TRAINING DYNAMICS OF LARGE STATE-TRANSITION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Jointly learning input and output embeddings is common practice in modern neural sequence models, yet it remains unclear when gradient-based training can successfully optimize these coupled components to achieve accurate predictions. We study this question in a softmax state-transition model with a learned state embedding and prediction head, trained from Gaussian initialization using exact population gradients. We consider a model with V states and embedding dimension $d < V$, where the ground-truth transition law's row-centered log-probability table has low rank and bounded entries. We measure prediction quality by relative prediction gain (RPG), normalized so that uniform prediction has gain zero and optimal prediction has gain one. We provide analysis for both untied and tied models. With an untied model and weight decay of order $1/V$, we show that perturbed gradient descent achieves RPG above $7/8$ in polynomial time when $d \geq C\sqrt{V}$. With probability at least $1 - e^{-cd}$ where $c > 0$, this holds at every iterate in the final $1 - 1/d$ fraction of the run while perturbations remain at a fixed scale. For tied embeddings, we prove training guarantees on the observed current states, and under the assumption of ground-truth feature coverage and bounded learned rows, we prove prediction guarantees on unseen current states. We present numerical experiments with practical parameter choices that qualitatively illustrate persistent training gain and tied-model transfer.

1 INTRODUCTION

Understanding why gradient-based training produces useful neural networks remains a longstanding challenge. We study this through learning a state-transition law $P(s' \mid s)$ on V states. Storing the full table requires V^2 entries. Low-dimensional state embeddings $\phi(s), \psi(s') \in \mathbb{R}^d$ instead predict $\widehat{P}(s' \mid s) \propto \exp\big(\phi(s)^\top \psi(s')\big)$, with normalization over s'. This uses $2Vd$ parameters, or Vd when the input and output embeddings are tied, meaning $\phi = \psi$ (Press & Wolf, 2017; Inan et al., 2017). For $d < V$, this substantially reduces storage. Equivalently, this is a next-token prediction model with a one-token context, illustrating how learned embeddings handle a large discrete input space.

Distributed representations have a long history in neural language modeling (Bengio et al., 2000). Exponential-family models provide a related low-rank view of categorical data, with dedicated global optimization algorithms (Guo & Schuurmans, 2008). Nonconvex optimization theory has identified favorable low-rank loss landscapes (Ge et al., 2017) and developed gradient-based methods for escaping saddles (Jin et al., 2019). Recent theory also studies next-token training in linear and attention models (Li et al., 2024; Huang et al., 2024; Thrampoulidis, 2024). Our interest is in why joint gradient-based training can learn a good predictor while retaining a low-dimensional factorization, ordinary Gaussian initialization, cross-entropy loss, and either tied or untied embeddings. The ground-truth law may be any member of a full-support, fixed-rank, bounded-logit class with a nonvanishing information gap. The tied results additionally require a Gram representation. Our main question is:

What are the training dynamics of state-transition models with a large state space and cross-entropy loss?

To answer this question, we must specify what successful training means. A stationary point need not give good predictions, because even a related binary Markov-transformer model can converge to local or global minima depending on the data and initialization (Makkuva et al., 2024). Conversely,

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.

a model can be useful without attaining the optimal loss. We therefore measure prediction quality by relative prediction gain (RPG): the fraction of the uniform predictor's excess loss removed by training, with the true transition law as the reference optimum. Uniform prediction has gain zero, the true law has gain one, and gain $7/8$ means that training removes seven eighths of the uniform predictor's excess cross-entropy. RPG measures prediction loss. It enables us to study both training progress and prediction on unseen current states. In the tied model, an unseen state's embedding is still updated through its role as an output label, providing a mechanism for transfer.

Our contributions are as follows.

- **Untied training: high gain with sublinear width.** For every target in our full-support, fixed-rank class, we jointly train both embedding factors from Gaussian initialization using cross-entropy, weight decay, and fixed-step perturbed gradient descent. Already at width $d \geq C\sqrt{V} < V$, prediction gain exceeds $7/8$ with probability at least $1 - e^{-cd}$ at *every* iterate in $[T/d, T]$ for a predeclared polynomial horizon T. Thus the guarantee concerns the training trajectory itself, not merely a favorable stationary point or a selected checkpoint.
- **Tied training and transfer.** When only $N = \Theta(V) < V$ current states are observed, jointly trained tied embeddings achieve observed-state gain $1 - O(V/d^2)$ throughout the same final window at comparable sublinear width. Tying also updates unseen-state embeddings through their role as output labels. With a stronger width and decay choice, centered coverage, and a fixed unseen-state information gap, we obtain gain $1 - O\big(V^4/[Nd^3(V-N)]\big)$ on the *entire* unseen set at every final-window iterate whose learned rows are uniformly bounded.
- **Proof technique.** We note that the softmax cross-entropy has limited total curvature across logit directions. By rotating the low-rank target among embedding directions, we find one where that curvature is small. In the region reached by training, a low-gain point with a small gradient must therefore have a direction of negative curvature. We then control the training steps between repeated visits to approximate stationary points, extending high gain from those visits to most training steps.

We also conduct synthetic experiments to show that high training gain persists over long runs and that tied embeddings can also predict well on unseen states. These patterns remain visible under sampled-transition training, although our theorem does not cover that setting.

2 Related Work

Low-rank softmax optimization. The softmax bottleneck concerns the representational capacity of low-rank logits (Yang et al., 2018; Ganea et al., 2019). We assume a realizable, fixed-rank ground truth and ask whether training reaches useful predictions. Global-landscape results for deep linear networks (Laurent & von Brecht, 2018), low-rank recovery (Ge et al., 2017), and factored nuclear-norm regularization (Li et al., 2017) use width or loss-curvature conditions specific to their settings. Ouyang et al. (2025) give more general connections between factor second-order stationarity and the nuclear-norm-regularized problem. Guo & Schuurmans (2008) study low-rank exponential-family models, including multinomial data, through a global-optimization reformulation rather than the dynamics of the original factor updates. For our loss, we use its Hessian trace to certify *prediction gain* at localized approximate stationary points. We do not claim that all local minima are global. A direct uniform-curvature comparison and the trace-based refinement are given in Appendix B.4. Neural-collapse analysis also studies jointly learned features and classifiers under cross-entropy and decay, but with one-hot targets and a different feature-width regime (Zhu et al., 2021).

Perturbations and weight tying. Perturbation methods can reach approximate second-order stationary points in nonconvex objectives (Ge et al., 2015; Jin et al., 2019). Closer to factor training, Liu et al. (2021) analyze noisy gradient descent for matrix factorization under squared loss, with a finite-window guarantee in the rank-one case. Their noise perturbs the factors before gradient evaluation, and their analysis does not address categorical prediction gain or transfer to unseen states. We use the always-perturbed algorithm and descent concentration bound of Jin et al. (2019); our model-specific steps turn repeated visits into gain for every iterate in a final time window. Weight tying shares input and output embeddings in language models (Press & Wolf, 2017; Inan et al., 2017). Our tied training bound needs no coverage assumption. The unseen-state bound additionally requires centered coverage and, for the complete unseen set, a bound on the learned embedding rows.

Next-token training. We also include next-token prediction literature as a more general case. Related theory studies attention-based next-token prediction (Li et al., 2024; Huang et al., 2024), linear predictors with fixed context embeddings (Thrampoulidis, 2024), and structured language or binary Markov tasks (Huang et al., 2025; Varre et al., 2025; Makkuva et al., 2024). Some of these models include attention and longer contexts, but their guarantees do not cover the combination studied here: $d < V$, jointly learned input and output embeddings, Gaussian initialization, and a fixed-rank target distribution. We prove finite-time prediction gain throughout a final window of joint factor training. A more detailed comparison is listed in Table 1.

Table 1: Ingredients explicitly analyzed in representative results: factorized optimization (Li et al., 2017; Ouyang et al., 2025), noisy matrix factorization (Liu et al., 2021), neural collapse (Zhu et al., 2021), weight tying (Press & Wolf, 2017; Inan et al., 2017), binary Markov transformers (Makkuva et al., 2024), and perturbed optimization (Jin et al., 2019). "Partial" means a related ingredient is analyzed in a different setting: weight tying and binary Markov transformers analyze related factorized losses, while noisy matrix factorization perturbs factors before gradient evaluation rather than adding noise to the gradient. "No" means the ingredient is not part of the cited result. The last row summarizes our untied and tied theorems together.

Work	Factorized loss analysis	Tied input/output	Full-support transition law	Persistent-noise analysis	Gain on unseen states
Factorized optimization	Yes	No	No	No	No
Noisy matrix factorization	Yes	No	No	Partial	No
Neural collapse	Yes	No	No	No	No
Weight tying	Partial	Yes	No	No	No
Binary Markov transformers	Partial	Yes	Yes	No	No
Perturbed optimization	No	No	No	Yes	No
This work	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

3 Problem Setup

Let $[V] = \{1, \dots, V\}$ be the state space. The ground-truth transition table $Q \in \mathbb{R}^{V \times V}$ has all positive entries and row sums all equal 1; row $q_i^\top$ is the distribution of the next state given current state i. All softmax operations below act row-wise. The untied model predicts $P^{E,U} = \mathrm{softmax}(EU^\top)$, with $E, U \in \mathbb{R}^{V \times d}$. The tied model predicts $P^X = \mathrm{softmax}(XX^\top)$, with $X \in \mathbb{R}^{V \times d}$. We write $\Theta = (E, U)$ or $\Theta = X$ accordingly. Thus the numbers of trainable parameters are $2Vd$ and Vd. The dimension $d < V$ is the actual embedding dimension throughout the analysis. Sets use calligraphic letters, except for standard spaces and $[V]$; matrices use uppercase letters. Dimensions and fixed bounds retain their conventional notation.

For a nonempty set $\mathcal{A} \subseteq [V]$ of current states, define population and Bayes (optimal) risk:

$$\mathcal{R}_{\mathcal{A}}(\Theta) = -\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \sum_{j=1}^{V} q_{ij} \log P_{ij}^{\Theta}, \qquad \mathcal{H}_{\mathcal{A}} = -\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \sum_{j=1}^{V} q_{ij} \log q_{ij}.$$

While we call $\mathcal{R}_{\mathcal{A}}(\Theta)$ population risk, we treat Q as a fixed target transition table, which need not be interpreted as an underlying data-generating law. The risk $\mathcal{R}_{\mathcal{A}}$ uses uniform weights over the current states in $\mathcal{A}$. It also describes full-batch training on a dataset-induced transition table with those same weights, whenever that table satisfies our assumptions on Q.

Definition 3.1 (Relative prediction gain (RPG))**.** When $\log V - \mathcal{H}_{\mathcal{A}} > 0$, define

$$g_{\mathcal{A}}(\Theta) = \frac{\log V - \mathcal{R}_{\mathcal{A}}(\Theta)}{\log V - \mathcal{H}_{\mathcal{A}}} = 1 - \frac{|\mathcal{A}|^{-1} \sum_{i \in \mathcal{A}} D_{\mathrm{KL}}(q_i \| P_{i,:}^{\Theta})}{\log V - \mathcal{H}_{\mathcal{A}}}. \tag{1}$$

The uniform predictor has gain zero and the Bayes predictor has gain one. For example, gain $7/8$ means that training removes at least seven eighths of the uniform predictor's excess risk.

Ground-truth class. Since adding the same constant to every logit in a row leaves its softmax unchanged, define the ground truth's row-centered log-probability table by $L(Q) = \log Q - \frac{1}{V}(\log Q)\mathbf{1}\mathbf{1}^\top$. We consider the class

$$\mathcal{Q}_{r,B,\Delta_0} = \left\{ Q : \begin{array}{l} Q_{ij} > 0,\ Q\mathbf{1} = \mathbf{1},\ \operatorname{rank} L(Q) \le r, \\ \|L(Q)\|_\infty \le B,\ \log V - \mathcal{H}_{[V]} \ge \Delta_0 \end{array} \right\}.$$

Here r is a fixed positive integer and $B, \Delta_0 > 0$ are fixed independently of V. The matrix infinity norm is the maximum absolute entry. We note that centering identifies the rank relevant to softmax realizability, and it does not change the model or the risk. When $r \le d$, the ground truth is representable by the untied model. Bounded logits permit nonuniform next-state laws, but keep each probability positive. The information floor excludes an asymptotically uniform prediction problem.

Training recipe. All entries of Θ_0 are independent $\mathcal{N}(0, 1/d)$ variables. We train all entries simultaneously on the weight-decayed objective and use perturbed gradient descent (Jin et al., 2019) as follows.

$$\mathcal{F}_{\lambda,\mathcal{A}}(\Theta) = \mathcal{R}_{\mathcal{A}}(\Theta) + \frac{\lambda}{2}\|\Theta\|_{\mathrm{F}}^2, \tag{2}$$

$$\Theta_{t+1} = \Theta_t - h(\nabla \mathcal{F}_{\lambda,\mathcal{A}}(\Theta_t) + \xi_t), \qquad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{D_\Theta} I_{D_\Theta}\right). \tag{3}$$

Here D_Θ is the number of parameters and $\|(E,U)\|_{\mathrm{F}}^2 = \|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2$. The perturbations are independent of initialization. The decay λ, step size h, noise scale σ, and maximum step T are deterministic choices made before initialization. The step and noise scale stay fixed throughout the run. Unless stated otherwise, constants in our bounds may depend on the fixed assumptions of the corresponding theorem, but are uniform in V, N, d, and T.

Technical challenges. Although the model is simple, analyzing it while retaining the basic ingredients of practical training is not. The difficulty is their combination: $d < V$, *cross-entropy loss, ordinary Gaussian initialization, and simultaneous gradient-based training of all factors*. Specifically, (i) Low-dimensional factors make optimization nonconvex. (ii) The cross-entropy Hessian depends on the learned probabilities and can become nearly singular. (iii) Random initialization provides no prescribed alignment or local-convergence guarantee. Therefore, analyses based on fixed embeddings, staged training, or convex logit reformulations (Thrampoulidis, 2024; Huang et al., 2024; Zhao et al., 2024) do not directly establish a finite-time prediction guarantee under this combination. Moreover, generic saddle escape analysis (Jin et al., 2019) is not automatically a prediction guarantee. A saddle-based analysis must still quantitatively relate stationarity to low prediction loss in the region actually reached by training, and control the intervening iterates. We establish this connection for fixed-step perturbed GD, with noise present at every update. Thus the relevant comparison is a finite-time, high-probability prediction guarantee under these joint conditions, not merely a landscape characterization or an asymptotic convergence result for a related model.

4 Training Gain of Untied Model

We first train the untied model on all current states, $\mathcal{A} = [V]$. Our main result gives a finite-time guarantee for an entire interval of the original perturbed-gradient trajectory.

We are interested in whether an embedding dimension much smaller than V can support useful prediction over many consecutive training steps, despite the nonconvex factorization and perturbations at every update. The theorem below gives a finite-time answer: with high probability, the untied model removes more than seven eighths of the prediction-loss gap between uniform prediction and the ground truth, averaged over current states, at every iterate after the first $1/d$ fraction of a polynomial-length run. Note that the guarantee is not limited to a selected checkpoint, but it does not assert convergence or persistence beyond the stated horizon.

Theorem 4.1 (Untied constant gain)**.** *Fix r, B, Δ_0. There are positive constants c, C, V_0 such that, for every $V \ge V_0$, $Q \in \mathcal{Q}_{r,B,\Delta_0}$, and $\max\{2r, C\sqrt{V}\} \le d < V - 1$, setting $\lambda = \frac{\Delta_0}{8B\sqrt{r}\,V}$ allows deterministic choices of $h, \sigma > 0$ and an integer T divisible by d for which*

$$\Pr\left\{ g_{[V]}(\Theta_t) > \frac{7}{8} \text{ for every integer } t \in [T/d, T] \right\} \ge 1 - e^{-cd}.$$

The quantities h^{-1}, σ^{-1}, and T are bounded by polynomials in V. The probability is over both initialization and perturbations.

4.1 Proof Overview

The proof contains three main steps. The key idea is to use a trace bound on categorical curvature to show that every sufficiently accurate approximate second-order stationary point in a localized region has high prediction gain. We first establish that the training trajectory remains in this region. We then apply the landscape argument at repeated approximate-SOSP visits and propagate the resulting gain guarantee to the intervening iterates. Appendix B gives the parameter choices and complete proof.

Definition 4.2 (Approximate SOSP)**.** Let f be an objective whose Hessian is ρ-Lipschitz on the region under consideration. We call Θ an ϵ-approximate second-order stationary point (SOSP) of f if

$$\|\nabla f(\Theta)\| \le \epsilon, \qquad \nabla^2 f(\Theta) \succeq -\sqrt{\rho\epsilon}\, I.$$

Here $f = \mathcal{F}_{\lambda,[V]}$ is the untied regularized objective.

Step 1: Localize the trajectory. Positive weight decay creates a region where the squared parameter norm is at most a constant multiple of V. Leaving this region requires a substantial increase in the objective Gaussian initialization lies inside this region with high probability (lemma B.1). A concentration bound for the noisy updates shows that the objective cannot rise by more than a fixed small amount during each block of updates . When combined with Step 3, this bound shows that, with high probability, no iterate reaches the objective level required to leave the region before time T.

Step 2: Convert approximate stationarity into gain. We prove by contradiction. Consider a point in the localized region that has gain at most $7/8$ and a small gradient. Assume this point is an accurate approximate SOSP, we compare it with factors of slightly shrunken ground-truth logits whose total squared norm is $O(V)$. lemmas D.1 and B.2 shows that this comparator improves the regularized objective by a fixed margin. Meanwhile, the Hessian of each categorical row loss has trace at most one, so the total curvature budget is at most V. Rotating the low-rank comparator among the d embedding directions then produces a factor direction whose softmax-curvature cost is $O(V/d^2 + 1/d)$ (lemma B.3).

Since cross-entropy is a convex function of the logits, together with the Hessian identity, it converts the fixed objective excess into a negative second-order contribution along this direction. The small-gradient assumption controls the remaining radial error (lemmas D.3 and B.4). When $d \ge C\sqrt{V}$, the negative contribution dominates both this error and the softmax-curvature cost. Along the resulting direction, the normalized Hessian curvature is at most $-c/V$ for some fixed $c > 0$. The relevant Hessian-Lipschitz bound is $O(\sqrt{V})$, so we choose the SOSP accuracy ϵ to be a sufficiently small constant times $V^{-5/2}$. Then $\sqrt{\rho\epsilon} \le c/(2V)$: an ϵ-approximate SOSP would have curvature at least $-c/(2V)$ in every direction, contradicting the direction just found. Thus, a low-gain point with a small gradient cannot be an accurate approximate SOSP.

Step 3: Refresh the gain throughout the final window. We partition T into d consecutive time blocks. Applying the perturbed-gradient result of Jin et al. (2019) to each block shows that, with high probability, every block contains an iterate that is an ϵ-approximate SOSP. At every such visit, Step 2 places the objective below the gain threshold by a fixed margin. A within-block objective-increase bound preserves enough of this margin until the next visit, where the margin is then restored. Induction over the blocks therefore gives gain above $7/8$ for every integer $t \in [T/d, T]$.This block decomposition is an analytical device used to organize the proof.

Remark: The proof shows that low-gain points with small gradient have a direction of negative curvature. It does not classify critical points whose gain is already above the target. In particular, theorem 4.1 does not assert that all local minima are global.

5 Tied Model Training and Conditional Transfer

Let $\mathcal{S} \subset [V]$ be the observed current states, with $N = |\mathcal{S}| = \Theta(V)$, and let $\mathcal{U} = [V] \setminus \mathcal{S}$ be the full unseen set. We assume $d < N < V$. Every state remains a possible output label during training,

including the states in $\mathcal{U}$. Hence, tying causes the embedding of an unseen input state to be updated through its output role. For an untied model, the corresponding input row does not receive a loss gradient from $\mathcal{R}_{\mathcal{S}}$. We establish two results: a training guarantee on the observed states and, under additional conditions, a prediction guarantee on the complete unseen set.

5.1 Tied Model Training

Definition 5.1 (Tied-compatible base class)**.** For fixed $B_Y \geq 1$, let $\mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ consist of tables $Q \in \mathcal{Q}_{r,B,\Delta_0}$ that admit $Q = \mathrm{softmax}(YY^\top)$ for some $Y \in \mathbb{R}^{V \times r_Y}$ with $r_Y \leq r$ and $\max_i \|y_i\|_2^2 \leq B_Y$, where y_i is the column-vector form of row i of Y.

Tying the input and output embeddings restricts which transition laws the model can represent. For the class above, the next theorem shows that training achieves a gain close to one on the observed states throughout a long final portion of the run. The gap from gain one scales as V/d^2, so the guarantee strengthens as the embedding width grows beyond $\Omega(\sqrt{V})$.

Theorem 5.2 (Tied training)**.** *Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ and suppose $\log V - \mathcal{H}_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$. For sufficiently large V and $\max\{2r, C\sqrt{V}\} \leq d < N$, there is a choice $\lambda = C_\lambda/d^2$ and fixed $h, \sigma > 0$ with polynomially bounded inverses, together with a polynomial run length T divisible by d, such that, with probability at least $1 - e^{-cd}$, $g_{\mathcal{S}}(X_t) \geq 1 - \frac{CV}{d^2 \delta_{\mathrm{tr}}}$, for every integer $t \in [T/d, T]$.*

5.1.1 Proof Overview

The proof follows the same three-step structure as the untied argument in section 4. Localize the trajectory, obtain a loss guarantee at approximate SOSPs, then propagate this guarantee throughout the final window. The key difference is that the weight decay absorbs an energy-dependent curvature term arising from the tied factorization.

Gaussian initialization bounds the initial objective. Since cross entropy is bounded below by $\mathcal{H}_{\mathcal{S}}$, the penalty creates an objective barrier outside a deterministic ball. Modifying the objective outside this ball provides the global smoothness bounds required by the fixed-noise argument. The objective-increase control keeps the modified and original training trajectory inside the ball and identical under the same noise lemma C.1).

We express $K = XX^\top$ and $K^\star = YY^\top$ for the learned and ground-truth Gram matrices. A tied-specific packing argument constructs a direction Z with $ZZ^\top = K^\star$ and softmax-curvature cost $O(\|X\|_{\mathrm{F}}^2/d^2)$ (lemma C.2). It places the ground-truth features in directions with both small embedding energy and small categorical curvature, using sign averaging to cancel cross terms.

