# ACTION-CONDITIONED PREDICTIVE LEARNING WITH CAUSAL SELF-ATTENTION: IDENTIFICATION, INITIALIZATION, AND PLANNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study whether causal self-attention trained on action-conditioned outcomes can learn predictive models that support near-optimal decisions without optimal-action supervision. For a one-sparse logistic model on fixed binary history features, we characterize identification by positive feature-disagreement probabilities. Under this condition, population gradient flow and gradient descent from uniform attention converge, with prediction bounds uniform over all feature combinations. A logarithmic-size design attains the identification lower bound, with finite-sample guarantees from independent exploratory episodes. For a complementary multi-head model, we show that uniform-attention initialization yields tangent features of degree at most two and makes population training stationary for pure higher-order parity targets. A target-independent nonzero score initialization yields positive expected kernel eigenvalues, with interaction-order expansions and explicit bounds at every nonzero scale. These target-eigenvalue guarantees extend to finite-width population and empirical training, including bounded perturbations of paired initialization. A separate finite reachable-set analysis treats general full-support outcome distributions under representation and coverage conditions. Exact planning with the Bayes predictive model is Bayes-optimal, and uniform predictive error yields explicit value-suboptimality bounds. Six experiments examine sparse prediction, finite-sample and finite-width effects, higher-order initialization, and planning under the learned predictive model, and corroborate the predicted initialization spectrum.

## 1 INTRODUCTION

Many sequential decision-making systems must predict the possible consequences of different actions from an expanding interaction history before deciding what to do. Model-based approaches typically learn how the environment responds to actions and use the learned model for planning or action selection (Hafner et al., 2019). Causal self-attention provides a mechanism for selectively aggregating information from preceding positions when forming the current prediction (Vaswani et al., 2017). Accurate prediction of action outcomes, however, does not by itself imply that decisions based on the learned model will be near-optimal; the effect of model error on the performance of the resulting policy requires a separate analysis (Ávila Pires & Szepesvári, 2016). We study outcome-only predictive supervision in finite-horizon, history-dependent stochastic processes in which the latent task is drawn from a prior distribution and policies choose actions solely from the observable history.

Existing theory establishes sparse-token selection and recovery of relevant input locations or causal dependencies (Wang et al., 2024b; Nichani et al., 2024; Marion et al., 2025; Wen et al., 2025). Attention optimization has also been studied through neural tangent kernels and nonlinear training dynamics (Hron et al., 2020; Wu et al., 2023; Gao et al., 2024; Song et al., 2024). Transition prediction and language-model-based planning have direct precedents (Ildiz et al., 2024; Wang et al., 2024a; He et al., 2024). In particular, He et al. (2024) analyze predictive world models and planning within a broader pretraining framework. Predictive-state and minimal-information-state theories characterize sufficient internal states from controlled input–output behavior (Littman et al., 2001; Sakcak et al., 2024). Our decision-theoretic reduction uses standard Bayesian conditioning and



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.

simulation arguments; the learning results specify conditions under which outcome-only supervision suffices.

Our main distinctions concern training distributions and initialization. The one-sparse analysis identifies the exact disagreement condition for a common input law and derives optimization and episode-sampling guarantees under that condition. For multiple independently varying tokens, we prove an initialization-specific obstruction and its spectral resolution in the same softmax architecture. General gradient failures and the relation between gradient learning and tangent kernels are established topics (Shalev-Shwartz et al., 2017; Malach et al., 2021; Chizat et al., 2019). Ustaomeroglu & Qu (2025) study interaction order for linear attention and higher-order architectures. Here the restriction concerns the Jacobian at uniform attention. A change of initialization yields an explicit interaction-order spectrum, nonasymptotic eigenvalue bounds, and finite-width population and empirical learning guarantees. The empirical result allows bounded perturbations of paired initialization; it does not require zero initial outputs or coverage of the entire input cube. Further related work is discussed in Appendix P.

We also establish finite reachable-set guarantees for general conditional distributions under representation and coverage conditions, with Bayes value-suboptimality bounds from external dynamic programming. Six experiments examine sparse prediction, finite-width and coverage effects, higher-order initialization, and planning under the learned predictive model (Section 6). *Proofs and additional technical details are provided in the appendix.*

## 2 Controlled Processes, Predictive Outcomes, and Training Data

Let $\mathcal{E}$ be a finite task set with prior $\nu$. The observation and action sets $\mathcal{O}$ and $\mathcal{A}$ are finite, the reward set $\mathcal{R}$ is a finite subset of $[0,1]$, and the interaction horizon is $H$. Each task $e \in \mathcal{E}$ is specified by an initial observation distribution $P_e^0$ on $\mathcal{O}$ and a history-dependent stochastic kernel $P_e(\cdot \mid h, a)$ on $\mathcal{Y} = \mathcal{R} \times \mathcal{O}$.

At stage $k$, the agent observes $o_k$, chooses an action $a_k$, receives a reward $r_k$, and observes $o_{k+1}$. The corresponding decision history is $h_k = (o_1, a_1, r_1, \ldots, o_{k-1}, a_{k-1}, r_{k-1}, o_k)$, and the next outcome is $y_k = (r_k, o_{k+1})$. Thus, $P_e(\cdot \mid h_k, a_k)$ gives the conditional distribution of the next reward and observation under task $e$. We work directly with observable histories and do not assume access to a latent Markov state. Every finite-horizon partially observable Markov decision process (POMDP) induces such a history-dependent process after marginalizing its latent states.

For probability distributions $p$ and $q$ on $\mathcal{Y}$, let $\mathrm{TV}(p,q) = \frac{1}{2}\sum_{y\in\mathcal{Y}} |p(y) - q(y)|$ denote their total variation distance. The prior predictive initial observation distribution is $P_\nu^0(o) = \sum_{e\in\mathcal{E}} \nu(e) P_e^0(o)$. For a history $h_k$, define the task-specific likelihood by $\Lambda_e(h_k) = P_e^0(o_1)\prod_{t=1}^{k-1} P_e(y_t \mid h_t, a_t)$, with the empty product equal to one. Whenever the denominator is positive, the Bayes posterior is $\nu(e \mid h_k) = \frac{\nu(e)\Lambda_e(h_k)}{\sum_{e'\in\mathcal{E}} \nu(e')\Lambda_{e'}(h_k)}$, and the Bayes predictive kernel is $P_\nu(y \mid h, a) = \sum_{e\in\mathcal{E}} \nu(e \mid h) P_e(y \mid h, a)$.

A randomized history policy $\pi$ assigns a distribution $\pi(\cdot \mid h)$ over actions and does not observe the task identity. We use $P_e^0$, $P_e$, and $P_\nu$ for the distributions specified by the model, and $\Pr_\pi$ for probabilities under the joint interaction process induced by the prior, the task-specific stochastic kernels, and the policy. The Bayes value of $\pi$ is $V_\nu^\pi = \mathbb{E}_{\Pr_\pi}\left[\sum_{k=1}^H r_k\right]$. The Bayes-optimal value is $V_\nu^* = \max_\pi V_\nu^\pi$, where the maximum is over task-independent history policies, and the Bayes value suboptimality is $\mathrm{SubOpt}_H(\pi) = V_\nu^* - V_\nu^\pi$. When $\pi$ is learned from data or depends on random initialization, we condition on the resulting policy and evaluate it on an independent test episode.

**Proposition 1** (Posterior invariance under task-independent policies)**.** *For every task-independent history policy $\pi$ and every positive-probability history,* $\Pr_\pi(e \mid h_k) = \nu(e \mid h_k)$. *If* $\pi(a \mid h_k) > 0$, *then* $\Pr_\pi(y \mid h_k, a) = P_\nu(y \mid h_k, a)$.

Define the supported history sets recursively. Let $\mathcal{H}_1 = \{(o) : P_\nu^0(o) > 0\}$. For each $k \in \{1, \ldots, H\}$, let $\mathcal{C}_k = \mathcal{H}_k \times \mathcal{A}$ and define $\mathcal{H}_{k+1} = \{(h, a, r, o') : (h, a) \in \mathcal{C}_k,\ y = (r, o') \in \mathcal{Y},\ P_\nu(y \mid h, a) > 0\}$. The finite reachable history–action set is $\mathcal{C} = \bigcup_{k=1}^H \mathcal{C}_k = \{c_1, \ldots, c_m\}$.

A stochastic kernel $\mathsf{P}$ on $\mathcal{C}$ is supported on the history tree if, for every $k \in \{1, \dots, H\}$, every $(h, a) \in \mathcal{C}_k$, and every $y \in \mathcal{Y}$, the condition $\mathsf{P}(y \mid h, a) > 0$ implies $(h, a, y) \in \mathcal{H}_{k+1}$. Every stochastic kernel used in dynamic programming below is required to be supported on this history tree.

