energy estimate in Subsection E.13.4. The finite-data theorem and its four-arm consequence are stated in the reassignment section; the population proposition by itself makes no empirical ordering claim.

E.12.1 Proof of the common-window margin

We prove the population margin and record the full-space identities needed for its finite-data transfer. All parameter norms and Hessians below are in the original physical $25B + 9$ coordinates unless we explicitly restrict to the invariant population tangent space. Constants may depend on fixed $B, T, \mathcal{I}$ and reset bounds. They are uniform over the preselected finite depth set and do not depend on the cell laws or epoch count. We use the inherited coefficient geometry and stage estimates Lemmas E.3 and E.21, Subsection E.2.1, and Equation (51) and the full-space derivative estimates established in Subsections E.11.2 and E.11.3.

E.12.2 Population comparison on one window

For $P_{\mathrm{e}} \in \{1, B\}$, let $\bar{\Theta}_n^{P_{\mathrm{e}}}$ denote the population trajectory from reset under (22); let $q = B/P_{\mathrm{e}}$. On a population trajectory the sources have a common linear-history predictor and risk $F_{\rho,T}(u) \geq V_T$. Write $n_{P_{\mathrm{e}},1}$ for its first-stage checkpoint from (35). Choose the lower window constant A in (352) large enough that every legal N follows both $n_{1,1}$ and $n_{B,1}$, and large enough for the four source-contrast modes. The upper endpoint is a fixed multiple of these checkpoint scales.

At its own checkpoint, each reference has

$$g = p\rho + O(\rho^3 \ell_\rho^m), \quad y = (1-p)\rho + O(\rho^3 \ell_\rho^m), \quad \|\nu\| = O(\rho^2 \ell_\rho^m), \quad \mathcal{E}_1 \leq C\rho^8 \ell_\rho^m, \tag{354}$$

by Lemma E.3, with $u_0 = g, u_1 = gy$ in the regular chart. Here and below m denotes a finite logarithmic exponent that may increase. The ordinary checkpoint has $q = B \geq 2$, so the second role is still unlearned; every complementary raw coordinate, including t_2, is $O(\rho^2 \ell_\rho^m)$. The full-entrance checkpoint has $q = 1$, $D_{\mathrm{blk}} = BL \geq 2$, and its risk already obeys

$$R_{B,B,L}(\bar{\Theta}^B_{n_{B,1}}) - V_T \leq C\rho^8 \ell_\rho^m \tag{355}$$

by the matched case of (40).

We extend each reference from its own checkpoint; we do not restart its profile recursion at reset or identify the two reference parameters. Stop either path at the first exit from a slightly enlarged gate box. The checkpoint excess is $O(\rho^6 \ell_\rho^m)$ for both paths. Descent and Cauchy–Schwarz on any continuation of at most $O(\rho^{-2}\ell_\rho/\eta)$ updates give

$$\sum_n \eta \|\nabla R_{P_{\mathrm{e}}}(\bar{\Theta}_n^{P_{\mathrm{e}}})\|^2 \leq C\rho^6 \ell_\rho^m, \qquad \sum_n \eta \|\nabla R_{P_{\mathrm{e}}}(\bar{\Theta}_n^{P_{\mathrm{e}}})\| \leq C\rho^2 \ell_\rho^m. \tag{356}$$

Thus no path exits the fixed gate box, and the complementary coordinates stay in an $O(\rho^2 \ell_\rho^m)$ tube. This coarse path estimate alone does not preserve the ordinary risk coefficient, so we also propagate the low-profile deficit.

Put $a = (u_0, u_1)$, $b = (u_2, \ldots, u_{T-1})$ and define

$$H_1(\Theta) = \min_{v_0, v_1} F_{\rho,T}(v_0, v_1, b(\Theta)), \qquad \mathcal{E}_1(\Theta) = F_{\rho,T}(u(\Theta)) - H_1(\Theta). \tag{357}$$

The physical conditional-profile lemma Lemma E.21 applies to the actual unrestricted population step. In the regular chart, $D_{(g,y)}(u_0, u_1) = \left(\begin{smallmatrix} 1 & 0 \\ y & g \end{smallmatrix}\right)$ has least singular value at least $c\rho$, since $g = p\rho + O(\rho^3 \ell_\rho^m)$ and p stays interior. The physical chart metric in (56) is uniformly nondegenerate for fixed P_{e}. Moreover, the stage-one row-scaled cancellation (53)–(55) gives $\|J_{b,f} J_{a,f}^{-1}\| = O(\rho \ell_\rho^m)$, while the coefficient profile cross matrix in Lemma E.21 has the same order. Thus the effective profile Jacobian retains least singular value at least $c\rho$, and the physical chart metric gives

$$\|\nabla_\Theta \mathcal{E}_1\|^2 \geq \frac{c\rho^2}{P_{\mathrm{e}}} \mathcal{E}_1. \tag{358}$$

Here is the forcing bound needed to close this estimate. On the stopped tube, the coefficient-gradient expansion (45), the first-stage role orders, and the exact finite filter imply

$$|F_j(a, b)| \leq C\rho^3 \ell_\rho^m, \qquad \|D_\Theta u_j\| \leq C\rho^2 \ell_\rho^m, \qquad j \geq 2. \tag{359}$$

For $P_{\mathrm{e}} = 1$, the first bound at $j = 2$ is attained in order because $u_2 = O(\rho^4 \ell_\rho^m)$ while $u_2^* = p(1-p)^2\rho^3 + O(\rho^5)$; for $P_{\mathrm{e}} = B$ it follows from the same coefficient expansion. The derivative bound differentiates the complete filter, including the propagated input term, and sums repeated shared-block occurrences using the contractive geometric majorant of Subsection E.11.2 and Equation (318). At the conditional low minimizer, $\mathcal{E}_1 = O(\rho^8 \ell_\rho^m)$ moves (u_0, u_1) by only $O(\rho^4 \ell_\rho^m)$. The lag-decaying coefficient Hessian (47) changes each high score by at most $O(\rho^5 \ell_\rho^m)$, so the $O(\rho^3 \ell_\rho^m)$ score bound remains valid there. Consequently

$$\|\nabla H_1\| = \left\| \sum_{j \geq 2} F_j(a_*(b), b) D_\Theta u_j \right\| \leq C\rho^5 \ell_\rho^m. \tag{360}$$

Combining (358), (360), and (189) gives, on every stopped prefix,

$$\mathcal{E}_{1,n+1} \le \left(1 - \frac{c\eta\rho^2}{P_{\mathrm{e}}}\right)\mathcal{E}_{1,n} + C\eta\rho^{10}\ell_\rho^m. \tag{361}$$

Start this recurrence at $n_{P_{\mathrm{e}},1}$, where (354) holds. Its solution over the full fixed-multiple window remains $\mathcal{E}_1 = O(\rho^8\ell_\rho^m)$; choosing a larger bootstrap constant closes the first-exit argument. This checkpoint-started profile bootstrap, together with (356), holds for both $P_{\mathrm{e}} = 1$ and $P_{\mathrm{e}} = B$.

For ordinary training, $q = B \ge 2$ and the second temporal role stays $t_2 = O(\rho^2\ell_\rho^m)$ throughout the continuation. The degree-two part of the complete filter therefore gives

$$u_2^O = O(\rho^4\ell_\rho^m), \qquad u_2^* = p(1-p)^2\rho^3 + O(\rho^5). \tag{362}$$

Rows $j \ge 3$ contribute only $O(\rho^8\ell_\rho^m)$ to risk at this checkpoint: their coefficient errors are at most order ρ^4 in the weighted norm, and their cross-curvature with row two has the extra lag factor $O(\rho)$. Profiling u_0, u_1 and using (46)–(47) now yields

$$R_{B,1,L}(\bar\Theta_N^1) - V_T = \frac{w_2}{2}\{p(1-p)^2\}^2\rho^6 + O(\rho^7\ell_\rho^m) = a_T(p)\rho^6 + O(\rho^7\ell_\rho^m). \tag{363}$$

For full entrance $P_{\mathrm{e}} = B$, the risk is nonincreasing after its own first-stage checkpoint and is bounded below by V_T on the population invariant trajectory. Thus (355) gives, uniformly throughout the common continuation,

$$0 \le R_{B,B,L}(\bar\Theta_N^B) - V_T \le C\rho^8\ell_\rho^m. \tag{364}$$

Subtracting (364) from (363) proves the population margin $a_T(p)\rho^6 + O(\rho^7\ell_\rho^m)$ at every common checkpoint.