At a sufficiently accurate approximate SOSP, Hessian testing and convexity in the Gram matrix bound the regularized objective by its ground-truth value plus this energy-dependent error and small stationarity errors. Choosing $\lambda = C_\lambda/d^2$ with C_λ sufficiently large lets the penalty absorb the energy-dependent term. Since the ground-truth energy is $O(V)$, this yields $\|X\|_{\mathrm{F}}^2 = O(V)$ and training excess loss $O(V/d^2)$.

Finally, the fixed-noise argument supplies an accurate SOSP visit in each block of a single continuous run, and also gives a control of objective increases within the block. Each visit restores a margin below the target objective threshold, which maintains the bound until the next visit. The loss guarantee therefore holds at every iterate in the final window, rather than only at selected SOSPs. Dividing by the training information gap gives theorem 5.2. Appendix C.1 provides the complete argument.

5.2 Conditional Transfer to Unseen States

Training on $\mathcal{S}$ does not by itself determine predictions on $\mathcal{U}$. We therefore impose a coverage condition on the ground-truth representation.

Definition 5.3 (Coverage for transfer)**.** A representative Y from definition 5.1 satisfies centered coverage over $\mathcal{S}$ if it has linearly independent columns and for some fixed $\kappa > 0$, we have

$$Y^\top \mathbf{1} = 0, \qquad \frac{1}{N} \sum_{i \in \mathcal{S}} (y_i - \bar{y}_{\mathcal{S}})(y_i - \bar{y}_{\mathcal{S}})^\top \succeq \kappa^2 I, \qquad \bar{y}_{\mathcal{S}} = \frac{1}{N} \sum_{i \in \mathcal{S}} y_i.$$

The coverage condition says that the observed ground-truth embeddings capture every ground-truth feature direction. This lets information from the observed rows constrain predictions when unseen states appear as inputs. The theorem below uses a decay schedule chosen for this transfer argument and a stronger width requirement than tied training alone. It bounds gain on the observed states throughout the final window and, at any of those iterates with uniformly bounded learned embeddings, also bounds gain on the complete unseen set. The learned-embedding bound is an additional condition: the optimization analysis does not establish it, so the unseen-state conclusion is conditional.

Theorem 5.4 (Conditional tied transfer)**.** *Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ admit a representative Y satisfying definition 5.3. Suppose $\log V - \mathcal{H}_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$, and $\log V - \mathcal{H}_{\mathcal{U}} \geq \delta_{\mathrm{te}} > 0$. For sufficiently large V, let $\max\{2r, CV/N^{1/3}\} \leq d < N$, and $\lambda = C_\lambda \sqrt{\frac{V}{Nd^3}}$. There are deterministic h, σ, T of polynomial size, with polynomially bounded inverse scales, and an event of probability at least $1 - e^{-cd}$ on which every integer $t \in [T/d, T]$ satisfies*

$$g_{\mathcal{S}}(X_t) \geq 1 - \frac{CV^3}{Nd^3\delta_{\mathrm{tr}}}.$$

Moreover, at every such iterate satisfying $\max_{j\in[V]} \|x_j(t)\|_2^2 \leq R_0$ for a fixed R_0, the complete unseen set satisfies

$$g_{\mathcal{U}}(X_t) \geq 1 - \frac{CV^4}{Nd^3(V-N)\delta_{\mathrm{te}}}. \tag{4}$$

5.2.1 Proof Overview

The proof contains three main steps. The key idea is to use positive-semidefinite completion to extend control of the Gram matrix on the observed rows to the entire matrix. First, training gives an objective comparison with the ground truth for the entire final window. Second, coverage turns this comparison into training-loss and trace estimates. Finally, under the learned-row bound, these two forms of control determine the full Gram matrix and yield the unseen-state gain bound. Appendix C.2 gives the parameter choices and complete proof.

Step 1: Obtain a final window objective comparison. An approximate SOSP provides only local derivative information, whereas the transfer argument requires quantitative control relative to the ground truth. Let $\mu = \lambda V$ denote the effective regularization scale for a factor with $O(V)$ energy. Under the transfer decay schedule, μ^2 is of order $V^3/(Nd^3)$. The tied packing argument compares every accurate approximate SOSP with the ground-truth Gram matrix while incurring a curvature error proportional to $\|X\|_{\mathrm{F}}^2/d^2$ (lemma C.2). The chosen weight decay absorbs this energy-dependent error, and the remaining approximation terms are of order μ^2. The localization, repeated-visit, and refresh arguments then give the same regularized-objective comparison for every iterate in the final window (lemma D.7). This is a comparison of regularized objectives; converting it into a prediction-loss bound requires the coverage argument in Step 2.

Step 2: Convert the objective comparison into observed control. A regularized-objective comparison does not immediately bound the prediction loss, because an increase in loss could be offset by a decrease in factor energy. The centered coverage condition in definition 5.3 ensures that every ground-truth feature direction is sufficiently represented among the observed states. This yields a trace certificate that controls changes in factor energy through the error in the observed logits (lemma C.6). Combining this certificate with the objective comparison gives training excess loss of order μ^2 and controls the trace of the learned Gram matrix relative to the ground truth (lemma C.7). These are the two estimates needed for the completion argument in Step 3.

Step 3: Transfer the observed control to unseen states. The ground-truth logits are bounded by assumption, and the uniform bound on the learned embeddings also keeps the learned logits bounded. For each observed row, consider the straight-line interpolation in logit space between the ground-truth and learned logits. Both endpoints, and hence every intermediate logit vector, have uniformly bounded coordinates. Its softmax probabilities are therefore between c/V and C/V, so the categorical Hessian has curvature comparable to $1/V$ on the centered logit space. The lower curvature bound converts the training-loss estimate from Step 2 into control of the Gram-matrix error on the observed rows.

The observed rows alone do not determine an arbitrary learned Gram matrix. The completion argument decomposes the remaining error into a part controlled by the observed rows and a positive-semidefinite residual. The Frobenius norm of this residual is bounded by its trace, which is controlled in Step 2. These estimates together control the full Gram matrix (lemma C.8). The corresponding upper curvature bound then converts the full Gram-matrix estimate into a prediction-loss bound on the unseen rows. Substituting the transfer decay scale and dividing by the training and test information gaps gives the two gain bounds in theorem 5.4.

6 Numerical Experiments

We present two complementary experimental studies. Population-gradient experiments examine persistent training gain, tied-model transfer, and the learned-row condition used in our transfer result. A larger sampled-transition SGD sweep tests whether these qualitative behaviors persist under stochastic training and across larger state spaces, teacher ranks, and embedding widths. Both studies are qualitative stress tests rather than direct verifications of the theorem's asymptotic constants.

6.1 Population-Gradient Training and Transfer

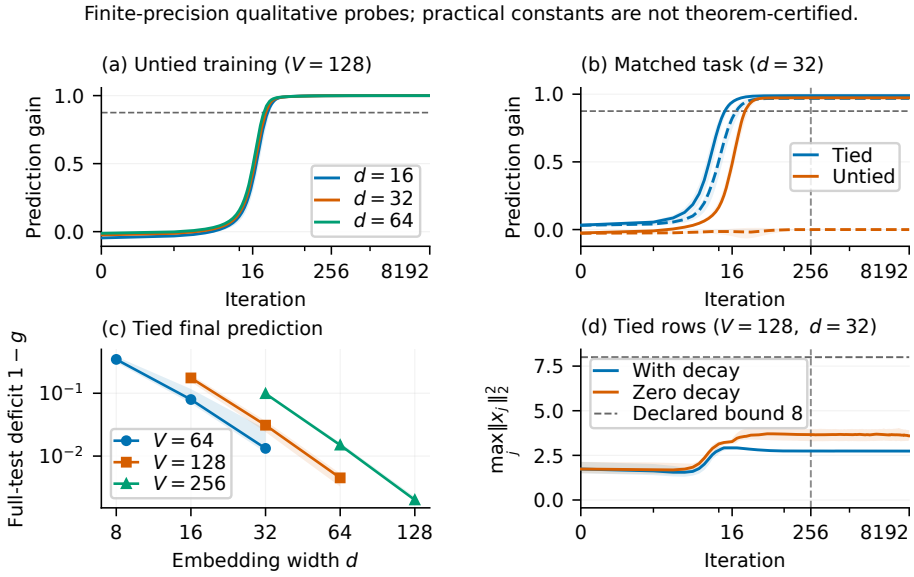

Figure 1: Training and transfer in the finite-precision experiments. Curves show medians and shaded full ranges over three ground truths and five initializations each. (a) Untied training gain at $V = 128$. (b) Tied and untied models on the same observed/unseen split at $V = 128, d = 32$ (solid: training; dashed: test). (c) Final tied test deficit ($1-$ gain) as width and its prescribed decay change together. (d) Maximum tied embedding-row norm squared. The vertical lines in (b,d) mark T/d, not a change to the updates. No trajectories are smoothed; statistics use every update.

We use ground truths with rank-four centered logits bounded by 3, $V \in \{64, 128, 256\}$, and $d \in \{V/8, V/4, V/2\}$. Untied training uses all input states; tied training uses $N = 3V/4$ states. Every run uses Gaussian initialization, full population gradients evaluated in FP32, fixed step $h = V/4$, and $T = 8192$ updates. The perturbation scale is fixed at $\sigma = 0.01/\sqrt{V}$. We report all 375 runs, including paired no-noise and no-decay controls. Appendix E gives the full protocol, data, and checks.

Persistent gain and tied transfer. All 135 full-untied main runs maintain gain above $7/8$ at every $t \in [T/d, T]$; the minimum over these runs and times is 0.9951. For tied training, 120 of 135 main runs remain above the same $7/8$ reference level. The exception is the narrowest setting, $V = 64, d = 8$, whose final training gains range from 0.8610 to 0.8753. On the matched split in

figure 1(b), final tied test gains range from 0.9638 to 0.9758, while the untied model's unseen gain remains near zero. All tied runs satisfy the predeclared row bound $\max_j \|x_j(t)\|^2 \le 8$ throughout the measured late window. These results show that high gain persists throughout late training. On the matched transfer problem, the tied model also predicts well on unseen current states while remaining within the bounded-row regime used by our transfer analysis.

Gradient-perturbation and weight-decay controls. The paired controls show that removing the Gaussian gradient perturbations yields comparable gain, indicating that the strong performance on these instances does not rely on the injected perturbations. Removing weight decay lowers final training prediction error, reflecting a tradeoff between regularization and training fit. In figure 1(c), width and prescribed weight decay vary together, so the curves reflect their joint effect.

6.2 Scaling Optimization and Transfer with Sampled SGD

To test whether our qualitative findings extend beyond exact population gradients, we evaluate sampled-transition training over larger state spaces, teacher ranks, and embedding widths. Figure 2 follows nine (V, d) settings at teacher rank $r = 8$ from 16,384 to 262,144 SGD updates. At larger state counts, some untied runs begin well below $7/8$ but continue to improve with training. By the final checkpoint, every one of the 78 $r = 8$ runs exceeds $7/8$ RPG: the lowest individual gains are 0.934 for untied training, 0.982 for tied training, and 0.968 for tied unseen states. Larger width does not always give higher gain at a fixed training budget due to limited training steps. These sampled-transition SGD results are qualitative probes of the setting studied by our perturbed population-gradient theorems. At larger teacher ranks, some configurations remain below $7/8$ while still improving; we do not interpret those checkpoints as plateaus. Section F gives the full protocol and the wider configuration grid.

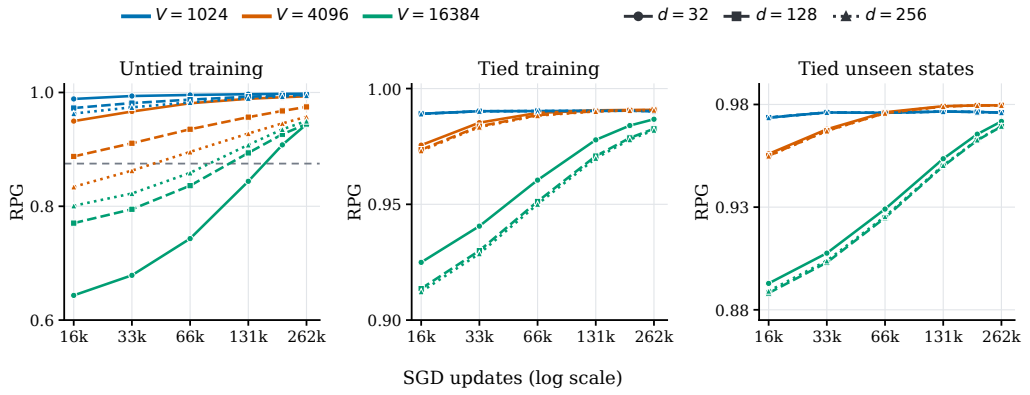

Figure 2: Relative prediction gain during sampled-transition SGD at teacher rank $r = 8$. Each panel contains nine curves: $V \in \{1024, 4096, 16384\}$ (colors) and $d \in \{32, 128, 256\}$ (line and marker styles). Lines join medians of three initialization seeds at each checkpoint, without smoothing. Training uses all current states in the untied model and $3V/4$ observed states in the tied model; tied unseen gain uses the remaining $V/4$ states. The horizontal axis is logarithmic and the RPG scales differ by panel. The dashed horizontal reference is $7/8$ wherever visible.

7 Conclusion

In this work, we investigated the training dynamics of state-transition models for one-step next-token prediction under cross-entropy loss. We first used the bounded trace of categorical curvature to show that accurate approximate second-order stationary points reached by the untied model have high prediction gain. Combining this result with localization and repeated visits, we proved that perturbed gradient descent with positive weight decay achieves constant gain in polynomial time and maintains it throughout a prescribed final interval of iterates. We then extended the training guarantee to tied embeddings without requiring coverage of the observed states. Under additional coverage and learned-row conditions, we also obtained a prediction guarantee for the complete unseen set. Our experiments illustrate training persistence and the possible transfer benefit of tying.

AI USE STATEMENT

In this work, we used generative AI tools to assist in proving mathematical claims, including checking and refining intermediate proof steps. We did not use generative AI tools for other research tasks requiring disclosure where applicable. Additionally, we used generative AI tools to aid and polish the writing of the manuscript, including improving clarity, grammar, and presentation.

All AI-assisted mathematical reasoning was independently reviewed and verified by the authors, and all AI-assisted writing was reviewed and edited to ensure that it accurately reflects the authors' intended arguments and findings. We take responsibility for the final content of this work, including all text, mathematical claims, proofs, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text states the model assumptions and training algorithms, and the appendices provide complete proofs and parameter choices. Section E and section F document the synthetic data construction, experimental protocols, and evaluation procedures. The accompanying `experiments/` directory provides the code, dependencies, fixed configurations, and instructions to rerun the experiments and regenerate the reported figures and table. No external dataset is required.

REFERENCES

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In T. Leen, T. Dietterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

Octavian Ganea, Sylvain Gelly, Gary Bécigneul, and Aliaksei Severyn. Breaking the softmax bottleneck via learnable monotonic pointwise non-linearities. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2073–2082, 2019. URL https://proceedings.mlr.press/v97/ganea19a.html.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Conference on Learning Theory*, volume 40, pp. 797–842, 2015. URL https://proceedings.mlr.press/v40/Ge15.html.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1233–1242, 2017. URL https://proceedings.mlr.press/v70/ge17a.html.

Yuhong Guo and Dale Schuurmans. Efficient global optimization for exponential family pca and low-rank matrix factorization. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1100–1107, 2008. doi: 10.1109/ALLERTON.2008.4797683.

Ruiquan Huang, Yingbin Liang, and Jing Yang. Non-asymptotic convergence of training transformers for next-token prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80634–80673, 2024. URL https://arxiv.org/abs/2409.17335.

Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25516–25549, 2025. URL https://proceedings.mlr.press/v267/huang25s.html.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling, 2017. URL https://arxiv.org/abs/1611.01462.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019. URL https://arxiv.org/abs/1902.04811v2. arXiv version 2.

Fritz John. *Extremum Problems with Inequalities as Subsidiary Conditions*, pp. 197–215. Springer Basel, Basel, 2014. ISBN 978-3-0348-0439-4. doi: 10.1007/978-3-0348-0439-4_9. URL https://doi.org/10.1007/978-3-0348-0439-4_9.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2902–2907, 2018. URL https://proceedings.mlr.press/v80/laurent18a.html.

Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*, 2017. URL https://arxiv.org/abs/1704.01265.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 685–693, 2024. URL https://proceedings.mlr.press/v238/li24f.html.

Tianyi Liu, Yan Li, Song Wei, Enlu Zhou, and Tuo Zhao. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1891–1899. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/liu21e.html.

Ashok Vardhan Makkuva, Marco Bondaschi, Chanakya Ekbote, Adway Girish, Alliot Nagle, Hyeji Kim, and Michael Gastpar. Local to global: Learning dynamics and effect of initialization for transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.03072.

Wenqing Ouyang, Ting Kei Pong, and Man-Chung Yue. Burer–Monteiro factorizability of nuclear norm regularized optimization. *arXiv preprint arXiv:2505.00349*, 2025. URL https://arxiv.org/abs/2505.00349v1. Version 1.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163. Association for Computational Linguistics, 2017. URL https://aclanthology.org/E17-2025/.

Christos Thrampoulidis. Implicit optimization bias of next-token prediction in linear models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.18551.

Aditya Varre, Gizem Yüce, and Nicolas Flammarion. Learning in-context n-grams with transformers: Sub-n-grams are near-stationary points. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 60924–60963, 2025. URL https://proceedings.mlr.press/v267/varre25a.html.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkwZSG-CZ.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qyilOnIRHI.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html.

PROOF ORGANIZATION

The appendices follow the order of the main results and make each step of the proof overviews in Sections 4 and 5 explicit. Appendix A first restates the models and training setup, then introduces the additional notation needed in the proofs.

Untied training. Appendix B proves theorem 4.1 in the three steps of Section 4.1: localize the trajectory, exclude low-gain approximate SOSPs using the categorical trace bound, and refresh the gain throughout the final window. The last step gives the fixed step size, perturbation scale, and polynomial horizon and also couples the localized and original recursions. A short comparison with a generic factorized optimality bound follows the proof.

Tied training. Appendix C.1 proves theorem 5.2 using the same localization, approximate-SOSP, and refresh structure. Its tied-specific middle step uses weight decay to absorb the energy-dependent error in the ground-truth Gram-matrix comparison.

Conditional transfer. Appendix C.2 follows the three steps of Section 5.2.1: obtain the final-window objective comparison, use centered coverage to derive observed-loss and trace control, and use positive-semidefinite completion to control the full Gram matrix and the unseen rows. The learned-row assumption is used only in this last deterministic step.

Appendix D collects the factor, localization, and perturbation lemmas used by these proofs. Appendix E gives the complete experimental protocol and settings.

Throughout the appendices, $\Delta_{\mathcal{A}} = \log V - \mathcal{H}_{\mathcal{A}}$ and $\epsilon_{\mathcal{A}} = \mathcal{R}_{\mathcal{A}} - \mathcal{H}_{\mathcal{A}}$ denote the information gap and prediction excess loss. For the untied proof, $\mu = \Delta_0/(8B\sqrt{r})$ is fixed; in the tied proofs, $\mu = \lambda V$ denotes the effective regularization scale. Unless a dependence is displayed explicitly, constants depend only on the fixed assumptions.

A SETUP FOR THE PROOFS

This appendix restates the model and training setup, then introduces the additional notation required by the proofs. Let $[V] = \{1, \dots, V\}$ be the current-state space and the set of possible next states. Untied training uses all current-state rows; tied training may instead use $N = |\mathcal{S}| < V$ observed rows, with unseen set $\mathcal{U} = [V] \setminus \mathcal{S}$. Every conditional distribution still has all V states as possible outputs.

Define

$$\mathcal{P}_V = \{p \in \mathbb{R}_+^V : p^\top \mathbf{1} = 1\}, \qquad \operatorname{softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^V e^{z_k}},$$

and let $q_i \in \operatorname{int} \mathcal{P}_V$ be the conditional next-state law given current state i. Collect these row laws in the transition table $Q \in \mathbb{R}^{V \times V}$, with $Q_{i,:} = q_i^\top$. For any logit table Z, let $P(Z) \in \mathbb{R}^{V \times V}$ have rows $P(Z)_{i,:} = \operatorname{softmax}(Z_{i,:})^\top$; when Z is fixed, write $P = P(Z)$. Softmax acts rowwise when applied to a matrix. For $p, q \in \operatorname{int} \mathcal{P}_V$, define

$$\mathcal{H}(p) = -\sum_{j=1}^V p_j \log p_j, \qquad D_{\mathrm{KL}}(q\|p) = \sum_{j=1}^V q_j \log \frac{q_j}{p_j}.$$

All logarithms are natural. Put

$$\mathsf{J} = I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top, \qquad L(Q)_{i,:} = (\log q_i)^\top \mathsf{J}, \qquad r_\star = \operatorname{rank} L(Q). \tag{5}$$

The untied model predicts $P^{E,U} = \operatorname{softmax}(EU^\top)$. Center the raw output factor only for analysis:

$$E, U \in \mathbb{R}^{V \times d}, \qquad U_{\mathrm{c}} := \mathsf{J}U, \qquad Z(E,U) := EU_{\mathrm{c}}^\top = EU^\top \mathsf{J}, \qquad d < V. \tag{6}$$

Rowwise softmax is unchanged by this centering, so $P^{E,U} = P(Z(E,U))$ and $Z(E,U)\mathbf{1} = 0$. The factor U_{c} is not trained separately: the actual parameters remain (E,U), of dimension $D_\Theta = 2Vd$.

The tied model instead trains $X \in \mathbb{R}^{V\times d}$ and predicts $P^X = \mathrm{softmax}(XX^\top)$, with $D_\Theta = Vd$. In either model we write $\Theta = (E, U)$ or $\Theta = X$, respectively, and P^Θ for its prediction table.

