TRAINING DYNAMICS OF LARGE STATE-TRANSITION MODELS

Yee Man Choi
University of Waterloo

Fengchi Wang
University of Waterloo

Ruiquan Huang
University of Kentucky

ABSTRACT

Next-token prediction relies on learned token embeddings whose dimension d is much smaller than the vocabulary size V. Existing analyses of training dynamics either fix the embeddings with $d \geq V$ or train them only in restricted settings. It remains unclear whether jointly training the embeddings by gradient-based methods from random initialization leads to accurate prediction when $d < V$. In this paper, we study this question in the simplest setting where it is already nontrivial, namely next-token prediction with a one-token context, which we refer to as a state-transition model. The model consists of an input embedding and an output embedding, either untied or tied, and is trained from Gaussian initialization by perturbed gradient descent with weight decay on the population cross-entropy loss. We consider ground-truth transition laws whose centered log-probability table has low rank and bounded entries, and measure the prediction quality by the relative prediction gain (RPG). It is zero for the uniform predictor and one for the optimal predictor. For the untied model, we show that when $d = \Omega(\sqrt{V})$, with probability at least $1 - e^{-cd}$ for some constant c, the RPG is above $7/8$ at every iterate in the final $1 - 1/d$ fraction of training iterations. For the tied model trained on a subset of states, we prove that the RPG on the observed states is $1 - O(V/d^2)$, and under a coverage condition on the ground truth and a bound on the learned embeddings, we further prove a prediction guarantee on the unseen states. Our analysis is based on the bounded trace of the Hessian of the softmax cross-entropy loss, which implies that the approximate second-order stationary points reached by training have high prediction gain. Numerical experiments with practical parameter choices illustrate the persistent training gain and the transfer of the tied model.

1 INTRODUCTION

Next-token prediction (NTP) is the training objective of modern language models. A basic ingredient of such models is the learned token embedding (Bengio et al., 2000): each of the V tokens in the vocabulary has an input embedding and an output embedding in $\mathbb{R}^d$, where d is much smaller than V and the two embeddings are often tied (Press & Wolf, 2017; Inan et al., 2017). Recent theory has made substantial progress on the training dynamics of transformers for NTP and related sequence tasks, with a focus on how attention layers are learned. To make the analysis tractable, these works take the token embeddings as given and linearly independent, which requires $d \geq V$ (Li et al., 2024; Huang et al., 2024; 2025; Varre et al., 2025; Yang et al., 2024), or train them only in restricted settings, such as a binary vocabulary (Makkuva et al., 2024) or a single gradient step (Wu et al., 2026), or characterize the global minimizers under $d \geq V$ without analyzing the training (Zhao et al., 2024). As a result, it remains open whether training the embeddings by gradient-based methods from random initialization leads to accurate prediction when $d < V$.

In this paper, we study this question in a nontrivial simplified setting, namely NTP with a one-token context. In this case, the model has no attention, and consists of an input embedding $\phi(s) \in \mathbb{R}^d$ of the current token s and an output embedding $\psi(s') \in \mathbb{R}^d$ of the next token s'. The predicted distribution is $\widehat{P}(s' \mid s) \propto \exp(\phi(s)^\top \psi(s'))$. In other words, the model learns a state-transition law $P(s' \mid s)$ over V states with $2Vd$ parameters, or Vd parameters when the two embeddings are tied. We thus refer to it as a **state-transition model**. Although the model is simple, its training is far from trivial: the loss is nonconvex because the two embeddings are jointly trained with $d < V$, and random

initialization provides no alignment with the ground truth. Moreover, since embeddings are part of every language model, these difficulties are also present in models with attention. Resolving them is a necessary step towards analyzing such models. This motivates the following question:

Can low-dimensional embeddings, jointly trained from random initialization, accurately predict the next state (token)?

To answer this question, we must specify what successful training means. Since the embeddings can be transformed without changing the predictions, they are not identifiable, and we measure the success of training by the prediction quality of the trained model. A stationary point may not give good predictions, because even a binary Markov-transformer model can converge to local minima depending on the data and initialization (Makkuva et al., 2024). Conversely, a model can be useful without attaining the optimal loss. We therefore measure prediction quality by relative prediction gain (RPG; Figure 1): 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. RPG is defined for any set of current states, so we can also study the prediction on unseen 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.

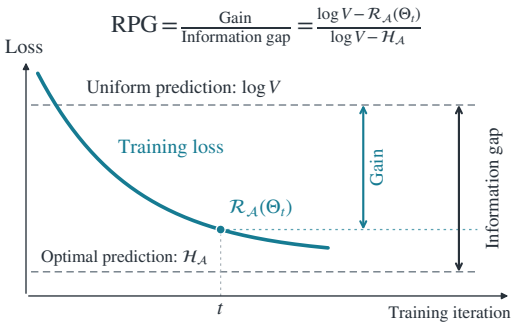


Figure 1: Illustration of relative prediction gain (RPG).

Our contributions are as follows.

- **Untied training: high gain with sublinear embedding dimension.** For every ground truth in our full-support, fixed-rank class, we jointly train both embeddings from Gaussian initialization using cross-entropy, weight decay, and fixed-step perturbed gradient descent. When the embedding dimension satisfies $d \in [C\sqrt{V}, V)$, the prediction gain exceeds $7/8$ with probability at least $1 - e^{-cd}$ at *every* iterate in $[T/d, T]$, for a polynomial number of iterations T fixed in advance. Thus, the guarantee holds for the whole final window of the training trajectory.
- **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 a comparable embedding dimension. Tying also updates unseen-state embeddings through their role as output labels. With a larger embedding dimension and a different weight decay, a coverage condition, 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 iterate in the final window whose learned embeddings have uniformly bounded norms.
- **Proof technique.** The softmax cross-entropy has limited total curvature across logit directions. By rotating the low-rank ground truth 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. For squared-loss matrix factorization, vanilla gradient descent converges globally from a sufficiently small random initialization when

the factor width equals the target rank (Ye & Du, 2021). 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. Related recommender-system work establishes local convergence of projected factor updates from a warm start (Sankagiri et al., 2025), or statistical guarantees for convex multinomial-logit estimators (Negahban et al., 2018). For our loss, we use the trace of its Hessian to guarantee *prediction gain* at approximate stationary points. 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 the cross-entropy loss 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 prediction. Our work is motivated by the theory of next-token prediction, which studies attention-based models (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). Many of these works include attention and longer contexts, but fix the token embeddings or train them only in restricted settings (Wu et al., 2026; Zhao et al., 2024). In contrast, we jointly train the input and output embeddings with $d < V$ from Gaussian initialization, and prove that high prediction gain is maintained throughout a final window of 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 that a related ingredient is analyzed in a different setting, i.e., weight tying and binary Markov transformers analyze related factorized losses, and noisy matrix factorization perturbs the factors before the gradient evaluation, while we add noise to the gradient. 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 $[V] = \{1, \ldots, V\}$ 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 $E, U \in \mathbb{R}^{V\times d}$, where the i-th rows of E and U are the input embedding $\phi(i)$ and the output embedding $\psi(i)$ of state i, respectively. 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. Throughout, the embedding dimension satisfies $d < V$. We use calligraphic letters for sets, except for standard spaces and $[V]$, and uppercase letters for matrices.

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

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

Note that the risk $\mathcal{R}_{\mathcal{A}}$ assigns uniform weights to the current states in $\mathcal{A}$. Although we call it the population risk, Q is only required to be a fixed transition table that satisfies our assumptions. For instance, Q can be the empirical transition table of a dataset, in which case $\mathcal{R}_{\mathcal{A}}$ is the full-batch training loss with uniform weights on the current states.

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

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

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

Ground-truth class. Adding the same constant to every logit in a row leaves its softmax unchanged. For any real matrix Z with V columns, define its row-centering by $\mathcal{L}(Z) = Z\left(I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top\right)$. Thus the ground truth's centered logits are $\mathcal{L}(\log Q)$, where the logarithm is entrywise. We consider the class

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

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 impose the rank condition on the centered logits, because centering does not change the model or the risk, and the rank determines whether the model can represent the ground truth. In particular, when $r \le d$, the ground truth is representable by the untied model. The bound on the logits keeps every transition probability positive. The last condition, which lower bounds the information gap $\log V - \mathcal{H}_{[V]}$, excludes a trivial uniform prediction problem.

Training algorithm. 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.

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

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 weight decay λ, step size h, perturbation scale σ, and number of iterations T are deterministic choices made before initialization. The step size and the perturbation 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. As discussed in section 1, the difficulty comes from the combination of $d < V$*, cross-entropy loss, Gaussian initialization, and simultaneous gradient-based training of all embeddings*. Specifically, (i) the low-dimensional embeddings make the optimization nonconvex; (ii) the Hessian of the cross-entropy loss depends on the learned probabilities and can become nearly singular; and (iii) random initialization provides no alignment with the ground truth and no 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

apply to our setting. Moreover, the generic saddle-escaping analysis (Jin et al., 2019) only guarantees that the algorithm visits approximate second-order stationary points, which does not imply accurate prediction by itself. To obtain a prediction guarantee, we need to (a) relate stationarity to low prediction loss in the region that is actually reached by training, and (b) control the iterates between consecutive visits to such stationary points.

4 Training Gain of Untied Model

In this section, we study the untied model trained on all current states, i.e., $\mathcal{A} = [V]$. We are interested in whether an embedding dimension much smaller than V can support useful prediction over many consecutive training steps, despite the nonconvex objective and perturbations at every update. The theorem below gives a finite-time answer: with high probability, the training achieves at least 7/8 RPG within sufficiently long intervals. Roughly speaking, the training loss is "close" to the optimal loss, and therefore the prediction is accurate.

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 the weight decay $\lambda = \frac{\Delta_0}{8B\sqrt{r\,V}}$, there exist a step size h, a perturbation scale σ, and a number of iterations T such that*

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

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

Theorem 4.1 states that, with high probability, every iterate after the first $1/d$ fraction of training, including the final one, has RPG above $7/8$. Moreover, an embedding dimension of order $\sqrt{V}$, which is much smaller than V, is sufficient.

4.1 Proof Overview

Definition 4.2 (Approximate SOSP (Jin et al., 2019))**.** 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

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

Throughout, we take $f = \mathcal{F}_{\lambda,[V]}$, i.e., the training objective.

At a high level, the goal is to prove that the dynamics of Θ will enter into a low-loss region. The difficulty is that Θ may fall into an approximate SOSP. Because otherwise, either the gradient $\nabla\mathcal{F}(\Theta)$ is sufficiently large to make the loss decrease, or the Hessian $\nabla^2\mathcal{F}(\Theta)$ has a sufficiently negative eigenvalue that allows Θ to escape. To handle this difficulty, the key idea is to show that approximate SOSPs "of interest" have high prediction gain. To this end, we construct three main steps. First, we establish that the training trajectory of Θ remains in a "good" region, namely localize the training trajectory. Second, we show that in this region, every approximate SOSP must have high prediction gain. Finally, the conclusion follows from combining repeated approximate-SOSP visits and induction on the error control. Appendix B gives the parameter choices and complete proof.

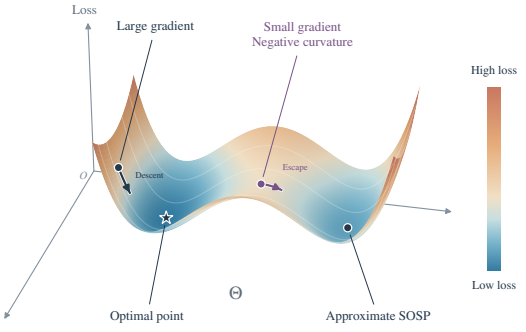


Figure 2: Illustration of an approximate SOSP. Our Step 2 shows that all approximate SOSPs "of interest" must have low loss.

Step 1: Localize the training trajectory. Positive weight decay creates a region where the squared parameter norm is at most $O(V)$. Leaving this region requires a substantial increase in the objective

loss. More importantly, concentration inequalities guarantee that, with high probability, Gaussian initialization lies inside this region (lemma B.1), and the perturbed gradient descent ensures that the objective cannot rise by more than a fixed small amount during training. Therefore, with high probability, the training trajectory remains in the region where the parameter norm is bounded.

Step 2: Convert approximate stationarity into gain. In this step, we show that every ϵ-approximate SOSP Θ in the localized region satisfies $\mathcal{F}(\Theta) < \mathcal{H}_{[V]} + \Delta_0/8 - 2b$ for some constant $b > 0$, which implies $g_{[V]}(\Theta) > 7/8$. To this end, we first construct a comparator $\Theta_\star$, whose logits are the slightly shrunken ground-truth logits, such that $\|\Theta_\star\|_{\mathrm{F}}^2 = O(V)$ and $\mathcal{F}(\Theta_\star) \le \mathcal{H}_{[V]} + \Delta_0/8 - 4b$ (lemma B.2). Suppose, for contradiction, that $\mathcal{F}(\Theta) \ge \mathcal{H}_{[V]} + \Delta_0/8 - 2b$, so that $\mathcal{F}(\Theta) - \mathcal{F}(\Theta_\star) \ge 2b$. The key step lies in analyzing the Hessian $\nabla^2\mathcal{F}(\Theta)$ along a direction $\dot\Theta$ obtained by rotating $\Theta_\star$ among the d embedding directions, which satisfies $\|\dot\Theta\|_{\mathrm{F}}^2 = O(V)$ and (lemmas B.3 and B.4)

$$\nabla^2\mathcal{F}(\Theta)[\dot\Theta,\dot\Theta] \le \underbrace{O\Big(\frac{V}{d^2} + \frac{1}{d}\Big)}_{\le b \text{ for } d \ge C\sqrt{V}} + \underbrace{\langle \nabla\mathcal{F}(\Theta), \Theta\rangle}_{\le \epsilon\|\Theta\|_{\mathrm{F}} \le b} - 2\underbrace{\big(\mathcal{F}(\Theta) - \mathcal{F}(\Theta_\star)\big)}_{\ge 2b} \le -2b.$$

Here, the first term bounds the curvature of the softmax, using that the Hessian of the loss of each row with respect to its logits has trace at most one, and the last two terms follow from the convexity of the loss in the logits and $\|\Theta\|_{\mathrm{F}} = O(\sqrt{V})$. Since $\|\dot\Theta\|_{\mathrm{F}}^2 = O(V)$, we obtain $\lambda_{\min}(\nabla^2\mathcal{F}(\Theta)) \le \nabla^2\mathcal{F}(\Theta)[\dot\Theta,\dot\Theta]/\|\dot\Theta\|_{\mathrm{F}}^2 \le -2b/\|\dot\Theta\|_{\mathrm{F}}^2 \le -c/V$ for some constant $c > 0$. However, an ϵ-approximate SOSP satisfies $\lambda_{\min}(\nabla^2\mathcal{F}(\Theta)) \ge -\sqrt{\rho\epsilon}$, where $\sqrt{\rho\epsilon} \le c/(2V)$ since $\rho = O(\sqrt{V})$ and ϵ is a sufficiently small constant times $V^{-5/2}$. This is a contradiction.

Step 3: Refresh the gain throughout the final window. Finally, we extend the guarantee in Step 2 from approximate SOSPs to all iterates in the final window. We partition the T iterations into d blocks, which are only used in the analysis. By the result of Jin et al. (2019) on perturbed gradient descent, with high probability, every block contains an ϵ-approximate SOSP, and the objective increases by at most b within each block, i.e., $\mathcal{F}(\Theta_v) - \mathcal{F}(\Theta_u) \le b$ for any $u < v$ in the same block (lemma D.6). Combining this with Step 2 and using induction over the blocks, we obtain $\mathcal{F}(\Theta_t) < \mathcal{H}_{[V]} + \Delta_0/8$, and hence $g_{[V]}(\Theta_t) > 7/8$, for every integer $t \in [T/d, T]$.

5 Tied Model Training and Conditional Transfer

In this section, we study the training dynamics when the input and output embeddings are tied. 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. 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 $B_Y \ge 1$, 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 $r_Y \le r$ and $\max_i \|y_i\|_2^2 \le B_Y$, 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.

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 $\max\{2r, C\sqrt{V}\} \le d < N$, setting the weight decay $\lambda = C_\lambda/d^2$, there exist a step size h, a perturbation scale σ, and a number of iterations T divisible by d such that, with probability at least $1 - e^{-cd}$, $g_{\mathcal{S}}(X_t) \ge 1 - \frac{CV}{d^2\delta_{\mathrm{tr}}}$, for every integer $t \in [T/d, T]$. The quantities h^{-1}, σ^{-1}, and T are bounded by polynomials in V.*

Theorem 5.2 indicates that the gap from gain one scales as V/d^2, so the guarantee strengthens as the embedding dimension grows beyond $\Omega(\sqrt{V})$.

5.1.1 PROOF OVERVIEW

The proof follows the similar structure as the untied argument in section 4, i.e., we localize the trajectory, obtain a loss guarantee at approximate SOSPs, and then propagate this guarantee throughout the final window. The key difference is that the weight decay absorbs a curvature term arising from the tied embeddings, which depends on the norm $\|X\|_{\mathrm{F}}^2$ of the embeddings, referred to as their *energy*.

Step 1: Localize the training trajectory. 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 analysis of perturbed gradient descent. The control of the objective increase keeps the modified and original training trajectories inside the ball and identical under the same noise (lemma C.1).

Step 2: Convert approximate stationarity into gain. We show that every sufficiently accurate approximate SOSP X in the ball satisfies $\|X\|_{\mathrm{F}}^2 = O(V)$ and $\mathcal{F}_{\lambda,\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} = O(V/d^2)$. Similarly to Step 2 in section 4.1, the key step lies in analyzing the Hessian $\nabla^2 \mathcal{F}_{\lambda,\mathcal{S}}(X)$ along a direction that represents the ground-truth Gram matrix $YY^\top$ and has small softmax curvature. Together with the convexity of the loss in the Gram matrix, this gives $\mathcal{F}_{\lambda,\mathcal{S}}(X) \lesssim \mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 + C\|X\|_{\mathrm{F}}^2/d^2$ up to small stationarity errors (lemma C.2), where the last term depends on the energy of X. Since $\mathcal{F}_{\lambda,\mathcal{S}}(X) \geq \mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|X\|_{\mathrm{F}}^2$, choosing $\lambda = C_\lambda/d^2$ with a sufficiently large C_λ lets the weight decay penalty absorb this term, which yields $\|X\|_{\mathrm{F}}^2 = O(V)$, since $\|Y\|_{\mathrm{F}}^2 = O(V)$, and then the claim (lemma C.3).

Step 3: Refresh the gain throughout the final window. Finally, as in Step 3 of section 4.1, we partition the iterations into d blocks. With high probability, every block contains an accurate approximate SOSP, and the objective increases by at most $V/(4d^2)$ within each block. Combining this with Step 2 and using induction over the blocks, we obtain $\mathcal{R}_{\mathcal{S}}(X_t) - \mathcal{H}_{\mathcal{S}} \leq \mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \mathcal{H}_{\mathcal{S}} = O(V/d^2)$ for every integer $t \in [T/d, T]$. 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 weight decay chosen for this transfer argument and requires a larger embedding dimension than tied training alone.

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 $\lambda = C_\lambda\sqrt{\frac{V}{Nd^3}}$. Then, there exist a step size h, a perturbation scale σ, and a number of iterations T, with h^{-1}, σ^{-1}, and T bounded by polynomials in V, such that, with probability at least $1 - e^{-cd}$, 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 $\max_{j\in[V]} \|x_j(t)\|_2^2 \le R_0$ *for a fixed* R_0*, the complete unseen set satisfies*

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

Theorem 5.4 bounds the gain on the observed states throughout the final window and, at any of those iterates with uniformly bounded learned embeddings, also the gain on the complete unseen set. The bound on the learned embeddings is an additional condition that our optimization analysis does not establish, so the conclusion on unseen states is conditional.

5.2.1 Proof Overview

The proof consists of three main steps. The key idea is that, since the Gram matrix $XX^\top$ is positive semidefinite, its rows indexed by the observed states together with its trace control the entire matrix. First, we compare the objective at every iterate in the final window with the objective at the ground truth. Second, we use the coverage condition to turn this comparison into bounds on the training loss and on the trace of the Gram matrix. Finally, under the bound on the learned embeddings, these two bounds control the entire Gram matrix, which yields the gain bound on the unseen states. Appendix C.2 gives the parameter choices and complete proof.

Step 1: Compare the objective with the ground truth in the final window. We show that, with high probability, every iterate X_t in the final window satisfies $\mathcal{F}_{\lambda,\mathcal{S}}(X_t) \le \mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 + C(\lambda V)^2$, where the first two terms are the objective at the ground truth, and $(\lambda V)^2$ is of order $V^3/(Nd^3)$ under the weight decay in theorem 5.4. This follows from the same three steps as in section 5.1.1. Specifically, by Step 2 there, every accurate approximate SOSP satisfies this comparison up to the energy-dependent term $C\|X\|_{\mathrm{F}}^2/d^2$ (lemma C.2), which is absorbed by the weight decay. Then, by Steps 1 and 3 there, the comparison extends to every iterate in the final window (lemma C.4).

Step 2: Bound the training loss and the trace of the Gram matrix. Let $K = XX^\top$ and $K^\star = YY^\top$ be the learned and ground-truth Gram matrices. By Step 1, every iterate X in the final window satisfies $\{\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}\} + \frac{\lambda}{2}\{\operatorname{tr} K - \operatorname{tr} K^\star\} \le C(\lambda V)^2$, which does not immediately bound the excess loss $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}$, because $\operatorname{tr} K - \operatorname{tr} K^\star$ can be negative. Using the coverage condition, we establish the following trace certificate (lemmas C.6 and C.7):

$$\operatorname{tr} K - \operatorname{tr} K^\star \ge -CV\Big(\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} + \sqrt{\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}}\Big).$$

Substituting it into the above inequality, and using that λV is small, we obtain $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} = O((\lambda V)^2)$ and $\operatorname{tr} K - \operatorname{tr} K^\star \le O(\lambda V^2)$.

Step 3: Transfer the bounds to unseen states. Finally, we transfer the bounds in Step 2 to the unseen states at any iterate X in the final window with $\max_j \|x_j\|_2^2 \le R_0$. In this case, all logits are bounded, so the softmax probabilities are of order $1/V$, and the excess loss of each row is equivalent, up to constants, to $1/V$ times the squared error of its centered logits. Hence, the training loss controls the error of K on the observed rows. The remaining error consists of a part determined by the observed rows and a positive semidefinite residual, whose norm is at most its trace. This gives (lemma C.8)

$$\|K - K^\star\|_{\mathrm{F}} \lesssim \sqrt{NV\big(\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}\big)} + \max\{\operatorname{tr} K - \operatorname{tr} K^\star, 0\} = O(\lambda V^2).$$

For the unseen rows, the same equivalence gives $\mathcal{R}_{\mathcal{U}}(X) - \mathcal{H}_{\mathcal{U}} \lesssim \|K - K^\star\|_{\mathrm{F}}^2/((V-N)V)$, which is $O((\lambda V)^2 V/(V-N))$ (lemma C.9). Substituting the choice of λ and dividing by the training and test information gaps gives the two gain bounds in theorem 5.4.

6 Numerical Experiments

In this section, we present two sets of experiments. The first set uses population gradients, and examines the persistence of the training gain, the transfer of the tied model, and the bounded-embedding condition in theorem 5.4. The second set uses SGD with sampled transitions at a larger scale, and tests whether these behaviors persist under stochastic training and across larger state spaces, ground-truth ranks, and embedding dimensions. Both studies use practical parameter values.

6.1 POPULATION(FULL BATCH)-GRADIENT TRAINING

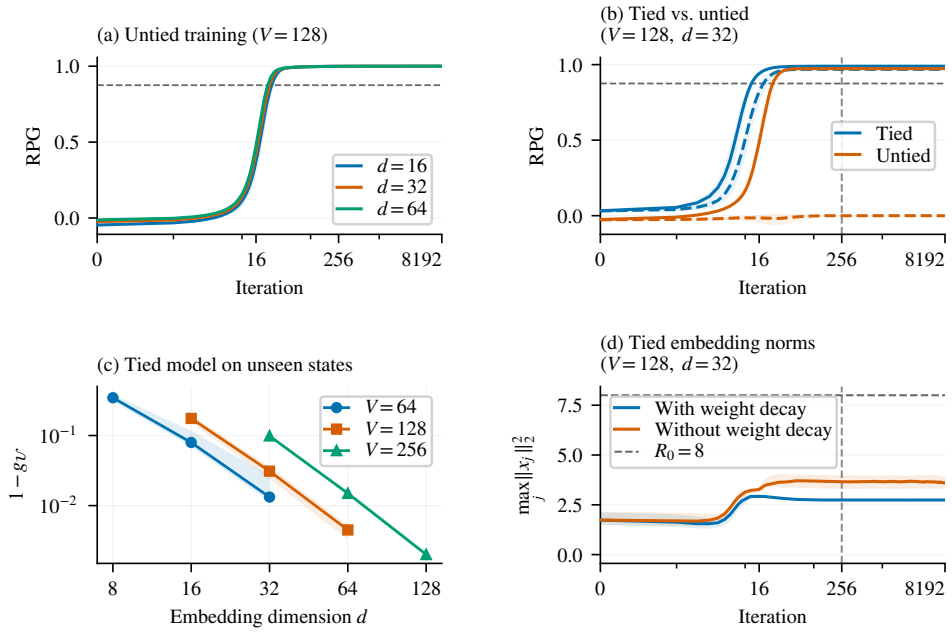


Figure 3: Training and transfer in the population-gradient experiments. Curves show medians and shaded full ranges over three ground truths and five initializations each. (a) Untied training gain at $V = 128$. (b) Tied and untied models on the same observed/unseen split at $V = 128, d = 32$ (solid: training; dashed: test). (c) Final tied test deficit ($1-$ gain) as the embedding dimension and the corresponding weight decay change together. (d) Maximum tied embedding-row norm squared. The vertical lines in (b,d) mark T/d. 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 current states; tied training uses $N = 3V/4$ observed 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.