An exploratory episode first draws $e \sim \nu$ and $o_1 \sim P_e^0$, and then follows a randomized history policy $\beta$. The behavior policy is task-independent, meaning that its action distribution depends on the observed history but not on the unobserved task $e$: $\Pr_\beta(a_k = a \mid h_k = h, e) = \beta(a \mid h)$. We assume that $\beta$ has full support on the reachable history–action set, so $\beta(a \mid h) > 0$ for every $(h, a) \in \mathcal{C}$.

**Lemma 2** (Reachability and exploratory coverage)**.** *For any task-independent history policy, the stage-$k$ history belongs to $\mathcal{H}_k$ almost surely for every $k$, and every positive-probability outcome from a context in $\mathcal{C}_k$ produces a history in $\mathcal{H}_{k+1}$. If a task-independent behavior policy $\beta$ assigns positive probability to every action at every supported history, each history–action pair $c_i \in \mathcal{C}$ is visited with positive probability in an episode. Let $\kappa(i)$ be the unique stage such that $c_i \in \mathcal{C}_{\kappa(i)}$, and denote this visitation probability by $\rho_i = \Pr_\beta\big((h_{\kappa(i)}, a_{\kappa(i)}) = c_i\big) > 0$.*

**Proposition 3** (Passive-data minimax impossibility)**.** *There are two horizon-one tasks with the same distribution over observation-only token sequences and different unique optimal actions. Any learner that receives only samples from this passive distribution has worst-case expected value suboptimality at least $1/2$.*

## 3 Sparse Dependence and Generalization Beyond the Training Input Support

Let $\mathcal{Q} = \{(k, a) : 1 \le k \le H,\ a \in \mathcal{A}\}$ and $n_{\mathrm{q}} = |\mathcal{Q}|$. A fixed feature map assigns every observable decision history $h$ a binary feature vector $z(h) = (z_1(h), \dots, z_P(h)) \in \{-1, 1\}^P$, where $P \ge 2$ and each coordinate $z_j$ is a fixed deterministic function of the history. For a history $h \in \mathcal{H}_k$ and action $a \in \mathcal{A}$, the serializer provides the model with the input $(u, z(h))$, where $u = (k, a)$. The complete raw history and any information not represented by $u$ or $z(h)$ are not included in the model input. The same feature map and the same coordinate meanings are shared across all queries $u \in \mathcal{Q}$; only the relevant-coordinate index $j_\star(u)$ may vary with the query.

**Assumption 4** (One-sparse logistic regression model)**.** *We use a query $u = (k, a)$ to index the conditional outcome-prediction problem associated with stage $k$ and action $a$. For each query, the conditional outcome distribution depends on a single feature coordinate indexed by $j_\star(u) \in \{1, \dots, P\}$, where the relevant-coordinate map $j_\star : \mathcal{Q} \to \{1, \dots, P\}$ is fixed but unknown to the learner. Let $\lambda > 0$ be known, and let $\mathrm{sigmoid}(x) = (1 + \exp(-x))^{-1}$.*

*For every query $u = (k, a)$, let $y_u^+, y_u^- \in \mathcal{Y}$ be two distinct known outcomes. For every binary feature vector $z \in \{-1, 1\}^P$, define the target conditional outcome distribution $P^\star_{\mathrm{feat}}(\cdot \mid u, z)$ by $P^\star_{\mathrm{feat}}(y_u^+ \mid u, z) = \mathrm{sigmoid}(\lambda z_{j_\star(u)})$, $P^\star_{\mathrm{feat}}(y_u^- \mid u, z) = 1 - P^\star_{\mathrm{feat}}(y_u^+ \mid u, z)$, and $P^\star_{\mathrm{feat}}(y \mid u, z) = 0$ for every $y \notin \{y_u^+, y_u^-\}$. We refer to $\lambda$ as the logistic regression coefficient.*

*For every supported history $h \in \mathcal{H}_k$ and action $a \in \mathcal{A}$, with $u = (k, a)$, the Bayes predictive kernel satisfies $P_\nu(\cdot \mid h, a) = P^\star_{\mathrm{feat}}(\cdot \mid u, z(h))$. For every $(u, z) \in \mathcal{Q} \times \{-1, 1\}^P$, we use the shorthand $P_\nu(\cdot \mid u, z) := P^\star_{\mathrm{feat}}(\cdot \mid u, z)$. Throughout the sparse-dependence analysis in Section 3 and Appendices A and F, the outcomes $y_u^+$ and $y_u^-$ are represented by the labels $+1$ and $-1$, respectively. This correspondence is known to the model and the external planner.*

For each candidate feature coordinate $j \in \{1, \dots, P\}$ and sign $s \in \{-1, +1\}$, define the binary feature vector $z^{j,s} \in \{-1, +1\}^P$ by $z_j^{j,s} = s$ and $z_\ell^{j,s} = -s$ for every $\ell \ne j$.

**Assumption 5** (Input distribution independent of the relevant-coordinate map)**.** *For every query $u \in \mathcal{Q}$, the model uses the indexed family $(z^{j,s})_{j \in \{1,\dots,P\},\, s \in \{-1,+1\}}$. We sample the index triple $(u, j, s)$ uniformly from $\mathcal{Q} \times \{1, \dots, P\} \times \{-1, +1\}$, so every index triple has probability $1/(2Pn_{\mathrm{q}})$. Given $(u, j, s)$, the model receives the serialized input $(u, z^{j,s})$; the pair $(j, s)$ is a sampling index and is not included in the model input. If two indexed feature vectors coincide, their multiplicities are retained in the induced training distribution. This convention includes the case $P = 2$, in which some indexed feature vectors coincide. Any two raw histories that produce the same serialized input have the same conditional outcome distribution. The sampling-index distribution, the serializer, and the set of raw histories mapped to each serialized input do not depend on $j_\star$.*

Proposition 17 realizes this design within a finite controlled process. Proposition 18 gives an exact one-head causal-attention realization of the following model. The model is parameterized by an attention score matrix $B \in \mathbb{R}^{n_q \times P}$, with attention weights $\alpha_j(B;u) = \exp(B_{u,j})/\sum_r \exp(B_{u,r})$, summary $r_B(u,z) = \sum_j \alpha_j(B;u) z_j$, and prediction $\widehat{P}_B(+1 \mid u,z) = \mathrm{sigmoid}(\lambda r_B(u,z))$. For a planning context $(h,a) \in \mathcal{C}_k$ with $u = (k,a)$, this binary prediction defines the learned predictive kernel by $\widehat{P}_B(y_u^+ \mid h,a) = \widehat{P}_B(+1 \mid u, z(h))$ and $\widehat{P}_B(y_u^- \mid h,a) = 1 - \widehat{P}_B(y_u^+ \mid h,a)$, with zero probability assigned to every other outcome.

Set $p_\lambda = \mathrm{sigmoid}(\lambda)$. For $q \in [0,1]$ and $p \in (0,1)$, define the binary cross-entropy between Bernoulli distributions with success probabilities $q$ and $p$ by $\mathrm{ce}_{\mathrm{bin}}(q,p) = -q \log p - (1-q)\log(1-p)$. The population binary cross-entropy loss (Murphy, 2012) is $\mathcal{L}_{\mathrm{sel}}(B) = (2Pn_q)^{-1} \sum_{u,j,s} \mathrm{ce}_{\mathrm{bin}}(P_\nu(+1 \mid u, z^{j,s}), \widehat{P}_B(+1 \mid u, z^{j,s}))$. Define $\mathcal{L}^*_{\mathrm{sel}} = (2Pn_q)^{-1} \sum_{u,j,s} \mathrm{Ent}(P_\nu(\cdot \mid u, z^{j,s}))$, where Ent denotes Shannon entropy (Shannon, 2001) and $0 \log 0 = 0$. The quantity $\mathcal{L}^*_{\mathrm{sel}}$ is the infimum of $\mathcal{L}_{\mathrm{sel}}$ over $B \in \mathbb{R}^{n_q \times P}$, although it is not attained by any finite score matrix. For every $B \in \mathbb{R}^{n_q \times P}$, the difference $\mathcal{L}_{\mathrm{sel}}(B) - \mathcal{L}^*_{\mathrm{sel}}$ is the uniform average of the conditional Kullback–Leibler divergences over all $u, j, s$. Training starts from $B_0 = 0$.