E.12.3 Full-space Hessian propagation and the empirical tube

For each method $a \in \{O, F, R\}$, let $\bar\Theta_n^a$ be its own population reference, and write $P_{\mathrm{e}a} = 1$ for O and $P_{\mathrm{e}a} = B$ for F, R. The two full-entrance references coincide; the ordinary reference is generally different. Define the conditional prediction residual

$$\epsilon_n^a = \max_{s,t,x_{1:t}} |\tanh z_t^{(s)}(\bar\Theta_n^a; x) - \mu_t(x_{1:t})|, \qquad \mu_t = \mathbb{E}[X_{t+1} \mid X_{1:t}]. \tag{365}$$

The inherited first-stage fast–slow estimates, extended from each $n_{P_{\mathrm{e}a},1}$ by the stopped continuation above, imply

$$\epsilon_n^a \le C\{\rho e^{-c\eta n} + \rho^2 e^{-c\eta\rho^2 n/P_{\mathrm{e}a}} + \rho^3\ell_\rho^m\}, \quad \sum_{n<N}\eta\epsilon_n^a \le C, \quad \epsilon_N^a \le C\rho^3\ell_\rho^m. \tag{366}$$

The lower window constant is chosen large enough for the last bound. The first term includes zero-head startup; it is not removed by resetting the time origin at the first-stage checkpoint.

The reset has a full physical gate box: all hidden-state gates remain fixed, block state Jacobians have norm at most $\gamma < 1$, and this box allows unpaired neurons, nonperiodic blocks, both additive biases, embedding, and head coordinates to move. Differentiating the actual shared-block recurrence and summing its repeated occurrences gives bounded physical logit derivatives through order four, uniformly in finite L; see (318). In the full parameter space the population Hessian along reference a is

$$H_n^a = \mathbb{E}\left[\operatorname{sech}^2(z_n) D z_n D z_n^{\mathsf{T}} + (\tanh z_n - \mu_t) D^2 z_n\right]. \tag{367}$$

The first term is positive semidefinite. The complete second term, including every normal and endpoint direction, has norm at most $C\epsilon_n^a$, so $H_n^a \succeq -C\epsilon_n^a I$ and $\|H_n^a\| \le C$. For small fixed η, each factor $I - \eta H_n^a$ has norm at most $1 + C\eta\epsilon_n^a$. Therefore every ordered full-space propagator satisfies

$$\left\|\prod_{j=u}^{v}(I - \eta H_j^a)\right\| \le \exp\left(C\sum_{j=u}^{v}\eta\epsilon_j^a\right) \le C. \tag{368}$$

No positive-curvature assumption is made on the full physical Hessian.

Let $h_n^a = \Theta_n^a - \bar{\Theta}_n^a$. On the full gate box, every single-sequence loss gradient and its first two physical derivatives are bounded. For ordinary training, $\|\widehat{Q}_{\mathrm{pool},b} - Q\|_1 \le \delta_D$; for either full-entrance schedule, each assigned source cell has the same bound. Hence the empirical gradient differs from its population field by at most $C\delta_D$ at every update, without an unbiasedness assumption conditional on the past. Taylor expansion at the reference gives

$$h_{n+1}^a = (I - \eta H_n^a)h_n^a - \eta d_n^a + r_n^a, \quad \|d_n^a\| \le C\delta_D, \quad \|r_n^a\| \le C\eta(\|h_n^a\|^2 + \delta_D\|h_n^a\|). \tag{369}$$

Stop at the first exit from a smaller full-space gate box. Applying (368) to (369) and taking a maximum over the stopped prefix yields

$$d_* \le Ct_N\delta_D + Ct_N(d_*^2 + \delta_D d_*). \tag{370}$$

If $\delta_D(1+t_N)^2 \le c$ with c sufficiently small, the first-exit bootstrap closes and gives, for all three uninterrupted empirical paths,

$$\max_{n\le N}\|\Theta_n^a - \bar{\Theta}_n^a\| \le Ct_N\delta_D, \qquad a \in \{O, F, R\}. \tag{371}$$

The estimate includes every update segment and does not reset parameters at epoch boundaries. It does not use empirical pairing or block periodicity.

E.12.4 Conditional cross-entropy identity

The full-space logit derivative bound and (371) give $b_N^a := \sup_{s,t,x}|z_t^{(s)}(\Theta_N^a; x) - z_t^{(s)}(\bar{\Theta}_N^a; x)| \le Ct_N\delta_D$. For a binary target $Y \in \{-1, 1\}$, $\ell(z, Y) = \log(2\cosh z) - Yz$. Conditional on the observed prefix, exact Taylor expansion gives

$$\mathbb{E}[\ell(\bar{z} + \Delta z, Y) - \ell(\bar{z}, Y) \mid X_{1:t}] = (\tanh\bar{z} - \mu_t)\Delta z \tag{372}$$

$$+ \int_0^1 (1-v)\,\mathrm{sech}^2(\bar{z} + v\Delta z)(\Delta z)^2\,dv. \tag{373}$$

Averaging this identity over prefixes, positions, and the appropriate sources supplies the risk-transfer step. The sharp bound needed for the finite-data sample range is (400); the population residual at the endpoint is (366). No empirical path is assumed to be paired or block-periodic.

For concentration, one cell contains $m = M/(BJ_{\mathrm{ep}})$ independent complete sequences from an alphabet of size $K_T = 2^{T+1}$. The dual representation of the ℓ_1 distance as a maximum over 2^{K_T} sign vectors and Hoeffding's inequality, union-bounded over BJ_{ep} cells, give

$$\Pr(\delta_D > d) \le BJ_{\mathrm{ep}}\,2^{K_T}\exp\left(-\frac{Md^2}{2BJ_{\mathrm{ep}}}\right), \tag{374}$$

This bound counts independent sequences, not token positions.

E.13 Cross-entropy energy bounds for empirical training

This section sharpens the full-parameter data-response estimate used in Subsections E.11.1 and E.12.1. It concerns the original $25B + 9$ Euclidean coordinates, the paired reset, and mean GD, at fixed B, T, P_{e}, L in the first slow-stage window. A schedule may assign the exogenous cell laws to sources differently at each update; no stochastic independence between the update and the current parameters is used. Put $t_N = \eta N$, $\mathfrak{T} = 1 + t_N$, and $\delta = \max_{s,b}\|\widehat{Q}_{s,b} - Q\|_1$. Constants below may depend on the fixed model, compact interior range of p, and preselected finite depth set, but not on the cell values, schedule, or N. The population reference for a given P_{e}, L is denoted $\bar{\Theta}_n$.

E.13.1 Finite-output geometry and the stopped empirical path

Collect every source, scored prefix, and its logit into the finite map $z(\Theta)$. Give its range the inner product weighted by the true prefix probabilities and by $1/(P_{\mathrm{e}}T)$. Since T is fixed and every prefix has positive probability uniformly on the compact HMM parameter range, this norm is equivalent to an ordinary finite-dimensional norm. In these coordinates the fresh-data risk is $R(\Theta) = \Phi(z(\Theta))$, where, with $A(v) = \log(2\cosh v)$,

$$\Phi(z) = \frac{1}{P_{\mathrm{e}}T}\sum_{s,t,x_{1:t}} Q(x_{1:t})\{A(z_{s,t,x}) - \mu_t(x_{1:t})z_{s,t,x}\}. \tag{375}$$

The physical empirical loss at update n, under the interpolation of every cell law $Q + u(\widehat{Q}_{s,b} - Q)$, has the exact form

$$R_{n,u}(\Theta) = \Phi(z(\Theta)) + u\Psi_n(z(\Theta)), \qquad 0 \le u \le 1. \tag{376}$$

The signed functional Ψ_n is determined by that update's cell assignment, including every shared-block occurrence and endpoint. It is affine in the complete-sequence cell-law deviations. Consequently, on the fixed gate box of Subsection E.11.3,

$$\lambda I \preceq D^2\Phi \preceq I, \quad \|D^j\Psi_n\| \le C\delta \quad (1 \le j \le 4), \quad \|D^j z\| \le C \quad (1 \le j \le 4). \tag{377}$$

Here $\lambda > 0$ is a bounded-logit constant; it is output-space curvature, not positive definiteness of the physical Hessian. In particular, the empirical forcing is precisely $Dz(\Theta)^\top D\Psi_n(z(\Theta))$. Uniform bounds on Dz include the complete occurrence sum (331).