Main results. We examine three behaviors predicted by our theorems. (i) *The training gain remains high throughout the final window* $[T/d, T]$ (Theorems 4.1 and 5.2). The gain stays above $7/8$ in all 135 runs of the untied model and in 120 of the 135 runs of the tied model. The exceptions are the runs with the smallest embedding dimension ($V = 64$, $d = 8$). - (ii) *The tied model transfers to unseen states* (Theorem 5.4). When both models are trained on the same observed states, the final gain of the tied model on the unseen states is above 0.96, while that of the untied model remains near zero. - (iii) *The condition on the learned embeddings in Theorem 5.4 holds in the experiments.* The squared norms of the learned embeddings stay below the bound $R_0 = 8$, which is fixed in advance, in all runs of the tied model.

Effects of perturbation and weight decay. The paired control runs 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 3(c), the embedding dimension and the weight decay vary together, so the curves reflect their joint effect.

6.2 Scaling Optimization with Sampled SGD

To test whether our qualitative findings extend beyond exact population gradients, we evaluate sampled-transition training over larger state spaces, ground-truth ranks, and embedding dimensions. Figure 4 follows nine (V, d) settings at ground-truth rank $r = 8$ from 16,384 to 262,144 SGD updates. For larger V, 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.At larger ground-truth ranks, some configurations remain below $7/8$ at the final checkpoint, but are still improving. Section F gives the full protocol and the wider configuration grid.

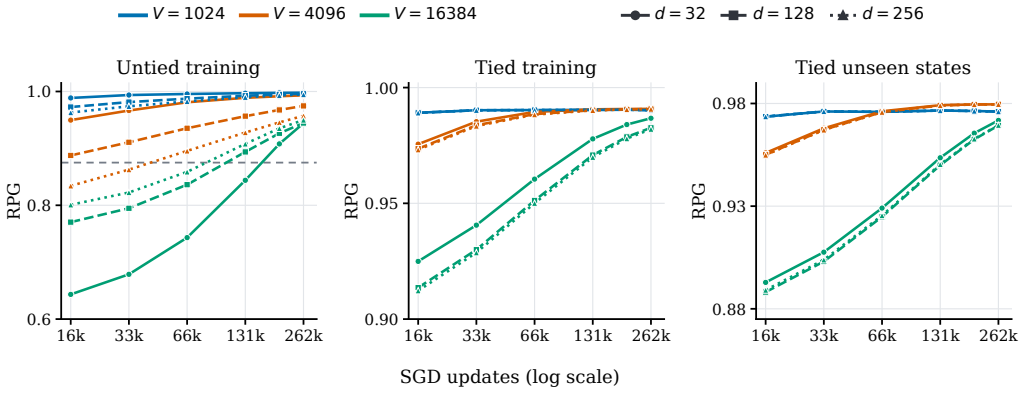


Figure 4: Relative prediction gain during sampled-transition SGD at ground-truth rank $r = 8$. Each panel contains nine curves: $V \in \{1024, 4096, 16384\}$ (colors) and $d \in \{32, 128, 256\}$ (line and marker styles). Lines join medians of three initialization seeds at each checkpoint, without smoothing. 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 state-transition models, i.e., next-token prediction with a one-token context, in which low-dimensional embeddings are jointly trained from random initialization. We first used the bounded trace of the Hessian of the cross-entropy loss to show that accurate approximate second-order stationary points reached by the untied model have high prediction gain. Combining this result with the localization of the trajectory and repeated visits to such points, we proved that perturbed gradient descent with weight decay achieves constant gain in polynomial time and maintains it throughout a final window of training. We then extended this guarantee to tied embeddings, and, under additional conditions on the coverage and the learned embeddings, obtained a prediction guarantee for the complete unseen set. Our experiments illustrate the persistence of the training gain and the transfer benefit of tying. An important future direction is to extend our analysis to models with attention.

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.

Sahand Negahban, Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Learning from comparisons and choices. *Journal of Machine Learning Research*, 19(40):1–95, 2018.

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/.

Suryanarayana Sankagiri, Jalal Etesami, and Matthias Grossglauser. Recommendations with sparse comparison data: Provably fast convergence for nonconvex matrix factorization. *arXiv preprint arXiv:2502.20033*, 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.

Diyuan Wu, Aleksandr Shevchenko, Samet Oymak, and Marco Mondelli. Attention with trained embeddings provably selects important tokens. *Advances in Neural Information Processing Systems*, 38:28956–28997, 2026.

Hongru Yang, Bhavya Kailkhura, Zhangyang Wang, and Yingbin Liang. Training dynamics of transformers to recognize word co-occurrence via gradient flow analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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.

Tian Ye and Simon S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1429–1439, 2021.

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 are organized as follows. Appendix A restates the models and the training algorithm, and introduces the additional notation used in the proofs. Appendix B proves theorem 4.1 for the untied model. Appendix C.1 proves theorem 5.2 for the tied model, and Appendix C.2 proves theorem 5.4 on the transfer to unseen states. Appendix D collects the auxiliary results that are shared by these proofs. Finally, Appendices E and F provide the details of the experiments. Each proof follows the three steps of the corresponding proof overview in the main text, which we summarize below.

Untied training. The proof of theorem 4.1 follows the three steps in section 4.1. In Step 1, we show that the Gaussian initialization lies in a ball of squared radius $O(V)$, and that the objective outside this ball is larger than its initial value by a constant. In Step 2, we show that every sufficiently accurate approximate SOSP in this ball has gain above $7/8$. The key fact is that the Hessian of the cross-entropy loss of each row with respect to its logits has trace at most one. In Step 3, we specify the step size, the perturbation scale, and the number of iterations, and prove by induction over d blocks of iterations that the trajectory stays in the ball and that the gain is above $7/8$ at every iterate in the final window. After the proof, we compare our analysis with a generic bound for factorized optimization.

Tied training. The proof of theorem 5.2 has the same three steps. The main difference is in Step 2, where the comparison with the ground-truth Gram matrix has an error term proportional to $\|X\|_{\mathrm{F}}^2/d^2$, and we use the weight decay to absorb this term.

Conditional transfer. The proof of theorem 5.4 follows the three steps in section 5.2.1. In Step 1, we compare the objective at every iterate in the final window with the objective at the ground truth. In Step 2, we use the coverage condition to turn this comparison into bounds on the training loss and on the trace of the learned Gram matrix. In Step 3, we use these two bounds and the positive semidefiniteness of the Gram matrix to control the entire Gram matrix, and thus the loss on the unseen states. The bound on the learned embeddings is only used in Step 3, which is deterministic.

A SETUP FOR THE PROOFS

In this section, we restate the models and the training algorithm in section 3, and introduce the additional notation that is used in the proofs.

Notation. All logarithms are natural, and the softmax function is applied to a matrix row-wise. For probability vectors $p, q \in \mathbb{R}^V$ with positive entries and a logit vector $z \in \mathbb{R}^V$, we define

$$\mathrm{softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^V e^{z_k}}, \qquad \mathrm{KL}(q\|p) = \sum_{j=1}^V q_j \log \frac{q_j}{p_j}, \qquad \mathrm{CE}(q,z) = \log \sum_{j=1}^V e^{z_j} - q^\top z.$$

Recall that $\mathcal{L}$ is the row-centering operator. In particular, the centered logits of the ground truth are

$$\mathcal{L}(\log Q) = (\log Q)\left(I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top\right), \qquad \mathcal{L}(\log Q)\mathbf{1} = 0.$$

We refer to the squared Frobenius norm of the parameters, e.g., $\|\Theta\|_{\mathrm{F}}^2$, simply as their squared norm, which is also called the energy in the main text.

Models. The untied model has parameters

$$\Theta = (E, U), \qquad E, U \in \mathbb{R}^{V \times d}, \qquad P^{E,U} = \mathrm{softmax}(EU^\top), \qquad d < V.$$

Its centered logits are $\mathcal{L}(EU^\top)$, which have the same row-wise softmax as $EU^\top$. The tied model has parameters $\Theta = X \in \mathbb{R}^{V \times d}$, with logits $XX^\top$, centered logits $\mathcal{L}(XX^\top)$, and $P^X = \mathrm{softmax}(XX^\top)$. Thus, the number of parameters is $D_\Theta = 2Vd$ for the untied model and $D_\Theta = Vd$ for the tied model. In both models, every conditional distribution is over all V states. The untied model is trained on all current states, while the tied model may be trained only on a subset $\mathcal{S} \subset [V]$ of current states, where $N = |\mathcal{S}|$ and $\mathcal{U} = [V] \setminus \mathcal{S}$.

Risk and gain. For any nonempty $\mathcal{A} \subseteq [V]$, the risk, the Bayes risk, and their difference are given by

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

$$\mathcal{R}_\mathcal{A}(\Theta) - \mathcal{H}_\mathcal{A} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathrm{KL}(q_i \| P^\Theta_{i,:}) \geq 0.$$

We refer to $\mathcal{R}_\mathcal{A}(\Theta) - \mathcal{H}_\mathcal{A}$ as the excess loss, and to $\log V - \mathcal{H}_\mathcal{A}$ as the information gap. Whenever $\log V - \mathcal{H}_\mathcal{A} > 0$, the relative prediction gain is

$$g_\mathcal{A}(\Theta) = \frac{\log V - \mathcal{R}_\mathcal{A}(\Theta)}{\log V - \mathcal{H}_\mathcal{A}} = 1 - \frac{\mathcal{R}_\mathcal{A}(\Theta) - \mathcal{H}_\mathcal{A}}{\log V - \mathcal{H}_\mathcal{A}}.$$

By definition, the uniform predictor has gain zero and the ground-truth predictor has gain one. Note that the gain only depends on the prediction loss, and does not include the weight decay penalty.

Ground-truth classes. We restate the classes of ground-truth transition tables in the main text.

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

$$\mathrm{rank}\, \mathcal{L}(\log Q) \leq r, \qquad \|\mathcal{L}(\log Q)\|_\infty \leq B, \qquad \log V - \mathcal{H}_{[V]} \geq \Delta_0,$$

where $\|\cdot\|_\infty$ denotes the largest absolute entry of a matrix.

When $r \leq d$, every ground truth in $\mathcal{Q}_{r,B,\Delta_0}$ can be represented by the untied model. The bound on the centered logits implies that every transition probability is within constant multiples of $1/V$, and the condition on the information gap excludes the case where the uniform predictor is asymptotically optimal.

Definition A.2 (Tied-compatible class)**.** For a fixed constant $B_Y \geq 1$, 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

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

where y_i denotes the i-th row of Y as a column vector.

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 $\kappa > 0$,

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

The coverage condition is only used in the proof of theorem 5.4, and is not needed for theorem 5.2.

Training algorithm. All entries of the initial parameters Θ_0 are independent $\mathcal{N}(0, 1/d)$ random variables. The regularized objective is

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

For the untied model trained on all states, we write $\mathcal{F}_\lambda = \mathcal{F}_{\lambda,[V]}$, and denote its global minimum by $\mathcal{F}^*_\lambda = \min_{E,U} \mathcal{F}_\lambda(E,U)$. All parameters are trained by perturbed gradient descent:

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

The perturbations $\{\xi_t\}$ are independent of the initialization. The weight decay λ, the step size h, the perturbation scale σ, and the number of iterations T are deterministic and chosen before training, and h and σ are fixed throughout training. In the proof of theorem 4.1, we use $\lambda = \Delta_0/(8BV\sqrt{r})$ and $\mathcal{A} = [V]$. In the proofs of theorems 5.2 and 5.4, the weight decay is specified by each theorem, and the quantity λV always refers to the weight decay of the theorem being proved.

Two facts on the untied objective. We end this section with two facts on the untied objective $\mathcal{F}_\lambda$. They are not used in the proofs of the main theorems. Instead, they relate the factorized objective to a nuclear-norm regularized problem over the logit matrix, which is the formulation used in the comparison with existing results in Appendix B.4. The first fact is that the mean of the rows of the output embedding matrix U does not affect the predictions, and is decoupled from the rest of the parameters.

Lemma A.4 (Decoupling of the output mean)**.** *For $U \in \mathbb{R}^{V\times d}$,*

$$\begin{aligned} U &= \left(I_V - \frac{\mathbf{1}\mathbf{1}^\top}{V}\right) U + \frac{\mathbf{1}\mathbf{1}^\top U}{V}, \\ \|U\|_{\mathrm{F}}^2 &= \left\| \left(I_V - \frac{\mathbf{1}\mathbf{1}^\top}{V}\right) U \right\|_{\mathrm{F}}^2 + \frac{\|\mathbf{1}^\top U\|_2^2}{V}. \end{aligned} \tag{2}$$

Consequently,

$$\mathcal{F}_{\lambda,[V]}(E,U) = \mathcal{F}_{\lambda,[V]}\left(E, \left(I_V - \frac{\mathbf{1}\mathbf{1}^\top}{V}\right) U\right) + \frac{\lambda}{2V}\|\mathbf{1}^\top U\|_2^2.$$

The predictions and the dynamics of the centered factors are independent of $\mathbf{1}^\top U$. Under equation (1)*,*

$$\mathbf{1}^\top U_{t+1} = (1 - h\lambda)\mathbf{1}^\top U_t - h\mathbf{1}^\top \xi_t^U, \tag{3}$$

where ξ_t^U is the component of ξ_t that corresponds to U. Moreover, at every critical point with $\lambda > 0$, we have $\mathbf{1}^\top U = 0$.

Proof. The two components of U in equation (2) are orthogonal in the Frobenius inner product, which gives the second identity in equation (2). Moreover, replacing U by $(I_V - \mathbf{1}\mathbf{1}^\top/V)U$ subtracts the same constant from all entries in each row of $EU^\top$, which does not change the row-wise softmax. Combining these two facts gives the decomposition of the objective. Next, since each row of $P^{E,U} - Q$ sums to zero, the gradient of the prediction loss with respect to U satisfies $\mathbf{1}^\top \nabla_U \mathcal{R}_{[V]}(\Theta) = 0$. Thus, multiplying the update of U in equation (1) by $\mathbf{1}^\top$ from the left gives equation (3). The remaining statements follow from the decomposition of the objective and equation (3). □

The second fact is the following reformulation of the global minimum.

Remark A.5 (Bayes optimum and regularized optimum)**.** The Bayes risk is $\mathcal{H}_{[V]}$, and a predictor is Bayes-optimal if and only if $P^{E,U} = Q$, or equivalently, $\mathcal{R}_{[V]}(\Theta) = \mathcal{H}_{[V]}$ and $g_{[V]}(\Theta) = 1$. Thus, if $d \geq \operatorname{rank} \mathcal{L}(\log Q)$, the minimum of the risk over the untied model without weight decay is $\mathcal{H}_{[V]}$. On the other hand, the global minimum of the regularized objective satisfies

$$\begin{aligned} \mathcal{F}^*_\lambda &= \min_{E,U\in\mathbb{R}^{V\times d}} \mathcal{F}_\lambda(E,U) \\ &= \min_{\substack{M\in\mathbb{R}^{V\times V},\ M\mathbf{1}=0 \\ \operatorname{rank} M \leq d}} \left\{ \frac{1}{V}\sum_{i=1}^V \mathrm{CE}(q_i, M_{i,:}) + \lambda\|M\|_* \right\}, \end{aligned}$$

where $\|M\|_*$ is the nuclear norm of M. The second equality follows from lemma A.4 and the identity

$$\min_{EU^\top = M} \frac{\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2}{2} = \|M\|_*.$$

When $\log V - \mathcal{H}_{[V]} > 0$ and $\lambda > 0$, every global minimizer of $\mathcal{F}_\lambda$ has positive excess loss and gain strictly smaller than one, as shown in remark B.7.

B PROOF OF THEOREM 4.1

In this section, we present the full proof of theorem 4.1. The proof consists of three main steps.

Step 1 (Appendix B.1). We show that, with high probability, the Gaussian initialization has squared norm at most $3V$ and bounded loss (lemma B.1). Due to the weight decay, the objective outside a ball of squared radius $O(V)$ is then larger than its initial value by a constant. This allows us to replace the objective outside the ball by a globally smooth extension (lemma D.3), to which the results on perturbed gradient descent apply.

Step 2 (Appendix B.2). We show that every sufficiently accurate approximate SOSP in the ball has objective below $\mathcal{H}_{[V]} + \Delta_0/8$, which implies gain above $7/8$. To this end, we construct a comparator whose objective is below this level by a fixed margin (lemma B.2), and a direction that represents the comparator and has small softmax curvature (lemma B.3). Then, we show that any point in the ball whose objective is not below this level has either a large gradient or a direction of negative curvature (lemma B.4 and corollary B.5).

Step 3 (Appendix B.3). We partition the iterations into d blocks. By the results of Jin et al. (2019) (theorem D.4 and lemma D.5), with high probability, each block contains an approximate SOSP, and the objective cannot increase much within a block (lemma D.6). We then prove by induction over the blocks that the trajectory stays in the ball and that the gain is above $7/8$ at every iterate after the first block.

Notation and parameters. Throughout this section, $\Theta = (E, U)$, $\mathcal{F}_\lambda = \mathcal{F}_{\lambda,[V]}$, and $\mathcal{R} = \mathcal{R}_{[V]}$. All $2Vd$ entries of Θ are updated by equation (1). The weight decay is

$$\lambda = \frac{\Delta_0}{8BV\sqrt{r}},$$

and we define $t_\star = 1 - \Delta_0/(8B^2)$, which is the factor by which the ground-truth logits are shrunk in the comparator of Step 2. Unless otherwise stated, the constants hidden in $O(\cdot)$ do not depend on V, and their dependence on the fixed parameters (B, r, Δ_0) is made explicit through the 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 $d \geq C_{\mathrm{init}} \log(eV)$,*

$$\Pr\left\{\|E_0\|_{\mathrm{F}}^2 + \|U_0\|_{\mathrm{F}}^2 \leq 3V,\ \|E_0 U_0^\top\|_\infty \leq 1\right\} \geq 1 - e^{-c_0 d}. \tag{4}$$

On this event,

$$\mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} \leq 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}}, \qquad \mathcal{F}_\lambda(\Theta_0) - \mathcal{F}_\lambda^* \leq 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}}.$$

Proof. All entries of E_0, U_0 are independent $\mathcal{N}(0, 1/d)$. Hence $d\left(\|E_0\|_{\mathrm{F}}^2 + \|U_0\|_{\mathrm{F}}^2\right)$ is chi-square with $2Vd$ degrees of freedom. A standard Chernoff bound gives $\Pr\{\|E_0\|_{\mathrm{F}}^2 + \|U_0\|_{\mathrm{F}}^2 > 3V\} \leq e^{-cdV}$ for a universal constant $c > 0$. Applying the same bound to each row of E_0 and taking a union bound yields

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

Conditional on E_0, each entry $(E_0 U_0^\top)_{ij}$ is Gaussian with mean zero and variance $\|E_{0,i,:}\|_2^2/d$. Thus, whenever $\max_i \|E_{0,i,:}\|_2^2 \leq 2$, each variance is at most $2/d$. Gaussian tail bounds and a union bound over all V^2 entries give

$$\Pr\left\{\|E_0 U_0^\top\|_\infty > 1,\ \max_i \|E_{0,i,:}\|_2^2 \leq 2\right\} \leq 2V^2 e^{-d/4}.$$

Combining these estimates, the failure probability in equation (4) is at most

$$e^{-cVd} + Ve^{-cd} + 2V^2e^{-d/4} \le e^{-c_0 d},$$

provided $d \ge C_{\text{init}} \log(eV)$ for a sufficiently large universal C_{init} and a sufficiently small universal $c_0 > 0$.

On the event in equation (4), all initial logits lie in $[-1, 1]$, so each row's cross-entropy loss is at most $\log V + 2$. Hence $\mathcal{R}(\Theta_0) \le \log V + 2$. Using $\lambda = \frac{\Delta_0}{8BV\sqrt{r}}$ and $\mathcal{H}_{[V]} \ge \log V - B$ from equation (6), we obtain

$$\begin{aligned}\mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} &= \mathcal{R}(\Theta_0) - \mathcal{H}_{[V]} + \frac{\lambda}{2}\left(\|E_0\|_{\mathrm{F}}^2 + \|U_0\|_{\mathrm{F}}^2\right) \\ &\le B + 2 + \frac{\Delta_0}{16BV\sqrt{r}} \times 3V \le 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}}.\end{aligned}$$

Finally, $\mathcal{R}(\Theta_0) \ge \mathcal{H}_{[V]}$ and the regularization term is nonnegative, so $\mathcal{F}_\lambda^* \ge \mathcal{H}_{[V]}$. Therefore,

$$\mathcal{F}_\lambda(\Theta_0) - \mathcal{F}_\lambda^* \le \mathcal{F}_\lambda(\Theta_0) - \mathcal{H}_{[V]} \le 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}}.$$

□

On the initialization event in lemma B.1, any $\Theta = (E, U)$ satisfying

$$\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2 \ge \left(6 + \frac{32B\sqrt{r}(B+3)}{\Delta_0}\right) V$$

has regularized loss at least four above its initial value. Indeed, $\mathcal{R}_{[V]}(\Theta) \ge \mathcal{H}_{[V]}$ and $\lambda = \Delta_0/(8BV\sqrt{r})$ give

$$\begin{aligned}\mathcal{F}_\lambda(\Theta) &\ge \mathcal{H}_{[V]} + \frac{\lambda}{2}\left(\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2\right) \\ &\ge \mathcal{H}_{[V]} + 2B + 6 + \frac{3\Delta_0}{8B\sqrt{r}} \ge \mathcal{F}_\lambda(\Theta_0) + 4.\end{aligned} \tag{5}$$

Since $\lambda > 0$, the objective is coercive and attains its global minimum. Every global minimizer lies strictly inside this ball: its objective is at most $\mathcal{F}_\lambda(\Theta_0)$, whereas equation (5) holds on and outside the boundary. The initialization also lies inside because its squared norm is at most $3V$. Thus lemma D.3 gives a smooth extension that agrees with $\mathcal{F}_\lambda$ inside the ball and preserves the exterior loss barrier. Step 3 will show that the noisy trajectory stays inside.

The extension has global gradient- and Hessian-Lipschitz bounds of orders $O(1)$ and $O(\sqrt{V})$, respectively, with constants depending only on (B, r, Δ_0). To verify these orders, the first three derivatives of population cross-entropy with respect to logits have norms $O(V^{-1/2})$, $O(V^{-1})$, and $O(V^{-1})$. On the factor ball above, the first two derivatives of $\Theta \mapsto EU^\top$ have norms $O(\sqrt{V})$ and $O(1)$. A radial clipping map with a fixed relative buffer has first, second, and third derivative norms $O(1)$, $O(V^{-1/2})$, and $O(V^{-1})$. The chain rule gives the stated global smoothness bounds.

B.2 Step 2: Convert Approximate Stationarity into Gain

In this step, we show that every sufficiently accurate approximate SOSP in the ball $\|\Theta\|_{\mathrm{F}}^2 \le C_E V$ satisfies $\mathcal{F}_\lambda(\Theta) < \mathcal{H}_{[V]} + \Delta_0/8$, where C_E is a constant that will be specified in Step 3. Since $\mathcal{R}_{[V]}(\Theta) \le \mathcal{F}_\lambda(\Theta)$ and $\log V - \mathcal{H}_{[V]} \ge \Delta_0$, any such point has gain above $7/8$.

The proof is based on the Hessian of the objective. Recall from lemma D.2 in Appendix D.2 that, for any direction $(\dot{E}, \dot{U})$ with the same dimensions as (E, U),

$$\begin{aligned}\nabla^2\mathcal{F}_\lambda(E,U)[(\dot{E},\dot{U}),(\dot{E},\dot{U})] = &\underbrace{\frac{1}{V}\sum_{i=1}^{V} \mathrm{Var}_{j\sim P_{i,:}^{E,U}}\left[(\dot{E}U^\top + E\dot{U}^\top)_{ij}\right]}_{\text{softmax curvature term}} \\ &+ \underbrace{\frac{2}{V}\langle P^{E,U} - Q, \dot{E}\dot{U}^\top\rangle}_{\text{bilinear term}} + \underbrace{\lambda(\|\dot{E}\|_{\mathrm{F}}^2 + \|\dot{U}\|_{\mathrm{F}}^2)}_{\text{weight decay term}}.\end{aligned}$$

The softmax curvature term is nonnegative, while the bilinear term can be negative. Our goal is to find a direction along which the sum of the three terms is negative, whenever the objective at Θ is not sufficiently below $\mathcal{H}_{[V]} + \Delta_0/8$ and the gradient at Θ is small. We proceed in three parts. First, we construct a comparator $(E_\star, U_\star)$ whose objective is below $\mathcal{H}_{[V]} + \Delta_0/8$ by a fixed margin (lemma B.2). Second, we construct a direction $(\dot{E}, \dot{U})$ with $\dot{E}\dot{U}^\top = E_\star U_\star^\top$ whose softmax curvature term is small (lemma B.3). Third, we show that along this direction, the bilinear term and the weight decay term are negative enough to dominate the softmax curvature term (lemma B.4).