**Theorem 6** (Population gradient-flow convergence for attention-based feature selection)**.** *Under Assumptions 4 and 5, population gradient flow has a unique global solution. Every score row has one relevant score and one common irrelevant score. Let $g(t)$ be their difference, let $\alpha_\star(t) = e^{g(t)}/(e^{g(t)} + P - 1)$, let $\varepsilon_{\mathrm{sel}}(t) = 1 - \alpha_\star(t)$ denote the total attention mass assigned to the irrelevant coordinates, and let $m_\lambda = \mathrm{sigmoid}(\lambda)(1 - \mathrm{sigmoid}(\lambda))$. Then $g(t)$ is strictly increasing, $g(t) \to \infty$, and $\varepsilon_{\mathrm{sel}}(t) \le \left[\frac{P^2}{(P-1)^2} + \frac{8\lambda^2 m_\lambda t}{n_q P (P-1)^2}\right]^{-1/2}$. For every query and every binary vector $z \in \{-1,1\}^P$, $\mathrm{TV}(P_\nu(\cdot \mid u,z), \widehat{P}_{B_t}(\cdot \mid u,z)) \le \lambda \varepsilon_{\mathrm{sel}}(t)/2$ and $D_{\mathrm{KL}}(P_\nu(\cdot \mid u,z) \,\|\, \widehat{P}_{B_t}(\cdot \mid u,z)) \le \lambda^2 \varepsilon_{\mathrm{sel}}(t)^2/2$. The excess population loss satisfies $\mathcal{L}_{\mathrm{sel}}(B_t) - \mathcal{L}^*_{\mathrm{sel}} \le \lambda^2 \varepsilon_{\mathrm{sel}}(t)^2/2$.*

Theorem 19 gives the same bound for full-batch gradient descent after replacing $t$ by $\eta_{\mathrm{gd}} n$, for any positive step size. This conclusion follows from monotone score-gap updates in the one-sparse binary model. The next result shows that this property extends beyond the symmetric design.

**General training distributions.** For each query, let $D_u$ be a distribution on $\{-1,1\}^P$, let $w_u > 0$ with $\sum_u w_u = 1$, and train on $\mathcal{L}_D(B) = \sum_u w_u \mathbb{E}_{D_u}[\mathrm{ce}_{\mathrm{bin}}(P_\nu(+1 \mid u, Z), \widehat{P}_B(+1 \mid u, Z))]$. Define $\kappa_u = \min_{j \ne j_\star(u)} \Pr_{D_u}\{Z_j \ne Z_{j_\star(u)}\}$.

**Theorem 7** (General input distributions)**.** *Under Assumption 4, the relevant coordinate for query $u$ is identifiable from its conditional outcome probabilities under a common input distribution $D_u$ if and only if $\kappa_u > 0$. If this holds for every query, population gradient flow and full-batch gradient descent on $\mathcal{L}_D$, initialized at zero, satisfy $\varepsilon_u(\tau) \le \left[\frac{P^2}{(P-1)^2} + \frac{6 w_u \lambda^2 m_\lambda \kappa_u \tau}{P^2(P-1)}\right]^{-1/2}$, where $\varepsilon_u = 1 - \alpha_{j_\star(u)}$, and $\tau = t$ for gradient flow or $\tau = \eta_{\mathrm{gd}} n$ for gradient descent with any $\eta_{\mathrm{gd}} > 0$. The uniform prediction bounds of Theorem 6 hold with $\varepsilon_{\mathrm{sel}}$ replaced by $\varepsilon_u$; the planning bound uses $\max_u \varepsilon_u$.*

Appendix R proves the theorem and gives finite-sample guarantees for empirical gradient flow from independent exploratory episodes, with explicit dependence on query visitation and feature disagreement probabilities. It also constructs a fixed design with $\lceil \log_2 P \rceil$ inputs, matching the identifiability lower bound. The preceding results provide separate bounds for the sign-paired design used in the experiments.

**Finite conditional samples.** For the symmetric design, Theorem 22 treats empirical gradient flow with $N_{\mathrm{sel}}$ independent outcomes at each indexed input. With $\varepsilon_{\mathrm{stat}} = \sqrt{\log(2Pn_q/\delta)/(4N_{\mathrm{sel}})}$ and $r_{\mathrm{stat}} = 4P\varepsilon_{\mathrm{stat}}/(\lambda m_\lambda) \le 1/2$, its uniform prediction bound is controlled by $\max_u \varepsilon_u(t) \le \max\left\{\left[\frac{P^2}{(P-1)^2} + \frac{4\lambda^2 m_\lambda t}{n_q P^3}\right]^{-1/2}, r_{\mathrm{stat}}\right\}$ with probability at least $1 - \delta$. Appendix A gives the complete sampling model and explicit sample and training requirements. Corollary 45 extends empirical gradient flow to ordinary independent episodes under positive query visitation and feature disagreement.

**Corollary 8** (Planning Beyond the Training Input Support)**.** *Let $\widehat{\pi}_t^{\mathrm{sel}}$ be an exact dynamic-programming policy under the learned predictive kernel $\widehat{P}_{B_t}$ obtained by population gradient flow, and let $\widehat{\pi}_n^{\mathrm{gd}}$ be an exact dynamic-programming policy under the learned predictive kernel $\widehat{P}_{B^{(n)}}$ obtained by full-batch gradient descent. Then* $\mathrm{SubOpt}_H(\widehat{\pi}_t^{\mathrm{sel}}) \le \lambda H(H+1)\varepsilon_{\mathrm{sel}}(t)/2$ *and* $\mathrm{SubOpt}_H(\widehat{\pi}_n^{\mathrm{gd}}) \le \lambda H(H+1)\varepsilon_{\mathrm{sel}}^{(n)}/2$.

*Remark* 9 (Scope of the sparse result). The feature map is fixed, the feature dimension does not grow at test time, and the network learns one relevant coordinate for each stage–action query. The result does not form a recurrent predictive state or prove length extrapolation to larger feature sets.

## 4 Multi-Head Causal Self-Attention on a Finite Reachable History–Action Set

The complementary result allows arbitrary conditional outcome distributions with full support on $\mathcal{Y}$ at every context in the finite reachable history–action set. Write $K = |\mathcal{Y}|$ and index the coordinates of $\mathbb{R}^K$ by the outcomes $y \in \mathcal{Y}$. For target distributions $q$ and model-predicted distribution $p$ on $\mathcal{Y}$, define $\mathrm{CE}(q,p) = -\sum_{y\in\mathcal{Y}} q(y)\log p(y)$, $\mathrm{Ent}(q) = -\sum_{y\in\mathcal{Y}} q(y)\log q(y)$, and $D_{\mathrm{KL}}(q\,\|\,p) = \sum_{y\in\mathcal{Y}} q(y)\log(q(y)/p(y))$. For $f \in \mathbb{R}^K$, define $\mathrm{softmax}(f)_y = \exp(f_y)/\sum_{b\in\mathcal{Y}}\exp(f_b)$.

For each history–action pair $c_i$, let $q_i = P_\nu(\cdot \mid c_i)$ denote its Bayes predictive distribution. Because every episode contains exactly $H$ stage-specific history–action pairs, $\sum_i \rho_i = H$. The normalized visitation weights $d_i = \rho_i/H$ therefore satisfy $\sum_i d_i = 1$. Assume that every $q_i$ has full support on $\mathcal{Y}$. Since $\mathcal{C}$ and $\mathcal{Y}$ are finite and every $\rho_i$ is positive, define $\eta = \min_{i,y} q_i(y) > 0$ and $d_{\min} = \min_i d_i > 0$. For an exploratory episode, let $i(k)$ be the index of the context $(h_k, a_k)$. The population cross-entropy loss for next-outcome prediction induced by exploratory episodes is $\mathcal{L}(f) = \mathbb{E}_{\mathrm{Pr}_\beta}[H^{-1}\sum_{k=1}^H -\log \mathrm{softmax}(f_{i(k)})_{y_k}] = \sum_i d_i \,\mathrm{CE}(q_i, \mathrm{softmax}(f_i))$. It is minimized when $\mathrm{softmax}(f_i) = q_i$ for every $i$, with minimum $\mathcal{L}^* = \sum_i d_i \,\mathrm{Ent}(q_i)$.

A fixed serializer and shared embedding matrix $W$ map context $c_i$ to token–position embeddings $(x_{i,1}, \ldots, x_{i,\ell_i})$, with $\ell_i \le L$ and $\|x_{i,j}\| \le R$. Write $u_i = x_{i,\ell_i}$. One causal attention head has weights $\alpha_{i,j}(A) = \exp(u_i^\top A x_{i,j})/\sum_s \exp(u_i^\top A x_{i,s})$, summary $s_i(A) = \sum_j \alpha_{i,j}(A) x_{i,j}$, and output $\varphi_i(A,v) = v^\top s_i(A)$. The final query attends to its prefix including itself. For $\omega = (A, v)$ use $\|\omega\|^2 = \|A\|_F^2 + \|v\|^2$.

The network contains $M$ independent head pairs. Draw independent Rademacher vectors $\epsilon_r \in \{-1,1\}^d$ and $b_r \in \{-1,1\}^K$, set $v_r^0 = \sigma_v \epsilon_r$, and initialize both copies at $\omega_r^0 = (0, v_r^0)$. Keep $b_r$ fixed and train all score matrices and value vectors. The logits are $f_i(\theta) = \dfrac{1}{\sqrt{2M}} \displaystyle\sum_{r=1}^{M} b_r[\varphi_i(\omega_{r,+}) - \varphi_i(\omega_{r,-})]$, so $f_i(\theta_0) = 0$. Let $\boldsymbol{\Theta}_A$ be the score-parameter contribution to the neural tangent kernel of the stacked logits. At zero scores, define $\bar{x}_i = \ell_i^{-1}\sum_j x_{i,j}$, $C_i = \ell_i^{-1}\sum_j (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^\top$, and $[Z_{\mathrm{ctx}}]_{i,:} = (u_i \otimes \mathrm{vec}(C_i))^\top$. We use $\sigma_{\min}(Z) = \sqrt{\lambda_{\min}(ZZ^\top)}$, which is positive exactly when $Z$ has full row rank. Appendix B specifies the architecture and Appendix C derives its Jacobian.

