FROM POLICY EVALUATION TO OFF-POLICY CONTROL: PROVABLE EMERGENCE OF IN-CONTEXT Q-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers can implement reinforcement-learning algorithms in context, yet how off-policy control emerges within Transformers remains poorly understood. We study Q-learning, a foundational model-free control algorithm that combines off-policy bootstrapping with greedy policy improvement. We prove that a two-layer attention-only Transformer can implement a finite-temperature form of Q-learning, where the temperature $\tau > 0$ controls the approximation to greedy action selection, recovering window Q-learning as $\tau \to 0$. We train the structured Transformer on randomly generated MDP tasks using output-only supervision. Experiments show that the theoretically predicted parameter structure emerges during training, including the internal greedy action-selection mechanism, and that the trained Transformer achieves performance comparable to Q-learning on unseen tasks. Our theoretical analysis further shows that the equivalent realizations form a solution manifold and that, under suitable task-diversity conditions, the training iterates converge locally to this manifold at an exponential rate. For the random-task model analyzed here, the local worst-case number of gradient steps needed to reduce the distance to the manifold by a factor ϵ is $\Theta\left(\tau^{-5}\log(1/\varepsilon)\right)$, quantifying the increasing optimization cost as the learned update approaches window Q-learning.

1 INTRODUCTION

In-context reinforcement learning (ICRL) offers an alternative to the conventional paradigm of reinforcement learning. Rather than adapting to a new task through parameter updates, a pretrained agent can adjust its behavior directly from the interaction history provided in its context. This phenomenon has motivated an algorithmic interpretation of ICRL: a Transformer may not merely map a trajectory to a prediction, but may execute a reinforcement-learning algorithm within its forward pass. Recent theoretical work supports this view. Transformers have been shown to implement temporal-difference (TD) methods for policy evaluation, and more recent results extend this perspective to policy-improvement methods such as SARSA and actor-critic (Wang et al., 2024; 2025; Xie et al., 2026a;b; Liang & Lai, 2026).

A central regime remains largely unexplored: off-policy control. Existing policy-improvement constructions focus primarily on on-policy learning, where the behavior policy and the policy being improved are tightly coupled (Liang & Lai, 2026). This coupling is restrictive in real applications where an agent must learn from previously collected, heterogeneous, or suboptimal experience rather than from trajectories generated by its current policy. Off-policy control asks whether useful policy improvement can still be carried out when these two roles are decoupled. Q-learning provides a canonical setting for studying this problem. Given a transition (s, a, r, s'), its target is $r+\gamma \max_{a'} Q(s', a')$, which does not depend on the successor action selected by the behavior policy. Thus, subject to sufficient state-action coverage, Q-learning can use experience generated by a suboptimal behavior policy while improving toward a greedy target policy. This makes Q-learning a natural testbed for a more demanding form of in-context adaptation: whether a Transformer can use off-policy experience to carry out control, rather than merely evaluate or improve the policy that generated its context.

This transition is substantially more than a change of update rule. At a conceptual level, it marks the passage from Bellman evaluation to Bellman optimality. An in-context implementation of Q-learning must compare competing successor actions, resolve the resulting action selection, and then use that decision to construct the bootstrap target. We show that there exists a two-layer

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

Transformer, combining masked softmax attention with linear attention, that jointly implements next-state action selection and the corresponding window semi-gradient Q-learning update. In this sense, the architecture of transformer does not merely reproduce the numerical output of Q-learning; its internal computation mirrors the decomposition of the algorithm itself.

What kind of pretraining can give rise to such a Transformer? We develop a gradient-based training algorithm in which a finite-temperature soft window-Q update serves as the teacher for each sampled context. Here, the softmax temperature $\tau > 0$ controls the approximation to greedy action selection, with standard window hard-max Q-learning recovered as $\tau \to 0$. Importantly, within the prescribed tokenization and attention mask, training supervises only the output update and imposes no numerical targets on the internal parameter blocks to implement either window Q-learning or action selection. Nevertheless, the constructed parameter structure emerges under training, recovering the underlying computation, including the internal action-selection mechanism. On unseen tasks, the trained Transformer achieves performance comparable to window Q-learning with the same step size.

To understand why a Q-learning implementation can emerge under this training procedure, we provide a theoretical analysis of its convergence. The nonlinear dependence of softmax attention on the action-score parameters makes the analysis substantially more intricate than in linear-attention settings. Meanwhile, the Q-learning realization is inherently non-identifiable, with different parameter configurations implementing the same update. This makes the relevant optimization target a manifold of functionally equivalent Q-learning realizations rather than a single parameter vector. We characterize this realization manifold and show that, under verifiable conditions on the training distribution, the discrete-time training dynamics converge locally to this manifold at an exponential rate. The analysis further reveals a quantitative difficulty in approaching greedy Q-learning: for the random-task model considered here, the local worst-case number of gradient steps needed to reduce the distance to the manifold by a factor ϵ is $\Theta\left(\tau^{-5}\log(1/\varepsilon)\right)$, quantifying the increasing optimization cost as the learned update approaches hard-max Q-learning.

Our results make three concrete advances. First, we give an explicit attention-only construction that jointly realizes successor-action comparison and the window Q-learning update. Second, we design a gradient-based training procedure and characterize a smooth manifold of functionally equivalent Q-learning realizations, showing local exponential convergence of the training dynamics to this manifold. Third, we quantify how the softmax temperature governs the tradeoff between approximating greedy action selection and optimization difficulty. Together, these results extend the mechanistic analysis of ICRL from policy evaluation and on-policy control to off-policy Q-learning.

2 Related work

Transformer-based agents have been studied extensively for sequential decision making. Decision Transformer and Trajectory Transformer formulate reinforcement learning as sequence modeling (Chen et al., 2021; Janner et al., 2021), while Algorithm Distillation and related ICRL methods show that pretrained sequence models can adapt to unseen tasks through interaction histories without test-time parameter updates (Laskin et al., 2023; Lee et al., 2023; Grigsby et al., 2024). On the theoretical side, Lin et al. (2024) establish provable ICRL under supervised pretraining and show that Transformers can approximate classical online decision-making algorithms.

Recent work has taken a mechanistic view of ICRL, asking which RL algorithms can be implemented by Transformer forward computation and whether the corresponding algorithmic parameters can emerge through pretraining. Wang et al. (2024; 2025) show that linear-attention Transformers can implement TD updates for policy evaluation, while Xie et al. (2026b) extend this result to standard softmax attention via a kernel-weighted TD construction. Beyond policy evaluation, Liang & Lai (2026) show that Transformers can realize semi-gradient SARSA and actor-critic updates for policy improvement. However, analogous results for off-policy control, including classical Q-learning, where trajectory data are coupled with subsequent action selection, remain largely unexplored. We study the emergence of Q-learning and the convergence of pretraining, with particular emphasis on the effect of the action-selection mechanism on convergence. Unlike Liang & Lai (2026), we analyze discrete-time training iterates and provide verifiable conditions for convergence.

Off-policy learning is particularly relevant to ICRL when context trajectories are offline, heterogeneous, or generated by suboptimal behavior policies. In such settings, the quality of the learned policy

need not be tied to that of the behavior generating the trajectories, making off-policy value-based learning particularly relevant. Dong et al. (2025) study ICRL from suboptimal historical trajectories and propose an advantage-weighted policy-learning framework to improve upon the behavior policies in the data. Scalable In-Context Q-Learning (Liu et al., 2026) combines context-conditioned value estimation, dynamic programming, and policy extraction to enable iterative improvement from suboptimal offline experience. In-Context Compositional Q-Learning (Xu et al., 2026) instead formulates Q-function estimation as contextual inference and uses retrieved transitions to infer local Q-functions for offline RL. Our work differs fundamentally from these studies by taking a mechanistic view of how Transformers realize Q-learning in ICRL, characterizing both the forward-pass implementation of the Q-learning update and the emergence of the corresponding parameter structure through training.

3 Background

3.1 Transformers and Self-Attention

All vectors are treated as column vectors. We denote by I_n the $n \times n$ identity matrix over $\mathbb{R}$, by $0_{m\times n}$ the $m \times n$ all-zero matrix, and by 0_m the m-dimensional zero vector. For vectors x and y of the same dimension, we use $\langle x, y\rangle = x^\top y$ to denote their Euclidean inner product.

Let $\mathbf{H} = [h_1, \ldots, h_n] \in \mathbb{R}^{D\times n}$ denote a prompt representation, where each column $h_i \in \mathbb{R}^D$ corresponds to one token. The standard single-head softmax self-attention (Vaswani et al., 2017) processes the prompt via $\mathrm{SoftAttn}_{\mathbf{P},\mathbf{V}}(\mathbf{H}) \triangleq (\mathbf{VH})\,\mathrm{softmax}\big(\mathbf{H}^\top\mathbf{PH}\big)$, where $\mathbf{V} \in \mathbb{R}^{D\times D}$ is the value projection matrix, $\mathbf{P} := \mathbf{W}_K^\top\mathbf{W}_Q \in \mathbb{R}^{D\times D}$ collects the key and query projection parameters, the conventional scaling factor in scaled dot-product attention is absorbed into $\mathbf{P}$, and the softmax is applied independently to each query column, normalizing the attention scores over all key positions.

As a widely studied alternative to softmax self-attention, linear self-attention replaces the softmax-normalized attention weights with unnormalized bilinear attention scores (Ahn et al. (2023), Gatmiry et al. (2024), Zhang et al. (2024)). Its formulation is given by $\mathrm{LinAttn}_{\mathbf{P},\mathbf{V}}(\mathbf{H}) \triangleq \frac{1}{N}(\mathbf{VH})(\mathbf{H}^\top\mathbf{PH})$, where the factor $\frac{1}{N}$ normalizes the aggregation with respect to the prompt length.

To model a multi-layer Transformer structure, we stack L attention layers with residual connections. Let $\mathbf{H}^{(0)} = \mathbf{H} \in \mathbb{R}^{D\times n}$ denote the initial input prompt representation. For each layer $\ell \in \{1, 2, \ldots, L\}$, the layer-wise update rule reads

$$\mathbf{H}^{(\ell)} := \mathbf{H}^{(\ell-1)} + \mathrm{Attn}_{\mathbf{P}^{(\ell)},\mathbf{V}^{(\ell)}}\left(\mathbf{H}^{(\ell-1)}\right) \in \mathbb{R}^{D\times n}, \tag{1}$$

where Attn denotes either $\mathrm{SoftAttn}$ or $\mathrm{LinAttn}$, and $\{\mathbf{P}^{(\ell)}, \mathbf{V}^{(\ell)}\}$ are the layer-specific projection parameters at layer ℓ. After stacking L attention layers, the final Transformer output is defined as

$$\mathrm{TF}_L\big(\mathbf{H}; \{\mathbf{P}^{(\ell)}, \mathbf{V}^{(\ell)}\}_{\ell=1}^{L}\big) := f_{\mathrm{read}}\big(\mathbf{H}^{(L)}\big), \tag{2}$$

where $f_{\mathrm{read}} \colon \mathbb{R}^{D\times n} \to \mathbb{R}^{D_{\mathrm{out}}}$ denotes a fixed readout map. Equation (2) establishes the notation convention we adopt for defining the output of an L-layer Transformer. Specifically, the attention mechanism produces a matrix, while we require a vector output for policy improvement. Following prior works (Xie et al. (2026b), Liang & Lai (2026),Wang et al. (2024)), we take the bottom-right vector of length d of the output matrix as this vector.

3.2 Reinforcement Learning and Q-learning

We consider an infinite-horizon discounted Markov decision process (MDP) $\mathcal{E} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(\cdot \mid s, a) \in \mathcal{P}(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. Here, $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$. At time t, an agent in state S_t selects an action $A_t \sim \pi(\cdot \mid S_t)$ according to a policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, receives a reward $R_{t+1} = R(S_t, A_t, S_{t+1})$, and transitions to a successor state $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. The discounted return from time t is defined as $G_t := \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$. The state-value and action-value functions associated with a policy π are respectively defined as $v_\pi(s) := \mathbb{E}_\pi[G_t \mid S_t = s]$, $q_\pi(s, a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$. The objective of reinforcement learning is to find a policy that maximizes the expected discounted return.

Q-learning (Watkins, 1989; Watkins & Dayan, 1992) is a classical off-policy TD control algorithm for estimating the optimal action-value function. In the tabular setting, an estimate $Q_t(s, a)$ is maintained explicitly for each state-action pair. Given a transition $(S_t, A_t, R_{t+1}, S_{t+1})$, Q-learning updates the value of the visited state-action pair according to

$$Q_{t+1}(S_t, A_t) = (1 - \alpha)Q_t(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(S_{t+1}, a') \right], \tag{3}$$

where $\alpha \in (0, 1]$ is the learning rate. All other state-action values remain unchanged. To handle large state-action spaces, we consider linear function approximation (Tsitsiklis & Van Roy, 1997; Melo et al., 2008). Given a feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and a parameter vector $\mathbf{w}_t \in \mathbb{R}^d$ at time t, the action-value function is approximated by $q_{\mathbf{w}_t}(s, a) := \mathbf{w}_t^\top \boldsymbol{\phi}(s, a)$. Given a transition $(S_t, A_t, R_{t+1}, S_{t+1})$, the semi-gradient Q-learning update is

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} \mathbf{w}_t^\top \boldsymbol{\phi}(S_{t+1}, a') - \mathbf{w}_t^\top \boldsymbol{\phi}(S_t, A_t) \right) \boldsymbol{\phi}(S_t, A_t). \tag{4}$$

The tabular setting is recovered as a special case when $\mathcal{S}$ and $\mathcal{A}$ are finite, $d = |\mathcal{S}||\mathcal{A}|$, and each state-action pair (s, a) is assigned a distinct one-hot feature vector $\boldsymbol{\phi}(s, a) = \mathbf{e}_{(s,a)}$. In this case, each coordinate of $\mathbf{w}_t$ corresponds to a single state-action value $Q_t(s, a)$, and the above semi-gradient update reduces exactly to the standard tabular Q-learning update.

4 Transformer can implement In-context soft Q-learning

We consider a Boltzmann-smoothed Q-learning update that replaces the hard maximum over successor actions with a Boltzmann-weighted average. Given the current parameter $\mathbf{w}$, trajectories $(s_0, a_0, r_1, \ldots, s_{n-1}, a_{n-1}, r_n, s_n)$ can be obtained by interacting with environment under its behavioral policy, where behavioral policy is a behavior policy π_b, which may differ from greedy target policy. Let $\mathbf{z}$ denote a context consisting of an n-transition window, parameter $\mathbf{w}$, and the features of all candidate successor actions. For $i = 0, \ldots, n-1$, define $\phi_i = \phi(s_i, a_i)$, $\phi_{i,a}^+ = \phi(s_{i+1}, a)$, $a \in \mathcal{A}$ and

$$\pi_{i,a}^\tau = \frac{\exp(\mathbf{w}^\top \phi_{i,a}^+ / \tau)}{\sum_b \exp(\mathbf{w}^\top \phi_{i,b}^+ / \tau)}, \qquad \bar{\phi}_i^{+,\tau} = \sum_a \pi_{i,a}^\tau \phi_{i,a}^+. \tag{5}$$

The corresponding batched semi-gradient update is $\mathbf{w}_{\text{soft}}^{(\tau)} = \mathbf{w} + \frac{\alpha}{n} \sum_{i=0}^{n-1} \phi_i (r_{i+1} + \gamma \mathbf{w}^\top \bar{\phi}_i^{+,\tau} - \mathbf{w}^\top \phi_i)$. Here, $\mathbf{w}^\top \bar{\boldsymbol{\phi}}_i^{+,\tau}$ is the Boltzmann-weighted average of the successor action values, which replaces the hard maximum $\max_{a \in \mathcal{A}} \mathbf{w}^\top \boldsymbol{\phi}_{i,a}^+$ in Q-learning. This soft backup recovers the Q-learning backup in the zero-temperature limit $\tau \downarrow 0$; see Lemma A.1.

We now give an explicit Transformer construction and show that its forward pass implements the soft Q-learning update. Each token is partitioned into six blocks, $[\Phi; G; R; \Psi; U; W]$ of dimensions $[d, d, 1, d, d, d+1]$, respectively. The prompt consists of three types of tokens: transition tokens, which encode the observed transitions; candidate-action tokens, which encode the features of candidate actions at the successor states; and a query token, which carries the current parameter vector. Letting $\widetilde{\mathbf{w}} = [1; \mathbf{w}]$, for each transition $i = 0, \ldots, n-1$ and candidate action a, we define

$$t_i = [\phi_i; 0; r_{i+1}; 0; \mathbf{w}; 0], \quad c_{i,a} = [0; 0; 0; \phi_{i,a}^+; 0; 0], \quad p = [0; 0; 0; 0; 0; \tilde{\mathbf{w}}] \tag{6}$$

Thus, the token dimension is $D = 5d + 2$, and the prompt contains $N = n(m + 1) + 1$ tokens. For each transition i, we group the candidate-action tokens as $\mathbf{C}_i = [c_{i,a^1}, c_{i,a^2}, \ldots, c_{i,a^m}]$, so that $\mathbf{C}_i$ contains exactly the candidate-action tokens associated with all actions available at the successor state s_{i+1}. The resulting input prompt is

$$\mathbf{H}^{(0)} = \big[t_0, \mathbf{C}_0, t_1, \mathbf{C}_1, \ldots, t_{n-1}, \mathbf{C}_{n-1}, p\big]. \tag{7}$$

Our Transformer consists of two successive attention layers. The first layer uses single-head softmax attention to compute a Boltzmann-weighted aggregation over candidate successor actions, thereby implementing a smooth approximation to the maximization operator in Q-learning. The second layer then uses linear attention to realize the corresponding update. Specifically, the first layer is defined as

$$\mathbf{H}^{(1)} = \mathbf{H}^{(0)} + \big(\mathbf{V}_1 \mathbf{H}^{(0)}\big) \mathrm{softmax}\Big((\mathbf{H}^{(0)})^\top \mathbf{P}_1 \mathbf{H}^{(0)} + \mathbf{M} \Big). \tag{8}$$

We use $\star$ to denote an arbitrary matrix block of the corresponding dimension. These blocks need not vanish and may take arbitrary values, since they do not affect the entries relevant to the construction under the prescribed attention mask and token structure. We choose

$$\mathbf{P}_1 = \begin{bmatrix} \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \\ 0_{d\times d} & \star & 0_{d\times 1} & \star & \mathbf{A}_\theta & \star \\ \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \end{bmatrix}, \quad \mathbf{V}_1 = \begin{bmatrix} \star & \star & \star & \mathbf{D}_\Phi & \star & 0 \\ \star & \star & \star & \mathbf{L} & \star & 0 \\ \star & \star & \star & \mathbf{D}_R & \star & 0 \\ \star & \star & \star & 0 & \star & 0 \\ \star & \star & \star & \mathbf{D}_U & \star & 0 \\ \star & \star & \star & 0 & \star & \mathbf{E}_\rho \end{bmatrix} \tag{9}$$

where $\mathbf{A}_\theta = \tau^{-1} I_d$, $\mathbf{D}_\Phi, \mathbf{D}_U \in \mathbb{R}^{d\times d}$, $\mathbf{L} \in \mathrm{GL}(d)$, $\mathbf{D}_R \in \mathbb{R}^{1\times d}$, and $\mathbf{E}_\rho = \mathbf{Q} - I_{d+1}$. Here $\mathbf{Q} = \begin{bmatrix} \rho & b^\top \\ 0_{d\times 1} & I_d \end{bmatrix}$, $b \in \mathbb{R}^d$, $\rho \neq 0$, and $\mathrm{GL}(d)$ denotes the set of invertible $d \times d$ matrices.

We define the attention mask $\mathbf{M} \in \{0, -\infty\}^{N\times N}$, where rows index keys and columns index queries. The mask is chosen such that each transition token t_i attends only to its corresponding candidate-action tokens $\{c_{i,a}\}_{a\in\mathcal{A}}$, while each $c_{i,a}$ and the query token p attend only to p. All other pairs are masked out, ensuring that each t_i aggregates only its corresponding successor action candidates.

The second layer employs single-head linear attention:

$$\mathbf{H}^{(2)} = \mathbf{H}^{(1)} + \frac{1}{N}(\mathbf{V}_2\mathbf{H}^{(1)})\left((\mathbf{H}^{(1)})^\top \mathbf{P}_2 \mathbf{H}^{(1)}\right). \tag{10}$$

We choose

$$\mathbf{P}_2 = \begin{bmatrix} \star & \star & \star & \star & \star & \mathbf{B}_- \\ \star & \star & \star & \star & \star & \mathbf{B}_+ \\ \star & \star & \star & \star & \star & \mathbf{B}_r \\ \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \mathbf{B}_U \\ \star & \star & \star & \star & \star & \star \end{bmatrix}, \mathbf{V}_2 = \begin{bmatrix} \star_{(4d+2)\times d} & \star_{(4d+2)\times(d)} & \star_{(4d+2)\times(3d+2)} \\ \mathbf{C} & \mathbf{C}_G & 0_{d\times(3d+2)} \end{bmatrix}. \tag{11}$$

where $\mathbf{B}_- = -c\mathbf{J}\mathbf{Q}^{-1}$, $\mathbf{B}_+ = c\mathbf{L}^{-\top}[-\mathbf{D}_R^\top, \gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top\Omega]\mathbf{Q}^{-1}$, $\mathbf{B}_r = ce_0^\top\mathbf{Q}^{-1}$, $\mathbf{B}_U = c\Omega\mathbf{J}\mathbf{Q}^{-1}$, $\mathbf{C} = (\alpha/c)I_d$, and $\mathbf{C}_G = (-\alpha/c)\mathbf{D}_\Phi\mathbf{L}^{-1}$. Here, $\Omega^\top = -\Omega$ is arbitrary and may, in particular, be nonzero, $\mathbf{J} = [0, I_d]$, $e_0^\top = [1, 0]$, and $c \neq 0$.

Theorem 1 (Forward pass as soft Q-learning)**.** *Consider two-layer Transformer defined above, with masked softmax attention* (8) *parameterized by* $\mathbf{P}_1, \mathbf{V}_1$ *in the first layer, and linear attention* (10) *parameterized by* $\mathbf{P}_2, \mathbf{V}_2$ *in the second layer. For the input prompt* $\mathbf{H}^{(0)}$ *in* (7)*, the following hold.*

(i) *The Transformer exactly implements the soft Q-learning update:*

$$\mathbf{w}_{\mathrm{TF}}^{(\tau)} := \mathrm{TF}_\tau(\mathbf{H}^{(0)}) = \mathbf{w} + \frac{\alpha}{n}\sum_{i=0}^{n-1}\phi_i\left[r_{i+1} + \gamma\mathbf{w}^\top\bar{\phi}_i^{+,\tau} - \mathbf{w}^\top\phi_i\right],$$

where $\bar{\phi}_i^{+,\tau} = \sum_{a\in\mathcal{A}}\pi_{i,a}^\tau\phi_{i,a}^+$, $\pi_{i,a}^\tau = \exp(\mathbf{w}^\top\phi_{i,a}^+/\tau)/\sum_{b\in\mathcal{A}}\exp(\mathbf{w}^\top\phi_{i,b}^+/\tau)$.

(ii) *For each* i*, define* $q_{i,a} := \mathbf{w}^\top\boldsymbol{\phi}_{i,a}^+$, $q_i^\star := \max_{a\in\mathcal{A}} q_{i,a}$*, and* $\mathcal{A}_i^\star := \arg\max_{a\in\mathcal{A}} q_{i,a}$*. Let* $m := |\mathcal{A}| < \infty$ *and* $K_i := |\mathcal{A}_i^\star|$*. Suppose that* $\|\phi(s,a)\|_2 \leqslant B_\phi$ *and* $|q_{\mathbf{w}}(s,a)| \leqslant B_q$ *for all state-action pairs. For every* i *such that* $\mathcal{A}_i^\star \neq \mathcal{A}$*, define the action gap* $\Delta_i := q_i^\star - \max_{a\notin\mathcal{A}_i^\star} q_{i,a} > 0$*. Then* $\lim_{\tau\downarrow 0}\mathbf{w}_{\mathrm{TF}}^{(\tau)} = \mathbf{w}_Q$ *and*

$$\left\|\mathbf{w}_Q - \mathbf{w}_{\mathrm{TF}}^{(\tau)}\right\|_2 \leqslant 2\alpha\gamma B_q B_\phi \sum_{i:\,\mathcal{A}_i^\star\neq\mathcal{A}} \frac{m - K_i}{nK_i} e^{-\Delta_i/\tau},$$

where $\mathbf{w}_Q = \mathbf{w} + \frac{\alpha}{n}\sum_{i=0}^{n-1}\phi_i\left[r_{i+1} + \gamma\max_{a\in\mathcal{A}}\mathbf{w}^\top\phi_{i,a}^+ - \mathbf{w}^\top\phi_i\right]$ *is window Q-learning update.*

The proof of Theorem 1 is provided in Appendix A, together with the corresponding result for the tabular setting. Moreover, if each maximizer is unique and $\Delta_{\min} := \min_i \Delta_i > 0$, then $\|\mathbf{w}_Q - \mathbf{w}_{\mathrm{TF}}^{(\tau)}\|_2 \leqslant 2\alpha\gamma B_q B_\phi(m-1)e^{-\Delta_{\min}/\tau}$, $m > 1$, showing that the Transformer update converges exponentially fast to the standard window Q-learning update as $\tau \downarrow 0$. The theorem thus establishes that a two-layer Transformer can exactly implement soft Q-learning in its forward pass and recover standard Q-learning in the zero-temperature limit. We call this in-context Q-learning.

Algorithm 1 IC-Q-Learning Transformer Training with Independent-Context Minibatches

Require: Transformer parameters $\boldsymbol{\theta}^{(0)} = (\mathbf{P}_1^{(0)}, \mathbf{V}_1^{(0)}, \mathbf{P}_2^{(0)}, \mathbf{V}_2^{(0)})$, trajectory window length n, number of optimizer updates U, minibatch size B, ϵ-greedy parameter ϵ, Transformer learning rate η, Q-learning step size α, and temperature τ.

for $k = 0, \dots, U-1$ **do**
  for $b = 1, \dots, B$ **do**
    Independently sample a new MDP $\mathcal{M}_b^{(k)}$, feature map $\phi_b^{(k)}$, $\mathbf{w}_b^{(k)}$, and initial state s_b.
    Estimate $q_{\mathbf{w}_b}^{(k,b)}(s,a) = (\mathbf{w}_b)^\top \phi_b^{(k)}(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Sample an n-step trajectory $(s_{b,0}, a_{b,0}, r_{b,1}, s_{b,1}, a_{b,1}, \dots, r_{b,n}, s_{b,n})$ from $\mathcal{M}_b^{(k)}$ according to the ϵ-greedy behavior policy derived from $q_{\mathbf{w}}^{(k,b)}$.
    Formulate $\mathbf{H}_b$ as in (7) using the sampled trajectory, the candidate features $\{\phi_b^{(k)}(s_{b,i+1}, a) : a \in \mathcal{A}\}$, and the current parameter $\mathbf{w}_b$.
    Compute $\widetilde{\mathbf{H}}_b = \mathbf{H}_b + (\mathbf{V}_1 \mathbf{H}_b)\,\mathrm{softmax}\left((\mathbf{H}_b)^\top \mathbf{P}_1 \mathbf{H}_b + \mathbf{M}\right)$.
    Compute $\mathbf{w}_{\mathrm{TF},b} = \left[\widetilde{\mathbf{H}}_b + \frac{1}{N}(\mathbf{V}_2 \widetilde{\mathbf{H}}_b)\left((\widetilde{\mathbf{H}}_b)^\top \mathbf{P}_2 \widetilde{\mathbf{H}}_b\right)\right]_{-d:,-1}$.
    Compute the teacher update $\mathbf{w}_{\mathrm{soft},b}^{(\tau)}$ according to soft Q-learning.
  end for
  Compute $\widehat{\mathcal{L}}_B^{(k)} = \frac{1}{2B}\sum_{b=1}^{B} \|\mathbf{w}_{\mathrm{TF},b} - \mathbf{w}_{\mathrm{soft},b}^{(\tau)}\|_2^2$.
  Update $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} - \eta \nabla_\theta \widehat{\mathcal{L}}_B^{(k)}$
end for
return trained Transformer parameters $\boldsymbol{\theta}$.

5 IC-Q-Learning algorithm and training dynamics analysis

5.1 IC-Q-Learning Algorithm and Training Dynamics

We now design an ICRL training procedure for the proposed soft Q-learning Transformer, inspired by prior ICRL training frameworks (Laskin et al., 2023; Liang & Lai, 2026), as summarized in Algorithm 1. The training objective follows the teacher-mimicking principle: for each sampled context, the Transformer is trained to predict the parameter obtained after one soft Q-learning update. In contrast to directly supervising the internal attention matrices, the loss is imposed only on the final parameter prediction. Hence, the Transformer is not explicitly provided with the analytical parameter construction derived in the preceding section.

Fix integers $d \geqslant 1$, $m \geqslant 2$, $n \geqslant 1$ and constants $\alpha > 0$, $0 < \gamma < 1$, $\tau > 0$. Let $\mathbf{z} := \left(\mathbf{w}, \{\phi_i, r_{i+1}, \phi_{i,j}^+ : 0 \leqslant i < n,\ j \in [m]\}\right) \sim \mathcal{D}_n$. Let $\mathbf{w}_{\mathrm{soft}}^{(\tau)}(\mathbf{z})$ denote the corresponding soft Q-learning target. We define the population teacher-mimicking objective as

$$\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{z}}\left[\frac{1}{2}\| \mathrm{TF}_\theta(\mathbf{H}(\mathbf{z})) - \mathbf{w}_{\mathrm{soft}}^{(\tau)}(\mathbf{z})\|_2^2\right]. \tag{12}$$

where the expectation is taken jointly over the training-MDP distribution, random feature maps, behavior trajectories, and independently sampled input Q-function parameters.

We optimize the population objective in (12) using independently sampled training contexts. At each optimization step, we draw a minibatch $\mathbf{z}_1, \dots, \mathbf{z}_B \overset{\text{i.i.d.}}{\sim} \mathcal{D}_n$, where each context is generated from an independently sampled MDP and feature map. The empirical loss is $\widehat{\mathcal{L}}_B(\boldsymbol{\theta}) := \frac{1}{2B}\sum_{b=1}^{B} \| \mathrm{TF}_\theta(\mathbf{H}(\mathbf{z}_b)) - \mathbf{w}_{\mathrm{soft}}^{(\tau)}(\mathbf{z}_b)\|_2^2$. Here, B controls the amount of averaging across independently sampled training contexts within each parameter update, and hence the stochasticity of the resulting gradient estimate. Further details are provided in Proposition B.1.

Write $\mathcal{K} = (\Phi, R, U)$ and $\mathcal{I} = (\Phi, G, R, U)$. Let W_{out} denote the last d coordinates of W. Let $\boldsymbol{\theta}_{\mathrm{out}}$ collect all entries of $\boldsymbol{\theta} = (\mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2)$ outside the following 16 blocks: $\mathbf{P}_{1,\Psi\mathcal{K}}$, $\mathbf{V}_{1,\mathcal{I}\Psi}$, $\mathbf{V}_{1,WW}$, $\mathbf{P}_{2,\mathcal{I}W}$, $\mathbf{V}_{2,W_{\mathrm{out}}\mathcal{I}}$. Under the full-matrix population gradient flow $\dot{\boldsymbol{\theta}} = -\nabla\mathcal{L}(\boldsymbol{\theta})$, if $\boldsymbol{\theta}_{\mathrm{out}}(0) = 0$, then $\boldsymbol{\theta}_{\mathrm{out}}(t) = 0$ for all $t \geqslant 0$ (Proposition B.2). Consequently, the training dynamics are entirely determined by the retained blocks. We refer to these parameter blocks as valid parameter blocks.

However, the resulting parameterization remains highly redundant, with additional degrees of freedom associated with $\mathbf{D}_\Phi$, $\mathbf{D}_U$, ρ, b, c, $\mathbf{L}$, and Ω. To isolate the essential parameter blocks, we define $\boldsymbol{\theta}_{\sup} := (\mathbf{A}_{\boldsymbol{\theta}}, \mathbf{L}, \mathbf{B}_-, \mathbf{B}_+, \mathbf{B}_r, \mathbf{C}) \in \Theta_{\sup}$. For any $\mathbf{L} \in \mathrm{GL}(d)$ and $c \neq 0$, let

$$\bar{\boldsymbol{\theta}}_\tau(\mathbf{L}, c) = \left(\tau^{-1} I_d,\ \mathbf{L},\ [0, -cI_d],\ [0, c\gamma \mathbf{L}^{-\top}],\ [c, 0],\ \frac{\alpha}{c} I_d\right). \tag{13}$$

These six blocks form an inclusion-minimal sufficient support within the specified architecture: they realize the target update in Theorem 1 on all admissible prompts, whereas no proper subset does so when every other block is fixed to zero (Proposition B.3). We call these six parameter blocks the support parameter blocks.

5.2 Empirical Analysis

Each MDP has $n_S = 9$ states and $n_A = m = 4$ actions, with discount $\gamma = 0.5$. For each MDP, the initial-state distribution P_0 and transition kernel $P(\cdot \mid s, a)$ are drawn independently from Dirichlet$(1, \ldots, 1)$ (uniform over the probability simplex on n_S states), and the rewards $r(a, s') \sim$ Unif$(-1, 1)$ are sampled i.i.d. per (a, s'). For each MDP, the state-action feature map $\phi\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ ($d = n_S n_A = 36$) and the initial weight $\mathbf{w} \in \mathbb{R}^d$ have entries drawn i.i.d. from Unif$(-1, 1)$. The behavior policy is ϵ-greedy ($\epsilon = 0.1$) with $Q(s, a) = \phi(s, a)^\top \mathbf{w}$, and the soft window Q-learning target $\mathbf{w}^{(\tau)}_{\text{soft}}$ is computed from the trajectory window with step size $\alpha = 0.2$.

The Transformer uses embedding dimension $D = 5d + 2 = 182$. The prompt supplies $\mathbf{w}$, current features ϕ_i, rewards r_{i+1}, and all next-state candidate features $\phi^+_{i,a}$. Tabular tests replace features by $\phi(s, a) = e_{(s-1)m+a}$ and are explicitly separated from dense-feature tests. Training is performed over 46.40 million tasks per model and five paired seeds. Adam uses cosine learning-rate decay from the first update; the default temperature, batch size and context length are $\tau = 0.2$, $B = 32$ and $n = 20$. By Proposition B.2, the sixteen valid parameter blocks are initialized using Xavier normal initialization with gain 0.1. All remaining entries of the two-layer attention construction are initialized to zero, as they remain fixed during training. We conduct the following experiments, with full details provided in Appendix C.1.

Experiments show that parameter blocks other than the sixteen valid parameter blocks stay at zero as training progresse, confirming Proposition B.2. We evaluate the learned update using three complementary metrics: Soft error measures how accurately the Transformer realizes the soft Q update, Hard error measures its discrepancy from standard Q-learning, and Greedy agreement measures how often it selects the same greedy action as Q-learning. The learned Transformer closely approximates the soft Q-learning update and achieves nearly 100% greedy agreement.

Figure 1 presents the final learned matrices $\mathbf{P}_1$, $\mathbf{V}_1$, $\mathbf{P}_2$, and $\mathbf{V}_2$ obtained after training. Remarkably, the learned matrices recover the parameter structure predicted in Theorem 1. In particular, the diagonal entries of $\mathbf{A}_\theta$ in $\mathbf{P}_1$ approach $1/\tau$, indicating that the Transformer has learned the internal action-selection computation required by the soft Q-learning update. This provides direct evidence of parameter emergence. To quantitatively assess parameter recovery, we examine several construction relations. The learned parameters closely approximate analytical relations, and the task-exposure curves show improved recovery as training progresses. See Appendix C.1 for details of the heatmaps.

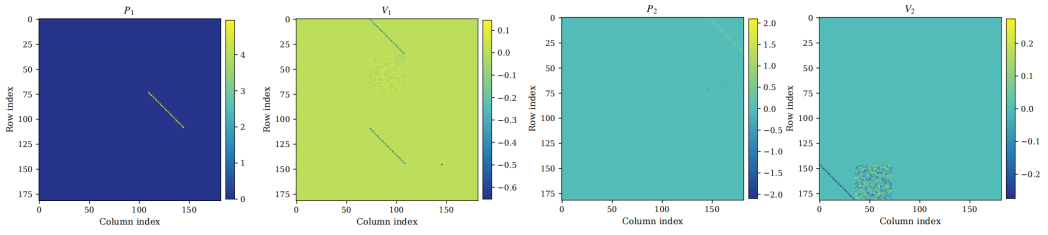


Figure 1: Final learned $\mathbf{P}_1$, $\mathbf{V}_1$, $\mathbf{P}_2$ and $\mathbf{V}_2$ matrices after training at $\tau = 0.2$, $B = 32$, and $n = 20$. the non-zero structure matches the theoretical constructions $(\mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2)$ in (9) and (11).

The left panel of Figure 2 deploys frozen dense-feature-trained Transformers on unseen tabular MDPs and measures exact policy returns under a common interaction budget. Comparisons include hard and

soft window Q-learning, a model-informed Oracle, and uniform Random. The Transformer achieves returns close to those of the soft window teacher and a higher sample-mean return than the hard window-Q baseline using the same fixed step size.