Let Θ_n^u be the actual interpolated GD path from the common reset, and write $e_n = \Theta_n^u - \bar{\Theta}_n$ and $d_n = z(\Theta_n^u) - z(\bar{\Theta}_n)$. All estimates in this subsection are stopped on the full-space gate box. Let $J_x = Dz(\Theta_n^u)$ and $J_y = Dz(\bar{\Theta}_n)$. The gradient difference is

$$g_n = J_x^\top\{D\Phi(z_x) - D\Phi(z_y)\} + (J_x - J_y)^\top D\Phi(z_y) + uJ_x^\top D\Psi_n(z_x), \qquad e_{n+1} = \quad e_n - \eta g_n. \tag{378}$$

Taylor's theorem gives $J_x e_n = d_n + O(\|e_n\|^2)$. The conditional prediction residual of the population path, denoted ϵ_n in (314), bounds the weighted output gradient $\|D\Phi(z_y)\| \le C\epsilon_n$. Output strong convexity, Young's inequality, and (377) therefore give

$$\langle e_n, g_n\rangle \ge c\|d_n\|^2 - C(\epsilon_n + \delta)\|e_n\|^2 - C\|e_n\|^4 - C\delta^2, \tag{379}$$

$$\|g_n\|^2 \le C\{\|d_n\|^2 + \epsilon_n^2\|e_n\|^2 + \delta^2\}. \tag{380}$$

For clarity, the nonlinear term from $J_x e_n - d_n$ is first bounded by $C\|e_n\|^2\|d_n\|$ and then absorbed into the output quadratic, leaving $C\|e_n\|^4$; replacing d_n by $J_x e_n$ at the outset would lose this cancellation. Choose the fixed step-size ceiling small enough that the $\eta^2\|g_n\|^2$ term in the exact squared GD update absorbs into its negative $\eta\|d_n\|^2$ term. Then

$$\|e_{n+1}\|^2 \le [1 + C\eta(\epsilon_n + \delta)]\|e_n\|^2 - c\eta\|d_n\|^2 + C\eta\|e_n\|^4 + C\eta\delta^2. \tag{381}$$

The inherited estimate (314) has $\sum_{n<N}\eta\epsilon_n \le C$. Stop also at $\max_{k\le n}\|e_k\|^2 = K\mathfrak{T}\delta^2$. On a stopped prefix the quartic term is at most $CK\mathfrak{T}\delta^2\|e_n\|^2$, and the two additional exponents in discrete Gronwall are bounded by $C\delta t_N + CKt_N\mathfrak{T}\delta^2$. Fix K above the resulting constant and then take $\delta\mathfrak{T} \le c_0$ with c_0 sufficiently small. Since $e_0 = 0$, the first exit cannot occur. The resulting bound, uniformly in $u \in [0,1]$ and in the exogenous schedule, is

$$\sup_{n\le N}\|e_n\|^2 + \sum_{n<N}\eta\|d_n\|^2 \le C\mathfrak{T}\delta^2. \tag{382}$$

The displacement is smaller than the fixed gate margin as $\mathfrak{T}\delta \to 0$, closing the physical first-exit argument. This argument does not assume that the empirical path remains paired or that each changing minibatch objective descends.

Along any such path, put $J_n = Dz(\Theta_n^u)$ and $H_n^u = D^2_\Theta R_{n,u}(\Theta_n^u)$. The chain rule gives

$$H_n^u = J_n^\top A_n J_n + S_n, \qquad \lambda I/2 \preceq A_n \preceq CI, \qquad \|S_n\| \le C(\epsilon_n + \|d_n\| + \delta) =: \kappa_n. \tag{383}$$

The lower bound on A_n follows by perturbing $D^2\Phi$ by $uD^2\Psi_n$; the term S_n contracts $D^2 z$ with the empirical conditional output residual. Cauchy–Schwarz and (382) imply

$$\sum_{n<N}\eta\kappa_n \le C\left(1 + \sqrt{t_N}\Big(\sum_{n<N}\eta\|d_n\|^2\Big)^{1/2} + t_N\delta\right) \le C. \tag{384}$$

This holds over every interpolated path, not only at $u = 0$.

E.13.2 A DISCRETE INPUT–OUTPUT ENERGY INEQUALITY

We record the precise linear estimate used for data derivatives. Let J_n have uniformly bounded norm, let $H_n = J_n^\top A_n J_n + S_n$, with $\lambda I/2 \preceq A_n \preceq CI$, $\|S_n\| \le C$ and $\sum_{n<N} \eta\|S_n\| \le C_0$. For $0 < \eta \le \eta_0$, where η_0 depends on these uniform bounds, suppose

$$Y_{n+1} = (I - \eta H_n)Y_n - \eta(J_n^\top a_n + b_n). \tag{385}$$

There is no assumption that b_n lies in the output Jacobian range. Provided the displayed right-hand side is finite, one has

$$\|Y\|_\infty + \left(\sum_{n<N} \eta\|J_n Y_n\|^2\right)^{1/2} \le C_{C_0}\left[\|Y_0\| + \left(\sum_{n<N} \eta\|a_n\|^2\right)^{1/2} + \sum_{n<N} \eta\|b_n\|\right]. \tag{386}$$

In particular, both the upper bound on A_n and the integrated bound on S_n are part of the lemma; small η is required for its discrete, rather than continuous-time, conclusion.

To prove it, split $Y = Y^{(a)} + Y^{(b)}$, assigning Y_0 to the first part. For $Y^{(a)}$, expand its exact squared recurrence. The term $-2\eta\langle J_n Y_n^{(a)}, a_n\rangle$ is absorbed by the $J_n^\top A_n J_n$ energy and $C\eta\|a_n\|^2$; bounded J_n, A_n, S_n and small η absorb the $\eta^2\|H_n Y_n^{(a)} + J_n^\top a_n\|^2$ increment. The only multiplicative term left is $1 + C\eta\|S_n\|$, whose product is at most e^{CC_0}. This proves (386) for $Y^{(a)}$. For $Y^{(b)}$, each homogeneous factor has norm at most $1 + C\eta\|S_n\|$: its positive-semidefinite part is nonexpansive when $\eta \le \eta_0$. Variation of constants gives $\|Y^{(b)}\|_\infty \le C_{C_0}\sum \eta\|b_n\|$. The same squared-energy expansion, now retaining the cross term as $2\eta\|Y_n^{(b)}\|\|b_n\|$, gives

$$\sum \eta\|J_n Y_n^{(b)}\|^2 \le C\left(\|Y^{(b)}\|_\infty^2 \sum \eta\|S_n\| + \|Y^{(b)}\|_\infty \sum \eta\|b_n\| + (\sum \eta\|b_n\|)^2\right).$$

Here the last term also controls the exact $\eta^2\|b_n\|^2$ increments because $\sum \eta^2\|b_n\|^2 \le (\sum \eta\|b_n\|)^2$. This proves the lemma without a pointwise bound on b_n.