For a nonempty set $\mathcal{A} \subseteq [V]$ of current states, the population prediction loss and Bayes entropy are

$$\mathcal{R}_\mathcal{A}(\Theta) = -\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\sum_{j=1}^{V} q_{ij}\log P^\Theta_{ij}, \qquad \mathcal{H}_\mathcal{A} = -\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\sum_{j=1}^{V} q_{ij}\log q_{ij}.$$

For a row law q and logit vector z, we write $\mathrm{CE}(q, z) = \log\sum_j e^{z_j} - q^\top z$ for their cross-entropy. When $\Delta_\mathcal{A} := \log V - \mathcal{H}_\mathcal{A} > 0$, the relative prediction gain is

$$g_\mathcal{A}(\Theta) = \frac{\log V - \mathcal{R}_\mathcal{A}(\Theta)}{\Delta_\mathcal{A}} = 1 - \frac{|\mathcal{A}|^{-1}\sum_{i\in\mathcal{A}} D_{\mathrm{KL}}(q_i\|P^\Theta_{i,:})}{\Delta_\mathcal{A}}.$$

The uniform predictor has gain zero, the Bayes predictor has gain one, and the weight-decay penalty is not part of this prediction metric.

For all-state training, write

$$\mathcal{R} = \mathcal{R}_{[V]}, \qquad g = g_{[V]}, \qquad \Delta_{[V]} = \log V - \mathcal{H}_{[V]}. \tag{7}$$

Definition A.1 (Base class)**.** Fix a positive integer r and constants $B, \Delta_0 > 0$ independent of V. The class $\mathcal{Q}_{r,B,\Delta_0}$ consists of transition tables with positive entries and row sums one such that

$$\mathrm{rank}\, L(Q) \le r, \qquad \|L(Q)\|_\infty \le B, \qquad \Delta_{[V]} \ge \Delta_0.$$

Here $\|L(Q)\|_\infty$ is the largest absolute entry. When $r \le d$, the Bayes table is representable by the untied model. The bound on centered logits keeps conditional probabilities within constant multiples of $1/V$, while the information floor excludes an asymptotically uniform prediction problem.

Definition A.2 (Tied-compatible class)**.** For a fixed row-energy bound $B_Y \ge 1$, the tied-compatible subclass $\mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ consists of tables $Q \in \mathcal{Q}_{r,B,\Delta_0}$ admitting a representative

$$Q = \mathrm{softmax}(YY^\top), \qquad Y \in \mathbb{R}^{V\times r_Y}, \quad r_Y \le r, \quad \max_i \|y_i\|_2^2 \le B_Y.$$

Here y_i is the column-vector form of row i of Y.

Definition A.3 (Centered coverage for transfer)**.** A representative Y in definition A.2 satisfies centered coverage on $\mathcal{S}$ if its columns are linearly independent and, for a fixed $\kappa > 0$,

$$Y^\top\mathbf{1} = 0, \qquad \bar{y}_\mathcal{S} = \frac{1}{N}\sum_{i\in\mathcal{S}} y_i, \qquad \frac{1}{N}\sum_{i\in\mathcal{S}}(y_i - \bar{y}_\mathcal{S})(y_i - \bar{y}_\mathcal{S})^\top \succeq \kappa^2 I.$$

This condition is used for transfer, not for tied training alone.

For untied factors, write

$$\mathcal{E}(E, U) = \|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2.$$

All entries of the actual factors at initialization are independent $\mathcal{N}(0, 1/d)$ variables. For either model and a nonempty training set $\mathcal{A}$, the regularized objective is

$$\mathcal{F}_{\lambda,\mathcal{A}}(\Theta) = \mathcal{R}_\mathcal{A}(\Theta) + \frac{\lambda}{2}\|\Theta\|_{\mathrm{F}}^2, \tag{8}$$

where $\|(E, U)\|_{\mathrm{F}}^2 = \mathcal{E}(E, U)$. For untied all-state training, write $\mathcal{F}_\lambda = \mathcal{F}_{\lambda,[V]}$; for tied training on $\mathcal{S}$, retain $\mathcal{F}_{\lambda,\mathcal{S}}$ unless a proof declares a local shorthand.

The noisy update trains every entry of the raw factors:

$$\Theta_{t+1} = \Theta_t - h\{\nabla\mathcal{F}_{\lambda,\mathcal{A}}(\Theta_t) + \xi_t\}, \qquad \xi_t \overset{\mathrm{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{D_\Theta} I_{D_\Theta}\right). \tag{9}$$

The perturbations are independent of initialization. The decay, step size, noise scale, and finite horizon are deterministic choices made before initialization; h and σ remain fixed throughout the run. We may evaluate $\mathcal{R}_\mathcal{A}$ and $g_\mathcal{A}$ on a centered logit table instead of factors, with the same meaning.

Subtracting the output mean changes no prediction. The next identity separates this loss-invisible channel from the actual trained factor; it does not project or modify the optimization algorithm.

Lemma A.4 (The output-mean channel decouples)**.** *Let $h_0 = \mathbf{1}/\sqrt{V}$ and $a = h_0^\top U \in \mathbb{R}^{1\times d}$. Then*

$$U = U_{\mathrm{c}} + h_0 a, \qquad \|U\|_{\mathrm{F}}^2 = \|U_{\mathrm{c}}\|_{\mathrm{F}}^2 + \|a\|_2^2, \tag{10}$$

and

$$\mathcal{F}_{\lambda,[V]}(E,U) = \mathcal{R}_{[V]}(EU_{\mathrm{c}}^\top) + \frac{\lambda}{2}(\|E\|_{\mathrm{F}}^2 + \|U_{\mathrm{c}}\|_{\mathrm{F}}^2) + \frac{\lambda}{2}\|a\|_2^2.$$

Consequently the prediction trajectory and the (E, U_{c}) dynamics do not depend on a. Under the noisy update equation (9)*,*

$$a_{t+1} = (1 - h\lambda)a_t - h\, h_0^\top \xi_t^U; \tag{11}$$

At every critical point with $\lambda > 0$, $a = 0$.

Proof. The two summands in equation (10) are Frobenius orthogonal, and $\mathsf{J}h_0 = 0$. This proves the objective split. Projecting the U update onto h_0 gives equation (11); the remaining statements follow directly. □

Definition A.5 (Bayes, rank-constrained, and regularized optima)**.** We distinguish three optimization targets throughout. The *Bayes-risk value* is

$$\mathcal{R}^*_{\mathrm{Bayes}} := \mathcal{H}_{[V]}.$$

When $\Delta_{[V]} > 0$, a predictor is *Bayes-optimal* exactly when $P(Z) = Q$, equivalently $\mathcal{R}(Z) = \mathcal{H}_{[V]}$ and $g(Z) = 1$. The best unregularized prediction risk within the centered rank-d model is

$$\mathcal{R}^*_d := \min_{\substack{Z\mathbf{1}=0 \\ \operatorname{rank} Z \le d}} \mathcal{R}(Z).$$

Finally, the *regularized global value* is

$$\begin{aligned}\mathcal{F}^*_\lambda &:= \min_{E,U} \mathcal{F}_\lambda(E,U) \\ &= \min_{\substack{Z\mathbf{1}=0 \\ \operatorname{rank} Z \le d}} \{\mathcal{R}(Z) + \lambda\|Z\|_*\}.\end{aligned}$$

The second equality follows from lemma A.4 and the balanced-factor identity $\min_{Z=EU_{\mathrm{c}}^\top}(\|E\|_{\mathrm{F}}^2 + \|U_{\mathrm{c}}\|_{\mathrm{F}}^2)/2 = \|Z\|_*$. Thus a "global minimizer of $\mathcal{F}_\lambda$" always means a minimizer of the weight-decayed training objective, not a Bayes-risk minimizer. In the base class, $L(Q)$ is feasible whenever $d \ge r_\star$, so $\mathcal{R}^*_d = \mathcal{R}^*_{\mathrm{Bayes}} = \mathcal{H}_{[V]}$; nevertheless, when $\Delta_{[V]} > 0$ and $\lambda > 0$, every $\mathcal{F}_\lambda$-global minimizer has positive prediction excess risk and hence gain strictly below one. We prove this distinction in remark B.7.

For the untied proof, abbreviate its positive decay by

$$\mu = \frac{\Delta_0}{8B\sqrt{r}}, \qquad \lambda = \frac{\mu}{V}.$$

All untied updates use equation (9) with $\mathcal{A} = [V]$.

B Proof of Theorem 4.1

The proof has three steps. In Step 1, lemma B.1 bounds the initial energy and loss. Positive decay then gives an objective barrier, and lemma D.4 supplies a smooth extension on which the optimization tools apply. In Step 2, lemma B.2 constructs a comparison point below the target loss; lemmas B.3 and B.4 show that an accurate SOSP inside the energy region must also lie below that target. In Step 3, the visit and objective-increase bounds (theorem D.5 and lemma D.6) keep the trajectory inside the region and refresh its prediction guarantee in every block. The final induction proves the theorem for the original updates.

Write $\mathcal{F}_\lambda = \mathcal{F}_{\lambda,[V]}$, $\mathcal{R} = \mathcal{R}_{[V]}$ and $g = g_{[V]}$. For analysis only, put $U_{\mathrm{c}} = \mathsf{J}U$ and $Z = EU^\top\mathsf{J}$. All $2Vd$ entries of the raw factors are still trained by equation (9). The decay is

$$\mu = \frac{\Delta_0}{8B\sqrt{r}}, \qquad \lambda = \frac{\mu}{V}. \tag{12}$$

Throughout, set

$$\Gamma := B\sqrt{r}, \qquad \theta := \frac{\Delta_0}{8B^2}, \qquad t_\star := 1-\theta, \qquad m_\star := \frac{\Delta_0^2}{128B^2}.$$

Every unqualified asymptotic bound in this section has a constant independent of V; any dependence on the fixed parameters (B, r, Δ_0) is stated through named constants below.

B.1 Step 1: Localize the Trajectory

We first bound the energy and loss at initialization. This fixes a deterministic objective budget before the perturbations are drawn.

Lemma B.1 (Ordinary-Gaussian initialization)**.** *Under the initialization in section A, there is a universal $c_0 > 0$ and a universal $C_{\mathrm{init}} \geq 1$ such that, whenever $d \geq C_{\mathrm{init}} \log(eV)$,*

$$\Pr\left\{\mathcal{E}_0 \leq 3V,\ \|Z_0\|_\infty \leq 1\right\} \geq 1 - e^{-c_0 d}. \tag{13}$$

On this event,

$$\mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} \leq A_0, \qquad \mathcal{F}_\lambda(\Theta_0) - \mathcal{F}_\lambda^* \leq A_0, \qquad A_0 := 2B + 2 + 3\mu.$$

Proof. All entries of both raw factors are independent $\mathcal{N}(0, 1/d)$. Hence $d\mathcal{E}_0$ is chi-square with $2Vd$ degrees of freedom. A standard moment-generating-function Chernoff bound gives $\Pr\{\mathcal{E}_0 > 3V\} \leq e^{-cdV}$. The same estimate and a union bound over the rows of E_0 give

$$\Pr\left\{\max_i \|E_{0,i,:}\|_2^2 > 2\right\} \leq Ve^{-cd}. \tag{14}$$

Conditional on a row $E_{0,i,:}$, the variable $Z_{0,ij}$ is centered Gaussian. Its variance is at most $2/d$ on the complementary event in equation (14), since $\mathsf{J}_{jj} = 1 - 1/V$. Gaussian tails and a union bound over V^2 entries give

$$\Pr\left\{\|Z_0\|_\infty > 1,\ \max_i \|E_{0,i,:}\|_2^2 \leq 2\right\} \leq 2V^2 e^{-d/4}.$$

If $d \geq C_{\mathrm{init}} \log(eV)$ for a sufficiently large universal C_{init}, these bounds imply equation (13) after decreasing a universal $c_0 > 0$. If $\|Z_0\|_\infty \leq 1$, every row loss is at most $\log V + 2$. Also equation (18) gives $\mathcal{H}_{[V]} \geq \log V - B$. Therefore

$$\mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} \leq B + 2 + \frac{3\mu}{2} \leq A_0.$$

Finally $\mathcal{F}_\lambda^* \geq \mathcal{H}_{[V]}$. □

The next calculation turns this budget into an energy bound. Leaving the ball below requires an objective increase of at least four. The smooth extension preserves this barrier; Step 3 will show that the noisy trajectory does not cross it. Use this initialization and the predeclared recursion equation (9). Put

$$C_{\mathrm{wall}} := \frac{2(A_0+4)}{\mu} = 6 + \frac{32B\sqrt{r}(B+3)}{\Delta_0},$$

and

$$b_{\mathrm{heat}} := \min\left\{\frac{m_\star}{4}, 1\right\}.$$

By lemma B.1, with probability at least $1 - e^{-c_0 d}$,

$$\mathcal{E}_0 \leq 3V, \qquad \|Z_0\|_\infty \leq 1, \qquad \mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} \leq A_0. \tag{15}$$

Outside the ball $\mathcal{E} < C_{\mathrm{wall}} V$,

$$\mathcal{F}_\lambda \geq \mathcal{H}_{[V]} + \frac{\lambda}{2} C_{\mathrm{wall}} V = \mathcal{H}_{[V]} + A_0 + 4 \geq \mathcal{F}_\lambda(\Theta_0) + 4. \tag{16}$$

The ball also contains a balanced global factorization: comparison with $L(Q)$ gives $\|Z_\lambda\|_* \leq \|L(Q)\|_* \leq \Gamma V$, hence balanced energy at most $2\Gamma V < C_{\mathrm{wall}} V$. We may therefore use the fixed-relative-radius extension in lemma D.4. For the remaining finite-time proof, set $C_E = C_{\mathrm{wall}}$ in

$\eta_{\mathrm{hg}}, \kappa_{\mathrm{hg}}, \epsilon_{\mathrm{hg}}$. Let $\ell_{\mathrm{ext}}, \rho_{\mathrm{ext}} \geq 1$ be deterministic global Lipschitz upper bounds for the extension's gradient and Hessian, respectively.

For completeness, on this $O(\sqrt{V})$ factor ball the derivative scales are

$$\|G\|_{\mathrm{F}} = O(V^{-1/2}), \qquad \|D_Z^2 \mathcal{R}\|_{\mathrm{op}} = O(V^{-1}), \qquad \|D_Z^3 \mathcal{R}\|_{\mathrm{op}} = O(V^{-1}),$$

while $\|D_\Theta Z\| = O(\sqrt{V})$ and $\|D_\Theta^2 Z\| = O(1)$. A fixed-relative-radius radial clipping map has first, second, and third derivatives $O(1)$, $O(V^{-1/2})$, and $O(V^{-1})$. The chain rule therefore permits the smooth localization to be chosen with constants $C_\ell = C_\ell(B, r, \Delta_0) \geq 1$ and $C_\rho = C_\rho(B, r, \Delta_0) \geq 1$ such that

$$\ell_{\mathrm{ext}} \leq C_\ell, \qquad \rho_{\mathrm{ext}} \leq C_\rho \sqrt{V}. \tag{17}$$

B.2 Step 2: Convert Approximate Stationarity into Gain

We need a comparison point whose regularized objective is strictly below the level corresponding to gain $7/8$. The bounded-factorization lemma in Appendix D.1 supplies the required $O(V)$ energy. Apply this lemma to $M = L(Q)$ and write

$$L(Q) = XY^\top = \sum_{k=1}^{r_0} x_k y_k^\top, \qquad r_0 \leq r, \qquad Y^\top \mathbf{1} = 0.$$

Here x_k, y_k are the kth columns of X, Y, respectively.

Scaling these factors down slightly reduces their penalty to first order, while their prediction loss changes only to second order. The following lemma quantifies the resulting strict margin.

Lemma B.2 (Shrunken-ground-truth comparator)**.** *For every ground truth in $\mathcal{Q}_{r,B,\Delta_0}$ with $r_0 := \operatorname{rank} L(Q) \leq d$,*

$$\Delta_{[V]} \leq \min\left\{B, \frac{B^2}{2}\right\}. \tag{18}$$

In particular $0 < \theta \leq 1/16$. The particular factorization

$$E_\star = \sqrt{t_\star} X, \qquad U_\star = \sqrt{t_\star} Y$$

of $t_\star L(Q)$, zero-padded from r_0 to d hidden coordinates, satisfies

$$\widehat{\mathcal{F}}_\star := \mathcal{R}(t_\star L(Q)) + \frac{\lambda}{2}(\|E_\star\|_{\mathrm{F}}^2 + \|U_\star\|_{\mathrm{F}}^2) \leq \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - m_\star. \tag{19}$$

Proof. For row i, put

$$\psi_i(s) = \log\left(\frac{1}{V}\sum_j e^{sL(Q)_{ij}}\right).$$

Then

$$\psi_i''(s) = \mathrm{Var}_{\mathrm{softmax}(sL(Q)_{i,:})}(L(Q)_{i,:}) \leq B^2.$$

The row information relative to uniform is

$$\mathrm{KL}(q_i \| \mathbf{1}/V) = \psi_i'(1) - \psi_i(1) = \int_0^1 s\psi_i''(s)\, ds \leq \frac{B^2}{2}.$$

Centeredness and Jensen's inequality also give the bound B: $\psi_i(1) \geq 0$, while $\psi_i'(1) \leq \max_j L(Q)_{ij} \leq B$. Averaging proves equation (18). Taylor's theorem gives

$$\mathrm{KL}(q_i \| \mathrm{softmax}(t_\star L(Q)_{i,:})) \leq \frac{B^2}{2}(1 - t_\star)^2.$$

Consequently,

$$\mathcal{R}(t_\star L(Q)) - \mathcal{H}_{[V]} \leq \frac{B^2\theta^2}{2} = m_\star. \tag{20}$$

By equation (55), the factor penalty is at most

$$\lambda t_\star \Gamma V = \mu t_\star \Gamma = t_\star \frac{\Delta_0}{8}.$$

Finally, $\theta\Delta_0/8 = \Delta_0^2/(64B^2) = 2m_\star$. Adding the penalty to equation (20) proves equation (19). Because $Y^\top \mathbf{1} = 0$, this comparator uses the admissible raw output factor $U_\star$ with zero mean channel. □

The strict margin $m_\star$ is the finite-time slack missing from the unshrunk Bayes-logit comparator. The shrinkage is only in the proof comparator; neither the ground truth nor the training objective is changed.

The objective margin alone does not control a factor Hessian. We next place the comparator in embedding directions for which the softmax Hessian contribution is small. The general trace-packing calculation is proved in lemma D.2; here we apply it to categorical loss. At a raw model factor pair (E, U), put $U_{\mathrm{c}} = \mathsf{J}U$, so $Z = EU_{\mathrm{c}}^\top$. Put

$$P = \mathrm{softmax}(Z), \qquad C_i = \mathrm{diag}(p_i) - p_i p_i^\top, \qquad \mathcal{E} = \|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2.$$

Recall that $C_i \succeq 0$ and $\mathrm{tr}\, C_i \le 1$.

Lemma B.3 (Low-energy, low-Fisher packing)**.** *Suppose $d \ge 2r$ and $\mathcal{E} \le C_E V$. There are feasible factor variations $A = \dot{E}$ and $C = \dot{U}$, with $C^\top \mathbf{1} = 0$, such that*

$$AC^\top = t_\star L(Q), \tag{21}$$

$$\|A\|_{\mathrm{F}}^2 + \|C\|_{\mathrm{F}}^2 \le 2\Gamma V, \tag{22}$$

$$\frac{1}{V}\sum_i (AU_{\mathrm{c}}^\top + EC^\top)_i^\top C_i (AU_{\mathrm{c}}^\top + EC^\top)_i \le K_E(d, V), \tag{23}$$

$$K_E(d, V) := \frac{8\Gamma r C_E}{d}\left(\frac{4V}{d} + 1\right). \tag{24}$$

Proof. Apply lemma D.2 with $n = m = V$, current output factor U_{c}, $H_i = C_i$, $\tau = 1$, and comparator factors $\sqrt{t_\star}X, \sqrt{t_\star}Y$. The raw energy bound also bounds $\|E\|_{\mathrm{F}}^2 + \|U_{\mathrm{c}}\|_{\mathrm{F}}^2$. Moreover $t_\star x_{ik}^2 \le \Gamma$ and $t_\star y_k^\top C_i y_k \le \Gamma$, since $y_k^\top C_i y_k \le \sum_j p_{ij} y_{jk}^2 \le \Gamma$. The abstract curvature bound is therefore exactly equations (23) and (24), while its product and norm identities give equations (21) and (22). More precisely, the returned variations satisfy

$$\|A\|_{\mathrm{F}}^2 + \|C\|_{\mathrm{F}}^2 = \|E_\star\|_{\mathrm{F}}^2 + \|U_\star\|_{\mathrm{F}}^2 = t_\star(\|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2). \tag{25}$$

Finally, the construction in lemma D.2 places each column of C in the span of the columns of Y. Since $Y^\top \mathbf{1} = 0$, $C^\top \mathbf{1} = 0$, so this is a feasible raw-output variation with no mean-channel component. □

We now combine this direction with the exact factor derivatives from lemma D.3. Convexity makes the objective gap contribute a negative second derivative, and a small gradient controls the radial error. This gives the contradiction needed in the main-text overview.

Lemma B.4 (Robust exclusion above the target sublevel)**.** *Fix an energy body $\mathcal{E} \le C_E V$, suppose $d \ge 2r$, and assume*

$$d^2 \ge \frac{160\Gamma r C_E}{m_\star} V. \tag{26}$$

Every point in that body with

$$\mathcal{F}_\lambda(E, U) - \mathcal{H}_{[V]} \ge \frac{\Delta_0}{8} - \frac{m_\star}{2}$$

satisfies

$$\|\nabla \mathcal{F}_\lambda(E, U)\|_{\mathrm{F}} > \frac{\eta_{\mathrm{hg}}}{\sqrt{V}} \quad \textit{or} \quad \lambda_{\min}(\nabla^2 \mathcal{F}_\lambda(E, U)) \le -\frac{\kappa_{\mathrm{hg}}}{V}, \tag{27}$$

where

$$\eta_{\mathrm{hg}} := \frac{m_\star}{4\sqrt{C_E}}, \qquad \kappa_{\mathrm{hg}} := \frac{m_\star}{4\Gamma}.$$

Proof. Suppose the gradient alternative in equation (27) fails. Radial testing gives

$$r_{\mathrm{rad}} := \langle G, Z\rangle + \frac{\lambda}{2}\mathcal{E}, \qquad |r_{\mathrm{rad}}| \le \frac{1}{2}\|\nabla\mathcal{F}_\lambda\|_{\mathrm{F}}\sqrt{\mathcal{E}} \le \frac{m_\star}{8}. \tag{28}$$

Use the variation supplied by Lemma B.3. Convexity gives

$$\langle G, t_\star L(Q)\rangle \le \mathcal{R}(t_\star L(Q)) - \mathcal{R}(Z) + \langle G, Z\rangle.$$

Together with equations (19), (21), (25) and (28), the non-Fisher terms in equation (59) obey

$$\begin{aligned} 2\langle G, AC^\top\rangle + \lambda(\|A\|_{\mathrm{F}}^2 + \|C\|_{\mathrm{F}}^2) &\le 2r_{\mathrm{rad}} - 2\{\mathcal{F}_\lambda(E,U) - \widehat{\mathcal{F}}_\star\} \\ &\le 2|r_{\mathrm{rad}}| - m_\star \le -\frac{3m_\star}{4}. \end{aligned}$$

Because $d < V$, one has $4V/d + 1 \le 5V/d$. Hence equations (23) and (26) make the Fisher term at most $m_\star/4$. Thus

$$\mathscr{Q}_{\mathcal{F}}(A, C) \le -\frac{m_\star}{2}.$$

Since equation (22) bounds the squared perturbation norm by $2\Gamma V$, its Rayleigh quotient is at most $-m_\star/(4\Gamma V) = -\kappa_{\mathrm{hg}}/V$. □

The visit theorem uses one accuracy parameter for both the gradient and the Hessian. The following consequence translates the preceding alternative into that convention.