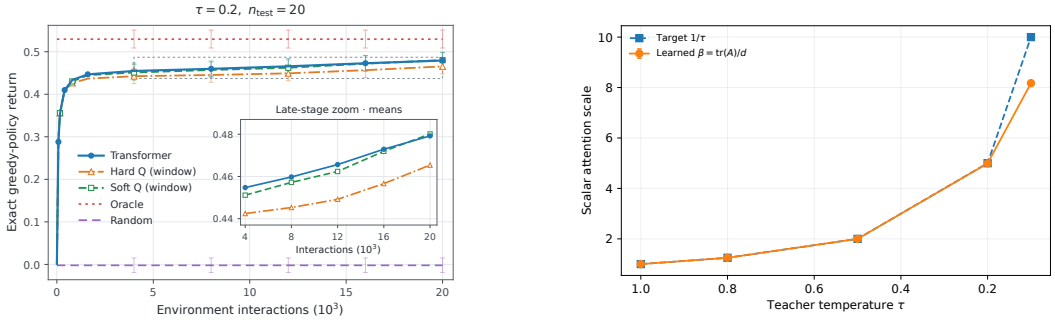


Figure 2: Left: Performance of Transformers on unseen tabular MDPs (sixteen valid parameter blocks). Right: The learned scalar projection β_{att} and its target $1/\tau$ (six support parameter blocks).

The effect of the action-selection operator. To separately study how temperature affects the emergence of parameters, we analyze the six-block parameterization, which retains the complete soft Q-learning update mechanism and is inclusion-minimal (Proposition B.3). All six blocks remain jointly trainable under output-only supervision. Training is performed over 20.48 million tasks per model. At a given temperature τ, the trained Transformer is shown to approximate the soft Q-learning update and produce emergent parameters, verifying Proposition B.3. Decreasing τ brings the soft Q update closer to standard Q-learning, but increases the approximation error between the Transformer and the soft update. With a fixed training budget, the recovery error of the predicted parameter structure increases gradually as τ decreases. The most pronounced discrepancy appears in the scalar attention scale $\beta_{\text{att}} := \text{tr}(\mathbf{A}_\theta)/d$. As shown in the right panel of Figure 2, the learned scale falls below its theoretical target $1/\tau$ at low temperatures. This implies that smaller τ requires greater task exposure for accurate parameter recovery. We further examine the effects of batch size and training-context length. See Appendix C.2 for details.

5.3 Theoretical Analysis

In this section, we provide a theoretical account of the above empirical findings by explaining how parameter emergence arises from the pretraining algorithm. Our primary focus is to investigate the effect of the action-selection mechanism on convergence. To this end, we restrict optimization to the support parameter blocks, reducing parameter redundancy and isolating the Transformer's Q-learning mechanism. Define the optimal parameter manifold as $\mathcal{M}_\tau := \{\bar{\boldsymbol{\theta}}_\tau(\mathbf{L}, c) : \mathbf{L} \in \text{GL}(d),\ c \neq 0\}$, we study whether deterministic gradient descent $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_k)$ converges to $\mathcal{M}_\tau$ and characterize how its convergence rate and iteration complexity depend on τ.

Let $\mathcal{V}$ be a vector space with $\dim \mathcal{V} = 4d^2 + 3d$ and $\zeta := (E_C, e_q, e_-, E_-, e_+, E_+, E_A) \in \mathcal{V}$, where the four E-symbols are $d \times d$ matrices and the three lowercase e-symbols are vectors in $\mathbb{R}^d$. Let the teacher-weighted covariance be $\Sigma_i^\tau = \sum_{a=1}^m \pi_{i,a}^\tau (\phi_{i,a}^+ - \bar{\phi}_i^{+,\tau})(\phi_{i,a}^+ - \bar{\phi}_i^{+,\tau})^\top \in \mathbb{R}^{d\times d}$ and

$$\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta] = \frac{1}{n}\sum_{i=0}^{n-1}\Big\{E_C\phi_i\delta_i^\tau + \phi_i\big[\phi_i^\top(e_- + E_-\mathbf{w}) + (\bar{\phi}_i^{+,\tau})^\top(e_+ + E_+\mathbf{w}) + r_{i+1}e_q^\top\mathbf{w} + \mathbf{w}^\top\Sigma_i^\tau E_A\mathbf{w}\big]\Big\}. \tag{14}$$

where $\delta_i^\tau(\mathbf{z}) = r_{i+1} - \mathbf{w}^\top\phi_i + \gamma\mathbf{w}^\top\bar{\phi}_i^{+,\tau}(\mathbf{z})$. In these coordinates, the adjoint $\mathcal{B}_{\tau,n}(\mathbf{z})^* : \mathbb{R}^d \to \mathcal{V}$ is represented by the transpose of $\mathcal{B}_{\tau,n}(\mathbf{z})$. We define the $\dim\mathcal{V} \times \dim\mathcal{V}$ data Gram matrix as $\mathsf{H}_n(\tau) = \mathbb{E}[\mathcal{B}_{\tau,n}(\mathbf{z})^*\mathcal{B}_{\tau,n}(\mathbf{z})]$, and denote its smallest eigenvalue by $\chi = \lambda_{\min}(\mathsf{H}_n(\tau))$.

We now state the richness assumptions on the training-MDP distribution

Assumption 1. Let Π denote the class of behavior policies used to generate training trajectories. There exist positive constants B_w, B_ϕ, B_r such that, for every $\pi \in \Pi$, $\mathcal{D}_n$ is independent of $\boldsymbol{\theta}$ and

(i) Boundedness. $\|\mathbf{w}\|_2 \leqslant B_w$, $\|\phi_i\|_2, \|\phi_{i,a}^+\|_2 \leqslant B_\phi$, $|r_{i+1}| \leqslant B_r$, $\mathcal{D}_n$-a.s.

(ii) Joint observable richness. For the fixed (n, τ), $\mathsf{H}_n(\tau) \succ 0$, meaning $\langle \zeta, \mathsf{H}_n(\tau)\zeta \rangle > 0$ for every $\zeta \neq 0$. Equivalently, $\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta] = 0$, $\mathcal{D}_n$-a.s. $\Longrightarrow \zeta = 0$.

Assumption 1 holds for a broad class of task populations with a nonzero absolutely continuous component and sufficient trajectory support, including the distribution used in our experiments. Explicit sufficient conditions and their verification are given in Appendix E.

Fix a nonempty relatively compact open set $\mathbf{\Omega} \Subset \mathrm{GL}(d) \times (\mathbb{R} \setminus \{0\})$. Its closure $\overline{\mathbf{\Omega}}$ is compact and remains inside $\mathrm{GL}(d) \times (\mathbb{R} \setminus \{0\})$. Define its images $\mathcal{U}_\tau = \{\bar{\boldsymbol{\theta}}_\tau(\mathbf{L}, c) : (\mathbf{L}, c) \in \mathbf{\Omega}\}$, $\overline{\mathcal{U}}_\tau = \{\bar{\boldsymbol{\theta}}_\tau(\mathbf{L}, c) : (\mathbf{L}, c) \in \overline{\mathbf{\Omega}}\}$. The set $\overline{\mathcal{U}}_\tau$ is compact at each fixed τ.

At a point $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$, the tangent space $T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ consists of derivatives at zero of smooth curves lying in $\mathcal{M}_\tau$ and passing through $\bar{\boldsymbol{\theta}}$. The normal space is its orthogonal complement, $N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau = (T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau)^\perp$, define the tubular neighborhood of $\mathcal{U}_\tau$ as

$$\mathcal{T} := \mathcal{T}_r(\mathcal{U}_\tau) = \{\bar{\boldsymbol{\theta}} + \Delta : \bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau,\ \Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau,\ \|\Delta\| < r\}. \tag{15}$$

A perturbation $\mathbf{U} \in \Theta_{\mathrm{sup}}$ has blocks denoted by $\Delta A, \ldots, \Delta C$ in the same order as (13). Let κ denote the minimum local curvature normal to $\mathcal{M}_\tau$, defined by

$$\kappa := \inf_{\bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau, \mathbf{U} \perp T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau, \|\mathbf{U}\|=1} \left\langle \mathbf{U}, \nabla^2 \mathcal{L}(\bar{\boldsymbol{\theta}})\mathbf{U} \right\rangle > 0. \tag{16}$$

Let $T_\eta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta \nabla \mathcal{L}(\boldsymbol{\theta})$ and $d(\boldsymbol{\theta}) := \mathrm{dist}(\boldsymbol{\theta}, \mathcal{M}_\tau)$, define

$$K_{\mathrm{GD}}(\varepsilon; \boldsymbol{\theta}_0) := \min\left\{K \in \mathbb{N}_0 : d\big(T_\eta^K(\boldsymbol{\theta}_0)\big) \leqslant \varepsilon d(\boldsymbol{\theta}_0)\right\}.$$

Theorem 2. *Under Assumption 1, there exist a neighborhood $\mathcal{T}$, a constant $M_J > 0$ such that, for any $\varepsilon \in (0, 1/2]$, $0 < \eta \leqslant \frac{1}{2M_J^2}$, and $\boldsymbol{\theta}_0 \in \mathcal{A}_\tau := \{\boldsymbol{\theta} \in \mathcal{T} : d(\boldsymbol{\theta}) > 0, \frac{4M_J}{\sqrt{\kappa}} d(\boldsymbol{\theta}) < \mathrm{dist}(\boldsymbol{\theta}, \mathcal{T}^c)\}$,*

(i) The gradient descent $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_k)$, converges to some $\boldsymbol{\theta}_\infty \in \mathcal{U}_\tau$, and $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_\infty\| \leqslant \frac{4M_J}{\sqrt{\kappa}} d(\boldsymbol{\theta}_0)(1 - \frac{1}{4}\eta\kappa)^{k/2}$, where $1 - \frac{1}{4}\eta\kappa \in (0, 1)$.

(ii) $K_{GD}(\varepsilon; \boldsymbol{\theta}_0) = O(\frac{1}{\eta\kappa} \log \frac{1}{\varepsilon})$ for every $\boldsymbol{\theta}_0 \in \mathcal{A}_\tau$, and $\sup_{\boldsymbol{\theta}_0 \in \mathcal{A}_\tau} K_{GD}(\varepsilon; \boldsymbol{\theta}_0) = \Theta_{\mathbf{\Omega}}(\frac{1}{\eta\kappa} \log \frac{1}{\varepsilon})$.

In particular, under our experimental setting, we obtain the following corollary.

Corollary 3. *Let feature table $\Phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, reward table $R : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and an input weight $\mathbf{w}$, $\mathcal{N} \subset \mathbb{N}$ be a nonempty finite set of context lengths. Suppose $n_S, d, m \geqslant 2$, P_0 and all transition rows are mutually independent $\mathrm{Dirichlet}(1, \ldots, 1)$, independent of $(\mathbf{w}, \Phi, R)$, whose entries are mutually independent $\mathrm{Unif}(-1, 1)$. The behavior policy is ϵ_b-greedy with uniform exploration, $0 < \varepsilon_b \leqslant 1$. There exists $\tau_* > 0$ such that, for $0 < \tau \leqslant \tau_*$, $n \in \mathcal{N}$ and η satisfying Theorem 2, $K_{GD}(\varepsilon; \boldsymbol{\theta}_0) = O(\frac{1}{\eta\tau^5} \log \frac{1}{\varepsilon})$ for every $\boldsymbol{\theta}_0 \in \mathcal{A}_\tau$, and $\sup_{\boldsymbol{\theta}_0 \in \mathcal{A}_\tau} K_{GD}(\varepsilon; \boldsymbol{\theta}_0, \tau, n) = \Theta_{\mathbf{\Omega}, \mathcal{N}}(\frac{1}{\eta\tau^5} \log \frac{1}{\varepsilon})$.*

The proofs of Theorem 2 and Corollary 3 are provided in Appendices D and E. Corollary 3 shows that as τ decreases, the local curvature diminishes, slowing convergence and requiring more iterations to achieve the same optimization accuracy, consistent with our empirical results.

6 Discussion and Limitations

This work takes a step toward white-boxing in-context reinforcement learning for off-policy control. We show that an attention-only Transformer can implement a finite-temperature Q-learning update in its forward pass and recover standard window Q-learning as the temperature approaches zero. We further show that the corresponding parameter structure can emerge through teacher-mimicking training, and establish local exponential convergence to a manifold of functionally equivalent soft Q-learning realizations, with particular attention to how action selection in soft Q-updates affects parameter emergence. These results provide a mechanistic account of how Q-learning can be represented and learned within Transformer computation.

Our analysis has several limitations. First, our theoretical analysis is restricted to the support parameter blocks because the valid parameter blocks exhibit substantial redundancy; extending the convergence analysis to all valid blocks remains future work. Second, we do not analyze minibatch SGD, which would be needed to characterize the dependence of convergence on minibatch size.

AI USE STATEMENT

Generative AI tools were used to assist with code review, experimental implementation and analysis, verification of mathematical derivations, and linguistic refinement of the manuscript. In Appendix E, AI tools contributed key proof steps that were adopted by the authors. The core ideas of the other proofs were developed by the authors, with AI assistance in checking mathematical arguments, carrying out partial algebraic calculations, and identifying and correcting errors. The authors checked the proofs step by step and executed and tested the AI-assisted programs. All numerical results reported in the paper were obtained from executable simulation or model-evaluation code and its saved outputs. All AI-assisted content was reviewed and verified by the authors, who retain full responsibility for the accuracy, validity, and final presentation of this work.

ETHICS STATEMENT

This work investigates the representation and learning of Q-learning updates in structured Transformers through mathematical analysis and experiments on synthetically generated Markov decision processes. The study does not involve human participants, personal or sensitive data, or real-world deployment. We do not identify specific ethical concerns arising from the reported experiments. The results apply under the stated architectural and mathematical assumptions and do not constitute a safety guarantee for real-world reinforcement-learning deployment.

REPRODUCIBILITY STATEMENT

The Transformer architecture, prompt construction, attention mask, and teacher-mimicking objective are specified in Sections 4 and 5 and Algorithm 1. The assumptions and proofs supporting the theoretical results are provided in Appendices A, B, D, and E. Section 5.2 and Appendix C describe the synthetic task generation, training configurations, evaluation protocols, and definitions of the reported metrics. Appendix C also reports results across five paired random seeds and describes the uncertainty estimates used in the experiments. An anonymized implementation is provided in the supplementary material.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Juncheng Dong, Moyang Guo, Ethan X. Fang, Zhuoran Yang, and Vahid Tarokh. In-context reinforcement learning from suboptimal historical data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14021–14039. PMLR, 2025.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=o8AaRKbP9K.

Jake Grigsby, Linxi Fan, and Yuke Zhu. AMAGO: Scalable in-context reinforcement learning for adaptive agents. In *International Conference on Learning Representations*, 2024.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1273–1286, 2021.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations*, 2023.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2013. doi: 10.1007/978-1-4419-9982-5.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haodong Liang and Lifeng Lai. Transformers provably implement in-context reinforcement learning with policy improvement. *arXiv preprint arXiv:2605.05755*, 2026.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, 2024.

Jinmei Liu, Fuhong Liu, Zhenhong Sun, Jianye Hao, Huaxiong Li, Bo Wang, Daoyi Dong, Chunlin Chen, and Zhi Wang. Scalable in-context q-learning. In *International Conference on Learning Representations*, 2026.

Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 664–671, 2008. doi: 10.1145/1390156.1390240.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jiuqi Wang, Ethan H Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers learn temporal difference methods for in-context reinforcement learning. In *ICML 2024 Workshop on In-Context Learning*, 2024.

Jiuqi Wang, Rohan Chandra, and Shangtong Zhang. Towards provable emergence of in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, 1989.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698.

Zixuan Xie, Xinyu Liu, Rohan Chandra, and Shangtong Zhang. Convergence and emergence of in-context reinforcement learning with chain of thought, 2026a. URL `https://arxiv.org/abs/2605.07123`.

Zixuan Xie, Xinyu Liu, Claire Chen, Shuze Daniel Liu, Rohan Chandra, and Shangtong Zhang. Beyond linear attention: Softmax transformers implement in-context reinforcement learning. *arXiv preprint arXiv:2605.07333*, 2026b.

Qiushui Xu, Yuhao Huang, Yushu Jiang, Wenliang Zheng, Lei Song, Jinyu Wang, and Jiang Bian. In-context compositional q-learning for offline reinforcement learning. In *International Conference on Learning Representations*, 2026.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

A TWO-LAYER TRANSFORMER FOR SOFT Q-LEARNING

Softmax weights at finite temperature τ are not exactly equivalent to $\max$. Recall (5). Lemma A.1 states that they can approximate $\max$ arbitrarily well, with the true hard-max limit recovered only as the softmax temperature $\tau \to 0$.

Lemma A.1. *For each i, define $q_{i,a} := \mathbf{w}^\top \boldsymbol{\phi}^+_{i,a}$, $q^\star_i := \max_{a\in\mathcal{A}} q_{i,a}$, and $\mathcal{A}^\star_i := \arg\max_{a\in\mathcal{A}} q_{i,a}$. Let $m := |\mathcal{A}| < \infty$ and $K_i := |\mathcal{A}^\star_i|$. Then, as $\tau \downarrow 0$, $\pi^\tau_{i,a} \to \frac{1}{K_i}\mathbf{1}\{a \in \mathcal{A}^\star_i\}$. If $\mathcal{A}^\star_i \neq \mathcal{A}$, define the action gap $\Delta_i := q^\star_i - \max_{a\notin\mathcal{A}^\star_i} q_{i,a} > 0$, then $1 - \sum_{a\in\mathcal{A}^\star_i} \pi^\tau_{i,a} \leqslant [(m-K_i)/K_i]e^{-\Delta_i/\tau}$.*

Proof of Lemma A.1. Fix i. Dividing the numerator and denominator of $\pi^\tau_{i,a}$ by $\exp(q^\star_i/\tau)$ gives

$$\pi^\tau_{i,a} = \frac{\exp(-(q^\star_i - q_{i,a})/\tau)}{\sum_{b\in\mathcal{A}} \exp(-(q^\star_i - q_{i,b})/\tau)}.$$

For every $a \in \mathcal{A}^\star_i$, we have $q^\star_i - q_{i,a} = 0$, whereas for $a \notin \mathcal{A}^\star_i$, $q^\star_i - q_{i,a} > 0$. Hence, as $\tau \to 0$, $\sum_{b\in\mathcal{A}} \exp(-(q^\star_i - q_{i,b})/\tau) \to K_i$, which yields $\pi^\tau_{i,a} \to 1/K_i$, for $a \in \mathcal{A}^\star_i$, $\pi^\tau_{i,a} \to 0$, for $a \notin \mathcal{A}^\star_i$. Therefore, $\sum_{a\in\mathcal{A}^\star_i} \pi^\tau_{i,a} \to 1$. Now suppose $\mathcal{A}^\star_i \neq \mathcal{A}$ and define $\Delta_i = q^\star_i - \max_{a\notin\mathcal{A}^\star_i} q_{i,a} > 0$. Then

$$1 - \sum_{a\in\mathcal{A}^\star_i} \pi^\tau_{i,a} = \frac{\sum_{a\notin\mathcal{A}^\star_i} e^{-(q^\star_i - q_{i,a})/\tau}}{K_i + \sum_{a\notin\mathcal{A}^\star_i} e^{-(q^\star_i - q_{i,a})/\tau}}.$$

Since $q^\star_i - q_{i,a} \geqslant \Delta_i$ for all $a \notin \mathcal{A}^\star_i$,

$$1 - \sum_{a\in\mathcal{A}^\star_i} \pi^\tau_{i,a} \leqslant \frac{m - K_i}{K_i} e^{-\Delta_i/\tau}.$$

When $K_i = 1$, this reduces to $1 - \pi^\tau_{i,a^\star_i} \leqslant (m-1)e^{-\Delta_i/\tau}$. □

We proceed to prove Theorem 1. Our Transformer architecture is two-layered, where the first layer adopts single-head softmax attention to produce Boltzmann-weighted aggregation across all candidate successor actions. To facilitate the proof, we first formally characterize the output behavior of this layer in Lemma A.2, before establishing the full result of Theorem 1.

We first recall the first-layer attention mask. Let T_i, $C_{i,a}$, and P_o denote the positions of t_i, $c_{i,a}$, and p, respectively. For each transition query T_i, only its associated candidate-action tokens are unmasked:

$$\mathbf{M}_{C_{j,a},T_i} = \begin{cases} 0, & j = i, \\ -\infty, & j \neq i, \end{cases} \qquad \mathbf{M}_{T_j,T_i} = \mathbf{M}_{P_o,T_i} = -\infty.$$

For each candidate-action query and for the parameter query, only the parameter token p is unmasked: $\mathbf{M}_{P_o,C_{i,a}} = 0$, $\mathbf{M}_{P_o,P_o} = 0$, and all remaining entries are set to $-\infty$. Hence, each transition token t_i attends exactly to its own candidate-action set $\{c_{i,a}\}_{a\in\mathcal{A}}$.

Lemma A.2 (First layer computes the Boltzmann-weighted successor feature)**.** *Consider the single-head masked softmax attention layer in* (8) *with mask* $\mathbf{M}$ *and parameters* $\mathbf{P}_1, \mathbf{V}_1$ *specified in* (9)*. Define*

$$\zeta_Q = (\mathbf{Q} - I_{d+1})\widetilde{\mathbf{w}} = \begin{bmatrix} \rho - 1 + b^\top \mathbf{w} \\ 0_d \end{bmatrix}, \qquad \widetilde{\mathbf{w}}_Q = \mathbf{Q}\widetilde{\mathbf{w}} = \begin{bmatrix} \rho + b^\top \mathbf{w} \\ \mathbf{w} \end{bmatrix}.$$

For the input prompt $H^{(0)}$ in (7)*, the output is*

$$\mathbf{H}^{(1)} = \big[t^{(1)}_0, \mathbf{C}^{(1)}_0, t^{(1)}_1, \mathbf{C}^{(1)}_1, \ldots, t^{(1)}_{n-1}, \mathbf{C}^{(1)}_{n-1}, p^{(1)}\big],$$

where $\mathbf{C}^{(1)}_i = [c^{(1)}_{i,a}]_{a\in\mathcal{A}}$ and

$$t^{(1)}_i = \begin{bmatrix} \phi_i + \mathbf{D}_\Phi \bar{\phi}^{+,\tau}_i \\ \mathbf{L}\bar{\phi}^{+,\tau}_i \\ r_{i+1} + \mathbf{D}_R \bar{\phi}^{+,\tau}_i \\ 0_d \\ \mathbf{w} + \mathbf{D}_U \bar{\phi}^{+,\tau}_i \\ 0_{d+1} \end{bmatrix}, \quad c^{(1)}_{i,a} = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ \phi^+_{i,a} \\ 0_d \\ \zeta_Q \end{bmatrix}, \quad p^{(1)} = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \widetilde{\mathbf{w}}_Q \end{bmatrix}. \tag{A.1}$$

Proof. Write $\mathbf{H}^{(0)} = [h_1, \ldots, h_N]$. Let $\bar{\mathbf{A}} := \mathrm{softmax}\left((\mathbf{H}^{(0)})^\top \mathbf{P}_1 \mathbf{H}^{(0)} + \mathbf{M}\right)$, where the softmax is applied columnwise over the key positions. Thus, if h_k and h_q denote the tokens at key position k and query position q, respectively, then

$$\bar{\mathbf{A}}_{k,q} = \frac{\exp(h_k^\top \mathbf{P}_1 h_q + \mathbf{M}_{k,q})}{\sum_{\ell=1}^N \exp(h_\ell^\top \mathbf{P}_1 h_q + \mathbf{M}_{\ell,q})}. \tag{A.2}$$

Since the kth column of $\mathbf{V}_1\mathbf{H}^{(0)}$ is $\mathbf{V}_1 h_k$, the qth column of the attention output satisfies $\left[(\mathbf{V}_1\mathbf{H}^{(0)})\bar{\mathbf{A}}\right]_{:,q} = \sum_{k=1}^N \bar{\mathbf{A}}_{k,q}\mathbf{V}_1 h_k$. Therefore, using the residual connection, the output corresponding to query token h_q is

$$h_q^{(1)} = h_q + \sum_{k=1}^N \bar{\mathbf{A}}_{k,q}\mathbf{V}_1 h_k. \tag{A.3}$$

We first consider a transition token t_i. By construction of the mask, the query t_i can attend only to its associated candidate-action tokens $\{c_{i,a}\}_{a\in\mathcal{A}}$. Hence, the only finite attention scores in the column corresponding to t_i are $c_{i,a}^\top \mathbf{P}_1 t_i$, $a \in \mathcal{A}$.

Recall that $t_i = [\phi_i; 0_d; r_{i+1}; 0_d; \mathbf{w}; 0_{d+1}]$, $c_{j,a} = [0_d; 0_d; 0; \phi_{j,a}^+; 0_d; 0_{d+1}]$. Since $c_{j,a}$ is nonzero only in its Ψ block, the inner product $c_{j,a}^\top \mathbf{P}_1 t_i$ depends only on the Ψ block of $P_1 t_i$. From the fourth block row of $\mathbf{P}_1$, $(\mathbf{P}_1 t_i)_\Psi = 0_{d\times d}\phi_i + \star 0_d + 0_{d\times 1} r_{i+1} + \star 0_d + \mathbf{A}_\theta \mathbf{w} + \star 0_{d+1}$.

All terms except the fifth vanish, so $(\mathbf{P}_1 t_i)_\Psi = \mathbf{A}_\theta \mathbf{w}$. Since $\mathbf{A}_\theta = \tau^{-1} I_d$, we obtain $(\mathbf{P}_1 t_i)_\Psi = \tau^{-1}\mathbf{w}$. Therefore, for any candidate token $c_{j,a}$, $c_{j,a}^\top \mathbf{P}_1 t_i = (\phi_{j,a}^+)^\top (\mathbf{P}_1 t_i)_\Psi = \frac{1}{\tau}\mathbf{w}^\top \phi_{j,a}^+$. The arbitrary $\star$ blocks in $\mathbf{P}_1$ do not affect the relevant attention scores. Although $\mathbf{P}_1 t_i$ may have nonzero components in other blocks, the candidate token $c_{j,a}$ has zeros in all of those coordinates and therefore removes them in the inner product.

The mask now restricts the query t_i to the candidates with the same transition index i. Hence, $M_{C_{i,a},T_i} = 0$, and $M_{k,T_i} = -\infty$ for every other key position k. It follows from (A.2) that

$$\bar{\mathbf{A}}_{C_{i,a},T_i} = \frac{\exp(c_{i,a}^\top \mathbf{P}_1 t_i)}{\sum_{b\in\mathcal{A}} \exp(c_{i,b}^\top \mathbf{P}_1 t_i)} = \frac{\exp(\mathbf{w}^\top \phi_{i,a}^+/\tau)}{\sum_{b\in\mathcal{A}} \exp(\mathbf{w}^\top \phi_{i,b}^+/\tau)} = \pi_{i,a}^\tau. \tag{A.4}$$

All other attention weights in the T_ith column are zero.

We next compute the corresponding value vectors. Since $c_{i,a}$ is nonzero only in its Ψ block, only the fourth block column of $\mathbf{V}_1$ contributes to $\mathbf{V}_1 c_{i,a}$. From the prescribed form of $\mathbf{V}_1$, this block column is

$$(\mathbf{V}_1)_{:,\Psi} = \begin{bmatrix} \mathbf{D}_\Phi \\ \mathbf{L} \\ \mathbf{D}_R \\ 0_{d\times d} \\ \mathbf{D}_U \\ 0_{(d+1)\times d} \end{bmatrix}, \quad \mathbf{V}_1 c_{i,a} = \begin{bmatrix} \mathbf{D}_\Phi \phi_{i,a}^+ \\ \mathbf{L}\phi_{i,a}^+ \\ \mathbf{D}_R \phi_{i,a}^+ \\ 0_d \\ \mathbf{D}_U \phi_{i,a}^+ \\ 0_{d+1} \end{bmatrix}$$

Hence, the attention contribution to the transition query t_i is

$$\begin{aligned} \left[(\mathbf{V}_1\mathbf{H}^{(0)})\bar{\mathbf{A}}\right]_{:,T_i} &= \sum_{a\in\mathcal{A}} \bar{\mathbf{A}}_{C_{i,a},T_i}\mathbf{V}_1 c_{i,a} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \mathbf{V}_1 c_{i,a} \\ &= \begin{bmatrix} \mathbf{D}_\Phi \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+ \\ \mathbf{L} \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+ \\ \mathbf{D}_R \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+ \\ 0_d \\ \mathbf{D}_U \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+ \\ 0_{d+1} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} \\ \mathbf{L}\bar{\phi}_i^{+,\tau} \\ \mathbf{D}_R \bar{\phi}_i^{+,\tau} \\ 0_d \\ \mathbf{D}_U \bar{\phi}_i^{+,\tau} \\ 0_{d+1} \end{bmatrix}, \end{aligned} \tag{A.5}$$

where we used $\bar{\phi}_i^{+,\tau} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+$ in the last equality.

Finally, applying the residual connection gives

$$t_i^{(1)} = t_i + \left[(\mathbf{V}_1\mathbf{H}^{(0)})\bar{\mathbf{A}}\right]_{:,T_i} = \begin{bmatrix} \phi_i \\ 0_d \\ r_{i+1} \\ 0_d \\ \mathbf{w} \\ 0_{d+1} \end{bmatrix} + \begin{bmatrix} \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} \\ \mathbf{L}\bar{\phi}_i^{+,\tau} \\ \mathbf{D}_R \bar{\phi}_i^{+,\tau} \\ 0_d \\ \mathbf{D}_U \bar{\phi}_i^{+,\tau} \\ 0_{d+1} \end{bmatrix} = \begin{bmatrix} \phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} \\ \mathbf{L}\bar{\phi}_i^{+,\tau} \\ r_{i+1} + \mathbf{D}_R \bar{\phi}_i^{+,\tau} \\ 0_d \\ \mathbf{w} + \mathbf{D}_U \bar{\phi}_i^{+,\tau} \\ 0_{d+1} \end{bmatrix}. \tag{A.6}$$

It remains to compute the candidate-action tokens and the parameter token. By the mask construction, each candidate-action query $c_{i,a}$ and the parameter query p can attend only to p. There is exactly one unmasked key in each of these columns, so

$$\bar{\mathbf{A}}_{P_o,C_{i,a}} = \frac{\exp(p^\top \mathbf{P}_1 c_{i,a})}{\exp(p^\top \mathbf{P}_1 c_{i,a})} = 1, \qquad \bar{\mathbf{A}}_{P_o,P_o} = \frac{\exp(p^\top \mathbf{P}_1 p)}{\exp(p^\top \mathbf{P}_1 p)} = 1.$$

All other attention weights in these columns are zero.

Recall that $p = [0_d; 0_d; 0; 0_d; 0_d; \widetilde{\mathbf{w}}]$. The last block column of $\mathbf{V}_1$, corresponding to the W block, is zero except for its last block $\mathbf{Q} - I_{d+1}$. Hence

$$\mathbf{V}_1 p = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ (\mathbf{Q} - I_{d+1})\widetilde{\mathbf{w}} \end{bmatrix} = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \zeta_Q \end{bmatrix},$$

where, explicitly, $(\mathbf{Q} - I_{d+1})\widetilde{\mathbf{w}} = \begin{bmatrix} \rho - 1 & b^\top \\ 0_{d\times 1} & 0_{d\times d} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \rho - 1 + b^\top \mathbf{w} \\ 0_d \end{bmatrix} = \zeta_Q$.

Therefore, the attention outputs for all candidate-action queries and for the parameter query are $\left[(\mathbf{V}_1\mathbf{H}^{(0)})\bar{\mathbf{A}}\right]_{:,C_{i,a}} = \mathbf{V}_1 p$, $\left[(\mathbf{V}_1\mathbf{H}^{(0)})\bar{\mathbf{A}}\right]_{:,P_o} = \mathbf{V}_1 p$. Using the residual connection once again,

$$c_{i,a}^{(1)} = c_{i,a} + \mathbf{V}_1 p = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ \phi_{i,a}^+ \\ 0_d \\ 0_{d+1} \end{bmatrix} + \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \zeta_Q \end{bmatrix} = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ \phi_{i,a}^+ \\ 0_d \\ \zeta_Q \end{bmatrix}, \quad p^{(1)} = p + \mathbf{V}_1 p = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \widetilde{\mathbf{w}} \end{bmatrix} + \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \zeta_Q \end{bmatrix} = \begin{bmatrix} 0_d \\ 0_d \\ 0 \\ 0_d \\ 0_d \\ \widetilde{\mathbf{w}}_Q \end{bmatrix},$$

where $\widetilde{\mathbf{w}} + \zeta_Q = \big(I_{d+1} + \mathbf{Q} - I_{d+1}\big)\widetilde{\mathbf{w}} = \mathbf{Q}\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}_Q$.

Thus, $\mathbf{H}^{(1)} = \left[t_0^{(1)}, \mathbf{C}_0^{(1)}, t_1^{(1)}, \mathbf{C}_1^{(1)}, \ldots, t_{n-1}^{(1)}, \mathbf{C}_{n-1}^{(1)}, p^{(1)}\right]$, with all three types of tokens as stated. In particular, the candidate-action tokens and the parameter token remain unchanged for every $\mathbf{w}$ precisely when $\rho = 1$ and $b = 0$, that is, when $\mathbf{Q} = I_{d+1}$. □

With the above lemma established, we proceed to prove Theorem 1.

Proof of Theorem 1. (i) By Lemma A.2, after the first attention layer, each transition token takes the form

$$t_i^{(1)} = \begin{bmatrix} \phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} \\ \mathbf{L}\bar{\phi}_i^{+,\tau} \\ r_{i+1} + \mathbf{D}_R \bar{\phi}_i^{+,\tau} \\ 0_d \\ \mathbf{w} + \mathbf{D}_U \bar{\phi}_i^{+,\tau} \\ 0_{d+1} \end{bmatrix}, \qquad \bar{\phi}_i^{+,\tau} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau \phi_{i,a}^+.$$

The candidate-action tokens and the parameter token have the forms $c_{i,a}^{(1)} = [0_d; 0_d; 0; \phi_{i,a}^+; 0_d; \zeta_Q]$, $p^{(1)} = [0_d; 0_d; 0; 0_d; 0_d; \widetilde{\mathbf{w}}_Q]$. In particular, the Φ and G blocks of these latter tokens are still zero, and the only possibly nonzero block of $p^{(1)}$ is its W block.

We now consider the second linear-attention layer. Write $\mathbf{H}^{(1)} = [h_1^{(1)}, \dots, h_N^{(1)}]$. From $\mathbf{H}^{(2)} = \mathbf{H}^{(1)} + \frac{1}{n}(\mathbf{V}_2\mathbf{H}^{(1)})\big((\mathbf{H}^{(1)})^\top \mathbf{P}_2 \mathbf{H}^{(1)}\big)$, the update to the parameter token is

$$\Delta p := p^{(2)} - p^{(1)} = \frac{1}{n}\sum_{j=1}^{N} \mathbf{V}_2 h_j^{(1)} \big((h_j^{(1)})^\top \mathbf{P}_2 p^{(1)}\big). \tag{A.7}$$

Indeed, this is simply the column of the matrix product corresponding to the query token $p^{(1)}$: the jth entry of the corresponding score column is $(h_j^{(1)})^\top \mathbf{P}_2 p^{(1)}$, and the jth value column is $\mathbf{V}_2 h_j^{(1)}$.

We only need the last d coordinates of Δp, since the readout extracts these coordinates of the parameter token. Recall $\Pi_{\text{out}} = [0_{d\times(4d+2)}, I_d]$, $\mathbf{C} = \frac{\alpha}{c} I_d$, $\mathbf{C}_G = -\mathbf{C}\mathbf{D}_\Phi \mathbf{L}^{-1}$. By the block structure of $\mathbf{V}_2$, for any token h,

$$\Pi_{\text{out}} \mathbf{V}_2 h = \mathbf{C} h_\Phi + \mathbf{C}_G h_G, \tag{A.8}$$

where h_Φ and h_G denote its Φ and G blocks. The arbitrary entries in the first $4d+2$ rows of $\mathbf{V}_2$ do not appear in this expression. Therefore,

$$\begin{aligned}
\Pi_{\text{out}} \mathbf{V}_2 t_i^{(1)} &= \mathbf{C}\big(\phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau}\big) + \mathbf{C}_G \mathbf{L} \bar{\phi}_i^{+,\tau} \\
&= \mathbf{C}\phi_i + \mathbf{C}\mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} - \mathbf{C}\mathbf{D}_\Phi \mathbf{L}^{-1}\mathbf{L}\bar{\phi}_i^{+,\tau} \\
&= \mathbf{C}\phi_i + \mathbf{C}\mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} - \mathbf{C}\mathbf{D}_\Phi \bar{\phi}_i^{+,\tau} = \mathbf{C}\phi_i = \frac{\alpha}{c}\phi_i.
\end{aligned} \tag{A.9}$$

Here the second equality uses $\mathbf{C}_G = -\mathbf{C}\mathbf{D}_\Phi \mathbf{L}^{-1}$, and the third uses $\mathbf{L}^{-1}\mathbf{L} = I_d$.

On the other hand, both $c_{i,a}^{(1)}$ and $p^{(1)}$ have zero Φ and G blocks. Applying (A.8) gives

$$\Pi_{\text{out}} \mathbf{V}_2 c_{i,a}^{(1)} = \mathbf{C} 0_d + \mathbf{C}_G 0_d = 0_d, \qquad \Pi_{\text{out}} \mathbf{V}_2 p^{(1)} = \mathbf{C} 0_d + \mathbf{C}_G 0_d = 0_d. \tag{A.10}$$

Thus, only the transition tokens contribute to the last d coordinates of the parameter update. The additional W block ζ_Q in each candidate token does not affect this conclusion, since the last d rows of the W block column of $\mathbf{V}_2$ are zero. No assumption that the candidate or parameter scores vanish is needed: their projected value vectors already vanish.