**Proposition 10** (Rank of the context feature matrix and Jacobian Gram Matrix for the Attention Score Matrices)**.** *If $r = \mathrm{rank}(W)$, then $\mathrm{rank}(Z_{\mathrm{ctx}}) \le r^2(r+1)/2$. At initialization, the expected Gram matrix of the Jacobian with respect to the attention score matrices is $\boldsymbol{G}_A = \sigma_v^2 Z_{\mathrm{ctx}} Z_{\mathrm{ctx}}^\top$. Hence the condition $\sigma_{\min}(Z_{\mathrm{ctx}}) \ge \zeta > 0$ gives $\boldsymbol{G}_A \succeq \gamma \boldsymbol{I}_m$ with $\gamma = \sigma_v^2\zeta^2$.*

The two claims are proved in Propositions 25 and 30. The rank condition is computable from the serializer and shared embedding matrix. Appendix C proves the rank and Gram identities, generic full row rank conditional on an embedding witness, and an explicit one-hot sufficient condition.

Write $E(t) = \mathcal{L}(f(\theta_t)) - \mathcal{L}^*$ and $E_0 = E(0) = \log K - \sum_i d_i \,\mathrm{Ent}(q_i)$. Let $\gamma = \sigma_v^2\zeta^2$, $p_* = \exp[-\log K/(\eta d_{\min})]$, $\mu = d_{\min} K p_*^2$, and $c = \gamma\mu/2$. Appendix Q defines $C_{\mathrm{score}}, r_M, r_*$ and the explicit width requirements used below.

**Theorem 11** (Finite-Width Population Convergence)**.** *Assume the full-support condition above and $\sigma_{\min}(Z_{\mathrm{ctx}}) \ge \zeta > 0$. Fix $\delta \in (0,1)$ and assume $M \ge 8KmC_{\mathrm{score}}^2\gamma^{-1}\log(mK/\delta)$ and $r_M \le r_*/2$. Then, with probability at least $1-\delta$ over the random initialization $\{(\epsilon_r, b_r)\}_{r=1}^M$, population gradient*

*flow exists for all $t \geq 0$. With $P_K = \boldsymbol{I}_K - K^{-1}\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{P} = \boldsymbol{I}_m \otimes P_K$, its neural tangent kernel contribution from the attention score matrices satisfies $\boldsymbol{P}\boldsymbol{\Theta}_A(t)\boldsymbol{P} \succeq \frac{\gamma}{4}\boldsymbol{P}$, for all $t \geq 0$. For every $r \in \{1, \ldots, M\}$ and $s \in \{+, -\}$, $\sup_{t\geq 0} \left\|\omega_{r,s}(t) - \omega_r^0\right\|_2 \leq r_M$, and the excess cross-entropy satisfies $E(t) \leq E_0 e^{-ct}$ for all $t \geq 0$.*

The theorem follows from Theorems 32, 34, and 35 in the appendix.

**Theorem 12** (Trajectory-level finite-sample consequence)**.** *Let $N$ independent exploratory episodes be observed, and let $\rho_{\min}$ be the minimum episode-level visitation probability over $\mathcal{C}$. Under the conditions of Theorem 40, with probability at least $1 - 2\delta$ over the training episodes and random initialization, the empirical excess cross-entropy decays exponentially and $\max_i \mathrm{TV}\left(q_i, \widehat{P}_t(\cdot \mid c_i)\right) \leq \frac{K}{2}\sqrt{\frac{\log(4mK/\delta)}{N\rho_{\min}}} + \sqrt{\frac{\widehat{E}_0}{2\underline{d}}}e^{-\underline{c}t/2}$, where $\underline{d} = \rho_{\min}/(2H)$ and $\underline{c}$ is the empirical exponential convergence-rate constant defined in Appendix G. Any policy $\widehat{\pi}_t$ obtained by exact dynamic programming under $\widehat{P}_t$ satisfies $\mathrm{SubOpt}_H(\widehat{\pi}_t) \leq H(H+1)\left[\frac{K}{2}\sqrt{\frac{\log(4mK/\delta)}{N\rho_{\min}}} + \sqrt{\frac{\widehat{E}_0}{2\underline{d}}}e^{-\underline{c}t/2}\right]$.*

This is a direct consequence of Theorem 40 in the appendix.

### 4.1 Initialization and Higher-Order Interactions

The representation condition has a specific limitation for histories containing independently varying tokens. Consider $z \in \{-1, 1\}^P$, serialized as $P$ binary tokens followed by a fixed query. The shared one-hot embedding is $x_j(z) = e_{j,z_j}$ and $u = e_0$ in dimension $2P + 1$. Write $\chi_S(z) = \prod_{j\in S} z_j$, and let $\mathcal{V}_2 = \mathrm{span}\{\chi_S : |S| \leq 2\}$. Inputs are uniform on the binary cube. All attention score matrices and value vectors are trained, with the paired initialization and output scaling defined above.

**Proposition 13** (Uniform-attention initialization)**.** *At $A = 0$, every scalar coordinate of the full parameter Jacobian belongs to $\mathcal{V}_2$. This inclusion also holds after any fixed linear transformation of the shared embeddings. For the one-hot embedding, the expected scalar score kernel and the expected full scalar kernel have range exactly $\mathcal{V}_2$, hence rank $1 + P + \binom{P}{2}$. If $P \geq 3$, $|S| \geq 3$, and $q(1 \mid z) = \frac{1}{2} + \delta_{\mathrm{p}}\chi_S(z),\ 0 < \delta_{\mathrm{p}} < \frac{1}{2}$, then every finite-width paired network initialized at zero score matrices is stationary under population gradient flow and exact population gradient descent. Its prediction remains uniform, with total variation error $\delta_{\mathrm{p}}$ at every input. Appendix T.7 gives a one-stage planning consequence.*

To examine whether this failure requires a different representation, retain the same tokens and network and change only the score initialization. Independently for each head pair, draw $\xi_{\mathrm{sc}} \in \{-1, 1\}^P$ uniformly and set $A^\tau(\xi_{\mathrm{sc}}) = \tau e_0 \left(\sum_{j=1}^{P} \xi_{\mathrm{sc},j}(e_{j,1} - e_{j,-1})\right)^\top$. The token scores are $\tau\xi_{\mathrm{sc},j}z_j$, the query score is zero, and paired logits remain zero. Let $k_A^\tau$ and $k_v^\tau$ be the expected scalar-head Jacobian kernels for score and value parameters. Their operators use the uniform measure: $(\mathcal{T}g)(z) = 2^{-P}\sum_{z'} k(z, z')g(z')$.

**Theorem 14** (Initialization spectrum)**.** *The Walsh functions diagonalize both operators, with eigenvalues $\lambda_s^A(\tau)$ and $\lambda_s^v(\tau)$ depending only on $s = |S|$. For fixed $P$ and $2 \leq s \leq P$, as $\tau \to 0$, $\lambda_s^A(\tau) = \frac{\sigma_v^2\binom{s}{2}((s-1)!)^2}{2(P+1)^{2s}}\tau^{2s-4} + O_{P,s,\sigma_v}(\tau^{2s-2})$, and $\lambda_s^v(\tau) = s((s-1)!)^2\tau^{2s-2}/(2(P+1)^{2s}) + O_{P,s}(\tau^{2s})$. Both expected kernels are positive definite for every $\tau \neq 0$.*

Appendix S proves both results and specifies the normalization. Increasing width alone cannot remove the zero-score failure. Moreover, for the binary-margin eigenvalue $\kappa_s = 2(\lambda_s^A + \lambda_s^v)$, a nonasymptotic bound gives, for every $\tau \neq 0$ and $2 \leq s \leq P$, $\kappa_s \geq \underline{\kappa}_s(\tau) := \frac{\sigma_v^2\binom{s}{2}((s-1)!)^2 e^{-4|\tau|}\sinh^{2s-4}(|\tau|)}{(1 + Pe^{|\tau|})^{2s}}$. The proof uses a constant-sign integral for a Walsh coefficient, rather than a dimension-dependent

Taylor remainder (Proposition 49). Thus the bound is explicit in initialization amplitude, dimension, and interaction order.