E.13.3 FIRST, SECOND, AND THIRD DATA RESPONSES

Write $U_n = \partial_u \Theta_n^u$, $V_n = \partial_u^2 \Theta_n^u$, and $W_n = \partial_u^3 \Theta_n^u$. All start at zero because the reset is data-independent. All displayed parameter derivatives are in the original physical coordinates. Differentiating the true GD update gives

$$U_{n+1} = (I - \eta H_n^u)U_n - \eta J_n^\top D\Psi_n, \tag{387}$$

$$V_{n+1} = (I - \eta H_n^u)V_n - \eta\{D^3 R_{n,u}[U_n, U_n] + 2D^2(\Psi_n \circ z)U_n\}, \tag{388}$$

$$W_{n+1} = (I - \eta H_n^u)W_n - \eta\{D^4 R_{n,u}[U_n, U_n, U_n] + 3D^3 R_{n,u}[U_n, V_n] + 3D^3(\Psi_n \circ z)[U_n, U_n] + 3D^2(\Psi_n \circ z)V_n\}. \tag{389}$$

All functions in these equations are evaluated at Θ_n^u. Apply (386) to (387), taking $a_n = D\Psi_n$ and $b_n = 0$. Because $\sum \eta\|D\Psi_n\|^2 \le C t_N \delta^2$,

$$\|U\|_\infty + \|JU\|_{L^2_\eta} \le C\mathfrak{T}^{1/2}\delta. \tag{390}$$

For the second response, the forcing in braces in (388) is exactly $J_n^\top a_n^{(2)} + b_n^{(2)}$. Writing $z_j = D^j z$ and $\Phi_{n,u} = \Phi + u\Psi_n$, its components are

$$a_n^{(2)} = D^3\Phi_{n,u}[JU, JU] + D^2\Phi_{n,u}\, z_2[U, U] + 2D^2\Psi_n\, JU, \tag{391}$$

$$b_n^{(2)} = 2z_2[U, \cdot]^\top D^2\Phi_{n,u}\, JU + z_3[U, U, \cdot]^\top D\Phi_{n,u} + 2z_2[U, \cdot]^\top D\Psi_n. \tag{392}$$

The middle term of $b^{(2)}$ uses the empirical output residual, whose norm is at most κ_n; it cannot be replaced by an arbitrary constant without losing a time power. For example, $\sum \eta\|U\|\|JU\| \le \|U\|_\infty\sqrt{t_N}\|JU\|_{L^2_\eta}$. The bounded output derivatives, (384), and (390) give term by term

$$\|a^{(2)}\|_{L^2_\eta} + \|b^{(2)}\|_{L^1_\eta} \le C\mathfrak{T}^{3/2}\delta^2. \tag{393}$$

Thus the repaired discrete lemma yields

$$\|V\|_\infty + \|JV\|_{L^2_\eta} \le C\mathfrak{T}^{3/2}\delta^2. \tag{394}$$

For completeness, the third-response forcing can be verified without dropping normal or endpoint derivatives. Repeated chain rule in (389) places a leading $J_n^\top$ on the outermost output derivative and leaves all terms in which a derivative hits J_n in $b_n^{(3)}$. The $D^3R[U,V]$ terms in $a_n^{(3)}$ are linear combinations of $D^3\Phi_{n,u}[JU,JV]$ and $D^2\Phi_{n,u}z_2[U,V]$; their $b_n^{(3)}$ terms are combinations of $z_2[U,\cdot]^\top D^2\Phi_{n,u}JV$, $z_2[V,\cdot]^\top D^2\Phi_{n,u}JU$, and $z_3[U,V,\cdot]^\top D\Phi_{n,u}$. The $D^4R[U,U,U]$ terms additionally involve $D^4\Phi_{n,u}[JU,JU,JU]$, $D^3\Phi_{n,u}[JU,z_2[U,U]]$, $D^2\Phi_{n,u}z_3[U,U,U]$, the corresponding z_2 and z_3 transpose terms, and $z_4[U,U,U,\cdot]^\top D\Phi_{n,u}$. The two explicit Ψ_n terms have the same product-rule families with one output derivative bounded by $C\delta$ instead of by a bounded derivative of $\Phi_{n,u}$. These are all partitions of the three directional derivatives among the outer output function and z; in particular, z_4 is included.

Using the response bounds (390) and (394), the pointwise inequality $\|JU\|_\infty \le C\|U\|_\infty$, and $\sum\eta\|D\Phi_{n,u}\| \le C$ from (384), the four forcing families satisfy

	$\|a\|_{L^2_\eta}$	$\|b\|_{L^1_\eta}$
$D^3R[U,V]$	$C\mathfrak{T}^{5/2}\delta^3$	$C\mathfrak{T}^{5/2}\delta^3$
$D^4R[U,U,U]$	$C\mathfrak{T}^{2}\delta^3$	$C\mathfrak{T}^{5/2}\delta^3$
$D^3(\Psi_n\circ z)[U,U]$	$C\mathfrak{T}^{3/2}\delta^3$	$C\mathfrak{T}^{2}\delta^3$
$D^2(\Psi_n\circ z)V$	$C\mathfrak{T}^{3/2}\delta^3$	$C\mathfrak{T}^{5/2}\delta^3$

(395)

For instance, $\|z_2[U,V]\|_{L^2_\eta} \le C\sqrt{t_N}\|U\|_\infty\|V\|_\infty \le C\mathfrak{T}^{5/2}\delta^3$. The $z_4^\top D\Phi_{n,u}$ term is bounded in L^1_η by $C\|U\|_\infty^3\sum\eta\kappa_n$. Applying (386) once more yields

$$\|W\|_\infty + \|JW\|_{L^2_\eta} \le C\mathfrak{T}^{5/2}\delta^3. \tag{396}$$

E.13.4 Endpoint risk and ordinary transfer

Let $\mathcal{L}(u) = R(\Theta_N^u)$ be fresh-data risk. The exact endpoint chain rule is

$$\mathcal{L}'''(u) = D^3R[U_N,U_N,U_N] + 3D^2R[U_N,V_N] + DR[W_N]. \tag{397}$$

Write $G_N = \|\nabla R(\bar\Theta_N)\|$. Bounded Hessian and (382) give $\|\nabla R(\Theta_N^u)\| \le G_N + C\sqrt{\mathfrak{T}}\delta$. Therefore

$$|\mathcal{L}'''(u)| \le C\mathfrak{T}^2\delta^3 + CG_N\mathfrak{T}^{5/2}\delta^3 + C\mathfrak{T}^3\delta^4. \tag{398}$$

For the fixed-P_{e} rotation reference, $G_N \le C\rho^5\ell_\rho^{m_\star}$ by (317); the weaker bound $G_N \le C\rho^3\ell_\rho^{m_\star}$ also follows directly from its terminal conditional residual and bounded Dz. In the prescribed $\mathfrak{T} = O(\rho^{-2}\ell_\rho)$ window, $G_N\sqrt{\mathfrak{T}} = o(1)$. Moreover $\mathfrak{T}\delta \le c_0$. Thus, uniformly over the entire data interpolation,

$$\sup_{0\le u\le 1}|\mathcal{L}'''(u)| \le C\mathfrak{T}^2\delta^3. \tag{399}$$

Taylor's integral remainder at $u=0$ is at most the same order. This supplies the cubic error $C\mathfrak{T}^2\delta^3$ used in the second-order fixed-versus-rotation calculation (327). The quadratic coefficient and its $C\rho\ell_\rho^{m_0}\delta^2$ approximation are unchanged. This statement by itself does not compute the quadratic for a new schedule; that calculation must also preserve the zero-mode and shared-endpoint terms.

For ordinary $P_{\mathrm{e}} = 1$ and full entrance $P_{\mathrm{e}} = B$, use each arm's own population reference and the residual bound (366). On the event $\delta\mathfrak{T} \le c_0$, bounded Dz and (382) give a terminal logit displacement at most $C\sqrt{\mathfrak{T}}\delta$. The exact conditional-CE expansion (373), with no assertion that the empirical predictor is linear-history, gives for every arm a

$$|R_a(\Theta_N^a) - R_a(\bar\Theta_N^a)| \le C\{\rho^3\ell_\rho^{m_1}\sqrt{\mathfrak{T}}\delta + \mathfrak{T}\delta^2\}. \tag{400}$$

Combined with (363) and (364), this preserves the same-window ordinary minus fixed full-entrance coefficient $a_T(p)\rho^6 + o_{\mathbb{P}}(\rho^6)$ under the sample range below. It is a ρ^6-precision statement, not a joint M^{-1}-precision ordinary-versus-reassignment expansion.