It remains to compute the scalar score $(t_i^{(1)})^\top \mathbf{P}_2 p^{(1)}$. Since $p^{(1)} = [0_d; 0_d; 0; 0_d; 0_d; \widetilde{\mathbf{w}}_Q]$, $\widetilde{\mathbf{w}}_Q = \mathbf{Q}\widetilde{\mathbf{w}} = \mathbf{Q}\begin{bmatrix}1\\ \mathbf{w}\end{bmatrix}$, only the last block column of $\mathbf{P}_2$ contributes to $\mathbf{P}_2 p^{(1)}$. Because $\mathbf{Q}$ is invertible, $\mathbf{Q}^{-1}\widetilde{\mathbf{w}}_Q = \mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}} = \begin{bmatrix}1\\ \mathbf{w}\end{bmatrix}$. From the definition of the specified blocks, we obtain

$$\mathbf{B}_- \widetilde{\mathbf{w}}_Q = -cJ\mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{w}} = -c[0_{d\times 1}, I_d]\begin{bmatrix}1\\ \mathbf{w}\end{bmatrix} = -c\mathbf{w}, \tag{A.11}$$

$$\begin{aligned}
\mathbf{B}_+ \widetilde{\mathbf{w}}_Q &= c\mathbf{L}^{-\top}\big[-\mathbf{D}_R^\top, \gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega\big]\mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{w}} \\
&= c\mathbf{L}^{-\top}\big[-\mathbf{D}_R^\top, \gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega\big]\begin{bmatrix}1\\ \mathbf{w}\end{bmatrix} \\
&= c\mathbf{L}^{-\top}\left[-\mathbf{D}_R^\top + (\gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega)\mathbf{w}\right],
\end{aligned} \tag{A.12}$$

$$\mathbf{B}_r \widetilde{\mathbf{w}}_Q = ce_0^\top \mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{w}} = c[1, 0_{1\times d}]\begin{bmatrix}1\\ \mathbf{w}\end{bmatrix} = c, \tag{A.13}$$

$$\mathbf{B}_U \widetilde{\mathbf{w}}_Q = c\Omega J\mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{w}} = c\Omega[0_{d\times 1}, I_d]\begin{bmatrix}1\\ \mathbf{w}\end{bmatrix} = c\Omega\mathbf{w}. \tag{A.14}$$

The fourth and sixth block rows of $\mathbf{P}_2 p^{(1)}$ may contain arbitrary terms arising from the corresponding $\star$ blocks. However, the Ψ and W blocks of $t_i^{(1)}$ are both zero, so these terms do not contribute to the bilinear score. Consequently,

$$\begin{aligned}
(t_i^{(1)})^\top \mathbf{P}_2 p^{(1)} =& \big(\phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau}\big)^\top \mathbf{B}_- \widetilde{\mathbf{w}}_Q + \big(\mathbf{L}\bar{\phi}_i^{+,\tau}\big)^\top \mathbf{B}_+ \widetilde{\mathbf{w}}_Q + \big(r_{i+1} + \mathbf{D}_R \bar{\phi}_i^{+,\tau}\big)\mathbf{B}_r \widetilde{\mathbf{w}}_Q \\
&+ \big(\mathbf{w} + \mathbf{D}_U \bar{\phi}_i^{+,\tau}\big)^\top \mathbf{B}_U \widetilde{\mathbf{w}}_Q \\
=& -c\big(\phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau}\big)^\top \mathbf{w} + c\big(\mathbf{L}\bar{\phi}_i^{+,\tau}\big)^\top \mathbf{L}^{-\top}\left[-\mathbf{D}_R^\top + (\gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega)\mathbf{w}\right] \\
&+ c\big(r_{i+1} + \mathbf{D}_R \bar{\phi}_i^{+,\tau}\big) + c\big(\mathbf{w} + \mathbf{D}_U \bar{\phi}_i^{+,\tau}\big)^\top \Omega\mathbf{w}.
\end{aligned} \tag{A.15}$$

We simplify these terms one by one. First, $-c\big(\phi_i + \mathbf{D}_\Phi \bar{\phi}_i^{+,\tau}\big)^\top \mathbf{w} = -c\phi_i^\top \mathbf{w} - c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_\Phi^\top \mathbf{w}$. Next, since $\mathbf{L}^\top \mathbf{L}^{-\top} = I_d$,

$$\begin{aligned} &c\big(\mathbf{L}\bar{\phi}_i^{+,\tau}\big)^\top \mathbf{L}^{-\top}\left[-\mathbf{D}_R^\top + (\gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega)\mathbf{w}\right] \\ &\quad = c(\bar{\phi}_i^{+,\tau})^\top \mathbf{L}^\top \mathbf{L}^{-\top}\left[-\mathbf{D}_R^\top + (\gamma I_d + \mathbf{D}_\Phi^\top - \mathbf{D}_U^\top \Omega)\mathbf{w}\right] \\ &\quad = -c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_R^\top + c\gamma(\bar{\phi}_i^{+,\tau})^\top \mathbf{w} + c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_\Phi^\top \mathbf{w} - c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_U^\top \Omega \mathbf{w}. \end{aligned}$$

The remaining two terms satisfy

$$\begin{aligned} c\big(r_{i+1} + \mathbf{D}_R \bar{\phi}_i^{+,\tau}\big) &= c r_{i+1} + c\mathbf{D}_R \bar{\phi}_i^{+,\tau}, \\ c\big(\mathbf{w} + \mathbf{D}_U \bar{\phi}_i^{+,\tau}\big)^\top \Omega \mathbf{w} &= c\mathbf{w}^\top \Omega \mathbf{w} + c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_U^\top \Omega \mathbf{w}. \end{aligned}$$

The terms involving $\mathbf{D}_\Phi$, $\mathbf{D}_R$, and $\mathbf{D}_U$ therefore cancel in pairs:

$$\begin{aligned} &- c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_\Phi^\top \mathbf{w} + c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_\Phi^\top \mathbf{w} = 0, \\ &- c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_R^\top + c\mathbf{D}_R \bar{\phi}_i^{+,\tau} = 0, \\ &- c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_U^\top \Omega \mathbf{w} + c(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_U^\top \Omega \mathbf{w} = 0. \end{aligned}$$

For the second cancellation, we used the fact that $\mathbf{D}_R \bar{\phi}_i^{+,\tau}$ is a scalar, so it equals its transpose $(\bar{\phi}_i^{+,\tau})^\top \mathbf{D}_R^\top$. Moreover, since $\Omega^\top = -\Omega$,

$$\mathbf{w}^\top \Omega \mathbf{w} = (\mathbf{w}^\top \Omega \mathbf{w})^\top = \mathbf{w}^\top \Omega^\top \mathbf{w} = -\mathbf{w}^\top \Omega \mathbf{w},$$

and therefore $\mathbf{w}^\top \Omega \mathbf{w} = 0$. Substituting these identities into (A.15) yields

$$\begin{aligned} (t_i^{(1)})^\top \mathbf{P}_2 p^{(1)} &= -c\phi_i^\top \mathbf{w} + c\gamma(\bar{\phi}_i^{+,\tau})^\top \mathbf{w} + c r_{i+1} \\ &= c\big(r_{i+1} + \gamma \mathbf{w}^\top \bar{\phi}_i^{+,\tau} - \mathbf{w}^\top \phi_i\big). \end{aligned} \tag{A.16}$$

Define $\delta_i^{(\tau)} := r_{i+1} + \gamma \mathbf{w}^\top \bar{\phi}_i^{+,\tau} - \mathbf{w}^\top \phi_i$. Then $(t_i^{(1)})^\top \mathbf{P}_2 p^{(1)} = c\delta_i^{(\tau)}$.

Therefore, the last d coordinates of the second-layer update are

$$\begin{aligned} \Pi_{\text{out}} \Delta p &= \frac{1}{n}\sum_{j=1}^{N} \Pi_{\text{out}} \mathbf{V}_2 h_j^{(1)} \big((h_j^{(1)})^\top \mathbf{P}_2 p^{(1)}\big) = \frac{1}{n}\sum_{i=0}^{n-1} \Pi_{\text{out}} \mathbf{V}_2 t_i^{(1)} \big((t_i^{(1)})^\top \mathbf{P}_2 p^{(1)}\big) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} \left(\frac{\alpha}{c}\phi_i\right) c\delta_i^{(\tau)} = \frac{\alpha}{n}\sum_{i=0}^{n-1} \phi_i \delta_i^{(\tau)}. \end{aligned} \tag{A.17}$$

The second equality follows from (A.10), and the third uses (A.9) and (A.16).

Since the W block of the first-layer parameter token is $\widetilde{\mathbf{w}}_Q = [\rho + b^\top \mathbf{w}; \mathbf{w}]$, its last d coordinates are

$$\Pi_{\text{out}} p^{(1)} = J\widetilde{\mathbf{w}}_Q = J\mathbf{Q}\widetilde{\mathbf{w}} = J\widetilde{\mathbf{w}} = [0_{d\times 1}, I_d] \begin{bmatrix} 1 \\ \mathbf{w} \end{bmatrix} = \mathbf{w}.$$

Here $J\mathbf{Q} = J$ follows directly from the prescribed form of $\mathbf{Q}$. The residual connection thus gives

$$\begin{aligned} \mathrm{TF}_\tau(\mathbf{H}^{(0)}) = \Pi_{\text{out}} p^{(2)} &= \Pi_{\text{out}} p^{(1)} + \Pi_{\text{out}} \Delta p = \mathbf{w} + \frac{\alpha}{n}\sum_{i=0}^{n-1} \phi_i \delta_i^{(\tau)} \\ &= \mathbf{w} + \frac{\alpha}{n}\sum_{i=0}^{n-1} \phi_i \left[r_{i+1} + \gamma \mathbf{w}^\top \bar{\phi}_i^{+,\tau} - \mathbf{w}^\top \phi_i\right]. \end{aligned}$$

The readout extracts precisely these last d coordinates of $p^{(2)}$. Hence we have proved part (i). Notice that this argument does not prescribe the first coordinate of the W block of $p^{(2)}$; that coordinate is outside the readout.

(ii) Recall that $q_{i,a} := \mathbf{w}^\top \phi_{i,a}^+$, $q_i^\star := \max_{a\in\mathcal{A}} q_{i,a}$, $\mathcal{A}_i^\star := \arg\max_{a\in\mathcal{A}} q_{i,a}$, $K_i := |\mathcal{A}_i^\star|$. The soft backup value is $\bar{q}_i^{(\tau)} := \mathbf{w}^\top \bar{\phi}_i^{+,\tau} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau q_{i,a}$. Since $\sum_{a\in\mathcal{A}} \pi_{i,a}^\tau = 1$ and $q_{i,a} = q_i^\star$ for every $a \in \mathcal{A}_i^\star$,

$$q_i^\star - \bar{q}_i^{(\tau)} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau q_i^\star - \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau q_{i,a} = \sum_{a\in\mathcal{A}} \pi_{i,a}^\tau (q_i^\star - q_{i,a}) = \sum_{a\notin\mathcal{A}_i^\star} \pi_{i,a}^\tau (q_i^\star - q_{i,a}). \tag{A.18}$$

In particular, $q_i^\star - \bar{q}_i^{(\tau)} \geqslant 0$. Since $|q_{i,a}| \leqslant B_q$ and $|q_i^\star| \leqslant B_q$, we have $0 \leqslant q_i^\star - q_{i,a} \leqslant |q_i^\star| + |q_{i,a}| \leqslant 2B_q$. Therefore, $0 \leqslant q_i^\star - \bar{q}_i^{(\tau)} \leqslant 2B_q \sum_{a \notin \mathcal{A}_i^\star} \pi_{i,a}^\tau$. If $\mathcal{A}_i^\star \neq \mathcal{A}$, Lemma A.1 gives

$$\sum_{a \notin \mathcal{A}_i^\star} \pi_{i,a}^\tau = 1 - \sum_{a \in \mathcal{A}_i^\star} \pi_{i,a}^\tau \leqslant \frac{m - K_i}{K_i} e^{-\Delta_i/\tau}. \tag{A.19}$$

Hence

$$0 \leqslant q_i^\star - \bar{q}_i^{(\tau)} \leqslant 2B_q \frac{m - K_i}{K_i} e^{-\Delta_i/\tau}. \tag{A.20}$$

If $\mathcal{A}_i^\star = \mathcal{A}$, then all actions have the same value and $q_i^\star - \bar{q}_i^{(\tau)} = 0$ for every $\tau > 0$.

The difference between the hard Q-learning update and the Transformer update is therefore

$$\mathbf{w}_Q - \mathbf{w}_{\mathrm{TF}}^{(\tau)} = \frac{\alpha}{n} \sum_{i=0}^{n-1} \phi_i \left[(r_{i+1} + \gamma q_i^\star - \mathbf{w}^\top \phi_i) - (r_{i+1} + \gamma \bar{q}_i^{(\tau)} - \mathbf{w}^\top \phi_i) \right] = \frac{\alpha\gamma}{n} \sum_{i=0}^{n-1} \phi_i \big(q_i^\star - \bar{q}_i^{(\tau)} \big).$$

By the triangle inequality and $\|\phi_i\|_2 \leqslant B_\phi$,

$$\begin{aligned} \big\| \mathbf{w}_Q - \mathbf{w}_{\mathrm{TF}}^{(\tau)} \big\|_2 &\leqslant \frac{\alpha\gamma}{n} \sum_{i=0}^{n-1} \|\phi_i\|_2 \big| q_i^\star - \bar{q}_i^{(\tau)} \big| \leqslant \frac{\alpha\gamma B_\phi}{n} \sum_{i:\, \mathcal{A}_i^\star \neq \mathcal{A}} \big(q_i^\star - \bar{q}_i^{(\tau)} \big) \\ &\leqslant \frac{2\alpha\gamma B_q B_\phi}{n} \sum_{i:\, \mathcal{A}_i^\star \neq \mathcal{A}} \frac{m - K_i}{K_i} e^{-\Delta_i/\tau}. \end{aligned} \tag{A.21}$$

The second inequality uses the nonnegativity of the backup gap and the fact that this gap is zero whenever $\mathcal{A}_i^\star = \mathcal{A}$; the last inequality uses (A.20).

Since the number of transitions is finite and each $\Delta_i > 0$ in the displayed sum, the right-hand side converges to zero as $\tau \downarrow 0$. Thus $\lim_{\tau \downarrow 0} \mathbf{w}_{\mathrm{TF}}^{(\tau)} = \mathbf{w}_Q$. In particular, if $m \geqslant 2$ and every maximizer is unique, then $K_i = 1$ and $1 - \pi_{i,a_i^\star}^\tau \leqslant (m-1) e^{-\Delta_i/\tau}$. Since there are finitely many transitions, $\Delta_{\min} := \min_{0 \leqslant i \leqslant n-1} \Delta_i > 0$. Consequently, $e^{-\Delta_i/\tau} \leqslant e^{-\Delta_{\min}/\tau}$, and hence

$$\begin{aligned} \big\| \mathbf{w}_Q - \mathbf{w}_{\mathrm{TF}}^{(\tau)} \big\|_2 &\leqslant \frac{2\alpha\gamma B_q B_\phi (m-1)}{n} \sum_{i=0}^{n-1} e^{-\Delta_i/\tau} \leqslant \frac{2\alpha\gamma B_q B_\phi (m-1)}{n} n e^{-\Delta_{\min}/\tau} \\ &= 2\alpha\gamma B_q B_\phi (m-1) e^{-\Delta_{\min}/\tau}. \end{aligned}$$

If $m = 1$, all backup gaps are zero and $\mathbf{w}_{\mathrm{TF}}^{(\tau)} = \mathbf{w}_Q$ for every $\tau > 0$. This proves the claimed exponential approximation bound. □

We next consider whether emergence of tabular Q-learning algorithms can be achieved within Transformers. Let $e_{(s,a)}$ denote the one-hot basis vector for state-action pair (s, a).

Corollary A.3 (Transformer implements tabular Q-learning)**.** *Under the conditions of the preceding theorem, let $\mathcal{S}$ and $\mathcal{A}$ be finite, and consider $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\mathbf{w}_{(s,a)} = Q(s,a)$. The two-layer Transformer constructed in the preceding theorem implements the batch tabular Q-learning update up to softmax approximation error.*

Proof. According to Theorem 1, the batch-update formula using n samples is

$$\mathrm{TF}_\tau(\mathbf{H}^{(0)}; \mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2) = \mathbf{w} + \frac{\alpha}{n} \sum_{i=0}^{n-1} e_{(s_i,a_i)} \Big[r_{i+1} + \gamma \sum_{b \in \mathcal{A}} \pi_{i,b}^\tau \mathbf{w}_{(s_{i+1},b)} - \mathbf{w}_{(s_i,a_i)} \Big],$$

where $\pi_{i,a}^\tau := \exp\big(Q(s_{i+1},a)/\tau\big) / \sum_{b \in \mathcal{A}} \exp\big(Q(s_{i+1},b)/\tau\big)$. Only the weight components corresponding to visited (s_i, a_i) samples are updated; all remaining coordinates are multiplied by zero and thus remain unchanged.

The equivalent update rule for the Q-values reads

$$[\mathrm{TF}_\tau(\mathbf{H}^{(0)}; \mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2)]_{(s,a)} = Q(s,a) + \frac{\alpha}{n} \sum_{i:(s_i,a_i)=(s,a)} \big[r_{i+1} + \gamma \sum_{b \in \mathcal{A}} \pi_{i,b}^\tau Q(s_{i+1},b) - Q(s_i,a_i) \big].$$

Notably, only state-action pairs (s, a) appearing within the mini-batch receive updates; all other entries stay fixed. □

B TRAINING ALGORITHM

B.1 INDEPENDENT-CONTEXT GRADIENT VARIANCE

Let $\ell(\boldsymbol{\theta}; \mathbf{z}) = \frac{1}{2}\|\hat{u}_\theta(\mathbf{z}) - u_\tau(\mathbf{z})\|_2^2$, $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\sim \mathcal{D}_n}[\ell(\boldsymbol{\theta}; \mathbf{z})]$. Fix $\boldsymbol{\theta}$, and let $\mathbf{z}_1, \ldots, \mathbf{z}_B \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ be independent of $\boldsymbol{\theta}$. Define $g(\boldsymbol{\theta}; \mathbf{z}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{z})$, $\hat{g}_B(\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^B g(\boldsymbol{\theta}; \mathbf{z}_b)$.

Proposition B.1 (Independent-context minibatch gradient)**.** *Assume $\mathbb{E}\|g(\boldsymbol{\theta}; \mathbf{z})\|_2^2 < \infty$ and that differentiation and expectation can be interchanged. Then*

$$\mathbb{E}[\hat{g}_B(\boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \nabla \mathcal{L}(\boldsymbol{\theta}), \tag{B.1}$$

$$\mathrm{Cov}(\hat{g}_B(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = \frac{1}{B}\mathrm{Cov}(g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta}), \tag{B.2}$$

$$\mathbb{E}\Big[\|\hat{g}_B(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \Big| \boldsymbol{\theta}\Big] = \frac{1}{B}\mathbb{E}\Big[\|g(\boldsymbol{\theta}; \mathbf{z}) - \nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \Big| \boldsymbol{\theta}\Big]. \tag{B.3}$$

Equivalently, for every deterministic direction v, $\mathrm{Var}(v^\top \hat{g}_B(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = \frac{1}{B}\mathrm{Var}(v^\top g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta})$. Thus independent-context minibatching preserves the population-gradient direction in expectation while reducing context-specific gradient fluctuations by a factor $1/B$.

Proof. Let $\mu(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta}]$. By the interchange of differentiation and expectation,

$$\mu(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{z})] = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(\boldsymbol{\theta}; \mathbf{z})] = \nabla \mathcal{L}(\boldsymbol{\theta}).$$

Therefore, $\mathbb{E}[\hat{g}_B(\boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \frac{1}{B}\sum_{b=1}^B \mathbb{E}[g(\boldsymbol{\theta}; \mathbf{z}_b) \mid \boldsymbol{\theta}] = \mu(\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$, which proves (B.1). Now define the centered gradients $\xi_b := g(\boldsymbol{\theta}; \mathbf{z}_b) - \mu(\boldsymbol{\theta})$. Conditioned on $\boldsymbol{\theta}$, the random vectors $\xi_1, \ldots, \xi_B$ are independent and identically distributed, with $\mathbb{E}[\xi_b \mid \boldsymbol{\theta}] = 0$. Moreover, $\hat{g}_B(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{B}\sum_{b=1}^B \xi_b$. Hence

$$\mathrm{Cov}(\hat{g}_B \mid \boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{1}{B}\sum_{b=1}^B \xi_b\right)\left(\frac{1}{B}\sum_{b'=1}^B \xi_{b'}\right)^\top \middle| \boldsymbol{\theta}\right] = \frac{1}{B^2}\sum_{b=1}^B \mathbb{E}[\xi_b \xi_b^\top \mid \boldsymbol{\theta}] + \frac{1}{B^2}\sum_{b\neq b'} \mathbb{E}[\xi_b \xi_{b'}^\top \mid \boldsymbol{\theta}].$$

For $b \neq b'$, conditional independence and centering give $\mathbb{E}[\xi_b \xi_{b'}^\top \mid \boldsymbol{\theta}] = \mathbb{E}[\xi_b \mid \boldsymbol{\theta}]\mathbb{E}[\xi_{b'}^\top \mid \boldsymbol{\theta}] = 0$. Thus $\mathrm{Cov}(\hat{g}_B \mid \boldsymbol{\theta}) = \frac{B}{B^2}\mathrm{Cov}(g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta}) = \frac{1}{B}\mathrm{Cov}(g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta})$, which proves (B.2). Taking the trace of the covariance identity yields $\mathbb{E}\left[\|\hat{g}_B - \nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \middle| \boldsymbol{\theta}\right] = \frac{1}{B}\mathbb{E}\left[\|g(\boldsymbol{\theta}; \mathbf{z}) - \nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 \middle| \boldsymbol{\theta}\right]$, which is (B.3). Finally, for any deterministic v,

$$\mathrm{Var}(v^\top \hat{g}_B \mid \boldsymbol{\theta}) = v^\top \mathrm{Cov}(\hat{g}_B \mid \boldsymbol{\theta}) v = \frac{1}{B} v^\top \mathrm{Cov}(g(\boldsymbol{\theta}; \mathbf{z}) \mid \boldsymbol{\theta}) v,$$

giving the directional variance identity. □

B.2 INVARIANT PARAMETER GROUPS

Proposition B.2 (Invariant parameter groups)**.** *Assume that the context distribution $\mathcal{D}_n$ is independent of θ and satisfies the boundedness condition in Assumption 1 (i).Let $\boldsymbol{\theta}_{\mathrm{out}}$ collect all entries of $\boldsymbol{\theta} = (\mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2)$ outside the following 16 blocks: $\mathbf{P}_{1,\Psi\mathcal{K}}$, $\mathbf{V}_{1,\mathcal{I}\Psi}$, $\mathbf{V}_{1,WW}$, $\mathbf{P}_{2,\mathcal{I}W}$, $\mathbf{V}_{2,W_{\mathrm{out}}\mathcal{I}}$. Under the full-matrix population gradient flow $\dot{\boldsymbol{\theta}} = -\nabla \mathcal{L}(\boldsymbol{\theta})$, if $\boldsymbol{\theta}_{\mathrm{out}}(0) = 0$, then $\boldsymbol{\theta}_{\mathrm{out}}(t) = 0$ for all $t \geqslant 0$. The minibatch SGD is entirely determined by the retained blocks.*

Proof. Let $\mathcal{S}_{16} = \{\boldsymbol{\theta} : \boldsymbol{\theta}_{\mathrm{out}} = 0\}$. We will prove the pointwise identity

$$\nabla_{\boldsymbol{\theta}_{\mathrm{out}}} \ell(\boldsymbol{\theta}; \mathbf{z}) = 0 \qquad \text{for every context } \mathbf{z} \text{ and every } \boldsymbol{\theta} \in \mathcal{S}_{16}, \tag{B.4}$$

where $\ell(\boldsymbol{\theta}; \mathbf{z}) = \frac{1}{2}\|f_{\boldsymbol{\theta}}(\mathbf{z}) - y_\tau(\mathbf{z})\|_2^2$, $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}_n}\ell(\boldsymbol{\theta}; \mathbf{z})$,$y_\tau(\mathbf{z}) = \mathbf{w}_{\mathrm{soft}}^{(\tau)}(\mathbf{z})$. Here the context and teacher target are fixed when differentiating with respect to $\boldsymbol{\theta}$. Throughout the proof, derivatives are taken in the full space of the four matrices. The condition $\boldsymbol{\theta} \in \mathcal{S}_{16}$ is imposed only when evaluating these derivatives. In particular, we do not infer a zero normal derivative merely from a formula for the output restricted to $\mathcal{S}_{16}$.

Fix $\mathbf{z}$ and $\boldsymbol{\theta} \in \mathcal{S}_{16}$. Recall the input tokens t_i, $c_{i,a}$, p. Thus t_i, $c_{i,a}$, and p are supported in $\mathcal{K}$, Ψ, and W, respectively. Here a vector being supported in a coordinate set means that all of its coordinates outside that set are zero. The first-layer mask allows t_i to attend only to $\{c_{i,a} : a \in \mathcal{A}\}$. Write

$$u_{i,a} = c_{i,a}^\top \mathbf{P}_1 t_i, \qquad \pi_{i,a} = \frac{e^{u_{i,a}}}{\sum_{b\in\mathcal{A}} e^{u_{i,b}}}, \qquad \bar{c}_i = \sum_{a\in\mathcal{A}} \pi_{i,a} c_{i,a}. \tag{B.5}$$

The probabilities $\pi_{i,a}$ are the student's current attention weights; no teacher parameter value is imposed on $\mathbf{P}_1$. Since each $c_{i,a}$ is supported in Ψ, so is $\bar{c}_i$. Every candidate query and the parameter query have exactly one allowed key, namely p. Their attention weights are identically one, for arbitrary full matrices $\mathbf{P}_1$ and $\mathbf{V}_1$. Consequently, the first-layer tokens satisfy

$$h_i = t_i + \mathbf{V}_1 \bar{c}_i, \qquad \widehat{c}_{i,a} = c_{i,a} + \mathbf{V}_1 p, \qquad q = p + \mathbf{V}_1 p. \tag{B.6}$$

On $\mathcal{S}_{16}$, the only possibly nonzero blocks of $\mathbf{V}_1$ are $\mathbf{V}_{1,\mathcal{I}\Psi}$ and $\mathbf{V}_{1,WW}$. Therefore $\operatorname{supp}(\mathbf{V}_1\bar{c}_i) \subseteq \mathcal{I}$, $\operatorname{supp}(\mathbf{V}_1 p) \subseteq W$. Because $\mathcal{K} \subseteq \mathcal{I}$ and the coordinate sets $\mathcal{I}, \Psi, W$ are pairwise disjoint, (B.6) gives

$$\operatorname{supp}(h_i) \subseteq \mathcal{I}, \qquad \operatorname{supp}(\widehat{c}_{i,a}) \subseteq \Psi \cup W, \qquad \operatorname{supp}(q) \subseteq W. \tag{B.7}$$

Neither $q = p$ nor $\widehat{c}_{i,a} = c_{i,a}$ is required. In particular, $\mathbf{V}_{1,WW}$ can be any $(d+1)\times(d+1)$ matrix.

Let $\Pi = [0_{d\times(4d+2)}, I_d]$, $\widetilde{\mathbf{V}}_2 = \Pi\mathbf{V}_2$. Write $h_1, \ldots, h_N$ for all first-layer tokens, in their original order, with $h_N = q$. For arbitrary full matrices, the second-layer readout is

$$f_{\boldsymbol{\theta}}(\mathbf{z}) = \Pi q + \frac{1}{n}\sum_{j=1}^{N} (\widetilde{\mathbf{V}}_2 h_j)(h_j^\top \mathbf{P}_2 q). \tag{B.8}$$

To compute its full differential, define $v_j = \widetilde{\mathbf{V}}_2 h_j$ and $\sigma_j = h_j^\top \mathbf{P}_2 q$. The product rule gives

$$d(v_j\sigma_j) = \sigma_j\big(\Pi\, d\mathbf{V}_2\, h_j + \widetilde{\mathbf{V}}_2\, dh_j\big) + v_j\big((dh_j)^\top \mathbf{P}_2 q + h_j^\top (d\mathbf{P}_2) q + h_j^\top \mathbf{P}_2\, dq\big). \tag{B.9}$$

For $j = N$, one has $h_N = q$ and $dh_N = dq$; thus this formula includes the parameter token's key, value, and query roles.

We now evaluate this differential at $\boldsymbol{\theta} \in \mathcal{S}_{16}$. Since $\mathbf{V}_2$ has nonzero entries only in $(W_{\mathrm{out}}, \mathcal{I})$, $\widetilde{\mathbf{V}}_2$ has nonzero columns only in $\mathcal{I}$. Similarly, $\mathbf{P}_2$ has nonzero entries only in $(\mathcal{I}, W)$, and q is supported in W, so $\operatorname{supp}(\mathbf{P}_2 q) \subseteq \mathcal{I}$. For every candidate token and for the parameter token considered as a key/value token, (B.7) therefore implies

$$v_j = \widetilde{\mathbf{V}}_2 h_j = 0_d, \qquad \sigma_j = h_j^\top \mathbf{P}_2 q = 0. \tag{B.10}$$

Substituting (B.10) into (B.9), every term vanishes:

$$d(v_j\sigma_j) = 0 \quad \text{for every full-matrix perturbation } d\boldsymbol{\theta}.$$

This step uses the vanishing of *both* factors. Vanishing of the summand alone would not suffice to conclude that its derivative is zero. Although an arbitrary perturbation may create nonzero candidate values or scores, their product has no first-order contribution at the current point in $\mathcal{S}_{16}$.

For a transition token, write $a_i = \widetilde{\mathbf{V}}_2 h_i$,$s_i = h_i^\top \mathbf{P}_2 q$, $e = f_{\boldsymbol{\theta}}(\mathbf{z}) - y_\tau(\mathbf{z})$. Consequently, the full differential evaluated on $\mathcal{S}_{16}$ is

$$\begin{aligned} df_{\boldsymbol{\theta}}(\mathbf{z}) = \Pi\, dq &+ \frac{1}{n}\sum_{i=0}^{n-1} s_i\big(\Pi\, d\mathbf{V}_2\, h_i + \widetilde{\mathbf{V}}_2\, dh_i\big) \\ &+ \frac{1}{n}\sum_{i=0}^{n-1} a_i\big((dh_i)^\top \mathbf{P}_2 q + h_i^\top (d\mathbf{P}_2) q + h_i^\top \mathbf{P}_2\, dq\big). \end{aligned} \tag{B.11}$$

The sample-loss differential is $d\ell = e^\top df_{\boldsymbol{\theta}}(\mathbf{z})$.

Holding the first-layer tokens fixed, the terms involving $d\mathbf{P}_2$ in (B.11) give $d_{\mathbf{P}_2}\ell = \frac{1}{n}\sum_{i=0}^{n-1}(e^\top a_i) h_i^\top (d\mathbf{P}_2) q$. Using the Frobenius inner product $\langle A, B\rangle_F = \operatorname{tr}(A^\top B)$ and $h_i^\top (d\mathbf{P}_2) q = \langle h_i q^\top, d\mathbf{P}_2\rangle_F$, we obtain $\nabla_{\mathbf{P}_2}\ell(\boldsymbol{\theta}; \mathbf{z}) = \frac{1}{n}\sum_{i=0}^{n-1}(e^\top a_i) h_i q^\top$. Each h_i is supported

in $\mathcal{I}$ and q is supported in W. Thus every outer product $h_i q^\top$ has nonzero entries only in rows $\mathcal{I}$ and columns W, and $\operatorname{supp}(\nabla_{\mathbf{P}_2}\ell) \subseteq (\mathcal{I}, W)$. Entrywise, for a row coordinate r and column coordinate s, $(\nabla_{\mathbf{P}_2}\ell)_{rs} = \frac{1}{n}\sum_i (e^\top a_i)(h_i)_r q_s = 0$ if $r \notin \mathcal{I}$ or $s \notin W$.

Likewise, the terms involving $d\mathbf{V}_2$ give

$$d_{\mathbf{V}_2}\ell = \frac{1}{n}\sum_{i=0}^{n-1} s_i e^\top \Pi (d\mathbf{V}_2) h_i = \left\langle \frac{1}{n}\sum_{i=0}^{n-1} s_i \Pi^\top e\, h_i^\top, d\mathbf{V}_2 \right\rangle_F .$$

Hence $\nabla_{\mathbf{V}_2}\ell(\boldsymbol{\theta}; \mathbf{z}) = \frac{1}{n}\sum_{i=0}^{n-1} s_i \Pi^\top e\, h_i^\top$. The vector $\Pi^\top e$ is supported in W_{out}, while h_i is supported in $\mathcal{I}$. Therefore $\operatorname{supp}(\nabla_{\mathbf{V}_2}\ell) \subseteq (W_{\text{out}}, \mathcal{I})$. In particular, every row outside W_{out}, including the first row of the W block, has zero gradient; within W_{out}, the columns in Ψ and W also have zero gradient. To propagate the loss gradient through the first layer, collect the dh_i and dq terms in (B.11). For each transition token, $d_{h_i}\ell = \frac{1}{n}\left[s_i e^\top \widetilde{\mathbf{V}}_2\, dh_i + (e^\top a_i)(dh_i)^\top \mathbf{P}_2 q\right] = \lambda_i^\top dh_i$, where

$$\lambda_i = \frac{1}{n}\left[s_i \widetilde{\mathbf{V}}_2^\top e + (e^\top a_i)\mathbf{P}_2 q\right]. \tag{B.12}$$

Both $\widetilde{\mathbf{V}}_2^\top e$ and $\mathbf{P}_2 q$ are supported in $\mathcal{I}$, so $\operatorname{supp}(\lambda_i) \subseteq \mathcal{I}$.

For the parameter token, its residual contribution is $e^\top \Pi\, dq$, and its query contribution is $n^{-1}\sum_i (e^\top a_i) h_i^\top \mathbf{P}_2\, dq$. Thus

$$\lambda_q = \Pi^\top e + \frac{1}{n}\sum_{i=0}^{n-1} (e^\top a_i)\mathbf{P}_2^\top h_i, \qquad d_q\ell = \lambda_q^\top dq. \tag{B.13}$$

There is no additional key/value contribution from $h_N = q$, because its summand has zero differential by (B.10). Since $\mathbf{P}_2$ is supported in $(\mathcal{I}, W)$, its transpose maps h_i into W. Also $\Pi^\top e$ is supported in $W_{\text{out}} \subseteq W$. Consequently, $\operatorname{supp}(\lambda_q) \subseteq W$. Each candidate token appears in the second layer only as a key/value token in its own summand. That summand has zero differential, so

$$\nabla_{\widehat{c}_{i,a}}\ell = 0. \tag{B.14}$$

Here q and the candidate tokens are treated as separate intermediate variables; their shared dependence on $\mathbf{V}_1 p$ is included in the next step.

Gradient of $\mathbf{V}_1$. When varying $\mathbf{V}_1$ and keeping $\mathbf{P}_1$ fixed, the attention weights and $\bar{c}_i$ are fixed. Differentiating (B.6) yields $dh_i = (d\mathbf{V}_1)\bar{c}_i$, $d\widehat{c}_{i,a} = (d\mathbf{V}_1)p$, $dq = (d\mathbf{V}_1)p$. By the chain rule, including all three token types,

$$d_{\mathbf{V}_1}\ell = \sum_{i=0}^{n-1} \lambda_i^\top (d\mathbf{V}_1)\bar{c}_i + \sum_{i,a} (\nabla_{\widehat{c}_{i,a}}\ell)^\top (d\mathbf{V}_1)p + \lambda_q^\top (d\mathbf{V}_1)p = \sum_{i=0}^{n-1} \lambda_i^\top (d\mathbf{V}_1)\bar{c}_i + \lambda_q^\top (d\mathbf{V}_1)p,$$

where the candidate terms vanish by (B.14). It follows that $\nabla_{\mathbf{V}_1}\ell(\boldsymbol{\theta}; \mathbf{z}) = \sum_{i=0}^{n-1} \lambda_i \bar{c}_i^\top + \lambda_q p^\top$. For the first sum, λ_i is supported in $\mathcal{I}$ and $\bar{c}_i$ in Ψ, so its support is contained in $(\mathcal{I}, \Psi)$. For the second term, both λ_q and p are supported in W, so its support is contained in (W, W). Hence

$$\operatorname{supp}(\nabla_{\mathbf{V}_1}\ell) \subseteq (\mathcal{I}, \Psi) \cup (W, W). \tag{B.15}$$

Equivalently, the only possibly nonzero gradient blocks of $\mathbf{V}_1$ are (Φ, Ψ), (G, Ψ), (R, Ψ), (U, Ψ), (W, W). Every entry outside these five blocks therefore has zero sample gradient.

Gradient of $\mathbf{P}_1$, including the softmax derivative. Only transition queries have nonconstant first-layer attention weights. For a candidate query or the parameter query, the unique allowed-key weight is $e^u/e^u = 1$, whose derivative with respect to every entry of $\mathbf{P}_1$ is zero.