**Finite-width learning of a fixed parity.** The smallest eigenvalue over the entire Boolean cube can be much smaller than the eigenvalue of the target. The following result uses only the latter. Let $\chi_S(z) = \prod_{j\in S} z_j$, $s = |S|$, and let $\lambda_s = \lambda_s^A(\tau) + \lambda_s^v(\tau) > 0$ be its eigenvalue for the expected full scalar head kernel under the nonzero initialization above. The corresponding binary-margin kernel has eigenvalue $\kappa_s = 2\lambda_s$.

**Theorem 15** (Finite-width population learning from a target eigenvalue)**.** *Use the paired network and initialization above. Let $q(z) = 1/2 + \rho_{\mathrm{p}}\chi_S(z)$ with $0 < \rho_{\mathrm{p}} < 1/2$, and set $a_{\mathrm{p}} = \log((1/2 + \rho_{\mathrm{p}})/(1/2 - \rho_{\mathrm{p}}))$. Train all score and value parameters by population gradient flow on binary cross-entropy. For $\varepsilon, \delta \in (0,1)$, put $R_s^2 = 2a_{\mathrm{p}}^2/(\kappa_s\delta)$ and $T = 2R_s^2/\varepsilon$. Under the explicit width condition in equation* (216)*, with probability at least $1-\delta$ over initialization, the solution exists on $[0,T]$, every head moves by at most $1/2$, and $\mathcal{L}(\theta_T) - \mathcal{L}^* \le \varepsilon,\ \mathbb{E}_z[(\widehat{p}_T(z) - q(z))^2] \le \varepsilon/2$, where $\widehat{p}_t(z) = \mathrm{sigmoid}(f_1(z;\theta_t) - f_2(z;\theta_t))$. The derivative constants and width requirements are collected in Appendix T.1.*

The proof controls distance to a comparison point built from the target eigenfunction (Appendix T).

**Independent noisy outcomes and nonzero initial outputs.** Theorem 51 extends the result to unprojected empirical gradient flow on independent inputs and Bernoulli outcomes. It allows a starting point $\theta_{\mathrm{init}} = \theta_0 + U$ with $\|U\| \le r_{\mathrm{init}}$ and each head displaced by at most $1/4$. Under the explicit sample and width conditions in equations (247)–(248), population excess loss is at most $\varepsilon$ at time $T_{\mathrm{emp}} = 4R_{\mathrm{emp}}^2/\varepsilon$, where $R_{\mathrm{emp}} = \max\{1, 2a_{\mathrm{p}}/\sqrt{\kappa_s\delta} + r_{\mathrm{init}}\}$. A sufficient sample size is of order $(C_0 R_{\mathrm{emp}} + a_{\mathrm{p}})^2 \log(8/\delta)/\varepsilon^2$. It requires neither exact conditional probabilities nor complete input coverage. The nonzero-output extension is local in parameter space, not an unrestricted initialization result.

**Complexity and its limitation.** For $s \ge 2$ and $\tau \ne 0$, substituting $\underline{\kappa}_s(\tau)$ into either theorem makes every resource bound explicit. Equation (212) also gives a uniform small-scale bound with explicit dimension and order dependence. However, the trace bound $\sum_{j=0}^{P} \binom{P}{j}\kappa_j \le 2(\sigma_v^2 + 1)$ limits these guarantees at growing order. For $s = \lfloor P/2 \rfloor$, their kernel-norm dependence can be exponential in $P$ (Appendix S.6). Neither theorem asserts uniform prediction error or a lower bound on nonlinear training time.

## 5 External Planning and Bayes Value Suboptimality

The learned predictor is used by an external planner. Proposition 36 shows that the original task mixture and the history process with initial law $P_\nu^0$ and predictive kernel $P_\nu$ have the same trajectory law for every task-independent policy. Backward induction under $P_\nu$ is therefore Bayes-optimal. Appendix E gives the recursion and the complete planning statements.

For two kernels supported on the same history tree and using the same initial law, a uniform total-variation error $\varepsilon$ implies $|V_{\mathsf{P}}^\pi - V_{\widehat{P}}^\pi| \le H(H+1)\varepsilon/2$ for every policy. An optimizer under $\widehat{P}$ consequently has value gap at most $H(H+1)\varepsilon$ (Lemma 37). This standard reduction yields the following consequence of the finite reachable-set theorem.

**Theorem 16** (Causal self-attention training to Bayes value suboptimality)**.** *Under the conditions of Theorem 11, let $\widehat{\pi}_t$ be an exact dynamic-programming policy for $\widehat{P}_t(\cdot \mid c_i) = \mathrm{softmax}(f_i(\theta_t))$. With probability at least $1-\delta$, $\mathrm{SubOpt}_H(\widehat{\pi}_t) \le H(H+1)\sqrt{E_0/(2d_{\min})}\, e^{-ct/2}$.*

The average-error guarantees for a fixed parity in Section 4.1 are separate; they do not supply this uniform multistep planning conclusion.

# 6 EXPERIMENTS AND DISCUSSION

Experiments 1 and 2 examine sparse feature selection, Experiments 3 and 4 finite-width learning on finite context sets, Experiment 5 initialization and higher-order learning, and Experiment 6 external planning under the learned predictive model. We use FP64 CUDA arithmetic with deterministic algorithms and TF32 disabled. Experiment 1 is deterministic; the remaining experiments use ten seeds per training setting, with 100 initializations per representation–width setting in Experiment 3. Reported variability is the sample standard deviation across seeds. In Experiments 1–4, prediction errors are maximized over queries and feature vectors in the sparse setting, and over contexts in the multi-head setting. Protocols and numerical checks are given in Appendix U. All experiments were run on an NVIDIA GeForce RTX 4090 GPU.

## 6.1 EXPERIMENT 1: POPULATION FEATURE SELECTION

We use $\lambda = 1$, $n_{\mathrm{q}} = 4$, $P \in \{4, 8, 12, 16\}$, and the sign-paired design of Assumption 5. Training starts from $B = 0$ and uses exact conditional probabilities on $2P$ feature vectors per query. Evaluation enumerates all $2^P$ vectors per query, including the $2^P - 2P$ vectors outside the training support. Appendix Table 1 reports population gradient flow at $t = 50{,}000$. For $P = 16$, relevant attention grows from $0.0625$ to $0.97156$, and maximum unseen-input TV falls from $0.41024$ to $0.01056$. Larger feature sets retain more error at the same training time; generalization is evaluated at fixed dimension within each run. The gradient-descent comparisons and all 96 recorded flow and descent checkpoints are reported in Appendix U.2; each satisfies the corresponding attention, TV, and KL bounds.

## 6.2 EXPERIMENT 2: FINITE CONDITIONAL SAMPLES

Using the same design with $P \in \{4, 8\}$, we draw $N_{\mathrm{sel}} \in \{256, 1024, 4096, 16384, 65536, 262144\}$ independent outcomes per indexed input and train empirical gradient flow to $t = 50{,}000$. The maximum symmetrized-frequency error $\max_{u,k} |\widehat{t}_{u,k} - t_{u,k}|$ decreases approximately as $N_{\mathrm{sel}}^{-1/2}$: log–log fits to the six group means give slopes $-0.518$ and $-0.506$. At the largest sample count, mean maximum TV is $0.00448$ for $P = 4$ and $0.00738$ for $P = 8$, approaching the population values in Table 1 at the same time, not an asymptotic statistical floor. The full sample-size comparison appears in Table 2 in the appendix. All checkpoints satisfying the sample-size condition of Theorem 22 satisfy its bounds; Appendix U.3 reports the qualifying settings and complete counts.

## 6.3 EXPERIMENT 3: REPRESENTATION AND KERNEL STABILITY

We train the paired multi-head model on ten contexts with full-support binary targets and uniform weights $d_i = 0.1$, using embedding dimension 16, $\sigma_v = 0.25$, and $M \in \{16, 64, 256\}$ head pairs. Both representation settings have $\mathrm{rank}(Z_{\mathrm{ctx}}) = 10$, but their singular-value margins are $\zeta = 0.31427$ and $0.05727$. Appendix Table 3 distinguishes relative operator-norm distance to the expected kernel at initialization from subsequent drift relative to each run's initial kernel.

Initialization error and head movement decrease approximately as $M^{-1/2}$, while kernel drift also decreases with width. The initialization inequality $\lambda_{\min}(\boldsymbol{\Theta}_A(0)) \geq \gamma/2$ holds in all 100 initializations at $M = 256$ for each representation. The centered kernel inequality $\boldsymbol{P}\boldsymbol{\Theta}_A(t)\boldsymbol{P} \succeq (\gamma/4)\boldsymbol{P}$ holds numerically at all 540 training checkpoints. Yet full row rank does not imply rapid finite-time optimization: the smaller representation margin accompanies maximum TV near $0.20$, compared with $0.01$–$0.02$ at scale 1. Increasing width improves kernel stability without monotonically improving fixed-time prediction. These comparisons support the role of representation and width, but do not isolate $\gamma$ as the sole determinant of the training rate.