E.13.5 Cell concentration and the sufficient sample exponents

Each of the $P_{\mathrm{e}}J_{\mathrm{ep}}$ exogenous cells contains $m = M/(P_{\mathrm{e}}J_{\mathrm{ep}})$ independent complete HMM sequences. The sequence alphabet has 2^{T+1} elements. The same sign-vector union bound as in (374) gives, with probability at least $1-\alpha$,

$$\delta^2 \le \frac{H_\alpha}{M}, \qquad H_\alpha = 2P_{\mathrm{e}}J_{\mathrm{ep}}\left[2^{T+1}\log 2 + \log\frac{P_{\mathrm{e}}J_{\mathrm{ep}}}{\alpha}\right]. \tag{401}$$

Hence the deterministic cubic remainder and the inherited quadratic approximation are bounded on this event by

$$C\rho\ell_\rho^{m_0}\frac{H_\alpha}{M} + C\mathfrak{T}^2\left(\frac{H_\alpha}{M}\right)^{3/2}, \qquad \mathfrak{T}\sqrt{H_\alpha/M} \le c_0. \tag{402}$$

For $0 \le a < 1$, fixed $C_{J_{\mathrm{ep}}} \ge 2$, $\varepsilon > 0$, $2 \le J_{\mathrm{ep}} \le C_{J_{\mathrm{ep}}}\rho^{-a}$, $P_{\mathrm{e}}J_{\mathrm{ep}} \mid M$, and $M \ge \rho^{-(8+3a+\varepsilon)}$, choose $\alpha = \rho^2$. At fixed B, T, P_{e}, L, $H_\alpha = O(\rho^{-a}\,\mathrm{polylog}(1/\rho))$ and $\mathfrak{T} = O(\rho^{-2}\ell_\rho)$. The two errors multiplied by M are bounded respectively by

$$O(\rho^{1-a}\,\mathrm{polylog}(1/\rho)), \qquad O(\rho^{\varepsilon/2}\,\mathrm{polylog}(1/\rho)), \tag{403}$$

which tend to zero. The tube condition follows as well. Moreover (400) is $o_{\mathbb{P}}(\rho^6)$: the first term is $O_{\mathbb{P}}(\rho^{6+a+\varepsilon/2}\,\mathrm{polylog}(1/\rho))$ and the second is $O_{\mathbb{P}}(\rho^{6+2a+\varepsilon}\,\mathrm{polylog}(1/\rho))$. For each fixed exogenous partition, the concentration event is common to all its preselected complete-epoch checkpoints and finite depth choices. For several preselected legal divisors P_{e}, apply the bound to each chosen partition and union-bound the finitely many events (equivalently, allocate the desired failure probability among them). This is a joint-probability remainder result; weak convergence of the leading quadratic alone supplies no joint expectation limit or growing-B, T uniformity.

E.14 Data reassignment across entrances

The following finite-data result quantifies how exogenous reassignment of complete-sequence shards changes the fresh-data risk of the same physical network. Its leading term is a weighted variance across entrance assignments.

Matched protocols. Fix $P_{\mathrm{e}} \mid B$ and partition $M = P_{\mathrm{e}}J_{\mathrm{ep}}m$ independent *complete sequences* exogenously into cells $D_{s,b}$, $0 \le s < P_{\mathrm{e}}$, $0 \le b < J_{\mathrm{ep}}$. Write $\widehat{Q}_{s,b}$ for a cell's empirical sequence law. An epoch has J_{ep} updates; at update $n = eJ_{\mathrm{ep}} + b$, schedule π assigns cell $D_{\pi_e(s),b}$ to entrance s. Its empirical objective and update are

$$\begin{aligned} \widehat{R}^{\pi}_{e,b}(\Theta) &= \frac{1}{P_{\mathrm{e}}}\sum_{s=0}^{P_{\mathrm{e}}-1}\int \ell_s(\Theta; X)\, d\widehat{Q}_{\pi_e(s),b}(X), \\ \Theta^{\pi}_{n+1} &= \Theta^{\pi}_n - \eta\nabla\widehat{R}^{\pi}_{e,b}(\Theta^{\pi}_n). \end{aligned} \tag{404}$$

Here ℓ_s averages the T causal next-token binary cross-entropies from source s. The fixed schedule F has $\pi_e(s) = s$; the cyclic schedule C has $\pi_e(s) = s + e \pmod{P_{\mathrm{e}}}$; and the random schedule U has $\pi_0 = \mathrm{id}$, followed by independent uniform whole-shard permutations, independent of the observed sequence values. Schedule C is the rotating assignment of Subsection E.11. All schedules use the same reset, all $25B + 9$ physical parameters, the same L, η, N, and the same union $\bigcup_s D_{s,b}$ at each update. Each sequence is used once per epoch and reused in later epochs. Test risk averages fresh-sequence source losses, not predictions. Define $\delta_D = \max_{s,b}\|\widehat{Q}_{s,b} - Q\|_1$ and write $\mathcal{T}_\pi = R_{P_{\mathrm{e}}}(\Theta^{\pi}_N)$ for the fresh-sequence source-average test risk of schedule π.

A four-entrance example. With four entrances, cyclic reassignment sends each shard through all four starting points over four successive epochs; fixed binding leaves its starting point unchanged. Each epoch still contains J_{ep} updates, and all schedules process the same sequence union at each update. Only the shard-to-entrance assignment changes.

Let $N = E_{\mathrm{ep}}J_{\mathrm{ep}}$ be a complete-epoch checkpoint and $\ell_\rho = \log(1/\rho)$. The prediction task, finite-width network, paired reset and physical Euclidean update are those of Subsection E.1.1, Assumption E.1, and Equation (22). Risks $\mathcal{T}_F, \mathcal{T}_C, \mathcal{T}_U$ use the source average at P_{e} entrances. When $P_{\mathrm{e}} = B$, also let $\mathcal{T}_O$ be the ordinary-loop fresh-sequence loss at entrance zero, as in Equation (351).

Theorem E.40 (Finite-data certificate for exogenous reassignment)**.** *Fix finite $B \geq 2$, $T \geq 4$, a compact $\mathcal{I} \Subset (0,1)$ for p, the paired reset, and a finite L. There are constants $A_- > 0$, $A_+ > A_-$, $\eta_0, \rho_0, c, C > 0$, and a finite m_0 such that the following holds for every legal $P_{\mathrm{e}} \mid B$, $0 < \eta \leq \eta_0$, $0 < \rho \leq \rho_0$, and complete-epoch $N = E_{\mathrm{ep}} J_{\mathrm{ep}}$ in the common first-stage window*

$$A_- \rho^{-2} \ell_\rho \leq t_N = \eta N \leq A_+ \rho^{-2} \ell_\rho. \tag{405}$$

For any preselected epoch-permutation sequence independent of the cell values, if $\delta_D(1+t_N) \leq c$, all interpolating empirical paths remain in the full physical fixed-gate region. Their fresh-risk difference from fixed binding satisfies

$$\mathcal{T}_F - \mathcal{T}_\pi = \mathcal{B}_D^\pi + \mathrm{Rem}_D^\pi, \qquad |\mathrm{Rem}_D^\pi| \leq C\rho \ell_\rho^{m_0} \delta_D^2 + C(1+t_N)^2 \delta_D^3. \tag{406}$$

The constants are uniform over the finite divisor set $P_{\mathrm{e}} \mid B$ and the preselected finite L; they may depend on fixed $B, T, \mathcal{I}$ and reset bounds. In particular, this is a full-parameter GD result for the stated physical network, not a feature-training surrogate.