For a transition query, differentiating (B.5) gives

$$d\pi_{i,a} = \pi_{i,a}(du_{i,a} - \sum_{b \in \mathcal{A}} \pi_{i,b}\, du_{i,b}), \tag{B.16}$$

$$d\bar{c}_i = \sum_{a \in \mathcal{A}} c_{i,a}\, d\pi_{i,a} = \sum_{a \in \mathcal{A}} \pi_{i,a} c_{i,a}\, du_{i,a} - \bar{c}_i \sum_{b \in \mathcal{A}} \pi_{i,b}\, du_{i,b} = \sum_{a \in \mathcal{A}} \pi_{i,a}(c_{i,a} - \bar{c}_i)\, du_{i,a}. \tag{B.17}$$

When only $\mathbf{P}_1$ varies, $dh_i = \mathbf{V}_1\, d\bar{c}_i$, while $dq = 0$ and $d\widehat{c}_{i,a} = 0$. Therefore

$$d_{\mathbf{P}_1}\ell = \sum_{i=0}^{n-1} \lambda_i^\top \mathbf{V}_1\, d\bar{c}_i = \sum_{i=0}^{n-1}\sum_{a\in\mathcal{A}} \pi_{i,a}\lambda_i^\top \mathbf{V}_1 (c_{i,a} - \bar{c}_i)\, du_{i,a}.$$

Define the scalar $\kappa_{i,a} := \pi_{i,a}\lambda_i^\top \mathbf{V}_1(c_{i,a} - \bar{c}_i)$. Since the input tokens are fixed, $du_{i,a} = c_{i,a}^\top (d\mathbf{P}_1) t_i = \langle c_{i,a} t_i^\top, d\mathbf{P}_1\rangle_F$. Consequently, $\nabla_{\mathbf{P}_1}\ell(\boldsymbol{\theta}; \mathbf{z}) = \sum_{i=0}^{n-1}\sum_{a\in\mathcal{A}} \kappa_{i,a} c_{i,a} t_i^\top$. Each $c_{i,a}$ is supported in Ψ and each input transition token t_i is supported in $\mathcal{K} = (\Phi, R, U)$. Hence

$$\operatorname{supp}(\nabla_{\mathbf{P}_1}\ell) \subseteq (\Psi, \mathcal{K}). \tag{B.18}$$

Explicitly, only $\mathbf{P}_{1,\Psi\Phi}$, $\mathbf{P}_{1,\Psi R}$, and $\mathbf{P}_{1,\Psi U}$ can have nonzero gradients. The input t_i, rather than the updated token h_i, appears here because the first-layer attention scores are computed from the original prompt.

Population gradients and zero coordinates. The calculations above hold for every fixed context z. Taking expectations gives the full-matrix population gradient formulas, evaluated at any $\boldsymbol{\theta} \in \mathcal{S}_{16}$:

$$\nabla_{\mathbf{P}_1}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}}\left[\sum_{i,a} \kappa_{i,a} c_{i,a} t_i^\top\right], \quad \nabla_{\mathbf{V}_1}\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}}\left[\sum_i \lambda_i \bar{c}_i^\top + \lambda_q p^\top\right],$$
$$\nabla_{\mathbf{P}_2}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\mathbb{E}_z\left[\sum_i (e^\top a_i) h_i q^\top\right], \quad \nabla_{\mathbf{V}_2}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\mathbb{E}_{\mathbf{z}}\left[\sum_i s_i \Pi^\top e\, h_i^\top\right].$$

Their supports are therefore contained in

$$\begin{aligned} &\operatorname{supp}(\nabla_{\mathbf{P}_1}\mathcal{L}) \subseteq (\Psi, \mathcal{K}), \quad \operatorname{supp}(\nabla_{\mathbf{V}_1}\mathcal{L}) \subseteq (\mathcal{I}, \Psi) \cup (W, W), \\ &\operatorname{supp}(\nabla_{\mathbf{P}_2}\mathcal{L}) \subseteq (\mathcal{I}, W), \quad \operatorname{supp}(\nabla_{\mathbf{V}_2}\mathcal{L}) \subseteq (W_{\text{out}}, \mathcal{I}). \end{aligned} \tag{B.19}$$

These are exactly the 16 retained blocks. Thus

$$\nabla_{\boldsymbol{\theta}_{\text{out}}}\mathcal{L}(\boldsymbol{\theta}) = 0 \quad \text{whenever } \boldsymbol{\theta}_{\text{out}} = 0. \tag{B.20}$$

No cancellation between different contexts is used. In particular, the zero-gradient conclusion does not require a density, symmetry, or richness condition on the training distribution. The boundedness condition in Assumption 1 (i) justifies differentiation under the expectation: on each compact parameter set, the fixed-mask softmax and the two attention layers have uniformly bounded derivatives of every fixed order on these bounded inputs. The teacher target is bounded as well. Thus the sample loss and its first two parameter derivatives admit an integrable bound there, so $\mathcal{L}$ is C^2 and the displayed expectation formulas are valid.

From zero normal gradients to invariance of the flow. It remains to show that (B.20) holds along the entire training trajectory, rather than only at initialization. Write $\boldsymbol{\theta} = (x, y)$, where x collects the retained coordinates and $y = \boldsymbol{\theta}_{\text{out}}$. Then the full gradient-flow system is

$$\dot{x} = -\nabla_x \mathcal{L}(x, y), \qquad \dot{y} = -\nabla_y \mathcal{L}(x, y). \tag{B.21}$$

By (B.20), $\nabla_y \mathcal{L}(x, 0) = 0$ for every x, not merely for the initial value $x(0)$. Consider the restricted system $\dot{\bar{x}} = -\nabla_x \mathcal{L}(\bar{x}, 0)$, $\bar{x}(0) = x(0)$, $\bar{y}(t) \equiv 0$. It satisfies $\dot{\bar{y}}(t) = 0 = -\nabla_y \mathcal{L}(\bar{x}(t), 0)$, so $(\bar{x}(t), \bar{y}(t))$ solves the full system (B.21) with the same initial condition as $(x(t), y(t))$. Since $\mathcal{L}$ is C^2, its gradient is locally Lipschitz, and the initial-value problem has a unique local solution. The full solution therefore coincides with the restricted solution. Repeating this uniqueness argument along the trajectory gives $y(t) = 0$ throughout its maximal interval of existence.

For completeness, this interval contains every $t \geqslant 0$. Along a gradient-flow solution, using the product Frobenius norm, $\frac{d}{dt}\mathcal{L}(\boldsymbol{\theta}(t)) = \langle \nabla\mathcal{L}(\boldsymbol{\theta}(t)), \dot{\boldsymbol{\theta}}(t)\rangle = -\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\|^2 = -\|\dot{\boldsymbol{\theta}}(t)\|^2$. Integrating and using $\mathcal{L} \geqslant 0$ yields $\int_0^t \|\dot{\boldsymbol{\theta}}(s)\|^2\, ds = \mathcal{L}(\boldsymbol{\theta}(0)) - \mathcal{L}(\boldsymbol{\theta}(t)) \leqslant \mathcal{L}(\boldsymbol{\theta}(0))$. By the triangle inequality and Cauchy–Schwarz,

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \leqslant \int_0^t \|\dot{\boldsymbol{\theta}}(s)\|\, ds \leqslant \sqrt{t}\left(\int_0^t \|\dot{\boldsymbol{\theta}}(s)\|^2\, ds\right)^{1/2} \leqslant \sqrt{t\,\mathcal{L}(\boldsymbol{\theta}(0))}.$$

If the maximal existence time were finite, this bound would keep the trajectory inside a compact subset of the full parameter space up to that time. The locally Lipschitz vector field then permits continuation, contradicting maximality. Thus the solution is global, and $\boldsymbol{\theta}_{\mathrm{out}}(t) = 0$ for all $t \geqslant 0$.

For any batch $\mathbf{z}_1, \ldots, \mathbf{z}_B$, at every $\boldsymbol{\theta} \in \mathcal{S}_{16}$, $\nabla_{\boldsymbol{\theta}_{\mathrm{out}}} \widehat{\mathcal{L}}_B(\boldsymbol{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \nabla_{\boldsymbol{\theta}_{\mathrm{out}}} \ell(\boldsymbol{\theta}; \mathbf{z}_b) = 0$. Hence an update $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta_k \nabla \widehat{\mathcal{L}}_B(\boldsymbol{\theta}^k)$ gives $\boldsymbol{\theta}^{k+1}_{\mathrm{out}} = \boldsymbol{\theta}^k_{\mathrm{out}} - \eta_k \, 0 = 0$ whenever $\boldsymbol{\theta}^k_{\mathrm{out}} = 0$. Induction proves invariance at every discrete iteration, without projecting the gradients or explicitly freezing these coordinates. □

B.3 Support parameter blocks

Denote the six blocks by $A = (\mathbf{P}_1)_{\Psi,U}$, $L = (\mathbf{V}_1)_{G,\Psi}$, $\mathbf{B}_- = (\mathbf{P}_2)_{\Phi,W}$, $\mathbf{B}_+ = (\mathbf{P}_2)_{G,W}$, $B_r = (\mathbf{P}_2)_{R,W}$, $C = (\mathbf{V}_2)_{W_{\mathrm{out}},\Phi}$. Let $\mathcal{S}_6 = \{\mathbf{A}, \mathbf{L}, \mathbf{B}_-, \mathbf{B}_+, \mathbf{B}_r, \mathbf{C}\}$. For $\mathcal{S} \subseteq \mathcal{S}_6$, let $\Theta(\mathcal{S})$ be the parameter space in which the blocks in $\mathcal{S}$ are unrestricted and every other entry of the four matrices is zero. Set $\Theta_6 = \Theta_{\sup} = \Theta(\mathcal{S}_6)$. Let $\mathcal{H}$ denote the class of prompts of the form (7). Define

$$F_\tau(\mathbf{H}) = \mathbf{w} + \frac{\alpha}{n} \sum_{i=1}^{n} \phi_i \big(r_i + \gamma (\mu_i^\tau)^\top \mathbf{w} - \phi_i^\top \mathbf{w} \big). \tag{B.22}$$

where $\mu_i(\mathbf{A}) = \sum_{a=1}^{m} \frac{\exp(\psi_{ia}^\top \mathbf{A}\mathbf{w})}{\sum_{b=1}^{m} \exp(\psi_{ib}^\top \mathbf{A}\mathbf{w})} \psi_{ia}$, and $\mu_i^\tau = \mu_i(\tau^{-1} I_d)$.

Proposition B.3 (Sufficiency and inclusion-minimality of the six-block support)**.** *Let $d, n \geqslant 1$, $m \geqslant 2$, $\alpha > 0$, $0 < \gamma < 1$, and $0 < \tau < \infty$. Then*

$$\exists\, \boldsymbol{\theta} \in \Theta_6 \quad \forall \mathbf{H} \in \mathcal{H} : \quad \mathrm{TF}_{\boldsymbol{\theta}}(\mathbf{H}) = F_\tau(\mathbf{H}), \tag{B.23}$$

$$\forall\, \mathcal{S} \subsetneq \mathcal{S}_6 \quad \forall \boldsymbol{\theta} \in \Theta(\mathcal{S}) \quad \exists \mathbf{H} \in \mathcal{H} : \quad \mathrm{TF}_{\boldsymbol{\theta}}(\mathbf{H}) \neq F_\tau(\mathbf{H}). \tag{B.24}$$

Thus the six-block support realizes the target update exactly, whereas none of its proper subsets does so on all original prompts.

Proof. For arbitrary parameters in Θ_6, the first-layer transition token is $t_i^{(1)} = [\phi_i; L\mu_i(A); r_i; 0_d; \mathbf{w}; 0_{d+1}]$. Indeed, the mask restricts transition i to its own successor candidates, whose scores are $\psi_{ia}^\top \mathbf{A}\mathbf{w}$. The candidates and final query are unchanged, since each attends only to the final query p and $\mathbf{V}_1 p = 0$. In the second layer, $\mathbf{P}_2 p$ is supported on Φ, G, R. Consequently, the candidate and final-query scores against p are zero; only the transition tokens contribute. Their scores are $\sigma_i = \phi_i^\top \mathbf{B}_- \tilde{\mathbf{w}} + \mu_i(\mathbf{A})^\top \mathbf{L}^\top \mathbf{B}_+ \tilde{\mathbf{w}} + r_i \mathbf{B}_r \tilde{\mathbf{w}}$, and their value vectors in the output coordinates are $\mathbf{C}\phi_i$. Including the residual readout gives

$$\mathrm{TF}_{\boldsymbol{\theta}}(\mathbf{H}) - \mathbf{w} = \frac{1}{n} \sum_{i=1}^{n} C\phi_i \left(\phi_i^\top \mathbf{B}_- \tilde{\mathbf{w}} + \mu_i(A)^\top L^\top \mathbf{B}_+ \tilde{\mathbf{w}} + r_i \mathbf{B}_r \tilde{\mathbf{w}} \right). \tag{B.25}$$

Sufficiency. For any $c \neq 0$ and any $\mathbf{L} \in \mathrm{GL}(d)$, choose $\mathbf{A} = \tau^{-1} I_d$, $\mathbf{B}_- = -c\mathbf{J}$, $\mathbf{B}_r = c e_0^\top$, $\mathbf{B}_+ = \gamma c \mathbf{L}^{-\top} \mathbf{J}$, $\mathbf{C} = \frac{\alpha}{c} I_d$. Then $\mathbf{J}\tilde{\mathbf{w}} = \mathbf{w}$, $e_0^\top \tilde{\mathbf{w}} = 1$, and $L^\top \mathbf{B}_+ = \gamma c \mathbf{J}$. Thus $\sigma_i = c\big(r_i + \gamma(\mu_i^\tau)^\top \mathbf{w} - \phi_i^\top \mathbf{w}\big)$, $\mathbf{C}\phi_i = \frac{\alpha}{c}\phi_i$. Equation (B.25) becomes (B.22), proving (B.23).

Necessity of each block within this support. Fix a unit vector $e \in \mathbb{R}^d$. In every counterexample below, take $\phi_1 = e$ and $\phi_i = 0$ for $i \geqslant 2$. Accordingly only transition 1 contributes to either update, regardless of the fixed context length n. These inputs can be realized by a trajectory that moves from an initial state to a second state and subsequently selects a zero-feature action there. The successor feature lists used below are either all zero or $(e, 0, \ldots, 0)$, so such an action exists because $m \geqslant 2$. Set the later rewards to zero. This construction uses the original prompt format; it does not require independent transition tokens or a particular sampling distribution.

If $\mathbf{C} = 0$, equation (B.25) gives $\mathrm{TF}_{\boldsymbol{\theta}}(\mathbf{H}) = \mathbf{w}$ for every input. Choose $\mathbf{w} = 0$, $r_1 = 1$, and all first-transition successor features zero. The target increment is $\alpha e/n \neq 0$, which cannot be reproduced; With $\mathbf{B}_r = 0$, the output in (B.25) does not depend on any reward. Fix $\mathbf{w} = 0$ and the first-transition successor features at zero, and compare otherwise identical inputs with $r_1 = 0$ and $r_1 = 1$. Their target outputs differ by $\alpha e/n$, while the model outputs coincide. No fixed choice of the other blocks can fit both inputs; Set $\mathbf{w} = e$, $r_1 = 0$, and all first-transition successor features

to zero. Then $\mu_1(\mathbf{A}) = 0$ for every $\mathbf{A}$. If $\mathbf{B}_- = 0$, the model increment is zero, but the target increment is $-\alpha e/n \neq 0$; If $\mathbf{L} = 0$ or $\mathbf{B}_+ = 0$, then $\mathbf{L}^\top \mathbf{B}_+ = 0$, so the output is independent of the successor features. Fix $\mathbf{w} = e$ and $r_1 = 0$, and compare successor lists $(0, \dots, 0)$ and $(e, 0, \dots, 0)$. The target outputs differ by $\frac{\alpha\gamma}{n} \frac{\exp(1/\tau)}{\exp(1/\tau)+m-1} e \neq 0$. Again the model outputs coincide, so they cannot both equal the targets; If $\mathbf{A} = 0$, attention weights are uniform. Holding the features and rewards fixed, $\mu_i(0)$ is independent of $\mathbf{w}$, and the increment in (B.25) is an affine function of $\mathbf{w}$. Choose the first-transition successor list $(e, 0, \dots, 0)$, $r_1 = 0$, and $\mathbf{w} = te$, with t varying in an interval about zero. The target increment is $g(t) = \frac{\alpha}{n} e\big[\gamma t p(t) - t\big]$, $p(t) = \frac{\exp(t/\tau)}{\exp(t/\tau)+m-1}$. Since $p'(0) = (m-1)/(\tau m^2)$, $g''(0) = \frac{2\alpha\gamma(m-1)}{n\tau m^2} e \neq 0$. Thus the target increment is not affine in t, whereas the model increment is affine. They cannot agree for all these inputs.

Every proper subset of $\mathcal{S}_6$ omits at least one of these blocks. The corresponding case applies irrespective of the values of the remaining blocks and proves (B.24). □

C Experimental reproducibility and implementation details

We evaluate parameter recovery and cross-task generalization in a structured two-layer Q-learning Transformer trained by output-only supervision. All models use cosine learning-rate decay from the first Adam update. Batch-size and context-length comparisons further reveal distinct effects on parameter recovery and functional accuracy under fixed task budgets. Frozen models transfer to 512 unseen tabular tasks, achieving returns close to their soft teachers.

Table 1: Notation. A task is one independently sampled MDP/feature realization; a context is a valid trajectory window generated in that task. In the training, one task supplies one context.

Symbol	Meaning and default
S, m, d	State count $S = 9$, action count $m = 4$, Q-feature dimension $d = 36$.
$\mathbf{w}, \boldsymbol{\theta}$	Q parameter $\mathbf{w} \in \mathbb{R}^d$; distinct Transformer parameter vector $\boldsymbol{\theta}$.
τ	Teacher temperature.
$\alpha, \gamma, \epsilon, \eta$	Q-update step 0.2, discount 0.5, behavior exploration 0.1, Adam learning rate.
B, U	Independent contexts per optimizer update; number of optimizer updates.
n	Transitions within one context, default 20; not minibatch size B.
$N_{\text{task}}, N_{\text{obs}}$	Tasks/contexts seen: $N_{\text{task}} = BU$; supplied transitions: $N_{\text{obs}} = nBU$.
D, N_{tok}	Token width $D = 5d + 2 = 182$; tokens per prompt $N_{\text{tok}} = n(m+1) + 1$.

Task budget and optimization. Each independently sampled task contributes one context. Let B denote the number of contexts per optimizer update, n the number of transitions within a context, and U_B the number of optimizer updates. At update $k = 0, \dots, U_B - 1$, let $m_k = kB$ be the number of tasks already consumed. The learning rate is $\eta(m_k) = 10^{-6} + \frac{10^{-3}-10^{-6}}{2}\left[1 + \cos(\pi \frac{m_k}{N_{\text{task}}})\right]$. All blocks follow this schedule from the first update at 10^{-3}, without warm-up, restarts, or an additional phase. Its minimum is 10^{-6} at the budget boundary; the last actual update occurs at $m_k = N_{\text{task}} - B$ and uses a slightly larger value. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_{\text{Adam}} = 10^{-8}$, no weight decay, and float32 parameters and moments.

The study uses five paired seeds for each configuration. For experiments on valid parameter blocks, the cosine schedule terminates at 20.48×10^6 tasks. Training continues at a learning rate of 10^{-6} to 25.60×10^6 tasks, then 10^{-4}, with the rate lowered to 10^{-5} beyond 37.76×10^6 tasks. The rate is finally switched to 10^{-6}, ending at 46.40×10^6 tasks per seed. For experiments on structured parameter blocks, the cosine schedule terminates at 20.48×10^6 tasks.

C.1 Experiments on Valid Parameter Blocks

Evaluating the accuracy of approximate Q updates. We adopt three metrics, namely Soft error err_{soft}, Hard error err_{hard}, and Greedy agreement, to quantify how closely the Transformer proposed

in this work approximates the Q-learning update rule. Specifically,

$$\text{err}_{\text{soft}} = \sqrt{\frac{\sum_{\mathbf{z}\in\mathcal{Z}} \|\hat{u}(\mathbf{z}) - u_\tau(\mathbf{z})\|_2^2}{\sum_{\mathbf{z}\in\mathcal{Z}} \|u_\tau(\mathbf{z})\|_2^2}}, \qquad \text{err}_{\text{hard}} = \sqrt{\frac{\sum_{\mathbf{z}} \|\hat{u}(\mathbf{z}) - u_{\max}(\mathbf{z})\|_2^2}{\sum_{\mathbf{z}} \|u_{\max}(\mathbf{z})\|_2^2}}. \tag{C.1}$$

where $u_\tau(\mathbf{z}) = \frac{\alpha}{n}\sum_{i=0}^{n-1} \phi_i[r_{i+1} + \gamma V_{\tau,i} - \mathbf{w}^\top\phi_i]$, $u_{\max}(\mathbf{z}) = \frac{\alpha}{n}\sum_{i=0}^{n-1} \phi_i[r_{i+1} + \gamma \max_a q^+_{i,a} - \mathbf{w}^\top\phi_i]$, and $q^+_{i,a} = \mathbf{w}^\top\phi_{i,a}$, $V_{\tau,i} = \sum_a \pi_i q^+_{i,a}$. The two greedy metrics quantify action agreement at distinct granularities: $\text{Greedy}_{\text{tr}}$ reports the fraction of individual transitions for which the model's attention-selected action matches the teacher's greedy action, i.e., the action with the maximum Q-value; $\text{Greedy}_{\text{ctx}}$ is a stricter metric, defined as the fraction of contexts where such a match holds across all 20 transitions within the context. The Teacher bias is obtained by replacing the numerator of err_{hard} with $\sum_{\mathbf{z}} \|u_\tau - u_{\max}\|^2$. Teacher bias is deterministic on the common corpus and is not assigned an artificial seed confidence interval.

Update errors and greedy agreement are reported as percentages. The identity/zero-update predictor yields a 100% update error: the residual $\mathbf{w}$ cannot mask an unlearned update. $\pm$ denotes a 95% Student-t interval across trained seeds.

Table 2 (dense features) demonstrates close reproduction of the soft-Q update by the model, yielding 0.01256% NRMSE and 99.990% transition-level greedy agreement; by contrast, the one-hot feature setting (Table 3) exhibits noticeably weaker performance, attaining 0.07950% NRMSE and 99.989% transition-level greedy agreement.

Table 2: Dense held-out contexts: update errors and greedy agreement

Tasks (10^6)	Soft error (%)	Hard error (%)	$\text{Greedy}_{\text{tr}}$ (%)	$\text{Greedy}_{\text{ctx}}$ (%)
0.32	9.0157 ± 0.6508	9.0465 ± 0.6716	95.314 ± 0.102	56.587 ± 0.776
5.12	7.1287 ± 0.2225	7.1792 ± 0.2118	96.708 ± 0.157	67.205 ± 0.895
10.24	4.4694 ± 0.6952	4.5328 ± 0.6750	98.246 ± 0.087	80.769 ± 1.045
20.48	0.0657 ± 0.0002	0.6655 ± 0.0019	99.920 ± 0.005	98.840 ± 0.082
25.60	0.0625 ± 0.0002	0.6633 ± 0.0020	99.927 ± 0.003	98.936 ± 0.035
37.76	0.7393 ± 0.0557	0.9834 ± 0.0444	99.654 ± 0.012	96.003 ± 0.197
45.76	0.0846 ± 0.0107	0.6468 ± 0.0044	99.971 ± 0.002	99.636 ± 0.046
46.40	0.0126 ± 0.0005	0.6394 ± 0.0008	99.990 ± 0.004	99.861 ± 0.045

Table 3: One-hot held-out contexts: update errors and greedy agreement

Tasks (10^6)	Soft error (%)	Hard error (%)	$\text{Greedy}_{\text{tr}}$ (%)	$\text{Greedy}_{\text{ctx}}$ (%)
0.32	55.7678 ± 4.5493	57.4641 ± 4.8244	95.752 ± 0.151	66.636 ± 0.899
5.12	44.0258 ± 1.9607	45.6216 ± 1.9971	97.076 ± 0.085	75.928 ± 0.833
10.24	27.8956 ± 4.6804	29.2407 ± 4.7820	98.406 ± 0.067	86.035 ± 0.483
20.48	0.4427 ± 0.0066	5.1372 ± 0.0071	99.985 ± 0.002	99.834 ± 0.023
25.60	0.4181 ± 0.0073	5.1137 ± 0.0097	99.987 ± 0.003	99.875 ± 0.022
37.76	4.5953 ± 0.3353	6.7774 ± 0.2200	99.681 ± 0.012	97.070 ± 0.077
45.76	0.5257 ± 0.0639	4.8381 ± 0.0171	99.978 ± 0.004	99.783 ± 0.020
46.40	0.0795 ± 0.0037	4.8091 ± 0.0029	99.989 ± 0.004	99.890 ± 0.039

Emergence of Parameters. To distinguish scalar calibration from nonscalar distortion, let $\beta_{att} = \text{tr}(\mathbf{A}_\theta)/d$. Frobenius orthogonality of $\mathbf{A}_\theta - \beta_{att} I$ and I gives

$$E_A^2 = \underbrace{(\tau\beta_{att} - 1)^2}_{E^2_{A,\text{scale}}} + \underbrace{\frac{\tau^2}{d}\|\mathbf{A}_\theta - \beta_{att} I\|_F^2}_{E^2_{A,\text{shape}}}.$$

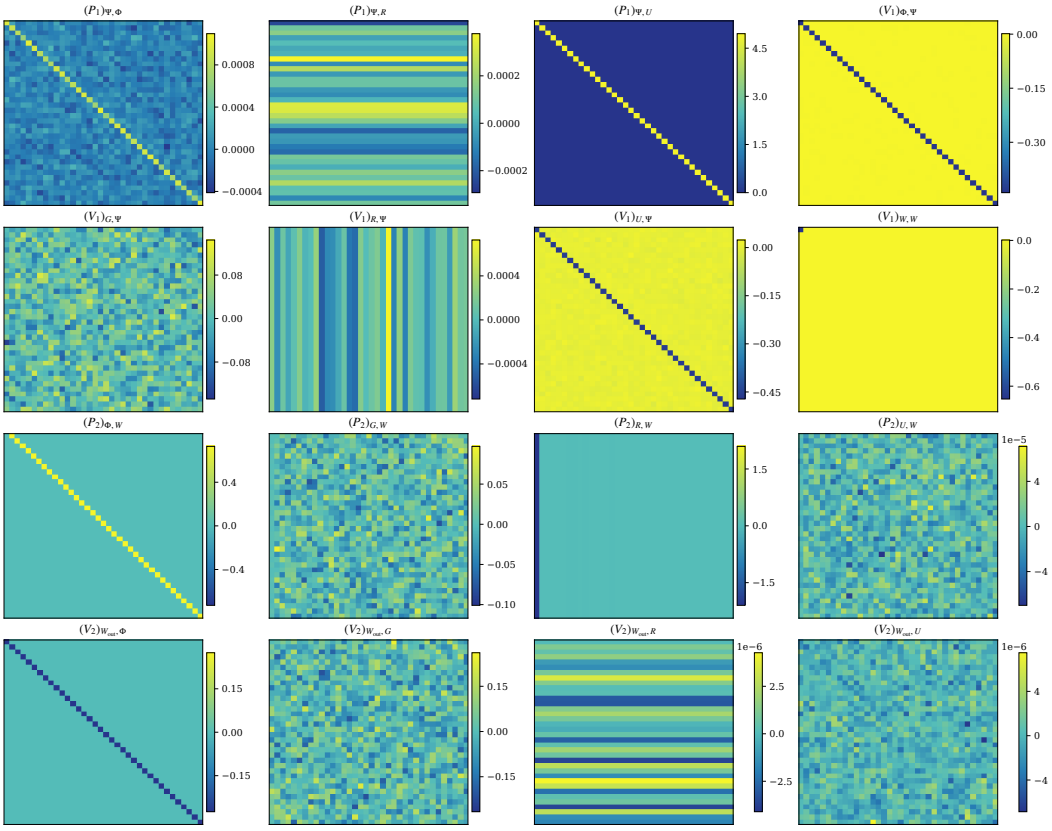


Figure 3: Terminal detail of all 16 initialized blocks for seed 0. Each block has its own color range, shared across seeds for that block. This view complements the full matrices; it does not replace structural zeros by cropped or normalized values in the full-matrix figures.

When $\beta_{att} > 0$, $1/\beta_{att}$ summarizes the effective temperature of the scalar projection; it is not an exact temperature when nonscalar components remain. A visible diagonal in a heatmap is therefore insufficient evidence of inverse-temperature recovery. As can be seen in Table 4, both residual components decrease substantially toward the end of training, demonstrating the emergence of an action-selection mechanism. Figures 3 presents heatmaps of the key matrices under $\tau = 0.2$, revealing that the training algorithm successfully induces the constructed parameters to emerge.

Table 4: Attention residual and scale across training.

Tasks (10^6)	E_A	β_{att}	$E_{A,\text{shape}}$
0.32	0.729464 ± 0.000924	1.353730 ± 0.004596	0.017515 ± 0.000474
5.12	0.435038 ± 0.001273	2.828350 ± 0.006350	0.024805 ± 0.000613
10.24	0.303830 ± 0.000436	3.483203 ± 0.002233	0.016895 ± 0.000379
20.48	0.081884 ± 0.000192	4.590583 ± 0.000960	0.000274 ± 0.000009
25.60	0.077982 ± 0.000189	4.610089 ± 0.000944	0.000148 ± 0.000007
37.76	0.023377 ± 0.000188	4.885535 ± 0.001009	0.004731 ± 0.000092
45.76	0.009809 ± 0.000093	4.951000 ± 0.000462	0.000428 ± 0.000016
46.40	0.009781 ± 0.000091	4.951105 ± 0.000457	0.000185 ± 0.000008

Generalization to unseen tabular tasks We evaluate frozen Transformers on a common bank of 512 independent unseen MDPs. Deployment features are one-hot, $\phi(s, a) = e_{4s+a}$, whereas training

uses dense features. These experiments assess dense-to-tabular transfer, not direct tabular training. Each learner recursively updates its own Q parameter $\mathbf{w}$ and generates its own ϵ-greedy actions and successor states. The Transformer parameters $\boldsymbol{\theta}$ remain fixed, with no teacher parameters, actions, or states injected into its rollout. The primary comparison uses $\mathbf{w}_0 = 0$, $n_{\text{test}} = 20$, and 20,000 environment interactions per learner.

Every temperature comparison includes the Transformer, hard and soft window-Q teachers, Oracle, and Random. The window teachers, or their hard-max counterpart, maintain a fixed $\mathbf{w}$ for each trajectory segment. Standard online Q-learning instead updates after every transition; we report both a constant step size of 0.2 and a visit-count step size of $1/N_{sa}$. Oracle uses the true transition and reward model to compute an optimal policy and is therefore a privileged upper reference. Random uses a uniform action policy rather than a greedy policy induced by a random Q table.

We evaluate the exact discounted return $J(\pi) = P_0^\top (I - \gamma P_\pi)^{-1} r_\pi$, $r(s,a) = \sum_{s'} P(s' \mid s,a) R(a,s')$, where P_π and r_π are the policy-induced transition matrix and expected reward vector. For learned methods and Q-learning baselines, π is the current greedy policy; Oracle and Random use their specified policies. Thus, J is a policy-evaluation metric.

Table 5: Exact greedy-policy return after 20,000 interactions on 512 unseen MDPs.

Method	Mean $J(\pi)$	95% interval
Transformer	0.479225	$[0.461365, 0.497690]$
Soft window-Q	0.480136	$[0.462090, 0.498943]$
Hard window-Q	0.465381	$[0.447874, 0.483684]$
Online Q, constant	0.449844	$[0.426708, 0.473394]$
Online Q, visit count	0.522897	$[0.502016, 0.544255]$
Oracle	0.529490	$[0.508685, 0.550915]$
Random	-0.002155	$[-0.019109, 0.015196]$

Figure 4 deploys frozen dense-feature-trained Transformers on unseen tabular MDPs and measures exact policy returns under a common interaction budget. Comparisons include hard and soft window Q-learning, a model-informed Oracle, and uniform Random. The Transformer achieves returns close to those of the soft window teacher and a higher sample-mean return than the hard window-Q baseline using the same fixed step size.

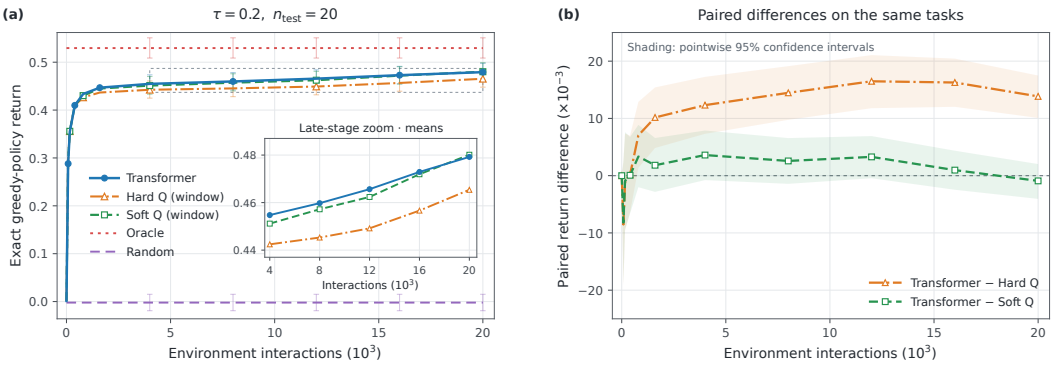


Figure 4: Frozen dense-feature-trained Transformers on unseen tabular MDPs and measures exact policy returns under a common interaction budget.

C.2 Experiments on Structured Parameter Blocks

Evaluating the accuracy of approximate Q updates. Learning error across different **temperatures** τ. Table 6 gives an analysis of the impacts of τ. The $\tau = 0.1$ model has a soft-update error of 0.07575% despite its appreciable parameter residual; the $\tau = 1$ model closely matches its soft teacher but has a large hard-update error, reflecting the finite-temperature teacher discrepancy. Here,

Greedy agreement quantifies the fraction of contexts where the argmax action from the attention module matches the optimal maximizing next action of the Q-learning target.

Table 6: Update accuracy on 8,192 common dense-feature contexts.

τ	Soft error (%)	Hard error (%)	Teacher bias (%)	Greedy (%)
1	0.00283 ± 0.00107	10.03718 ± 0.00020	10.0370	99.9985
0.8	0.00261 ± 0.00071	7.114048 ± 0.000049	7.1140	99.9974
0.5	0.00276 ± 0.00078	3.21972 ± 0.00013	3.2198	99.9962
0.2	0.00393 ± 0.00019	0.63950 ± 0.00014	0.6385	99.9990
0.1	0.07575 ± 0.00025	0.25509 ± 0.00040	0.1960	99.9707

Table 7 illustrates how **different batch sizes** B affect the approximation error. The soft error decreases first and then rises with increasing batch size, minimized at $B = 64$. Greedy agreement stays near 100%. The fitted effective temperature $\widehat{\tau}_{\text{eff}} = 1/\beta_{att}$, with $\beta_{att} = \text{tr}(\mathbf{A}_{\boldsymbol{\theta}})/d > 0$ acting as a scalar projection diagnostic, remains around 0.2.

Table 7: Final checkpoints on 8,192 new shared contexts.

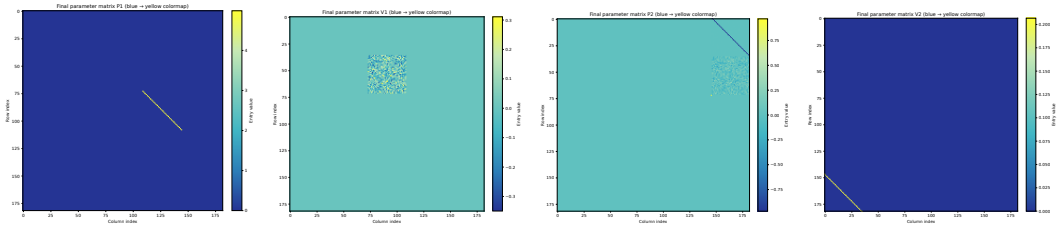
B	Soft error (%)	Hard error (%)	Greedy (%)	$\widehat{\tau}_{\text{eff}}$
1	0.0202 ± 0.0008	0.6348 ± 0.0007	99.994	0.200113
2	0.0135 ± 0.0013	0.6347 ± 0.0004	99.995	0.200128
4	0.0110 ± 0.0014	0.6349 ± 0.0004	99.995	0.200161
8	0.0077 ± 0.0009	0.6351 ± 0.0002	99.999	0.200223
16	0.0052 ± 0.0005	0.6352 ± 0.0003	99.999	0.200299
32	0.0041 ± 0.0003	0.6355 ± 0.0001	100.000	0.200373
64	0.0036 ± 0.0001	0.6357 ± 0.0001	100.000	0.200459
128	0.0045 ± 0.0001	0.6367 ± 0.0001	100.000	0.200784
256	0.0123 ± 0.0001	0.6424 ± 0.0003	99.998	0.202525

To compare the effects of different optimizations on experimental results, we also evaluated the following methods. 1. Constant Adam: All six blocks retain learning rate 10^{-3}, with moments $(0.9, 0.999)$ and numerical constant 10^{-8}; 2. Uniform cosine decay: separately labeled optimizer change. After the common 100,000-step prefix, all block learning rates decrease from 10^{-3} to 10^{-5}. For an update indexed by $j \in \{0, \ldots, 219999\}$ after the fork, the rate for block k is $\eta_k(j) = 10^{-5} + \frac{\eta_{k,0} - 10^{-5}}{2}[1 + \cos(\frac{\pi j}{219999})]$; 3. Low-rate extension: a larger, adaptively chosen budget. After observing the completed uniform-cosine runs and the aggressive schedule's training/validation overscaling, we continue all five uniform-cosine endpoints to 20.48 million tasks. Every block uses $\eta(j) = 10^{-6} + \frac{10^{-5} - 10^{-6}}{2}[1 + \cos(\frac{\pi j}{319999})]$, $j = 0, \ldots, 319999$. Table 8 reports three metrics under $\tau = 0.2$ with above learning rate schedules. Each dense-feature experiment uses $B = 32$ and $n = 20$.

Table 8: Results of Soft error, Hard error, and Greedy agreement with three learning rate schedules.