## 6.4 EXPERIMENT 4: TRAJECTORY COVERAGE AND EMPIRICAL FITTING

We train $M = 64$ head pairs on independent length-two episodes with ten contexts, using behavior-action probabilities $b \in \{0.5, 0.1, 0.02\}$ and empirical visitation weights. Appendix U.5 specifies the stopping rule and retains budget-exhausted runs.

Let $D(t)$ be maximum TV against the true conditionals over all contexts, and $F(t)$ maximum TV against empirical conditionals over visited contexts. Under full coverage, $S = \max_i \mathrm{TV}(q_i, \widehat{q}_i)$ gives $D(t) \le S + F(t)$, an upper bound rather than an additive identity. Appendix Table 4 separates common-time prediction from prediction at each run's stopping time $T$.

All 40 runs at $b = 0.5$ reach the fitting tolerance, and sampling and terminal prediction errors decrease with $N$. All 80 runs at $b \in \{0.1, 0.02\}$ exhaust the budget. At $b = 0.02$ and $N = 65{,}536$, full coverage coexists with mean fitting error $0.3319$ and prediction error $0.3928$, despite mean empirical excess loss of only $1.17 \times 10^{-4}$. Thus small visitation-weighted loss does not ensure accurate prediction at every context, and more episodes need not resolve incomplete fitting within the available training time. This distinguishes coverage from optimization in the error decomposition of Theorem 12.

### 6.5 Experiment 5: Initialization and higher-order learning

We use the model of Section 4.1 with $\rho_{\mathrm{p}} = 0.2$ and $\sigma_v = 0.25$. Training results below are ten-seed means at $t = 200$; population errors average over the Boolean cube. For $P \in \{4, 6, 8, 12\}$ and $3 \le s \le \min\{6, P\}$, exhaustive Walsh calculations at $\tau \in \{0.05, 0.1, 0.2\}$ give exponent errors below $0.047$ relative to $2s - 4$ and $2s - 2$ for the score and value kernels. At $P = 6$, zero-score population training retains excess cross-entropy $0.082283$, whereas $\tau = 4, M = 1024$ reduces it to $0.072567$ and $0.076896$ for $s = 3, 4$, respectively. At $s = 3, \tau = 4$, increasing $M$ from 64 to 1024 reduces mean maximum per-head movement from $0.1881$ to $0.0462$.

**Finite samples and generalization.** With $\tau = 4, M = 256$, we train on $n_{\mathrm{p}} \in \{32, 128, 512, 2048\}$ independent noisy examples for $P \in \{6, 8\}$ and $s \in \{3, 4\}$. Nonlinear and full score-and-value linearized models differ in mean population excess loss by at most $1.61 \times 10^{-3}$ across recorded checkpoints. At $P = 6, n_{\mathrm{p}} = 2048$, excess losses are $0.07437$ and $0.07838$ for $s = 3, 4$, below the initial $0.08228$. However, empirical fitting need not improve population prediction. For $P = 8, n_{\mathrm{p}} = 1024$, mean coverage exceeds $98\%$, yet unseen-input mean squared errors reach $0.04583$ and $0.04272$, respectively, compared with $0.04$ initially.

**Perturbed initialization.** At $P = 6, s = 3, n_{\mathrm{p}} = 128$, perturbation radii $r_{\mathrm{init}} \in \{0, 0.5, 1, 2\}$ yield mean target Walsh coefficients of the predicted probabilities between $0.01250$ and $0.01252$. Mean maximum per-head distances from the paired reference initialization range from $0.3381$ to $0.3707$. Mean excess losses differ by less than $7 \times 10^{-5}$ across radii but remain near $0.1068$. Thus bounded nonzero-output perturbations preserve the observed target-component learning without eliminating finite-sample overfitting.

### 6.6 Experiment 6: Planning and Bayes value recovery

We use a two-stage problem in which the greedy root action has value $1.1$, while the Bayes-optimal action has value $1.185$. With $M = 128$, all ten runs recover the complete Bayes-optimal policy by $t = 205$ and retain it thereafter, giving zero terminal value suboptimality. At $t = 205$, mean maximum TV remains $0.155$, so policy recovery precedes high-accuracy conditional estimation. For master seed 0, matched-time checks with $dt \in \{0.05, 0.025, 0.0125\}$ give identical policies and values at the recorded training times.

## 7 Limitations

The one-sparse analysis assumes a fixed binary feature map, one relevant coordinate per query, and positive visitation and feature disagreement. The finite reachable-set results require full context coverage, full-support outcomes, and a singular-value condition, with finite-sample costs that may grow poorly with horizon. Higher-order guarantees assume uniform binary inputs and fixed parity targets, control average finite-time error under gradient flow, and allow only bounded perturbations of paired initialization. Their sufficient complexity may be exponential at growing order. Planning is external, and the theory neither covers deep-Transformer convergence nor learns features from raw histories.

### AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs, and to support the verification of portions of the experimental code. All AI-assisted mathematical work and experimental code were independently verified by the authors against the stated assumptions, derivations, and reported results, and all references identified with AI assistance were checked against the original publications. The authors take full responsibility for the final content of this work.

### REPRODUCIBILITY STATEMENT

Complete proofs of the theoretical results are provided in the appendix. Appendix U reports experimental protocols, implementation details, hardware and software configurations, random seeds, evaluation procedures, and numerical results. The supplementary material contains the source code used to reproduce the experiments and numerical evaluations reported in the paper.

## REFERENCES

Bernardo Ávila Pires and Csaba Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 121–151, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/avilapires16.html.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin F. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Xi-Ren Cao and Xianping Guo. Partially observable markov decision processes with reward information: Basic ideas and models. *IEEE Transactions on Automatic Control*, 52(4):677–681, 2007. doi: 10.1109/TAC.2007.894520.

Thomas Chen, Tengyu Ma, and Zhiyuan Li. Non-asymptotic length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Elvis Dohmatob. Understanding softmax attention layers:\\exact mean-field analysis on a toy problem. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 31305–31336. Curran Associates, Inc., 2025. doi: 10.52202/085713-1051. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2cf63608d264a3c65e68e3cff860c152-Paper-Conference.pdf.

Odilon Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *International Conference on Learning Representations*, volume 2026, pp. 87068–87100, 2026.

Ting-Han Fan and Peter Ramadge. A contraction approach to model-based reinforcement learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 325–333. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/fan21a.html.

Cheng Gao, Yuan Cao, Zihao Li, Yihan He, Mengdi Wang, Han Liu, Jason M. Klusowski, and Jianqing Fan. Global convergence in training large-scale transformers. In *Proceedings of the 38th*

*International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Noah Golowich, Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. The role of sparsity for length generalization in LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19809–19840. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/golowich25a.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Jianliang He, Siyu Chen, Fengzhuo Zhang, and Zhuoran Yang. From words to actions: unveiling the theoretical underpinnings of llm-driven autonomous systems. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. URL http://www.jstor.org/stable/2282952?

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4376–4386. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hron20a.html.

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

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

M. Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: unveiling the dynamics of generative transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Alexander Kozachinskiy, Tomasz Steifer, and Przemyslaw Andrzej Walega. Parity, sensitivity, and transformers. *ArXiv*, abs/2602.05896, 2026. URL https://api.semanticscholar.org/CorpusID:285303924.

Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561. MIT Press, 2001.

Eran Malach, Pritish Kamath, Emmanuel Abbe, and Nathan Srebro. Quantifying the benefit of using differentiable learning over tangent kernels. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7379–7389. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/malach21a.html.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DVlPp7Jd7P.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Basak Sakcak, Kalle G. Timperi, Vadim Weinstein, and Steven M. LaValle. A mathematical characterization of minimally sufficient robot brains. *The International Journal of Robotics Research*, 43(9):1342–1362, 2024. doi: 10.1177/02783649231198898.

Matteo Saponati, Pascal Josef Sager, Pau Vilimelis Aceituno, Thilo Stadelmann, and Benjamin F Grewe. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in transformer training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52958–52994. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/saponati25a.html.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, pp. 3067–3075. PMLR, 2017.

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001. ISSN 1559-1662. doi: 10.1145/584091.584093. URL https://doi.org/10.1145/584091.584093.

Wei Shi and Yuan Cao. Towards understanding transformers in learning random walks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZacSviw4em.

Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. *Advances in Neural Information Processing Systems*, 37: 92317–92351, 2024.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4): 389–434, August 2012. ISSN 1615-3375.

Muhammed Ustaomeroglu and Guannan Qu. A theoretical study of (Hyper) self-attention through the lens of interactions: Representation, training, generalization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60657–60710. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ustaomeroglu25a.html.

Aditya Varre, Gizem Yüce, and Nicolas Flammarion. Learning in-context $n$-grams with transformers: Sub-$n$-grams are near-stationary points. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60924–60963. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/varre25a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Siwei Wang, Yifei Shen, Shi Feng, Haoran Sun, Shang-Hua Teng, and Wei Chen. Alpine: unveiling the planning capability of autoregressive learning in language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9798331314385.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AmEgWDhmTr.