Part 1: A comparator below the target level. We first construct the comparator. A natural candidate is a factorization of the ground-truth logits $\mathcal{L}(\log Q)$, whose prediction loss equals $\mathcal{H}_{[V]}$. By lemma D.1 in Appendix D.1, such a factorization exists with squared norm at most $2B\sqrt{r}\,V$, so that its weight decay penalty is at most $\Delta_0/8$ under our choice of λ. However, this only shows that its objective is at most $\mathcal{H}_{[V]} + \Delta_0/8$, without any margin. To obtain a margin, we shrink the logits by the factor $t_\star = 1 - \Delta_0/(8B^2) < 1$. The shrinkage reduces the penalty by an amount that is linear in $1 - t_\star$, while it increases the prediction loss only by an amount that is quadratic in $1 - t_\star$. The following lemma makes this precise.

Lemma B.2 (Comparator with shrunken ground-truth logits)**.** *For every ground truth in $\mathcal{Q}_{r,B,\Delta_0}$ with* $\operatorname{rank}\mathcal{L}(\log Q) \le d$*,*

$$(\log V - \mathcal{H}_{[V]}) \le \min\left\{B, \frac{B^2}{2}\right\}. \tag{6}$$

In particular, $0 < 1 - t_\star \le 1/16$*. Moreover, there exist factors $E_\star, U_\star \in \mathbb{R}^{V\times d}$ satisfying*

$$E_\star U_\star^\top = t_\star \mathcal{L}(\log Q), \qquad U_\star^\top \mathbf{1} = 0,$$
$$\max_i \|(E_\star)_{i,:}\|_2^2 \le t_\star B\sqrt{r}, \qquad \max_j \|(U_\star)_{j,:}\|_2^2 \le t_\star B\sqrt{r},$$

where all but $\operatorname{rank}\mathcal{L}(\log Q)$ *columns of $E_\star$ and $U_\star$ are zero. These factors satisfy*

$$\mathcal{F}_\lambda(E_\star, U_\star) = \mathcal{R}_{[V]}(E_\star, U_\star) + \frac{\lambda}{2}(\|E_\star\|_{\mathrm{F}}^2 + \|U_\star\|_{\mathrm{F}}^2) \le \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - \frac{\Delta_0^2}{128B^2}. \tag{7}$$

Proof. We first prove equation (6). For every row i and every real number s, the second derivative of the following function of s is a variance, which is at most B^2 because the centered logits are bounded by B:

$$\frac{d^2}{ds^2}\log\left(\frac{1}{V}\sum_j e^{s\mathcal{L}(\log Q)_{ij}}\right) = \operatorname{Var}_{j\sim\operatorname{softmax}(s\mathcal{L}(\log Q)_{i,:})}(\mathcal{L}(\log Q)_{ij}) \le B^2.$$

Since $\mathcal{L}(\log Q)$ is row-centered, both this function and its first derivative vanish at $s = 0$. Thus, integrating by parts, the KL divergence between q_i and the uniform distribution satisfies

$$\mathrm{KL}(q_i\|\mathbf{1}/V) = \int_0^1 s\frac{d^2}{ds^2}\log\left(\frac{1}{V}\sum_j e^{s\mathcal{L}(\log Q)_{ij}}\right) ds \le \frac{B^2}{2}.$$

On the other hand, since $\mathcal{L}(\log Q)$ is row-centered, Jensen's inequality implies that the logarithm in the following display is nonnegative, and thus

$$\mathrm{KL}(q_i\|\mathbf{1}/V) = \sum_j q_{ij}\mathcal{L}(\log Q)_{ij} - \log\left(\frac{1}{V}\sum_j e^{\mathcal{L}(\log Q)_{ij}}\right) \le \max_j \mathcal{L}(\log Q)_{ij} \le B.$$

Since $\log V - \mathcal{H}_{[V]} = \frac{1}{V}\sum_i \mathrm{KL}(q_i\|\mathbf{1}/V)$, averaging the above two bounds over i proves equation (6). In particular, $\Delta_0 \le \log V - \mathcal{H}_{[V]} \le B^2/2$ gives $0 < 1 - t_\star = \Delta_0/(8B^2) \le 1/16$.

Next, we construct the comparator. We apply lemma D.1 to the matrix $t_\star\mathcal{L}(\log Q)$, which has rank at most r, entries bounded by $t_\star B$, and zero row sums. This gives factors with $\operatorname{rank}\mathcal{L}(\log Q)$ columns that satisfy the product identity, the centering condition, and the bounds on the row norms in the lemma. Appending zero columns gives $E_\star, U_\star \in \mathbb{R}^{V\times d}$.

It remains to prove equation (7). For the prediction loss, we view $\mathrm{KL}(q_i \| \operatorname{softmax}(t\mathcal{L}(\log Q)_{i,:}))$ as a function of t. This function and its first derivative vanish at $t = 1$, and its second derivative is the same variance as above, which is at most B^2. Thus, Taylor's theorem gives

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

Consequently,

$$\mathcal{R}_{[V]}(E_\star, U_\star) - \mathcal{H}_{[V]} \le \frac{B^2(1-t_\star)^2}{2} = \frac{\Delta_0^2}{128B^2}. \tag{8}$$

For the weight decay penalty, the bounds on the row norms of $E_\star$ and $U_\star$ imply that $\frac{\lambda}{2}(\|E_\star\|_{\mathrm{F}}^2 + \|U_\star\|_{\mathrm{F}}^2)$ is at most

$$\lambda t_\star B\sqrt{r}\, V = t_\star \frac{\Delta_0}{8}.$$

Since $(1 - t_\star)\Delta_0/8 = \Delta_0^2/(64B^2)$, adding this penalty to equation (8) proves equation (7). □

We remark that the margin $\frac{\Delta_0^2}{128B^2}$ in equation (7) is what allows the objective to increase slightly between two visits to approximate SOSPs in Step 3.

Part 2: A direction with small softmax curvature. Next, we construct the direction $(\dot{E}, \dot{U})$. Since the comparator has at most r nonzero columns, we can place each of its columns along any one of d orthonormal directions in the embedding space without changing the product $E_\star U_\star^\top$. We refer to this construction as packing. The key observation is that the total softmax curvature over all embedding directions is limited. Specifically, the Hessian of the cross-entropy loss of row i with respect to its logits is the covariance matrix $\operatorname{diag}(P_{i,:}^{E,U}) - (P_{i,:}^{E,U})^\top P_{i,:}^{E,U}$, which is positive semidefinite and has trace at most one. Hence, the sum of the traces over the V rows is at most V, and most of the d directions have softmax curvature of order V/d relative to their squared norm. The following lemma chooses r such directions, along which the squared norm of Θ is also small.

Lemma B.3 (Packing the comparator into directions with small softmax curvature)**.** *Suppose $d \ge 2r$ and $\|\Theta\|_{\mathrm{F}}^2 \le C_E V$. Then, there exists a direction $(\dot{E}, \dot{U})$ with $\dot{U}^\top \mathbf{1} = 0$ such that*

$$\dot{E}\dot{U}^\top = t_\star \mathcal{L}(\log Q), \tag{9}$$

$$\|\dot{E}\|_{\mathrm{F}}^2 + \|\dot{U}\|_{\mathrm{F}}^2 \le 2B\sqrt{r}V, \tag{10}$$

$$\frac{1}{V}\sum_i \operatorname{Var}_{j \sim P_{i,:}^{E,U}}\left((\dot{E}U^\top + E\dot{U}^\top)_{ij}\right) \le \frac{8Br^{3/2}C_E}{d}\left(\frac{4V}{d} + 1\right). \tag{11}$$

Proof. We first select the directions in the embedding space. Let $w_1, \ldots, w_d$ be an orthonormal eigenbasis of $U^\top U$. Then, the vectors $Uw_j/\|Uw_j\|_2$ with $Uw_j \neq 0$ are orthonormal in $\mathbb{R}^V$. For a vector $v \in \mathbb{R}^V$, we write $\operatorname{Var}_{P_{i,:}^{E,U}}(v)$ for the variance of its coordinate v_j when j is drawn from the distribution $P_{i,:}^{E,U}$. We have

$$\sum_{j=1}^{d}\left(\|Uw_j\|_2^2 + \|Ew_j\|_2^2\right) = \|U\|_{\mathrm{F}}^2 + \|E\|_{\mathrm{F}}^2 \le C_E V,$$

$$\sum_{j:\, Uw_j \neq 0} \frac{\sum_{i=1}^V \operatorname{Var}_{P_{i,:}^{E,U}}(Uw_j)}{\|Uw_j\|_2^2} \le \sum_{i=1}^V \left(1 - \sum_{\ell=1}^V (P_{i\ell}^{E,U})^2\right) \le V,$$

where the first line follows from the orthonormality of $\{w_j\}$ and the assumption $\|\Theta\|_{\mathrm{F}}^2 \le C_E V$, and the second line follows from the fact that the sum of the quadratic forms of a covariance matrix over orthonormal vectors is at most its trace. Therefore, at most $d/4$ indices j violate the first of the following two bounds, and at most $d/4$ indices violate the second one:

$$\|Uw_j\|_2^2 + \|Ew_j\|_2^2 \le \frac{4C_E V}{d}, \qquad \sum_{i=1}^V \operatorname{Var}_{P_{i,:}^{E,U}}(Uw_j) \le \frac{4V}{d}\|Uw_j\|_2^2.$$

Here, the second bound holds trivially when $Uw_j = 0$. Thus, at least $d/2 \geq r$ indices satisfy both bounds.

Next, we construct the direction. Let $(E_\star, U_\star)$ be the comparator in lemma B.2, whose nonzero columns are the first $\operatorname{rank}\mathcal{L}(\log Q)$ columns. We choose distinct indices $j_1, \ldots, j_{\operatorname{rank}\mathcal{L}(\log Q)}$ that satisfy both bounds above, and signs $\varepsilon_k \in \{-1, 1\}$ that will be specified later. Define

$$\dot{E} = \sum_{k=1}^{\operatorname{rank}\mathcal{L}(\log Q)} \varepsilon_k (E_\star)_{:,k} w_{j_k}^\top, \qquad \dot{U} = \sum_{k=1}^{\operatorname{rank}\mathcal{L}(\log Q)} \varepsilon_k (U_\star)_{:,k} w_{j_k}^\top.$$

Since $\{w_{j_k}\}$ are orthonormal and $\varepsilon_k^2 = 1$, for every choice of the signs we have $\dot{E}\dot{U}^\top = E_\star U_\star^\top = t_\star \mathcal{L}(\log Q)$ and

$$\|\dot{E}\|_{\mathrm{F}}^2 + \|\dot{U}\|_{\mathrm{F}}^2 = \|E_\star\|_{\mathrm{F}}^2 + \|U_\star\|_{\mathrm{F}}^2 \leq 2t_\star B\sqrt{r}V. \tag{12}$$

In addition, $U_\star^\top \mathbf{1} = 0$ implies $\dot{U}^\top \mathbf{1} = 0$. This proves equations (9) and (10) and $\dot{U}^\top \mathbf{1} = 0$.

It remains to choose the signs such that equation (11) holds. By the bounds on the row norms of the comparator in lemma B.2, for every column k and every row i,

$$(E_\star)_{ik}^2 \leq B\sqrt{r}, \qquad \operatorname{Var}_{j \sim P_{i,:}^{E,U}}((U_\star)_{jk}) \leq \sum_j P_{ij}^{E,U} (U_\star)_{jk}^2 \leq B\sqrt{r}.$$

By the definition of $(\dot{E}, \dot{U})$, the i-th row of $\dot{E}U^\top + E\dot{U}^\top$ is the sum over k of $\varepsilon_k\{(E_\star)_{ik} U w_{j_k} + (Ew_{j_k})_i (U_\star)_{:,k}\}^\top$. Suppose that the signs are independent and uniformly distributed on $\{-1, 1\}$. Then, in the expectation of the softmax curvature term over the signs, the cross terms between different columns vanish, and the expectation is the sum of the contributions of the individual columns. Using $\operatorname{Var}(a+b) \leq 2\operatorname{Var}(a) + 2\operatorname{Var}(b)$, the contribution of column k is bounded as

$$\begin{aligned}
&\frac{1}{V}\sum_{i=1}^{V} \operatorname{Var}_{P_{i,:}^{E,U}}((E_\star)_{ik} U w_{j_k} + (Ew_{j_k})_i (U_\star)_{:,k}) \\
&\quad \leq \frac{2B\sqrt{r}}{V}\left\{\sum_{i=1}^{V} \operatorname{Var}_{P_{i,:}^{E,U}}(U w_{j_k}) + \|Ew_{j_k}\|_2^2\right\} \leq \frac{8B\sqrt{r}\, C_E}{d}\left(\frac{4V}{d} + 1\right),
\end{aligned}$$

where the last inequality follows from the two bounds satisfied by the index j_k. Since there are at most r columns, the expectation of the softmax curvature term is at most the right-hand side of equation (11). Therefore, there exists a choice of the signs for which equation (11) holds. □

We note that lemma B.3 only uses an upper bound on the traces of the covariance matrices. Thus, it also holds when some rows of $P^{E,U}$ are nearly deterministic and have very small curvature.

Part 3: Large gradient or negative curvature. We now evaluate the Hessian along the direction in lemma B.3. By equation (11), the softmax curvature term is small when $d \geq C\sqrt{V}$. For the bilinear term and the weight decay term, we use the convexity of the cross-entropy loss in the logits to relate them to the gap between the objective at Θ and the objective at the comparator, which is positive when the objective at Θ is large. The remaining error is the inner product between the gradient and Θ, which is small when the gradient is small. This leads to the following lemma.

Lemma B.4 (Large gradient or negative curvature above the target level)**.** *Fix a constant $C_E > 0$. Suppose $d \geq 2r$ and*

$$d^2 \geq \frac{20480 B^3 r^{3/2} C_E}{\Delta_0^2} V. \tag{13}$$

Then, every point Θ with $\|\Theta\|_{\mathrm{F}}^2 \leq C_E V$ and

$$\mathcal{F}_\lambda(\Theta) - \mathcal{H}_{[V]} \geq \frac{\Delta_0}{8} - \frac{\Delta_0^2}{256 B^2}$$

satisfies

$$\|\nabla \mathcal{F}_\lambda(\Theta)\|_{\mathrm{F}} > \frac{\Delta_0^2}{512 B^2 \sqrt{C_E V}} \quad \textit{or} \quad \lambda_{\min}(\nabla^2 \mathcal{F}_\lambda(\Theta)) \leq -\frac{\Delta_0^2}{512 B^3 \sqrt{r}\, V}. \tag{14}$$

Proof. Suppose that the first inequality in equation (14) does not hold, i.e., $\|\nabla \mathcal{F}_\lambda(\Theta)\|_{\mathrm{F}} \le \Delta_0^2/(512B^2\sqrt{C_E V})$. We show that the second inequality must hold.

First, we bound the inner product between the gradient and Θ. By equation (34), we have $\langle \nabla \mathcal{F}_\lambda(\Theta), \Theta\rangle = \frac{2}{V}\langle P^{E,U} - Q, EU^\top\rangle + \lambda\|\Theta\|_{\mathrm{F}}^2$. Thus, the Cauchy–Schwarz inequality and $\|\Theta\|_{\mathrm{F}} \le \sqrt{C_E V}$ give

$$\left|\frac{1}{V}\langle P^{E,U} - Q, EU^\top\rangle + \frac{\lambda}{2}\|\Theta\|_{\mathrm{F}}^2\right| \le \frac{1}{2}\|\nabla \mathcal{F}_\lambda(\Theta)\|_{\mathrm{F}}\|\Theta\|_{\mathrm{F}} \le \frac{\Delta_0^2}{1024B^2}. \tag{15}$$

Next, let $(\dot{E}, \dot{U})$ be the direction in lemma B.3. Since the cross-entropy loss is convex in the logit matrix, and its gradient with respect to the logits at $EU^\top$ is $(P^{E,U} - Q)/V$, we have $\mathcal{R}_{[V]}(E_\star, U_\star) \ge \mathcal{R}_{[V]}(\Theta) + \frac{1}{V}\langle P^{E,U} - Q, E_\star U_\star^\top - EU^\top\rangle$. Since $E_\star U_\star^\top = t_\star \mathcal{L}(\log Q)$, this is equivalent to

$$\frac{1}{V}\langle P^{E,U} - Q, t_\star \mathcal{L}(\log Q)\rangle \le \mathcal{R}_{[V]}(E_\star, U_\star) - \mathcal{R}_{[V]}(\Theta) + \frac{1}{V}\langle P^{E,U} - Q, EU^\top\rangle.$$

We now bound the sum of the bilinear term and the weight decay term in equation (35) along $(\dot{E}, \dot{U})$:

$$\begin{aligned}
&\frac{2}{V}\langle P^{E,U} - Q, \dot{E}\dot{U}^\top\rangle + \lambda(\|\dot{E}\|_{\mathrm{F}}^2 + \|\dot{U}\|_{\mathrm{F}}^2) \\
&\quad\overset{(a)}{\le} 2\left\{\frac{1}{V}\langle P^{E,U} - Q, EU^\top\rangle + \frac{\lambda}{2}\|\Theta\|_{\mathrm{F}}^2\right\} - 2\{\mathcal{F}_\lambda(\Theta) - \mathcal{F}_\lambda(E_\star, U_\star)\} \\
&\quad\overset{(b)}{\le} \frac{\Delta_0^2}{512B^2} - \frac{\Delta_0^2}{128B^2} = -\frac{3\Delta_0^2}{512B^2},
\end{aligned}$$

where (a) follows from equations (9) and (12) and the convexity inequality above, and (b) follows from equation (15), the assumption $\mathcal{F}_\lambda(\Theta) \ge \mathcal{H}_{[V]} + \Delta_0/8 - \Delta_0^2/(256B^2)$, and equation (7). Finally, we bound the softmax curvature term. Since $d < V$, we have $4V/d + 1 \le 5V/d$. Thus, equations (11) and (13) imply that the softmax curvature term is at most $40Br^{3/2}C_E V/d^2 \le \Delta_0^2/(512B^2)$. Combining the above bounds with equation (35), we obtain

$$\nabla^2 \mathcal{F}_\lambda(\Theta)[(\dot{E}, \dot{U}), (\dot{E}, \dot{U})] \le -\frac{\Delta_0^2}{256B^2}.$$

Since the squared norm of $(\dot{E}, \dot{U})$ is at most $2B\sqrt{r}V$ by equation (10), the minimum eigenvalue of the Hessian is at most $-\Delta_0^2/(512B^3\sqrt{r}\,V)$, which is the second inequality in equation (14). □

Recall that an ϵ-approximate SOSP is defined with a single accuracy parameter ϵ for both the gradient and the Hessian. The following corollary restates lemma B.4 in terms of ϵ.

Corollary B.5 (Approximate SOSPs in the ball have small objective)**.** *Let $\rho \ge 1$ be an upper bound on the Lipschitz constant of the Hessian, and suppose*

$$0 < \epsilon \le \min\left\{\frac{\Delta_0^2}{512B^2\sqrt{C_E V}}, \frac{\Delta_0^4}{2^{20}\rho B^6 r V^2}, 1\right\}. \tag{16}$$

Then, under equation (13)*, every point Θ with $\|\Theta\|_{\mathrm{F}}^2 \le C_E V$ and*

$$\|\nabla \mathcal{F}_\lambda(\Theta)\| \le \epsilon, \qquad \nabla^2 \mathcal{F}_\lambda(\Theta) \succeq -\sqrt{\rho\epsilon}\, I$$

satisfies

$$\mathcal{F}_\lambda(\Theta) < \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - \frac{\Delta_0^2}{256B^2}.$$

Proof. We prove the claim by contradiction. Suppose that $\mathcal{F}_\lambda(\Theta) \ge \mathcal{H}_{[V]} + \Delta_0/8 - \Delta_0^2/(256B^2)$. Then, by lemma B.4, one of the two inequalities in equation (14) holds. The first one contradicts $\|\nabla \mathcal{F}_\lambda(\Theta)\| \le \epsilon$ and the first bound in equation (16). The second one contradicts $\nabla^2 \mathcal{F}_\lambda(\Theta) \succeq -\sqrt{\rho\epsilon}\, I$, because the second bound in equation (16) implies $\sqrt{\rho\epsilon} \le \Delta_0^2/(1024B^3\sqrt{r}\,V)$. □

In summary, Step 2 shows that every ϵ-approximate SOSP in the ball has objective below $\mathcal{H}_{[V]} + \Delta_0/8$ by the margin $\Delta_0^2/(256B^2)$.

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

In this step, we combine Step 1 and Step 2 to prove theorem 4.1. The idea is as follows. By Step 2, whenever the trajectory visits an approximate SOSP in the ball, the objective is below $\mathcal{H}_{[V]} + \Delta_0/8$ by the margin $\Delta_0^2/(256B^2)$. We partition the iterations into d blocks, and choose the step size, the perturbation scale, and the length of the blocks such that, with high probability, each block contains a visit to an approximate SOSP, and the objective increases by at most half of this margin within the block (lemma D.6 in Appendix D.4). Then, the first block establishes the margin at its last iterate, and each later block keeps the objective below $\mathcal{H}_{[V]} + \Delta_0/8$ and restores the margin at its last iterate. Since the results on perturbed gradient descent require global smoothness, they are applied to the smooth extension in Step 1. The induction also shows that the trajectory never leaves the ball, so that it coincides with the trajectory under the extension. For convenience, we restate theorem 4.1 below.

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 the weight decay $\lambda = \Delta_0/(8B\sqrt{r}\,V)$, there exist a step size h, a perturbation scale σ, and a number of iterations T divisible by d such that*

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

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

Proof. **Parameter choices.** Let $\ell, \rho \geq 1$ be the Lipschitz constants of the gradient and the Hessian of the smooth extension in Step 1, which satisfy $\ell = O(1)$ and $\rho = O(\sqrt{V})$ uniformly over the class $\mathcal{Q}_{r,B,\Delta_0}$. Let C_E be the constant in Step 1, i.e.,

$$C_E = 6 + \frac{32B\sqrt{r}(B+3)}{\Delta_0},$$

so that, by equation (5), the ball $\|\Theta\|_{\mathrm{F}}^2 < C_E V$ contains the initialization, and the objective outside the ball is at least $\mathcal{F}_\lambda(\Theta_0) + 4$. For the accuracy, we choose

$$\epsilon = c_{\mathrm{acc}} V^{-5/2},$$

where $0 < c_{\mathrm{acc}} = c_{\mathrm{acc}}(B, r, \Delta_0) \leq 1$ is a sufficiently small constant such that equation (16) holds. Such a constant exists since $\rho/\sqrt{V}$ is uniformly bounded. For the embedding dimension, we choose $C_{\mathrm{wid}} = C_{\mathrm{wid}}(B, r, \Delta_0) \geq 1$ sufficiently large such that, for all $V \geq 2$,

$$C_{\mathrm{wid}}\sqrt{V} \geq \max\left\{ 2r, C_{\mathrm{init}} \log(eV), \sqrt{\frac{20480B^3 r^{3/2} C_E V}{\Delta_0^2}} \right\}.$$

Such a choice exists because $\log(eV)/\sqrt{V}$ is bounded. Then, the condition $d \geq C_{\mathrm{wid}}\sqrt{V}$ guarantees that both lemma B.1 and corollary B.5 are applicable.

Next, we specify the quantities in lemma D.6. We set

$$D_\Theta = 2Vd, \qquad A = 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}},$$
$$\delta_c = \frac{\Delta_0^2}{512B^2}, \qquad \zeta = \frac{1}{4}e^{-d}.$$

Here, A is the bound on the initial objective gap in lemma B.1, and δ_c is half of the margin in corollary B.5. These quantities satisfy the assumptions of lemma D.6, i.e., $A \geq 1$, $0 < \epsilon \leq 1$, and $0 < \delta_c \leq 1/512$, where the last one is due to $\Delta_0 \leq B$. The step size, the perturbation scale, and the number of iterations are chosen as

$$\begin{aligned} \iota &= C_{\mathrm{sched}} d, & h &= (\iota^9 \ell)^{-1}, \quad \sigma = \iota\epsilon, \\ T_{\mathrm{blk}} &= \left\lceil C_J \ell \left(2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}} \right) \epsilon^{-2} \iota^{15} \right\rceil, & T &= dT_{\mathrm{blk}}, \end{aligned} \tag{17}$$

where T_{blk} is the length of each block. With the above choices, the lower bound on ι required in equation (38) is $O(d + \log(eV)) = O(d)$, and $\log(T_{\mathrm{blk}} + 1) = O(\log(eV) + \log C_{\mathrm{sched}})$. Hence, a sufficiently large constant $C_{\mathrm{sched}} = C_{\mathrm{sched}}(B, r, \Delta_0)$ satisfies both this lower bound and $4(T_{\mathrm{blk}} + 1)^2 e^{-\iota} \le e^{-d}/4$. Note that all the above quantities are deterministic and chosen before training, and the same h and σ are used in all blocks.

Coupling with the smooth extension. We condition on the event in equation (4). Let $\widetilde{\mathcal{F}}_\lambda$ be the smooth extension of $\mathcal{F}_\lambda$ in Step 1. Recall that $\widetilde{\mathcal{F}}_\lambda$ has the same minimum as $\mathcal{F}_\lambda$, coincides with $\mathcal{F}_\lambda$ inside the ball $\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2 < C_E V$, and satisfies

$$\widetilde{\mathcal{F}}_\lambda \ge \mathcal{H}_{[V]} + 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}} + 4 \quad \text{outside this ball.} \tag{18}$$

We define an auxiliary trajectory $\{\widetilde{\Theta}_t\}$, which runs perturbed gradient descent on $\widetilde{\mathcal{F}}_\lambda$ with the same initialization and the same perturbations $\{\xi_t\}$ as the original trajectory in equation (1):

$$\widetilde{\Theta}_0 = \Theta_0, \qquad \widetilde{\Theta}_{t+1} = \widetilde{\Theta}_t - h\{\nabla \widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) + \xi_t\}.$$

Since the gradients of $\widetilde{\mathcal{F}}_\lambda$ and $\mathcal{F}_\lambda$ coincide inside the ball, a simple induction on t shows that $\widetilde{\Theta}_t = \Theta_t$ as long as the auxiliary trajectory has not left the ball.