Corollary B.5 (Prediction loss at approximate SOSPs)**.** *For any Hessian-Lipschitz upper bound $\bar\rho \ge 1$, set*

$$\epsilon_{\mathrm{hg}}(\bar\rho) := \min\left\{\eta_{\mathrm{hg}} V^{-1/2}, \frac{\kappa_{\mathrm{hg}}^2}{4\bar\rho V^2}, 1\right\}. \tag{29}$$

Under equation (26)*, every point in the energy body satisfying*

$$\|\nabla\mathcal{F}_\lambda\| \le \epsilon, \qquad \nabla^2\mathcal{F}_\lambda \succeq -\sqrt{\bar\rho\epsilon}\, I, \qquad 0 < \epsilon \le \epsilon_{\mathrm{hg}}(\bar\rho)$$

obeys

$$\mathcal{F}_\lambda(E,U) < \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - \frac{m_\star}{2}.$$

Proof. Otherwise lemma B.4 applies. Its gradient alternative contradicts equation (29), while its curvature alternative contradicts $\sqrt{\bar\rho\epsilon} \le \kappa_{\mathrm{hg}}/(2V)$. □

B.3 Step 3: Refresh the Gain throughout the Final Window

We use the always-perturbed visit theorem and descent concentration bound of Jin et al. (2019), stated with their parameter specialization in Appendix D.4. A successful visit gives an objective margin of $m_\star/2$. We choose the schedule so that every forward objective increase within a block is at most $b_{\mathrm{heat}} \le m_\star/4$. This is the margin-refresh argument of lemma D.8. The blocks share their endpoints and do not change the updates.

Accuracy and width. Both constants are independent of V. In equation (32) below, choose a fixed accuracy coefficient $c_{\mathrm{acc}} = c_{\mathrm{acc}}(B, r, \Delta_0) > 0$ satisfying

$$0 < c_{\mathrm{acc}} \le \min\left\{\eta_{\mathrm{hg}}, \frac{\kappa_{\mathrm{hg}}^2}{4C_\rho}, 1\right\};$$

then $c_{\mathrm{acc}} V^{-5/2} \le \epsilon_{\mathrm{hg}}(\rho_{\mathrm{ext}})$ for every $V \ge 1$. Choose a finite width constant $C_{\mathrm{wid}} = C_{\mathrm{wid}}(B, r, \Delta_0) \ge 1$ large enough that, for every $V \ge 2$,

$$C_{\mathrm{wid}}\sqrt{V} \ge \max\left\{2r, C_{\mathrm{init}}\log(eV), \sqrt{\frac{160\Gamma r C_{\mathrm{wall}}}{m_\star}V}\right\}. \tag{30}$$

Such a constant exists because $\sup_{V\ge 2}\log(eV)/\sqrt{V} < \infty$. Increase it once and choose a finite $V_{\mathrm{hg}} = V_{\mathrm{hg}}(B, r, \Delta_0)$ so that $C_{\mathrm{wid}}\sqrt{V} < V$ and the probability absorption below holds for all $V \ge V_{\mathrm{hg}}$. There is then a universal $c_{\mathrm{u}} > 0$ such that, whenever $C_{\mathrm{wid}}\sqrt{V} \le d < V$,

$$e^{-c_0 d} + \frac{d}{2}e^{-d} \le e^{-c_{\mathrm{u}} d}. \tag{31}$$

Step size, noise scale, and horizon. Assume the ground truth belongs to $\mathcal{Q}_{r,B,\Delta_0}$, fix $V \geq V_{\mathrm{hg}}$, and suppose $C_{\mathrm{wid}}\sqrt{V} \leq d < V-1$. Initialize the two raw factors $\Theta_0 = (E_0, U_0) \in \mathbb{R}^{V\times d} \times \mathbb{R}^{V\times d}$ by the ordinary Gaussian law in section A. Run equation (9) with $\Theta_t = (E_t, U_t)$ and $Z_t = E_t U_t^\top$ J. The schedule below contains only the displayed instance and class parameters. Choose

$$\epsilon = c_{\mathrm{acc}} V^{-5/2} \leq \epsilon_{\mathrm{hg}}(\rho_{\mathrm{ext}}). \tag{32}$$

There is a finite constant $C_{\mathrm{sched}} = C_{\mathrm{sched}}(B, r, \Delta_0) \geq 1$, independent of V, for which the predeclared choice

$$\iota = C_{\mathrm{sched}} d \tag{33}$$

satisfies

$$\iota \geq C_\iota \left[1 + \log\left(\frac{4e^{d+1} D_\Theta(\ell_{\mathrm{ext}} + \rho_{\mathrm{ext}} + 1)(A_0+1)}{\epsilon}\right) + \frac{A_0+1}{b_{\mathrm{heat}}} + b_{\mathrm{heat}}^{-1/6}\right]. \tag{34}$$

The remaining predeclared parameters are

$$\begin{aligned} h &= (\iota^9 \ell_{\mathrm{ext}})^{-1}, & \sigma &= \iota\epsilon, \\ T_{\mathrm{blk}} &= \left\lceil C_J \ell_{\mathrm{ext}} A_0 \epsilon^{-2} \iota^{15} \right\rceil, & T &= d T_{\mathrm{blk}}. \end{aligned} \tag{35}$$

Increase C_{sched} once, if necessary, so that this same choice also satisfies

$$4(T_{\mathrm{blk}}+1)^2 e^{-\iota} \leq \frac{1}{4} e^{-d}. \tag{36}$$

This is possible because equation (33) dominates the logarithm of the polynomial expression in equation (35). Finally define

$$C_T := 1 + C_J C_\ell A_0 c_{\mathrm{acc}}^{-2} C_{\mathrm{sched}}^{15}.$$

Thus $C_T = C_T(B, r, \Delta_0)$ is independent of V. Consequently,

$$T_{\mathrm{blk}} \leq C_T V^5 d^{15}, \qquad T \leq C_T V^5 d^{16} \leq C_T V^{21}.$$

Moreover,

$$\Pr\left\{\text{there exists } t \in [T_{\mathrm{blk}}, T] : g(Z_t) \leq \frac{7}{8}\right\} \leq e^{-c_0 d} + \frac{d}{2} e^{-d} \leq e^{-c_{\mathrm{u}} d}. \tag{37}$$

This probability is only over the ordinary Gaussian initialization and the prescribed perturbations; the instance and every displayed schedule quantity are fixed beforehand. The same h and σ are used at every step: T_{blk} is a proof block length and T is the reported finite horizon, not a stopping time. On the same event,

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbf{1}\left\{g(Z_t) > \frac{7}{8}\right\} > 1 - \frac{1}{d}. \tag{38}$$

In fact the pointwise statement in equation (37) is stronger: every iterate in the final $1 - 1/d$ fraction of the run has high gain.

These choices complete the three steps. We restate the main result and prove the localization and refresh induction on a single probability event.

Theorem B.6 (Restatement of Theorem 4.1: Untied constant gain)**.** *Fix r, B, Δ_0. There are positive constants c, C, V_0 such that, for every $V \geq V_0$, $Q \in \mathcal{Q}_{r,B,\Delta_0}$, and $\max\{2r, C\sqrt{V}\} \leq d < V-1$, setting $\lambda = \Delta_0/(8B\sqrt{r}\,V)$ allows deterministic choices of $h, \sigma > 0$ and an integer T divisible by d for which*

$$\Pr\left\{g_{[V]}(\Theta_t) > \frac{7}{8} \text{ for every integer } t \in [T/d, T]\right\} \geq 1 - e^{-cd}.$$

The quantities h^{-1}, σ^{-1} and T are bounded by polynomials in V. The probability is over both initialization and perturbations.

Proof. Put

$$q_{\mathrm{hg}} := \mathcal{H}_{[V]} + \frac{\Delta_0}{8}, \qquad \zeta := \frac{1}{4} e^{-d},$$

and condition on the initialization event equation (15). Use the single globally smooth localized objective $\widetilde{\mathcal{F}}_\lambda$ supplied by lemma D.4. Its construction and equation (16) give the stronger exterior bound

$$\widetilde{\mathcal{F}}_\lambda \geq \mathcal{H}_{[V]} + A_0 + 4 \quad \text{outside the wall ball.} \tag{39}$$

On the same probability space, define an auxiliary trajectory $\widetilde{\Theta}_0 = \Theta_0$ and

$$\widetilde{\Theta}_{t+1} = \widetilde{\Theta}_t - h\{\nabla \widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) + \xi_t\},$$

using exactly the perturbations in the actual recursion equation (9). The global optimization inputs are applied first to this auxiliary trajectory. Whenever it stays in the wall ball, equality of the two objectives there and induction on t give $\widetilde{\Theta}_t = \Theta_t$.

Partition the auxiliary noisy trajectory, without restarting the actual algorithm, into the d blocks

$$[0, T_{\mathrm{blk}}],\ [T_{\mathrm{blk}}, 2T_{\mathrm{blk}}], \ldots, [(d-1)T_{\mathrm{blk}}, dT_{\mathrm{blk}}].$$

Fix one block and condition on the complete history before its left endpoint. Future perturbations remain iid Gaussian. Applying lemma D.6 to every deterministic start/end pair in that block and using equation (36) gives, except with conditional probability ζ,

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_v) - \widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_u) \leq C_{\mathrm{heat}} \left(\frac{A_0}{\iota} + \frac{\epsilon^2}{\ell_{\mathrm{ext}} \iota^6} \right) \leq b_{\mathrm{heat}} \tag{40}$$

simultaneously for all integer pairs $u < v$ in the block. Here $C_{\mathrm{heat}} > 0$ is the universal constant obtained by substituting the schedule into equation (60). The last inequality follows from the two nonlogarithmic slack terms in equation (34). Whenever the auxiliary objective gap at the block start is at most A_0, theorem D.5, applied conditionally to the same future noises, also visits an ϵ-SOSP during the block except with conditional probability ζ.

In the first block, equation (40) with $u = 0$ keeps the auxiliary objective below $\mathcal{F}_\lambda(\Theta_0) + 1$. Thus equation (39) prevents any auxiliary iterate from lying outside the wall ball. The same-noise induction therefore couples the auxiliary and actual recursions throughout the block. At the visited time τ_0, corollary B.5 with $C_E = C_{\mathrm{wall}}$ gives

$$\mathcal{F}_\lambda(\Theta_{\tau_0}) < q_{\mathrm{hg}} - \frac{m_\star}{2}.$$

If $\tau_0 < T_{\mathrm{blk}}$, apply equation (40) to the pair $(\tau_0, T_{\mathrm{blk}})$; if $\tau_0 = T_{\mathrm{blk}}$, the visited-time bound itself is stronger. In either case,

$$\mathcal{F}_\lambda(\Theta_{T_{\mathrm{blk}}}) < q_{\mathrm{hg}} - \frac{m_\star}{4}. \tag{41}$$

Now suppose the two trajectories are coupled through jT_{blk} and equation (41) holds with T_{blk} replaced by jT_{blk}. Since

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_{jT_{\mathrm{blk}}}) - \inf \widetilde{\mathcal{F}}_\lambda < q_{\mathrm{hg}} - \mathcal{H}_{[V]} = \frac{\Delta_0}{8} \leq A_0,$$

the visit and heating results may first be invoked conditionally for the auxiliary next block. On their success, every time t in that block satisfies

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) \leq \mathcal{F}_\lambda(\Theta_{jT_{\mathrm{blk}}}) + b_{\mathrm{heat}} < q_{\mathrm{hg}}.$$

Since $\Delta_0 \leq \Delta_{[V]} \leq B$, this level lies strictly below the exterior bound in equation (39); hence no auxiliary iterate exits the wall and the same-noise induction couples the two trajectories throughout the block. Moreover,

$$\mathcal{R}(Z_t) - \mathcal{H}_{[V]} \leq \mathcal{F}_\lambda(\Theta_t) - \mathcal{H}_{[V]} < \frac{\Delta_0}{8} \leq \frac{\Delta_{[V]}}{8},$$

so $g(Z_t) > 7/8$ throughout the block. Its now-coupled SOSP visit restores the margin $q_{\mathrm{hg}} - m_\star/2$. Heating from the visit to the endpoint restores equation (41) at $(j+1)T_{\mathrm{blk}}$ when the visit occurs earlier, while a visit at the endpoint gives the stronger bound directly. Induction proves the pointwise gain statement on all $d - 1$ blocks after the first.

No independence between block-success events is needed. Their conditional failure probabilities are at most 2ζ, so the tower property and a union bound over d blocks give total optimization failure at most $2d\zeta = de^{-d}/2$. Add the initialization failure and use equation (31). The schedule together with equation (17) gives the two time bounds. Finally, at most the first T_{blk} indices are not certified, which proves equation (38). □

B.4 Comparison with a Generic Factorized Bound

The improvement over a direct use of generic factorized optimality theory can be seen already at an exact SOSP of the positive-decay untied objective. Balance identifies its factor penalty with $\lambda\|Z\|_*$, where $Z = EU^\top$; suppose its factor energy is at most $C_E V$, and write $f(Z) = \mathcal{R}(Z) + \lambda\|Z\|_*$. Theorem 4.1 of Ouyang et al. (2025), combined with convexity and stationarity, gives

$$f(Z) - f(M) \le L\sigma_d(Z)\|M\|_*, \qquad L = \sup_{\|H\|_{\mathrm{F}}=1,\,\mathrm{rank}\,H=2} \nabla^2\mathcal{R}(Z)[H,H]$$

for a centered comparator M. Here $L \le 1/(2V)$ for categorical cross-entropy, $\sigma_d(Z) \le C_E V/(2d)$, and a bounded fixed-rank ground-truth comparator has $\|M\|_* \le B\sqrt{r}\,V$. This route therefore gives only an $O(V/d)$ comparison error and does not certify constant gain at $d \asymp \sqrt{V}$.

Our trace-packing argument uses the sum of the categorical Hessian traces rather than only a uniform curvature bound. It produces the $O(V/d^2 + 1/d)$ term in equation (23), which is small enough for the strict-margin argument at $d \ge C\sqrt{V}$. This comparison explains the gain in the sufficient width bound; it is not a lower bound on the width required by another method.

Remark B.7 (Prediction loss versus the regularized objective)**.** A global minimizer of $\mathcal{F}_\lambda$ cannot be Bayes-optimal. When $\lambda > 0$ and $\Delta_{[V]} > 0$, the centered Bayes table $L(Q)$ is nonzero, and the derivative of $\mathcal{R}(tL(Q)) + \lambda t\|L(Q)\|_*$ at $t = 1$ equals $\lambda\|L(Q)\|_* > 0$. Moving to $t < 1$ therefore lowers the regularized objective. The theorem accordingly guarantees prediction gain rather than convergence to the Bayes predictor.

C Tied Training and Transfer

C.1 Proof of Theorem 5.2

The proof has three steps. Step 1 uses lemma C.1 to bound the initial objective and construct a smooth extension with an exterior objective barrier. Step 2 compares an accurate SOSP with the ground-truth Gram: lemma C.2 gives the comparison direction, and lemma C.3 uses weight decay to absorb its energy-dependent error. Step 3 applies the fixed-noise block bound (lemma D.7) repeatedly. Lemma C.4 proves the resulting comparison for every final-window iterate, after which we specialize the decay and restate the main theorem.

We keep the effective decay $\mu = \lambda V$ explicit in the supporting lemmas. This lets the transfer proof use the same argument with its different decay, without repeating the optimization proof.

Let $\mathcal{S} \subset [V]$ index observed input states and let $\mathcal{U} = [V] \setminus \mathcal{S}$ index unseen input states. Throughout, assume

$$|\mathcal{S}| = N = \Theta(V), \qquad |\mathcal{U}| = m = V - N, \qquad 1 \le d < N < V,$$

where the constants in $N = \Theta(V)$ are fixed. In particular $d < V - 1$. All V tokens, including those in $\mathcal{U}$, remain output classes in every observed conditional distribution. Write $P_{\mathcal{S}}$ for row selection and set

$$\mathsf{J} = I_V - \mathbf{1}\mathbf{1}^\top/V, \qquad H_N = I_N - \mathbf{1}_N\mathbf{1}_N^\top/N.$$

The ground-truth distributions q_i have positive entries, and their canonical logits are $L(Q)_{i,:} = (\log q_i)^\top \mathsf{J}$. The tied learner is

$$X \in \mathbb{R}^{V\times d}, \qquad K = XX^\top, \qquad L = K\mathsf{J}, \qquad p_i(X) = \mathrm{softmax}(L_i).$$

Its Vd entries are all optimized. Our matrix ∞-norm is the maximum absolute entry, $\|A\|_1 = \sum_{ij}|A_{ij}|$, and $\|A\|_{2\to\infty} = \max_i \|A_{i,:}\|_2$.

For any nonempty current-state set $\mathcal{A} \subseteq [V]$, use the common notation $\mathcal{R}_\mathcal{A}, \mathcal{H}_\mathcal{A}, g_\mathcal{A}$ and the proof abbreviations

$$\mathcal{H}_\mathcal{A} = |\mathcal{A}|^{-1}\sum_{i\in\mathcal{A}}\mathcal{H}(q_i), \qquad \Delta_\mathcal{A} = \log V - \mathcal{H}_\mathcal{A},$$

$$\epsilon_\mathcal{A}(X) = \mathcal{R}_\mathcal{A}(X) - \mathcal{H}_\mathcal{A} = |\mathcal{A}|^{-1}\sum_{i\in\mathcal{A}} D_{\mathrm{KL}}(q_i\|p_i(X)), \quad g_\mathcal{A}(X) = 1 - \epsilon_\mathcal{A}(X)/\Delta_\mathcal{A} \quad (\Delta_\mathcal{A} > 0).$$

Thus $\mathcal{H}_{\mathcal{S}}$ and $\mathcal{H}_{\mathcal{U}}$ denote the scalar optimal cross-entropies on the observed and unseen states. The nontriviality requirements are $\Delta_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$ and, when full-unseen gain is discussed, $\Delta_{\mathcal{U}} \geq \delta_{\mathrm{te}} > 0$. They are average information floors, not information floors for every individual state. For $\lambda = \mu/V$,

$$\mathcal{F}_{\lambda,\mathcal{S}}(X) = \mathcal{H}_{\mathcal{S}} + \epsilon_{\mathcal{S}}(X) + \frac{\lambda}{2}\|X\|_{\mathrm{F}}^2,$$
$$\Phi_{\lambda,\mathcal{S}}(K) = \frac{1}{N}\sum_{i\in\mathcal{S}} \mathrm{CE}(q_i, (K\mathsf{J})_i) + \frac{\lambda}{2}\operatorname{tr} K, \qquad K \succeq 0.$$

Here $\mathcal{F}_{\lambda,\mathcal{S}}(X) = \mathcal{R}_{\mathcal{S}}(X) + \lambda\|X\|_{\mathrm{F}}^2/2$, and $P^{\Theta}_{i,:} = p_i(X)^\top$ for $\Theta = X$. We abbreviate the two objectives by $\mathcal{F}$ and Φ within proofs. The factorized global value is $\min_X \mathcal{F}(X)$; its equality to the unrestricted PSD global value is neither assumed nor needed.

Fix a row-energy bound $B_Y \geq 1$ and a rank bound $r \geq 1$. Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ and choose a Gram representative Y with linearly independent columns:

$$K^\star = YY^\top, \qquad L(Q) = K^\star\mathsf{J}, \qquad Y \in \mathbb{R}^{V\times r_Y}, \quad 1 \leq r_Y \leq r, \quad \max_i \|y_i\|_2^2 \leq B_Y. \tag{42}$$

Such a representative is always available: delete redundant latent directions from any realization in definition A.2 while preserving its Gram matrix. This also preserves every row norm. The information floor $\Delta_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$ is retained. This definition does not require $Y^\top\mathbf{1} = 0$ or observed coverage. The Gram bound is a quantitative realizability requirement, not a consequence of arbitrary algebraic PSD representability. It implies a uniform canonical-logit bound $\|L(Q)\|_\infty \leq 2B_Y$. We do not claim approximate realizability without an additional approximation-error analysis.

Put $C_p = 40B_Y r$ and $\nu_0 = V/d^2$. Throughout the optimization argument assume $2r_Y \leq d < N < V$, $d \geq C\log(eV)$, $\lambda = \mu/V \geq 2C_p/d^2$, and $0 < \mu \leq 1$. Initialize $(X_0)_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1/d)$ and run

$$X_{t+1} = X_t - h\{\nabla\mathcal{F}_{\lambda,\mathcal{S}}(X_t) + \xi_t\}, \qquad \mathrm{vec}(\xi_t) \overset{\mathrm{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{Vd} I_{Vd}\right).$$

The step and noise scale below are fixed for the entire run and selected in advance. The proof blocks do not restart the recursion.

C.1.1 Step 1: Localize the Trajectory

We first control the loss at ordinary Gaussian initialization. The positive penalty then gives an objective barrier at a deterministic radius. The following lemma also provides the global smoothness bounds needed to apply the perturbation results. The radius depends on μ; the sharper $O(V)$ energy estimate at successful visits comes in Step 2.