Diyuan Wu, Aleksandr Shevchenko, Samet Oymak, and Marco Mondelli. Attention with trained embeddings provably selects important tokens. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28956–28997. Curran Associates, Inc., 2025. doi: 10.52202/085713-0972. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/29be2340edb8b224728248d4bb1ea9d4-Paper-Conference.pdf.

Yongtao Wu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. On the convergence of encoder-only shallow transformers. *Advances in Neural Information Processing Systems*, 36:52197–52237, 2023.

# A Additional Results for the One-Sparse Logistic Regression Model

**Proposition 17** (Finite controlled-process realization)**.** *For every $H \geq 1$, every nonempty finite action set $\mathcal{A}$, every $P \geq 2$, and every $\lambda > 0$, there exist a finite observation set with at most $2^P + 2P + 1$ elements, a finite reward set, an initial observation distribution, a serializer, a full-support behavior policy, a rule that retains an exploratory episode according only to its initial observation, and a rule that selects one of the $H$ stages of each retained episode uniformly at random. These objects and rules can be chosen independently of the relevant-coordinate map. For every map $j_\star : \mathcal{Q} \to \{1, \dots, P\}$, there exists a reward–observation stochastic kernel (Cao & Guo, 2007) on these fixed spaces that satisfies Assumption 4. At every stage, each binary feature vector $z \in \{-1, 1\}^P$ is realized by at least one supported history. Moreover, retaining exploratory episodes according to the fixed retention rule and then selecting one of the $H$ stages of each retained episode according to the fixed stage-selection rule yields exactly the training input distribution in Assumption 5.*

**Proposition 18** (One-head causal-attention realization)**.** *For every attention score matrix $B \in \mathbb{R}^{n_{\mathrm{q}} \times P}$, the map $(u, z) \mapsto r_B(u, z)$ is exactly realized by a one-head causal self-attention layer with fixed query embeddings, fixed feature-coordinate embeddings, a strict causal mask, a fixed scalar value projection, and one trainable bilinear attention score matrix whose free entries are identified with $B$. Under the natural coordinate identification of this trainable matrix with $\mathbb{R}^{n_{\mathrm{q}} \times P}$, Euclidean gradient flow and full-batch gradient descent induce exactly the score dynamics analyzed below.*

**Theorem 19** (Full-batch gradient descent under the training input distribution)**.** *Under Assumptions 4 and 5, let $\eta_{\mathrm{gd}} > 0$ be the gradient-descent step size, set $B^{(0)} = 0$, and define*

$$B^{(n+1)} = B^{(n)} - \eta_{\mathrm{gd}} \nabla_B \mathcal{L}_{\mathrm{sel}}(B^{(n)}). \tag{1}$$

*Every row preserves one relevant score and one common irrelevant score. If $\varepsilon_{\mathrm{sel}}^{(n)}$ is the total attention mass on irrelevant coordinates, then*

$$\varepsilon_{\mathrm{sel}}^{(n)} \leq \left[ \frac{P^2}{(P-1)^2} + \frac{8\eta_{\mathrm{gd}} \lambda^2 m_\lambda n}{n_{\mathrm{q}} P (P-1)^2} \right]^{-1/2}. \tag{2}$$

*The distributional and population-loss bounds of Theorem 6 hold with $\varepsilon_{\mathrm{sel}}(t)$ replaced by $\varepsilon_{\mathrm{sel}}^{(n)}$.*

**Corollary 20** (Explicit sparse training and sampling requirements)**.** *Fix $\varepsilon_{\mathrm{val}} > 0$ and let $L_\delta = \log(2Pn_{\mathrm{q}}/\delta)$. Population gradient flow guarantees $\mathrm{SubOpt}_H(\widehat{\pi}_t^{\mathrm{sel}}) \le \varepsilon_{\mathrm{val}}$ whenever*

$$t \ge \frac{n_{\mathrm{q}} P(P-1)^2}{8\lambda^2 m_\lambda} \max\left\{0, \frac{\lambda^2 H^2 (H+1)^2}{4\varepsilon_{\mathrm{val}}^2} - \frac{P^2}{(P-1)^2}\right\}. \tag{3}$$

*For full-batch gradient descent, the same conclusion holds after dividing the right-hand side by $\eta_{\mathrm{gd}}$ and replacing $t$ by $n$. For empirical gradient flow, the deterministic condition $r_{\mathrm{stat}} \le 1/2$ holds, and with probability at least $1-\delta$ the conclusion $\limsup_{t\to\infty} \mathrm{SubOpt}_H(\widehat{\pi}_t^{\mathrm{emp}}) \le \varepsilon_{\mathrm{val}}$ holds, whenever*

$$N_{\mathrm{sel}} \ge \max\left\{\frac{16P^2 L_\delta}{\lambda^2 m_\lambda^2}, \frac{P^2 H^2 (H+1)^2 L_\delta}{m_\lambda^2 \varepsilon_{\mathrm{val}}^2}\right\}. \tag{4}$$

*The total number of conditional outcome samples is $2Pn_{\mathrm{q}}N_{\mathrm{sel}}$.*

### A.1 Finite conditional samples for the symmetric design

**Corollary 21** (Generative-Model Access at the Training Inputs)**.** *Under the controlled process of Proposition 17, for every relevant-coordinate map $j_\star$, every query $u = (\ell, a)$, and every $(j, s)$, there exists a supported stage-$\ell$ history $h_{u,j,s}$ satisfying $z(h_{u,j,s}) = z^{j,s}$. Consequently, independent generative-model queries at $(h_{u,j,s}, a)$ produce independent outcomes with common distribution $P_\nu(\cdot \mid u, z^{j,s})$.*

By Corollary 21, fix an integer $N_{\mathrm{sel}} \ge 1$, denoting the number of independent conditional outcome samples collected for each indexed sampling point $(u, j, s)$. The following conditional-sampling model is well defined on supported history–action pairs. For each $u \in \mathcal{Q}$, $j \in \{1, \dots, P\}$, and $s \in \{-1, +1\}$, observe mutually independent outcomes $Y_{u,j,s}^{(1)}, \dots, Y_{u,j,s}^{(N_{\mathrm{sel}})}$ drawn from $P_\nu(\cdot \mid u, z^{j,s})$, and define the empirical positive-outcome frequency

$$\widehat{q}_{u,j,s} = N_{\mathrm{sel}}^{-1} \sum_{b=1}^{N_{\mathrm{sel}}} \mathbf{1}\{Y_{u,j,s}^{(b)} = +1\}. \tag{5}$$

The empirical binary cross-entropy loss is

$$\widehat{\mathcal{L}}_{\mathrm{sel}}(B) = (2Pn_{\mathrm{q}})^{-1} \sum_{u,j,s} \mathrm{ce}_{\mathrm{bin}}(\widehat{q}_{u,j,s}, \widehat{P}_B(+1 \mid u, z^{j,s})). \tag{6}$$

Empirical gradient flow starts from $B_0 = 0$ and satisfies $\dot{B}_t = -\nabla_B \widehat{\mathcal{L}}_{\mathrm{sel}}(B_t)$.

For each pair $(u, k)$, define the symmetrized empirical target probability

$$\widehat{t}_{u,k} = \frac{1}{2}[\widehat{q}_{u,k,+1} + 1 - \widehat{q}_{u,k,-1}] \tag{7}$$

and its population counterpart

$$t_{u,k} = \frac{1}{2}[P_\nu(+1 \mid u, z^{k,+1}) + P_\nu(-1 \mid u, z^{k,-1})]. \tag{8}$$

Thus, $\widehat{t}_{u,k}$ averages the empirical frequency of outcome $+1$ on $z^{k,+1}$ and the empirical frequency of outcome $-1$ on $z^{k,-1}$. Under Assumption 4, $t_{u,k} = p_\lambda$ when $k = j_\star(u)$ and $t_{u,k} = 1 - p_\lambda$ otherwise. Fix a failure-probability parameter $\delta \in (0, 1)$ and define the uniform sampling-error radius $\varepsilon_{\mathrm{stat}} = \sqrt{\log(2Pn_{\mathrm{q}}/\delta)/(4N_{\mathrm{sel}})}$ and the resulting statistical-neighborhood radius for the irrelevant-coordinate attention mass $r_{\mathrm{stat}} = 4P\varepsilon_{\mathrm{stat}}/(\lambda m_\lambda)$.