One-block estimate. We partition the iterations $\{0, 1, \dots, T\}$ into d blocks of length T_{blk}, where consecutive blocks share their boundary iterate. Consider any block, and condition on the history up to the first iterate of the block. Suppose that the objective gap $\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) - \mathcal{F}_\lambda^*$ at the first iterate of the block is at most $A = 2B + 2 + 3\Delta_0/(8B\sqrt{r})$. Then, by lemma D.6, with conditional probability at least $1 - e^{-d}/2$, the block contains an iterate that is an ϵ-approximate SOSP of $\widetilde{\mathcal{F}}_\lambda$, and

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_v) - \widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_u) \le \frac{\Delta_0^2}{512B^2} \quad \text{for every integer pair } u < v \text{ in the block.} \tag{19}$$

We say that a block is *successful* if both events occur. Since equation (19) holds for all pairs of iterates in the block, it can be applied with u being the random iteration at which the approximate SOSP is visited. In the following, we show by induction that, as long as all blocks so far are successful, the condition on the objective gap holds at the first iterate of the next block, and the two trajectories coincide.

Base case: the first block. By lemma B.1, the objective gap at Θ_0 is at most A, and thus the one-block estimate applies to the first block. Suppose that the first block is successful. Applying equation (19) with $u = 0$, every auxiliary iterate in the block satisfies $\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) < \mathcal{F}_\lambda(\Theta_0) + 1 \le \mathcal{H}_{[V]} + A + 1$, which is strictly smaller than the lower bound in equation (18). Hence, every auxiliary iterate in the block lies inside the ball, and $\widetilde{\Theta}_t = \Theta_t$ for all t in the block. In particular, the ϵ-approximate SOSP visited by the auxiliary trajectory, say at iteration τ_0, is also an ϵ-approximate SOSP of the original objective $\mathcal{F}_\lambda$ in the ball. By corollary B.5, we have

$$\mathcal{F}_\lambda(\Theta_{\tau_0}) < \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - \frac{\Delta_0^2}{256B^2}.$$

Applying equation (19) with $u = \tau_0$ and $v = T_{\mathrm{blk}}$, we obtain

$$\mathcal{F}_\lambda(\Theta_{jT_{\mathrm{blk}}}) < \mathcal{H}_{[V]} + \frac{\Delta_0}{8} - \frac{\Delta_0^2}{512B^2} \qquad (j = 1). \tag{20}$$

If $\tau_0 = T_{\mathrm{blk}}$, equation (20) follows directly from the bound at τ_0.

Induction step: the later blocks. Suppose that the first j blocks are successful for some $1 \le j < d$. The induction hypothesis consists of two parts: (H1) the two trajectories stay inside the ball and coincide up to iteration jT_{blk}; (H2) equation (20) holds at iteration jT_{blk}. Both parts hold for $j = 1$ by the base case. We show that, if the $(j+1)$-th block is also successful, then (H1) and (H2) hold with j replaced by $j + 1$, and the gain is above $7/8$ at every iterate in the $(j+1)$-th block.

First, by (H2) and $\mathcal{F}_\lambda^* \ge \mathcal{H}_{[V]}$, the objective gap at the first iterate of the $(j+1)$-th block satisfies

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_{jT_{\mathrm{blk}}}) - \mathcal{F}_\lambda^* < \frac{\Delta_0}{8} \le 2B + 2 + \frac{3\Delta_0}{8B\sqrt{r}}.$$

Thus, the one-block estimate applies to the $(j+1)$-th block conditionally on the history. Suppose that this block is successful. Then, by equation (19) with $u = jT_{\mathrm{blk}}$ and (H2), every iteration t in the block satisfies

$$\widetilde{\mathcal{F}}_\lambda(\widetilde{\Theta}_t) \le \mathcal{F}_\lambda(\Theta_{jT_{\mathrm{blk}}}) + \frac{\Delta_0^2}{512B^2} < \mathcal{H}_{[V]} + \frac{\Delta_0}{8}.$$

Since $\Delta_0 \le B$, the right-hand side is smaller than the lower bound in equation (18). Hence, the auxiliary trajectory stays inside the ball, and $\widetilde{\Theta}_t = \Theta_t$ for all t in the block, which proves (H1) for $j+1$. Moreover, since the weight decay penalty is nonnegative,

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

which implies $g_{[V]}(\Theta_t) > 7/8$ for all t in the block. Finally, by the same argument as in the base case, i.e., applying corollary B.5 at the approximate SOSP visited in this block and then equation (19), we conclude that equation (20) holds with j replaced by $j+1$, which proves (H2) for $j+1$. By induction, if all d blocks are successful, then $g_{[V]}(\Theta_t) > 7/8$ for every integer $t \in [T_{\mathrm{blk}}, T]$.

Failure probability and running time. Conditional on the initialization event and on the success of all previous blocks, each block fails with probability at most $e^{-d}/2$. Thus, by the tower property and the union bound, which do not require the blocks to be independent, we have

$$\Pr\left\{\exists t \in [T_{\mathrm{blk}}, T] : g_{[V]}(\Theta_t) \le \frac{7}{8}\right\} \le e^{-c_0 d} + \frac{d}{2}e^{-d} \le e^{-c_{\mathrm{u}} d},$$

for a universal constant $c_{\mathrm{u}} > 0$, provided that $V \ge V_0$ for a sufficiently large $V_0 = V_0(B, r, \Delta_0)$. Since $T_{\mathrm{blk}} = T/d$, this proves the probability bound in the theorem with $c = c_{\mathrm{u}}$ and $C = C_{\mathrm{wid}}$. For the running time, equation (17) and $\ell = O(1)$ give

$$h^{-1} = O(d^9), \qquad \sigma^{-1} = O(V^{5/2}/d), \qquad T_{\mathrm{blk}} \le C_T V^5 d^{15}, \qquad T \le C_T V^5 d^{16} \le C_T V^{21},$$

where C_T and the hidden constants depend only on (B, r, Δ_0). Thus, h^{-1}, σ^{-1}, and T are bounded by polynomials in V. We also note that, on the same event, only the first T_{blk} iterates may have gain below $7/8$. Hence, the fraction of iterates with gain above $7/8$ satisfies

$$\frac{1}{T+1}\sum_{t=0}^{T} \mathbf{1}\left\{g_{[V]}(\Theta_t) > \frac{7}{8}\right\} \ge 1 - \frac{T_{\mathrm{blk}}}{dT_{\mathrm{blk}} + 1} > 1 - \frac{1}{d}.$$

□

B.4 Comparison with a Generic Factorized Bound

In this subsection, we compare our analysis in Step 2 with a direct application of the generic theory for factorized optimization (Ouyang et al., 2025), and explain why the latter requires a larger embedding dimension. For simplicity, consider an exact SOSP $\Theta = (E, U)$ of the untied objective $\mathcal{F}_\lambda$ with $\lambda > 0$ and $\|\Theta\|_{\mathrm{F}}^2 \le C_E V$. By lemma D.2, at such a point, we have $\mathbf{1}^\top U = 0$, and the weight decay penalty equals $\lambda\|EU^\top\|_*$. Then, for any comparator matrix M with $M\mathbf{1} = 0$, Theorem 4.1 of Ouyang et al. (2025), together with the convexity of the loss and the stationarity of Θ, gives

$$\begin{aligned}
&\mathcal{F}_\lambda(\Theta) - \left\{\frac{1}{V}\sum_i \mathrm{CE}(q_i, M_{i,:}) + \lambda\|M\|_*\right\} \\
&\quad\le \left\{\frac{1}{V}\sup_{\|H\|_{\mathrm{F}}=1,\,\mathrm{rank}\,H=2} \sum_i \mathrm{Var}_{j\sim P_{i,:}^{E,U}}(H_{ij})\right\}\sigma_d(EU^\top)\|M\|_* \\
&\quad\le \frac{\sigma_d(EU^\top)\|M\|_*}{2V}.
\end{aligned}$$

Here, the expression in the braces is the largest curvature of the loss along a rank-two direction in the logit space, which is at most $1/(2V)$ for categorical distributions. Moreover, we have $\sigma_d(EU^\top) \le C_E V/(2d)$, and when the comparator is given by the ground-truth logits, which have rank at most r and entries bounded by B, we have $\|M\|_* \le B\sqrt{r}\,V$. Therefore, the above bound only gives a comparison error of $O(V/d)$, which is not small when $d \asymp \sqrt{V}$.

In contrast, our analysis uses the sum of the traces of the Hessians over all rows, instead of a uniform bound on the curvature along each direction. This leads to the bound $O(V/d^2 + 1/d)$ in equation (11), which is sufficiently small for the argument in Step 2 as long as $d \geq C\sqrt{V}$. We note that this comparison only explains the improvement in our sufficient condition on d, and it is not a lower bound on the embedding dimension required by other methods.

Remark B.7 (Prediction loss versus the regularized objective)**.** We note that a global minimizer of $\mathcal{F}_\lambda$ is not Bayes-optimal. Indeed, when $\lambda > 0$ and $\log V - \mathcal{H}_{[V]} > 0$, the centered ground-truth logits $\mathcal{L}(\log Q)$ are nonzero, and the derivative of $t \mapsto V^{-1} \sum_i \mathrm{CE}(q_i, t\mathcal{L}(\log Q)_{i,:}) + \lambda t \|\mathcal{L}(\log Q)\|_*$ at $t = 1$ equals $\lambda \|\mathcal{L}(\log Q)\|_* > 0$. Thus, slightly shrinking the ground-truth logits decreases the regularized objective. For this reason, theorem 4.1 is stated as a guarantee on the prediction gain, and not as the convergence to the Bayes predictor.

C Tied Training and Transfer

C.1 Proof of Theorem 5.2

In this section, we present the full proof of theorem 5.2. The proof consists of three main steps, which are parallel to those in Appendix B.

Step 1 (Appendix C.1.1). We bound the objective at the Gaussian initialization, and construct a smooth extension of the objective that coincides with it in a ball and is large outside the ball (lemma C.1).

Step 2 (Appendix C.1.2). We compare the objective at an approximate SOSP with the objective at the ground truth. Specifically, we construct a direction that represents the ground-truth Gram matrix, along which the softmax curvature term of the Hessian is at most $O(\|X\|_{\mathrm{F}}^2/d^2)$ (lemma C.2). Then, we use the weight decay to absorb this term, which gives $\|X\|_{\mathrm{F}}^2 = O(V)$ and a comparison error of $O(V/d^2)$ (lemma C.3).

Step 3 (Appendix C.1.3). We apply the one-block estimate (lemma D.6) to each of the d blocks of iterations, and prove by induction that the comparison in Step 2 holds at every iterate in the final window (lemma C.4). Finally, we specify the weight decay and prove theorem 5.2.

We remark that the lemmas in this section are stated for a general weight decay λ, with the bounds expressed in terms of λV. Thus, they can also be used in the proof of theorem 5.4 in Appendix C.2, where a different λ is chosen.

Notation. Recall that $\mathcal{S} \subset [V]$ is the set of observed current states, and $\mathcal{U} = [V] \setminus \mathcal{S}$ is the set of unseen states. Throughout this section, we assume

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

where the constants in $N = \Theta(V)$ are fixed. In particular, $d < V - 1$. Note that all V states, including those in $\mathcal{U}$, are possible next states of every observed current state. For a matrix A, the subscripts $_{\mathcal{S},:}$ and $_{:,\mathcal{S}}$ select the rows and the columns indexed by $\mathcal{S}$, respectively. In particular, $(I_V)_{:,\mathcal{S}}$ consists of the columns of I_V indexed by $\mathcal{S}$, and multiplying an N-row matrix by it from the left extends the matrix by zero rows outside $\mathcal{S}$. We use $\|A\|_\infty$ to denote the maximum absolute entry of A, and $\|A\|_1 = \sum_{ij} |A_{ij}|$. For the tied model, we write x_i for the i-th row of X as a column vector, and $K = XX^\top$ for the learned Gram matrix, so that $P^X = \mathrm{softmax}(XX^\top) = \mathrm{softmax}(\mathcal{L}(K))$. All Vd entries of X are trained. The risk, the gain, and the objective are defined as in Appendix A. In particular,

$$\mathcal{R}_{\mathcal{A}}(X) - \mathcal{H}_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathrm{KL}(q_i \| P^X_{i,:}), \qquad \mathcal{F}_{\lambda,\mathcal{S}}(X) = \mathcal{R}_{\mathcal{S}}(X) + \frac{\lambda}{2} \|X\|_{\mathrm{F}}^2.$$

We assume that the information gap on the observed states satisfies $\log V - \mathcal{H}_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$ and, when the gain on the unseen states is considered, $\log V - \mathcal{H}_{\mathcal{U}} \geq \delta_{\mathrm{te}} > 0$. Note that these conditions are imposed on the averages over $\mathcal{S}$ and $\mathcal{U}$, and not on each individual state.

Ground truth. Let $Q \in \mathcal{Q}^{\mathrm{tied}}_{r,B,\Delta_0}$ with the constants $B_Y \geq 1$ and $r \geq 1$ in definition A.2. We fix a representative Y of Q with linearly independent columns, i.e.,

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

Such a representative always exists. Indeed, if the columns of a representative Y in definition A.2 are linearly dependent, we can write $Y = Y'R^\top$, where Y' has linearly independent columns and R has orthonormal columns. Then, Y' has the same Gram matrix and the same row norms as Y. The bound on the row norms implies $\|\mathcal{L}(\log Q)\|_\infty \leq 2B_Y$. Note that the centering condition $Y^\top \mathbf{1} = 0$ and the coverage condition are not required in this section. The objective at the ground-truth embedding matrix Y, padded with zero columns, is $\mathcal{H}_\mathcal{S} + \frac{\lambda}{2}\|Y\|_\mathrm{F}^2$, which serves as the reference value in the comparisons below.

Parameters and algorithm. Let $C_p = 40B_Y r$. Throughout this section, we assume $2r_Y \leq d < N < V$, $d \geq C\log(eV)$, $\lambda \geq 2C_p/d^2$, and $0 < \lambda V \leq 1$. The entries of X_0 are independent $\mathcal{N}(0, 1/d)$ random variables, and the embeddings are updated by

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

The step size h and the perturbation scale σ are chosen before training and fixed throughout training.

C.1.1 Step 1: Localize the Trajectory

In this step, we show that the Gaussian initialization has bounded objective with high probability, and that the objective is large outside a ball with a deterministic radius R_w, due to the weight decay. Since the results on perturbed gradient descent require global smoothness, we also construct a smooth extension of the objective outside the ball. We note that the radius R_w in the following lemma depends on λV, and can be much larger than $\sqrt{V}$. A sharper bound $\|X\|_\mathrm{F}^2 = O(V)$ at approximate SOSPs will be established in Step 2.

Lemma C.1 (Initialization and smooth extension)**.** *For $d \geq C\log(eV)$, with probability at least $1 - e^{-cd}$,*

$$\max_i \|(X_0)_i\|_2^2 \leq 2, \qquad \|X_0\|_\mathrm{F}^2 \leq 2V, \qquad \mathcal{F}_{\lambda,\mathcal{S}}(X_0) - \mathcal{H}_\mathcal{S} \leq 2B_Y + 8 + (\lambda V). \tag{21}$$

Moreover, let $A = 4B_Y + 20 + C_p^{-1}$ and

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

Then, there exists a smooth function $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}} : \mathbb{R}^{Vd} \to \mathbb{R}$ that coincides with $\mathcal{F}_{\lambda,\mathcal{S}}$ on the ball $\|X\|_\mathrm{F} \leq R_w$, has the same global minimum as $\mathcal{F}_{\lambda,\mathcal{S}}$, and satisfies $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}} > \mathcal{H}_\mathcal{S} + A + 4$ outside the ball. Moreover, the Lipschitz constants $\ell_{\mathrm{ext}}, \rho_{\mathrm{ext}} \geq 1$ of its gradient and its Hessian can be chosen to be polynomial in V, N, and $(\lambda V)^{-1}$.

Proof. We first prove equation (21). For each row i, the random variable $d\|(X_0)_i\|_2^2$ follows the χ_d^2 distribution. Thus, the Chernoff bound and the union bound over the V rows give

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

for a universal constant $c_0 > 0$. Since $d \geq C\log(eV)$, the right-hand side is at most e^{-cd} after adjusting the constants. On this event, we have $\|X_0\|_\mathrm{F}^2 \leq 2V$, and the entries of $X_0X_0^\top$ are at most two in absolute value. Thus, the entries of the centered logits are at most four in absolute value, which implies $\mathcal{R}_\mathcal{S}(X_0) \leq \log V + 8$. Moreover, since the entries of $YY^\top$ lie in $[-B_Y, B_Y]$, we have $q_{ij} \leq e^{2B_Y}/V$, and thus $\log V + \sum_j q_{ij}\log q_{ij} \leq 2B_Y$ for every i, i.e., $\log V - \mathcal{H}_\mathcal{S} \leq 2B_Y$. Combining these two bounds with the penalty $\frac{\lambda}{2}\|X_0\|_\mathrm{F}^2 \leq \lambda V$ gives the last bound in equation (21).

Next, we show that the objective is large outside the ball. Since the cross-entropy loss is lower bounded by $\mathcal{H}_\mathcal{S}$ and the weight decay penalty is coercive, a global minimizer of $\mathcal{F}_{\lambda,\mathcal{S}}$ exists. Comparing it with the ground-truth embedding matrix Y padded with zero columns, whose squared norm is at

most $B_Y V$, we conclude that every global minimizer has squared norm at most $B_Y V < R_w^2$, and thus lies in the ball. On the other hand, for any X with $\|X\|_{\mathrm{F}} \geq R_w$, the definition of C_w gives $\mathcal{F}_{\lambda,\mathcal{S}}(X) \geq \mathcal{H}_{\mathcal{S}} + \lambda R_w^2/2 > \mathcal{H}_{\mathcal{S}} + A + 4$.

We now construct the extension. Let $s : [0, \infty) \to [0, 2R_w]$ be a smooth function such that $s(a) = a$ for $a \leq R_w$, $s(a) \geq R_w$ for $a \geq R_w$, and s is constant for $a \geq 3R_w$. Such a function can be chosen with its k-th derivative bounded by a constant multiple of R_w^{1-k}. For $X \neq 0$, we define

$$\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(X) = \mathcal{F}_{\lambda,\mathcal{S}}\left(s(\|X\|_{\mathrm{F}})\frac{X}{\|X\|_{\mathrm{F}}}\right),$$

and set $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(0) = \mathcal{F}_{\lambda,\mathcal{S}}(0)$. By construction, $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ coincides with $\mathcal{F}_{\lambda,\mathcal{S}}$ on the ball $\|X\|_{\mathrm{F}} \leq R_w$. Outside the ball, $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(X)$ is the value of $\mathcal{F}_{\lambda,\mathcal{S}}$ at a point with norm at least R_w, which is larger than $\mathcal{H}_{\mathcal{S}} + A + 4$. Since all global minimizers of $\mathcal{F}_{\lambda,\mathcal{S}}$ lie in the ball, $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ has the same global minimum.

It remains to bound the Lipschitz constants. For any directions $Z, W \in \mathbb{R}^{V\times d}$, the first and second variations of the centered logits of the observed rows satisfy

$$\|\mathcal{L}(XZ^\top + ZX^\top)_{\mathcal{S},:}\|_{\mathrm{F}} \leq 2\|X\|_{\mathrm{F}}\|Z\|_{\mathrm{F}},$$
$$\|\mathcal{L}(ZW^\top + WZ^\top)_{\mathcal{S},:}\|_{\mathrm{F}} \leq 2\|Z\|_{\mathrm{F}}\|W\|_{\mathrm{F}}.$$

In addition, the first, second, and third derivatives of the average cross-entropy loss with respect to the logits have norms at most $\sqrt{2/N}$, $1/(2N)$, and C/N, respectively. Thus, by the chain rule, on the ball of radius $2R_w$, which contains the image of the map $X \mapsto s(\|X\|_{\mathrm{F}})X/\|X\|_{\mathrm{F}}$, the Lipschitz constants of the gradient and the Hessian of $\mathcal{F}_{\lambda,\mathcal{S}}$ are at most $C(1 + R_w^2/N) + \lambda$ and $C(1 + (R_w^3 + R_w)/N)$, respectively. Applying the chain rule again to the composition with this map, whose derivatives are bounded as above, we conclude that ℓ_{ext} and ρ_{ext} are polynomial in V, N, and $(\lambda V)^{-1}$. □

C.1.2 Step 2: Convert Approximate Stationarity into Gain

In this step, we compare the objective at an approximate SOSP with the objective at the ground truth, i.e., $\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2$. The argument is parallel to Step 2 in Appendix B.2, with two differences. First, since the input and output embeddings are tied, a perturbation of X changes both of them, and thus the gradient and the Hessian take a different form. Second, the softmax curvature term along the constructed direction is proportional to $\|X\|_{\mathrm{F}}^2/d^2$, where $\|X\|_{\mathrm{F}}^2$ is not known to be $O(V)$ in advance. We handle the second issue in lemma C.3 by using the weight decay.

The following lemma provides the gradient and the Hessian of the tied objective, and constructs the direction by the same packing argument as in lemma B.3.

Lemma C.2 (Derivatives of the tied objective and comparison with the ground truth)**.** *Let e_i denote the ith coordinate vector in $\mathbb{R}^V$ and let x_i denote row i of X as a column vector. For every direction Z,*

$$\nabla\mathcal{F}_{\lambda,\mathcal{S}}(X) = \frac{1}{N}\sum_{i\in\mathcal{S}}\sum_{j=1}^{V}(P_{ij}^X - q_{ij})(e_i x_j^\top + e_j x_i^\top) + \lambda X,$$
$$\nabla^2\mathcal{F}_{\lambda,\mathcal{S}}(X)[Z,Z] = \frac{1}{N}\sum_{i\in\mathcal{S}}\mathrm{Var}_{P_{i,:}^X}((XZ^\top + ZX^\top)_{i,:}) + \frac{2}{N}\sum_{i\in\mathcal{S}}\sum_{j=1}^{V}(P_{ij}^X - q_{ij})(ZZ^\top)_{ij} + \lambda\|Z\|_{\mathrm{F}}^2,$$

$$\frac{1}{N}\sum_{i\in\mathcal{S}}\sum_{j=1}^{V}(P_{ij}^X - q_{ij})K_{ij} + \frac{\lambda}{2}\|X\|_{\mathrm{F}}^2 = \tfrac{1}{2}\langle\nabla\mathcal{F}_{\lambda,\mathcal{S}}(X), X\rangle. \tag{22}$$

We refer to the first term on the right-hand side of the second identity as the softmax curvature term. Under definition A.2, if $d \geq 2r_Y$, there exists a direction $Z \in \mathbb{R}^{V\times d}$ with $ZZ^\top = K^\star$ and

$\|Z\|_{\mathrm{F}}^2 \le B_Y V$, *along which the softmax curvature term is at most* $C_p\|X\|_{\mathrm{F}}^2/d^2$, *where* $C_p = 40B_Y r$. *Consequently, if* $\|\nabla \mathcal{F}_{\lambda,\mathcal{S}}(X)\|_{\mathrm{F}} \le \eta$ *and* $\nabla^2 \mathcal{F}_{\lambda,\mathcal{S}}(X) \succeq -\gamma I$, *then*

$$\mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \frac{C_p\|X\|_{\mathrm{F}}^2}{2d^2} + \frac{\eta\|X\|_{\mathrm{F}}}{2} + \frac{\gamma B_Y V}{2}. \tag{23}$$

Proof. We first prove equation (22). The first and second variations of $XX^\top$ along Z are $XZ^\top + ZX^\top$ and $2ZZ^\top$, respectively. Since the gradient of the cross-entropy loss of row i with respect to its logits is $P_{i,:}^X - q_i^\top$, and its Hessian is the covariance matrix of $P_{i,:}^X$, the chain rule gives the first two identities. Here, we use the fact that the variance is invariant to adding a constant, so that the centering of the logits does not affect the Hessian. The third identity follows from taking the inner product of the gradient with X, since $\langle e_i x_j^\top + e_j x_i^\top, X\rangle = 2K_{ij}$.