Lemma C.1 (Localization from ordinary Gaussian initialization)**.** *For $d \geq C\log(eV)$, with probability at least $1 - e^{-cd}$,*

$$\max_i \|(X_0)_i\|_2^2 \leq 2, \qquad \|X_0\|_{\mathrm{F}}^2 \leq 2V, \qquad \mathcal{F}(X_0) - \mathcal{H}_{\mathcal{S}} \leq 2B_Y + 8 + \mu. \tag{43}$$

Fix $A = 4B_Y + 20 + C_p^{-1}$ and

$$C_w = 1 + \max\{2, B_Y, 2(A+4)/\mu\}, \qquad R_w^2 = C_w V.$$

There exists a smooth $\widetilde{\mathcal{F}} : \mathbb{R}^{Vd} \to \mathbb{R}$ equal to $\mathcal{F}$ on $\|X\|_{\mathrm{F}} \leq R_w$, with the same global value as $\mathcal{F}$, satisfying $\widetilde{\mathcal{F}} > \mathcal{H}_{\mathcal{S}} + A + 4$ outside that ball. Its global gradient- and Hessian-Lipschitz constants $\ell_{\mathrm{ext}}, \rho_{\mathrm{ext}} \geq 1$ may be chosen polynomial in V, N, μ^{-1}.

Proof. The exponential moment of a χ^2_d variable and a union bound over V rows give

$$\Pr\{\max_i \|(X_0)_i\|^2 > 2\} \leq Ve^{-c_0 d}.$$

The logarithmic width assumption gives the stated probability after adjusting constants. On this event the raw Gram entries are at most two in absolute value, and its centered entries at most four. Moreover the ground-truth raw logits lie in $[-B_Y, B_Y]$, so $\log V - \mathcal{H}(q_i) \leq 2B_Y$. The conservative risk bound in equation (43) follows by adding the penalty.

Cross-entropy is bounded below by its Bayes entropy, and the positive quadratic penalty is coercive. Thus a factorized global minimizer exists. The ground truth padded with zero columns is an admissible comparison point of squared norm at most $B_Y V$, so some global minimizer also has squared norm at most $B_Y V$. This comparison does not constrain X_0. For $\|X\|_{\mathrm{F}} \geq R_w$, $\mathcal{F}(X) \geq \mathcal{H}_{\mathcal{S}} + \lambda R_w^2/2 > \mathcal{H}_{\mathcal{S}} + A + 4$.

For completeness choose a smooth scalar function $s : [0, \infty) \to [0, 2R_w]$ equal to r for $r \leq R_w$, with $s(r) \geq R_w$ for $r \geq R_w$, and constant for $r \geq 3R_w$. It can be chosen with derivatives bounded by the corresponding powers of R_w^{-1}. Define $\psi(X) = s(\|X\|)X/\|X\|$ for $X \neq 0$, with $\psi(0) = 0$, and $\widetilde{\mathcal{F}} = \mathcal{F} \circ \psi$. It has the claimed equality, wall and global value. To bound derivatives, write $L_{\mathrm{obs}}(X) = P_{\mathcal{S}} X X^\top \mathrm{J}$. Then

$$\|DL_{\mathrm{obs}}(X)[Z]\|_{\mathrm{F}} \leq 2\|X\|_{\mathrm{F}}\|Z\|_{\mathrm{F}}, \qquad \|D^2 L_{\mathrm{obs}}(X)[Z, W]\|_{\mathrm{F}} \leq 2\|Z\|_{\mathrm{F}}\|W\|_{\mathrm{F}}.$$

Average logit cross-entropy has first, second and third derivative norms at most $\sqrt{2/N}$, $1/(2N)$ and C/N. Thus on the image ball of radius $2R_w$ the factor gradient- and Hessian-Lipschitz constants are bounded by $C(1 + R_w^2/N) + \lambda$ and $C(1 + (R_w^3 + R_w)/N)$. The chain rule with ψ gives global polynomial bounds, as claimed. □

C.1.2 Step 2: Convert Approximate Stationarity into Gain

The tied factorization changes both the first and second derivatives: a variation in X changes the input and output embeddings together. We compute these derivatives and choose a direction representing the ground-truth Gram. The categorical trace bound keeps its Hessian cost proportional to $\|X\|_{\mathrm{F}}^2/d^2$.

Lemma C.2 (Factor derivatives and a packed ground-truth direction)**.** *Let $P(X) \in \mathbb{R}^{N \times V}$ have rows $p_i(X)^\top$, $i \in \mathcal{S}$. Put $\Gamma = (P(X) - Q_{\mathcal{S}})/N$ and $M = \mathrm{sym}(P_{\mathcal{S}}^\top \Gamma) + \lambda I_V/2$. For every direction Z,*

$$\begin{aligned} \nabla \mathcal{F}(X) &= 2MX, \\ \nabla^2 \mathcal{F}(X)[Z, Z] &= \frac{1}{N} \sum_{i \in \mathcal{S}} \mathrm{Var}_{p_i(X)}((XZ^\top + ZX^\top)_{i,:}) + 2\langle M, ZZ^\top \rangle, \\ \langle M, XX^\top \rangle &= \tfrac{1}{2}\langle \nabla \mathcal{F}(X), X \rangle. \end{aligned} \tag{44}$$

Under definition A.2, if $d \geq 2r_Y$, there is a direction $Z \in \mathbb{R}^{V \times d}$ with $ZZ^\top = K^\star$ and Fisher term at most $C_p\|X\|_{\mathrm{F}}^2/d^2$, where $C_p = 40B_Y r$. Consequently, if $\|\nabla \mathcal{F}(X)\|_{\mathrm{F}} \leq \eta$ and $\nabla^2 \mathcal{F}(X) \succeq -\gamma I$, then

$$\mathcal{F}(X) - \Phi(K^\star) \leq \frac{C_p\|X\|_{\mathrm{F}}^2}{2d^2} + \frac{\eta\|X\|_{\mathrm{F}}}{2} + \frac{\gamma B_Y V}{2}. \tag{45}$$

Proof. The variations of $XX^\top$ are $XZ^\top + ZX^\top$ and $2ZZ^\top$; the categorical covariance annihilates constants, so right-centering does not affect its quadratic form. This proves equation (44).

Take an orthonormal eigenbasis w_j of $X^\top X$, let $z_j = Xw_j$, $e_j = \|z_j\|^2$, and for $e_j > 0$ define

$$v_j = z_j/\sqrt{e_j}, \qquad h_j = \sum_{i \in \mathcal{S}} v_j^\top (\mathrm{Diag}(p_i) - p_i p_i^\top) v_j.$$

For $e_j = 0$ put $h_j = 0$. The nonzero v_j are orthonormal, so $\sum_j e_j = \|X\|_{\mathrm{F}}^2$ and $\sum_j h_j \leq N$. At least $d/2$ indices satisfy both $e_j \leq 4\|X\|_{\mathrm{F}}^2/d$ and $h_j \leq 4N/d$. Choose r_Y distinct such indices j_k. Writing the columns of Y as $y^{(k)}$, set $Z_\sigma = \sum_{k=1}^{r_Y} \sigma_k y^{(k)} w_{j_k}^\top$ for independent signs σ_k. Every choice has $Z_\sigma Z_\sigma^\top = YY^\top$ and $\|Z_\sigma\|_{\mathrm{F}}^2 \leq B_Y V$. Sign averaging cancels the Fisher cross terms. The kth remaining term is at most

$$\frac{2B_Y}{N} e_{j_k}(1 + h_{j_k}) \leq \frac{8B_Y\|X\|_{\mathrm{F}}^2}{Nd}(1 + 4N/d).$$

Indeed, for $F_i = \mathrm{Diag}(p_i) - p_i p_i^\top$, $y^{(k)\top} F_i y^{(k)} \leq B_Y$ and $|y_i^{(k)}|^2 \leq B_Y$; apply $\|a + b\|_{F_i}^2 \leq 2\|a\|_{F_i}^2 + 2\|b\|_{F_i}^2$. Summing and using $d < N$ gives $C_p\|X\|_{\mathrm{F}}^2/d^2$ for the expectation, hence for some sign choice. These signs specify a proof direction only; they do not alter initialization or the recursion.

Convexity of Φ gives

$$\langle M, K^\star\rangle \le \Phi(K^\star) - \mathcal{F}(X) + \langle M, XX^\top\rangle \le \Phi(K^\star) - \mathcal{F}(X) + \eta\|X\|_{\mathrm{F}}/2.$$

Insert this and the packed direction into the Hessian lower bound $-\gamma\|Z\|_{\mathrm{F}}^2 \le \nabla^2\mathcal{F}(X)[Z,Z]$ and rearrange. This proves equation (45) without referring to a regularized lifted minimizer. □

The last inequality still depends on the learned energy. We next use the penalty already present in $\mathcal{F}$ to move this term to the left. This first bounds the energy by $O(V)$ and then gives an objective comparison error of order V/d^2.

Lemma C.3 (Absorbing the factor-energy error)**.** *Let $\nu_0 = V/d^2$, $\lambda = \mu/V \ge 2C_p/d^2$, and let R_w be the radius in lemma C.1. If $\|X\|_{\mathrm{F}} \le R_w$, $\|\nabla\mathcal{F}(X)\|_{\mathrm{F}} \le \nu_0/(4R_w)$, and $\nabla^2\mathcal{F}(X) \succeq -\nu_0 I/(4B_Y V)$, then*

$$\|X\|_{\mathrm{F}}^2 \le \left(2B_Y + \frac{1}{2C_p}\right)V, \qquad \mathcal{F}(X) - \Phi(K^\star) \le (C_pB_Y + 1/2)\nu_0.$$

Proof. Writing $R = \|X\|_{\mathrm{F}}$, the last two terms in equation (45) total at most $\nu_0/4$. Put $v = \mathcal{F}(X) - \mathcal{H}_{\mathcal{S}} \ge \lambda R^2/2$. Because $\lambda \ge 2C_p/d^2$, comparison yields

$$v - \mu B_Y/2 \le \lambda R^2/4 + \nu_0/4 \le v/2 + \nu_0/4.$$

Hence $v \le \mu B_Y + \nu_0/2$ and

$$R^2 \le 2B_Y V + \frac{\nu_0 V}{\mu} \le \left(2B_Y + \frac{1}{2C_p}\right)V.$$

Inserting this improved energy bound back into equation (45) gives

$$\mathcal{F}(X) - \Phi(K^\star) \le (C_pB_Y + 1/2)\nu_0.$$

The large wall radius therefore does not enter the final comparison error with a factor μ^{-1}.

□

C.1.3 Step 3: Refresh the Gain throughout the Final Window

The comparison from Step 2 applies at accurate SOSP visits. We now choose the accuracies and a fixed-noise schedule so that one visit in each block leaves enough margin for every iterate before the next visit. The general block result is lemma D.7; its full proof appears in Appendix D.4. The following lemma proves the common event for the original tied recursion, including the coupling with the smooth extension.

Lemma C.4 (Final-window objective comparison)**.** *Let definition A.2 hold, with $\Delta_{\mathcal{S}} \ge \delta_{\mathrm{tr}} > 0$, $2r_Y \le d < N < V$ and $d \ge C\log(eV)$. Set*

$$C_p = 40B_Y r, \qquad \nu_0 = V/d^2, \qquad \lambda = \mu/V \ge 2C_p/d^2, \quad 0 < \mu \le 1, \qquad \nu = (C_pB_Y + 1)\nu_0.$$

There are deterministic $h, \sigma > 0$ and an integer T_{blk}, specified in the proof, for which

$$T = dT_{\mathrm{blk}} \le p_{\mathrm{time}}(V, N, d, \mu^{-1}), \tag{46}$$

where p_{time} is a fixed polynomial with class-dependent coefficients. With probability at least

$$1 - e^{-cd}, \tag{47}$$

over initialization and all perturbations, every integer $T_{\mathrm{blk}} \le t \le T$ satisfies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \Phi_{\lambda,\mathcal{S}}(K^\star) \le \nu, \qquad g_{\mathcal{S}}(X_t) \ge 1 - \frac{B_Y\mu/2 + \nu}{\delta_{\mathrm{tr}}}.$$

In particular the conclusion holds at the predeclared final iterate.

Proof. Use the wall and budget of lemma C.1, and put

$$\eta = \frac{\nu_0}{4R_w}, \qquad \gamma = \frac{\nu_0}{4B_Y V}, \qquad \varepsilon = \min\{\eta, \gamma^2/\rho_{\mathrm{ext}}, 1\},$$
$$\delta_c = \min\{\nu_0/4, 1\}, \qquad \zeta = \frac{e^{-d}}{8d}.$$

Apply lemma D.7 to $\widetilde{\mathcal{F}}$, with $p = Vd$, and run exactly dT_{blk} steps using its fixed h, σ.

At every successful visit inside the wall, these accuracies allow us to apply lemma C.3. Thus $\mathcal{F}(X) - \Phi(K^\star) \le (C_p B_Y + 1/2)\nu_0$, with no factor μ^{-1} in the comparison error.

On the initialization event, the first block starts below $\mathcal{H}_{\mathcal{S}} + 2B_Y + 8 + \mu$. Its simultaneous increase bound is at most $\delta_c \le 1$. Thus every auxiliary iterate in this block is below the exterior wall. The auxiliary and original recursions coincide step by step under the same perturbations. The visit belongs to the original objective, and its block endpoint satisfies

$$\mathcal{F}(X_{T_{\mathrm{blk}}}) - \Phi(K^\star) \le (C_p B_Y + 3/4)\nu_0.$$

Starting a subsequent block at such an endpoint, its simultaneous increase bound puts every iterate below $\Phi(K^\star) + (C_p B_Y + 1)\nu_0$. This is below the wall since $\mu \le 1$ and $\nu_0 \le \mu/(2C_p) \le 1/(2C_p)$. Its start also has gap at most A to $\inf \mathcal{F} \ge \mathcal{H}_{\mathcal{S}}$. The conditional block theorem therefore applies again. Each new visit refreshes the same endpoint bound; induction proves the common comparison bound throughout the final $d - 1$ blocks. The total conditional failure is at most $2d\zeta$, in addition to the initialization failure. Adjusting c gives equation (47).

All scales are polynomial or inverse-polynomial in the dimensions: $\mu \ge 2C_p V/d^2 > 2C_p/V$, $\nu_0^{-1} \le V$, and $\log(1/\zeta) = d + \log(8d)$. The wall, extension constants and inverse tolerances are polynomial, so the schedule gives equation (46). Finally $\Phi(K^\star) = \mathcal{H}_{\mathcal{S}} + \lambda \operatorname{tr} K^\star/2$ and the student penalty is nonnegative. Thus $\epsilon_{\mathcal{S}}(X_t) \le B_Y \mu/2 + \nu$, which proves the gain claim. □

For the training theorem, take $\mu = 2C_p V/d^2$. Its ground-truth penalty and comparison error are then both of order V/d^2. We restate the result and make this specialization explicit.

Theorem C.5 (Restatement of Theorem 5.2: Tied training)**.** *Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ and suppose $\log V - \mathcal{H}_{\mathcal{S}} \ge \delta_{\mathrm{tr}} > 0$. For sufficiently large V and $\max\{2r, C\sqrt{V}\} \le d < N$, there is a choice $\lambda = C_\lambda/d^2$ and fixed $h, \sigma > 0$ with polynomially bounded inverses, together with a polynomial run length T divisible by d, such that, with probability at least $1 - e^{-cd}$,*

$$g_{\mathcal{S}}(X_t) \ge 1 - \frac{CV}{d^2 \delta_{\mathrm{tr}}} \quad \textit{for every integer } t \in [T/d, T].$$

Proof. Choose $C_\lambda = 2C_p$. A sufficiently large fixed width constant in $d \ge C\sqrt{V}$ ensures $\mu = 2C_p V/d^2 \le 1$ and $d \ge C' \log(eV)$ for all sufficiently large V. The assumption $d \ge 2r$ implies $d \ge 2r_Y$. Apply lemma C.4 with this μ. On its event, every integer $t \in [T/d, T]$ satisfies

$$\epsilon_{\mathcal{S}}(X_t) \le B_Y \mu/2 + (C_p B_Y + 1)\frac{V}{d^2} = (2C_p B_Y + 1)\frac{V}{d^2}.$$

Dividing by $\Delta_{\mathcal{S}} \ge \delta_{\mathrm{tr}}$ proves the stated gain bound. The lemma's schedule has fixed h, σ, a horizon divisible by d, and polynomially bounded inverse scales and time. Its event includes the ordinary Gaussian initialization and all perturbations, with probability at least $1 - e^{-cd}$. □

C.2 Proof of Theorem 5.4

The proof has three steps. Step 1 specializes the final-window comparison in lemma C.4 to the transfer decay, giving error $O(\mu^2)$ with $\mu = \lambda V$. Step 2 uses lemma C.6 to relate changes in factor energy to observed-logit error, and lemma C.7 to obtain training excess loss $O(\mu^2)$ and trace control. Step 3 uses lemma C.8 and the bounded learned rows to turn these estimates into a full Gram-matrix bound. Lemma C.9 gives the deterministic prediction implication. We then restate the theorem and substitute the decay scale.

Retain the tied notation from Appendix C.1. Under definition A.3, the chosen representative additionally satisfies

$$Y^\top \mathbf{1} = 0, \qquad D = H_N Y_{\mathcal{S}}, \qquad \sigma_{r_Y}(D) \geq \kappa\sqrt{N}.$$

Thus $L(Q) = K^\star = YY^\top$. Its diagonal entries imply $\max_i \|y_i\|^2 \leq B$, so $|K^\star_{ij}| \leq B$. Coverage also gives $r_Y \kappa^2 N \leq \|D\|_{\mathrm{F}}^2 \leq NB$. Constants in this section may depend on the fixed class and coverage parameters and the fixed proportions in $N = \Theta(V)$, but not on V, N, d. Dependence on a learned-row bound is stated separately.

C.2.1 Step 1: Compare the Final-Window Iterates with the Ground Truth

The comparison must use the transfer decay, not the training-only choice $\lambda = 2C_p/d^2$. We choose it as follows. For the transfer theorem choose the shrinking tied decay

$$\mu = C_\mu \frac{V^{3/2}}{\sqrt{N}\, d^{3/2}}, \qquad \lambda = \mu/V, \qquad d \geq C_{\mathrm{wid}} \frac{V}{N^{1/3}}, \qquad 2r_Y \leq d < N < V. \tag{48}$$

Choose C_μ large enough and then C_{wid} large enough, both depending only on class parameters. For all sufficiently large admissible dimensions these choices ensure $\lambda \geq 2C_p/d^2$, $\nu \leq C\mu^2$, and $0 < \mu \leq \mu_0 \leq 1$, where the class-small μ_0 is fixed below. Indeed $\lambda d^2 = C_\mu\sqrt{Vd/N}$, $\nu/\mu^2 = (C_p B_Y + 1)Nd/(C_\mu^2 V^2)$, and $\mu \leq C_\mu C_{\mathrm{wid}}^{-3/2}$.

Here μ_0 is chosen small enough for lemma C.7. Lemma C.4, whose complete proof is in Appendix C.1.3, now gives deterministic fixed h, σ and a polynomial horizon $T = dT_{\mathrm{blk}}$. With probability at least $1 - e^{-cd}$, every integer $t \in [T/d, T]$ satisfies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \Phi_{\lambda,\mathcal{S}}(K^\star) \leq \nu = (C_p B_Y + 1)V/d^2 \leq C\mu^2.$$

All remaining arguments are deterministic at each such iterate. In particular, the learned-row condition will not be included in this probability event.

C.2.2 Step 2: Convert the Objective Comparison into Observed Control

A decrease in the factor penalty could offset an increase in prediction loss. We therefore need a bound on how much the learned Gram trace can decrease. Coverage supplies this bound through relations between ground-truth feature vectors.

Reconstructing feature relations from training data. Coverage lets us express each ground-truth vector as $y_i = Y_{\mathcal{S}}^\top a_i$. Although input i may be unseen, it remains an output label, so column i of $Q_{\mathcal{S}}$ is available. Centered realizability lets us recover the coefficients from this population table:

$$A^\star = (\log Q_{\mathcal{S}})\mathsf{J}, \quad W = H_N A^\star, \quad C_{\mathcal{S}} = W P_{\mathcal{S}}^\top H_N, \quad a_i = C_{\mathcal{S}}^\dagger W_{:,i} \in \mathbb{R}^N. \tag{49}$$

The logarithm is entrywise and $\dagger$ denotes the Moore–Penrose inverse. Here $A^\star$ contains observed centered logits, W subtracts their observed-row mean, and $C_{\mathcal{S}} = DD^\top$.

For a learned factor, write

$$E = P_{\mathcal{S}} X X^\top \mathsf{J} - A^\star, \qquad e_1(X) = \frac{\|E\|_1}{NV}, \qquad b_X = \frac{\|\mathsf{J}X\|_{\mathrm{F}}^2}{V}, \qquad u_i(X) = x_i - X_{\mathcal{S}}^\top a_i.$$

Here E is the observed-logit error, not an untied factor, and x_i is row i of X viewed as a column vector. The residual u_i measures whether the learned embeddings satisfy the ground truth's relation to observed rows.

Throughout this section, the assumptions of definitions A.2 and A.3 hold. Put $\mathscr{A}(H) = P_{\mathcal{S}} H \mathsf{J}$ on the symmetric matrix space, $\Pi = Y(Y^\top Y)^{-1}Y^\top$, and $Q_\perp = I_V - \Pi$. The map $\mathscr{A}$ selects and centers observed logit rows; Π projects onto the column space of Y and $Q_\perp$ onto its orthogonal complement. Its Frobenius adjoint is $\mathscr{A}^*(B_0) = \mathrm{sym}(P_{\mathcal{S}}^\top B_0 \mathsf{J})$, where $\mathrm{sym}(H) = (H + H^\top)/2$. The learner need not be centered or stationary.

The following lemma proves the feature relations and a trace identity. Its nonnegative residual is what prevents an uncontrolled reduction in the factor energy.