Condition	N_{task}	Soft error (%)	Hard error (%)	Greedy (%)
Constant Adam	10,240,000	3.242 ± 0.732	3.332 ± 0.717	98.100 ± 0.142
Uniform cosine	10,240,000	0.04028 ± 0.00419	0.64309 ± 0.00076	99.980 ± 0.006
Low-rate extension	20,480,000	0.00415 ± 0.00023	0.63428 ± 0.00012	100.000 ± 0.000

Emergence of Parameters. Figure 5 presents the final learned matrices $\mathbf{P}_1$, $\mathbf{V}_1$, $\mathbf{P}_2$, and $\mathbf{V}_2$ obtained after training with $\tau = 0.2$. Remarkably, the learned matrices recover the parameter structure predicted in Theorem 1. In particular, the diagonal entries of $\mathbf{A}_\theta$ in $\mathbf{P}_1$ approach $1/\tau$, indicating that the Transformer has learned the internal action-selection computation required by the soft Q-learning update. This provides direct evidence of parameter emergence.



(a) Learned $\mathbf{P}_1$ (b) Learned $\mathbf{V}_1$ (c) Learned $\mathbf{P}_2$ (d) Learned $\mathbf{V}_2$

Figure 5: Final learned $\mathbf{P}_1$, $\mathbf{V}_1$, $\mathbf{P}_2$ and $\mathbf{V}_2$ matrices after training at $\tau = 0.2$, $B = 32$, and $n = 20$. the non-zero structure matches the theoretical constructions $(\mathbf{P}_1, \mathbf{V}_1, \mathbf{P}_2, \mathbf{V}_2)$ in (9) and (11).

We define some evaluation metrics to demonstrate whether the final parameters emerge. Write $\mathbf{B}_- = [b_-, K_-]$ and $\mathbf{B}_+ = [b_+, K_+]$. The analytical construction defines a sufficient family satisfying $\mathbf{A}_\theta = \tau^{-1} I$, $K_- = -cI$, $\mathbf{B}_r = c[1, 0]$, $\mathbf{L}^\top K_+ = \gamma c I$, $c\mathbf{C} = \alpha I$, $b_- = 0$, $\mathbf{L}^\top b_+ = 0$, where $c \neq 0$ may be positive or negative. For each trained model, we estimate $\widehat{c} = ((\mathbf{B}_r)_0 - \operatorname{tr} K_-)/(d+1)$ separately for each model, allowing either sign. The product $\mathbf{L}^\top K_+$ removes the arbitrary invertible representation of the intermediate greedy feature. Likewise $\mathbf{L}^\top b_+$, rather than b_+ alone, is the appropriate product entering the output relation. Low residuals indicate agreement with these sufficient relations. The complete diagnostics used in every table are

$$E_{\mathbf{A}_\theta} = \frac{\|\tau \mathbf{A}_\theta - I\|_F}{\sqrt{d}}, \qquad E_- = \frac{\|K_- + \widehat{c} I\|_F}{|\widehat{c}|\sqrt{d}}, \tag{C.2}$$

$$E_r = \frac{\|\mathbf{B}_r - \widehat{c}[1, 0]\|_F}{|\widehat{c}|}, \qquad E_+ = \frac{\|\mathbf{L}^\top K_+ - \gamma \widehat{c} I\|_F}{\gamma |\widehat{c}| \sqrt{d}}, \tag{C.3}$$

$$E_C = \frac{\|\widehat{c}\mathbf{C} - \alpha I\|_F}{\alpha \sqrt{d}}, \qquad E_{-,0} = \frac{\|b_-\|_2}{|\widehat{c}|\sqrt{d}}, \tag{C.4}$$

$$E_{+,0} = \frac{\|\mathbf{L}^\top b_+\|_2}{\gamma |\widehat{c}| \sqrt{d}}, \qquad E_{\mathrm{TD}}^2 = E_-^2 + E_r^2 + E_+^2 + E_C^2 + E_{-,0}^2 + E_{+,0}^2. \tag{C.5}$$

and $E_{\mathrm{all}} = \sqrt{E_{\mathbf{A}_\theta}^2 + E_{\mathrm{TD}}^2}$, $\widehat{\tau}_{\mathrm{eff}} = 1/\beta_{att}$ with $\beta_{att} = \operatorname{tr}(\mathbf{A}_{\boldsymbol{\theta}})/d > 0$ is a scalar projection diagnostic, not an exact temperature when $\mathbf{A}_{\boldsymbol{\theta}}$ has nonscalar components. The scalar estimate is allowed to be negative. The full completed models have nondegenerate $\widehat{c}$; a normalization near zero is not interpreted as a zero residual. For arbitrary active blocks, E_{all} is not an established Euclidean distance to all minimizers. Figures 6 and 7 present heatmaps of the key matrices and their relations under $\tau = 0.2$, respectively, revealing that the training algorithm successfully induces the constructed parameters to emerge.

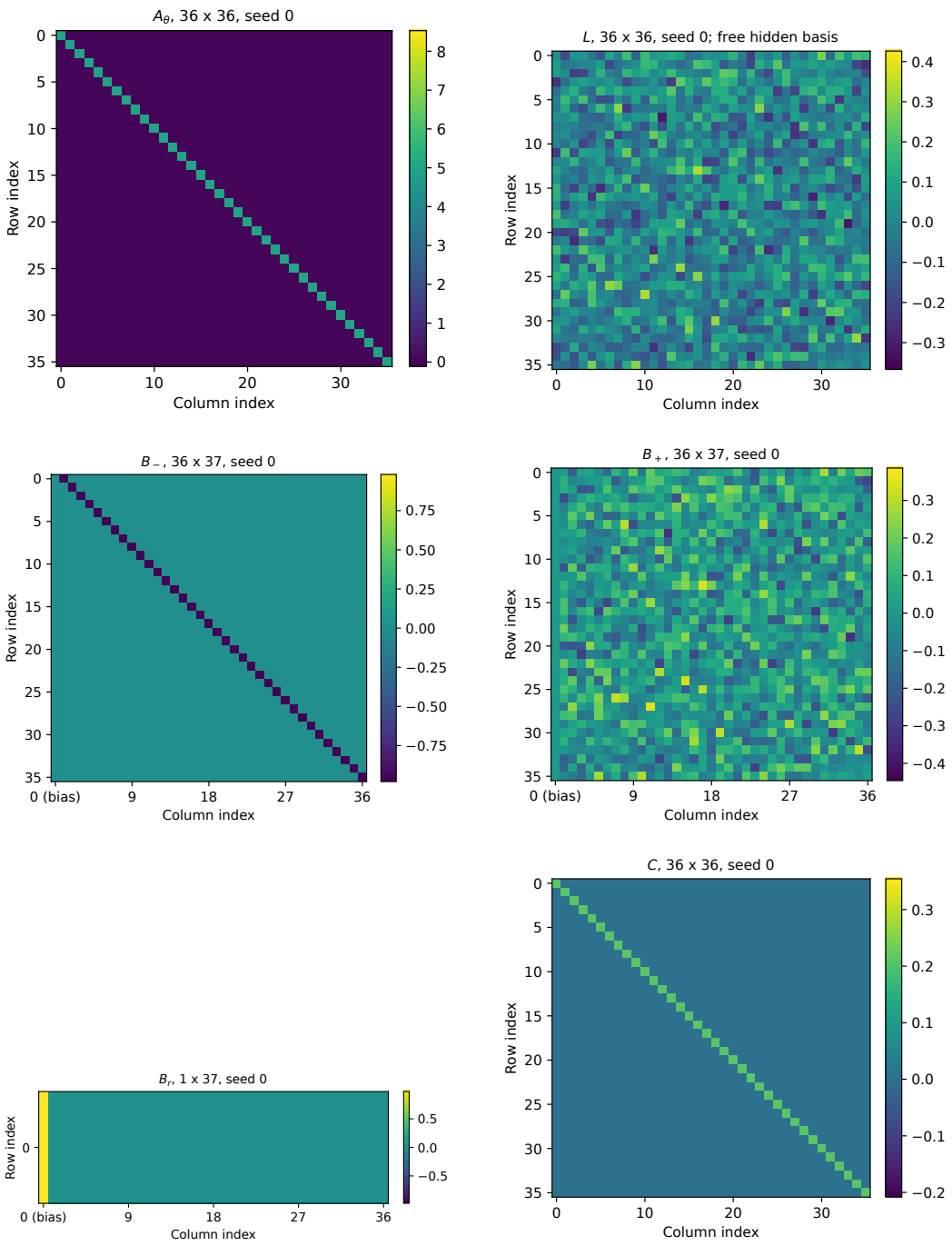


Figure 6: All six actual trained blocks, seed 0. $\mathbf{A}_{\boldsymbol{\theta}}, \mathbf{L}, \mathbf{C}$ are 36×36; $\mathbf{B}_-, \mathbf{B}_+$ are 36×37; $\mathbf{B}_r$ is 1×37. The fitted scalar is $\widehat{c} = 0.967893$. Dense $\mathbf{L}$ and $\mathbf{B}_+$ are compatible with a nontrivial invertible gauge and are not replaced by identity-shaped pictures.

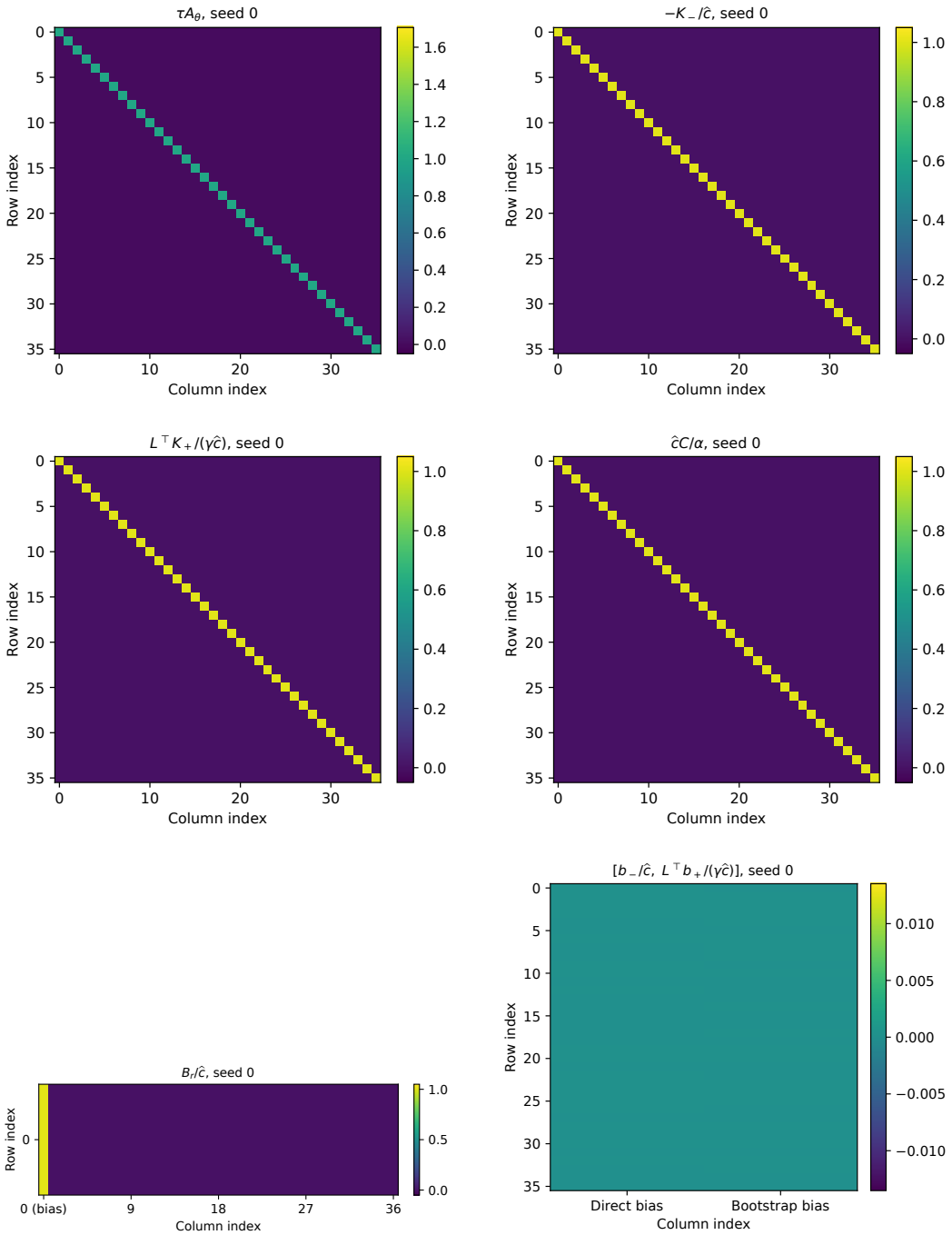


Figure 7: Normalized relations for the same model. The first four targets are I, the reward-row target is $[1, 0, \ldots, 0]$, and the two bias targets are zero. These are display-only diagnostics, not raw weights. Common limits are used for each named relation across seeds and methods, including the failed control.

Emergence of parameters and cumulative **task exposure**. Figure 8 tracks recovery along one training trajectory under growing cumulative task exposure. Reduced final TD residual shows attention-scale calibration dominates aggregate residual. Since task exposure and optimizer updates rise together, the curves cannot isolate task diversity's fixed-compute effect.

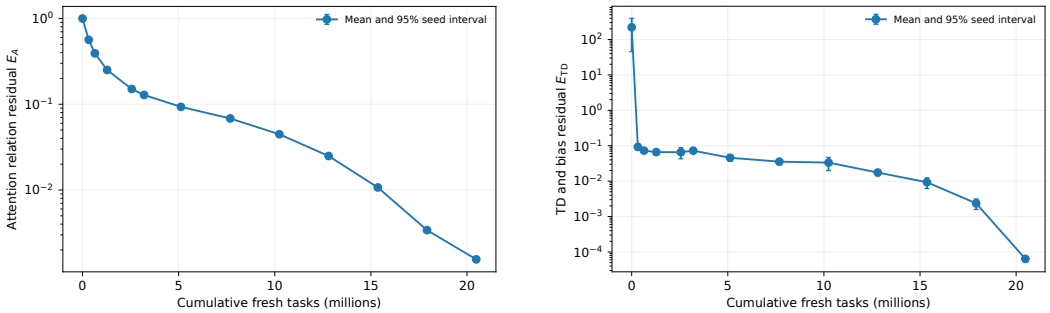


Figure 8: Parameter recovery versus cumulative fresh tasks at $B = 32$, $n = 20$, and $\tau = 0.2$. Left: attention residual E_A. Right: the aggregate of the six remaining TD and bias residuals, E_{TD}.

Temperature: Table 9 shows close agreement with the analytical relations at $\tau = 1, 0.8, 0.5$, and 0.2, while recovery remains incomplete at $\tau = 0.1$. At $\tau = 0.2$, the learned scalar projection is $\beta_{att} = 4.992236$ against a target of 5. At $\tau = 0.1$, it is 8.174396 against a target of 10, and $E_{\mathrm{all}} = 0.1826$. Figure 9 makes this scale discrepancy explicit rather than inferring recovery from matrix appearance alone. No temperature-specific tuning is performed; the tested values do not establish a critical temperature or an impossibility result for lower-temperature recovery.

Table 9: Final parameter recovery at 20.48 million tasks, $B = 32$, and $n = 20$. Entries are five-seed means; E_{all} includes a pointwise 95% interval half-width. Residuals are dimensionless. The scalar projection β_{att} is compared with $1/\tau$ without removing nonscalar distortion.

τ	$1/\tau$	β_{att}	E_A	E_{TD}	E_{all}
1	1	1.000000	3.808×10^{-5}	5.410×10^{-5}	$(6.624 \pm 2.102) \times 10^{-5}$
0.8	1.25	1.250000	3.150×10^{-5}	4.741×10^{-5}	$(5.715 \pm 1.049) \times 10^{-5}$
0.5	2	2.000002	2.424×10^{-5}	4.882×10^{-5}	$(5.466 \pm 1.323) \times 10^{-5}$
0.2	5	4.992236	1.553×10^{-3}	6.384×10^{-5}	$(1.554 \pm 0.017) \times 10^{-3}$
0.1	10	8.174396	0.1826	6.590×10^{-4}	0.18256 ± 0.00020

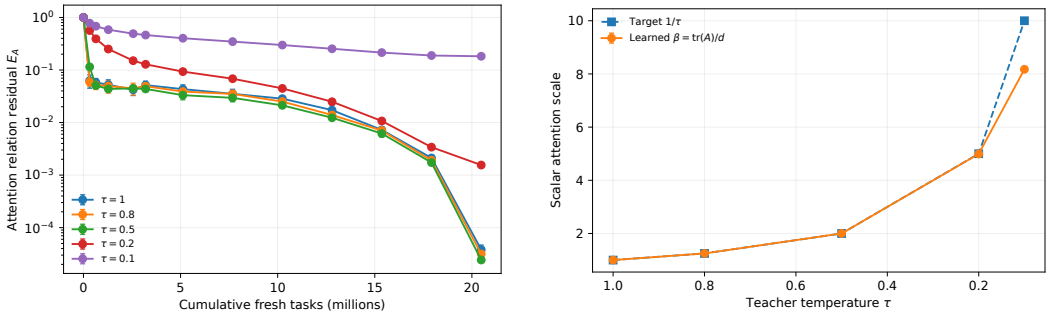


Figure 9: Temperature-dependent recovery at $B = 32$ and $n = 20$. Left: attention residual versus cumulative tasks, with five paired seeds per temperature. Right: the learned scalar projection β_{att} and its target $1/\tau$. The target is shown only for evaluation; no matrix target is supplied during training.

Batch size: scale calibration and matrix shape. We report both components in Figure 10. The lowest endpoint mean of E_{all} occurs at $B = 128$, with value 7.204×10^{-4}. By contrast, mean soft-update error continues to decrease up to $B = 64$ in Table 7. Thus, the best observed parameter recovery and the lowest observed functional error do not coincide in this sweep. At fixed N_{task}, reducing B increases the number of Adam updates, whereas increasing B averages more tasks in each update. Adam's moment memories are defined in update units, so their span in task units also changes. The

observed ranking is consequently specific to the shared schedule, numerical precision, architecture, and task budget; it does not establish a universally optimal batch size.

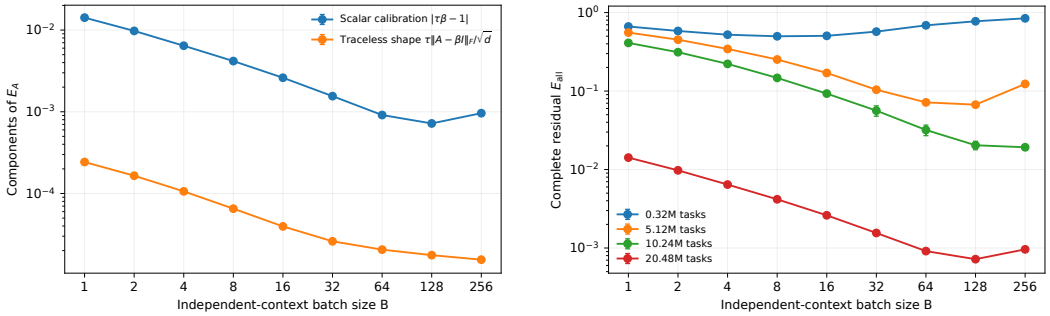


Figure 10: Batch-size effects under a fixed task budget. Left: attention scale and traceless-shape residuals, whose squares sum to E_A^2 for each model. Right: aggregate residuals at four checkpoints of the same training runs. Intermediate curves are not independently retrained short-budget experiments.

Training context length. For each training length, only the first n_{train} transitions of the common 80-step trajectory enter the loss. The teacher increment is recomputed from that prefix with normalization α/n_{train}. The $n_{\text{train}} = 20$ default uses the same paired generator and is shared across the three sweeps.

Table 10: Training-length comparison at $\tau = 0.2$ and $B = 32$. Every model receives 20.48 million tasks and 640,000 updates; supplied transitions differ. Residuals are five-seed means, with the 95% interval half-width for E_{all} in the final column. The default is shared across sweeps.

n_{train}	Transitions (M)	E_A	E_{TD}	E_{all}	95% interval half-width
1	20.48	0.0128	2.033×10^{-4}	0.0128	1.969×10^{-4}
5	102.4	2.155×10^{-3}	8.059×10^{-5}	2.156×10^{-3}	8.661×10^{-5}
10	204.8	1.699×10^{-3}	6.440×10^{-5}	1.700×10^{-3}	5.381×10^{-5}
20	409.6	1.553×10^{-3}	6.384×10^{-5}	1.554×10^{-3}	1.728×10^{-5}
40	819.2	1.739×10^{-3}	6.393×10^{-5}	1.740×10^{-3}	5.290×10^{-5}
80	1638.4	2.019×10^{-3}	7.050×10^{-5}	2.020×10^{-3}	3.846×10^{-5}

Table 11: Length-specific models on matched-length and common-$n_{\text{test}} = 20$ dense tests, and matched-length one-hot tests. Entries are update NRMSE percentages with pointwise 95% seed-interval half-widths. The tabular column measures feature transfer without one-hot training.

n_{train}	Dense matched (%)	Dense common 20 (%)	Tabular matched (%)
1	0.01893 ± 0.00032	0.01573 ± 0.00025	0.08467 ± 0.00095
5	0.00561 ± 0.00039	0.00525 ± 0.00038	0.01558 ± 0.00077
10	0.00431 ± 0.00043	0.00417 ± 0.00042	0.01284 ± 0.00049
20	0.00393 ± 0.00019	0.00393 ± 0.00019	0.01222 ± 0.00036
40	0.00367 ± 0.00020	0.00384 ± 0.00021	0.01413 ± 0.00060
80	0.00380 ± 0.00012	0.00411 ± 0.00012	0.01679 ± 0.00044

Among the tested lengths, $n_{\text{train}} = 20$ yields the lowest sample mean of E_{all}, 1.554×10^{-3} (Table 10). This is a fixed-task, fixed-update comparison, not a fixed-transition or fixed-compute comparison: $n = 80$ supplies 80 times as many transitions per task as $n = 1$. Context length also changes the distribution and variance of the window-averaged target. Table 11 therefore reports both matched-length tests and tests at the common length $n_{\text{test}} = 20$. The common-length evaluation holds the test window fixed when comparing models trained with different lengths. The one-hot results measure dense-to-tabular feature transfer, not performance after one-hot training.

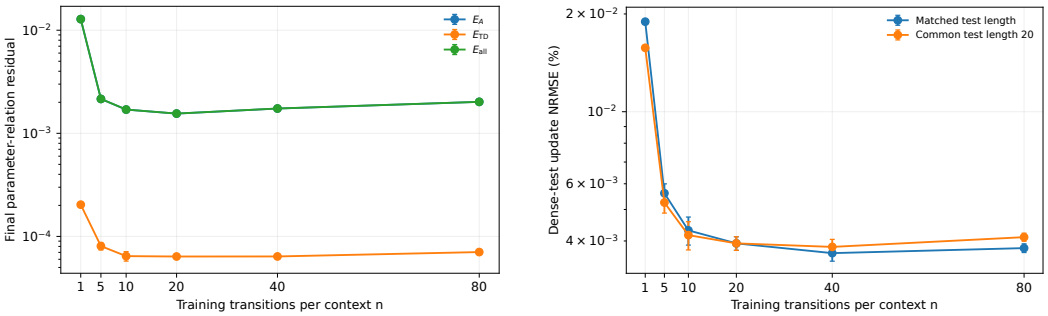


Figure 11: Training context length at $\tau = 0.2$ and $B = 32$, with 20.48 million tasks per model. Left: final parameter residuals. Right: update errors on matched-length and common-length dense-feature tests.

C.3 Generalization to unseen tabular tasks

Returns and uncertainty. At $\tau = 0.2$, the Transformer achieves mean return 0.4880, compared with 0.4861 for its soft teacher and 0.5374 for Oracle. The paired Transformer-minus-soft difference is 0.00188. This comparison does not resolve a return advantage and does not establish equivalence or noninferiority. The visit-count online Q-learning baseline reaches 0.5318, exceeding the Transformer means in the primary comparison. The evidence therefore supports useful self-fed transfer (Figure 12). Deployment intervals use 5,000 percentile-bootstrap replicates with independent resampling of the training-seed and test-task axes, preserving paired methods and complete trajectories.

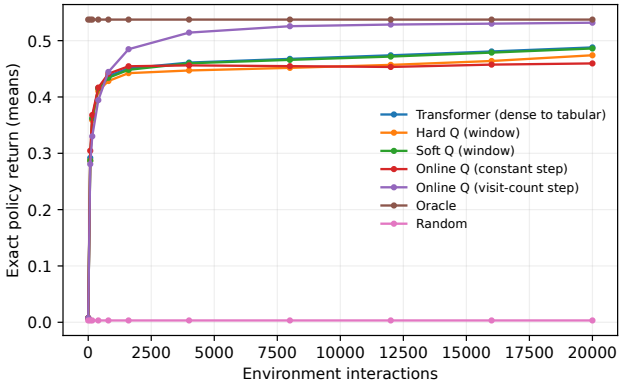


Figure 12: Online Q-learning is included in the comparisons; the variant with decaying visit-count learning rate yields the best performance.

Tabular learning curves across temperatures Table 12 and Figure 13 complete the five-temperature comparison; $\tau = 0.2$ appears in the main text. All curves use the same task bank, interaction axis, and evaluation protocol, and include Transformer, hard window-Q, soft window-Q, Oracle, and Random.

Table 12: Final returns on 512 unseen tabular tasks: $n_{\text{test}} = 20$, zero Q initialization.

τ	Transformer [95% CI]	Hard Q (window)	Soft Q (window)	Oracle	Random
1	0.5036 [0.4835, 0.5237]	0.4740	0.5036	0.5374	0.0033
0.8	0.5033 [0.4829, 0.5227]	0.4740	0.5025	0.5374	0.0033
0.5	0.5003 [0.4804, 0.5198]	0.4740	0.4996	0.5374	0.0033
0.2	0.4880 [0.4686, 0.5074]	0.4740	0.4861	0.5374	0.0033
0.1	0.4833 [0.4644, 0.5022]	0.4740	0.4795	0.5374	0.0033

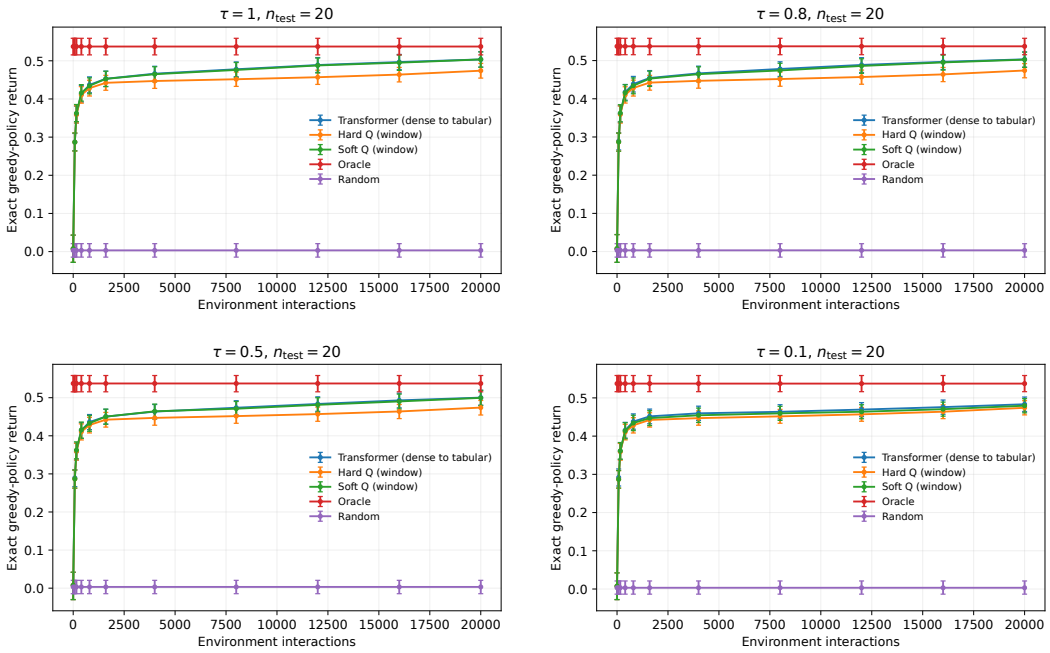


Figure 13: Tabular transfer at $\tau = 1$, $\tau = 0.8$, $\tau = 0.5$, and $\tau = 0.1$, with $n_{\text{test}} = 20$, zero Q initialization, and 20,000 interactions. Error bars denote pointwise 95% intervals. Transformer weights are frozen throughout. All experiments use the same 512 tasks and 20,000-interaction budget.

Training and deployment context lengths We distinguish three evaluations. The frozen-length test varies n_{test} for models trained with $n_{\text{train}} = 20$. The matched-length test deploys each length-specific model at its training length. The common-length test deploys all length-specific models at $n_{\text{test}} = 20$. Every learner receives 20,000 interactions, so a window-based learner performs $20,000/n_{\text{test}}$ updates. Varying the deployment length therefore changes both the window-averaging coefficient and the update frequency.

In the frozen-length test, $n_{\text{test}} = 5$ yields the highest observed Transformer mean return, 0.5194, whereas $n_{\text{test}} = 20$ yields 0.4880. The window teachers exhibit a similar dependence on deployment length. With common $n_{\text{test}} = 20$, the length-specific models instead have mean returns between 0.4868 and 0.4884. Thus, training-context effects should be distinguished from deployment update frequency; longer windows do not uniformly improve finite-budget control.

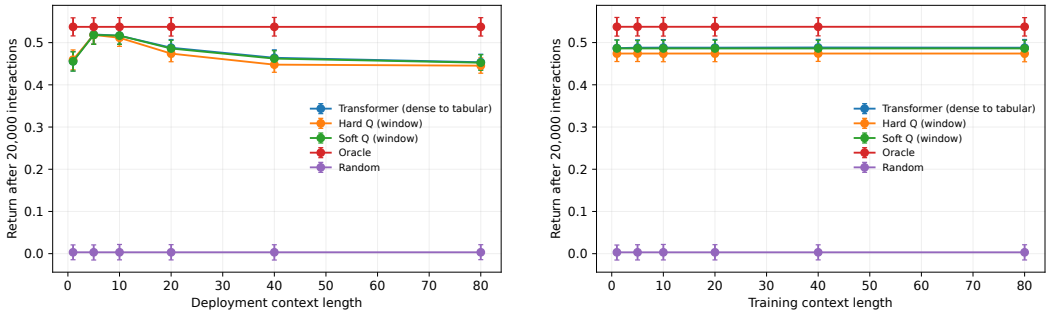


Figure 14: Separating training-length and deployment-length effects at $\tau = 0.2$. Left: models trained at $n_{\text{train}} = 20$ and deployed with different windows. Right: models trained at different lengths and all deployed at $n_{\text{test}} = 20$. Both evaluations use the same task bank and 20,000-interaction budget.

D THEORETICAL ANALYSIS

Fix integers $d \geqslant 1$, $m \geqslant 2$, $n \geqslant 1$ and real constants $\alpha > 0$, $0 < \gamma < 1$, $\tau > 0$. The meanings of these fixed quantities are given below. The temperature τ is fixed during optimization, let

$$\mathbf{z} = \big(\mathbf{w}, \{\phi_i, r_{i+1}, \psi_{i,a} : 0 \leqslant i < n,\ a \in [m]\}\big) \sim \mathcal{D}_n.$$

Here $\mathbf{w} \in \mathbb{R}^d$ is an input vector, held fixed within a context. The vector $\phi_i \in \mathbb{R}^d$ is the current state–action feature, $r_{i+1} \in \mathbb{R}$ is the observed reward, and $\psi_{i,a} \in \mathbb{R}^d$ is the feature of candidate action a at the successor state. These are allowed to be statistically dependent. The symbol $\mathcal{D}_n$ denotes their joint probability law, let

$$\pi^\tau_{i,a}(\mathbf{z}) = \frac{\exp(\mathbf{w}^\top \psi_{i,a}/\tau)}{\sum_{b=1}^m \exp(\mathbf{w}^\top \psi_{i,b}/\tau)}, \qquad \bar{\psi}^\tau_i(\mathbf{z}) = \sum_{a=1}^m \pi^\tau_{i,a}(\mathbf{z})\psi_{i,a}, \tag{D.1}$$

$$\delta^\tau_i(\mathbf{z}) = r_{i+1} - \mathbf{w}^\top \phi_i + \gamma \mathbf{w}^\top \bar{\psi}^\tau_i(\mathbf{z}), \qquad F_{\tau,n}(\mathbf{z}) = \frac{\alpha}{n} \sum_{i=0}^{n-1} \phi_i \delta^\tau_i(\mathbf{z}). \tag{D.2}$$

The probabilities $\pi^\tau_{i,a}$ average successor features; they are not the behavior policy that generated the trajectory. The scalar δ^τ_i is the specified temporal-difference target minus the current score. The vector $F_{\tau,n} \in \mathbb{R}^d$ is the update increment to imitate. Thus the teacher's full prediction is $\mathbf{w} + F_{\tau,n}$.

In this appendix, we explicitly restrict optimization to the six-block support of Proposition B.3, with every other block fixed at zero. The six-block support $\boldsymbol{\theta}_{\sup} = (\mathbf{A}_{\boldsymbol{\theta}}, \mathbf{L}, \mathbf{B}_-, \mathbf{B}_+, \mathbf{B}_r, \mathbf{C}) \in \Theta_{\sup}$ can be broken down into following blocks

$$\boldsymbol{\theta} = (A, L, b_-, K_-, b_+, K_+, \beta, q, C) \in \Theta, \qquad \dim \Theta = 5d^2 + 3d + 1. \tag{D.3}$$

where $A = \mathbf{A}_\theta$, $L = \mathbf{L}$, $\mathbf{B}_- = [b_-, K_-]$, $\mathbf{B}_+ = [b_+, K_+]$, $\mathbf{B}_r = [\beta, q^\top]$. The trainable matrix L is not assumed invertible; invertibility is imposed only on the exact-realization manifold.

$$p_{i,a}(A; \mathbf{z}) = \frac{\exp(\psi_{i,a}^\top A \mathbf{w})}{\sum_{b=1}^m \exp(\psi_{i,b}^\top A \mathbf{w})}, \quad \bar{\psi}_i(A; \mathbf{z}) = \sum_a p_{i,a}(A; \mathbf{z})\psi_{i,a}, \tag{D.4}$$

$$s_i(\boldsymbol{\theta}; \mathbf{z}) = \phi_i^\top (b_- + K_- \mathbf{w}) + (L\bar{\psi}_i(A; \mathbf{z}))^\top (b_+ + K_+ \mathbf{w}) + r_{i+1}(\beta + q^\top \mathbf{w}), \tag{D.5}$$

$$F_{\boldsymbol{\theta},n}(\mathbf{z}) = \frac{1}{n} \sum_{i=0}^{n-1} C\phi_i s_i(\boldsymbol{\theta}; \mathbf{z}). \tag{D.6}$$

Both full outputs add the same $\mathbf{w}$, so their difference is the increment error $e_{\boldsymbol{\theta}}(\mathbf{z}) := F_{\boldsymbol{\theta},n}(\mathbf{z}) - F_{\tau,n}(\mathbf{z})$. For the fixed context length n, we write the loss as $L_n(\theta) = \mathcal{L}(\theta)$, i.e.

$$L_n(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_n} \|F_{\boldsymbol{\theta},n}(\mathbf{z}) - F_{\tau,n}(\mathbf{z})\|_2^2. \tag{D.7}$$

We prove Theorem 2 in three parts. First, we characterize the local geometry and normal curvature of the optimal parameter manifold. Second, we establish a one-step contraction of gradient descent toward this manifold. Finally, we use these results to prove convergence to a single optimal parameter and derive the matching worst-case iteration complexity over the admissible initialization class.

D.1 GEOMETRY AND CURVATURE OF THE OPTIMAL PARAMETER MANIFOLD

A perturbation $\mathbf{U} \in \Theta$ has blocks denoted by $\Delta A, \ldots, \Delta C$ in the same order as (D.3).

$$\begin{aligned}\|\mathbf{U}\|_\Theta^2 =& \|\Delta A\|_F^2 + \|\Delta L\|_F^2 + \|\Delta b_-\|_2^2 + \|\Delta K_-\|_F^2 + \|\Delta b_+\|_2^2 \\ &+ \|\Delta K_+\|_F^2 + |\Delta\beta|^2 + \|\Delta q\|_2^2 + \|\Delta C\|_F^2.\end{aligned}$$

For a vector, $\|x\|_2^2 = \sum_j x_j^2$; for a matrix, $\|M\|_F^2 = \operatorname{tr}(M^\top M)$. The product inner product is the corresponding sum of vector dot products and matrix Frobenius inner products. We abbreviate $\|\mathbf{U}\|_\Theta$ to $\|\mathbf{U}\|$ when the space is unambiguous. Norms of operators are induced by their domain and codomain norms, not entrywise norms unless specified.

For a fixed context $\mathbf{z}$, let $J_{\boldsymbol{\theta},n}(\mathbf{z}) := D_{\boldsymbol{\theta}} F_{\boldsymbol{\theta},n}(\mathbf{z}) : \Theta \to \mathbb{R}^d$ denote the Jacobian of the Transformer output with respect to its parameters. Thus, for a parameter perturbation $\mathbf{U} \in \Theta$, $J_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U}] = \frac{d}{dt} F_{\boldsymbol{\theta}+t\mathbf{U},n}(\mathbf{z})\big|_{t=0}$. Equivalently, $F_{\boldsymbol{\theta}+t\mathbf{U},n}(\mathbf{z}) = F_{\boldsymbol{\theta},n}(\mathbf{z}) + tJ_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U}] + o(t)$. Hence $J_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U}]$ measures the first-order change of the Transformer output, on the particular context $\mathbf{z}$, when the parameters are perturbed in direction $\mathbf{U}$.