Next, we construct the direction Z by the same packing argument as in the proof of lemma B.3. Let $w_1, \dots, w_d$ be an orthonormal eigenbasis of $X^\top X$. Then, the vectors $Xw_j/\|Xw_j\|_2$ with $Xw_j \neq 0$ are orthonormal. Since the covariance matrix of each row has trace at most one, we have

$$\sum_{j=1}^{d} \|Xw_j\|_2^2 = \|X\|_{\mathrm{F}}^2, \qquad \sum_{j:Xw_j\neq 0} \sum_{i\in\mathcal{S}} \frac{\mathrm{Var}_{P_{i,:}^X}(Xw_j)}{\|Xw_j\|_2^2} \le N.$$

Consequently, at least $d/2$ indices j satisfy both of the following bounds:

$$\|Xw_j\|_2^2 \le \frac{4\|X\|_{\mathrm{F}}^2}{d}, \qquad \sum_{i\in\mathcal{S}} \mathrm{Var}_{P_{i,:}^X}(Xw_j) \le \frac{4N}{d}\|Xw_j\|_2^2.$$

Since $d \ge 2r_Y$, we can choose r_Y distinct indices $j_1, \dots, j_{r_Y}$ among them. Let $y^{(k)}$ be the k-th column of Y, and define $Z = \sum_{k=1}^{r_Y} \varepsilon_k y^{(k)} w_{j_k}^\top$ with signs $\varepsilon_k \in \{-1, 1\}$. For every choice of the signs, the orthonormality of $\{w_{j_k}\}$ gives $ZZ^\top = YY^\top = K^\star$ and $\|Z\|_{\mathrm{F}}^2 = \|Y\|_{\mathrm{F}}^2 \le B_Y V$. To choose the signs, suppose that they are independent and uniformly distributed. Then, in the expectation of the softmax curvature term $\frac{1}{N}\sum_{i\in\mathcal{S}} \mathrm{Var}_{P_{i,:}^X}((XZ^\top + ZX^\top)_{i,:})$, the cross terms between different columns vanish, and the contribution of the k-th column is at most

$$\frac{2B_Y}{N}\left\{\|Xw_{j_k}\|_2^2 + \sum_{i\in\mathcal{S}} \mathrm{Var}_{P_{i,:}^X}(Xw_{j_k})\right\} \le \frac{8B_Y\|X\|_{\mathrm{F}}^2}{Nd}(1 + 4N/d).$$

Here, we use $\mathrm{Var}(a+b) \le 2\,\mathrm{Var}(a) + 2\,\mathrm{Var}(b)$, $\mathrm{Var}_{P_{i,:}^X}(y^{(k)}) \le B_Y$, $|y_i^{(k)}|^2 \le B_Y$, and the two bounds satisfied by the index j_k. Summing over the $r_Y \le r$ columns and using $d < N$, the expectation of the softmax curvature term is at most $C_p\|X\|_{\mathrm{F}}^2/d^2$. Hence, there exists a choice of the signs for which the softmax curvature term is at most $C_p\|X\|_{\mathrm{F}}^2/d^2$.

Finally, we prove equation (23). Note that the function $K \mapsto N^{-1}\sum_{i\in\mathcal{S}} \mathrm{CE}(q_i, \mathcal{L}(K)_{i,:}) + \lambda\,\mathrm{tr}\,K/2$ is convex in the symmetric matrix K. It equals $\mathcal{F}_{\lambda,\mathcal{S}}(X)$ at $K = XX^\top$, and $\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2$ at $K = K^\star$. Thus, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in\mathcal{S}}\sum_{j=1}^{V}(P_{ij}^X - q_{ij})K_{ij}^\star + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 \\
&\overset{(a)}{\le} \mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 - \mathcal{F}_{\lambda,\mathcal{S}}(X) + \tfrac{1}{2}\langle\nabla\mathcal{F}_{\lambda,\mathcal{S}}(X), X\rangle \\
&\overset{(b)}{\le} \mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 - \mathcal{F}_{\lambda,\mathcal{S}}(X) + \eta\|X\|_{\mathrm{F}}/2,
\end{aligned}$$

where (a) follows from the first-order condition of convexity at $K = XX^\top$ and the last identity in equation (22), and (b) follows from $\|\nabla\mathcal{F}_{\lambda,\mathcal{S}}(X)\|_{\mathrm{F}} \le \eta$ and the Cauchy–Schwarz inequality. On the

other hand, applying the second identity in equation (22) to the direction Z constructed above, and using $ZZ^\top = K^\star$, $\|Z\|_{\mathrm{F}}^2 = \|Y\|_{\mathrm{F}}^2 \le B_Y V$, and $\nabla^2 \mathcal{F}_{\lambda,\mathcal{S}}(X) \succeq -\gamma I$, we obtain

$$-\gamma B_Y V \le \nabla^2 \mathcal{F}_{\lambda,\mathcal{S}}(X)[Z,Z] \le \frac{C_p \|X\|_{\mathrm{F}}^2}{d^2} + 2\left\{ \frac{1}{N} \sum_{i \in \mathcal{S}} \sum_{j=1}^{V} (P_{ij}^X - q_{ij}) K_{ij}^\star + \frac{\lambda}{2} \|Y\|_{\mathrm{F}}^2 \right\}.$$

Combining the above two inequalities and rearranging the terms proves equation (23). □

Note that the right-hand side of equation (23) depends on the squared norm $\|X\|_{\mathrm{F}}^2$ of the learned embeddings, which can be as large as R_w^2 in the ball. On the other hand, the objective on the left-hand side contains the weight decay penalty $\frac{\lambda}{2}\|X\|_{\mathrm{F}}^2$. Thus, when $\lambda \ge 2C_p/d^2$, the term $C_p\|X\|_{\mathrm{F}}^2/(2d^2)$ can be absorbed by the penalty. The following lemma shows that this first gives $\|X\|_{\mathrm{F}}^2 = O(V)$, and then a comparison error of order V/d^2.

Lemma C.3 (Approximate SOSPs have squared norm $O(V)$ and small objective)**.** *Let $\lambda \ge 2C_p/d^2$, and let R_w be the radius in lemma C.1. If $\|X\|_{\mathrm{F}} \le R_w$, $\|\nabla \mathcal{F}_{\lambda,\mathcal{S}}(X)\|_{\mathrm{F}} \le V/(4R_w d^2)$, and $\nabla^2 \mathcal{F}_{\lambda,\mathcal{S}}(X) \succeq -I/(4B_Y d^2)$, then*

$$\|X\|_{\mathrm{F}}^2 \le \left(2B_Y + \frac{1}{2C_p}\right) V, \qquad \mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le (C_p B_Y + 1/2)\frac{V}{d^2}.$$

Proof. We apply equation (23) with $\eta = V/(4R_w d^2)$ and $\gamma = 1/(4B_Y d^2)$. Since $\|X\|_{\mathrm{F}} \le R_w$, each of the last two terms in equation (23) is at most $V/(8d^2)$, i.e., their sum is at most $V/(4d^2)$. Thus, using $\|Y\|_{\mathrm{F}}^2 \le B_Y V$, we have

$$\begin{aligned} \mathcal{F}_{\lambda,\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} - \lambda V B_Y/2 &\overset{(a)}{\le} \lambda \|X\|_{\mathrm{F}}^2/4 + V/(4d^2) \\ &\overset{(b)}{\le} \tfrac{1}{2}\big(\mathcal{F}_{\lambda,\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}\big) + V/(4d^2), \end{aligned}$$

where (a) follows from equation (23) and $\lambda \ge 2C_p/d^2$, and (b) follows from $\mathcal{F}_{\lambda,\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \ge \lambda\|X\|_{\mathrm{F}}^2/2$. Rearranging the terms, we have $\mathcal{F}_{\lambda,\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le \lambda V B_Y + V/(2d^2)$ and

$$\|X\|_{\mathrm{F}}^2 \le 2B_Y V + \frac{V}{\lambda d^2} \le \left(2B_Y + \frac{1}{2C_p}\right) V.$$

Substituting this bound on $\|X\|_{\mathrm{F}}^2$ back into equation (23) gives

$$\mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le (C_p B_Y + 1/2)\frac{V}{d^2}.$$

Note that the final comparison error does not depend on the radius R_w, and thus has no factor of $(\lambda V)^{-1}$.

□

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

In this step, we extend the comparison in Step 2, which holds at approximate SOSPs, to every iterate in the final window. The argument is the same as Step 3 in Appendix B.3. We partition the iterations into d blocks, and apply the one-block estimate (lemma D.6 in Appendix D.4) to the smooth extension in Step 1. With high probability, each block contains a visit to an approximate SOSP, at which lemma C.3 applies, and the objective increases by at most $V/(4d^2)$ within the block. An induction over the blocks then gives the following lemma.

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

$$C_p = 40 B_Y r, \qquad \lambda \ge 2C_p/d^2, \quad 0 < \lambda V \le 1.$$

Then, there exist deterministic $h, \sigma > 0$ and an integer T_{blk}, which are specified in the proof, such that h^{-1} and σ^{-1} are bounded by polynomials in $(V, N, d, (\lambda V)^{-1})$, and

$$T = dT_{\mathrm{blk}} \le p_{\mathrm{time}}(V, N, d, (\lambda V)^{-1}), \tag{24}$$

where p_{time} is a fixed polynomial whose coefficients depend only on the parameters of the class. Moreover, with probability at least $1 - e^{-cd}$ over the initialization and the perturbations, every integer $T_{\mathrm{blk}} \le t \le T$ satisfies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \frac{(C_p B_Y + 1)V}{d^2},$$
$$g_{\mathcal{S}}(X_t) \ge 1 - \frac{B_Y(\lambda V)/2 + \frac{(C_p B_Y+1)V}{d^2}}{\delta_{\mathrm{tr}}}.$$

In particular, both bounds hold at the final iterate $t = T$.

Proof. The proof follows the same structure as the proof of theorem 4.1 in Appendix B.3.

Parameter choices. Let A, R_w, $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$, ℓ_{ext}, and ρ_{ext} be as in lemma C.1. We set

$$\epsilon = \min\left\{\frac{V}{4R_w d^2}, \frac{1}{16 B_Y^2 d^4 \rho_{\mathrm{ext}}}, 1\right\},$$
$$\delta_c = \min\{V/(4d^2), 1\}, \qquad \zeta = \frac{e^{-d}}{8d},$$

and let h, σ, and T_{blk} be given by lemma D.6 applied to $f = \widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ with $D_\Theta = Vd$, $\ell = \ell_{\mathrm{ext}}$, $\rho = \rho_{\mathrm{ext}}$, and the above $A, \epsilon, \delta_c, \zeta$. The algorithm runs $T = dT_{\mathrm{blk}}$ iterations with these fixed h and σ. Since $\sqrt{\rho_{\mathrm{ext}}\epsilon} \le 1/(4B_Y d^2)$, the choice of ϵ guarantees that every ϵ-approximate SOSP X of $\mathcal{F}_{\lambda,\mathcal{S}}$ with $\|X\|_{\mathrm{F}} \le R_w$ satisfies the assumptions of lemma C.3. Thus, at every such point,

$$\mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \left(C_p B_Y + \frac{1}{2}\right)\frac{V}{d^2}. \tag{25}$$

Coupling and one-block estimate. We condition on the event in equation (21). As in Appendix B.3, let $\{\widetilde{X}_t\}$ be the auxiliary trajectory that runs the same recursion on $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ with the same initialization $\widetilde{X}_0 = X_0$ and the same perturbations $\{\xi_t\}$. Since $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ coincides with $\mathcal{F}_{\lambda,\mathcal{S}}$ in the ball $\|X\|_{\mathrm{F}} \le R_w$, we have $\widetilde{X}_t = X_t$ as long as the auxiliary trajectory has not left the ball. We partition the iterations into d blocks of length T_{blk}, where consecutive blocks share their boundary iterate. By lemma D.6, conditional on the history up to the first iterate of a block, if the objective gap $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(\widetilde{X}_t) - \min_X \mathcal{F}_{\lambda,\mathcal{S}}(X)$ at this iterate is at most A, then with conditional probability at least $1 - 2\zeta$, the block is *successful*, i.e., it contains an ϵ-approximate SOSP of $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$, and $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}$ increases by at most $\delta_c \le V/(4d^2)$ between any two iterates $u < v$ in the block.

Base case: the first block. By equation (21) and $\lambda V \le 1$, we have $\mathcal{F}_{\lambda,\mathcal{S}}(X_0) \le \mathcal{H}_{\mathcal{S}} + 2B_Y + 9$. Since $\min_X \mathcal{F}_{\lambda,\mathcal{S}}(X) \ge \mathcal{H}_{\mathcal{S}}$ and $2B_Y + 9 \le A$, the one-block estimate applies to the first block. Suppose that the first block is successful. Then, every auxiliary iterate in the block satisfies $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(\widetilde{X}_t) \le \mathcal{F}_{\lambda,\mathcal{S}}(X_0) + \delta_c \le \mathcal{H}_{\mathcal{S}} + 2B_Y + 10$, which is smaller than $\mathcal{H}_{\mathcal{S}} + A + 4$. By lemma C.1, every auxiliary iterate in the block is inside the ball, and thus $\widetilde{X}_t = X_t$ for all t in the block. In particular, the approximate SOSP visited in the block is an approximate SOSP of $\mathcal{F}_{\lambda,\mathcal{S}}$ in the ball, at which equation (25) holds. Since the objective increases by at most $V/(4d^2)$ from this visit to the last iterate of the block, we obtain

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_{jT_{\mathrm{blk}}}) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \left(C_p B_Y + \frac{3}{4}\right)\frac{V}{d^2} \tag{26}$$

for $j = 1$.

Induction step: the later blocks. Suppose that the first j blocks are successful for some $1 \le j < d$, that the two trajectories coincide up to iteration jT_{blk}, and that equation (26) holds. Since $\|Y\|_{\mathrm{F}}^2 \le B_Y V$, $\lambda V \le 1$, and $V/d^2 \le \lambda V/(2C_p) \le 1/(2C_p)$, equation (26) implies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_{jT_{\mathrm{blk}}}) - \mathcal{H}_{\mathcal{S}} \le \frac{B_Y(\lambda V)}{2} + \left(C_p B_Y + \frac{3}{4}\right)\frac{V}{d^2} \le B_Y + \frac{3}{8C_p} \le A.$$

Thus, the one-block estimate applies to the $(j+1)$-th block. Suppose that this block is successful. Then, every iteration t in the block satisfies

$$\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(\widetilde{X}_t) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \left(C_p B_Y + \frac{3}{4}\right)\frac{V}{d^2} + \delta_c \le (C_p B_Y + 1)\,\frac{V}{d^2}.$$

By the same calculation as above, this implies $\widetilde{\mathcal{F}}_{\lambda,\mathcal{S}}(\widetilde{X}_t) \le \mathcal{H}_{\mathcal{S}} + B_Y + 1/(2C_p) < \mathcal{H}_{\mathcal{S}} + A + 4$. Hence, the auxiliary trajectory stays inside the ball, $\widetilde{X}_t = X_t$ for all t in the block, and the first bound in the lemma holds for all t in the block. Moreover, applying equation (25) at the approximate SOSP visited in this block, together with the bound on the increase of the objective, we conclude that equation (26) holds with j replaced by $j+1$. By induction, if all d blocks are successful, the first bound in the lemma holds for every integer $t \in [T_{\mathrm{blk}}, T]$.

Gain, failure probability, and running time. For the gain, since $\|Y\|_{\mathrm{F}}^2 \le B_Y V$ and the weight decay penalty of X_t is nonnegative, the first bound in the lemma implies

$$\mathcal{R}_{\mathcal{S}}(X_t) - \mathcal{H}_{\mathcal{S}} \le \frac{B_Y(\lambda V)}{2} + \frac{(C_p B_Y + 1)V}{d^2}.$$

Dividing both sides by $\log V - \mathcal{H}_{\mathcal{S}} \ge \delta_{\mathrm{tr}}$ gives the second bound in the lemma. For the probability, conditional on the initialization event and on the success of all previous blocks, each block fails with probability at most 2ζ. Thus, by the union bound, the total failure probability is at most the failure probability of the initialization plus $2d\zeta = e^{-d}/4$, which is at most e^{-cd} after adjusting the constant c. For the running time, note that $\lambda V \ge 2C_p V/d^2 > 2C_p/V$, $d^2/V \le V$, and $\log(1/\zeta) = d + \log(8d)$. Thus, R_w, ℓ_{ext}, ρ_{ext}, ϵ^{-1}, and δ_c^{-1} are polynomial in $(V, N, d, (\lambda V)^{-1})$, and so are h^{-1}, σ^{-1}, and T_{blk} by equation (38). This proves equation (24). □

We are now ready to prove theorem 5.2. We choose $\lambda = 2C_p/d^2$, under which both the weight decay penalty of the ground truth, i.e., $\frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2 \le B_Y(\lambda V)/2$, and the comparison error in lemma C.4 are of order V/d^2. For convenience, we restate the theorem below.

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 $\max\{2r, C\sqrt{V}\} \le d < N$, setting the weight decay $\lambda = C_\lambda/d^2$, there exist a step size h, a perturbation scale σ, and a number of iterations T divisible by d such that, with probability at least $1 - e^{-cd}$,*

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

The quantities h^{-1}, σ^{-1}, and T are bounded by polynomials in V.

Proof. Let $C_\lambda = 2C_p$, i.e., $\lambda = 2C_p/d^2$. We first verify the assumptions of lemma C.4. When the constant C in $d \ge C\sqrt{V}$ is sufficiently large, we have $\lambda V = 2C_p V/d^2 \le 1$, and $d \ge C' \log(eV)$ for all sufficiently large V. In addition, $d \ge 2r$ implies $d \ge 2r_Y$. Thus, lemma C.4 applies with this λ. With probability at least $1 - e^{-cd}$ over the initialization and the perturbations, every integer $t \in [T/d, T]$ satisfies

$$\big(\mathcal{R}_{\mathcal{S}}(X_t) - \mathcal{H}_{\mathcal{S}}\big) \le B_Y(\lambda V)/2 + (C_p B_Y + 1)\frac{V}{d^2} = (2C_p B_Y + 1)\frac{V}{d^2}.$$

Dividing both sides by $\log V - \mathcal{H}_{\mathcal{S}} \ge \delta_{\mathrm{tr}}$ proves the gain bound. Finally, by lemma C.4, h and σ are fixed, $T = dT_{\mathrm{blk}}$ is divisible by d, and h^{-1}, σ^{-1}, and T are bounded by polynomials in V, since $N < V$, $d < V$, and $(\lambda V)^{-1} = d^2/(2C_p V) < V$. □

C.2 Proof of Theorem 5.4

In this section, we present the full proof of theorem 5.4. The proof consists of three main steps.

Step 1 (Appendix C.2.1). We apply lemma C.4 with the weight decay in theorem 5.4, and show that, with high probability, the objective at every iterate in the final window exceeds the objective at the ground truth by at most $O((\lambda V)^2)$.

Step 2 (Appendix C.2.2). We convert this comparison of the objectives into a bound of $O((\lambda V)^2)$ on the training excess loss, and an upper bound on the trace of the learned Gram matrix. The key tool is an identity that relates the trace of the Gram matrix to the logits on the observed rows under the coverage condition (lemma C.6), together with a bound on the error of the logits in terms of the KL divergence (lemma C.7).

Step 3 (Appendix C.2.3). We show that the error of the entire Gram matrix is controlled by the error of the logits on the observed rows and the excess of the trace (lemma C.8). Combining this result with Step 2 and the bound on the learned embeddings, we bound the loss on the unseen states (lemma C.9), and then prove theorem 5.4.

We note that Steps 2 and 3 are deterministic, i.e., they hold for every X that satisfies the comparison in Step 1.

Notation and assumptions. We use the notation in Appendix C.1. In addition, by definition A.3, the representative Y satisfies

$$Y^\top \mathbf{1} = 0, \qquad \sigma_{r_Y}(Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top) \geq \kappa\sqrt{N},$$
$$\bar{y}_{\mathcal{S}} = \frac{1}{N}\sum_{i\in\mathcal{S}} y_i.$$

We collect some direct consequences of these conditions. First, since $Y^\top \mathbf{1} = 0$, the matrix $K^\star = YY^\top$ is row-centered, and thus $\mathcal{L}(\log Q) = K^\star$. Second, since the diagonal entries of $K^\star$ are $\|y_i\|_2^2$ and $\|\mathcal{L}(\log Q)\|_\infty \leq B$, we have $\max_i \|y_i\|_2^2 \leq B$, and hence $|K^\star_{ij}| \leq B$. Third, the coverage condition implies $r_Y \kappa^2 N \leq \|Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top\|_{\mathrm{F}}^2 \leq NB$. Throughout this section, the constants c and C may depend on the parameters of the class, the coverage parameter κ, and the constants in $N = \Theta(V)$, but not on V, N, or d. The dependence on the bound R_0 on the learned embeddings is stated explicitly.

C.2.1 Step 1: Compare the Objective with the Ground Truth in the Final Window

In this step, we specify the weight decay and apply lemma C.4. Recall that in theorem 5.4, the weight decay and the embedding dimension satisfy

$$\lambda = C_\lambda \sqrt{\frac{V}{Nd^3}}, \qquad d \geq C_{\mathrm{wid}} \frac{V}{N^{1/3}}, \qquad 2r_Y \leq d < N < V. \tag{27}$$

Note that this weight decay is different from $\lambda = 2C_p/d^2$ in the proof of theorem 5.2. We choose the constant C_λ sufficiently large, and then the constant C_{wid} sufficiently large, both depending only on the parameters of the class, such that the following three conditions hold:

$$\text{(i) } \lambda \geq \frac{2C_p}{d^2}, \qquad \text{(ii) } \frac{(C_pB_Y+1)V}{d^2} \leq C(\lambda V)^2, \qquad \text{(iii) } 0 < \lambda V \leq c_{\mathrm{decay}} \leq 1,$$

where c_{decay} is a sufficiently small constant that is required by lemma C.7 in Step 2. These conditions can be satisfied because

$$\lambda d^2 = C_\lambda \sqrt{\frac{Vd}{N}} \geq C_\lambda, \qquad \lambda V \leq C_\lambda C_{\mathrm{wid}}^{-3/2},$$
$$\frac{(C_pB_Y+1)V/d^2}{(\lambda V)^2} = \frac{(C_pB_Y+1)Nd}{C_\lambda^2 V^2} \leq \frac{C_pB_Y+1}{C_\lambda^2},$$

where we use $1 \leq d < N < V$ and $d \geq C_{\mathrm{wid}} V/N^{1/3}$.

Conditions (i) and (iii) are the assumptions on λ in lemma C.4, and its remaining assumptions are verified in the proof of the theorem below. Thus, lemma C.4 provides deterministic h, σ, and $T = dT_{\mathrm{blk}}$ such that, with probability at least $1 - e^{-cd}$, every integer $t \in [T/d, T]$ satisfies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \leq \frac{(C_pB_Y+1)V}{d^2} \leq C(\lambda V)^2,$$

where the second inequality is condition (ii). The remaining steps are deterministic, and apply to every such iterate. In particular, the bound on the learned embeddings in theorem 5.4 is not part of this probability event.

C.2.2 Step 2: Bound the Training Loss and the Trace of the Gram Matrix

In Step 1, we obtained an upper bound on $\mathcal{F}_{\lambda,\mathcal{S}}(X) - \big(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\big) = \{\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}\} + \frac{\lambda}{2}\{\operatorname{tr} K - \operatorname{tr} K^\star\}$. This does not immediately bound the excess loss $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}$, because the second term can be negative, i.e., an increase in the loss can be compensated by a decrease in the weight decay penalty. Therefore, we need to show that the trace of the learned Gram matrix K cannot be much smaller than that of $K^\star$, unless the logits on the observed rows have a large error. This is where the coverage condition is used.

Linear relations among the ground-truth embeddings. By the coverage condition, the centered embeddings of the observed states span $\mathbb{R}^{r_Y}$. Hence, the embedding y_i of every state $i \in [V]$, observed or unseen, is a linear combination of the embeddings of the observed states, i.e., $y_i = Y_{\mathcal{S}}^\top a_i$ for some coefficient vector $a_i \in \mathbb{R}^N$, whose coordinates are indexed by $s \in \mathcal{S}$. Specifically, we use the following coefficient vectors:

$$\begin{aligned} a_i = {} & \big[(I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)(\log Q_{\mathcal{S},\mathcal{S}})(I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)\big]^\dagger \\ & \cdot (I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)\mathcal{L}(\log Q_{\mathcal{S},:})_{:,i} \in \mathbb{R}^N. \end{aligned} \tag{28}$$

Here, the logarithm is entrywise, and † denotes the Moore–Penrose inverse. lemma C.6 below shows that these vectors satisfy $y_i = Y_{\mathcal{S}}^\top a_i$, and provides a closed form in terms of Y. We note that equation (28) only involves the rows of Q indexed by $\mathcal{S}$. This is possible because, although the state i may be unseen as a current state, it is still a possible next state, and thus the i-th column of $Q_{\mathcal{S},:}$ is available.

For the learned embeddings, the same linear relations do not hold in general, and we measure their violation by the residuals $x_i - X_{\mathcal{S}}^\top a_i$. The key identity in lemma C.6 states that the sum of the squared residuals equals the excess of the trace, $\operatorname{tr}(K - K^\star)$, minus a linear function of the error of the logits on the observed rows. Since the squared residuals are nonnegative, this identity lower bounds $\operatorname{tr}(K - K^\star)$ by the error of the logits on the observed rows, which is what we need. We refer to this identity as the trace certificate.

In the following, we use

$$Q_\perp = I_V - Y(Y^\top Y)^{-1}Y^\top$$

to denote the orthogonal projection onto the orthogonal complement of the column space of Y, and write $\operatorname{sym}(H) = (H + H^\top)/2$. Note that the results in this step hold for every $X \in \mathbb{R}^{V\times d}$, which is not required to be centered or to be a stationary point.