Lemma C.6 (Observable anchors and a trace certificate)**.** *The coefficients in equation* (49) *satisfy*

$$a_i = D(D^\top D)^{-1} y_i, \qquad a_i^\top Y_{\mathcal{S}} = y_i^\top, \qquad a_i^\top A^\star = K^\star_{i,:},$$
$$\|a_i\|_2 \le \frac{\sqrt{B}}{\kappa\sqrt{N}}, \qquad \|a_i\|_\infty \le \frac{2B}{\kappa^2 N}. \tag{50}$$

In particular $K^\star = W^\top C_{\mathcal{S}}^\dagger W$ is identified by $Q_{\mathcal{S}}$. There is a matrix $Y_{\mathrm{dual}} \in \mathbb{R}^{N\times V}$, with $\|Y_{\mathrm{dual}}\|_\infty \le C/V$ and $\|Y_{\mathrm{dual}}\|_{\mathrm{F}} \le C$, such that

$$Z = I_V - \mathscr{A}^* Y_{\mathrm{dual}} \succeq Q_\perp, \qquad ZY = 0, \qquad \sum_i \|u_i(X)\|_2^2 = \mathrm{tr}(K-K^\star) - \langle Y_{\mathrm{dual}}, \mathscr{A}(K-K^\star)\rangle. \tag{51}$$

Proof. Since Y is centered, $A^\star = Y_{\mathcal{S}} Y^\top$, $W = DY^\top$ and $C_{\mathcal{S}} = DD^\top$. The identities $(DD^\top)^\dagger D = D(D^\top D)^{-1}$ and $D^\top Y_{\mathcal{S}} = D^\top D$ prove the interpolation statements and the reconstruction formula. Coverage gives the stated 2-norm bound; the row bound $\|D_s\| \le 2\sqrt{B}$ gives the ∞-norm bound.

Set

$$C_* = P_{\mathcal{S}}^\top D(D^\top D)^{-1}, \qquad G = Y^\top Y, \qquad Y_{\mathrm{dual}} = 2(P_{\mathcal{S}} C_*) Y^\top - (P_{\mathcal{S}} C_*) G C_*^\top.$$

Then $C_*^\top Y = I_{r_Y}$, $C_*^\top \mathbf{1} = 0$ and $Y^\top \mathbf{1} = 0$. Thus $Y_{\mathrm{dual}} \mathsf{J} = Y_{\mathrm{dual}}$ and

$$\mathscr{A}^* Y_{\mathrm{dual}} = \mathrm{sym}(P_{\mathcal{S}}^\top Y_{\mathrm{dual}} \mathsf{J}) = C_* Y^\top + Y C_*^\top - C_* G C_*^\top, \qquad Z = (I_V - Y C_*^\top)^\top (I_V - Y C_*^\top).$$

Consequently $ZY = 0$. For $v \perp \mathrm{range}\, Y$, the vectors v and $YC_*^\top v$ are orthogonal, so $v^\top Z v \ge \|v\|^2$. Together with $ZY = 0$, this proves $Z \succeq Q_\perp$. The rows of $(I_V - YC_*^\top)X$ are exactly $u_i^\top$; taking its squared Frobenius norm proves equation (51). Finally $\|(C_*)_i\| \le 2\sqrt{B}/(\kappa^2 N)$ and $\|G\|_{\mathrm{op}} \le BV$, whence

$$\|Y_{\mathrm{dual}}\|_\infty \le \frac{4B}{\kappa^2 N} + \frac{4B^2 V}{\kappa^4 N^2} \le C/V.$$

Multiplying this bound by $\sqrt{NV}$ proves the Frobenius bound. □

The trace identity uses logit error, whereas the objective comparison uses KL loss. We next connect the two without assuming bounded learned logits. Combining this connection with the trace identity gives the training-loss and trace estimates needed for completion.

Lemma C.7 (Training KL controls centered-logit L^1 error)**.** *For every finite X, without a bound on its logits,*

$$e_1(X) \le 2e^{2B}\{\epsilon_{\mathcal{S}}(X) + \sqrt{2\epsilon_{\mathcal{S}}(X)}\}. \tag{52}$$

If also $\mathcal{F}(X) - \Phi(K^\star) \le \nu$, then, for sufficiently small class-dependent $\mu > 0$,

$$\epsilon_{\mathcal{S}}(X) \le C(\nu + \mu^2), \qquad (\mathrm{tr}\, K - \mathrm{tr}\, K^\star)_+ \le 2\nu V/\mu,$$
$$b_X \le B + 2\nu/\mu, \qquad \sum_i \|u_i(X)\|^2 \le CV(\mu + \sqrt{\nu} + \nu/\mu).$$

Proof. For one observed row put $\ell_j = \log(p_j/q_j)$ and $k = D_{\mathrm{KL}}(q\|p)$. Since $\log z \le z - 1$,

$$\sum_j q_j(\ell_j)_+ \le \sum_j (p_j - q_j)_+ = \mathrm{TV}(p,q) \le \sqrt{k/2}.$$

Here the last inequality is Pinsker's inequality. A short derivation is to aggregate onto $\{j : p_j \ge q_j\}$ using the log-sum inequality and then use $\mathrm{kl}(a\|b) \ge 2(a-b)^2$ for Bernoulli laws; the latter follows by integrating its second derivative in a, $1/[a(1-a)] \ge 4$. Since $\sum_j q_j \ell_j = -k$, we obtain $\sum_j q_j|\ell_j| \le k + \sqrt{2k}$. Ground-truth logits in $[-B, B]$ give $q_j \ge e^{-2B}/V$. The centered logit error is $E_i = \ell^\top \mathsf{J}$, so $\|E_i\|_1 \le 2\|\ell\|_1 \le 2Ve^{2B}(k + \sqrt{2k})$. Average over $\mathcal{S}$ and apply Jensen to prove equation (52).

The trace certificate and $Z \succeq 0$ give

$$\mathrm{tr}(K - K^\star) \ge \langle Y_{\mathrm{dual}}, E\rangle \ge -CVe_1(X).$$

Writing $\epsilon = \epsilon_{\mathcal{S}}(X)$, the comparison assumption is $\epsilon + \lambda \operatorname{tr}(K - K^\star)/2 \le \nu$, hence

$$\epsilon \le \nu + C\mu(\epsilon + \sqrt{\epsilon}).$$

Choose μ_0 so that the coefficient of ϵ on the right is at most $1/2$, and then apply Young's inequality to the square-root term. This gives $\epsilon \le C(\nu + \mu^2)$. Nonnegativity of ϵ proves the trace-excess bound, and $\|\mathsf{J}X\|_{\mathrm{F}}^2 \le \operatorname{tr} K$ proves the bound on b_X. Finally equation (51) gives

$$\sum_i \|u_i\|^2 \le 2\nu V/\mu + CVe_1 \le CV\{\nu/\mu + \nu + \mu^2 + \sqrt{\nu} + \mu\}.$$

Since $\mu \le 1$, the terms ν and μ^2 are absorbed into ν/μ and μ, respectively, proving the claim. □

With $\nu \le C\mu^2$, these bounds give $\epsilon_{\mathcal{S}} \le C\mu^2$ and $(\operatorname{tr} K - \operatorname{tr} K^\star)_+ \le C\mu V$ at every final-window iterate. They also give $e_1 \le C\mu$, $b_X \le C$, and $\sum_i \|u_i\|^2 \le C\mu V$. The main transfer theorem uses the training-loss and trace bounds in the next step.

C.2.3 Step 3: Transfer the Observed Control to Unseen States

Observed rows control perturbations with no component wholly outside the ground-truth span. For a general learned Gram, the remaining component is positive semidefinite, so its trace bounds its Frobenius norm. The trace certificate controls this trace. The next lemma combines these estimates without restricting the learner's rank.

Lemma C.8 (From observed logits and trace to the full Gram matrix)**.** *For every symmetric H satisfying $Q_\perp H Q_\perp = 0$,*

$$\|H\|_{\mathrm{F}} \le C\|\mathscr{A}(H)\|_{\mathrm{F}}. \tag{53}$$

For every $K \succeq 0$ satisfying $\operatorname{tr} K \le \operatorname{tr} K^\star + \tau$, $\tau \ge 0$, one has

$$\|K - K^\star\|_{\mathrm{F}} \le C\{\|\mathscr{A}(K - K^\star)\|_{\mathrm{F}} + \tau\}.$$

Neither assertion restricts the rank of K.

Proof. Every such tangent matrix has a representation $H = YV_0^\top + V_0 Y^\top$. For example, with $G = Y^\top Y$ one may take $V_0 = HYG^{-1} - \frac{1}{2}YG^{-1}(Y^\top HY)G^{-1}$; substituting it and using $Q_\perp H Q_\perp = 0$ verifies the representation. Put $C_0 = \mathsf{J}V_0$, $E_0 = H_N(C_0)_{\mathcal{S}}$, and $W_0 = H_N\mathscr{A}(H)$. Then $W_0 = DC_0^\top + E_0 Y^\top$. Set $\Omega = D^\dagger E_0$, $E_\perp = (I - DD^\dagger)E_0$, and $R = C_0 + Y\Omega^\top$. The two terms of $W_0 = DR^\top + E_\perp Y^\top$ are Frobenius orthogonal. Coverage and $\sigma_{r_Y}(Y) \ge \sigma_{r_Y}(D)$ give, with $e = \|\mathscr{A}(H)\|_{\mathrm{F}}$,

$$\|R\|_{\mathrm{F}} + \|E_\perp\|_{\mathrm{F}} \le Ce/\sqrt{N}.$$

On the other hand $E_0 = D\Omega + E_\perp = H_N R_{\mathcal{S}} - D\Omega^\top$, whence $\|\Omega + \Omega^\top\|_{\mathrm{F}} \le Ce/N$. It follows that

$$\mathsf{J}H\mathsf{J} = YC_0^\top + C_0 Y^\top = YR^\top + RY^\top - Y(\Omega + \Omega^\top)Y^\top$$

has Frobenius norm at most Ce, since $\|Y\|_{\mathrm{op}} \le \sqrt{BV}$ and $N = \Theta(V)$. Because Y is centered, $\mathbf{1}^\top H \mathbf{1} = 0$, so $H = \mathsf{J}H\mathsf{J} + \mathbf{1}b^\top + b\mathbf{1}^\top$ for $b = H\mathbf{1}/V$ with $b^\top \mathbf{1} = 0$. Consequently $\mathscr{A}(H) = P_{\mathcal{S}}\mathsf{J}H\mathsf{J} + \mathbf{1}_N b^\top$. This gives $\sqrt{N}\|b\| \le e + \|\mathsf{J}H\mathsf{J}\|_{\mathrm{F}} \le Ce$ and proves equation (53).

For the second assertion write $\Delta = K - K^\star$ and $D_0 = Q_\perp K Q_\perp \succeq 0$. The trace certificate gives

$$\operatorname{tr} D_0 \le \langle Z, K\rangle = \operatorname{tr}\Delta - \langle Y_{\mathrm{dual}}, \mathscr{A}(\Delta)\rangle \le \tau + C\|\mathscr{A}(\Delta)\|_{\mathrm{F}}.$$

The remainder $H = \Delta - D_0$ satisfies $Q_\perp H Q_\perp = 0$. Therefore equation (53), the contraction property of $\mathscr{A}$, and $\|D_0\|_{\mathrm{F}} \le \operatorname{tr} D_0$ yield

$$\|\Delta\|_{\mathrm{F}} \le C\|\mathscr{A}(\Delta)\|_{\mathrm{F}} + (C+1)\operatorname{tr} D_0 \le C\{\|\mathscr{A}(\Delta)\|_{\mathrm{F}} + \tau\}.$$

This uses an ordinary linear tangent decomposition, with no claim that the actual learner or an optimizer remains on the ground truth's PSD face. □

It remains to relate the Gram error to prediction loss. Bounded learned rows keep every probability on each ground-truth-to-learner logit segment between fixed multiples of $1/V$. The lower loss-Hessian bound controls the observed Gram rows, and the upper bound converts the full Gram estimate into unseen loss. The next lemma collects this deterministic implication, separately from the optimization event.

Lemma C.9 (From objective comparison to test prediction)**.** *Assume definitions A.2 and A.3, with $N = \Theta(V)$. Fix $C_0 > 0$ and let $\lambda = \mu/V$ for sufficiently small $\mu > 0$. Every factor X satisfying*

$$\mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\mu}{2V}\|Y\|_{\mathrm{F}}^2\right) \le C_0\mu^2$$

has $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le C\mu^2$. If also $\max_j \|x_j\|^2 \le R_0$, then

$$\|XX^\top - YY^\top\|_{\mathrm{F}} \le C\mu V, \qquad \mathcal{R}_{\mathcal{U}_0}(X) - \mathcal{H}_{\mathcal{U}_0} \le \frac{C\mu^2 V}{|\mathcal{U}_0|}$$

for every nonempty $\mathcal{U}_0 \subseteq \mathcal{U}$ simultaneously. Here C may depend on C_0 and, in the last two bounds, R_0.

Proof. Put $\nu = C_0\mu^2$. The assumed comparison is exactly the premise of lemma C.7, since $\Phi(K^\star) = \mathcal{H}_{\mathcal{S}} + \mu\|Y\|_{\mathrm{F}}^2/(2V)$. For sufficiently small μ, that lemma gives $\epsilon_{\mathcal{S}} \le C\mu^2$, $(\operatorname{tr} K - \operatorname{tr} K^\star)_+ \le 2\nu V/\mu$, and $b_X \le C$. This proves the training part without a learned-row assumption.

For the unseen-loss bound, suppose the learned-row condition holds. We use lower loss curvature to obtain observed-logit accuracy, lemma C.8 to recover the full Gram, and upper curvature to obtain unseen loss. Put $\Delta = XX^\top - K^\star$ and $e = \|P_{\mathcal{S}}\Delta\mathsf{J}\|_{\mathrm{F}}$. Ground truth raw logits lie in $[-B, B]$ and the learned raw logits lie in $[-R_0, R_0]$. Thus every ground-truth-to-learner softmax segment has atoms between c_{R_0}/V and C_{R_0}/V. For a centered vector v the categorical variance obeys

$$\frac{c_{R_0}}{V}\|v\|^2 \le \operatorname{Var}_p(v) = \min_a \sum_j p_j(v_j - a)^2 \le \frac{C_{R_0}}{V}\|v\|^2.$$

Taylor's integral formula and the training bound just proved give $c_{R_0}e^2/(2NV) \le \epsilon_{\mathcal{S}} \le C\mu^2$, so $e \le C_{R_0}\mu\sqrt{NV} \le C_{R_0}\mu V$. The comparison bound gives trace excess at most $\tau = 2\nu V/\mu \le C\mu V$. By lemma C.8, $\|\Delta\|_{\mathrm{F}} \le C_{R_0}\mu V$. The upper variance bound on any set of rows now gives

$$\epsilon_{\mathcal{U}_0}(X) \le \frac{C_{R_0}}{|\mathcal{U}_0|V}\|P_{\mathcal{U}_0}\Delta\mathsf{J}\|_{\mathrm{F}}^2 \le C_{R_0}\mu^2 V/|\mathcal{U}_0|.$$

This proves the claim simultaneously for all nonempty unseen sets.

□

We have obtained the two loss bounds on the same optimization event. We now restate the main theorem and convert them to RPG.

Theorem C.10 (Restatement of Theorem 5.4: Conditional tied transfer)**.** *Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ admit a representative Y satisfying definition A.3. Suppose $\log V - \mathcal{H}_{\mathcal{S}} \ge \delta_{\mathrm{tr}} > 0$ and $\log V - \mathcal{H}_{\mathcal{U}} \ge \delta_{\mathrm{te}} > 0$. For sufficiently large V, let $\max\{2r, CV/N^{1/3}\} \le d < N$ and $\lambda = C_\lambda\sqrt{V/(Nd^3)}$. There are deterministic h, σ, T of polynomial size, with polynomially bounded inverse scales, and an event of probability at least $1 - e^{-cd}$ on which every integer $t \in [T/d, T]$ satisfies*

$$g_{\mathcal{S}}(X_t) \ge 1 - \frac{CV^3}{Nd^3\delta_{\mathrm{tr}}}.$$

Moreover, at every such iterate satisfying $\max_{j\in[V]}\|x_j(t)\|_2^2 \le R_0$ for a fixed R_0, the complete unseen set satisfies

$$g_{\mathcal{U}}(X_t) \ge 1 - \frac{CV^4}{Nd^3(V-N)\delta_{\mathrm{te}}}.$$

The optimization analysis does not establish the learned-row bound. The constant C in the unseen-set conclusion may depend on R_0.

Proof. Choose $C_\lambda = C_\mu$ in equation (48), first large enough to ensure $\lambda \ge 2C_p/d^2$. Then choose the width constant sufficiently large that $\mu \le C_\mu C_{\mathrm{wid}}^{-3/2} \le \mu_0$. For sufficiently large V, the width also implies $d \ge C'\log(eV)$, and $d \ge 2r$ implies $d \ge 2r_Y$. Step 1 therefore applies lemma C.4 with $\nu \le C\mu^2$ and supplies the stated polynomial schedule and common event of probability at least $1 - e^{-cd}$.

Fix any integer $t \in [T/d, T]$ on this event. Lemma C.9 gives $\epsilon_{\mathcal{S}}(X_t) \le C\mu^2$. At every such iterate satisfying the learned-row bound, the same lemma with $\mathcal{U}_0 = \mathcal{U}$ gives $\epsilon_{\mathcal{U}}(X_t) \le C\mu^2 V/(V-N)$. Substituting $\mu^2 = C_\mu^2 V^3/(Nd^3)$ and dividing by $\Delta_{\mathcal{S}} \ge \delta_{\mathrm{tr}}$ and $\Delta_{\mathcal{U}} \ge \delta_{\mathrm{te}}$ proves both gain bounds. The implications are deterministic for every iterate on the common event; no probability for occurrence of the learned-row bound is used. □

D Auxiliary Results for the Main Proofs

This appendix collects the tools shared by the main theorem proofs. The factor and derivative lemmas support the approximate-SOSP comparisons, while the localization and perturbation lemmas support the first and third steps of the training proofs.

D.1 Bounded Factorizations and Trace Packing

The next lemma turns bounded entries and low rank into bounded factor rows. In particular, its total factor energy is linear in V. For this lemma put $\Gamma = B\sqrt{r}$.

Lemma D.1 (Rank versus max-norm). *If $M \in \mathbb{R}^{V\times V}$ has rank $r_0 \le r$ and $\|M\|_\infty \le B$, then there are $X, Y \in \mathbb{R}^{V\times r_0}$ such that*

$$M = XY^\top, \qquad \max_i \|X_{i,:}\|_2^2 \le \Gamma, \qquad \max_j \|Y_{j,:}\|_2^2 \le \Gamma. \tag{54}$$

If $M\mathbf{1} = 0$, the factors can be chosen with $Y^\top \mathbf{1} = 0$. Consequently,

$$\|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2 \le 2\Gamma V. \tag{55}$$

Proof. Take a full-rank factorization $M = A_0 B_0^\top$ in latent dimension r_0. Let $\mathcal{K}$ be the centrally symmetric convex hull of the rows b_j of B_0. Since B_0 has full column rank, $\mathcal{K}$ is full-dimensional. We use the symmetric Löwner–John ellipsoid theorem (John, 2014): if the minimum-volume ellipsoid containing a centrally symmetric convex body $K \subset \mathbb{R}^n$ is the Euclidean unit ball, then $n^{-1/2}\mathcal{B}_2 \subseteq K \subseteq \mathcal{B}_2$. Apply an invertible linear change of coordinates so that the minimum-volume ellipsoid containing $\mathcal{K}$ becomes the unit ball. The theorem then gives

$$r_0^{-1/2}\mathcal{B}_2 \subseteq \mathcal{K} \subseteq \mathcal{B}_2.$$

Here $\mathcal{B}_2$ is the Euclidean unit ball in $\mathbb{R}^{r_0}$. Apply the inverse-transpose coordinate change to the rows a_i of A_0, so all inner products remain M_{ij}. Now $\|b_j\|_2 \le 1$ and $|a_i^\top b_j| \le B$. Write $\mathcal{K}^\circ = \{z : \sup_{b\in\mathcal{K}} z^\top b \le 1\}$ for the polar set. Then $a_i/B \in \mathcal{K}^\circ$, while the first John inclusion gives $\mathcal{K}^\circ \subseteq \sqrt{r_0}\mathcal{B}_2$. Thus $\|a_i\|_2 \le B\sqrt{r_0}$. Rescale the two factors reciprocally so that both maximum row norms are at most $\sqrt{B\sqrt{r_0}} \le \sqrt{\Gamma}$. This proves equation (54).

The left factor has full column rank. If $M\mathbf{1} = 0$, then $X(Y^\top\mathbf{1}) = 0$, so $Y^\top\mathbf{1} = 0$. Summing the squared row bounds gives equation (55). □

We next give the general trace-packing argument used for the untied categorical loss. It chooses comparator directions whose factor energy and total row-Hessian contribution are both small.

Lemma D.2 (Trace-budget hidden-channel packing). *Fix a point of a row-separable convex loss $\mathcal{L}(Z) = n^{-1}\sum_{i=1}^n \ell_i(Z_i)$, and write $H_i = \nabla^2\ell_i(Z_i) \succeq 0$. Let $E \in \mathbb{R}^{n\times d}$ and $U \in \mathbb{R}^{m\times d}$ obey*

$$\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2 \le C_E n, \qquad \sum_{i=1}^n \operatorname{tr} H_i \le \tau n.$$

Suppose a comparator $M = XY^\top$ has rank $r_0 \le r$, where $X = [x_1, \dots, x_{r_0}]$, $Y = [y_1, \dots, y_{r_0}]$, and

$$\max_{i,k} x_{ik}^2 \le a, \qquad \max_{i,k} y_k^\top H_i y_k \le \nu.$$

If $d \ge 2r$, there are $A \in \mathbb{R}^{n\times d}$ and $C \in \mathbb{R}^{m\times d}$ satisfying

$$AC^\top = M, \qquad \|A\|_{\mathrm{F}}^2 + \|C\|_{\mathrm{F}}^2 = \|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2. \tag{56}$$

$$\frac{1}{n}\sum_{i=1}^n (AU^\top + EC^\top)_i^\top H_i (AU^\top + EC^\top)_i \le \frac{8rC_E}{d}\left(\frac{4a\tau n}{d} + \nu\right). \tag{57}$$

Proof. Choose an orthonormal eigenbasis $w_1, \dots, w_d$ of $U^\top U$ and put

$$u_j = \|Uw_j\|_2^2, \qquad e_j = \|Ew_j\|_2^2, \qquad s_j = u_j + e_j.$$

For $u_j > 0$, set $v_j = Uw_j/\sqrt{u_j}$ and $h_j = \sum_i v_j^\top H_i v_j$; set $h_j = 0$ if $u_j = 0$. The nonzero v_j are orthonormal, so

$$\sum_j s_j \le C_E n, \qquad \sum_j h_j \le \sum_i \operatorname{tr} H_i \le \tau n.$$

At most $d/4$ indices violate each of the bounds $s_j \le 4C_E n/d$ and $h_j \le 4\tau n/d$. At least $d/2 \ge r$ therefore satisfy both. Choose distinct such indices $j_1, \dots, j_{r_0}$ and independent uniform signs $\varepsilon_k \in \{-1, 1\}$, and put

$$A_\varepsilon = \sum_{k=1}^{r_0} \varepsilon_k x_k w_{j_k}^\top, \qquad C_\varepsilon = \sum_{k=1}^{r_0} \varepsilon_k y_k w_{j_k}^\top.$$

Orthonormality proves both identities in equation (56) for every sign choice. Averaging over signs cancels cross terms between distinct comparator columns. The contribution from column k is at most

$$\frac{2}{n}\{a u_{j_k} h_{j_k} + \nu e_{j_k}\} \le \frac{8C_E}{d}\left(\frac{4a\tau n}{d} + \nu\right).$$

Summing over at most r terms and choosing a sign realization no worse than the expectation proves equation (57). □

Thus the usable width is controlled by a Hessian trace budget, not by a uniform lower curvature bound. This distinction matters for saturated categorical rows, where restricted strong convexity can fail while the trace budget below remains valid.