Since our training objective is the population loss (D.7), we now introduce the population Jacobian. We therefore consider the Hilbert space

$$\mathcal{H}_n = L^2(\mathcal{D}_n; \mathbb{R}^d), \quad \langle f, h\rangle_{\mathcal{H}_n} = \mathbb{E}\langle f(\mathbf{z}), h(\mathbf{z})\rangle_2,$$

with norm $|f|^2_{\mathcal{H}_n} = \mathbb{E}|f(\mathbf{z})|_2^2$, and define the population Jacobian $\mathcal{J}_{\boldsymbol{\theta},n} : \Theta \to \mathcal{H}_n$, $(\mathcal{J}_{\boldsymbol{\theta},n}\mathbf{U})(\mathbf{z}) = J_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U}]$. Thus $|\mathcal{J}_{\boldsymbol{\theta},n}\mathbf{U}|^2_{\mathcal{H}_n} = \mathbb{E}|J_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U}]|_2^2$ measures the average squared first-order output change caused by the parameter perturbation $\mathbf{U}$ over the entire context distribution.

Let $\mathcal{J}^*_{\boldsymbol{\theta},n}$ denote the adjoint of the population Jacobian, defined by $\langle \mathcal{J}_{\boldsymbol{\theta},n}\mathbf{U}, f\rangle_{\mathcal{H}_n} = \langle \mathbf{U}, \mathcal{J}^*_{\boldsymbol{\theta},n} f\rangle_\Theta$. Equivalently, $\mathcal{J}^*_{\boldsymbol{\theta},n} f = \mathbb{E}\big[J_{\boldsymbol{\theta},n}(\mathbf{z})^* f(\mathbf{z})\big]$.

At an exact realization $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$, define the population Gauss-Newton operator $G(\bar{\boldsymbol{\theta}}) := \mathcal{J}^*_{\bar{\boldsymbol{\theta}},n}\mathcal{J}_{\bar{\boldsymbol{\theta}},n}$. For every parameter direction $\mathbf{U}$,

$$\langle \mathbf{U}, G(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle = |\mathcal{J}_{\bar{\boldsymbol{\theta}},n}\mathbf{U}|^2_{\mathcal{H}_n} = \mathbb{E}|J_{\bar{\boldsymbol{\theta}},n}(\mathbf{z})[\mathbf{U}]|_2^2.$$

Therefore $G(\bar{\boldsymbol{\theta}})$ quantifies how strongly the Transformer output responds, to first order, to perturbations of its parameters around an exact realization. The minimum normal curvature over the chosen target patch $\mathcal{U}_\tau \subset \mathcal{M}_\tau$ is defined by $\kappa := \inf_{\bar{\boldsymbol{\theta}}\in\mathcal{U}_\tau, \mathcal{U}_\tau \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau, |\mathbf{U}|=1} |\mathcal{J}_{\bar{\boldsymbol{\theta}},n}\mathbf{U}|^2_{\mathcal{H}_n}$.

The quantity κ has a direct optimization interpretation. A large κ means that every unit perturbation away from the exact-realization manifold produces a substantial change in the Transformer output, so the population loss has appreciable curvature in every identifiable parameter direction. A small κ means that at least one identifiable direction is nearly flat and is therefore harder to learn by training.

Throughout the remainder of Appendix D, n and τ are fixed. We use the abbreviations $J_\theta = J_{\theta,n}$, $G = G_n$, $\mathcal{H} = \mathcal{H}_n$, and $\mathsf{H} = \mathsf{H}_n(\tau)$. Below we mainly analyze the geometric and curvature information of the optimal parameter manifold, as detailed in Lemmas D.1, D.2 and D.3 below.

Structure of the optimal parameter manifold. The structure of the optimal parameter manifold establishes that every parameter on the manifold exactly realizes the target update and achieves zero loss, that the collection of such optimal parameters forms a smooth embedded manifold, and that the Hessian $\nabla^2 L_n(\bar{\boldsymbol{\theta}})$ is degenerate exactly along the tangent directions of this manifold, which means motion within $\mathcal{M}_\tau$ preserves zero loss, whereas every nonzero normal direction has strictly positive Hessian curvature.

Lemma D.1 (Structure of the optimal parameter manifold)**.** *For every $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}_\tau(L, c) \in \mathcal{M}_\tau$, $F_{\bar{\boldsymbol{\theta}},n}(\mathbf{z}) = F_{\tau,n}(\mathbf{z})$ for every admissible input $\mathbf{z}$, and*

(i) $L_n(\bar{\boldsymbol{\theta}}) = 0 = \min_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$, and $\mathcal{M}_\tau \subseteq \arg\min_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$.

(ii) $\mathcal{M}_\tau$ is a closed smooth embedded submanifold of Θ, $\dim(\mathcal{M}_\tau) = d^2 + 1$.

(iii) Under Assumption 1, $\ker G_n(\bar{\boldsymbol{\theta}}) = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ and $\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = G_n(\bar{\boldsymbol{\theta}})$, where $\ker T$ denotes the set of directions that the operator T maps to zero, i.e., the directions along which T has no effect.

Proof. (i) Since the population loss is $L_n(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\left[\|F_{\boldsymbol{\theta},n}(\mathbf{z}) - F_{\tau,n}(\mathbf{z})\|_2^2\right]$, Theorem 1 immediately gives $L_n(\bar{\boldsymbol{\theta}}) = 0$ and $L_n(\boldsymbol{\theta}) \geqslant 0$ for every $\boldsymbol{\theta} \in \Theta$. Therefore $L_n(\bar{\boldsymbol{\theta}}) = 0 = \min_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$. Since $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$ was arbitrary, $\mathcal{M}_\tau \subseteq \arg\min_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$.

(ii) Define $\mathcal{G} := GL(d) \times (\mathbb{R} \setminus \{0\})$. Since $GL(d) = \{L \in \mathbb{R}^{d\times d} : \det(L) \neq 0\}$, and the determinant map is continuous, $GL(d)$ is an open subset of $\mathbb{R}^{d\times d} \simeq \mathbb{R}^{d^2}$. Likewise, $\mathbb{R} \setminus \{0\}$ is open in $\mathbb{R}$. Therefore $\mathcal{G} \subset \mathbb{R}^{d^2+1}$ is an open smooth manifold of dimension $\dim(\mathcal{G}) = d^2 + 1$.

Define the parameterization $\Phi_\tau : \mathcal{G} \to \Theta$, $\Phi_\tau(L, c) := \bar{\boldsymbol{\theta}}_\tau(L, c)$. Then, by definition, $\mathcal{M}_\tau = \Phi_\tau(\mathcal{G})$. Matrix inversion $L \mapsto L^{-1}$ is smooth on $GL(d)$, while the scalar map $c \mapsto c^{-1}$ is smooth on $\mathbb{R} \setminus \{0\}$. Consequently, $(L, c) \mapsto c\gamma L^{-1}$, $(L, c) \mapsto \frac{\alpha}{c} I_d$ are smooth, and hence Φ_τ is a smooth map.

In order to prove $\mathcal{M}_\tau = \Phi_\tau(\mathcal{G})$ is an embedded submanifold of Θ, it suffices to verify that Φ_τ is an injective immersion which is a homeomorphism onto its image. Since $\Phi_\tau(L_1, c_1) = \Phi_\tau(L_2, c_2) \to (L_1, c_1) = (L_2, c_2)$, Φ_τ is injective. For a perturbation $(Z, \xi) \in \mathbb{R}^{d\times d} \times \mathbb{R} \simeq T_{(L,c)}\mathcal{G}$, where $T_{(L,c)}\mathcal{G}$ is the tangent space of $\mathcal{G}$ at the point (L, c), the differential of Φ_τ is

$$D\Phi_\tau(L, c)[Z, \xi] = \Big(0,\ Z,\ 0,\ -\xi I_d,\ 0, \gamma\xi L^{-\top} - c\gamma L^{-\top} Z^\top L^{-\top},\ \xi,\ 0,\ -\alpha\xi c^{-2} I_d\Big).$$

This follows from differentiating $L^{-\top} L^\top = I_d$ in the direction Z, which gives $D(L^{-\top})[Z] = -L^{-\top} Z^\top L^{-\top}$. Suppose $D\Phi_\tau(L, c)[Z, \xi] = 0$, we have $(Z, \xi) = (0, 0)$, which means $D\Phi_\tau(L, c)$ is injective at every $(L, c) \in \mathcal{G}$, i.e. Φ_τ is an immersion. For any $\boldsymbol{\theta} = (A, L, b_-, K_-, b_+, K_+, \beta, q, C) \in \mathcal{M}_\tau$, The inverse parameterization on $\mathcal{M}_\tau$ is simply $\Phi_\tau^{-1}(\boldsymbol{\theta}) = (L,\ \beta)$. Since coordinate projections are continuous, Φ_τ^{-1} is continuous on its image. Therefore, $\mathcal{M}_\tau = \Phi_\tau(\mathcal{G})$ is an embedded submanifold of Θ, and $\dim(\mathcal{M}_\tau) = d^2 + 1$ by the fact that an embedding preserves dimension.

It remains to show that $\mathcal{M}_\tau$ is closed as a subset of Θ. Consider an arbitrary convergent sequence $\boldsymbol{\theta}_j = \bar{\boldsymbol{\theta}}_\tau(L_j, c_j) \in \mathcal{M}_\tau$, $\boldsymbol{\theta}_j \to \boldsymbol{\theta}_\star$ in Θ. We show that $\boldsymbol{\theta}_\star \in \mathcal{M}_\tau$. Since convergence in Θ implies convergence of every parameter block, $L_j \to L_\star$. Since the seventh block satisfies $\beta_j = c_j$, $c_j \to c_\star := \beta_\star$. We need to verify that $c_\star \neq 0$ and $L_\star \in GL(d)$.

Suppose, for contradiction, that $c = 0$, then $c_j \to 0$. However, the ninth parameter block is $C_j = \frac{\alpha}{c_j} I_d$. Since $\alpha > 0$, $\|C_j\|_F = \frac{\alpha}{|c_j|}\|I_d\|_F = \frac{\alpha\sqrt{d}}{|c_j|} \to \infty$. This contradicts the assumed convergence of $\boldsymbol{\theta}_j$ in the finite-dimensional parameter space Θ, thus $c_\star \neq 0$.

Similarly, suppose that $L_\star$ is singular. Since $L_j \in GL(d)$, $L_j \to L_\star$, with $L_\star$ singular, the smallest singular value satisfies $\sigma_{\min}(L_j) \to 0$. Consequently, $\|L_j^{-1}\|_{\mathrm{op}} = \frac{1}{\sigma_{\min}(L_j)} \to \infty$. On the other hand, $c_j \to c_\star \neq 0$, so for sufficiently large j, $|c_j| \geqslant \frac{|c_\star|}{2} > 0$. The sixth parameter block is $K_{+,j} = c_j \gamma L_j^{-\top}$. Therefore $\|K_{+,j}\|_{\mathrm{op}} = \gamma|c_j|\,\|L_j^{-\top}\|_{\mathrm{op}} = \frac{\gamma|c_j|}{\sigma_{\min}(L_j)} \longrightarrow \infty$, contradicting convergence of the sixth block of $\boldsymbol{\theta}_j$, thus $L_\star \in GL(d)$. Since $L_\star \in GL(d)$ and $c_\star \neq 0$, the pair $(L_\star, c_\star)$ belongs to $\mathcal{G}$. Moreover, continuity of matrix inversion on $GL(d)$ gives $L_j^{-1} \to L_\star^{-1}$, and continuity of $c_\star \mapsto 1/c_\star$ gives $\frac{1}{c_j} \to \frac{1}{c_\star}$.

Therefore every block of $\boldsymbol{\theta}_j$ converges to the corresponding block of $\bar{\boldsymbol{\theta}}_\tau(L, c)$, so that $\boldsymbol{\theta}_\star = \bar{\boldsymbol{\theta}}_\tau(L_\star, c_\star) \in \mathcal{M}_\tau$. Hence every convergent sequence in $\mathcal{M}_\tau$ whose limit exists in Θ has its limit again in $\mathcal{M}_\tau$. Since Θ is finite dimensional, this proves that $\mathcal{M}_\tau$ is closed.

(iii) We prove $\ker G(\bar{\boldsymbol{\theta}}) = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ and $\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = G(\bar{\boldsymbol{\theta}}) := J_{\bar{\boldsymbol{\theta}}}^* J_{\bar{\boldsymbol{\theta}}}$ in the following four steps.

Step 1: Tangent space of the optimal parameter manifold $\mathcal{M}_\tau$.

Consider a smooth curve $t \mapsto (L(t), c(t))$ in $GL(d) \times (\mathbb{R} \setminus \{0\})$ such that $L(0) = L$, $c(0) = c$. Write $Z := \dot{L}(0)$, $\xi := \dot{c}(0)$. Differentiating $L(t)^{-1}L(t) = I_d$ at $t = 0$ gives

$$\frac{d}{dt} L(t)^{-\top}\Big|_{t=0} = -L^{-\top} Z^\top L^{-\top}.$$

Hence differentiating $\bar{\boldsymbol{\theta}}_\tau(L(t), c(t))$ gives

$$\begin{aligned}
&\Delta A = 0, \quad \Delta L = Z, \quad \Delta b_- = 0, \quad \Delta K_- = -\xi I_d,\\
&\Delta b_+ = 0, \quad \Delta q = 0, \quad \Delta\beta = \xi, \quad \Delta C = -\alpha\xi c^{-2} I_d,\\
&\Delta K_+ = \frac{d}{dt}\left[\gamma c(t) L(t)^{-\top}\right]_{t=0} = \gamma\xi L^{-\top} - c\gamma L^{-\top} Z^\top L^{-\top}.
\end{aligned}$$

Therefore

$$T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau = \Big\{\big(0, Z, 0, -\xi I_d, 0, \gamma\xi L^{-\top} - c\gamma L^{-\top} Z^\top L^{-\top}, \xi, 0, -\alpha\xi c^{-2} I_d\big) : Z \in \mathbb{R}^{d\times d},\ \xi \in \mathbb{R}\Big\}. \tag{D.8}$$

These are precisely the infinitesimal gauge directions.

Step 2: Sample Jacobian and collect the observable parameter combinations.

Recall $p_{i,a}(A)$ and $\bar{\psi}_i(A)$ defined in (D.4). Let $\Sigma_i^\tau = \sum_a \pi_{i,a}^\tau(\psi_{i,a} - \bar{\psi}_i^\tau)(\psi_{i,a} - \bar{\psi}_i^\tau)^\top$. The softmax derivative gives $Dp_{i,a}(\tau^{-1}I_d)[\Delta A] = \pi_{i,a}^\tau(\psi_{i,a} - \bar{\psi}_i^\tau)^\top \Delta A\, \mathbf{w}$. Consequently,

$$D\bar{\psi}_i(\tau^{-1}I_d)[\Delta A] = \sum_a \psi_{i,a} Dp_{i,a}(\tau^{-1}I_d)[\Delta A] = \Sigma_i^\tau \Delta A\, \mathbf{w}. \tag{D.9}$$

Recall $s_i(\boldsymbol{\theta}; \mathbf{z}) = \phi_i^\top(b_- + K_-\mathbf{w}) + b_+^\top L\bar{\psi}_i(A) + \mathbf{w}^\top K_+^\top L\bar{\psi}_i(A) + r_{i+1}(\beta + q^\top \mathbf{w})$ defined in (D.5). At $\bar{\theta} = \bar{\theta}_\tau(L, c)$, we have $b_+ = 0$ and $K_+ = c\gamma L^{-\top}$. For an arbitrary perturbation $\mathbf{U}$, write $\xi := \Delta\beta$. Using the product rule and (D.9), we obtain

$$\begin{aligned}\dot{s}_i[\mathbf{U}] =& \phi_i^\top(\Delta b_- + \Delta K_- \mathbf{w}) + (\bar{\psi}_i^\tau)^\top \left[L^\top \Delta b_+ + \left(L^\top \Delta K_+ + c\gamma \Delta L^\top L^{-\top}\right)\mathbf{w}\right] \\ &+ r_{i+1}\left(\xi + \Delta q^\top \mathbf{w}\right) + \gamma c\, \mathbf{w}^\top \Sigma_i^\tau \Delta A\, \mathbf{w}.\end{aligned} \tag{D.10}$$

Since $F_{\boldsymbol{\theta},n}(\mathbf{z}) = \frac{1}{n}\sum_i C\phi_i s_i(\boldsymbol{\theta}; \mathbf{z})$ and $C = \frac{\alpha}{c} I_d$, $s_i(\bar{\boldsymbol{\theta}}; \mathbf{z}) = c\delta_i^\tau$, the sample Jacobian satisfies

$$J_{\bar{\boldsymbol{\theta}}}(\mathbf{z})[\mathbf{U}] = \frac{1}{n}\sum_i \left[c\Delta C\, \phi_i \delta_i^\tau + \frac{\alpha}{c}\phi_i \dot{s}_i[\mathbf{U}]\right]. \tag{D.11}$$

Then we collect the observable parameter combinations. Define the linear map $Q_{L,c} : \Theta \to \mathcal{V}$, $Q_{L,c}\mathbf{U} = (E_C, e_q, e_-, E_-, e_+, E_+, E_A)$, by

$$E_C = \frac{c^2}{\alpha}\Delta C + \xi I_d, \quad e_q = \Delta q, \quad e_- = \Delta b_-, \quad E_- = \Delta K_- + \xi I_d, \tag{D.12}$$

$$e_+ = L^\top \Delta b_+, \quad E_A = \gamma c\, \Delta A, \quad E_+ = c\gamma \Delta L^\top L^{-\top} + L^\top \Delta K_+ - \gamma\xi I_d. \tag{D.13}$$

Subtracting $\xi\delta_i^\tau$ from (D.10), where $\delta_i^\tau = r_{i+1} - \mathbf{w}^\top\phi_i + \gamma \mathbf{w}^\top \bar{\psi}_i^\tau$, gives

$$\dot{s}_i[\mathbf{U}] - \xi\delta_i^\tau = \phi_i^\top(e_- + E_-\mathbf{w}) + (\bar{\psi}_i^\tau)^\top(e_+ + E_+\mathbf{w}) + r_{i+1}e_q^\top \mathbf{w} + \mathbf{w}^\top \Sigma_i^\tau E_A \mathbf{w}.$$

Likewise, $c\Delta C = \frac{\alpha}{c}(E_C - \xi I_d)$. Substitution into (D.11) shows that the terms proportional to $\xi\delta_i^\tau$ cancel exactly. Thus

$$J_{\bar{\boldsymbol{\theta}}}(\mathbf{z})[\mathbf{U}] = \frac{\alpha}{c}\mathcal{B}_{\tau,n}(\mathbf{z})[Q_{L,c}\mathbf{U}]. \tag{D.14}$$

Step 3: The kernels of $Q_{L,c}$ and $G_n(\bar{\boldsymbol{\theta}})$ are exactly the tangent space.

Suppose $Q_{L,c}\mathbf{U} = 0$. Then $E_A = e_q = e_- = e_+ = 0$. Since $\gamma > 0$, $c \neq 0$, and L is invertible, this implies $\Delta A = 0$, $\Delta q = 0$, $\Delta b_- = 0$, $\Delta b_+ = 0$. Moreover, $E_C = 0$ implies $\Delta C = -\alpha\xi c^{-2} I_d$, while $E_- = 0$ gives $\Delta K_- = -\xi I_d$. Finally, $E_+ = 0$ gives $L^\top \Delta K_+ = \gamma\xi I_d - c\gamma \Delta L^\top L^{-\top}$. Then $\Delta K_+ = \gamma\xi L^{-\top} - c\gamma L^{-\top}\Delta L^\top L^{-\top}$. Set $Z := \Delta L$. Together with the preceding identities, this is precisely the tangent form in (D.8). Hence $\ker Q_{L,c} \subset T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$.

Conversely, take any $\mathbf{U} \in T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$. Then for some Z and ξ, direct substitution of (D.8) into (D.12)-(D.13) gives $Q_{L,c}\mathbf{U} = 0$. Therefore $T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau \subset \ker Q_{L,c}$, which means

$$\ker Q_{L,c} = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau. \tag{D.15}$$

Since $\mathcal{H} = L^2(\mathcal{D}_n; \mathbb{R}^d)$ is the population output space, and $G(\bar{\boldsymbol{\theta}}) = J_{\bar{\boldsymbol{\theta}}}^* J_{\bar{\boldsymbol{\theta}}}$, then, for every perturbation $\mathbf{U}$, by (D.14), we have

$$\langle \mathbf{U}, G(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle = \|J_{\bar{\boldsymbol{\theta}}}\mathbf{U}\|_{\mathcal{H}}^2 = \frac{\alpha^2}{c^2}\,\mathbb{E}\left[\|\mathcal{B}_{\tau,n}(\mathbf{z})[Q_{L,c}\mathbf{U}]\|_2^2\right] = \frac{\alpha^2}{c^2}\left\langle Q_{L,c}\mathbf{U},\, H\, Q_{L,c}\mathbf{U}\right\rangle_{\mathcal{V}}, \tag{D.16}$$

where $H = \mathbb{E}[\mathcal{B}_{\tau,n}(\mathbf{z})^* \mathcal{B}_{\tau,n}(\mathbf{z})]$. By Assumption 1 (ii), $H \succ 0$. Therefore $\langle Q_{L,c}\mathbf{U}, HQ_{L,c}\mathbf{U}\rangle = 0$ if and only if $Q_{L,c}\mathbf{U} = 0$. Since $G(\bar{\boldsymbol{\theta}}) = J_{\bar{\boldsymbol{\theta}}}^* J_{\bar{\boldsymbol{\theta}}}$ is positive semi-definite,

$$\langle \mathbf{U}, G(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle = 0 \quad \Longleftrightarrow \quad G(\bar{\boldsymbol{\theta}})\mathbf{U} = 0.$$

Combining this with (D.16) and (D.15), we obtain

$$\ker G(\bar{\boldsymbol{\theta}}) = \ker Q_{L,c} = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau. \tag{D.17}$$

Step 4: Prove $\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = G(\bar{\boldsymbol{\theta}})$.

Recall that $L_n(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\|e_{\boldsymbol{\theta}}(\mathbf{z})\|_2^2$, where $e_{\boldsymbol{\theta}}(\mathbf{z}) := F_{\boldsymbol{\theta},n}(\mathbf{z}) - F_{\tau,n}(\mathbf{z})$. Under Assumption 1 (i), the data are bounded. On a sufficiently small bounded parameter neighborhood, softmax smoothness gives common integrable bounds on the predictor and its first two derivatives. Hence differentiation under the expectation is valid. For arbitrary directions $\mathbf{U}, \mathbf{V}$,

$$D^2 L_n(\boldsymbol{\theta})[\mathbf{U},\mathbf{V}] = \mathbb{E}\left[\langle J_{\boldsymbol{\theta}}(\mathbf{z})[\mathbf{U}], J_{\boldsymbol{\theta}}(\mathbf{z})[\mathbf{V}]\rangle\right] + \mathbb{E}\left[\left\langle e_{\boldsymbol{\theta}}(\mathbf{z}), D^2 F_{\boldsymbol{\theta},n}(\mathbf{z})[\mathbf{U},\mathbf{V}]\right\rangle\right]. \tag{D.18}$$

By part (i), at every correct parameter $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$ we have the samplewise identity $e_{\bar{\boldsymbol{\theta}}}(\mathbf{z}) = 0$. Therefore the second term in (D.18) vanishes identically, and

$$D^2 L_n(\bar{\boldsymbol{\theta}})[\mathbf{U},\mathbf{V}] = \mathbb{E}\left[\langle J_{\bar{\boldsymbol{\theta}}}(\mathbf{z})[\mathbf{U}], J_{\bar{\boldsymbol{\theta}}}(\mathbf{z})[\mathbf{V}]\rangle\right] = \langle \mathbf{U}, J_{\bar{\boldsymbol{\theta}}}^* J_{\bar{\boldsymbol{\theta}}} \mathbf{V}\rangle.$$

Hence, as operators,

$$\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = J_{\bar{\boldsymbol{\theta}}}^* J_{\bar{\boldsymbol{\theta}}} = G(\bar{\boldsymbol{\theta}}). \tag{D.19}$$

Combining (D.19) with (D.17) yields

$$\ker \nabla^2 L_n(\bar{\boldsymbol{\theta}}) = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau.$$

Equivalently, if $\mathbf{U} \notin T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$, then $\langle \mathbf{U}, \nabla^2 L_n(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle > 0$. In particular, for every nonzero normal direction $\mathbf{U} \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau \setminus \{0\}$, we have $\langle \mathbf{U}, \nabla^2 L_n(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle > 0$. □

Uniform normal curvature of the optimal parameter manifold By Lemma D.1, we have $\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = G(\bar{\boldsymbol{\theta}})$. It follows that the curvature κ defined in (16) characterizes the weakest second-order growth of the loss when moving away from the optimal manifold.

Lemma D.2 shows that the loss has uniformly positive curvature in all normal, non-gauge directions around the optimal parameter manifold, and that this curvature is quantitatively comparable to the data-richness constant: $\kappa \asymp \chi := \lambda_{\min}(\mathsf{H}_n(\tau))$, where $\mathsf{H} := \mathbb{E}\left[\mathcal{B}_{\tau,n}(\mathbf{z})^* \mathcal{B}_{\tau,n}(\mathbf{z})\right]$. Thus, stronger data identifiability leads to stronger local curvature and faster contraction toward the optimal manifold.

Recall that $\overline{\boldsymbol{\Omega}} \Subset GL(d) \times (\mathbb{R} \setminus \{0\}), \mathcal{U}_\tau := \left\{\bar{\boldsymbol{\theta}}_\tau(L,c) : (L,c) \in \boldsymbol{\Omega}\right\} \subset \mathcal{M}_\tau$.

Lemma D.2 (Uniform normal curvature)**.** *Under Assumptions 1, there exists constants $0 < c_\Omega \leqslant C_\Omega < \infty$ such that, uniformly over $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}_\tau(L,c) \in \mathcal{U}_\tau$, $c_\Omega \chi \leqslant \lambda_{\min}(\nabla^2 L_n(\bar{\boldsymbol{\theta}})|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}) \leqslant C_\Omega \chi$. Equivalently, $0 < c_\Omega \chi \leqslant \kappa \leqslant C_\Omega \chi$.*

Proof. Fix $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}_\tau(L,c) \in \mathcal{U}_\tau$. From (D.14), $J_{\bar{\boldsymbol{\theta}}}(\mathbf{z})[\mathbf{U}] = \frac{\alpha}{c}\mathcal{B}_{\tau,n}(\mathbf{z})[Q_{L,c}\mathbf{U}]$, where $Q_{L,c} : T_{\bar{\boldsymbol{\theta}}}\Theta \to \mathcal{V}$ is the linear map collecting the effective observable coefficients. Moreover, $\ker Q_{L,c} = T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$.

We first show that the restriction $Q_{L,c}|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau} : N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau \to \mathcal{V}$ is an isomorphism (a bijective linear map that preserves linear operations). Since $N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau \cap \ker Q_{L,c} = \{0\}$, the restriction is injective. Furthermore, $\dim N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau = (5d^2+3d+1) - (d^2+1) = 4d^2+3d$. The dimension of observable coefficient space is $\dim \mathcal{V} = 4d^2 + 3d$, thus the restriction is bijective. For completeness, one may construct an explicit right inverse $R_{L,c} : \mathcal{V} \to T_{\bar{\boldsymbol{\theta}}}\Theta$. Specifically, for $\zeta = (E_C, e_q, e_-, E_-, e_+, E_+, E_A) \in \mathcal{V}$, set

$$\begin{aligned}
&\Delta A = (\gamma c)^{-1} E_A, && \Delta L = 0, && \Delta\beta = 0,\\
&\Delta C = \alpha c^{-2} E_C, && \Delta b_- = e_-, && \Delta K_- = E_-,\\
&\Delta q = e_q, && \Delta b_+ = L^{-\top} e_+, && \Delta K_+ = L^{-\top} E_+.
\end{aligned}$$

Then $Q_{L,c}R_{L,c} = I_{\mathcal{V}}$. Let P_N denote the orthogonal projection onto $N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$. Since $Q_{L,c}$ vanishes on $T_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$, $Q_{L,c}P_N = Q_{L,c}$. Therefore $P_N R_{L,c}$ is the inverse of $Q_{L,c}|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}$.

Since $\overline{\boldsymbol{\Omega}}$ is compact and $L,\ L^{-1},\ c,\ c^{-1}$ are uniformly bounded on $\overline{\boldsymbol{\Omega}}$, the operator norms of $Q_{L,c}$ and $R_{L,c}$ are uniformly bounded. Hence there exist constants $0 < t_- \leqslant t_+ < \infty$ such that, for every $\bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau$ and every $\mathbf{U} \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$,

$$t_-\|\mathbf{U}\| \leqslant \|Q_{L,c}\mathbf{U}\|_{\mathcal{V}} \leqslant t_+\|\mathbf{U}\|. \tag{D.20}$$

where one may take $t_- = (\sup_{(L,c)\in\overline{\boldsymbol{\Omega}}} \|R_{L,c}\|_{\mathrm{op}})^{-1}$, $t_+ = \sup_{(L,c)\in\overline{\boldsymbol{\Omega}}} \|Q_{L,c}\|_{\mathrm{op}}$.

We now transfer Assumption 1 from coefficient space to parameter space. Since $\mathsf{H} \succeq \chi I_{\mathcal{V}}$, we can obtain $\langle Q_{L,c}\mathbf{U},\ \mathsf{H}Q_{L,c}\mathbf{U}\rangle \geqslant \chi\|Q_{L,c}\mathbf{U}\|_{\mathcal{V}}^2$ by (D.16). Thus, for every unit normal vector

$\mathbf{U} \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$, $\|\mathbf{U}\| = 1$, yields $\langle \mathbf{U}, G(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle \geqslant \frac{\alpha^2}{c^2}\chi t_-^2$ from (D.20). Let $c_+ := \max_{(L,c)\in\overline{\boldsymbol{\Omega}}} |c|$, then $\frac{1}{c^2} \geqslant \frac{1}{c_+^2}$, and hence $\langle \mathbf{U}, G(\bar{\boldsymbol{\theta}})\mathbf{U}\rangle \geqslant \frac{\alpha^2 t_-^2}{c_+^2}\chi$. Therefore, by taking $c_\Omega = \frac{\alpha^2 t_-^2}{c_+^2}$, we have

$$\lambda_{\min}(G(\bar{\boldsymbol{\theta}})|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}) \geqslant c_\Omega \chi. \tag{D.21}$$

Let $\zeta_* \in \mathcal{V}$, $\|\zeta_*\|_{\mathcal{V}} = 1$, be a unit eigenvector of H corresponding to its smallest eigenvalue $\mathsf{H}\zeta_* = \chi\zeta_*$. Since $Q_{L,c}|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}$ is invertible, define $\mathbf{U}_* = (Q_{L,c}|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau})^{-1}\zeta_*$. Then $Q_{L,c}\mathbf{U}_* = \zeta_*$. By the upper estimate in (D.20), $1 = \|Q_{L,c}\mathbf{U}_*\|_{\mathcal{V}} \leqslant t_+\|\mathbf{U}_*\|$, so $\|\mathbf{U}_*\| \geqslant \frac{1}{t_+}$. Using (D.16),

$$\langle \widehat{\mathbf{U}}_*, G(\bar{\boldsymbol{\theta}})\widehat{\mathbf{U}}_*\rangle = \frac{\alpha^2}{c^2}\frac{\langle \zeta_*, \mathsf{H}\zeta_*\rangle}{\|\mathbf{U}_*\|^2} = \frac{\alpha^2}{c^2}\frac{\chi}{\|\mathbf{U}_*\|^2} \leqslant \frac{\alpha^2 t_+^2}{c^2}\chi, \quad \widehat{\mathbf{U}}_* = \frac{\mathbf{U}_*}{\|\mathbf{U}_*\|}$$

Let $c_- := \min_{(L,c)\in\overline{\boldsymbol{\Omega}}} |c| > 0$. Then $\frac{1}{c^2} \leqslant \frac{1}{c_-^2}$, and therefore, by taking $C_\Omega = \frac{\alpha^2 t_+^2}{c_-^2}$,

$$\lambda_{\min}(G(\bar{\boldsymbol{\theta}})|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}) \leqslant C_\Omega \chi. \tag{D.22}$$

Finally, by Lemma D.1, $\nabla^2 L_n(\bar{\boldsymbol{\theta}}) = G(\bar{\boldsymbol{\theta}})$ for every $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$. Combining (D.21) and (D.22) gives

$$c_\Omega \chi \leqslant \lambda_{\min}(\nabla^2 L_n(\bar{\boldsymbol{\theta}})|_{N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau}) \leqslant C_\Omega \chi.$$

Taking the infimum over $\bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau$ and over unit normal directions yields $c_\Omega\chi \leqslant \kappa \leqslant C_\Omega\chi$. Since $\chi > 0$ by Assumption 1 and $c_\Omega > 0$, we conclude that $\kappa > 0$. □

Local tubular geometry and unique projection. Lemma D.3 provides a precise local notion of distance to the optimal parameter manifold.

For any parameter $\boldsymbol{\theta}$ sufficiently close to $\mathcal{M}_\tau$, define,

$$d(\boldsymbol{\theta}) := \operatorname{dist}(\boldsymbol{\theta}, \mathcal{M}_\tau) = \inf_{\boldsymbol{\vartheta}\in\mathcal{M}_\tau} \|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|.$$

Lemma D.3 guarantees that this distance is attained at a unique nearest point $\bar{\boldsymbol{\theta}} = \Pi(\boldsymbol{\theta})$. Moreover, the displacement $\Delta = \boldsymbol{\theta} - \Pi(\boldsymbol{\theta})$ is normal to the manifold, i.e. $d(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \Pi(\boldsymbol{\theta})\|$.

Lemma D.3 (Local tubular geometry and unique projection)**.** *There exists a constant $r_{\text{geo}} > 0$ such that the following statements hold.*

(i) For every $\bar{\boldsymbol{\theta}} \in \overline{\mathcal{U}}_\tau$ and $\Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ with $\|\Delta\| < r_{\text{geo}}$, the point $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \Delta$ has $\bar{\boldsymbol{\theta}}$ as its unique nearest point on the full manifold $\mathcal{M}_\tau$.

(ii) Define the normal tube $\mathcal{T}_{r_{\text{geo}}}(\mathcal{U}_\tau) = \{\bar{\boldsymbol{\theta}} + \Delta : \bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau,\ \Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau,\ \|\Delta\| < r_{\text{geo}}\}$. For every $\boldsymbol{\theta} \in \mathcal{T}_{r_{\text{geo}}}(\mathcal{U}_\tau)$, the nearest-point projection $\Pi(\boldsymbol{\theta})$ onto $\mathcal{M}_\tau$ is uniquely defined and satisfies $\boldsymbol{\theta} - \Pi(\boldsymbol{\theta}) \in N_{\Pi(\boldsymbol{\theta})}\mathcal{M}_\tau$ and $d(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \Pi(\boldsymbol{\theta})\|$. Moreover, $\mathcal{T}_{r_{\text{geo}}}(\mathcal{U}_\tau)$ is an open neighborhood of $\mathcal{U}_\tau$ in the ambient parameter space.

Proof. By Lemma D.1, $\mathcal{M}_\tau$ is a smooth closed embedded submanifold of the finite-dimensional Euclidean parameter space, and $\overline{\mathcal{U}}_\tau$ is compact. By the tubular neighborhood theorem (Lee, 2013, Theorem 6.24), every point sufficiently close to $\mathcal{M}_\tau$ can be uniquely represented as $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \Delta$, where $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$ and $\Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ is a sufficiently small normal displacement.

For each $\bar{\boldsymbol{\theta}} \in \overline{\mathcal{U}}_\tau$, there exist a neighborhood $\mathcal{V}_{\bar{\boldsymbol{\theta}}} \subset \mathcal{M}_\tau$ and a radius $r_{\bar{\boldsymbol{\theta}}} > 0$ such that the above normal representation remains unique for all normal perturbations of size smaller than $r_{\bar{\boldsymbol{\theta}}}$ from points in $\mathcal{V}_{\bar{\boldsymbol{\theta}}}$. Since $\overline{\mathcal{U}}_\tau$ is compact, finitely many such neighborhoods, say $\mathcal{V}_1, \ldots, \mathcal{V}_L$, cover $\overline{\mathcal{U}}_\tau$. Define $\mathcal{V} = \bigcup_{j=1}^L \mathcal{V}_j$ and $r_0 = \min_{1\leqslant j\leqslant L} r_j$. Since the minimum is taken over finitely many positive radii, $r_0 > 0$. Thus every normal perturbation of norm smaller than r_0 from a point in $\mathcal{V}$ lies in the region where the normal representation is unique.

(i) We choose the radius small enough so that the nearest point cannot leave $\mathcal{V}$. Since $\overline{\mathcal{U}}_\tau \subset \mathcal{V}$, $\overline{\mathcal{U}}_\tau$ is compact, and $\mathcal{M}_\tau \setminus \mathcal{V}$ is closed, we have $\delta := \operatorname{dist}(\overline{\mathcal{U}}_\tau, \mathcal{M}_\tau \setminus \mathcal{V}) > 0$. Choose $r_{\text{geo}} > 0$ such that $r_{\text{geo}} < \min\{r_0, \delta/2\}$. Now take $\bar{\boldsymbol{\theta}} \in \overline{\mathcal{U}}_\tau$ and $\Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ with $\|\Delta\| < r_{\text{geo}}$, and set $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \Delta$. Since $\mathcal{M}_\tau$ is closed, there exists at least one nearest point $\boldsymbol{\theta}^\star \in \mathcal{M}_\tau$ to $\boldsymbol{\theta}$. Because $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$,

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\| \leqslant \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| = \|\Delta\| < r_{\text{geo}}.$$

The triangle inequality then gives $\|\boldsymbol{\theta}^\star - \bar{\boldsymbol{\theta}}\| < 2r_{\mathrm{geo}} < \delta$. Hence $\boldsymbol{\theta}^\star \in \mathcal{V}$.