Lemma C.6 (Linear relations among the ground-truth embeddings and the trace certificate)**.** *The coefficients in equation* (28) *satisfy*

$$\begin{gathered} a_i = (Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top)(Y_{\mathcal{S}}^\top Y_{\mathcal{S}} - N\bar{y}_{\mathcal{S}}\bar{y}_{\mathcal{S}}^\top)^{-1}y_i, \\ a_i^\top Y_{\mathcal{S}} = y_i^\top, \qquad a_i^\top K^\star_{\mathcal{S},:} = K^\star_{i,:}, \\ \|a_i\|_2 \le \frac{\sqrt{B}}{\kappa\sqrt{N}}, \qquad \|a_i\|_\infty \le \frac{2B}{\kappa^2 N}. \end{gathered}$$

In particular, every entry of $K^\star$ is determined by $Q_{\mathcal{S},:}$ through

$$K^\star_{ij} = a_i^\top \mathcal{L}(\log Q_{\mathcal{S},:})_{:,j}.$$

Moreover, there exists 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, for every $X \in \mathbb{R}^{V\times d}$ with $K = XX^\top$,

$$\begin{aligned} & M_{\mathrm{dual}} = I_V - \operatorname{sym}((I_V)_{:,\mathcal{S}}\mathcal{L}(Y_{\mathrm{dual}})), \\ & M_{\mathrm{dual}} \succeq Q_\perp, \qquad M_{\mathrm{dual}}Y = 0, \\ & \sum_i \|x_i - X_{\mathcal{S}}^\top a_i\|_2^2 = \operatorname{tr}(K - K^\star) \\ & \qquad\qquad - \langle Y_{\mathrm{dual}}, \mathcal{L}(K - K^\star)_{\mathcal{S},:}\rangle. \end{aligned} \tag{29}$$

Proof. We first prove the properties of a_i. Since $Y^\top\mathbf{1} = 0$, we have $\mathcal{L}(\log Q) = YY^\top$, and thus

$$\mathcal{L}(\log Q_{\mathcal{S},:}) = Y_{\mathcal{S}}Y^\top.$$

Multiplying it by $I_N - \mathbf{1}_N \mathbf{1}_N^\top / N$ from the left gives $(Y_\mathcal{S} - \mathbf{1}_N \bar{y}_\mathcal{S}^\top) Y^\top$. Similarly, since the entries of $\log Q_{\mathcal{S},\mathcal{S}}$ differ from those of $Y_\mathcal{S} Y_\mathcal{S}^\top$ only by a constant in each row, centering $\log Q_{\mathcal{S},\mathcal{S}}$ on both sides gives

$$(Y_\mathcal{S} - \mathbf{1}_N \bar{y}_\mathcal{S}^\top)(Y_\mathcal{S} - \mathbf{1}_N \bar{y}_\mathcal{S}^\top)^\top.$$

Since $Y_\mathcal{S} - \mathbf{1}_N \bar{y}_\mathcal{S}^\top$ has full column rank by the coverage condition, the identity $(AA^\top)^\dagger A = A(A^\top A)^{-1}$ for a matrix A with full column rank, applied to equation (28), proves the closed form of a_i. Since

$$(Y_\mathcal{S} - \mathbf{1}_N \bar{y}_\mathcal{S}^\top)^\top Y_\mathcal{S} = Y_\mathcal{S}^\top Y_\mathcal{S} - N \bar{y}_\mathcal{S} \bar{y}_\mathcal{S}^\top,$$

the closed form gives $a_i^\top Y_\mathcal{S} = y_i^\top$. Multiplying this identity by $Y^\top$ from the right gives $a_i^\top K^\star_{\mathcal{S},:} = K^\star_{i,:}$ and the formula for $K^\star_{ij}$. For the norms, the coverage condition gives $\|a_i\|_2 \le \|y_i\|_2/(\kappa\sqrt{N}) \le \sqrt{B}/(\kappa\sqrt{N})$, and the bound $\|y_s - \bar{y}_\mathcal{S}\|_2 \le 2\sqrt{B}$ gives the bound on $\|a_i\|_\infty$. In addition, the closed form implies $a_i^\top \mathbf{1}_N = 0$ and $\sum_{i=1}^V a_i = 0$, which will be used below.

Next, we construct Y_{dual}. For $s \in \mathcal{S}$ and $j \in [V]$, define

$$(Y_{\mathrm{dual}})_{s,j} = \begin{cases} 2(a_j)_s - \displaystyle\sum_{i=1}^V (a_i)_s (a_i)_j, & j \in \mathcal{S}, \\ 2(a_j)_s, & j \notin \mathcal{S}. \end{cases}$$

By $a_i^\top \mathbf{1}_N = 0$ and $\sum_i a_i = 0$, every row of Y_{dual} sums to zero, and thus $\mathcal{L}(Y_{\mathrm{dual}}) = Y_{\mathrm{dual}}$. Expanding the square, for every $v \in \mathbb{R}^V$ we have

$$v^\top M_{\mathrm{dual}} v = \sum_{i=1}^V \left(v_i - \sum_{s \in \mathcal{S}} (a_i)_s v_s \right)^2.$$

We now verify the properties of M_{dual}. First, when v is a column of Y, the identity $a_i^\top Y_\mathcal{S} = y_i^\top$ shows that every summand is zero. Since the above identity also shows $M_{\mathrm{dual}} \succeq 0$, this implies $M_{\mathrm{dual}} Y = 0$. Second, by the closed form of a_i, the vector with entries $\sum_{s \in \mathcal{S}} (a_i)_s v_s$ lies in the column space of Y. Thus, for v orthogonal to the column space of Y, the cross term in the above sum of squares vanishes, which gives $v^\top M_{\mathrm{dual}} v \ge \|v\|_2^2$. Together with $M_{\mathrm{dual}} Y = 0$, this proves $M_{\mathrm{dual}} \succeq Q_\perp$. Third, applying the above identity to each column of X and summing over the columns gives $\langle M_{\mathrm{dual}}, K \rangle = \sum_i \|x_i - X_\mathcal{S}^\top a_i\|_2^2$. On the other hand, by the definition of M_{dual}, the symmetry of K, and $\langle M_{\mathrm{dual}}, K^\star \rangle = 0$, we have $\langle M_{\mathrm{dual}}, K \rangle = \mathrm{tr}(K - K^\star) - \langle Y_{\mathrm{dual}}, (K - K^\star)_{\mathcal{S},:} \rangle$. Since the rows of Y_{dual} sum to zero, $(K - K^\star)_{\mathcal{S},:}$ can be replaced by $\mathcal{L}(K - K^\star)_{\mathcal{S},:}$ in the last inner product. This proves the last identity in equation (29). Finally, the bound on $\|a_i\|_\infty$ gives

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

Multiplying this bound by $\sqrt{NV}$ gives $\|Y_{\mathrm{dual}}\|_{\mathrm{F}} \le C$. □

The trace certificate in equation (29) is stated in terms of the error of the logits on the observed rows, while the comparison in Step 1 is stated in terms of the loss, i.e., the KL divergence. The following lemma connects the two. Its first part bounds the ℓ^1 error of the centered logits by the KL divergence, which holds without any bound on the learned logits. Its second part combines this bound with the trace certificate and the comparison in Step 1. Among the four conclusions of the second part, the bounds on the training excess loss and on the trace are the ones used in Step 3.

Lemma C.7 (From the objective comparison to bounds on the training loss and the trace)**.** *For every $X \in \mathbb{R}^{V \times d}$, without any bound on its logits,*

$$\begin{aligned} \frac{\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_1}{NV} &\le 2e^{2B} \big(\mathcal{R}_\mathcal{S}(X) - \mathcal{H}_\mathcal{S} \big) \\ &\quad + 2e^{2B} \sqrt{2 \big(\mathcal{R}_\mathcal{S}(X) - \mathcal{H}_\mathcal{S} \big)}. \end{aligned} \tag{30}$$

Moreover, suppose that $\mathcal{F}_{\lambda,\mathcal{S}}(X) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \le \nu$ *for some* $\nu \ge 0$*, and that* $0 < \lambda V \le c_{\mathrm{decay}}$ *for a sufficiently small constant* c_{decay} *depending only on the parameters of the class. Then,*

$$\begin{aligned}
\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} &\le C(\nu + (\lambda V)^2),\\
(\operatorname{tr} K - \operatorname{tr} K^\star)_+ &\le 2\nu/\lambda,\\
\|(I_V - \mathbf{1}\mathbf{1}^\top/V)X\|_{\mathrm{F}}^2/V &\le B + 2\nu/(\lambda V),\\
\sum_i \|x_i - X_{\mathcal{S}}^\top a_i\|^2 &\le CV((\lambda V) + \sqrt{\nu} + \nu/(\lambda V)).
\end{aligned}$$

Proof. We first prove equation (30). Fix an observed row $i \in \mathcal{S}$. Using the inequality $\log z \le z - 1$, we have

$$\sum_j Q_{ij}\left(\log\frac{P_{ij}^X}{Q_{ij}}\right)_+ \le \sum_j (P_{ij}^X - Q_{ij})_+ \le \sqrt{\frac{\mathrm{KL}(q_i\|P_{i,:}^X)}{2}},$$

where the last inequality is Pinsker's inequality. Since $\sum_j Q_{ij}\log(P_{ij}^X/Q_{ij}) = -\mathrm{KL}(q_i\|P_{i,:}^X)$ and $|a| = 2a_+ - a$ for any real number a, we obtain

$$\sum_j Q_{ij}\left|\log\frac{P_{ij}^X}{Q_{ij}}\right| \le \mathrm{KL}(q_i\|P_{i,:}^X) + \sqrt{2\,\mathrm{KL}(q_i\|P_{i,:}^X)}.$$

Recall that the centered ground-truth logits lie in $[-B, B]$, which implies $Q_{ij} \ge e^{-2B}/V$. In addition, since $\log P_{i,:}^X$ and $\log Q_{i,:}$ differ from $K_{i,:}$ and $K_{i,:}^\star$ by constants, respectively, the error of the centered logits is the centered log-ratio of the probabilities:

$$\begin{aligned}
&\mathcal{L}(K - K^\star)_{i,:}\\
&\qquad = \mathcal{L}(\log P_{i,:}^X - \log Q_{i,:}).
\end{aligned}$$

Since centering increases the ℓ^1 norm by at most a factor of two, we have

$$\begin{aligned}
\|\mathcal{L}(K - K^\star)_{i,:}\|_1 &\le 2\sum_j \left|\log\frac{P_{ij}^X}{Q_{ij}}\right|\\
&\le 2Ve^{2B}\left\{\mathrm{KL}(q_i\|P_{i,:}^X) + \sqrt{2\,\mathrm{KL}(q_i\|P_{i,:}^X)}\right\}.
\end{aligned}$$

Averaging over $i \in \mathcal{S}$ and applying Jensen's inequality to the square root proves equation (30).

Next, we prove the second part. Since the left-hand side of the last identity in equation (29) is nonnegative, and $\|Y_{\mathrm{dual}}\|_\infty \le C/V$ with $N = \Theta(V)$, we have

$$\begin{aligned}
\operatorname{tr}(K - K^\star) &\ge \langle Y_{\mathrm{dual}}, \mathcal{L}(K - K^\star)_{\mathcal{S},:}\rangle\\
&\ge -\frac{C}{N}\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_1.
\end{aligned}$$

Note that $\|X\|_{\mathrm{F}}^2 = \operatorname{tr} K$ and $\|Y\|_{\mathrm{F}}^2 = \operatorname{tr} K^\star$, so the assumption on the objective is equivalent to $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} + \lambda\operatorname{tr}(K - K^\star)/2 \le \nu$. Combining it with the above bound and equation (30), we have

$$\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le \nu + C\lambda V\left\{\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} + \sqrt{\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}}}\right\}.$$

We choose c_{decay} such that the coefficient $C\lambda V$ is at most $1/2$, and apply Young's inequality to the term with the square root. This gives $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le C(\nu + (\lambda V)^2)$, which is the first bound. For the second bound, since $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \ge 0$, the assumption on the objective directly gives $\lambda\operatorname{tr}(K - K^\star)/2 \le \nu$. The third bound follows from $\|(I_V - \mathbf{1}\mathbf{1}^\top/V)X\|_{\mathrm{F}}^2 \le \operatorname{tr} K$, the second bound, and $\operatorname{tr} K^\star \le BV$. Finally, for the fourth bound, equation (29) gives

$$\begin{aligned}
\sum_i \|x_i - X_{\mathcal{S}}^\top a_i\|^2 &\le 2\nu/\lambda + \frac{C}{N}\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_1\\
&\le CV\{\nu/(\lambda V) + \nu + (\lambda V)^2 + \sqrt{\nu} + \lambda V\}.
\end{aligned}$$

Since $\lambda V \le 1$, the terms ν and $(\lambda V)^2$ are dominated by $\nu/(\lambda V)$ and λV, respectively, which completes the proof. □

In summary, applying lemma C.7 with $\nu \le C(\lambda V)^2$ from Step 1, every iterate $X = X_t$ in the final window satisfies

$$\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le C(\lambda V)^2, \qquad (\operatorname{tr} K - \operatorname{tr} K^\star)_+ \le C(\lambda V)V.$$

These are the two bounds used in Step 3. In addition, lemma C.7 also gives $\frac{\|\mathcal{L}(K-K^\star)_{\mathcal{S},:}\|_1}{NV} \le C(\lambda V)$, $\|(I_V - \mathbf{1}\mathbf{1}^\top/V)X\|_{\mathrm{F}}^2/V \le C$, and $\sum_i \|x_i - X_{\mathcal{S}}^\top a_i\|^2 \le C(\lambda V)V$, which are not needed in the proof of theorem 5.4.

C.2.3 Step 3: Transfer the Bounds to Unseen States

In this step, we show that the bounds on the observed rows in Step 2 control the entire Gram matrix, and thus the loss on the unseen states. The difficulty is that the rows of $K - K^\star$ indexed by $\mathcal{S}$ do not determine the whole matrix in general. To handle this, we decompose the error as

$$K - K^\star = H + D_0, \qquad D_0 = Q_\perp K Q_\perp,$$

where $Q_\perp$ is the orthogonal projection onto the orthogonal complement of the column space of Y. The first part H satisfies $Q_\perp H Q_\perp = 0$, i.e., it has no component that lies outside the column space of Y on both sides. We show that such a matrix is controlled by its centered rows indexed by $\mathcal{S}$. The second part D_0 is positive semidefinite because K is, so its Frobenius norm is at most its trace, and its trace is controlled by the trace certificate. The following lemma makes this argument precise.

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

$$\|H\|_{\mathrm{F}} \le C\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}. \tag{31}$$

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

$$\|K - K^\star\|_{\mathrm{F}} \le C\{\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}} + \tau\}.$$

Both statements hold regardless of the rank of K.

Proof. We first prove equation (31). Let H be a symmetric matrix with $Q_\perp H Q_\perp = 0$.

Representation of H. Such a matrix can be written as $H = YV_0^\top + V_0 Y^\top$ for some $V_0 \in \mathbb{R}^{V \times r_Y}$. For example, one can take

$$V_0 = HY(Y^\top Y)^{-1} - \tfrac{1}{2}Y(Y^\top Y)^{-1}(Y^\top H Y)(Y^\top Y)^{-1}.$$

Indeed, substituting this V_0 gives $YV_0^\top + V_0 Y^\top = H - Q_\perp H Q_\perp$, which equals H by assumption.

Decomposition of the observed rows. Since $Y^\top \mathbf{1} = 0$, we have $\mathcal{L}(H) = Y\big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)^\top + V_0 Y^\top$. Thus, after centering the rows indexed by $\mathcal{S}$, i.e., multiplying by $I_N - \mathbf{1}_N\mathbf{1}_N^\top/N$ from the left, the observed rows of $\mathcal{L}(H)$ become

$$(Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)\big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)^\top + (I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)\big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)_{\mathcal{S},:} Y^\top.$$

We decompose the $N \times r_Y$ matrix in the second term into its projection onto the column space of the centered observed embeddings $Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top$ and the orthogonal residual. Specifically, define

$$\begin{aligned}
\Omega &= (Y_{\mathcal{S}}^\top Y_{\mathcal{S}} - N\bar{y}_{\mathcal{S}}\bar{y}_{\mathcal{S}}^\top)^{-1}(Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top)^\top \\
&\qquad \cdot \big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)_{\mathcal{S},:}, \\
W_\perp &= (I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)\big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)_{\mathcal{S},:} - (Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top)\Omega, \\
V_1 &= (I_V - \mathbf{1}\mathbf{1}^\top/V)V_0 + Y\Omega^\top.
\end{aligned}$$

Here, $\Omega \in \mathbb{R}^{r_Y \times r_Y}$ is the coefficient matrix of the projection, $W_\perp \in \mathbb{R}^{N \times r_Y}$ is the residual, which satisfies $(Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top)^\top W_\perp = 0$, and $V_1 \in \mathbb{R}^{V \times r_Y}$ is obtained from $(I_V - \mathbf{1}\mathbf{1}^\top/V)V_0$ by adding the component $Y\Omega^\top$, which lies in the column space of Y. With these definitions, we have

$$\begin{aligned}
&(I_N - \mathbf{1}_N\mathbf{1}_N^\top/N)\mathcal{L}(H)_{\mathcal{S},:} \\
&\qquad = (Y_{\mathcal{S}} - \mathbf{1}_N\bar{y}_{\mathcal{S}}^\top)V_1^\top + W_\perp Y^\top
\end{aligned}$$

and the two terms on the right-hand side are orthogonal in the Frobenius inner product, because $(Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)^\top W_\perp = 0$. Therefore, each of them has Frobenius norm at most $\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}$. Since the smallest singular value of $Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top$ is at least $\kappa\sqrt{N}$ by the coverage condition, and $\sigma_{r_Y}(Y) \geq \sigma_{r_Y}(Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)$, we obtain

$$\|V_1\|_{\mathrm{F}} + \|W_\perp\|_{\mathrm{F}} \leq \frac{C}{\sqrt{N}} \|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}.$$

Bound on $\Omega + \Omega^\top$. By the definitions of $W_\perp$ and V_1,

$$\begin{aligned}(I_N - \mathbf{1}_N \mathbf{1}_N^\top/N)((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0)_{\mathcal{S},:} &= (Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)\Omega + W_\perp \\ &= (I_N - \mathbf{1}_N \mathbf{1}_N^\top/N)(V_1)_{\mathcal{S},:} \\ &\quad - (Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)\Omega^\top,\end{aligned}$$

which implies $(Y_{\mathcal{S}} - \mathbf{1}_N \bar{y}_{\mathcal{S}}^\top)(\Omega + \Omega^\top) = (I_N - \mathbf{1}_N \mathbf{1}_N^\top/N)(V_1)_{\mathcal{S},:} - W_\perp$. Using the coverage condition and the above bound on $\|V_1\|_{\mathrm{F}} + \|W_\perp\|_{\mathrm{F}}$, we obtain $\|\Omega + \Omega^\top\|_{\mathrm{F}} \leq C\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}/N$.

Bound on H. Substituting $(I_V - \mathbf{1}\mathbf{1}^\top/V)V_0 = V_1 - Y\Omega^\top$ into $(I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H) = Y\big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)^\top + \big((I_V - \mathbf{1}\mathbf{1}^\top/V)V_0\big)Y^\top$, we have

$$(I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H) = YV_1^\top + V_1 Y^\top - Y(\Omega + \Omega^\top)Y^\top,$$

which has Frobenius norm at most $C\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}$, since $\|Y\|_{\mathrm{op}} \leq \sqrt{BV}$ and $N = \Theta(V)$. It remains to control the part of H that is removed by the centering. Since $Y^\top \mathbf{1} = 0$, we have $\mathbf{1}^\top H \mathbf{1} = 0$, and thus

$$\begin{aligned}H &= (I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H) + \frac{\mathbf{1}(H\mathbf{1})^\top + (H\mathbf{1})\mathbf{1}^\top}{V}, \\ \mathcal{L}(H)_{\mathcal{S},:} &= ((I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H))_{\mathcal{S},:} + \frac{\mathbf{1}_N(H\mathbf{1})^\top}{V}.\end{aligned}$$

The second identity gives

$$\begin{aligned}\frac{\sqrt{N}}{V}\|H\mathbf{1}\|_2 &\leq \|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}} \\ &\quad + \|(I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H)\|_{\mathrm{F}} \\ &\leq C\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}}.\end{aligned}$$

Combining this bound with the first identity and the bound on $(I_V - \mathbf{1}\mathbf{1}^\top/V)\mathcal{L}(H)$ proves equation (31).

Bound on $K - K^\star$. Next, we prove the second statement. Let $D_0 = Q_\perp K Q_\perp \succeq 0$. Since $K \succeq 0$ and $M_{\mathrm{dual}} \succeq Q_\perp$ by lemma C.6, we have $\operatorname{tr} D_0 = \langle Q_\perp, K\rangle \leq \langle M_{\mathrm{dual}}, K\rangle$. Thus, the trace certificate in equation (29) gives

$$\begin{aligned}\operatorname{tr} D_0 &\leq \langle M_{\mathrm{dual}}, K\rangle \\ &= \operatorname{tr}(K - K^\star) \\ &\quad - \langle Y_{\mathrm{dual}}, \mathcal{L}(K - K^\star)_{\mathcal{S},:}\rangle \\ &\leq \tau + C\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}},\end{aligned}$$

where the last inequality follows from $\operatorname{tr} K \leq \operatorname{tr} K^\star + \tau$ and $\|Y_{\mathrm{dual}}\|_{\mathrm{F}} \leq C$. Let $H = (K - K^\star) - D_0$. Since $Q_\perp K^\star Q_\perp = 0$, we have $Q_\perp H Q_\perp = 0$, and thus equation (31) applies to H. Moreover, since selecting rows and applying the centering operator $\mathcal{L}$ do not increase the Frobenius norm, and $\|D_0\|_{\mathrm{F}} \leq \operatorname{tr} D_0$ for the positive semidefinite matrix D_0, we have $\|\mathcal{L}(H)_{\mathcal{S},:}\|_{\mathrm{F}} \leq \|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}} + \operatorname{tr} D_0$. Therefore,

$$\begin{aligned}\|K - K^\star\|_{\mathrm{F}} &\leq C\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}} + (C+1)\operatorname{tr} D_0 \\ &\leq C\{\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}} + \tau\}.\end{aligned}$$

This completes the proof. □

It remains to relate the error of the Gram matrix to the prediction loss. When the learned embeddings have bounded norms, both the learned logits and the ground-truth logits are bounded. Thus, on the line segment between them, every probability is between constant multiples of $1/V$, and the Hessian of the cross-entropy loss is bounded from below and from above by constant multiples of $1/V$ on centered vectors. We use the lower bound to convert the bound on the training loss in Step 2 into a bound on the error of the logits on the observed rows, and the upper bound to convert the error of the entire Gram matrix into a bound on the loss on the unseen rows. The following lemma summarizes the results of Steps 2 and 3. It is deterministic, i.e., it holds for every X that satisfies the comparison in Step 1.

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 choose $\lambda > 0$ with λV sufficiently small. Then, every $X \in \mathbb{R}^{V\times d}$ satisfying*

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

satisfies $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le C(\lambda V)^2$. If in addition $\max_j \|x_j\|^2 \le R_0$, then

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

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

Proof. **Training loss.** We apply lemma C.7 with $\nu = C_0(\lambda V)^2$. When λV is sufficiently small, it gives $\mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \le C(\lambda V)^2$, $(\operatorname{tr} K - \operatorname{tr} K^\star)_+ \le 2C_0\lambda V^2$, and $\|(I_V - \mathbf{1}\mathbf{1}^\top/V)X\|_{\mathrm{F}}^2/V \le C$. In particular, the bound on the training excess loss does not require the bound on the learned embeddings.

Error on the observed rows. Suppose in addition that $\max_j \|x_j\|^2 \le R_0$. Then, the learned logits $K_{ij} = x_i^\top x_j$ lie in $[-R_0, R_0]$, and the ground-truth logits $K^\star_{ij}$ lie in $[-B, B]$. Hence, for every row i and every logit vector on the line segment between $K^\star_{i,:}$ and $K_{i,:}$, the corresponding softmax probabilities are between c_{R_0}/V and C_{R_0}/V. For such a probability vector p and a centered vector v, the variance satisfies

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

Note that the gradient of $\mathrm{CE}(q_i, \cdot)$ vanishes at $K^\star_{i,:}$, and its Hessian is the covariance matrix of the corresponding softmax distribution. Thus, by Taylor's theorem with the integral remainder, the lower bound on the variance, and the bound on the training excess loss, we have

$$\begin{aligned}\frac{c_{R_0}}{2NV}\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}}^2 &\le \mathcal{R}_{\mathcal{S}}(X) - \mathcal{H}_{\mathcal{S}} \\ &\le C(\lambda V)^2.\end{aligned}$$

Consequently

$$\|\mathcal{L}(K - K^\star)_{\mathcal{S},:}\|_{\mathrm{F}} \le C_{R_0}\lambda V\sqrt{NV} \le C_{R_0}\lambda V^2.$$

Error of the entire Gram matrix. By the bound on the trace above, the assumption of lemma C.8 holds with $\tau = 2C_0\lambda V^2$. Thus, lemma C.8 gives $\|K - K^\star\|_{\mathrm{F}} \le C_{R_0}\lambda V^2$, which is the bound on $\|XX^\top - YY^\top\|_{\mathrm{F}}$ in the lemma.

Loss on the unseen rows. Finally, for any nonempty $\mathcal{U}_0 \subseteq \mathcal{U}$, applying Taylor's theorem again to the rows in $\mathcal{U}_0$, now with the upper bound on the variance, we obtain

$$\begin{aligned}\mathcal{R}_{\mathcal{U}_0}(X) - \mathcal{H}_{\mathcal{U}_0} &\le \frac{C_{R_0}}{|\mathcal{U}_0|V}\|\mathcal{L}(K - K^\star)_{\mathcal{U}_0,:}\|_{\mathrm{F}}^2 \\ &\le C_{R_0}(\lambda V)^2 V/|\mathcal{U}_0|.\end{aligned}$$

Since the bound on $\|K - K^\star\|_{\mathrm{F}}$ does not depend on $\mathcal{U}_0$, the result holds for all nonempty $\mathcal{U}_0 \subseteq \mathcal{U}$ simultaneously.

□

We are now ready to prove theorem 5.4, which we restate below for convenience.