The leading term has an interpretable variance form. Use the four-dimensional sequence statistic Z and matrix W_T from Equation (302). For each coordinate ν, center the cell means across shards, $d_{\nu,b,s} = \widehat{m}_{s,b,\nu} - {P_{\mathrm{e}}}^{-1} \sum_j \widehat{m}_{j,b,\nu}$, and let $\bar{d}_\nu^{J_{\mathrm{ep}}} = \sum_{b=0}^{J_{\mathrm{ep}}-1} \varpi_{\nu,b} d_{\nu,b}$ with the physical within-epoch weights $\varpi_{\nu,b}$ of Equation (304). Write $\|x\|_{P_{\mathrm{e}}}^2 = {P_{\mathrm{e}}}^{-1} \sum_s x_s^2$, let P_e be the permutation matrix defined by $(\mathsf{P}_e x)_s = x_{\pi_e(s)}$, and set

$$r_\nu = a_\nu^{J_{\mathrm{ep}}}, \qquad \omega_{\nu,e} = \frac{r_\nu^{E_{\mathrm{ep}}-1-e}}{\sum_{j=0}^{E_{\mathrm{ep}}-1} r_\nu^{E_{\mathrm{ep}}-1-j}}, \qquad \mathsf{Q}_{\nu,\pi} = \sum_{e=0}^{E_{\mathrm{ep}}-1} \omega_{\nu,e} \mathsf{P}_e. \tag{407}$$

Then the quadratic in Theorem E.40 is

$$\mathcal{B}_D^\pi = \frac{1}{2} \sum_{\nu=1}^{4} \frac{(1-a_\nu^N)^2}{w_\nu} \left(\|\bar{d}_\nu^{J_{\mathrm{ep}}}\|_{P_{\mathrm{e}}}^2 - \|\mathsf{Q}_{\nu,\pi} \bar{d}_\nu^{J_{\mathrm{ep}}}\|_{P_{\mathrm{e}}}^2 \right) \geq 0, \qquad w_\nu = (W_T)_{\nu\nu}. \tag{408}$$

Indeed, each P_e is an isometry and the weights are nonnegative and sum to one; more explicitly,

$$\|d\|_{P_{\mathrm{e}}}^2 - \left\| \sum_e \omega_e \mathsf{P}_e d \right\|_{P_{\mathrm{e}}}^2 = \frac{1}{2} \sum_{e,f} \omega_e \omega_f \|\mathsf{P}_e d - \mathsf{P}_f d\|_{P_{\mathrm{e}}}^2. \tag{409}$$

Thus the removed component is the weighted variance of a shard contrast across its assigned entrances. The proof below derives this formula from the four physical contrast modes and controls the common mode and shared endpoint terms; arbitrary permutations of the data do not permute the network blocks. See Subsections E.13 and E.14.1.

Corollary E.41 (Test-Risk Reduction under Data Reassignment)**.** *Fix finite B, T, L, compact-interior p, exact reset, $0 \leq \alpha < 1$, $\varepsilon > 0$, and $C_J \geq 2$. For sufficiently small $\rho > 0$ and $0 < \eta \leq \eta_0$, take $P_{\mathrm{e}} \mid B$, $2 \leq J_{\mathrm{ep}} \leq C_J \rho^{-\alpha}$, $P_{\mathrm{e}} J_{\mathrm{ep}} \mid M$, and $M \geq \rho^{-(8+3\alpha+\varepsilon)}$. Use the constants A_-, A_+, η_0 of Theorem E.40. Preselect a complete-epoch $N = E_{\mathrm{ep}} J_{\mathrm{ep}}$ in the certified common window $A_- \rho^{-2} \ell_\rho \leq \eta N \leq A_+ \rho^{-2} \ell_\rho$. Then, as $\rho \downarrow 0$,*

$$\mathcal{T}_F - \mathcal{T}_a = \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}), \quad a \in \{C, U\}, \qquad M\mathcal{A}_D \Rightarrow \frac{\chi^2_{4(P_{\mathrm{e}}-1)}}{2T}. \tag{410}$$

Here $a_T(p) = (T-2)p^2(1-p)^4/(2T)$ and $\mathcal{A}_D$ is the nonnegative whole-shard contrast quadratic in Equation (303). At $P_{\mathrm{e}} = 1$, the three schedules coincide and $\mathcal{A}_D = 0$; the displayed limit is the point mass at zero. For $P_{\mathrm{e}} > 1$, both gains are strictly positive with probability tending to one; the random-arm probability includes schedule randomness. At $P_{\mathrm{e}} = B > 1$, the same data and checkpoint also give

$$\mathcal{T}_O - \mathcal{T}_F = a_T(p)\rho^6 + o_{\mathbb{P}}(\rho^6), \qquad \mathbb{P}\{\mathcal{T}_O > \mathcal{T}_F > \max(\mathcal{T}_C, \mathcal{T}_U)\} \longrightarrow 1. \tag{411}$$

The same leading statistic for C and U yields $\mathcal{T}_C - \mathcal{T}_U = o_{\mathbb{P}}(M^{-1})$; this corollary does not resolve their ordering. The ρ^6 ordinary-to-fixed margin and the M^{-1} reassignment gain have different precision, so they do not give an M^{-1}-accurate expansion of ordinary versus either reassigned arm.

For $B = P_{\mathrm{e}} = 4$, $T = 10$, $J_{\mathrm{ep}} = 4$, and $p = 1/2$, there are $25B + 9 = 109$ physical parameters and $a_{10}(1/2) = 1/160$. Taking $\alpha = 0$, $\varepsilon = 1$, and $C_J = 4$ permits the sufficient range $M \geq \rho^{-9}$, with $16 \mid M$, and gives

$$\begin{aligned} \mathcal{T}_O - \mathcal{T}_F &= \frac{1}{160}\rho^6 + o_{\mathbb{P}}(\rho^6), \\ \mathcal{T}_F - \mathcal{T}_C &= \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}), \\ \mathcal{T}_F - \mathcal{T}_U &= \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}). \end{aligned} \tag{412}$$

In this example, $M\mathcal{A}_D \Rightarrow \chi^2_{12}/20$.

All statements keep B, T fixed and compare methods at the same finite L; they establish no cross-depth ordering, new small-L advantage, or uniform result for growing B or T. The displayed sample range is a sufficient asymptotic condition, not a practical sample-size threshold. The chi-square convergence is distributional and does not imply a joint expected-risk limit. In particular, no expectation claim for the joint $(M, \rho, J_{\mathrm{ep}})$ regime follows here. The energy and reassignment arguments appear below and in Subsections E.13 and E.14.1.

E.14.1 Proof of the reassignment certificate

We prove the reassignment extension for the same binary HMM, all $25B + 9$ physical parameters, paired reset, Euclidean source-mean GD, and fresh-data source-average risk as in Subsection E.11. Throughout B, T are fixed, $P_{\mathrm{e}} \mid B$, $q = B/P_{\mathrm{e}}$, L belongs to a preselected finite set, and $N = E_{\mathrm{ep}}J_{\mathrm{ep}}$ lies in (306). Constants are uniform over the finitely many legal P_{e} and this preselected depth set. Write $\delta = \delta_D$ and $\mathfrak{T} = 1 + \eta N$. The improved full-space interpolation and third-order risk bound used below are proved in Subsection E.13; the present proof supplies the schedule-dependent physical second-order coefficient. In particular, its gate condition is $\delta\mathfrak{T} \leq c$, and its cubic risk remainder is $C\mathfrak{T}^2\delta^3$.

Matched assignments and the common sector. Let π_e be any preselected permutation of $\{0, \ldots, P_{\mathrm{e}} - 1\}$, independent of the sequence values. At update $n = eJ_{\mathrm{ep}} + b$, source s receives cell $D_{\pi_e(s),b}$. The fixed arm has $\pi_e(s) = s$. Both arms therefore use exactly the same union $\bigcup_d D_{d,b}$ at each update, and each sequence is used once per epoch. No permutation of the physical backbone is made. The full network still has only the cyclic source symmetry U and projector Π of Subsection E.11.4.

For a cell-law perturbation, put $\bar{\Delta}_b = {P_{\mathrm{e}}}^{-1}\sum_d \Delta_{d,b}$ and $\widetilde{\Delta}_{d,b} = \Delta_{d,b} - \bar{\Delta}_b$. At the paired population reference, group averaging the source-s gradient of any law gives the same common physical vector for every s, by cyclic equivariance. Thus the common forcing is the gradient generated by $\bar{\Delta}_b$, independent of π_e, whereas

$$\Pi h^{-,\pi}_{e,b} = 0, \qquad h^{-,\pi}_{e,b} = {P_{\mathrm{e}}}^{-1}\sum_s D\ell_s(\bar{\Theta}_{eJ_{\mathrm{ep}}+b}; \widetilde{\Delta}_{\pi_e(s),b}). \tag{413}$$

This identity uses $\sum_s \widetilde{\Delta}_{\pi_e(s),b} = 0$; it does not assert an arbitrary-permutation symmetry of the network. Since the reference Hessian commutes with Π, the first data variation decomposes into a common response, identical in both arms, and a contrast response in $\Pi^{\perp}$. The invariant terminal gradient annihilates its first-order contrast term. The invariant Hessian has no common–contrast cross term. In the second data variation, a common–contrast forcing remains in $\Pi^{\perp}$, while the pure-common contribution is identical in the two arms. The pure-contrast contribution can return to the common sector, including head and embedding coordinates; it is retained and controlled below.