Since $\boldsymbol{\theta}^\star$ minimizes the Euclidean distance from $\boldsymbol{\theta}$ to $\mathcal{M}_\tau$, the first-order optimality condition implies $\boldsymbol{\theta} - \boldsymbol{\theta}^\star \in N_{\boldsymbol{\theta}^\star}\mathcal{M}_\tau$. Therefore $\boldsymbol{\theta}$ has two normal representations: $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \Delta$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^\star + (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$. Both base points lie in $\mathcal{V}$, and both normal displacements have norm smaller than r_0. By uniqueness of the tubular representation, we must have $\boldsymbol{\theta}^\star = \bar{\boldsymbol{\theta}}$. Thus $\bar{\boldsymbol{\theta}}$ is the unique nearest point to $\boldsymbol{\theta}$ on the full manifold $\mathcal{M}_\tau$. Consequently,

$$\Pi(\boldsymbol{\theta}) = \bar{\boldsymbol{\theta}}, \qquad \boldsymbol{\theta} - \Pi(\boldsymbol{\theta}) \in N_{\Pi(\boldsymbol{\theta})}\mathcal{M}_\tau, \qquad d(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \Pi(\boldsymbol{\theta})\|.$$

(ii) Since $\mathcal{U}_\tau$ is relatively open in $\mathcal{M}_\tau$, the normal perturbations of norm smaller than r_{geo} based at points of $\mathcal{U}_\tau$ form an open set in the tubular coordinates. The tubular-neighborhood map sends this set to an open subset of the ambient parameter space containing $\mathcal{U}_\tau$. Hence the corresponding normal tube is an open neighborhood of $\mathcal{U}_\tau$. □

D.2 A Direct Parameter-Distance Estimate for One GD Step

The identity $d(\boldsymbol{\theta}) = \|\Delta\|$ is essential for the convergence analysis, where $\Delta = \boldsymbol{\theta} - \Pi(\boldsymbol{\theta})$ is the normal displacement from the current parameter to its nearest point on the optimal manifold. The operator $\mathcal{J}_{\Pi(\boldsymbol{\theta}),n}[\Delta]$ describes the first-order change of the model output when the parameter moves from the nearest optimal point $\Pi(\boldsymbol{\theta})$ in the direction Δ. Since Δ is normal to the manifold, the normal-curvature bound in Lemma D.2 can be applied directly, giving

$$\left\|\mathcal{J}_{\Pi(\boldsymbol{\theta}),n}[\Delta]\right\|^2_{L^2(D_n)} \geqslant \kappa\|\Delta\|^2 = \kappa d(\boldsymbol{\theta})^2.$$

Lemma D.4 shows that one gradient-descent step moves the current parameter closer to the optimal manifold. We first prove that the gradient has a sufficiently large component in the direction that reduces the normal displacement from the manifold. We then control the overall gradient norm to ensure that the update does not overshoot. Combining these two estimates yields a one-step contraction of the distance to $\mathcal{M}_\tau$.

Lemma D.4 (One-step contraction toward the optimal manifold)**.** *Under Assumption 1,*

(i) There exist constants $M_J > 0$, $L_J > 0$, and $0 < \rho_0 < r_{\mathrm{geo}}$ such that, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T}_{\rho_0}(\mathcal{U}_\tau)$, $\|\mathcal{J}_{\boldsymbol{\theta},n}\|_{\mathrm{op}} \leqslant M_J$, and $\|\mathcal{J}_{\boldsymbol{\theta},n} - \mathcal{J}_{\boldsymbol{\theta}',n}\|_{\mathrm{op}} \leqslant L_J\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

(ii) Let $r := \min\left\{\rho_0, \frac{\sqrt{\kappa}}{4L_J}\right\}$, $\mathcal{T} := \mathcal{T}_r(\mathcal{U}_\tau)$. For any $\boldsymbol{\theta} \in \mathcal{T}$ and every step size $0 < \eta \leqslant 1/(2M_J^2)$, $d\left(\boldsymbol{\theta} - \eta\nabla L_n(\boldsymbol{\theta})\right)^2 \leqslant \left(1 - \frac{\eta\kappa}{4}\right) d(\boldsymbol{\theta})^2$, and $\eta\,\|\nabla L_n(\boldsymbol{\theta})\| \leqslant \frac{4M_J}{\sqrt{\kappa}}\left[d(\boldsymbol{\theta}) - d\left(\boldsymbol{\theta} - \eta\nabla L_n(\boldsymbol{\theta})\right)\right]$, where $1 - \eta\kappa/4 \in [7/8, 1)$.

Proof. (i) Write the parameter as $\boldsymbol{\theta} = (A, x)$, where x collects all parameter blocks other than A. Since $\overline{\mathcal{U}}_\tau$ is compact, there exists $M_0 < \infty$ such that $\|x\| \leqslant M_0$ on $\overline{\mathcal{U}}_\tau$. Choose $0 < \rho_0 < \min\{r_{\mathrm{geo}}, 1\}$ sufficiently small so that $\mathcal{T}_{\rho_0}(\mathcal{U}_\tau)$ and the line segment joining each point in this tube to its projection onto $\mathcal{M}_\tau$ are contained in

$$\mathcal{C} = \left\{(A, x) : A \in \mathbb{R}^{d\times d},\ \|x\| < M_0 + 1\right\}.$$

Although A is unrestricted in $\mathcal{C}$, it enters the model through the softmax logits. The derivatives of the softmax are uniformly bounded, the features and weights are bounded by Assumption 1, and all remaining parameter blocks are bounded on $\mathcal{C}$. Hence there exist constants $M_J > 0$ and $L_J > 0$ such that, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T}_{\rho_0}(\mathcal{U}_\tau)$,

$$\|\mathcal{J}_{\boldsymbol{\theta},n}\|_{\mathrm{op}} \leqslant M_J, \qquad \|\mathcal{J}_{\boldsymbol{\theta},n} - \mathcal{J}_{\boldsymbol{\theta}',n}\|_{\mathrm{op}} \leqslant L_J\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|. \tag{D.23}$$

Since κ is the minimum curvature over unit normal directions on the selected optimal patch, the first bound in (D.23) also gives $\kappa \leqslant M_J^2$. For a fixed gauge domain Ω and temperature-independent data bounds, the constants M_J and L_J can be chosen independently of τ. Indeed, the non-A coordinates of the realization patch do not depend on τ, while A enters the model only through the softmax logits, whose derivatives with respect to those logits are uniformly bounded.

(ii) Fix $\boldsymbol{\theta} \in \mathcal{T}_r(\mathcal{U}_\tau)$ and let $\bar{\boldsymbol{\theta}} = \Pi(\boldsymbol{\theta})$ and $\Delta = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$. By Lemma D.3, $\Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$ and $\|\Delta\| = d(\boldsymbol{\theta})$. If $d(\boldsymbol{\theta}) = 0$, then $\boldsymbol{\theta} \in \mathcal{M}_\tau$ and $\nabla L_n(\boldsymbol{\theta}) = 0$, so the conclusions are immediate. Hence assume

$d(\boldsymbol{\theta}) > 0$. Define $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta \nabla L_n(\boldsymbol{\theta})$ and $d_+ := d(\boldsymbol{\theta}^+)$. Since $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$ is a valid comparison point in the definition of d_+,

$$d_+^2 \leqslant \left\|\boldsymbol{\theta} - \eta \nabla L_n(\boldsymbol{\theta}) - \bar{\boldsymbol{\theta}}\right\|^2 = d(\boldsymbol{\theta})^2 - 2\eta \left\langle \nabla L_n(\boldsymbol{\theta}), \Delta \right\rangle + \eta^2 \|\nabla L_n(\boldsymbol{\theta})\|^2. \tag{D.24}$$

Thus, it is enough to show that the directional component $\langle \nabla L_n(\boldsymbol{\theta}), \Delta \rangle$ is uniformly positive and that the total gradient norm is controlled by this same quantity. We establish these two bounds below.

For $\langle \nabla L_n(\boldsymbol{\theta}), \Delta \rangle$. Since $\bar{\boldsymbol{\theta}} \in \mathcal{M}_\tau$, $F_{\bar{\boldsymbol{\theta}},n} = F_{\tau,n}$, and the fundamental theorem of calculus gives

$$F_{\boldsymbol{\theta},n} - F_{\tau,n} = \int_0^1 \mathcal{J}_{\bar{\boldsymbol{\theta}}+t\Delta,n}[\Delta]\, dt. \tag{D.25}$$

Moreover, for every parameter direction $\mathbf{U}$,

$$\langle \nabla L_n(\boldsymbol{\theta}), \mathbf{U} \rangle = \left\langle F_{\boldsymbol{\theta},n} - F_{\tau,n}, \mathcal{J}_{\boldsymbol{\theta},n}[\mathbf{U}] \right\rangle_{L^2(\mathcal{D}_n)}. \tag{D.26}$$

Because Δ is normal to $\mathcal{M}_\tau$, Lemma D.2 gives $\left\|\mathcal{J}_{\bar{\boldsymbol{\theta}},n}[\Delta]\right\|_{L^2(\mathcal{D}_n)} \geqslant \sqrt{\kappa}\, d(\boldsymbol{\theta})$. Using the Jacobian Lipschitz bound in (D.23),

$$\|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|_{L^2(\mathcal{D}_n)} \geqslant \left\|\mathcal{J}_{\bar{\boldsymbol{\theta}},n}[\Delta]\right\|_{L^2(\mathcal{D}_n)} - L_J \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \|\Delta\| \geqslant \sqrt{\kappa}\, d(\boldsymbol{\theta}) - L_J d(\boldsymbol{\theta})^2.$$

Since $d(\boldsymbol{\theta}) < r$ and $r \leqslant \sqrt{\kappa}/(4L_J)$,

$$\|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|_{L^2(\mathcal{D}_n)} \geqslant \frac{3}{4}\sqrt{\kappa}\, d(\boldsymbol{\theta}). \tag{D.27}$$

Using (D.25), write

$$F_{\boldsymbol{\theta},n} - F_{\tau,n} = \mathcal{J}_{\boldsymbol{\theta},n}[\Delta] + R, \tag{D.28}$$

where $R := \int_0^1 \left(\mathcal{J}_{\bar{\boldsymbol{\theta}}+t\Delta,n} - \mathcal{J}_{\boldsymbol{\theta},n}\right)[\Delta]\, dt$. By the Jacobian Lipschitz bound,

$$\|R\|_{L^2(\mathcal{D}_n)} \leqslant \int_0^1 L_J (1-t) \|\Delta\|^2\, dt = \frac{L_J}{2} d(\boldsymbol{\theta})^2 \leqslant \frac{1}{8}\sqrt{\kappa}\, d(\boldsymbol{\theta}). \tag{D.29}$$

Combining this with (D.27) gives $\|R\|_{L^2(\mathcal{D}_n)} \leqslant \frac{1}{6}\|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|_{L^2(\mathcal{D}_n)}$. Taking $\mathbf{U} = \Delta$ in (D.26) and using (D.28), we have

$$\langle \nabla L_n(\boldsymbol{\theta}), \Delta \rangle = \left\langle \mathcal{J}_{\boldsymbol{\theta},n}[\Delta] + R, \mathcal{J}_{\boldsymbol{\theta},n}[\Delta] \right\rangle_{L^2(\mathcal{D}_n)} \geqslant \frac{5}{6} \|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|^2_{L^2(\mathcal{D}_n)}. \tag{D.30}$$

Together with (D.27), this yields

$$\langle \nabla L_n(\boldsymbol{\theta}), \Delta \rangle \geqslant \frac{15}{32} \kappa d(\boldsymbol{\theta})^2. \tag{D.31}$$

For $\|\nabla L_n(\boldsymbol{\theta})\|$. By (D.28), $\|F_{\boldsymbol{\theta},n} - F_{\tau,n}\|_{L^2(\mathcal{D}_n)} \leqslant \frac{7}{6}\|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|_{L^2(\mathcal{D}_n)}$. Using (D.26), (D.23), and (D.30), we obtain

$$\|\nabla L_n(\boldsymbol{\theta})\|^2 \leqslant \frac{49}{30} M_J^2 \left\langle \nabla L_n(\boldsymbol{\theta}), \Delta \right\rangle. \tag{D.32}$$

We now return to the distance expansion (D.24). Using (D.32),

$$d_+^2 \leqslant d(\boldsymbol{\theta})^2 - \eta \left(2 - \frac{49}{30}\eta M_J^2\right) \left\langle \nabla L_n(\boldsymbol{\theta}), \Delta \right\rangle. \tag{D.33}$$

If $\eta \leqslant 1/(2M_J^2)$, then $2 - (49/30)\eta M_J^2 \geqslant 71/60 > 1$. Therefore, by (D.31),

$$d_+^2 \leqslant d(\boldsymbol{\theta})^2 - \eta \left\langle \nabla L_n(\boldsymbol{\theta}), \Delta \right\rangle \leqslant \left(1 - \frac{15}{32}\eta\kappa\right) d(\boldsymbol{\theta})^2 \leqslant \left(1 - \frac{\eta\kappa}{4}\right) d(\boldsymbol{\theta})^2. \tag{D.34}$$

Since $\kappa \leqslant M_J^2$ and $\eta \leqslant 1/(2M_J^2)$, the contraction factor satisfies $1 - \eta\kappa/4 \in [7/8, 1)$.

It remains to control the step length. From (D.33) and the same step-size condition,

$$\eta \left\langle \nabla L_n(\boldsymbol{\theta}), \Delta \right\rangle \leqslant d(\boldsymbol{\theta})^2 - d_+^2 \leqslant 2d(\boldsymbol{\theta}) \left(d(\boldsymbol{\theta}) - d_+\right), \tag{D.35}$$

where $d_+ \leqslant d(\boldsymbol{\theta})$ follows from (D.34). Furthermore, by (D.30), (D.27), and the bound $\|F_{\boldsymbol{\theta},n} - F_{\tau,n}\|_{L^2(\mathcal{D}_n)} \leqslant (7/6)\|\mathcal{J}_{\boldsymbol{\theta},n}[\Delta]\|_{L^2(\mathcal{D}_n)}$,

$$\frac{\|\nabla L_n(\boldsymbol{\theta})\|}{\langle \nabla L_n(\boldsymbol{\theta}), \Delta\rangle} \leqslant \frac{28M_J}{15\sqrt{\kappa}\, d(\boldsymbol{\theta})}. \tag{D.36}$$

Combining (D.35) and (D.36) yields

$$\eta\|\nabla L_n(\boldsymbol{\theta})\| \leqslant \frac{56M_J}{15\sqrt{\kappa}}\,(d(\boldsymbol{\theta}) - d_+) \leqslant \frac{4M_J}{\sqrt{\kappa}}\,[d(\boldsymbol{\theta}) - d\,(\boldsymbol{\theta} - \eta\nabla L_n(\boldsymbol{\theta}))]\,.$$

This proves both the one-step contraction and the step-length estimate. □

D.3 The proof of Theorem 2

Let $q_\eta := 1 - \eta\kappa/4$. By Lemma D.4, if $0 < \eta \leqslant 1/(2M_J^2)$, then $q_\eta \in (0, 1)$.

Convergence of the full parameter sequence and its limit. Let $d_k := d(\boldsymbol{\theta}_k)$. Assume $d_0 > 0$. Define $\Lambda_0 := \frac{4M_J}{\sqrt{\kappa}}\, d_0$. By the initialization condition, $\Lambda_0 < b_{\mathcal{T}}(\boldsymbol{\theta}_0)$, where $b_{\mathcal{T}}(\boldsymbol{\theta}) := \operatorname{dist}(\boldsymbol{\theta}, \mathcal{T}^c)$. Therefore the closed ball $\overline{B}(\boldsymbol{\theta}_0, \Lambda_0)$ is contained in $\mathcal{T}$.

We prove inductively that every iterate and every update segment remain in this ball and that $d_k \leqslant q_\eta^{k/2} d_0$. Suppose $\boldsymbol{\theta}_k \in \mathcal{T}$. By Lemma D.4,

$$d_{k+1} \leqslant \sqrt{q_\eta}\, d_k, \qquad \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\| \leqslant \frac{4M_J}{\sqrt{\kappa}}(d_k - d_{k+1}). \tag{D.37}$$

The first inequality gives $d_{k+1} \leqslant q_\eta^{(k+1)/2} d_0$.

Now consider any point $\boldsymbol{y} = \boldsymbol{\theta}_k + s(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k)$, with $0 \leqslant s \leqslant 1$, on the new update segment. Using (D.37) and the induction hypothesis,

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{\theta}_0\| &\leqslant \sum_{j=0}^{k-1} \|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\| + s\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\| \\ &\leqslant \frac{4M_J}{\sqrt{\kappa}} \sum_{j=0}^{k}(d_j - d_{j+1}) = \frac{4M_J}{\sqrt{\kappa}}(d_0 - d_{k+1}) \leqslant \Lambda_0.\end{aligned} \tag{D.38}$$

Hence the entire update segment remains in $\overline{B}(\boldsymbol{\theta}_0, \Lambda_0) \subset \mathcal{T}$. Every iterate remains in $\mathcal{T}$, and $d_k \leqslant q_\eta^{k/2} d_0$ holds for all $k \geqslant 0$.

The decay $d_k \to 0$ alone does not guarantee convergence of the full parameter sequence, since the iterates could in principle drift along the optimal manifold. The step-length estimate in (D.37) rules out this possibility. For integers $\ell > k$, summing the step-length bound gives

$$\sum_{j=k}^{\ell-1} \|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\| \leqslant \frac{4M_J}{\sqrt{\kappa}}(d_k - d_\ell) \leqslant \frac{4M_J}{\sqrt{\kappa}} d_k \leqslant \Lambda_0 q_\eta^{k/2}. \tag{D.39}$$

The right-hand side tends to zero as $k \to \infty$. Hence $(\boldsymbol{\theta}_k)$ is a Cauchy sequence. Since the parameter space is finite-dimensional, there exists $\boldsymbol{\theta}_\infty$ such that $\boldsymbol{\theta}_k \to \boldsymbol{\theta}_\infty$.

Letting $\ell \to \infty$ in (D.39) yields

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_\infty\| \leqslant \Lambda_0 q_\eta^{k/2}. \tag{D.40}$$

Together with $d_k \leqslant q_\eta^{k/2} d_0$, this proves geometric convergence both to the optimal manifold and to a single parameter limit. The distance function $d(\boldsymbol{\theta}) = \operatorname{dist}(\boldsymbol{\theta}, \mathcal{M}_\tau)$ is 1-Lipschitz. Since $\boldsymbol{\theta}_k \to \boldsymbol{\theta}_\infty$ and $d_k \to 0$, we obtain $d(\boldsymbol{\theta}_\infty) = 0$. Because $\mathcal{M}_\tau$ is closed, $\boldsymbol{\theta}_\infty \in \mathcal{M}_\tau$.

Moreover, all iterates lie in the closed ball $\overline{B}(\boldsymbol{\theta}_0, \Lambda_0)$, so the limit also lies in this ball and therefore in $\mathcal{T}$. Since $\mathcal{T} = \mathcal{T}_r(\mathcal{U}_\tau)$, there exist $\bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau$ and $\Delta \in N_{\bar{\boldsymbol{\theta}}}\mathcal{M}_\tau$, with $\|\Delta\| < r$, such that $\boldsymbol{\theta}_\infty = \bar{\boldsymbol{\theta}} + \Delta$.

By Lemma D.3, $\bar{\boldsymbol{\theta}}$ is the unique nearest point of $\boldsymbol{\theta}_\infty$ on $\mathcal{M}_\tau$. But $\boldsymbol{\theta}_\infty \in \mathcal{M}_\tau$, so $\boldsymbol{\theta}_\infty$ itself is a nearest point. Uniqueness therefore implies $\Delta = 0$ and $\boldsymbol{\theta}_\infty = \bar{\boldsymbol{\theta}} \in \mathcal{U}_\tau$. Hence there exist $(L_\infty, c_\infty) \in \boldsymbol{\Omega}$ such that $\boldsymbol{\theta}_\infty = \bar{\boldsymbol{\theta}}_\tau(L_\infty, c_\infty)$. By exact realization on $\mathcal{M}_\tau$, $F_{\boldsymbol{\theta}_\infty,n} = F_{\tau,n}$, and therefore $L_n(\boldsymbol{\theta}_\infty) = 0 = \min_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$.

Iteration-complexity bound. By $d_k \leqslant q_\eta^{k/2} d_0$, $d(\boldsymbol{\theta}_K) \leqslant q_\eta^{K/2} d_0$. Since $q_\eta = 1 - \eta\kappa/4$ and $1 - x \leqslant e^{-x}$, $d(\boldsymbol{\theta}_K) \leqslant \exp(-\frac{\eta\kappa K}{8}) d_0$. Thus $K \geqslant \frac{8}{\eta\kappa}\log(1/\varepsilon)$ is sufficient to guarantee $d(\boldsymbol{\theta}_K) \leqslant \varepsilon d_0$. Consequently, $K_{\mathrm{GD}}(\varepsilon; \boldsymbol{\theta}_0) \leqslant \left\lceil \frac{8}{\eta\kappa} \log \frac{1}{\varepsilon} \right\rceil$.

We prove the lower bound by considering a small perturbation along the normal direction with the smallest local curvature. This direction is the locally slowest contraction direction of gradient descent.

Choose an interior point $\bar{\boldsymbol{\theta}}^\circ \in \mathcal{U}_\tau$ and define

$$\ell := \lambda_{\min}\left(\nabla^2 L_n(\bar{\boldsymbol{\theta}}^\circ)\big|_{N_{\bar{\boldsymbol{\theta}}^\circ}\mathcal{M}_\tau}\right). \tag{D.41}$$

Thus, ℓ is the smallest Hessian curvature over the normal directions at $\bar{\boldsymbol{\theta}}^\circ$. By Lemma D.2, $\kappa \leqslant \ell \leqslant (C_\Omega/c_\Omega)\kappa$. At an optimal parameter, $\nabla^2 L_n(\bar{\boldsymbol{\theta}}^\circ) = \mathcal{J}^*_{\bar{\boldsymbol{\theta}}^\circ,n}\mathcal{J}_{\bar{\boldsymbol{\theta}}^\circ,n}$. Hence the local Jacobian bound gives $\ell \leqslant M_J^2$. Since the Hessian is symmetric and $\ker(\nabla^2 L_n(\bar{\boldsymbol{\theta}}^\circ)) = T_{\bar{\boldsymbol{\theta}}^\circ}\mathcal{M}_\tau$, the normal space is invariant under the Hessian. Therefore, there exists a unit vector $\mathbf{U}^\circ \in N_{\bar{\boldsymbol{\theta}}^\circ}\mathcal{M}_\tau$ such that

$$\nabla^2 L_n(\bar{\boldsymbol{\theta}}^\circ)\mathbf{U}^\circ = \ell\mathbf{U}^\circ. \tag{D.42}$$

Thus, $\mathbf{U}^\circ$ is a normal direction with the smallest local curvature.

For $t > 0$, define $\boldsymbol{\theta}_t = \bar{\boldsymbol{\theta}}^\circ + t\mathbf{U}^\circ$. For sufficiently small t, Lemma D.3 gives $d(\boldsymbol{\theta}_t) = t$. We first verify that these perturbations are admissible initializations. Let $b^\circ := b_{\mathcal{T}}(\bar{\boldsymbol{\theta}}^\circ) > 0$. Since the distance to $\mathcal{T}^c$ is 1-Lipschitz, $b_{\mathcal{T}}(\boldsymbol{\theta}_t) \geqslant b^\circ - t$. Thus, if t is sufficiently small so that $t < r$, $t < b^\circ/2$, and $t < b^\circ\sqrt{\kappa}/(8M_J)$, then

$$\frac{4M_J}{\sqrt{\kappa}} d(\boldsymbol{\theta}_t) < b_{\mathcal{T}}(\boldsymbol{\theta}_t). \tag{D.43}$$

Hence all sufficiently small nonzero $\boldsymbol{\theta}_t$ are admissible initializations.

We next determine how fast gradient descent contracts this perturbation. Define $T_\eta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta\nabla L_n(\boldsymbol{\theta})$. Since $\nabla L_n(\bar{\boldsymbol{\theta}}^\circ) = 0$, the point $\bar{\boldsymbol{\theta}}^\circ$ is a fixed point of T_η. Moreover, $DT_\eta(\bar{\boldsymbol{\theta}}^\circ) = I - \eta\nabla^2 L_n(\bar{\boldsymbol{\theta}}^\circ)$. Therefore, using (D.42), for every fixed finite integer K,

$$D(T_\eta^K)(\bar{\boldsymbol{\theta}}^\circ)\mathbf{U}^\circ = (1 - \eta\ell)^K \mathbf{U}^\circ. \tag{D.44}$$

Hence, to first order, the perturbation along $\mathbf{U}^\circ$ is reduced by the factor $(1 - \eta\ell)^K$ after K gradient-descent steps.

We now make this local linearization precise. Fix K and then let t tend to zero. Since L_n is twice continuously differentiable in the local regularity region, T_η is continuously differentiable there. For sufficiently small t, the first K iterates remain in this region, and

$$T_\eta^K(\boldsymbol{\theta}_t) = \bar{\boldsymbol{\theta}}^\circ + t(1 - \eta\ell)^K\mathbf{U}^\circ + R_K(t), \qquad \frac{\|R_K(t)\|}{t} \longrightarrow 0 \quad \text{as } t \downarrow 0. \tag{D.45}$$

Here K is fixed before taking the limit $t \downarrow 0$. Since $\ell \leqslant M_J^2$ and $\eta \leqslant 1/(2M_J^2)$, we have $0 < \eta\ell \leqslant 1/2$, and therefore $(1 - \eta\ell)^K > 0$. For sufficiently small t, Lemma D.3 gives $d(\bar{\boldsymbol{\theta}}^\circ + t(1 - \eta\ell)^K\mathbf{U}^\circ) = t(1 - \eta\ell)^K$. Because the distance to $\mathcal{M}_\tau$ is 1-Lipschitz, (D.45) implies

$$\lim_{t\downarrow 0} \frac{d(T_\eta^K(\boldsymbol{\theta}_t))}{d(\boldsymbol{\theta}_t)} = (1 - \eta\ell)^K. \tag{D.46}$$

Suppose that K iterations guarantee $d(T_\eta^K(\boldsymbol{\theta}_0)) \leqslant \varepsilon d(\boldsymbol{\theta}_0)$ for every admissible initialization. Since all sufficiently small $\boldsymbol{\theta}_t$ constructed above are admissible, (D.46) implies $(1 - \eta\ell)^K \leqslant \varepsilon$. Therefore, $K \geqslant \log(1/\varepsilon)/[-\log(1 - \eta\ell)]$. Since $0 < \eta\ell \leqslant 1/2$ and $-\log(1 - x) \leqslant 2x$ for $0 < x \leqslant 1/2$, we obtain

$$\sup_{\boldsymbol{\theta}^0 \in \mathcal{A}_\tau} K_{\mathrm{GD}}(\varepsilon; \theta^0) \geqslant \frac{1}{2\eta\ell} \log \frac{1}{\varepsilon} \geqslant \frac{c_\Omega}{2C_\Omega} \frac{1}{\eta\kappa} \log \frac{1}{\varepsilon}. \tag{D.47}$$

E VERIFICATION FOR A CONCRETE POPULATION MODEL

In this section, we provide simple sufficient conditions under which the richness condition in Assumption 1 holds, showing that it is satisfied by a broad class of task distributions, including our experimental setting. We then quantify its dependence on the temperature parameter for the experimental distribution.

Let a task consist of an initial distribution P_0, transition kernel P, feature table $\Phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, reward table $R(a, s')$, and an input weight $\mathbf{w}$. Given a task, a trajectory is generated by a fixed behavior policy π_b, and $\phi_i = \Phi(s_i, a_i)$, $\psi_{i,a} = \Phi(s_{i+1}, a)$, and $r_{i+1} = R(a_i, s_{i+1})$.

E.1 QUALITATIVE VERIFICATION OF OBSERVABLE RICHNESS

Theorem E.1 (Continuous-population verification)**.** *Under Assumption 1(i), assume that:*

(i) the finite-dimensional random vector formed by $\mathbf{w}$*, the entries of the feature* Φ*, and the entries of the reward* R *has a nonzero absolutely continuous component with respect to Lebesgue measure;*

(ii) there exist distinct states s_0, s_1 *and, when* $n > 1$*, distinct actions* a_x, a_z *such that, almost surely over tasks,* $P_0(s_0) > 0$*,* $P(s_1 \mid s_0, a_x) > 0$*,* $P(s_1 \mid s_1, a_z) > 0$*, and* $\pi_b(a_x \mid s_0) > 0$*,* $\pi_b(a_z \mid s_1) > 0$*. For* $n = 1$*, only* $P_0(s_0) > 0$*,* $P(s_1 \mid s_0, a_x) > 0$*, and* $\pi_b(a_x \mid s_0) > 0$ *are required.*

Then, for every fixed $n \geqslant 1$ *and* $\tau > 0$*, Assumption 1 (ii) holds.*

Proof. Fix $n \geqslant 1$ and $\tau > 0$. Recall that $\mathsf{H}_{\tau,n} = \mathbb{E}_{\mathbf{z}}[\mathcal{B}_{\tau,n}(\mathbf{z})^* \mathcal{B}_{\tau,n}(\mathbf{z})]$. Hence, for every coefficient vector ζ, $\langle \zeta, \mathsf{H}_{\tau,n}\zeta \rangle = \mathbb{E}[\|\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta]\|^2]$. Therefore, it is enough to prove that $\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta] = 0$ almost surely implies $\zeta = 0$. Let T_n denote the random discrete trajectory and $\mathcal{E}_n = \{T_n = w_n\}$ denote the calibration family. Since $\mathbb{E}|\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta]|^2 \geqslant \mathbb{E}\left[|\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta]|^2 \mathbf{1}_{\mathcal{E}_n}\right]$, it suffices to show that every $\zeta \neq 0$ is detected on $\mathcal{E}_n$, i.e., $\mathbb{E}\left[|\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta]|^2 \mathbf{1}_{\mathcal{E}_n}\right] > 0$.

Consider the deterministic calibration path ω_n consisting of $s_0 \xrightarrow{a_x} s_1$, followed, when $n > 1$, by $n-1$ repetitions of $s_1 \xrightarrow{a_z} s_1$. For $n = 1$, only the first transition is retained. Let X_n collect the continuous coordinates $W = \mathbf{w}$, $Y = \Phi(s_0, a_x)$, $\Psi_a = \Phi(s_1, a)$, $a \in [m]$, together with $R_0 = R(a_x, s_1)$ and, when $n > 1$, $R_1 = R(a_z, s_1)$. Define the subprobability measure $\nu_n(E) := \mathbb{P}\left(X_n \in E,\, T_n = \omega_n\right)$.

The calibration path ω_n is not a null event. Let μ denote the task distribution, and define $p_{\omega_n}(\Xi) := \Pr(T_n = \omega_n \mid \Xi)$, where Ξ denotes the sampled task variables. By assumption (ii), every factor in $p_{\omega_n}(\Xi)$ is strictly positive almost surely. Hence $p_{\omega_n}(\Xi) > 0$ for μ-almost every task, and $\Pr(T_n = \omega_n) = \mathbb{E}_\mu[p_{\omega_n}(\Xi)] > 0$. Let μ_X denote the marginal distribution of X_n, and define $\ell_n(x) := \mathbb{E}_\mu\big[p_{\omega_n}(\Xi) \mid X_n = x\big]$. Since $p_{\omega_n}(\Xi) > 0$ almost surely, we have $\ell_n(x) > 0$ for μ_X-almost every x. By the tower property, for every Borel set E, $\nu_n(E) = \Pr(X_n \in E,\, T_n = \omega_n) = \int_E \ell_n(x)\, \mu_X(dx)$. By assumption (i), the joint distribution of (w, Φ, R) has a nonzero absolutely continuous component. Since X_n is a coordinate subvector of (w, Φ, R), its marginal distribution μ_X also has a nonzero absolutely continuous component, denoted by μ_X^{ac}. The measure $\widetilde{\nu}_n(dx) := \ell_n(x)\, \mu_X^{\mathrm{ac}}(dx)$ is absolutely continuous with respect to Lebesgue measure. Moreover, it is nonzero because $\mu_X^{\mathrm{ac}} \neq 0$ and $\ell_n > 0$ μ_X^{ac}-almost everywhere. Since $\mu_X^{\mathrm{ac}} \leqslant \mu_X$, we also have $\widetilde{\nu}_n \leqslant \nu_n$. Therefore, ν_n has a nonzero absolutely continuous component.

Suppose that $\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta] = 0$ almost surely under the joint task-trajectory distribution. Restricting to the event $T_n = \omega_n$ gives $\mathcal{B}_{\tau,n}\big(\mathbf{z}_{\omega_n}(x)\big)[\zeta] = 0$ for ν_n-almost every x. Each component of the left-hand side is real analytic in the continuous coordinates x, since it is formed from finite sums, products, exponentials, and divisions by strictly positive softmax denominators. Because ν_n has a nonzero absolutely continuous component, each component vanishes on a set of positive Lebesgue measure. A nonzero real-analytic scalar function on $\mathbb{R}^q$ has a Lebesgue-null zero set. Hence every component must vanish identically. Therefore, $\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta] = 0$ implies

$$\mathcal{B}_{\tau,n}(\mathbf{z}_{\omega_n}(x))[\zeta] = 0, \text{ for all } x.$$

We now use this identity to identify the coefficient blocks in ζ.

Let $a_x = 1$, set $Y = y$, $\Psi_1 = x$, and $\Psi_a = 0$ for every $a \neq 1$. Since $a_z \neq a_x$ when $n > 1$, all later current-state features on the calibration path vanish. Writing $r = R_0$, $n\mathcal{B}_{\tau,n}[\zeta] = 0$ reduces to

$$0 = E_C y\big(r - \mathbf{w}^\top y + \gamma \mathbf{w}^\top \bar{\psi}\big) + y\Big[y^\top(e_- + E_-\mathbf{w}) + \bar{\psi}^\top(e_+ + E_+\mathbf{w}) + r\, e_q^\top \mathbf{w} + \mathbf{w}^\top \Sigma E_A \mathbf{w}\Big],$$

for every $x, y, \mathbf{w} \in \mathbb{R}^d$ and $r \in \mathbb{R}$.

The coefficient of r is $[E_C + (e_q^\top \mathbf{w})I_d]y$. Setting $\mathbf{w} = 0$ and varying y gives $E_C = 0$. The remaining identity $(e_q^\top \mathbf{w})y = 0$ for every $\mathbf{w}, y$ then gives $e_q = 0$. After substituting $E_C = e_q = 0$, the scalar multiplying y must vanish for every $y \neq 0$. Its dependence on y is affine, so its linear coefficient satisfies $e_- + E_-\mathbf{w} = 0$ for every $\mathbf{w}$. Setting $\mathbf{w} = 0$ gives $e_- = 0$, and varying $\mathbf{w}$ then yields $E_- = 0$. The remaining identity is

$$\bar{\psi}^\top(e_+ + E_+\mathbf{w}) + \mathbf{w}^\top \Sigma E_A \mathbf{w} = 0. \tag{E.1}$$

For the candidate feature set $(x, 0, \ldots, 0)$, define $p = \frac{\exp(\mathbf{w}^\top x/\tau)}{\exp(\mathbf{w}^\top x/\tau)+m-1}$. Then $\bar{\psi} = px$ and $\Sigma = p(1-p)xx^\top$. Since $m \geqslant 2$, we have $0 < p < 1$. Dividing (E.1) by p gives

$$x^\top(e_+ + E_+\mathbf{w}) + (1-p)(\mathbf{w}^\top x)(x^\top E_A \mathbf{w}) = 0. \tag{E.2}$$

Set $x = tv$ with $t \neq 0$ and divide (E.2) by t. The second term is $O(t)$ as $t \to 0$, while the first term becomes $v^\top(e_+ + E_+\mathbf{w})$. Letting $t \to 0$ gives $v^\top(e_+ + E_+\mathbf{w}) = 0$ for every $v, \mathbf{w}$. Hence $e_+ + E_+\mathbf{w} = 0$ for every $\mathbf{w}$, and therefore $e_+ = 0$ and $E_+ = 0$.

Equation (E.2) now reduces to

$$(\mathbf{w}^\top x)(x^\top E_A \mathbf{w}) = 0 \qquad \text{for every } x, \mathbf{w}. \tag{E.3}$$

Suppose, for contradiction, that $E_A\mathbf{w} \neq 0$ for some $\mathbf{w}$. Then $\mathbf{w} \neq 0$, and the two proper hyperplanes $\{x : \mathbf{w}^\top x = 0\}$ and $\{x : x^\top E_A \mathbf{w} = 0\}$ cannot cover $\mathbb{R}^d$. Hence there exists x for which both factors in (E.3) are nonzero, a contradiction. Therefore $E_A\mathbf{w} = 0$ for every $\mathbf{w}$, and thus $E_A = 0$. We have shown that all seven coefficient blocks vanish: $E_C = e_q = e_- = E_- = e_+ = E_+ = E_A = 0$. Hence $\zeta = 0$. □

Corollary E.2 (Verification under the experimental generator)**.** *Let $N_s \geqslant 2$ and $m \geqslant 2$. Suppose that P_0 and all transition rows are mutually independent* $\mathrm{Dirichlet}(1, \ldots, 1)$ *random vectors, independent of $(\mathbf{w}, \Phi, R)$, whose scalar entries are mutually independent* $\mathrm{Unif}(-1, 1)$ *random variables. Let the behavior policy be ϵ_b-greedy with uniform exploration for some $0 < \epsilon_b \leqslant 1$. Then the hypotheses of Theorem E.1 are satisfied. Consequently, for every fixed $n \geqslant 1$ and $\tau > 0$, Assumption 1 holds.*

Proof. We verify the two conditions of Theorem E.1. First, collect all scalar coordinates of $\mathbf{w}$, Φ, and R into a finite-dimensional random vector $X = (\mathbf{w}, \mathrm{vec}(\Phi), \mathrm{vec}(R)) \in \mathbb{R}^q$. By assumption, the coordinates of X are mutually independent $\mathrm{Unif}(-1, 1)$ random variables. Hence X has density $f_X(x) = 2^{-q}\mathbf{1}_{(-1,1)^q}(x)$ with respect to q-dimensional Lebesgue measure. In particular, the law of X is absolutely continuous, with a nonzero density on the open set $(-1, 1)^q$. Thus condition (i) of Theorem E.1 holds.