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 $\log V - \mathcal{H}_{\mathcal{S}} \geq \delta_{\mathrm{tr}} > 0$ 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 $\lambda = C_\lambda\sqrt{V/(Nd^3)}$. Then, there exist a step size h, a perturbation scale σ, and a number of iterations T, with h^{-1}, σ^{-1}, and T bounded by polynomials in V, such that, with probability at least $1 - e^{-cd}$, 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

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

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

Proof. We choose the constants C_λ and C_{wid} in equation (27) as in Step 1, so that $\lambda \geq 2C_p/d^2$ and $\lambda V \leq C_\lambda C_{\mathrm{wid}}^{-3/2} \leq c_{\mathrm{decay}}$. For sufficiently large V, the condition $d \geq C_{\mathrm{wid}}V/N^{1/3}$ implies $d \geq C'\log(eV)$, and $d \geq 2r$ implies $d \geq 2r_Y$. Thus, all assumptions of lemma C.4 hold. By Step 1, there exist deterministic h, σ, and T, which together with h^{-1} and σ^{-1} are bounded by polynomials in V since $N, d < V$ and $(\lambda V)^{-1} = C_\lambda^{-1}\sqrt{Nd^3/V^3} < V$, and an event of probability at least $1 - e^{-cd}$, on which every integer $t \in [T/d, T]$ satisfies

$$\mathcal{F}_{\lambda,\mathcal{S}}(X_t) - \left(\mathcal{H}_{\mathcal{S}} + \frac{\lambda}{2}\|Y\|_{\mathrm{F}}^2\right) \leq C(\lambda V)^2.$$

Fix any integer $t \in [T/d, T]$ on this event. By lemma C.9, we have $\mathcal{R}_{\mathcal{S}}(X_t) - \mathcal{H}_{\mathcal{S}} \leq C(\lambda V)^2$. If in addition $\max_{j\in[V]} \|x_j(t)\|_2^2 \leq R_0$, then lemma C.9 with $\mathcal{U}_0 = \mathcal{U}$ gives $\mathcal{R}_{\mathcal{U}}(X_t) - \mathcal{H}_{\mathcal{U}} \leq C(\lambda V)^2 V/(V-N)$. Substituting $(\lambda V)^2 = C_\lambda^2 V^3/(Nd^3)$, and dividing the two bounds by $\log V - \mathcal{H}_{\mathcal{S}} \geq \delta_{\mathrm{tr}}$ and $\log V - \mathcal{H}_{\mathcal{U}} \geq \delta_{\mathrm{te}}$, respectively, proves the two gain bounds. Note that the second bound is a deterministic implication at each iterate on the event, and the probability that the bound on the learned embeddings holds is not used. □

D Auxiliary Results for the Main Proofs

In this section, we collect the auxiliary results that are used in the proofs of the main theorems. Appendix D.1 shows that a low-rank matrix with bounded entries admits a factorization with bounded row norms, which is used to construct the comparator in lemma B.2. Appendix D.2 provides the gradient and the Hessian of the untied objective, which are used in Step 2 of Appendix B. Appendix D.3 constructs the smooth extension used in Step 1, and Appendix D.4 states the results on perturbed gradient descent used in Step 3.

D.1 Bounded Factorizations

The following lemma shows that a matrix with rank at most r and entries bounded by B admits a factorization whose rows have squared norms at most $B\sqrt{r}$. In particular, the total squared norm of the two factors is linear in V. This lemma is applied in lemma B.2 to construct the comparator. Note that, in this lemma only, X and Y denote generic factors of the matrix M.

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

$$M = XY^\top, \qquad \max_i \|X_{i,:}\|_2^2 \leq B\sqrt{r}, \qquad \max_j \|Y_{j,:}\|_2^2 \leq B\sqrt{r}. \tag{32}$$

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

$$\|X\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2 \leq 2B\sqrt{r}V. \tag{33}$$

Proof. If $r_0 = 0$, the claim holds trivially with empty factors. In the following, we assume $r_0 \geq 1$. Let $M = A_0 B_0^\top$ be a full-rank factorization with $A_0, B_0 \in \mathbb{R}^{V \times r_0}$, and let a_i and b_j denote the rows of A_0 and B_0, respectively. Let $\mathcal{K}$ be the convex hull of $\{\pm b_j\}_j$, which is a centrally symmetric convex body with nonempty interior, since B_0 has full column rank. We use the Löwner–John theorem for symmetric convex bodies (John, 2014): if the minimum-volume ellipsoid containing a centrally symmetric convex body $K \subset \mathbb{R}^n$ is the Euclidean unit ball $\mathcal{B}_2$, then $n^{-1/2}\mathcal{B}_2 \subseteq K \subseteq \mathcal{B}_2$. By an invertible linear change of coordinates applied to the rows of B_0, we can assume that the minimum-volume ellipsoid containing $\mathcal{K}$ is the unit ball in $\mathbb{R}^{r_0}$, so that

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

We apply the inverse transpose of this change of coordinates to the rows of A_0, so that all inner products $a_i^\top b_j = M_{ij}$ are preserved. Then, we have $\|b_j\|_2 \leq 1$ and $|a_i^\top b_j| \leq B$ for all i, j. Let $\mathcal{K}^\circ = \{z : \sup_{b \in \mathcal{K}} z^\top b \leq 1\}$ be the polar set of $\mathcal{K}$. The second bound implies $a_i/B \in \mathcal{K}^\circ$, and the inclusion $r_0^{-1/2}\mathcal{B}_2 \subseteq \mathcal{K}$ implies $\mathcal{K}^\circ \subseteq \sqrt{r_0}\mathcal{B}_2$. Thus, $\|a_i\|_2 \leq B\sqrt{r_0}$. Finally, we rescale the two factors by reciprocal constants such that the maximum row norms of both factors are at most $\sqrt{B\sqrt{r_0}} \leq \sqrt{B\sqrt{r}}$, and denote the resulting factors by X and Y. This proves equation (32).

For the remaining claims, note that X has full column rank. Thus, if $M\mathbf{1} = 0$, then $X(Y^\top \mathbf{1}) = 0$ implies $Y^\top \mathbf{1} = 0$. Summing the bounds on the squared row norms over the V rows of each factor gives equation (33). □

D.2 Untied Factor Derivatives

The following lemma provides the gradient and the Hessian of the untied objective, as well as some properties of its critical points. The Hessian identity in equation (35) is the starting point of Step 2 in Appendix B.2. Note that, although the cross-entropy loss is convex in the logits, the objective is not convex in (E, U), because the bilinear term in equation (35) can be negative.

Lemma D.2 (Gradient, balance, and Hessian identities)**.** *Fix $\lambda > 0$, and let $\mathcal{F}_\lambda = \mathcal{F}_{\lambda,[V]}$ be the untied objective with $\Theta = (E, U)$. Its gradient is given by*

$$\nabla_E \mathcal{F}_\lambda = \frac{1}{V}(P^{E,U} - Q)U + \lambda E, \qquad \nabla_U \mathcal{F}_\lambda = \frac{1}{V}(P^{E,U} - Q)^\top E + \lambda U. \tag{34}$$

For any direction $(\dot{E}, \dot{U})$ with the same dimensions as (E, U), the Hessian satisfies

$$\begin{aligned}\nabla^2 \mathcal{F}_\lambda(E, U)[(\dot{E}, \dot{U}), (\dot{E}, \dot{U})] = \frac{1}{V}\sum_{i=1}^{V} \mathrm{Var}_{j \sim P^{E,U}_{i,:}}\left[(\dot{E}U^\top + E\dot{U}^\top)_{ij}\right] \\ + \frac{2}{V}\langle P^{E,U} - Q, \dot{E}\dot{U}^\top\rangle + \lambda(\|\dot{E}\|_{\mathrm{F}}^2 + \|\dot{U}\|_{\mathrm{F}}^2).\end{aligned} \tag{35}$$

Here, the variance is over the index j drawn from the i-th row of $P^{E,U}$. We refer to the three terms on the right-hand side as the softmax curvature term, the bilinear term, and the weight decay term, respectively. The softmax curvature term is nonnegative. Moreover, at every critical point,

$$\begin{aligned}\mathbf{1}^\top U = 0, \qquad E^\top E = U^\top U, \\ \|EU^\top\|_* = \frac{1}{2}(\|E\|_{\mathrm{F}}^2 + \|U\|_{\mathrm{F}}^2), \\ \frac{1}{V}\langle P^{E,U} - Q, EU^\top\rangle + \lambda\|EU^\top\|_* = 0.\end{aligned}$$

Proof. The gradient of the loss with respect to the logits $EU^\top$ is $(P^{E,U} - Q)/V$, and thus the chain rule gives equation (34). For the Hessian, the first and second variations of $EU^\top$ along $(\dot{E}, \dot{U})$ are $\dot{E}U^\top + E\dot{U}^\top$ and $2\dot{E}\dot{U}^\top$, respectively, and the Hessian of the cross-entropy loss of row i with respect to its logits is the covariance matrix $\mathrm{diag}(P^{E,U}_{i,:}) - (P^{E,U}_{i,:})^\top P^{E,U}_{i,:}$. Combining these facts with the chain rule gives equation (35).

Next, consider a critical point. Multiplying the second equation in equation (34) by $\mathbf{1}^\top$ from the left, and using $(P^{E,U} - Q)\mathbf{1} = 0$ and $\lambda > 0$, we obtain $\mathbf{1}^\top U = 0$. Multiplying the first equation in

equation (34) by $E^\top$ from the left and taking the transpose, and comparing the result with the second equation multiplied by $U^\top$ from the left, we obtain $E^\top E = U^\top U$. Since the two Gram matrices coincide, their eigenvalues are the singular values of $EU^\top$, which gives the identity for the nuclear norm. Finally, taking the inner product of equation (34) with (E, U) and dividing by two gives the last identity. □

D.3 Smooth Localization

The results on perturbed gradient descent in Appendix D.4 require the gradient and the Hessian of the objective to be globally Lipschitz, while our objectives are only controlled on a ball. The following lemma constructs an extension that coincides with the objective in the ball, remains large outside the ball, and is globally smooth. We note that the extension is only used in the analysis, and the algorithm is run on the original objective.

Lemma D.3 (Smooth extension outside a ball)**.** *Suppose that the ball $\{\Theta : \|\Theta\|_{\mathrm{F}}^2 \le CV\}$ contains Θ_0 and a global minimizer of $\mathcal{F}_\lambda$, and that*

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

Then, for any constant $c > 0$ independent of V, there exists a smooth function $\widetilde{\mathcal{F}}_\lambda : \mathbb{R}^{D_\Theta} \to \mathbb{R}$ such that (i) $\widetilde{\mathcal{F}}_\lambda = \mathcal{F}_\lambda$ on the ball; (ii) $\widetilde{\mathcal{F}}_\lambda \ge \mathcal{F}_\lambda(\Theta_0) + 2$ outside the ball; (iii) $\widetilde{\mathcal{F}}_\lambda \ge \mathcal{F}_\lambda^*$ everywhere; and (iv) the gradient and the Hessian of $\widetilde{\mathcal{F}}_\lambda$ are globally Lipschitz. Moreover, if the derivatives of $\mathcal{F}_\lambda$ up to order three on the ball of radius $(1+c)\sqrt{CV}$ are polynomial in V, then the Lipschitz constants in (iv) are polynomial in V. Finally, if $\mathcal{F}_\lambda \ge M$ outside the ball for some constant M, then the extension can be chosen such that $\widetilde{\mathcal{F}}_\lambda \ge M$ outside the ball as well.

Proof. Let $R = \sqrt{CV}$. Let $s : [0, \infty) \to [0, \infty)$ be a C^∞ function such that $s(a) = a$ for $a \in [0, R]$, $s(a) \in [R, (1+c)R]$ for $a \ge R$, and s is constant for $a \ge (1+2c)R$. Such a function exists by a standard smoothing argument. Define

$$\widetilde{\mathcal{F}}_\lambda(\Theta) = \begin{cases} \mathcal{F}_\lambda(s(\|\Theta\|_{\mathrm{F}})\Theta/\|\Theta\|_{\mathrm{F}}), & \Theta \ne 0, \\ \mathcal{F}_\lambda(0), & \Theta = 0. \end{cases}$$

The map $\Theta \mapsto s(\|\Theta\|_{\mathrm{F}})\Theta/\|\Theta\|_{\mathrm{F}}$ is the identity on the ball $\|\Theta\|_{\mathrm{F}} \le R$, and maps every point outside the ball to the annulus $R \le \|\Theta\|_{\mathrm{F}} \le (1+c)R$. It is smooth at the origin because it is the identity in a neighborhood of the origin. Thus, $\widetilde{\mathcal{F}}_\lambda = \mathcal{F}_\lambda$ on the ball. Outside the ball, $\widetilde{\mathcal{F}}_\lambda(\Theta)$ is the value of $\mathcal{F}_\lambda$ at a point with squared norm at least CV, which is at least $\mathcal{F}_\lambda(\Theta_0) + 4$ by assumption, and in particular at least $\mathcal{F}_\lambda(\Theta_0) + 2$. For the same reason, any lower bound M of $\mathcal{F}_\lambda$ outside the ball is preserved. Moreover, since every value of $\widetilde{\mathcal{F}}_\lambda$ is a value of $\mathcal{F}_\lambda$, we have $\widetilde{\mathcal{F}}_\lambda \ge \mathcal{F}_\lambda^*$, with equality at the global minimizer in the ball.

It remains to bound the Lipschitz constants. Since c is a constant, the first, second, and third derivatives of the above map have norms $O(1)$, $O(R^{-1})$, and $O(R^{-2})$, respectively. Its image is contained in the ball of radius $(1+c)R$, on which the derivatives of $\mathcal{F}_\lambda$ up to order three are bounded. Thus, by the chain rule, the second and third derivatives of $\widetilde{\mathcal{F}}_\lambda$ are globally bounded, which gives the Lipschitz constants of the gradient and the Hessian. If the derivatives of $\mathcal{F}_\lambda$ are polynomial in V, so are these constants. □

D.4 Perturbed-Gradient Visits and Objective Increases

In this subsection, we state two results of Jin et al. (2019) on perturbed gradient descent, and combine them into the one-block estimate that is used in Step 3 of our proofs. We state the results for a generic objective $f : \mathbb{R}^{D_\Theta} \to \mathbb{R}$ with exact gradients, step size h, and perturbation scale σ. The first result (theorem D.4) shows that perturbed gradient descent visits an approximate SOSP within a bounded number of iterations. The second result (lemma D.5) controls how much the objective can increase over a given number of iterations.

Theorem D.4 (Visiting an approximate SOSP (Jin et al., 2019))**.** *Let $f : \mathbb{R}^{D_\Theta} \to \mathbb{R}$ have an ℓ-Lipschitz gradient and a ρ-Lipschitz Hessian, and let $f(\Theta_0) - \inf_\Theta f(\Theta) < \infty$. 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

$$\Theta_{t+1} = \Theta_t - h\{\nabla f(\Theta_t) + \xi_t\}, \qquad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_{D_\Theta}/D_\Theta),$$

with $h = \widetilde{\Theta}(1/\ell)$ and $\sigma = \widetilde{\Theta}(\epsilon)$ chosen as in Jin et al. (2019), visits a point satisfying

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

within $\widetilde{O}\big(\ell[f(\Theta_0) - \inf f]\epsilon^{-2}\big)$ iterations, with probability at least $1 - \zeta_{\mathrm{J}}$. The hidden factors are polylogarithmic in $D_\Theta, \ell, \rho, \epsilon^{-1}, \zeta_{\mathrm{J}}^{-1}$, and $f(\Theta_0) - \inf f$.

Note that theorem D.4 only guarantees one iterate that is an approximate SOSP, and does not control the other iterates. To this end, we use the following concentration inequality, which bounds the increase of the objective over any fixed interval of iterations.

Lemma D.5 (Descent concentration (Jin et al., 2019))**.** *Under the assumptions of theorem D.4, fix deterministic integers $t_0 \ge 0$ and $s \ge 1$. If $h \le 1/\ell$, Lemma 23 of Jin et al. (2019), again specialized to exact gradients, gives*

$$f(\Theta_{t_0+s}) - f(\Theta_{t_0}) \le -\frac{h}{8}\sum_{j=0}^{s-1} \|\nabla f(\Theta_{t_0+j})\|^2 + c_J h\sigma^2(h\ell s + \iota) \tag{36}$$

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

Proof. This is Lemma 23 of Jin et al. (2019), where the variance of the stochastic gradient is set to zero. The perturbation there is the isotropic Gaussian $\mathcal{N}(0, \sigma^2 I_{D_\Theta}/D_\Theta)$ used in theorem D.4, i.e., σ^2 is the total second moment of the perturbation. With these choices, the bound in that lemma and its failure probability are exactly equation (36). □

Specialization of the parameters. We now specify the parameters in the above results. Since we use exact gradients, the variance of the stochastic gradient in Jin et al. (2019) is zero. In this case, the quantity denoted by N in Appendix B of Jin et al. (2019), which should not be confused with our number of observed states, equals one. Let $C_J \ge 1$ be a universal constant that is sufficiently large to absorb the numerical constants in the parameter choices of Jin et al. (2019). Then, those choices reduce to

$$h = (\iota^9\ell)^{-1}, \qquad \sigma = \iota\epsilon.$$

Consequently, when the initial objective gap is at most A, the escape time and the scale of the objective decrease in Jin et al. (2019) give the following number of iterations for visiting an approximate SOSP:

$$A \max\left\{\frac{\iota/(h\sqrt{\rho\epsilon})}{\iota^{-5}\sqrt{\epsilon^3/\rho}}, \frac{1}{h\epsilon^2}\right\} = O(\ell A\epsilon^{-2}\iota^{15}). \tag{37}$$

This is the length of the blocks used below.

We now combine theorem D.4 and lemma D.5 into an estimate for one block of iterations. In our proofs, we need to bound the increase of the objective from the iteration at which an approximate SOSP is visited, which is random, to the end of the block. Since lemma D.5 only applies to deterministic intervals, we take a union bound over all pairs of iterations $u < v$ in the block, so that the bound holds for all pairs simultaneously.

Lemma D.6 (One-block estimate)**.** *Let $f : \mathbb{R}^{D_\Theta} \to \mathbb{R}$ be bounded from below, with ℓ-Lipschitz gradient and ρ-Lipschitz Hessian, where $\ell, \rho \ge 1$. Suppose that $f(\Theta_0) - \inf f \le A$ for some $A \ge 1$. For $0 < \epsilon, \delta_c \le 1$ and $0 < \zeta < 1/4$, let*

$$\iota \ge C\left[1 + \log\frac{eD_\Theta(\ell+\rho+1)(A+1)}{\epsilon\zeta} + \frac{A+1}{\delta_c} + \delta_c^{-1/6}\right],$$
$$h = (\iota^9\ell)^{-1}, \qquad \sigma = \iota\epsilon, \qquad T_{\mathrm{blk}} = \left\lceil C_J\ell A\epsilon^{-2}\iota^{15}\right\rceil, \tag{38}$$

where C is a sufficiently large universal constant, and ι is further enlarged, if necessary, such that $4(T_{\mathrm{blk}}+1)^2 e^{-\iota} \le \zeta$. Then, with probability at least $1 - 2\zeta$, the following two events occur for the iterates $\Theta_0, \ldots, \Theta_{T_{\mathrm{blk}}}$ of perturbed gradient descent: (i) there exists an iterate Θ_t with $\|\nabla f(\Theta_t)\| \le \epsilon$ and $\nabla^2 f(\Theta_t) \succeq -\sqrt{\rho\epsilon}I$; (ii) $f(\Theta_v) - f(\Theta_u) \le \delta_c$ for all $0 \le u < v \le T_{\mathrm{blk}}$. Moreover, the same statement holds for a block that starts from any iteration t_0, conditional on the history up to t_0, provided that $f(\Theta_{t_0}) - \inf f \le A$.

Proof. **Visit.** We apply theorem D.4 with exact gradients. As discussed above, the step size and the perturbation scale in Jin et al. (2019) reduce to those in equation (38), and by equation (37), an approximate SOSP is visited within $T_{\mathrm{blk}} = \lceil C_J \ell A \epsilon^{-2} \iota^{15} \rceil$ iterations with probability at least $1 - \zeta$. Here, the lower bound on ι in equation (38) absorbs the logarithmic dependence on the failure probability in theorem D.4.

Objective increase. Fix any deterministic pair $0 \le u < v \le T_{\mathrm{blk}}$, and let $s = v - u$. By lemma D.5, with probability at least $1 - 4e^{-\iota}$,

$$f(\Theta_v) - f(\Theta_u) \le -\frac{h}{8} \sum_{j=0}^{s-1} \|\nabla f(\Theta_{u+j})\|^2 + c_J h \sigma^2 (h\ell s + \iota).$$

With the choices in equation (38) and $s \le T_{\mathrm{blk}}$, the second term on the right-hand side is at most $C\{(A+1)/\iota + \epsilon^2/(\ell\iota^6)\} \le \delta_c$, where the last inequality is due to the lower bound on ι. Taking the union bound over at most $(T_{\mathrm{blk}} + 1)^2$ pairs, the failure probability is at most $4(T_{\mathrm{blk}} + 1)^2 e^{-\iota} \le \zeta$. Since the bound holds for all pairs simultaneously, it also holds when u is the random iteration at which the approximate SOSP is visited.

Conclusion. By the union bound, both events occur with probability at least $1 - 2\zeta$. For a block that starts from iteration t_0, conditioning on the history up to t_0 does not change the distribution of the future perturbations, which are independent Gaussian vectors. Thus, the same argument applies conditionally. Finally, we note that the additional requirement $4(T_{\mathrm{blk}} + 1)^2 e^{-\iota} \le \zeta$ is satisfied by some ι of polynomial size, since T_{blk} is polynomial in ι, whereas $e^{-\iota}$ decays exponentially. □

E Experimental Protocol and Complete Summaries

In this section, we provide the details of the population-gradient experiments in section 6. The experiments use the same models and the same perturbed gradient descent algorithm as in our theory, but with practical parameter values instead of the constants required by the proofs. All seeds and configurations were fixed before running the reported experiments, and all runs are reported. A preliminary check of the code used different seeds for the ground truths and the initializations, and did not change the reported configurations. No ground truths, runs, or checkpoints were selected based on the test results.

E.1 Ground truths, observed states, and initialization

Ground truths. For each $V \in \{64, 128, 256\}$, we generate three ground truths with seeds $s \in \{0, 1, 2\}$, using NumPy's default random generator with seed $10000 + 100V + s$. We first draw $V \times 4$ matrices with independent standard Gaussian entries, and normalize each row to unit norm. For the untied model trained on all states, we draw two such matrices, subtract the mean row from the second matrix, and divide it by its largest row norm. The logit matrix is three times the product of the first matrix and the transpose of the second matrix, and Q is obtained by applying the row-wise softmax to it. Since the logit matrix is row-centered, it equals $\mathcal{L}(\log Q)$. Thus, the centered logits have rank four and entries bounded by three. Note that the rank condition is on the centered logits, and not on Q itself.

For the tied model, and for the untied model that is trained on the same observed states as a baseline, we subtract the mean row from the first normalized matrix, and rescale it such that its largest row norm is $\sqrt{3}$. We denote the result by Y, and set $Q = \mathrm{softmax}(YY^\top)$. The set $\mathcal{S}$ consists of the first $3V/4$ indices of a random permutation drawn from the same generator, and $\mathcal{U} = [V] \setminus \mathcal{S}$ is fixed before training. Recall that all V states are possible next states, and "unseen" only refers to the role as a current state. For every ground truth, we check and save the rank and the bound of the centered logits, the smallest eigenvalue of the covariance matrix in definition A.3, and the information gaps. No ground truth is rejected. Over the nine tied ground truths, the smallest eigenvalue of the covariance matrix is 0.28957, and the smallest information gap over the observed and the unseen states is 0.47702.

Initialization. For each ground truth, we use five initialization seeds $a \in \{10, 11, 12, 13, 14\}$, with NumPy generator seed $1000000 + 1000s + a$. All entries of E and U, or of X, are drawn

independently from $\mathcal{N}(0, 1/d)$ and stored in single precision (FP32). The embeddings are not centered, balanced, projected, normalized, frozen, or aligned with the ground truth at any time. For the untied model, the gradients with respect to E and U are computed at the same current parameters before both of them are updated.

E.2 Training configurations and control runs

Configurations. The main grid consists of $d \in \{V/8, V/4, V/2\}$ for each of the three values of V and each of the two models, i.e., 18 settings with 15 runs each. All runs use $T = 8192$ iterations, step size $h = V/4$, and perturbation scale $\sigma = 0.01/\sqrt{V}$. The weight decay is

$$\lambda = \frac{1}{480V} \quad \text{(untied, all states)}, \qquad \lambda = 0.1\sqrt{\frac{V}{Nd^3}} \quad \text{(tied, and untied on the same split)}.$$

The first value is the weight decay in theorem 4.1 with $r = 4$, $B = 3$, and $\Delta_0 = 0.1$. The second value follows the scaling in theorem 5.4 with a practical constant. We note that the constants required by our proofs for the tied model are not satisfied in these experiments. Specifically, with $C_p = 40B_Y r = 480$, the conditions $\lambda \geq 2C_p/d^2$ and $\lambda V \leq 1$ would require $d^2 \geq 960V$, which does not hold for the embedding dimensions that we test. Thus, the tied experiments illustrate the behavior of the model beyond the range of constants covered by our theorems.

Training. The loss is the population cross-entropy loss, i.e., it sums over all V next states and averages uniformly over the current states used for training, and no transitions are sampled. The table Q is generated in double precision (FP64), and then rounded to FP32 and renormalized in each row for training. The checks of the ground truths and the final evaluation of the saved checkpoints use the original FP64 table. The weight decay penalty is $\lambda/2$ times the squared norm of all parameters, without any per-parameter normalization. At every update, we add a fresh Gaussian vector with covariance $\sigma^2 I/D_\Theta$ to the gradient, and then multiply the sum by h, where $D_\Theta = 2Vd$ for the untied model and $D_\Theta = Vd$ for the tied model. The perturbations are generated with seed $271828 + 100V + d$, independently of the initialization, and the same sequence of perturbations is shared by the paired runs with the same parameter dimensions. The perturbations are added throughout training, and there is no stopping rule based on the loss or the test data.