D.2 Untied Factor Derivatives

The next identities explain how an objective comparison in logits enters the factor Hessian. The bilinear term may be negative even though the categorical loss is convex in its logits. Let $h_0 = \mathbf{1}/\sqrt{V}$ and $C_i = \operatorname{diag}(p_i) - p_i p_i^\top$, where $p_i = \operatorname{softmax}((EU^\top)_{i,:})$ is viewed as a column vector.

Lemma D.3 (Gradient, balance, and Hessian identities)**.** *Fix $\lambda > 0$. Let $U_c = \mathsf{J}U$, $Z = EU_c^\top$, and $G = (P - Q)/V$. In the raw model coordinates (E, U),*

$$\nabla_E \mathcal{F}_\lambda = GU + \lambda E, \qquad \nabla_U \mathcal{F}_\lambda = G^\top E + \lambda U. \tag{58}$$

For raw variations (A, C), put $C_c = \mathsf{J}C$ and $D = AU_c^\top + EC_c^\top$. Write

$$\mathscr{Q}_{\mathcal{F}}(A, C) = \nabla^2 \mathcal{F}_\lambda(E, U)[(A, C), (A, C)].$$

The exact Hessian quadratic is

$$\mathscr{Q}_{\mathcal{F}}(A, C) = \frac{1}{V}\sum_i D_i^\top C_i D_i + 2\langle G, AC^\top\rangle + \lambda(\|A\|_{\mathrm{F}}^2 + \|C\|_{\mathrm{F}}^2). \tag{59}$$

The first term is nonnegative softmax curvature. The second comes from the bilinear factor product, and the last from weight decay. At every critical point,

$$h_0^\top U = 0, \qquad E^\top E = U^\top U, \qquad \|Z\|_* = \frac{1}{2}(\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2), \qquad \langle G, Z\rangle + \lambda\|Z\|_* = 0.$$

Proof. Because $G\mathbf{1} = 0$, one has $GU = GU_c$ and the loss part of the U-gradient lies in the centered output subspace. Differentiating $Z = EU_c^\top$ twice gives the first variation D above. In the second variation, $\langle G, AC_c^\top\rangle = \langle G, AC^\top\rangle$. Using the categorical covariance $C_i = \operatorname{diag}(p_i) - p_i p_i^\top$ gives equation (59).

At a critical point, projecting the second equation in equation (58) onto h_0 first gives $h_0^\top U = 0$ by $\lambda > 0$. Multiply the first equation in equation (58) on the left by $E^\top$, transpose, and compare it with the second equation multiplied on the left by $U^\top$. This gives $E^\top E = U^\top U$. The two factors are balanced, so their common Gram matrix contains the nonzero singular values of Z and yields the nuclear-norm identity. Finally, inner-product equation (58) with (E, U) and divide by two. □

D.3 Smooth Localization

The perturbation results assume global smoothness, while our factor objectives are controlled on an energy ball. The following construction extends the objective without changing it in that ball and preserves the objective barrier outside. It is an analysis construction, not an extra operation in the training rule.

Lemma D.4 (Smooth radial localization)**.** *Let a factor ball of squared radius CV contain Θ_0 and a global minimizer of $\mathcal{F}_\lambda$, and suppose*

$$\mathcal{F}_\lambda(\Theta) \geq \mathcal{F}_\lambda(\Theta_0) + 4 \qquad \textit{whenever } \|\Theta\|_2^2 \geq CV.$$

For any fixed relative-radius buffer, there is a smooth $\widetilde{\mathcal{F}}_\lambda : \mathbb{R}^{D_\Theta} \to \mathbb{R}$ which equals $\mathcal{F}_\lambda$ on the inner ball, is at least $\mathcal{F}_\lambda(\Theta_0) + 2$ outside it, satisfies $\widetilde{\mathcal{F}}_\lambda \geq \mathcal{F}_\lambda^*$, and has globally finite gradient- and Hessian-Lipschitz constants. If the derivatives of $\mathcal{F}_\lambda$ through order three on the buffered ball and the buffer radius are polynomial in V, those global constants are polynomial in V. More generally, if the assumed wall has the absolute lower bound $\mathcal{F}_\lambda \geq M$ outside the inner ball, the extension may be chosen with $\widetilde{\mathcal{F}}_\lambda \geq M$ there as well.

Proof. Let $R = \sqrt{CV}$ and fix $c > 0$, independent of V. A fixed relative buffer enlarges the radius by a dimension-independent factor. Choose a C^∞ scalar map $s : [0, \infty) \to [0, \infty)$ which equals the identity on $[0, R]$, remains in the fixed annulus $[R, (1 + c)R]$ after R, and is constant after $(1 + 2c)R$. It can be chosen so that all derivatives match at the junctions. Define the radial map

$$\Phi(\Theta) = \begin{cases} s(\|\Theta\|_2)\Theta/\|\Theta\|_2, & \Theta \neq 0, \\ 0, & \Theta = 0. \end{cases}$$

Then Φ is the identity on the inner ball and maps every exterior point to the closed buffered annulus. Put $\widetilde{\mathcal{F}}_\lambda = \mathcal{F}_\lambda \circ \Phi$. The assumed wall separation makes this composition at least $\mathcal{F}_\lambda(\Theta_0) + 4$, hence at least the stated two-unit margin, outside the inner ball. It is everywhere at least $\mathcal{F}_\lambda^*$, while the global minimizer inside the ball attains equality. The same construction preserves any stronger absolute lower bound M valid on the entire buffered exterior annulus.

The fixed relative buffer bounds the first derivative norm by $O(1)$, the second by $O(R^{-1})$, and the third by $O(R^{-2})$. Its image lies in the compact buffered ball, where the derivatives of $\mathcal{F}_\lambda$ through order three are bounded. The chain rule therefore bounds the second and third derivatives of the composition globally, giving the required Lipschitz constants. These bounds also preserve polynomial dependence on V. □

D.4 Perturbed-Gradient Visits and Objective Increases

The next two results are external inputs from Jin et al. (2019). We state them in the exact-gradient noise convention used in this paper. The visit theorem finds an approximate SOSP; the concentration inequality controls what can happen between visits. Their specialization to our fixed schedule and the block argument are proved below.

Theorem D.5 (Jin–et al. always-perturbed visit theorem)**.** *Let $f : \mathbb{R}^D \to \mathbb{R}$ have an ℓ-Lipschitz gradient and a ρ-Lipschitz Hessian, and let $\Delta_f = f(x_0) - \inf_x f(x) < \infty$. For every $\epsilon > 0$ and failure level $\zeta_{\mathrm{J}} > 0$, Algorithm 1 and Theorem 13 of Jin et al. (2019), specialized to exact gradients, show that*

$$x_{t+1} = x_t - h\{\nabla f(x_t) + \xi_t\}, \qquad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0, r_{\mathrm{J}}^2 I_D/D),$$

with the source's prescribed $h = \widetilde{\Theta}(1/\ell)$ and $r_{\mathrm{J}} = \widetilde{\Theta}(\epsilon)$, visits a point satisfying

$$\|\nabla f(x_t)\| \leq \epsilon, \qquad \nabla^2 f(x_t) \succeq -\sqrt{\rho\epsilon}\, I$$

within $\widetilde{O}(\ell\Delta_f\epsilon^{-2})$ iterations, with probability at least $1 - \zeta_{\mathrm{J}}$. The hidden factors are polylogarithmic in $D, \ell, \rho, \epsilon^{-1}, \zeta_{\mathrm{J}}^{-1}$, and Δ_f.

The visit guarantee alone does not control a whole interval of iterates. The following concentration inequality supplies the required objective-increase bound.

Lemma D.6 (Jin–et al. descent concentration)**.** *Under the assumptions of theorem D.5, fix deterministic integers $t_0 \geq 0$ and $s \geq 1$. If $h \leq 1/\ell$, Lemma 23 of Jin et al. (2019), again specialized to exact gradients, gives*

$$f(x_{t_0+s}) - f(x_{t_0}) \leq -\frac{h}{8}\sum_{j=0}^{s-1} \|\nabla f(x_{t_0+j})\|^2 + c_J h r_{\mathrm{J}}^2(h\ell s + \iota) \tag{60}$$

with probability at least $1 - 4e^{-\iota}$, where c_J is universal.

Source and specialization. This is Lemma 23 of Jin et al. (2019) with the stochastic-gradient variance set to zero. Its perturbation is the isotropic Gaussian $\mathcal{N}(0, r_{\mathrm{J}}^2 I_D/D)$ used in theorem D.5; thus r_{J}^2 is the total perturbation second moment. The remaining stochastic term and its failure bound are exactly equation (60). □

To match the notation of Jin et al. (2019), set its stochastic-gradient variance to zero. The auxiliary quantity denoted N in its Appendix B then equals one and is unrelated to our state count. Fix universal constants $C_\iota, C_J \geq 1$ large enough to absorb the numerical constants in its displayed exact-gradient choices. Those choices reduce to

$$h = (\iota^9\ell)^{-1}, \qquad r_{\mathrm{J}} = \iota\epsilon, \qquad T_{\mathrm{J}} = \frac{\iota}{h\sqrt{\rho\epsilon}}, \qquad \delta_{\mathrm{J}} = \iota^{-5}\sqrt{\frac{\epsilon^3}{\rho}}.$$

Consequently, for an objective-gap bound A_0,

$$A_0 \max\left\{\frac{T_{\mathrm{J}}}{\delta_{\mathrm{J}}}, \frac{1}{h\epsilon^2}\right\} = O(\ell A_0 \epsilon^{-2}\iota^{15}), \tag{61}$$

which is the horizon used below.

We now combine these two inputs. Requiring the increase bound for every forward pair allows it to be used from a random SOSP visit time, without changing the algorithm or conditioning on a future event.

Lemma D.7 (One block with a fixed perturbation scale)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be bounded below, with global smoothness constants $\ell, \rho \geq 1$, and $f(x_0) - \inf f \leq A$, where $A \geq 1$. For $0 < \varepsilon, \delta_c \leq 1$ and $0 < \zeta < 1/4$, take a sufficiently large universal constant in*

$$\iota \geq C\left[1 + \log\frac{ep(\ell+\rho+1)(A+1)}{\varepsilon\zeta} + \frac{A+1}{\delta_c} + \delta_c^{-1/6}\right],$$
$$h = (\iota^9\ell)^{-1}, \qquad \sigma = \iota\varepsilon, \qquad T_{\mathrm{blk}} = \left\lceil C_J \ell A \varepsilon^{-2}\iota^{15}\right\rceil, \tag{62}$$

and enlarge ι to ensure $4(T_{\mathrm{blk}}+1)^2 e^{-\iota} \leq \zeta$. With probability at least $1 - 2\zeta$, one block of the recursion visits a point with $\|\nabla f\| \leq \varepsilon$ and $\nabla^2 f \succeq -\sqrt{\rho\varepsilon}I$; simultaneously every forward pair of iterates in the block has objective increase at most δ_c. The statement holds conditionally from a random block start with the same objective budget and fresh future perturbations.

Proof. Apply theorem D.5 with the exact-gradient variance set to zero. The source parameter N_{J} is then one, and its displayed step and radius become those in equation (62). Enlarging ι absorbs the source's logarithmic failure parameter. Its escape time $\iota/(h\sqrt{\rho\varepsilon})$ and drop scale $\iota^{-5}\sqrt{\varepsilon^3/\rho}$ give the conservative visit horizon $C_J\ell A\varepsilon^{-2}\iota^{15}$, as in equation (61). For a deterministic starting time and an interval of length s, lemma D.6 bounds the increase by

$$-\frac{h}{8}\sum_{j=0}^{s-1}\|\nabla f(x_{t+j})\|^2 + c_J h\sigma^2(h\ell s + \iota)$$

except with probability $4e^{-\iota}$. For $s \leq T_{\mathrm{blk}}$ the positive term is at most $C\{(A+1)/\iota + \varepsilon^2/(\ell\iota^6)\} \leq \delta_c$. The pair condition permits a union bound over all start/end pairs; combining it with the visit event costs at most 2ζ. In particular the increase bound applies from the random visit time. Conditioning on a block start fixes only the past and preserves the future independent Gaussian law. Finally, the extra condition on ι admits a polynomial-size choice: T_{blk} is polynomial in ι, whereas $e^{-\iota}$ decays exponentially. □

Finally, the following elementary induction explains why repeated visits give a guarantee at every late iterate. It is independent of the model and does not require independence between block-success events.

Lemma D.8 (Refreshing the prediction margin)**.** *For an objective $\mathcal{F}$, partition $\Theta_0, \dots, \Theta_T$ into d blocks of length T_{blk}, sharing their boundary iterates, with $T = dT_{\mathrm{blk}}$. Fix $q \in \mathbb{R}$ and $b > 0$. Suppose initialization succeeds with probability at least $1 - \delta_0$. Conditional on this event and success of all preceding blocks, suppose each block succeeds with probability at least $1 - \delta$, where success means both*

$$\mathcal{F}(\Theta_\tau) \le q - 2b \text{ for some time } \tau \text{ in the block},$$
$$\mathcal{F}(\Theta_v) - \mathcal{F}(\Theta_u) \le b \text{ for all } u < v \text{ in the block}.$$

Then $\mathcal{F}(\Theta_t) \le q$ for every integer $t \in [T_{\mathrm{blk}}, T]$, with probability at least $1 - \delta_0 - d\delta$. Strict visits give a strict conclusion.

Proof. A successful visit puts the block endpoint at objective at most $q - b$. Starting the next block there, the increase bound keeps every iterate at most q, and its new visit restores the endpoint margin. Induction proves the claim. Conditional failure bounds give total failure at most $\delta_0 + d\delta$; block events need not be independent. □

E Experimental Protocol and Complete Summaries

The experiments measure behavior of the same raw-factor population models and fixed-perturbation optimizer family as the theory. Their practical constants do not instantiate the certified asymptotic inequalities. All seeds and configurations were fixed before the reported suite. A short software check used different ground-truth and initialization seeds; it did not change the reported parameter recipe. No ground truths, runs, states, or checkpoints were selected according to favorable test outcomes.

E.1 Ground Truths, Observed states, and Initialization

For each $V \in \{64, 128, 256\}$, we generate three ground truths with seeds $s \in \{0, 1, 2\}$. NumPy's default random generator uses seed $10000 + 100V + s$. Normalize each independent four-dimensional standard Gaussian row to unit norm. For the untied all-state problem, draw two such matrices A, B, center the rows of B across states, and divide B by its largest row norm. Set $L(Q) = 3AB^\top$ and $Q = \mathrm{softmax}(L(Q))$. These directed transitions have rank-four centered logits bounded by three; the rank condition is not on Q itself.

For tied and matched held-out untied problems, center the first normalized matrix across states and rescale it so that its largest row norm is $\sqrt{3}$. Call the result Y, and set $Q = \mathrm{softmax}(YY^\top)$. The first $3V/4$ indices of the generator's subsequent random permutation form $\mathcal{S}$; $\mathcal{U} = [V] \setminus \mathcal{S}$ is fixed before training. All V states remain output labels. Thus "unseen" refers only to the current/input role, not to an output label absent from training. The centered-logit ranks, bounds, observed covariance eigenvalues, and information gaps are checked and saved for every ground truth. There is no rejection-sampling step. Across the nine tied ground truths, the smallest observed covariance eigenvalue is 0.28957, and the smallest observed or unseen information gap is 0.47702.

For each ground truth use initialization seeds $a \in \{10, 11, 12, 13, 14\}$, with NumPy generator seed $1000000 + 1000s + a$. Draw every actual entry of E, U or X independently as $\mathcal{N}(0, 1/d)$ and store it in FP32. No learner factor is centered, balanced, projected, normalized, frozen, or aligned with the ground truth. The implementation computes all gradients from the same current factors before applying either untied update.

E.2 Optimization Recipe and Controls

The main grid has $d \in \{V/8, V/4, V/2\}$ for each of the three state counts and both models: 18 settings, with 15 runs each. Every run uses $T = 8192$, step $h = V/4$, and total gradient-perturbation scale $\sigma = 0.01/\sqrt{V}$. The decay is

$$\lambda = \frac{1}{480V} \quad \text{(full-untied)}, \qquad \lambda = 0.1\sqrt{\frac{V}{Nd^3}} \quad \text{(tied and matched held-out untied)}.$$

The first value is theorem 4.1 with $r = 4$, $B = 3$, and $\Delta_0 = 0.1$. The second uses the transfer scale with a practical prefactor. In particular, the tied proof's $C_p = 40B_Y r = 480$ and requirements $\lambda \geq 2C_p/d^2$, $\lambda V \leq 1$ would imply $d^2 \geq 960V$. The tested tied widths cannot satisfy these inequalities. Their results therefore illustrate behavior outside the certified constant regime, rather than verify or falsify those inequalities.

The loss sums over all output classes and averages uniformly over training input rows; no transitions are sampled. The generated FP64 table Q is rounded to FP32 and row-renormalized for training. Structural ground-truth checks and the independent checkpoint audit use the original FP64 table. Weight decay is the squared factor-norm sum times $\lambda/2$, without per-parameter normalization. At every update add a fresh Gaussian gradient vector of covariance $\sigma^2 I/D_\Theta$, then multiply the full gradient by h. Here $D_\Theta = 2Vd$ for untied and Vd for tied. The perturbation generator uses seed $271828 + 100V + d$, independently of factor initialization. Its sequence is shared across paired noisy optimizer arms with the same model shape. There is no noise cutoff or stopping rule based on loss or test data.

At $V = 128, d = 32$, we add three controls for each model: zero noise, zero decay, and both zero. These 90 runs retain the same step, horizon, ground truths, and initial factors. Fifteen additional untied runs use the tied ground truths, split, and parameter recipe to compare sharing on a matched transfer problem. In total there are 25 settings and 375 runs. Seeds are reused to pair settings; these are not 375 independent Bernoulli trials. The preliminary software check used $V = 64, d = 16$, 256 updates, ground-truth seed 90, and initialization seeds 91 and 92. It shares the formula for perturbation seeds and is not claimed to be an independent statistical validation set.

E.3 Measurements and Numerical Checks

At every update $t = 0, \ldots, T$, we record the training gain, the full-unseen gain when applicable, and the norm of the largest squared row. The late window includes both endpoints T/d and T. For each run, we compute its minimum gain, fraction of steps above $7/8$, and longest consecutive below-threshold interval in this window. Raw factors are saved at $0, T/d, T$; objective, energy per state, and largest scaled predicted probability are recorded every 16 updates. Test statistics for all-state training are explicitly absent from summaries. Divergent runs, if any, remain in the denominator, with nonfinite summaries stored as null; none occurred in this suite.

The closed-form gradients agree with independent automatic differentiation to maximum absolute error 1.12×10^{-16} in FP64 tests. A noise-variance check, a softmax-centering invariance check, and the held-out input-gradient identity are also tested. The audit recomputes every reported late-window statistic from the complete traces, verifies their SHA-256 hashes, and independently evaluates the original ground-truth risks in FP64 at all 1125 saved checkpoints. The largest gain discrepancy is 2.99×10^{-7}.

Runs used an NVIDIA RTX 3090, PyTorch 2.10.0+cu128, NumPy 2.2.6, and FP32 parameters, with deterministic algorithms enabled and TF32 disabled. The summed recorded training-loop wall-clock time was 40.7 seconds, including checkpoint transfers but excluding file compression and independent audits. This is a rough runtime, not a device-only benchmark: the host timer starts without a preceding CUDA synchronization. The executable protocol, dependency versions, raw traces, checkpoints, and figure commands are included in the accompanying package. Bitwise equivalence of GPU results on different hardware or library versions is not asserted.

The full-untied runs maintain high gain throughout the measured late window. Tied transfer improves in the wider tested settings, while the narrowest tied setting has lower training gain and the matched untied model has near-zero unseen gain. Observed row bounds give finite-instance examples of the transfer condition. Comparable performance without noise and lower final training error without decay provide no evidence about the frequency of the constructed bad wells. These experiments do not estimate an optimal width exponent or asymptotic failure rate.

Table 2: Every reported setting, with 15 runs each. "Min." is the minimum over all runs and every late-window step, not the final value. "Kept" counts runs whose entire late window has gain $> 7/8$ (training/test). "Max. row" is the largest squared row norm over the whole run and all factors; every run passes the predeclared late bound of eight. "Both" means decay and noise; Untied-H is the matched held-out control. Missing test entries mean that all input states were used for training. Values are rounded for display; threshold tests use unrounded traces.