Physical contrast system. For every source the derivative of a shared physical block is the sum over *all* of its occurrences,

$$D_{\theta_b} h^{(s)}_{D_{\mathrm{blk}}} = \sum_{j : i_{s,j} = b} A_{D_{\mathrm{blk}}-1} \cdots A_{j+1} G_j. \tag{414}$$

This is (331), not a derivative with independently trained copies of a repeated block. The terminal call uses $b(s) = sq - 1 \pmod B$ and has an empty suffix. Every other occurrence, including the remaining $L - 1$ uses of a terminal block, carries at least one state Jacobian. Their distinct suffix lengths sum to $O(\rho\ell^{m_\star}_\rho)$ by the geometric bound in Subsection E.11.5, uniformly in finite L. This proves (334) for $q = 1$ and $q > 1$ alike, with the four raw-coordinate features f_ν and their Gram

matrix (333). The nonterminal $B - P_{\mathrm{e}}$ blocks and the shared head/embedding are present in the error operator; they have not been discarded from GD.

Let $d_{\nu,b} \in \mathbb{R}^{P_{\mathrm{e}}}$ have entries

$$d_{\nu,b,d} = \widehat{m}_{d,b,\nu} - {P_{\mathrm{e}}}^{-1} \sum_{j} \widehat{m}_{j,b,\nu}, \qquad (\mathsf{P}_e d)_s = d_{\pi_e(s)}. \tag{415}$$

Every P_e is orthogonal and preserves the centered source subspace. The complete occurrence derivative yields the terminal forcing

$$(h^{-,\pi}_{e,b})_{b(s)} = -\frac{p\rho}{P_{\mathrm{e}}} \mathsf{F} (\mathsf{P}_e d_b)_s + \mathrm{err}_{e,b,s}, \tag{416}$$

where the full physical error, including nonterminal and endpoint coordinates, has the bound stated after (338). It holds uniformly in e and π_e because a permutation changes neither the maximum cell-law deviation nor the norm of a centered source vector. The endpoint component of the *leading* contrast forcing is zero by (413); its second-order feedback need not be zero.

On $\Pi^{\perp}$ the leading Hessian is exactly the terminal-orbit operator (335). Its four positive eigenvalues are $p^2\rho^2 v_\nu w_\nu / P_{\mathrm{e}}$ in each of the $P_{\mathrm{e}} - 1$ real source contrasts. The operator error (337) is a full-space estimate: the block–block logit second derivative carries the exact paired-head factor in (336), and all nonterminal/normal directions are included. Therefore its integrated norm is $O(\rho \ell_\rho^{m_\star})$. Likewise the forcing error in (416) contributes $O(\ell_\rho^{m_\star} \delta)$ to the terminal first variation. To see that these estimates are schedule-uniform, write the exact response by discrete Duhamel, insert the constant leading operator once, and bound each forcing vector by $C\rho\delta$. The leading stable convolution is $O(\delta/\rho)$; multiplying it by the integrated operator error gives $O(\ell_\rho^{m_\star}\delta)$, regardless of which orthogonal P_e occurs at a given epoch. The same argument controls the leading-null block modes. It does not require the P_e to commute with the physical cyclic action.

The regular-chart terminal gradient estimate for $q > 1$ is also needed. As proved in Subsection E.11.2, $u_0 = g, u_1 = gy$ exactly in that chart; holding g, y fixed makes their ν derivatives zero. The complete finite-history filter bounds the derivative of every higher lag, including repeated block occurrences, by $C\rho^2 \ell_\rho^{m_\star}$. Its coefficient scores then give $R_y, R_\nu = O(\rho^5 \ell_\rho^{m_\star})$, and the tracked fast root gives the same order for R_g. All population normal gradients vanish by the original sign, pair, and cyclic symmetries. Hence $\|g_N\| \le C\rho^5 \ell_\rho^{m_\star}$ in all physical coordinates, as in (317); no $q = 1$ shortcut is used. The pure-contrast second response retains shared-endpoint motion and obeys (329). Consequently its terminal linear risk contribution is at most $C\rho\ell_\rho^{m_\star}\delta^2$ by (330). The same bound applies for every exogenous permutation schedule because the contrast forcing norm is uniform. This is the required second-order shared-endpoint bridge.

Exact leading response and variance certificate. Set $a_\nu, r_\nu, \varpi_{\nu,b}$ as in (304) and define

$$\bar{d}_\nu^{J_{\mathrm{ep}}} = \sum_{b=0}^{J_{\mathrm{ep}}-1} \varpi_{\nu,b} d_{\nu,b}, \quad \omega_{\nu,e} = \frac{r_\nu^{E_{\mathrm{ep}}-1-e}}{\sum_{j=0}^{E_{\mathrm{ep}}-1} r_\nu^{E_{\mathrm{ep}}-1-j}}, \quad \mathsf{Q}_{\nu,\pi} = \sum_{e=0}^{E_{\mathrm{ep}}-1} \omega_{\nu,e} \mathsf{P}_e. \tag{417}$$

All weights are nonnegative and sum to one. In the normalized coordinate $f_\nu/\sqrt{v_\nu}$, the leading terminal response at source s satisfies

$$U_{\nu,s,n+1} = a_\nu U_{\nu,s,n} + \eta \frac{p\rho\sqrt{v_\nu}}{P_{\mathrm{e}}} (\mathsf{P}_e d_{\nu,b})_s, \qquad n = eJ_{\mathrm{ep}} + b. \tag{418}$$

Summing first over slots and then epochs, using $(1 - r_\nu) = (1 - a_\nu) \sum_b a_\nu^{J_{\mathrm{ep}}-1-b}$, gives the vector formula

$$U^{\pi,0}_{\nu,N} = \frac{1 - a_\nu^N}{p\rho\sqrt{v_\nu} w_\nu} \mathsf{Q}_{\nu,\pi} \bar{d}_\nu^{J_{\mathrm{ep}}}, \qquad U^{F,0}_{\nu,N} = \frac{1 - a_\nu^N}{p\rho\sqrt{v_\nu} w_\nu} \bar{d}_\nu^{J_{\mathrm{ep}}}. \tag{419}$$

The superscript 0 denotes only the constant-coefficient leading response, not a reduced training model. Inserting (419) into the physical terminal quadratic $\frac{1}{2}(U_N^-)^{\mathsf{T}} H_0^- U_N^-$ yields

$$\mathcal{B}_D^\pi = \frac{1}{2} \sum_{\nu=1}^{4} \frac{(1 - a_\nu^N)^2}{w_\nu} \left(\|\bar{d}_\nu^{J_{\mathrm{ep}}}\|^2_{P_{\mathrm{e}}} - \|\mathsf{Q}_{\nu,\pi} \bar{d}_\nu^{J_{\mathrm{ep}}}\|^2_{P_{\mathrm{e}}} \right), \qquad \|x\|^2_{P_{\mathrm{e}}} = {P_{\mathrm{e}}}^{-1} \sum_s x_s^2. \tag{420}$$

Indeed, the factor $1/P_{\mathrm{e}}$ in the physical Hessian and the sum over P_{e} terminal blocks produce precisely the source-average norm. For any vector d, expansion of a weighted variance proves

$$\|d\|_{P_{\mathrm{e}}}^2 - \|\mathsf{Q}_{\nu,\pi} d\|_{P_{\mathrm{e}}}^2 = \frac{1}{2} \sum_{e,f=0}^{E_{\mathrm{ep}}-1} \omega_{\nu,e}\omega_{\nu,f} \|\mathsf{P}_e d - \mathsf{P}_f d\|_{P_{\mathrm{e}}}^2 \geq 0. \tag{421}$$

This includes identity assignments, for which the certificate is zero, and shows why merely visiting every entrance is not a quantitative gain when the fitting weights are concentrated in one epoch. For cyclic $\pi_e(s) = s + e \pmod{P_{\mathrm{e}}}$, the source Fourier transform diagonalizes $\mathsf{Q}_{\nu,\pi}$; its squared nonzero-mode eigenvalues are $\gamma_{\nu k}$ in (305). Parseval therefore reduces (420) exactly to (307).