We next verify condition (ii). A $\mathrm{Dirichlet}(1, \ldots, 1)$ random vector is uniformly distributed on the probability simplex and lies in its relative interior almost surely. Therefore every coordinate is strictly positive almost surely. In particular, for any two distinct states s_0, s_1, $P_0(s_0) > 0$ a.s. and, for every action a, $P(s_1 \mid s_0, a) > 0$, $P(s_1 \mid s_1, a) > 0$ a.s.

It remains to check the behavior-policy probabilities. Because the behavior policy is ϵ_b-greedy with uniform exploration, every action receives at least its exploration probability: $\pi_b(a \mid s) \geqslant \frac{\epsilon_b}{m} > 0$ for every (s, a), independently of which action is greedy and of how ties are broken.

Hence, when $n > 1$, choose any distinct states s_0, s_1 and any distinct actions a_x, a_z. Then, almost surely, $P_0(s_0) > 0$, $P(s_1 \mid s_0, a_x) > 0$, $P(s_1 \mid s_1, a_z) > 0$, and $\pi_b(a_x \mid s_0) > 0$, $\pi_b(a_z \mid s_1) > 0$. Thus condition (ii) holds. For $n = 1$, only the first transition is needed, and the same argument gives $P_0(s_0) > 0$, $P(s_1 \mid s_0, a_x) > 0$, $\pi_b(a_x \mid s_0) > 0$ a.s. Therefore all hypotheses of Theorem E.1 are satisfied. Applying that theorem yields Assumption 1 for every fixed $n \geqslant 1$ and $\tau > 0$. □

E.2 Effect of Temperature on Convergence Rate

In this section, we establish the lower bound on $\kappa(\tau)$ and then construct a matching normal direction for the upper bound. By the normal-curvature comparison established in Lemma D.2, it is sufficient for the lower bound to control the smallest eigenvalue $\chi_{\tau,n} := \lambda_{\min}(\mathsf{H}_{\tau,n})$ of the coefficient-space Gram matrix. Equivalently, we need to prove a uniform lower bound of $\mathbb{E}_{\mathcal{D}_n} |\mathcal{B}_{\tau,n}(z)[\zeta]|_2^2$ for every coefficient direction ζ.

Lemma E.3 further reduces this task to the reference law U_n on the calibration path. Hence it suffices to establish the corresponding coercivity bound for $\mathbb{E}_{U_n}|\mathcal{B}_{\tau,n}(z_{\omega_n})[\zeta]|_2^2$. The only coefficient block whose response degenerates is the attention-score block in the first layer of Transformer. We therefore first isolate the successor-value and attention contribution $K_\tau = \bar{\Psi}_\tau^\top(e_+ + E_+W) + W^\top \Sigma_\tau E_A W$ and derive a quantitative lower bound for $\mathbb{E}_{U_n} K_\tau^2$, which shows in Lemma E.4. We then combine this estimate with orthogonal projections controlling the remaining coefficient blocks (E_C, e_q, e_-, E_-), which yields the lower bound of $\kappa(\tau)$.

Lemma E.3 (Calibration-path reduction)**.** *Under the experimental generator, let ω_n be the calibration path used in Theorem E.1, and let U_n denote the product-uniform law of the continuous coordinates appearing along this path. Then, for every coefficient direction ζ,*

$$\mathbb{E}_{\mathcal{D}_n} \|\mathcal{B}_{\tau,n}(z)[\zeta]\|_2^2 \geqslant p_n \, \mathbb{E}_{U_n} \|\mathcal{B}_{\tau,n}(z_{\omega_n})[\zeta]\|_2^2 ,$$

where

$$p_n = \frac{p_{\min}^n}{N_s^2} \frac{\Gamma(N_s)\Gamma(n)}{\Gamma(N_s+n-1)} > 0, \qquad p_{\min} := \frac{\epsilon_b}{m}.$$

Proof. We use the same calibration path as in Theorem E.1, $s_0 \xrightarrow{a_x} s_1$, $s_1 \xrightarrow{a_z} s_1$ for the remaining $n-1$ transitions, with the second part omitted when $n = 1$. Let U_n denote the product-uniform law of the continuous coordinates appearing along this path, and let $\mathbf{z}_{\omega_n}$ be the corresponding context.

Because the behavior policy selects every action with probability at least $p_{\min}$, the conditional probability of following the prescribed action sequence is at least $p_{\min}^n$. The remaining state-transition contribution is

$$Q_n = P_0(s_0)P(s_1 \mid s_0, a_x)P(s_1 \mid s_1, a_z)^{n-1},$$

where the final factor is omitted for $n = 1$. Hence the probability of the calibration path, conditional on the task, is bounded below by $p_{\min}^n Q_n$. Since the initial-state probability $P_0(s_0)$ and the transition probability $P(s_1 \mid s_0, a_x)$ each have mean $1/N_s$. The repeated self-transition probability $P(s_1 \mid s_1, a_z)$ has the $\mathrm{Beta}(1, N_s - 1)$ distribution, so

$$\mathbb{E}Q_n = \mathbb{E}P_0(s_0)\mathbb{E}P(s_1 \mid s_0, a_x)\mathbb{E}P(s_1 \mid s_1, a_z)^{n-1} = \frac{1}{N_s^2}\frac{\Gamma(N_s)\Gamma(n)}{\Gamma(N_s+n-1)}.$$

Under the experimental generator, the environment variables in Q_n are independent of the selected continuous coordinates $(\mathbf{w}, \Phi, R)$. Therefore, for every nonnegative measurable function f,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} f(\mathbf{Z}) &\geqslant \mathbb{E}_{\mathcal{D}_n}\left[f(\mathbf{Z})\mathbf{1}_{\{T_n=\omega_n\}}\right] = \mathbb{E}\left[f(\mathbf{z}_{\omega_n}(X_n))\mathbb{P}(T_n = \omega_n \mid \Xi)\right] \\ &\geqslant p_{\min}^n \mathbb{E}\left[f(\mathbf{z}_{\omega_n}(X_n))Q_n\right] = p_{\min}^n \mathbb{E}\left[f(\mathbf{z}_{\omega_n}(X_n))\right]\mathbb{E}[Q_n] \geqslant p_n \mathbb{E}_{U_n} f(\mathbf{z}_{\omega_n}).\end{aligned}$$

where $p_n = \frac{p_{\min}^n}{N_s^2}\frac{\Gamma(N_s)\Gamma(n)}{\Gamma(N_s+n-1)} > 0$. Let $f(z) = \|\mathcal{B}_{\tau,n}(z)[\zeta]\|^2$, we have the desired result. □

Let $\zeta = (E_C, e_q, e-, E_-, e_+, E_+, E_A)$ be the observable coefficient direction induced by $\mathbf{U}$. We decompose the attention coefficient E_A as $D := E_A = D_0 + aI_d$, $a := \frac{\mathrm{tr}(D)}{d}$, $\mathrm{tr}(D_0) = 0$. We first isolate the part of the response associated with the successor-feature and attention coefficients. Under the reference law U_n, let W denote the weight vector and $\Psi_1, \ldots, \Psi_m$ the successor features. Their contribution is summarized by $K_\tau := \bar{\Psi}\tau^\top(e^+ + E_+W) + W^\top\Sigma_\tau DW$.

Lemma E.4 shows that the temperature dependence of K_τ determines the weakest curvature contributed by the attention block and, in particular, reveals the τ^5 scaling of the scalar attention direction.

Lemma E.4. *Assume $d \geqslant 2$. Under the reference law U_n, there exist constants $b_{\mathrm{att}} > 0$ and $\tau_* > 0$, independent of τ and of the coefficient direction, such that for every $0 < \tau \leqslant \tau_*$,*

$$\mathbb{E}_{U_n} K_\tau^2 \geqslant b_{\mathrm{att}} \left(\|e_+\|^2 + \|E_+\|_F^2 + \tau^3\|D_0\|_F^2 + \tau^5 d a^2\right). \tag{E.4}$$

Proof. (i) We first lower-bound $\mathbb{E}[(W^\top \Sigma_\tau DW)^2]$. The difficulty is that D_0 may point in an arbitrary trace-free matrix direction. We first construct a finite collection of measurements $u^\top D_0 v$ that controls the entire matrix D_0. Define

$$\mathcal{F} = \{(e_i, e_j) : i \neq j\} \cup \left\{ \left(\frac{e_i - e_j}{\sqrt{2}}, \frac{e_i + e_j}{\sqrt{2}} \right), \left(\frac{e_i + e_j}{\sqrt{2}}, \frac{e_i - e_j}{\sqrt{2}} \right) : i < j \right\}.$$

The pairs (e_i, e_j) directly measure the off-diagonal entries of D_0, while the rotated pairs also measure differences between its diagonal entries. Since $\mathrm{tr}(D_0) = 0$, we have

$$\begin{aligned} \sum_{(u,v)\in\mathcal{F}} \big(u^\top D_0 v\big)^2 &= \sum_{i\neq j} (D_0)_{ij}^2 + \sum_{i<j} \frac{1}{2}\big((D_0)_{ii} - (D_0)_{jj}\big)^2 + \frac{1}{2}\big((D_0)_{ij} - (D_0)_{ji}\big)^2 \\ &= \sum_{i\neq j} (D_0)_{ij}^2 + \frac{d}{2}\sum_i (D_0)_{ii}^2 + \frac{1}{2}\sum_{i<j}\big((D_0)_{ij} - (D_0)_{ji}\big)^2 \\ &\geqslant \sum_{i\neq j} (D_0)_{ij}^2 + \sum_i (D_0)_{ii}^2 = \|D_0\|_F^2. \end{aligned} \tag{E.5}$$

Hence the finite family $\mathcal{F}$ detects every nonzero trace-free direction.

For each $(u, v) \in \mathcal{F}$, we construct a local configuration in which the attention response contains the measurement $u^\top D_0 v$. Choose fixed sufficiently small constants $s, t, b > 0$ and set $W = sv$, $V := \Psi_1 - \Psi_2 = tu$, and $M := (\Psi_1 + \Psi_2)/2 = bv$, with $\Psi_a = -bv$ for $a \geqslant 3$. Equivalently, $\Psi_1 = bv + (t/2)u$ and $\Psi_2 = bv - (t/2)u$.

Since $u^\top v = 0$, actions 1 and 2 have the same score, $W^\top \Psi_1 = W^\top \Psi_2 = sb$, whereas every action $a \geqslant 3$ has score $-sb$. Thus actions 1 and 2 are tied, while all remaining actions are separated from them by a fixed positive margin. Let $g := W^\top V = W^\top(\Psi_1 - \Psi_2)$ denote the score gap between actions 1 and 2. At the center, $g = 0$. In a sufficiently small neighborhood of each center, we use g, together with the remaining local coordinates υ, as coordinates. The corresponding reference density can be written as $j_\ell(\upsilon, g)\, d\upsilon\, dg$, with $0 < j_- \leqslant j_\ell(\upsilon, g) \leqslant j_+ < \infty$. Since only finitely many charts are used, these constants can be chosen uniformly.

We restrict each chart to the near-tie region $\tau \leqslant |g| \leqslant 2\tau$. In this region, actions 1 and 2 both receive nonnegligible softmax mass, whereas all other actions remain separated by a fixed score gap and contribute only exponentially small terms. Using the pairwise representation of Σ_τ, we obtain

$$W^\top \Sigma_\tau DW = \pi_1^\tau \pi_2^\tau\, g\, V^\top DW + R_\tau, |R_\tau| \leqslant Ce^{-g_0/\tau}\|D\|_F. \tag{E.6}$$

Since $D = D_0 + aI_d$ and $V^\top W = g$,

$$gV^\top DW = gV^\top D_0 W + ag^2. \tag{E.7}$$

The trace-free component therefore contains one factor of the small gap g, whereas the scalar component contains two. To make the dependence on τ explicit, set $g = \tau x$ and $a_e := \tau a$. Then

$$gV^\top D_0 W + ag^2 = \tau\left(xV^\top D_0 W + a_e x^2\right). \tag{E.8}$$

Squaring contributes a factor τ^2, while $dg = \tau\, dx$ contributes one additional factor τ. Hence the near-tie contribution has an overall factor τ^3. After removing this common factor, the corresponding quadratic forms converge to

$$G_0(D_0, a_e) = \sum_\ell \int_{Z_\ell} j_\ell(\upsilon, 0) \int_{1\leqslant|x|\leqslant 2} k(x)\left(x f_\ell(\upsilon; D_0) + a_e x^2\right)^2 dx\, d\upsilon, \tag{E.9}$$

where $k(x) := e^{2x}/(1+e^x)^4 > 0$ and $f_\ell(\upsilon; D_0) := V_\ell(\upsilon, 0)^\top D_0 W_\ell(\upsilon, 0)$. The role of G_0 is to rule out cancellation between the trace-free part D_0 and the scalar part aI_d. Suppose $G_0(D_0, a_e) = 0$. Since the integrand is nonnegative and the weights are strictly positive, we must have $xf_\ell(\upsilon; D_0) + a_e x^2 = 0$ throughout each chart. Evaluating this identity at x and $-x$, with $1 < x < 2$, gives $xf_\ell + a_e x^2 = 0, -xf_\ell + a_e x^2 = 0$. Hence $a_e = 0$ and $f_\ell = 0$. At the center of the ℓ-th chart, $V_\ell = tu_\ell$ and $W_\ell = sv_\ell$, and therefore $f_\ell(0; D_0) = st\, u_\ell^\top D_0 v_\ell = 0$. Since this holds for every $(u_\ell, v_\ell) \in \mathcal{F}$, (E.5) implies $D_0 = 0$. Thus G_0 is positive definite. Because the coefficient space is finite dimensional, compactness of its unit sphere gives a constant $\lambda_0 > 0$ such that

$$G_0(D_0, a_e) \geqslant \lambda_0\left(\|D_0\|_F^2 + da_e^2\right). \tag{E.10}$$

Moreover, after the rescaling $g = \tau x$ and $a = a_e/\tau$, the contribution of the remaining actions to the rescaled response is uniformly $O(\tau^{-2}e^{-g_0/\tau}) = o(1)$ on the coefficient unit sphere $\mathrm{tr}(D_0) = 0$, $\|D_0\|_F^2 + da_e^2 = 1$. The coordinate maps and densities also converge uniformly on the fixed chart domains, chosen with compact closures. Thus G_τ converges uniformly to G_0 on this unit sphere, equivalently in operator norm for the associated Gram operators. Hence, for all sufficiently small $\tau > 0$,

$$G_\tau(D_0, a_e) := \tau^{-3} \sum_\ell \int_{C_{\ell,\tau}} \left[W^\top \Sigma^\tau \left(D_0 + \frac{a_e}{\tau} I_d \right) W \right]^2 dP \geqslant \frac{\lambda_0}{2} \left(\|D_0\|_F^2 + da_e^2 \right). \quad \text{(E.11)}$$

Let $\mathcal{C}_{\ell,\tau}$ denote the near-tie region in the ℓ-th chart and define $\mathcal{E}_{\mathrm{tie}}(\tau) := \bigcup_\ell \mathcal{C}_{\ell,\tau}$. Since there are only $L := |\mathcal{F}|$ charts, their overlap multiplicity is at most L. Therefore, for every nonnegative function h,

$$\sum_\ell \int_{\mathcal{C}_{\ell,\tau}} h \, dP \leqslant L \, \mathbb{E}\left[h \, \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)} \right]. \quad \text{(E.12)}$$

Combining this finite-overlap bound with (E.11), and recalling that $a_e = \tau a$, gives

$$\mathbb{E}\left[\left(W^\top \Sigma_\tau D W \right)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)} \right] \geqslant b_3 \tau^3 \|D_0\|_F^2 + b_5 \tau^5 d a^2 - C e^{-g_0/\tau} \|D\|_F^2. \quad \text{(E.13)}$$

Thus the near-tie regions provide order-τ^3 control of the trace-free component D_0 and the weaker order-τ^5 control of the scalar component aI_d.

(ii) We next control cancellation with E_+. All expectations below are under the reference law U_n. Write $X_\tau := \bar{\Psi}_\tau^\mathsf{T} E_+ W$, $Y_\tau := W^\mathsf{T} \Sigma_\tau D W$, $Z_\tau := \bar{\Psi}_\tau^\mathsf{T} e_+$, where $D = D_0 + aI_d$ and $\mathrm{tr}(D_0) = 0$. In particular, $K_\tau = Z_\tau + X_\tau + Y_\tau$, $\|D\|_{\mathrm{F}}^2 = \|D_0\|_{\mathrm{F}}^2 + da^2$. We choose all constants independently of τ and the coefficient direction. The parameters d and m are fixed.

Choose fixed, sufficiently small positive numbers r_*, a_*, b_*, and consider the following d^2 configurations in the interior of the product-uniform support:

$$W = r_* e_j, \qquad \Psi_1 = b_* e_j, \qquad \Psi_a = -b_* e_j \quad (a \geqslant 2), \qquad j = 1, \ldots, d;$$
$$W = r_* e_j, \qquad \Psi_1 = a_* e_i + b_* e_j, \qquad \Psi_a = -b_* e_j \quad (a \geqslant 2), \qquad i \neq j.$$

At each configuration, action 1 has a fixed positive winning gap. Take sufficiently small, fixed open neighborhoods of these configurations whose closures lie inside the support, and denote their union by $\mathcal{E}_{\mathrm{sep}}$. On this union, every gap $W^\mathsf{T}(\Psi_1 - \Psi_a)$, $a \geqslant 2$, is bounded below by a common constant $g_{\mathrm{sep}} > 0$.

For $X_0(E) := \Psi_1^\mathsf{T} E W$, the center measurements are $b_* r_* E_{jj}$ and $a_* r_* E_{ij} + b_* r_* E_{jj}$. They jointly determine every entry of E. Hence the quadratic form $\mathbb{E}[X_0(E)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}]$ is positive definite: if it vanished, positivity of the reference density and continuity would force $X_0(E)$ to vanish on each neighborhood, including at every center, and therefore $E = 0$. The diagonal centers similarly show that $\mathbb{E}[(\Psi_1^\mathsf{T} e)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}]$ is positive definite in e. Compactness of the respective unit spheres therefore gives constants $\lambda_E, \lambda_e > 0$ such that

$$\mathbb{E}[X_0(E)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \geqslant \lambda_E \|E\|_{\mathrm{F}}^2, \quad \mathbb{E}[(\Psi_1^\mathsf{T} e)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \geqslant \lambda_e \|e\|^2. \quad \text{(E.14)}$$

All features and W are bounded. The uniform winning gap gives

$$\left\| \bar{\Psi}_\tau - \Psi_1 \right\| + \|\Sigma_\tau\|_{\mathrm{op}} \leqslant C e^{-g_{\mathrm{sep}}/\tau} \quad \text{on } \mathcal{E}_{\mathrm{sep}}. \quad \text{(E.15)}$$

For example, the total softmax mass outside action 1 is at most $(m-1)e^{-g_{\mathrm{sep}}/\tau}$, which bounds both quantities in (E.15) up to fixed constants. The reverse triangle inequality in the restricted L^2 norm, together with (E.14), implies that, for all sufficiently small τ,

$$\mathbb{E}[X_\tau^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \geqslant \frac{\lambda_E}{4} \|E_+\|_{\mathrm{F}}^2, \quad \mathbb{E}[Z_\tau^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \geqslant \frac{\lambda_e}{4} \|e_+\|^2. \quad \text{(E.16)}$$

These estimates hold simultaneously for all coefficient directions, because the L^2 approximation errors are bounded by $Ce^{-g_{\mathrm{sep}}/\tau}\|E_+\|_{\mathrm{F}}$ and $Ce^{-g_{\mathrm{sep}}/\tau}\|e_+\|$, respectively. Also, (E.15) yields $\mathbb{E}[Y_\tau^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \leqslant C_{\mathrm{sep}} e^{-2g_{\mathrm{sep}}/\tau} \|D\|_{\mathrm{F}}^2$. The elementary inequality $(u+v)^2 \geqslant \frac{1}{2}u^2 - v^2$ for $u, v \in \mathbb{R}$

follows by completing the square. Applying it with $u = X_\tau$ and $v = Y_\tau$, and weakening the exponential remainder if necessary, gives

$$\mathbb{E}[(X_\tau + Y_\tau)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] \geqslant b_E \|E_+\|_{\mathrm{F}}^2 - C_{\mathrm{sep}} e^{-g_{\mathrm{sep}}/\tau} \|D\|_{\mathrm{F}}^2, \qquad b_E := \frac{\lambda_E}{8} > 0. \tag{E.17}$$

Let $\mathcal{E}_{\mathrm{tie}}(\tau)$ be the union of the near-tie charts from part (i). Their bounded densities, fixed bounded coordinate domains, and gap restriction $\tau \leqslant |g| \leqslant 2\tau$ imply $\mathbb{P}(\mathcal{E}_{\mathrm{tie}}(\tau)) \leqslant C_t \tau$. Indeed, each chart contributes at most its upper density bound times its coordinate-domain volume times the gap-interval length 2τ; summing over the finitely many charts gives this bound regardless of overlap. Since $\left\|\bar{\Psi}_\tau\right\|, \|W\| \leqslant \sqrt{d}$,

$$\mathbb{E}[X_\tau^2 \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)}] \leqslant C_X \tau \|E_+\|_{\mathrm{F}}^2, \qquad C_X := d^2 C_t. \tag{E.18}$$

Rewrite (E.13), retaining its coefficient-independent constants, as

$$\mathbb{E}[Y_\tau^2 \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)}] \geqslant b_3 \tau^3 \|D_0\|_{\mathrm{F}}^2 + b_5 \tau^5 d a^2 - C_{\mathrm{tie}} e^{-g_0/\tau} \|D\|_{\mathrm{F}}^2. \tag{E.19}$$

Apply $(u+v)^2 \geqslant \frac{1}{2}u^2 - v^2$ with $u = Y_\tau$ and $v = X_\tau$. Equations (E.18) and (E.19) give

$$\mathbb{E}[(X_\tau + Y_\tau)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)}] \geqslant \frac{b_3}{2} \tau^3 \|D_0\|_{\mathrm{F}}^2 + \frac{b_5}{2} \tau^5 d a^2 - C_X \tau \|E_+\|_{\mathrm{F}}^2 - \frac{C_{\mathrm{tie}}}{2} e^{-g_0/\tau} \|D\|_{\mathrm{F}}^2. \tag{E.20}$$

No disjointness of the two regions is needed: their indicators sum to at most two. Consequently,

$$\mathbb{E}[(X_\tau + Y_\tau)^2] \geqslant \frac{1}{2}\big(\mathbb{E}[(X_\tau + Y_\tau)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{sep}}}] + \mathbb{E}[(X_\tau + Y_\tau)^2 \mathbf{1}_{\mathcal{E}_{\mathrm{tie}}(\tau)}]\big).$$

Set $c = \min\{g_0, g_{\mathrm{sep}}\} > 0$ and $C_0 = C_{\mathrm{sep}}/2 + C_{\mathrm{tie}}/4$. Combining the two regional lower bounds yields

$$\mathbb{E}[(X_\tau + Y_\tau)^2] \geqslant \left(\frac{b_E}{2} - \frac{C_X \tau}{2}\right) \|E_+\|_{\mathrm{F}}^2 + \frac{b_3}{4} \tau^3 \|D_0\|_{\mathrm{F}}^2 + \frac{b_5}{4} \tau^5 d a^2 - C_0 e^{-c/\tau} \big(\|D_0\|_{\mathrm{F}}^2 + d a^2\big). \tag{E.21}$$

Choose $\tau_* > 0$ smaller than the thresholds used above and in part (i), so that for every $0 < \tau \leqslant \tau_*$,

$$\tau \leqslant 1, \qquad C_X \tau \leqslant \frac{b_E}{2}, \qquad C_0 e^{-c/\tau} \leqslant \frac{\min\{b_3, b_5\}}{8} \tau^5. \tag{E.22}$$

Such a choice exists because $e^{-c/\tau}/\tau^5 \to 0$ as $\tau \downarrow 0$. Every quantity used to choose this threshold is independent of e_+, E_+, and D. Using $\tau^5 \leqslant \tau^3$ and (E.22) in (E.21) gives

$$\mathbb{E}[(X_\tau + Y_\tau)^2] \geqslant \frac{b_E}{4} \|E_+\|_{\mathrm{F}}^2 + \frac{b_3}{8} \tau^3 \|D_0\|_{\mathrm{F}}^2 + \frac{b_5}{8} \tau^5 d a^2. \tag{E.23}$$

The product-uniform law of $(W, \Psi_1, \ldots, \Psi_m)$ is invariant under simultaneous sign reversal. This transformation preserves all softmax logits and Σ_τ, reverses $\bar{\Psi}_\tau$, and makes Z_τ odd while $X_\tau + Y_\tau$ is even. Therefore $\mathbb{E}[Z_\tau(X_\tau + Y_\tau)] = 0$, $\mathbb{E}[K_\tau^2] = \mathbb{E}[Z_\tau^2] + \mathbb{E}[(X_\tau + Y_\tau)^2]$. The second estimate in (E.16) implies $\mathbb{E}[Z_\tau^2] \geqslant (\lambda_e/4) \|e_+\|^2$. Combining this with (E.23) proves

$$\mathbb{E}[K_\tau^2] \geqslant b_{\mathrm{att}} \left(\|e_+\|^2 + \|E_+\|_{\mathrm{F}}^2 + \tau^3 \|D_0\|_{\mathrm{F}}^2 + \tau^5 d a^2\right), \tag{E.24}$$

where one may take $b_{\mathrm{att}} = \min\left\{\frac{\lambda_e}{4}, \frac{b_E}{4}, \frac{b_3}{8}, \frac{b_5}{8}\right\} > 0$. Both b_{att} and τ_* are independent of the coefficient direction. They can also be chosen independently of n, since all quantities in this argument, including the near-tie construction in part (i), depend only on the same product-uniform marginal of $(W, \Psi_1, \ldots, \Psi_m)$ under every U_n. □

The proof of Corollary 3. (i) Since $\mathcal{B}_{\tau,n}(z)[\zeta] = 0$, $\mathcal{D}_n$-a.s. $\Longrightarrow \zeta = 0$ has already been established in Corollary E.2; This step is quantitative only. Lemma E.4 analyzed e_+, E_+, E_A, we now incorporate E_C, e_q, e_-, E_-.

On the calibration path, let $Y := \Phi(s_0, a_x)$, $R_0 := R(a_x, s_1)$, $Z_Y := YY^\top - \frac{1}{3} I_d$, and $v_\tau := W^\top \bar{\Psi}_\tau$. Conditional on all variables other than Y and R_0, the observable response can be written as

$$n\mathcal{B}_{\tau,n}(z_{\omega_n})[\zeta] = R_0 A_r(Y) + A_o(Y) + A_2(Y) + A_{\mathrm{const}}, \tag{E.25}$$

where $A_r(Y) = [E_C + (e_q^\top W) I_d]Y$, $A_o(Y) = \gamma v_\tau E_C Y + K_\tau Y$, $A_2(Y) = -E_C Z_Y W + Z_Y(e_- + E_- W)$, and $A_{\text{const}} = \frac{1}{3}(-E_C W + e_- + E_- W) + A_{\text{later}}$. Here A_{later} is the sum of the contributions from transitions $1, \dots, n-1$, and is zero when $n = 1$.

The calibration path has $s_0 \neq s_1$, and all later transitions use features associated with s_1. Hence, under the product reference law, A_{later} is independent of $Y = \Phi(s_0, a_x)$ and of R_0. Thus A_{const} is constant with respect to the conditional variables (Y, R_0).

The four terms in (E.25) are mutually orthogonal in $L^2(Y, R_0)$. Indeed, $\mathbb{E}R_0 = 0$ and R_0 independent on Y; A_o is odd in Y, i.e., $A_o(-Y) = -A_o(Y)$; Since $Z_{-Y} = Z_Y$, A_2 is even in Y and centered because $\mathbb{E}Z_Y = 0$; and A_{const} is constant in (Y, R_0). Odd and even functions are orthogonal under symmetric distributions. In particular, later transitions cannot cancel any of the first three components.

The uniform moments satisfy $\mathbb{E}[YY^\top] = \mathbb{E}[WW^\top] = \frac{1}{3} I_d$, $\mathbb{E}R_0^2 = \frac{1}{3}$, and $\mathbb{E}Z_Y^2 = \nu I_d$, $\nu := \frac{5d-1}{45} > 0$. For the reward projection,

$$\mathbb{E}\|R_0 A_r(Y)\|^2 = \frac{1}{9}\mathbb{E}\left\|E_C + (e_q^\top W) I_d\right\|_F^2 = \frac{1}{9}\|E_C\|_F^2 + \frac{d}{27}\|e_q\|^2, \tag{E.26}$$

where the cross term vanishes because $\mathbb{E}W = 0$.

For $0 < \delta < 1$, we use $\|x + y\|^2 \geqslant \delta\|y\|^2 - \frac{\delta}{1-\delta}\|x\|^2$. Since $|v_\tau| \leqslant d$,

$$\mathbb{E}\|A_o(Y)\|^2 \geqslant \frac{\delta_o d}{3}\,\mathbb{E}K_\tau^2 - \frac{\delta_o}{1-\delta_o}\frac{\gamma^2 d^2}{3}\|E_C\|_F^2. \tag{E.27}$$

Moreover, $Z_Y e_-$ is orthogonal to the terms linear in the centered variable W. Using $\mathbb{E}\|Z_Y E_- W\|^2 = \frac{\nu}{3}\|E_-\|_F^2$, $\mathbb{E}\|E_C Z_Y W\|^2 = \frac{\nu}{3}\|E_C\|_F^2$, we obtain

$$\mathbb{E}\|A_2(Y)\|^2 \geqslant \nu\|e_-\|^2 + \frac{\delta_2 \nu}{3}\|E_-\|_F^2 - \frac{\delta_2}{1-\delta_2}\frac{\nu}{3}\|E_C\|_F^2. \tag{E.28}$$

Choose $\delta_o = \min\left\{\frac{1}{2}, \frac{1}{48\gamma^2 d^2}\right\}$, $\delta_2 = \min\left\{\frac{1}{2}, \frac{1}{48\nu}\right\}$. Since $\delta/(1-\delta) \leqslant 2\delta$ for $\delta \leqslant 1/2$, each negative coefficient multiplying $\|E_C\|_F^2$ in (E.27) and (E.28) is at most $1/72$. Hence the coefficient $1/9$ in (E.26) retains at least $\frac{1}{12}$. Combining the orthogonal decomposition, (E.4), (E.26), (E.27), and (E.28), there exists $b_{\text{ref}} > 0$, independent of n, τ, and ζ, such that for all sufficiently small $0 < \tau \leqslant 1$,

$$\begin{aligned}\mathbb{E}_{U_n}\|\mathcal{B}_{\tau,n}(\mathbf{z}_{\omega_n})[\zeta]\|_2^2 \geqslant \frac{b_{\text{ref}}}{n^2}\Big(&\|E_C\|_F^2 + \|e_q\|^2 + \|e_-\|^2 + \|E_-\|_F^2 \\ &+ \|e_+\|^2 + \|E_+\|_F^2 + \tau^3\|D_0\|_F^2 + \tau^5 d\, a^2\Big).\end{aligned} \tag{E.29}$$

Since $\|D\|_F^2 = \|D_0\|_F^2 + da^2$ and $0 < \tau \leqslant 1$, the right-hand side is bounded below by $\frac{b_{\text{ref}}}{n^2}\tau^5\|\zeta\|_{\mathcal{V}}^2$. Using Lemma E.3,

$$\mathbb{E}_{\mathcal{D}_n}\|\mathcal{B}_{\tau,n}(\mathbf{z})[\zeta]\|_2^2 \geqslant \frac{p_n b_{\text{ref}}}{n^2}\tau^5\|\zeta\|_{\mathcal{V}}^2. \tag{E.30}$$

Therefore $\chi_{\tau,n} \geqslant \frac{p_n b_{\text{ref}}}{n^2}\tau^5$. By Lemma D.2, $\kappa(\tau, n) \geqslant c_\Omega \frac{p_n b_{\text{ref}}}{n^2}\tau^5$. Since $\mathcal{N}$ is finite, $c_{\text{exp}} := c_\Omega b_{\text{ref}} \min_{n\in\mathcal{N}} \frac{p_n}{n^2} > 0$, and hence $\kappa(\tau) \geqslant c_{\text{exp}}\tau^5$, $n \in \mathcal{N}$.

(ii) For the upper bound, it suffices to exhibit a single unit normal direction whose response is $O(\tau^5)$. Consider the pure scalar attention perturbation $\mathbf{U}_{\text{sc}} = (\frac{I_d}{\sqrt{d}}, 0, \dots, 0)$. Every tangent vector to $\mathcal{M}_\tau$ has zero A-block, so $\mathbf{U}_{\text{sc}} \in N_{\bar\theta}\mathcal{M}_\tau$ at every $\bar\theta \in \mathcal{M}_\tau$, and $\|\mathbf{U}_{\text{sc}}\| = 1$.

For a fixed successor state, define $g_{ab} = W^\top(\Psi_a - \Psi_b)$. This random variable has a uniformly bounded density. Conditional on $W \neq 0$, let $M_W := \max_j |W_j|$ and choose a coordinate attaining this maximum. Since the difference of two independent $\text{Unif}(-1, 1)$ variables has density bounded by $1/2$, convolution with the remaining independent coordinates yields $\|f_{g_{ab}|W}\|_\infty \leqslant \frac{1}{2M_W}$. Furthermore, $\mathbb{P}(M_W \leqslant t) = t^d$, $\mathbb{E}M_W^{-1} = \frac{d}{d-1}$, so, because $d \geqslant 2$, $\|f_{g_{ab}}\|_\infty \leqslant L_d := \frac{d}{2(d-1)} < \infty$.

Set $V_\tau := W^\top \Sigma_\tau W$. The pairwise covariance identity gives $V_\tau = \sum_{a<b} \pi_a^\tau \pi_b^\tau g_{ab}^2$. Moreover, $\pi_a^\tau \pi_b^\tau \leqslant e^{-|g_{ab}|/\tau}$. Indeed, if the larger of the two logits is q_a, the softmax denominator is at least $e^{2q_a/\tau}$ after multiplying the two probabilities.

Let $P_m := m(m-1)/2$. By Cauchy–Schwarz,

$$\mathbb{E}V_\tau^2 \leqslant P_m \sum_{a<b} \mathbb{E}\left[g_{ab}^4 e^{-2|g_{ab}|/\tau}\right] \leqslant 2L_d P_m^2 \int_0^\infty x^4 e^{-2x/\tau}\, dx = C_V \tau^5, \tag{E.31}$$

where $C_V < \infty$ depends only on d and m.

Direct differentiation in the direction $\mathbf{U}_{\mathrm{sc}}$ gives $J_{\bar{\theta}}(z)[\mathbf{U}_{\mathrm{sc}}] = \frac{\alpha\gamma}{n\sqrt{d}} \sum_{i=0}^{n-1} \phi_i V_{\tau,i}$. Using convexity of the squared norm and $\|\phi_i\|^2 \leqslant d$, $\mathbb{E}\|J_{\bar{\theta}}(\mathbf{z})[\mathbf{U}_{\mathrm{sc}}]\|_2^2 \leqslant \frac{\alpha^2\gamma^2}{n} \sum_{i=0}^{n-1} \mathbb{E}V_{\tau,i}^2$.

The visited successor state may depend on the task, so we do not condition on it. If $V_\tau(s)$ denotes the same scalar covariance computed using the feature table at a fixed state s, then

$$\mathbb{E}V_{\tau,i}^2 = \sum_{s\in\mathcal{S}} \mathbb{E}\left[\mathbf{1}_{\{s_{i+1}=s\}} V_\tau(s)^2\right] \leqslant \sum_{s\in\mathcal{S}} \mathbb{E}V_\tau(s)^2 \leqslant N_s C_V \tau^5.$$

Therefore $\mathbb{E}\|J_{\bar{\theta}}(\mathbf{z})[\mathbf{U}_{\mathrm{sc}}]\|_2^2 \leqslant \alpha^2\gamma^2 N_s C_V \tau^5$. Since $\kappa(\tau)$ is the minimum squared response over unit normal directions, $\kappa(\tau, n) \leqslant C_{\exp}\tau^5$, $C_{\exp} := \alpha^2\gamma^2 N_s C_V$, uniformly over all $n \geqslant 1$.

Finally, by the uniform local derivative bounds established in Lemma D.4, the same admissible step size η may be used throughout the small-temperature regime. Applying the parameter-distance complexity result of Theorem 2 and substituting $\kappa(\tau) = \Theta(\tau^5)$ yields $\sup_{\boldsymbol{\theta}^0 \in \mathcal{A}_\tau} K_{\mathrm{GD}}(\varepsilon; \boldsymbol{\theta}^0, \tau, n) = \Theta_{\boldsymbol{\Omega},\mathcal{N}}(\frac{1}{\eta\tau^5} \log \frac{1}{\varepsilon})$. □