Model	Arm	V	d	Min. train	Min. test	Kept tr./te.	Max. row
Untied	Both	64	8	0.9999	–	15/–	4.493
Tied	Both	64	8	0.8607	0.6150	0/0	3.246
Untied	Both	64	16	0.9998	–	15/–	4.271
Tied	Both	64	16	0.9762	0.8824	15/15	2.869
Untied	Both	64	32	0.9993	–	15/–	4.302
Tied	Both	64	32	0.9964	0.9728	15/15	3.124
Untied	Both	128	16	0.9998	–	15/–	4.313
Tied	Both	128	16	0.9228	0.8024	15/0	2.522
Untied	Both	128	32	0.9993	–	15/–	4.217
Tied	Both	128	32	0.9883	0.9636	15/15	3.020
Untied	Both	128	64	0.9978	–	15/–	4.227
Tied	Both	128	64	0.9982	0.9945	15/15	3.491
Untied	Both	256	32	0.9993	–	15/–	4.254
Tied	Both	256	32	0.9587	0.8962	15/15	2.694
Untied	Both	256	64	0.9976	–	15/–	4.210
Tied	Both	256	64	0.9941	0.9835	15/15	3.215
Untied	Both	256	128	0.9951	–	15/–	4.356
Tied	Both	256	128	0.9974	0.9925	15/15	3.733
Untied	Decay	128	32	0.9993	–	15/–	4.221
Tied	Decay	128	32	0.9884	0.9638	15/15	3.020
Untied	Noise	128	32	0.9992	–	15/–	4.306
Tied	Noise	128	32	0.9998	0.9754	15/15	4.041
Untied	Neither	128	32	0.9992	–	15/–	4.271
Tied	Neither	128	32	0.9998	0.9758	15/15	4.044
Untied-H	Both	128	32	0.9708	-0.0013	15/0	3.123

F Scaling with Sampled Transitions

We examine training and unseen-state prediction over a grid of state counts, teacher ranks, and embedding widths, using sampled-transition SGD. Figures 4–6 show all 12 choices of (V, r), with every valid width and all three initialization seeds retained.

Models and configuration grid. The state space is $[V] = \{1, \dots, V\}$. The untied model predicts $P = \mathrm{softmax}(EU^\top)$ with $E, U \in \mathbb{R}^{V \times d}$; the tied model predicts $P = \mathrm{softmax}(XX^\top)$ with $X \in \mathbb{R}^{V \times d}$. Softmax acts rowwise. We use $V = 256$, 1024, 4096, and 16384, teacher ranks $r = 8, 16, 32$, and candidate widths $d = 32, 128, 256, 1024$. Untied training observes all V current states. Tied training observes a fixed random set $\mathcal{S}$ with $N = |\mathcal{S}| = 3V/4$ and is evaluated on the unseen set $\mathcal{U} = [V] \setminus \mathcal{S}$ of size $V/4$. Thus the tied training/unseen counts are $192/64$, $768/256$, $3072/1024$, and $12288/4096$, respectively. All V states remain output labels: an unseen state's embedding is still updated through its output role. We retain exactly the widths with $r \le d < N$, using $N = V$ for the untied model. This gives 78 model configurations and 234 runs across initialization seeds $10, 11, 12$.

Ground-truth transitions. For untied ground truths, draw independent $V \times r$ Gaussian matrices A, B, normalize each row, and multiply A by three. Subtract the mean row from B, divide it by its largest row norm, and set $Q = \mathrm{softmax}(AB^\top)$. For tied ground truths, normalize the rows of a Gaussian $V \times r$ matrix, subtract their mean, and rescale the resulting matrix Y so that its largest row norm is $\sqrt{3}$; set $Q = \mathrm{softmax}(YY^\top)$. In exact arithmetic, these logits are centered, have rank at most r, and have entries bounded in absolute value by three. Ground-truth seed zero is used throughout, with feature-generator seed $10000 + 100V$ and tied-split permutation seed $73000 + V$. The ground truth, split, and sampling stream are shared across the three initializations of each configuration. Tied

and untied ground truths follow different constructions, so their panels are not a matched-teacher comparison.

Training. We initialize all entries independently as $\mathcal{N}(0, 1/d)$ and train every factor jointly. Each update draws 512 fresh pairs by sampling a current state uniformly from the observed set and then its next state from Q. Bounded-logit rejection sampling avoids storing the full transition table; the learner still uses the full V-class softmax for each input. We minimize cross-entropy with ordinary SGD and a squared-factor-norm penalty with coefficient $\lambda/2$, where $\lambda = 1/(120rV)$ for untied factors and $\lambda = 0.4/(rV)$ for tied embeddings. There is no added Gaussian perturbation, momentum, projection, or clipping. The first 8192 updates use a cosine learning-rate schedule from 64 to 0.25. We then continue the saved trajectories to 131072 updates with constant learning rate 0.25 and unchanged decay. Longer training also draws more fresh transitions; it does not repeat passes over a fixed dataset.

Measurements and numerical settings. For a state set $\mathcal{A}$, we report relative prediction gain $g_{\mathcal{A}} = (\log V - \mathcal{R}_{\mathcal{A}})/(\log V - \mathcal{H}_{\mathcal{A}})$, where $\mathcal{R}_{\mathcal{A}}$ is population cross-entropy and $\mathcal{H}_{\mathcal{A}}$ is the ground-truth conditional entropy, both averaged uniformly over $\mathcal{A}$. Weight decay is excluded from this metric. The uniform predictor has gain zero and the ground-truth predictor has gain one. Every plotted evaluation covers the complete training or unseen set, not the smaller state subsets used for intermediate monitoring. Continuation uses FP32 parameters, BF16 training autocast, and TF32 where enabled. Sampling and risk evaluation use FP32, with TF32 disabled for evaluation. The original prefix used a different numerical recipe, including FP64 sampling and risk accumulation. The plots therefore start at 16384 updates and also show 32768, 65536, and 131072 updates, all under the same continuation setting. Lines join checkpoint medians across the three seeds; shading spans their minimum and maximum. There is no smoothing or extrapolation. Width colors are fixed throughout, negative gains are retained, and the dashed line marks $7/8$. Vertical scales are shared within a row but may differ between rows.

Observed behavior. In each of the three metrics—untied training, tied training, and tied unseen prediction—all 39 configuration medians improve between 16384 and 131072 updates. Some configurations already have high gain early in this interval, while others continue improving below $7/8$. For example, at $V = 4096, r = 32, d = 256$, tied training gain rises from 0.711 to 0.932 and unseen gain from 0.340 to 0.752 (Figure 6). Unseen prediction can therefore lag behind training even when both improve. Width does not uniformly improve gain at a fixed training budget. These are observations for sampled-transition SGD, not a verification of the perturbed population-gradient theorem; low final gain alone does not show that a trajectory has converged to a low-gain local minimum.

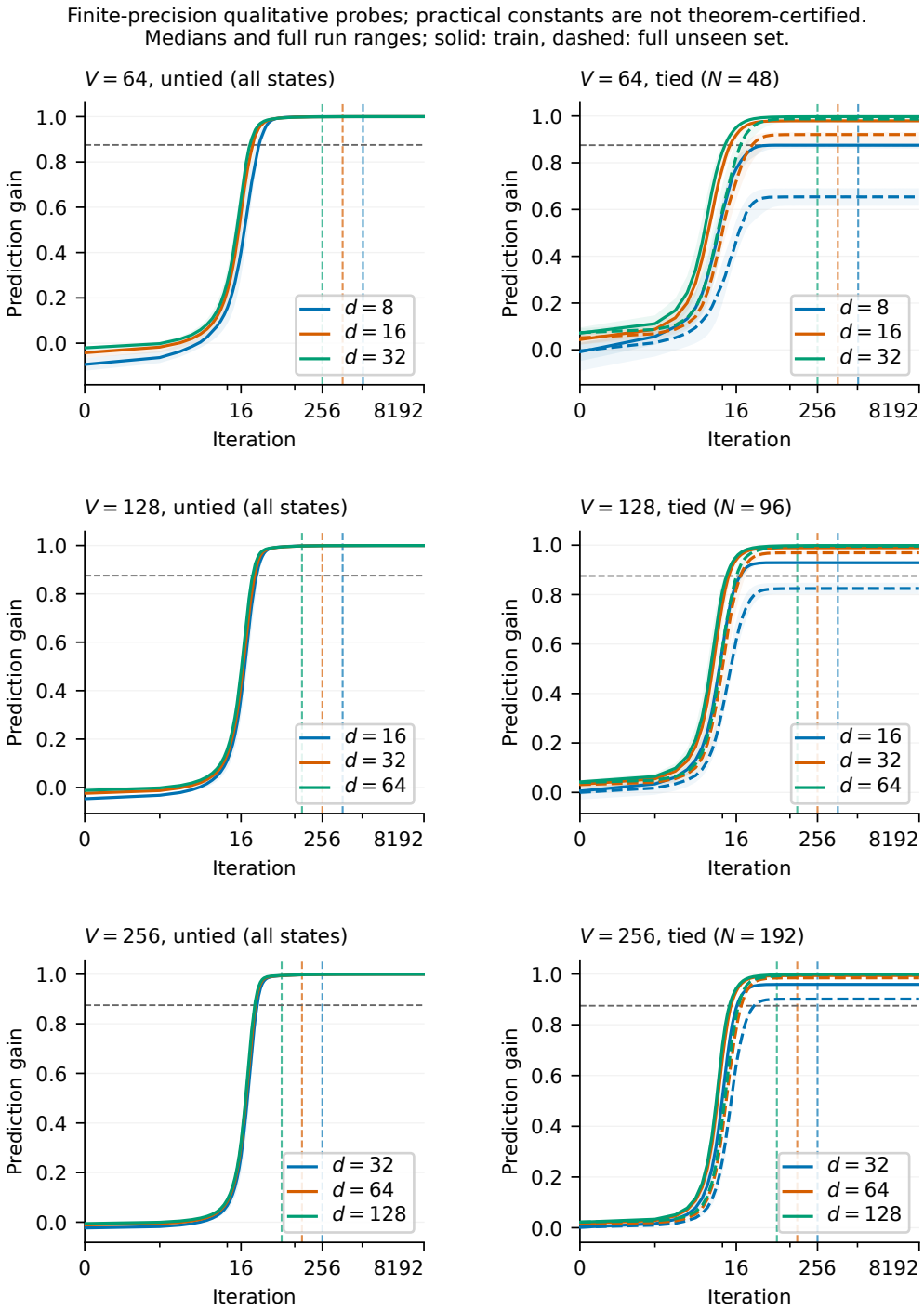

Figure 3: All main-grid trajectories, shown as medians and full ranges. Each panel contains all three widths. Untied curves use all input states; tied solid/dashed curves are training/full-unseen gain. The fixed perturbation law continues throughout every trajectory. A display subsample is used for drawing only; all reported minima and counts use every update.

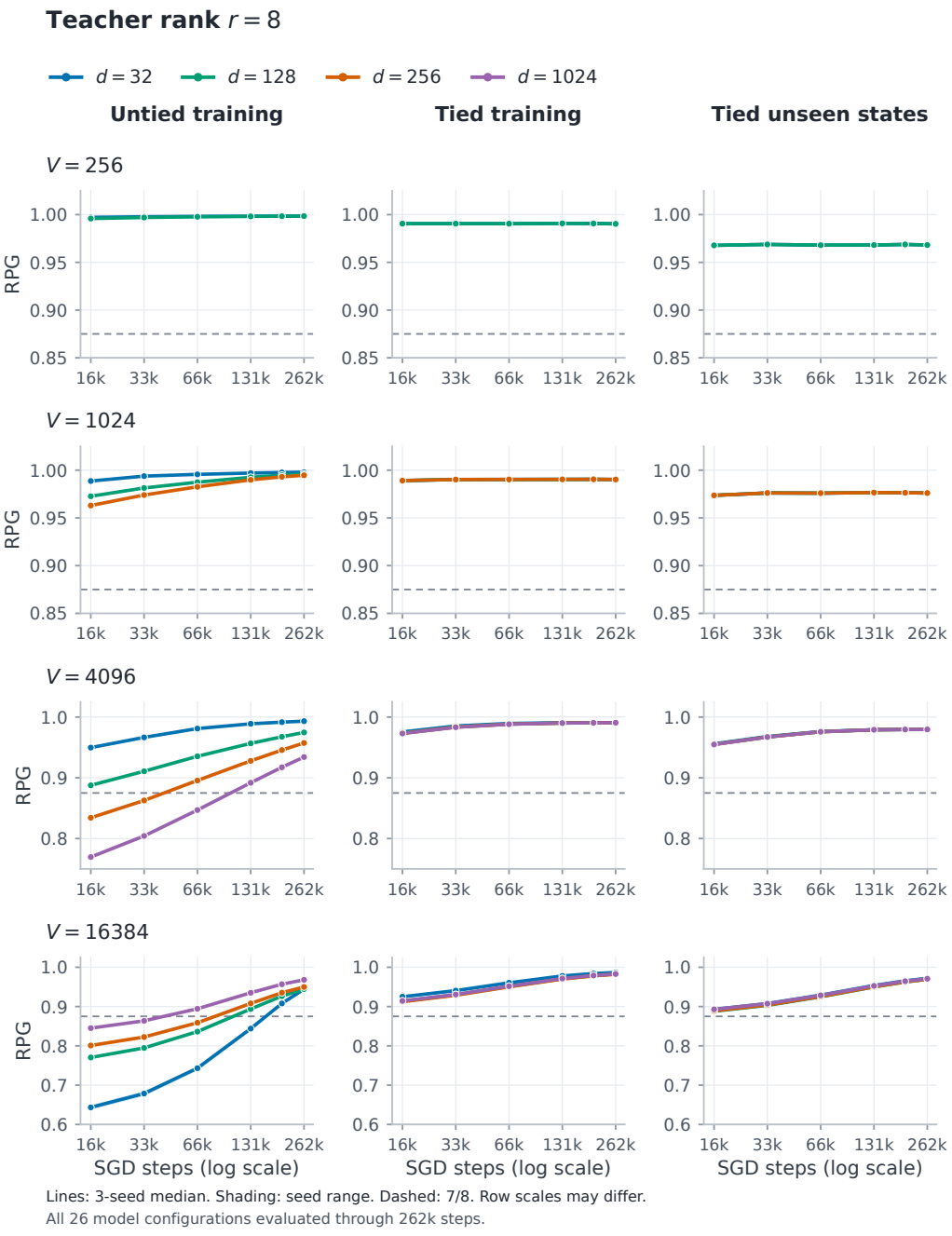

Figure 4: Sampled-transition SGD with $r = 8$. Rows correspond to $V = 256, 1024, 4096, 16384$, from top to bottom; columns show untied training, tied training, and tied unseen gain, from left to right. Every valid width and all three initialization seeds are included. Lines show seed medians and shading shows their full ranges. The dashed line marks $7/8$; row scales may differ.

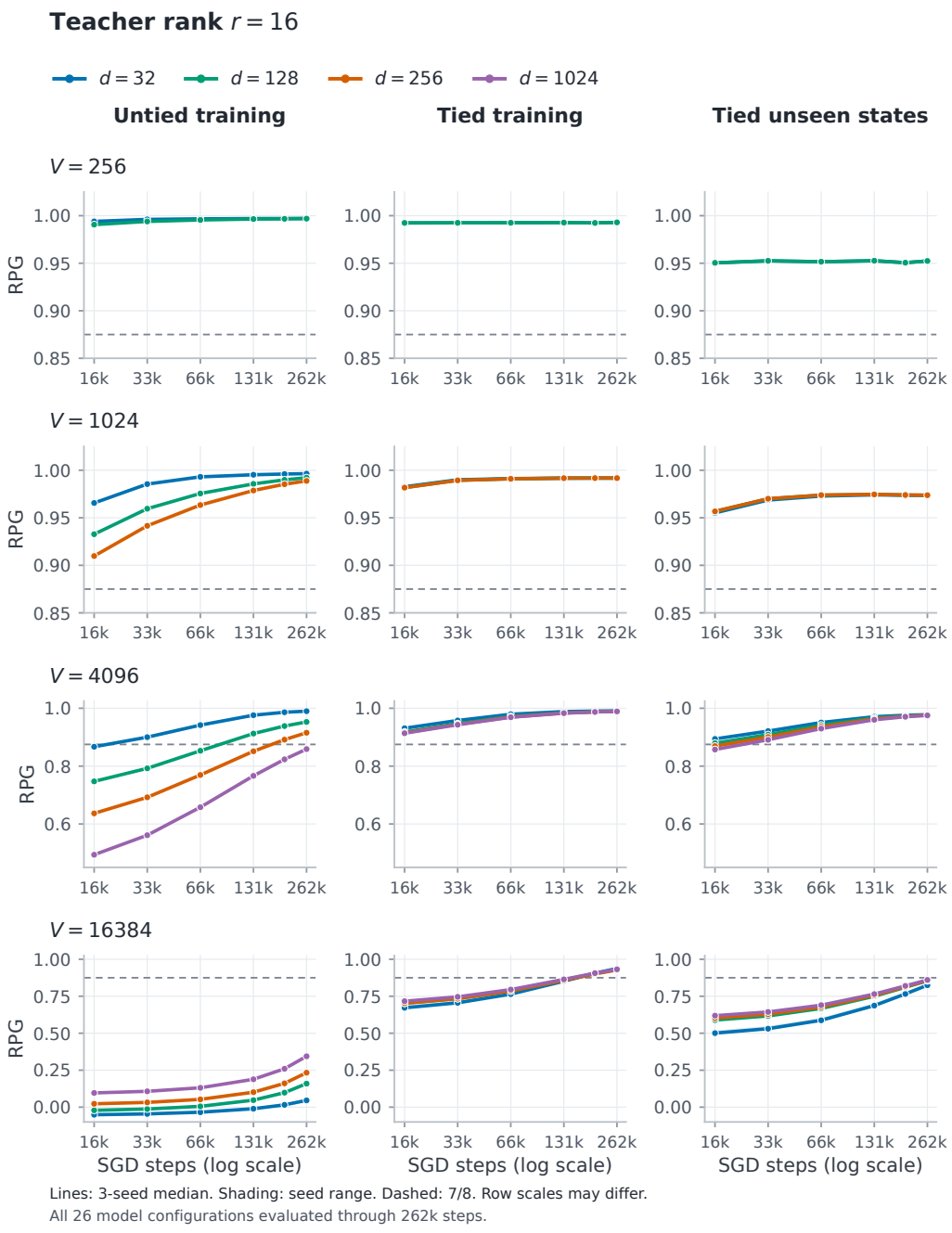

Figure 5: Sampled-transition SGD with $r = 16$, at $V = 256, 1024, 4096, 16384$ from top to bottom. Every valid width and all three initialization seeds are included. Columns and summary statistics follow Figure 4.

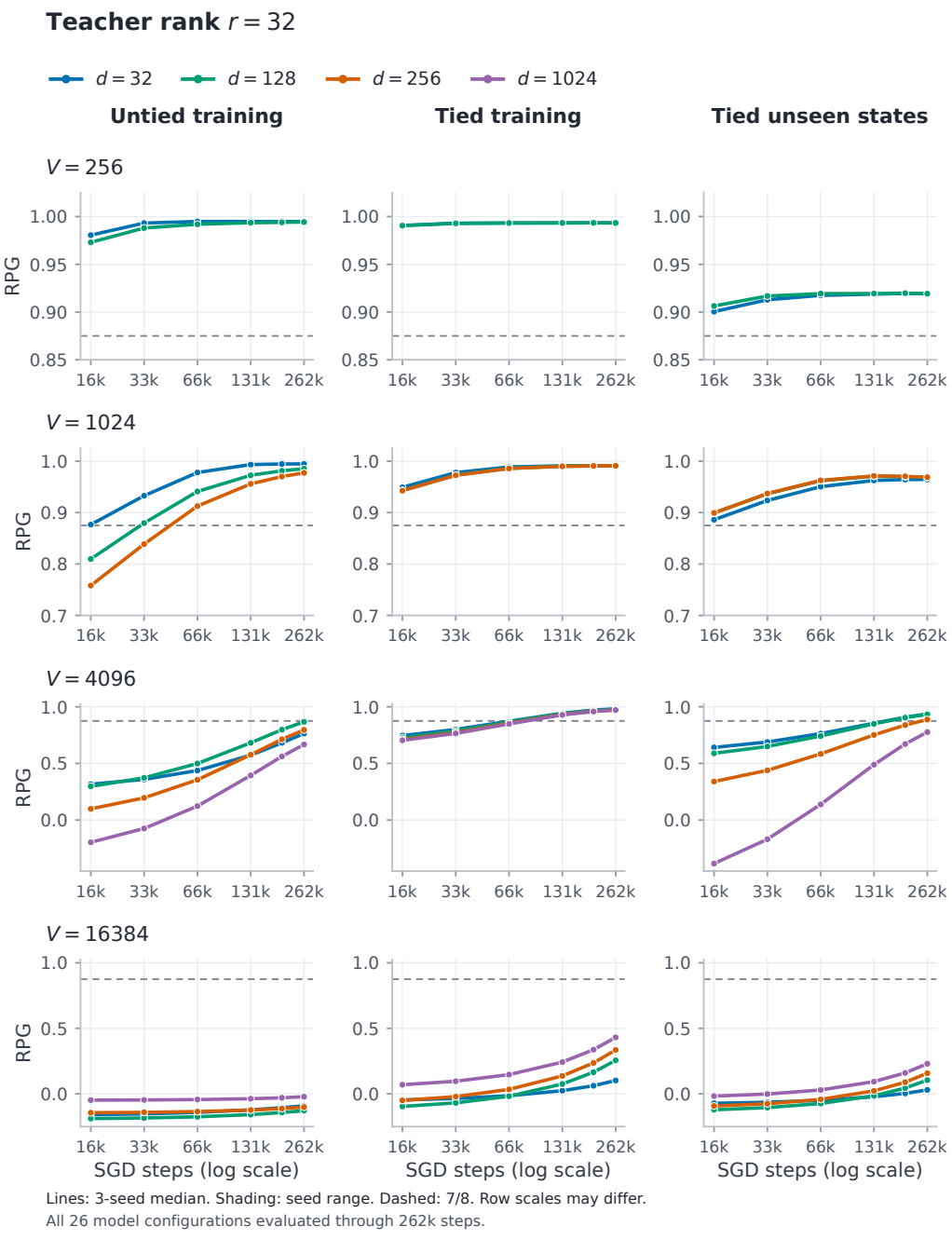

Figure 6: Sampled-transition SGD with $r = 32$, at $V = 256, 1024, 4096, 16384$ from top to bottom. Every valid width and all three initialization seeds are included. Negative gains are retained. Columns and summary statistics follow Figure 4.