Control runs. At $V = 128$ and $d = 32$, we add three control settings for each model: without perturbation, without weight decay, and without both. These 90 runs use the same step size, number of iterations, ground truths, and initializations as the main runs. In addition, 15 untied runs use the ground truths, the split, and the parameters of the tied model, in order to compare the tied and untied models on the same transfer problem. In total, there are 25 settings and 375 runs. Note that the seeds are shared across the settings so that the runs are paired, and thus the 375 runs are not independent trials. The preliminary check of the code used $V = 64$, $d = 16$, 256 updates, ground-truth seed 90, and initialization seeds 91 and 92, with the same formula for the seeds of the perturbations.

E.3 Measurements and numerical checks

Measurements. At every update $t = 0, \ldots, T$, we record the training gain, the gain on the unseen states when applicable, and the largest squared row norm of the embeddings. The final window is $[T/d, T]$, including both endpoints. For each run, we compute the minimum gain, the fraction of iterations with gain above $7/8$, and the longest interval of consecutive iterations with gain below $7/8$ in this window. The parameters are saved at $t = 0$, T/d, and T, and the objective, the squared norm per state, and the largest scaled predicted probability are recorded every 16 updates. When the model is trained on all states, there is no test gain to report. If a run diverges, it is still counted, and its statistics are recorded as missing. No run diverged in our experiments.

Numerical checks. The closed-form gradients agree with automatic differentiation, with a maximum absolute error of 1.12×10^{-16} in FP64. We also test the variance of the perturbations, the invariance of the softmax to the centering of the logits, and the gradient identity for the input embeddings of the unseen states. After the runs, all reported statistics in the final window are recomputed from the complete traces, whose SHA-256 hashes are verified, and the risks with respect to the original ground truths are evaluated independently in FP64 at all 1125 saved checkpoints. The largest discrepancy in the gain is 2.99×10^{-7}.

Table 2: Results of all settings, with 15 runs each. "Min." is the minimum gain over all runs and all iterations in the final window $[T/d, T]$. "Kept" is the number of runs whose gain is above $7/8$ throughout the final window (training/test). "Max. row" is the largest squared row norm of the embeddings over the entire run, which is below the bound of eight in all runs. In the column "Arm", "Both" means that both the weight decay and the perturbation are used, "Decay" and "Noise" mean that only one of them is used, and "Neither" means that none of them is used. Untied-H denotes the untied model trained on the ground truths and the split of the tied model. A missing test entry means that all current states are used for training. The values are rounded for display, and the comparisons with $7/8$ use the unrounded values.

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

Computing environment. The experiments were run on an NVIDIA RTX 3090 GPU with PyTorch 2.10.0+cu128 and NumPy 2.2.6, using FP32 parameters with deterministic algorithms enabled and TF32 disabled. The total wall-clock time of the training loops is 40.7 seconds, which includes the transfer of the checkpoints, and excludes the file compression and the checks after the runs. This number is only a rough estimate of the running time, since the timer is not synchronized with the GPU. The code, the versions of the dependencies, the raw traces, the checkpoints, and the commands for generating the figures are included in the supplementary material. The results may differ at the level of floating-point rounding on different hardware or library versions.

Summary of the results. Table 2 and figure 5 show that the untied model trained on all states maintains a high gain throughout the final window in all runs. For the tied model, the gain on the unseen states improves as the embedding dimension increases, while the setting with the smallest embedding dimension has a lower training gain. In contrast, the untied model trained on the same observed states has a gain close to zero on the unseen states. In all runs, the squared row norms of the embeddings are bounded, which shows that the condition on the learned embeddings in theorem 5.4 holds in these instances. Finally, the control runs achieve comparable gains without perturbations, and a lower final training error without weight decay. We note that these experiments are not designed to estimate the optimal dependence on the embedding dimension or the failure probability.

F SCALING WITH SAMPLED TRANSITIONS

In this section, we provide the details of the experiments with sampled transitions in section 6. We train the models by SGD over a grid of the number of states, the rank of the ground truth, and the embedding dimension. Figures 6–8 show the results of all 12 choices of (V, r), with every valid embedding dimension and all three initialization seeds.

Models and configurations. We use the untied model $P^{E,U} = \mathrm{softmax}(EU^\top)$ with $E, U \in \mathbb{R}^{V\times d}$ and the tied model $P^X = \mathrm{softmax}(XX^\top)$ with $X \in \mathbb{R}^{V\times d}$, as in section 3. We consider $V \in \{256, 1024, 4096, 16384\}$, ranks $r \in \{8, 16, 32\}$, and embedding dimensions $d \in \{32, 128, 256, 1024\}$. The untied model is trained on all V current states. The tied model is trained on 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 numbers of observed and unseen states are $192/64$, $768/256$, $3072/1024$, and $12288/4096$, respectively. As before, all V states are possible next states, so the embedding of an unseen state is still updated through its role as a next state. We only keep the embedding dimensions with $r \le d < N$, where $N = V$ for the untied model. This gives 78 configurations and 234 runs with initialization seeds 10, 11, and 12.

Ground truths. For the untied model, we draw two independent $V \times r$ Gaussian matrices and normalize each row. We multiply the first matrix by three, subtract the mean row from the second matrix, and divide the second matrix by its largest row norm. Then, Q is the row-wise softmax of the product of the first matrix and the transpose of the second matrix. For the tied model, we normalize the rows of a $V \times r$ Gaussian matrix, subtract the mean row, and rescale the resulting matrix Y such that its largest row norm is $\sqrt{3}$. Then, we set $Q = \mathrm{softmax}(YY^\top)$. In exact arithmetic, the logits are row-centered, have rank at most r, and have entries bounded by three in absolute value. We use the ground-truth seed zero in all configurations, where the generator of the ground truth has seed $10000 + 100V$, and the permutation that determines $\mathcal{S}$ has seed $73000 + V$. The ground truth, the set $\mathcal{S}$, and the stream of sampled transitions are shared by the three initializations of each configuration. Note that the ground truths of the tied and untied models are constructed differently, and thus the panels of the two models are not directly comparable.

Training. All entries of the embeddings are initialized independently from $\mathcal{N}(0, 1/d)$, and all embeddings are trained jointly. At each update, we draw 512 fresh transitions, where the current state is sampled uniformly from the set of observed states, and the next state is sampled from the corresponding row of Q. The next state is sampled by rejection sampling based on the bound on the logits, which avoids storing the full transition table, while the model still computes the softmax over all V states for each sampled current state. We minimize the cross-entropy loss by SGD with the weight decay penalty $\frac{\lambda}{2}\|\Theta\|_{\mathrm{F}}^2$, where $\lambda = 1/(120rV)$ for the untied model and $\lambda = 0.4/(rV)$ for the tied model. We do not use Gaussian perturbations, momentum, projection, or gradient clipping. The learning rate follows a cosine schedule from 64 to 0.25 in the first 8192 updates, and is fixed at 0.25 afterwards until 131072 updates, with the same weight decay. Since fresh transitions are drawn at every update, longer training uses more samples, instead of more passes over a fixed dataset. For $r = 8$, the curves in figure 4 are obtained by continuing the same runs to 262144 updates with the same learning rate and weight decay.

Measurements. We report the relative prediction gain $g_{\mathcal{A}}(\Theta)$ defined in section 3, where the risk $\mathcal{R}_{\mathcal{A}}(\Theta)$ is the population cross-entropy loss, and both the risk and the entropy $\mathcal{H}_{\mathcal{A}}$ are averaged uniformly over $\mathcal{A}$. As in the main text, the gain does not include the weight decay penalty. All reported gains are evaluated on the entire set of observed states or the entire set of unseen states, and not on the smaller subsets of states that are used for monitoring during training.

Numerical settings. After the first 8192 updates, the runs use FP32 parameters, BF16 automatic mixed precision for training, and TF32 where it is enabled. Sampling and risk evaluation use FP32, with TF32 disabled for the evaluation. The first 8192 updates used a different numerical setting, which includes FP64 sampling and risk evaluation. For this reason, the figures in this section start at 16384 updates, and show the checkpoints at 16384, 32768, 65536, and 131072 updates, which are all obtained under the same setting. The curves in figure 4 additionally include the checkpoints at 196608 and 262144 updates, which are evaluated in the same way. In the figures of this section, the

curves connect the medians over the three seeds at each checkpoint, and the shaded areas show the full ranges over the seeds, while figure 4 only shows the medians for readability. No smoothing or extrapolation is applied. Each embedding dimension has a fixed color in all panels, negative gains are shown as they are, and the dashed line marks $7/8$. The vertical scales are shared within each row, and may differ across rows.

Results. For each of the three metrics, i.e., the training gain of the untied model, the training gain of the tied model, and the gain of the tied model on the unseen states, the medians of all 39 configurations increase from 16384 updates to 131072 updates. Some configurations already achieve a high gain at 16384 updates, while the others are still below $7/8$ and keep improving. For example, at $V = 4096$, $r = 32$, and $d = 256$, the training gain of the tied model increases from 0.711 to 0.932, and its gain on the unseen states increases from 0.340 to 0.752 (Figure 8). This shows that the gain on the unseen states can lag behind the training gain, even when both are improving. We also observe that a larger embedding dimension does not always give a higher gain under a fixed number of updates. We note that these results are obtained by SGD with sampled transitions, which is different from the setting of our theorems. Moreover, a low gain at the last checkpoint alone does not imply that the run has converged to a local minimum with low gain.

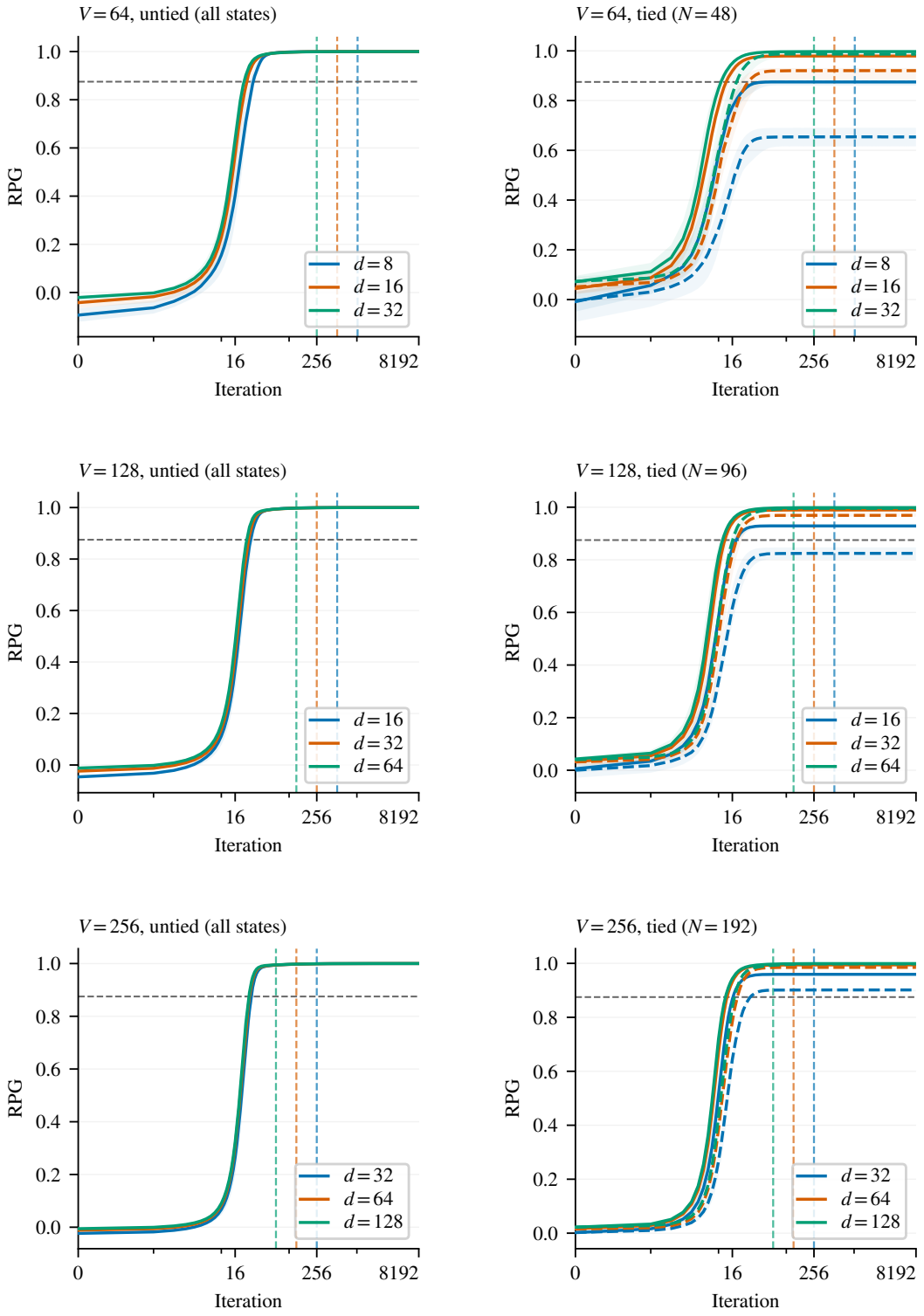


Figure 5: Trajectories of all settings in the main grid, shown as medians (curves) and full ranges (shaded areas). Each panel contains the three embedding dimensions. The untied model is trained on all current states. For the tied model, the solid and dashed curves show the training gain and the gain on the unseen states, respectively. The perturbations are added throughout training. The curves are subsampled for plotting only, and all reported statistics are computed over all updates.

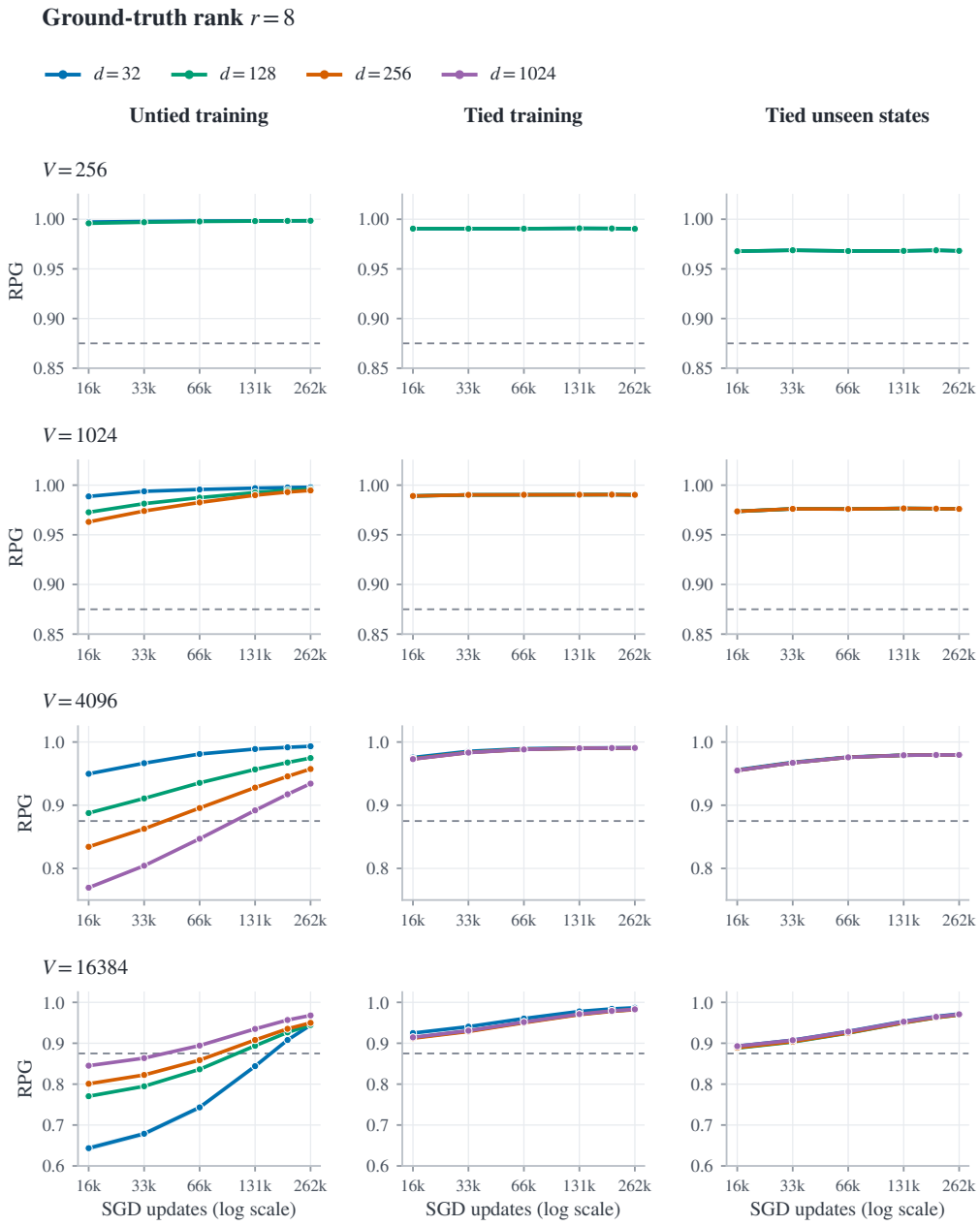


Figure 6: SGD with sampled transitions at rank $r = 8$. The rows correspond to $V = 256, 1024, 4096, 16384$ from top to bottom, and the columns show the training gain of the untied model, the training gain of the tied model, and the gain of the tied model on the unseen states, from left to right. All valid embedding dimensions and all three initialization seeds are included. The curves show the medians over the seeds, and the shaded areas show their full ranges. The dashed line marks $7/8$, and the vertical scales may differ across rows.

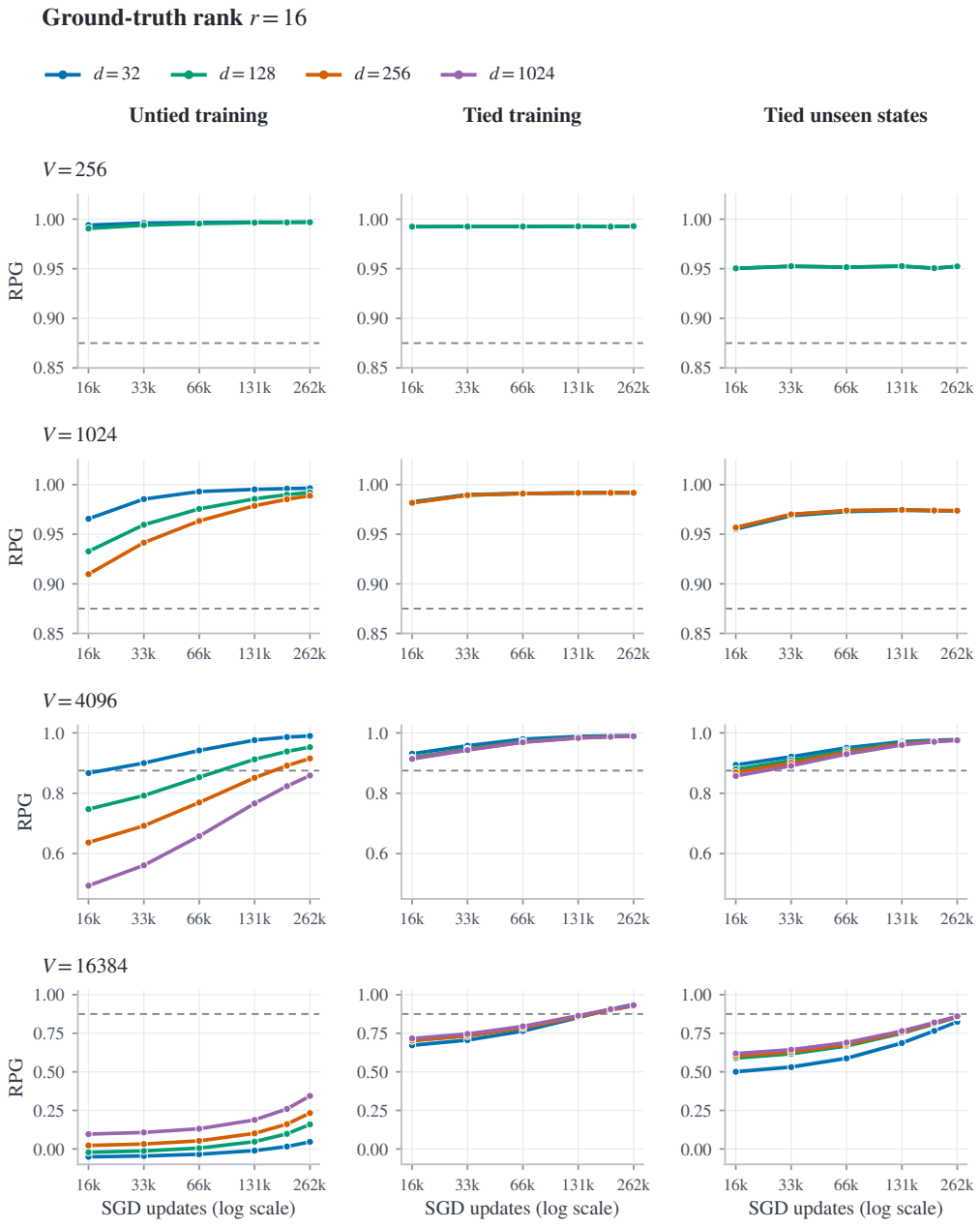


Figure 7: SGD with sampled transitions at rank $r = 16$, with $V = 256, 1024, 4096, 16384$ from top to bottom. All valid embedding dimensions and all three initialization seeds are included. The columns, curves, and shaded areas are the same as in Figure 6.

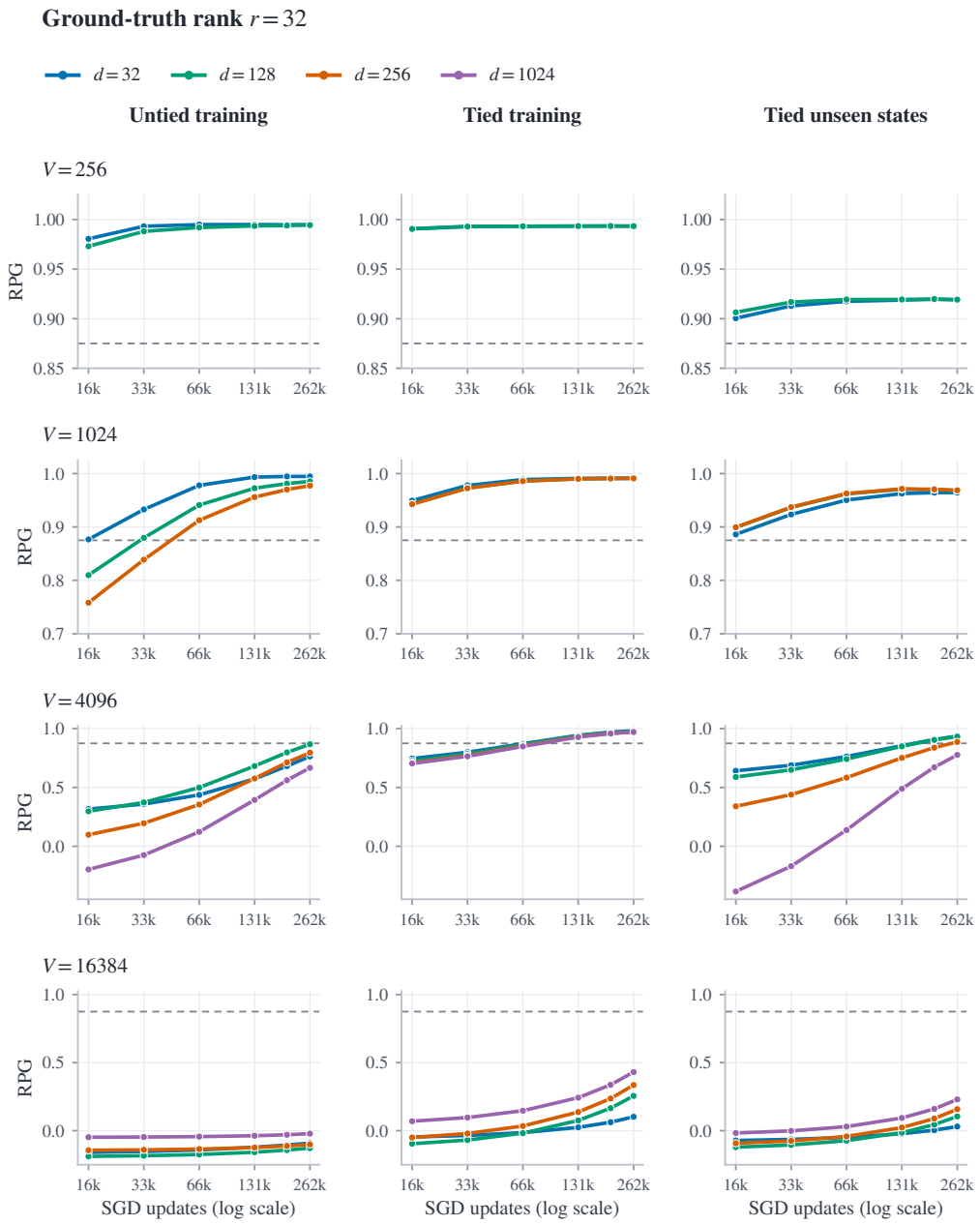


Figure 8: SGD with sampled transitions at rank $r = 32$, with $V = 256, 1024, 4096, 16384$ from top to bottom. All valid embedding dimensions and all three initialization seeds are included, and negative gains are shown as they are. The columns, curves, and shaded areas are the same as in Figure 6.