The common/contrast cancellations in (413), the schedule-uniform physical errors above, and the full-space risk expansion of Subsection E.13 now give

$$R_{P_{\mathrm{e}}}(\Theta_N^F) - R_{P_{\mathrm{e}}}(\Theta_N^\pi) = \mathcal{B}_D^\pi + \mathrm{Rem}_D^\pi, \qquad |\mathrm{Rem}_D^\pi| \leq C\rho\ell_\rho^{m_0}\delta^2 + C\mathfrak{T}^2\delta^3, \tag{422}$$

under $\delta\mathfrak{T} \leq c$. In particular, this is an actual trained-risk expansion, not only a linear-response surrogate. A dataset with $\mathcal{B}_D^\pi$ larger than the deterministic right side is certified to have a positive gap. A high-probability event that bounds δ can be intersected with this observed threshold; it does not give conditional coverage after selecting datasets by the threshold.

Random epochs and fast-epoch replacement. For the random protocol take $\mathsf{P}_0 = I$ and let P_e, $e \geq 1$, be independent uniform permutations, independent of the data. Every centered d has $\mathbb{E}_\pi[\mathsf{P}_e d] = 0$ for $e \geq 1$. Independence makes all cross terms in the squared weighted sum vanish, including those between epoch zero and later epochs. Orthogonality of each P_e gives the exact conditional identity

$$\mathbb{E}_\pi\|\mathsf{Q}_{\nu,\pi} d\|_{P_{\mathrm{e}}}^2 = \vartheta_{E_{\mathrm{ep}}}(r_\nu)\|d\|_{P_{\mathrm{e}}}^2, \quad \vartheta_{E_{\mathrm{ep}}}(r) = \frac{\sum_{j=0}^{E_{\mathrm{ep}}-1} r^{2j}}{(\sum_{j=0}^{E_{\mathrm{ep}}-1} r^j)^2} = \frac{1-r}{1+r}\frac{1+r^{E_{\mathrm{ep}}}}{1-r^{E_{\mathrm{ep}}}}. \tag{423}$$

Thus conditional on the cells, $\mathbb{E}_\pi\mathcal{B}_D^U$ equals (420) with the second squared norm replaced by $\vartheta_{E_{\mathrm{ep}}}(r_\nu)\|\bar{d}_\nu^{J_{\mathrm{ep}}}\|_{P_{\mathrm{e}}}^2$. The first epoch need not be randomized; its deterministic cross term still vanishes.

Let $0 \leq a < 1$, $2 \leq J_{\mathrm{ep}} \leq C_{J_{\mathrm{ep}}}\rho^{-a}$ and $M \geq \rho^{-(8+3a+\varepsilon)}$, with $P_{\mathrm{e}}J_{\mathrm{ep}} \mid M$ and fixed $\varepsilon > 0$. In the window (306), $a_\nu^N \leq \rho^2$ and $1 - r_\nu = O(\eta\rho^2 J_{\mathrm{ep}})$. Hence

$$\vartheta_{E_{\mathrm{ep}}}(r_\nu) = O(\eta\rho^2 J_{\mathrm{ep}}), \qquad \max_{k\neq 0}\gamma_{\nu k} = O((\eta\rho^2 J_{\mathrm{ep}})^2). \tag{424}$$

The second estimate is the cyclic finite-geometric-series bound from Subsection E.11.7; $1 - r_\nu^{E_{\mathrm{ep}}} = 1 - a_\nu^N$ stays away from zero. The normalized slot weights satisfy $\sum_b |\varpi_{\nu,b} - {J_{\mathrm{ep}}}^{-1}| = O(\eta\rho^2 J_{\mathrm{ep}})$. Consequently replacing $\bar{d}_\nu^{J_{\mathrm{ep}}}$ by the centered whole-shard mean and $(1 - a_\nu^N)^2$ by one changes the unsuppressed quadratic by at most

$$C\{\rho^2 + \eta\rho^2 J_{\mathrm{ep}}\}\delta^2. \tag{425}$$

The cyclic suppression costs at most $C(\eta\rho^2 J_{\mathrm{ep}})^2\delta^2$. For random epochs, conditional Markov applied to (423) shows that the suppressed quadratic is $o_{\mathbb{P}}(M^{-1})$: the unsuppressed whole-shard quadratic is $O_{\mathbb{P}}(M^{-1})$, the slot-flatness error is $o_{\mathbb{P}}(M^{-1})$, and $\vartheta_{E_{\mathrm{ep}}}(r_\nu) \to 0$. This probability is joint over data and random permutations; no pathwise uniform suppression over all permutation sequences is claimed.

For completeness, complete-sequence cell concentration gives $\delta^2 = O_{\mathbb{P}}(J_{\mathrm{ep}}\log(2P_{\mathrm{e}}J_{\mathrm{ep}})/M)$ as in (347). The gate condition holds because $\mathfrak{T}\delta = O_{\mathbb{P}}(\rho^{-2}\ell_\rho\sqrt{J_{\mathrm{ep}}\log(2P_{\mathrm{e}}J_{\mathrm{ep}})/M}) = o_{\mathbb{P}}(1)$. Multiplying the two terms in (422) by M gives respectively

$$\begin{aligned} M\rho\ell_\rho^{m_0}\delta^2 &= O_{\mathbb{P}}(\rho J_{\mathrm{ep}}\ell_\rho^{m_0}\log(2P_{\mathrm{e}}J_{\mathrm{ep}})) = o_{\mathbb{P}}(1), \\ M\mathfrak{T}^2\delta^3 &= O_{\mathbb{P}}\left(\frac{\rho^{-4}\ell_\rho^2 {J_{\mathrm{ep}}}^{3/2}\log^{3/2}(2P_{\mathrm{e}}J_{\mathrm{ep}})}{\sqrt{M}}\right) = o_{\mathbb{P}}(1). \end{aligned} \tag{426}$$

The second line is $O(\rho^{\varepsilon/2}\,\mathrm{polylog}(1/\rho))$; the first uses $a < 1$. Likewise, M times (425) is $O_{\mathbb{P}}((\rho^2 + \eta\rho^2 J_{\mathrm{ep}})J_{\mathrm{ep}}\log(2P_{\mathrm{e}}J_{\mathrm{ep}})) = o_{\mathbb{P}}(1)$, because $a < 1$. The same estimates apply to a preselected finite depth set, without asserting uniformity for growing L.

It follows that both cyclic C and random U assignments satisfy

$$R_{P_e}(\Theta_N^F) - R_{P_e}(\Theta_N^C) = \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}), \qquad R_{P_e}(\Theta_N^F) - R_{P_e}(\Theta_N^U) = \mathcal{A}_D + o_{\mathbb{P}}(M^{-1}). \quad (427)$$

The triangular-array complete-sequence CLT in (344)–(345) gives $M\mathcal{A}_D \Rightarrow \chi^2_{4(P_e-1)}/(2T)$. For fixed $P_e > 1$ this limit has no atom at zero. A union bound on the two $o_{\mathbb{P}}(M^{-1})$ errors therefore yields

$$\Pr\{R_F > \max(R_C, R_U)\} \longrightarrow 1, \qquad R_C - R_U = o_{\mathbb{P}}(M^{-1}). \quad (428)$$

The two reassignments are not strictly ordered by this result. Combining the $P_e = B$ case with the ordinary-to-fixed comparison in Subsection E.13 gives the four-arm statement at its separate $o_{\mathbb{P}}(\rho^6)$ ordinary margin. This does not produce an M^{-1}-accurate ordinary-to-reassignment expansion or a new small-L ordering. Finally, weak convergence in (427) gives no joint expectation statement; the fixed-J_{ep} sequential expectation argument remains a separate quantifier order as in (348).