**Theorem 22** (Finite-Sample Convergence of Attention-Based Feature Selection to a Statistical Neighborhood)**.** *Assume $r_{\mathrm{stat}} \le 1/2$. With probability at least $1-\delta$, the empirical gradient flow defined above exists for all $t \ge 0$, the relevant score remains no smaller than every irrelevant score, and, with $\varepsilon_u(t) = 1 - \alpha_{j_\star(u)}(B_t; u)$,*

$$\max_u \varepsilon_u(t) \le \max\left\{\left[\frac{P^2}{(P-1)^2} + \frac{4\lambda^2 m_\lambda t}{n_{\mathrm{q}} P^3}\right]^{-1/2}, r_{\mathrm{stat}}\right\}. \tag{9}$$

*For every query $u$ and every binary vector $z \in \{-1,1\}^P$, the total-variation and Kullback–Leibler bounds of Theorem 6 hold with $\varepsilon_{\mathrm{sel}}(t)$ replaced by $\varepsilon_u(t)$. Moreover,*

$$\mathcal{L}_{\mathrm{sel}}(B_t) - \mathcal{L}^*_{\mathrm{sel}} \le \lambda^2 (2n_{\mathrm{q}})^{-1} \sum_u \varepsilon_u(t)^2 \le \lambda^2 \max_u \varepsilon_u(t)^2/2. \tag{10}$$

Let $\widehat{\pi}_t^{\mathrm{emp}}$ denote any exact dynamic-programming policy under the learned predictive kernel $\widehat{P}_{B_t}$ obtained by empirical gradient flow.

**Proposition 23** (Lower Bound for Relevant-Coordinate Identification from Training Feature Vectors)**.** *Fix one query and let the training binary vectors $z^{(t)} \in \{-1,1\}^P$ be chosen independently of the relevant coordinate. If the population conditional probabilities distinguish all $P$ possible relevant coordinates, then $T \ge \lceil \log_2 P \rceil$.*

**Corollary 24** (Finite-Sample Value-Suboptimality Bound Beyond the Training Input Support)**.** *On the event of Theorem 22, exact planning under the learned predictive kernel obtained by empirical gradient flow satisfies*

$$\mathrm{SubOpt}_H(\widehat{\pi}_t^{\mathrm{emp}}) \le \lambda H(H+1) \max_u \varepsilon_u(t)/2. \tag{11}$$

*Its asymptotic error satisfies*

$$\limsup_{t\to\infty} \mathrm{SubOpt}_H(\widehat{\pi}_t^{\mathrm{emp}}) \le 2PH(H+1)\varepsilon_{\mathrm{stat}}/m_\lambda. \tag{12}$$

# B General Multi-Head Causal Self-attention on a Finite Reachable History–Action Set

The preceding section used a one-sparse logistic regression model and a fixed scalar value projection to isolate attention-score learning. We now allow arbitrary full-support outcome distributions and trainable value vectors. This greater generality requires the finite reachable history–action set to be covered by the training objective and leads to a near-initialization finite-width analysis.

## B.1 Shared Input Serialization

Let $\mathcal{T}$ be a finite token alphabet. A fixed serializer maps each context to $\mathrm{ser}(c_i) = (\tau_{i,1}, \dots, \tau_{i,\ell_i})$ with $\ell_i \le L$. The final token is the selected action. Define the finite index set of observed pairs consisting of a token and its absolute position

$$\Xi = \{(\tau_{i,j}, j) : 1 \le i \le m,\ 1 \le j \le \ell_i\},\ D = |\Xi|. \tag{13}$$

For $\xi \in \Xi$, let $e_\xi$ be its standard basis vector in $\mathbb{R}^D$. One fixed matrix $W \in \mathbb{R}^{d\times D}$ is shared by every context and every position. With $\xi_{i,j} = (\tau_{i,j}, j)$, define

$$x_{i,j} = W e_{\xi_{i,j}},\ X_i = (x_{i,1}, \dots, x_{i,\ell_i}),\ \max_{\xi\in\Xi} \|W e_\xi\|_2 \le R. \tag{14}$$

If $\xi_{i,j} = \xi_{i',j'}$, then $x_{i,j} = x_{i',j'}$. The embedding map $\xi \mapsto W e_\xi$ is shared across contexts and positions and has no trainable parameter indexed explicitly by the context index $i$. Nevertheless, the serialized index sequence $(\xi_{i,1}, \dots, \xi_{i,\ell_i})$ may uniquely identify a history–action pair; this dependence is represented explicitly by $Z_{\mathrm{ctx}}$. A head attends only to positions $j \le \ell_i$, so the architecture remains causal.

## B.2 One attention head

A head has a trainable score matrix $A \in \mathbb{R}^{d\times d}$ and a trainable value vector $v \in \mathbb{R}^d$. Its attention weights are

$$\alpha_{i,j}(A) = \frac{\exp\left(x_{i,\ell_i}^\top A x_{i,j}\right)}{\sum_{s=1}^{\ell_i} \exp\left(x_{i,\ell_i}^\top A x_{i,s}\right)}. \tag{15}$$

The causal self-attention summary and scalar head output are

$$s_i(A) = \sum_{j=1}^{\ell_i} \alpha_{i,j}(A) x_{i,j},\ \varphi_i(A, v) = v^\top s_i(A). \tag{16}$$

Write $\omega = (A, v)$ and use $\varphi_i(\omega) = \varphi_i(A, v)$. Equip the parameter pair with the product Euclidean norm $\|\omega\|_2^2 = \|A\|_F^2 + \|v\|_2^2$.

The directional derivative of the summary is

$$D_A s_i(A)[U] = \sum_{j=1}^{\ell_i} \alpha_{i,j}(A)\big(x_{i,j} - s_i(A)\big)\big(x_{i,\ell_i}^\top U x_{i,j}\big). \tag{17}$$

Consequently,

$$\nabla_A \varphi_i(A, v) = \sum_{j=1}^{\ell_i} \alpha_{i,j}(A)\big[v^\top (x_{i,j} - s_i(A))\big] x_{i,\ell_i} x_{i,j}^\top, \tag{18}$$

and

$$\nabla_v \varphi_i(A, v) = s_i(A). \tag{19}$$

### B.3 Paired multi-head network

Let $M$ be the number of independent head pairs and fix $\sigma_v > 0$. For each $r \in \{1, \dots, M\}$, draw $\epsilon_r \in \{-1, 1\}^d$ and $b_r \in \{-1, 1\}^K$. All coordinates of all vectors $\epsilon_r$ and $b_r$ are mutually independent Rademacher variables, and all initialization variables are independent of the episode data. Set

$$A_r^0 = 0,\ v_r^0 = \sigma_v \epsilon_r. \tag{20}$$

Thus

$$\mathbb{E}[v_r^0 (v_r^0)^\top] = \sigma_v^2 \boldsymbol{I}_d,\ \left\|v_r^0\right\|_2 = C_{\text{init}} := \sigma_v \sqrt{d}. \tag{21}$$

Set $\omega_r^0 = (A_r^0, v_r^0)$. For $s \in \{+, -\}$, write $\omega_{r,s} = (A_{r,s}, v_{r,s})$. After sampling, the output-layer weight vector $b_r$ remains fixed. The two trainable copies start from the same parameters,

$$\omega_{r,+}(0) = \omega_{r,-}(0) = \omega_r^0. \tag{22}$$

The logits at context $c_i$ are

$$f_i(\theta) = \frac{1}{\sqrt{2M}} \sum_{r=1}^{M} b_r \left[\varphi_i(\omega_{r,+}) - \varphi_i(\omega_{r,-})\right]. \tag{23}$$

Thus $f_i(\theta_0) = 0$ exactly. All trainable parameters are attention score matrices and trainable scalar value-projection vectors. The fixed output-layer weight vectors map scalar head outputs to token logits.

Training follows population gradient flow,

$$\dot{\theta}_t = -\nabla_\theta \mathcal{L}(f(\theta_t)). \tag{24}$$

The proof isolates the tangent contribution of the score matrices $A_{r,+}$ and $A_{r,-}$ in this architecture. The same lower bound also survives an independent additive parameter branch that leaves these score-parameter Jacobians unchanged, because the new branch contributes another positive semidefinite tangent term. This observation does not cover inserting normalization, feedforward blocks, residual transformations, or additional attention layers into the same computation path. Such changes alter the score Jacobian itself and require a separate analysis.

# C JACOBIAN WITH RESPECT TO THE ATTENTION SCORE MATRICES AND NEURAL TANGENT KERNEL LOWER BOUND

## C.1 CONTEXT DESIGN MATRIX AT $A = 0$

For context $c_i$, define the last-token query, the arithmetic mean of the embeddings, and the within-context covariance by

$$u_i = x_{i,\ell_i}, \bar{x}_i = \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} x_{i,j}, \tag{25}$$

$$C_i = \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^\top. \tag{26}$$

Define

$$Z_{\mathrm{ctx}}(W) \in \mathbb{R}^{m \times d^3}, [Z_{\mathrm{ctx}}(W)]_{i,:} = \big(u_i \otimes \mathrm{vec}(C_i)\big)^\top. \tag{27}$$

Let $S = \mathrm{col}(W)$ and $r = \mathrm{rank}(W)$. Every $u_i$ belongs to $S$, and every $C_i$ is symmetric with range contained in $S$.

Let $\mathbb{S}^r = \{M \in \mathbb{R}^{r \times r} : M^\top = M\}$ denote the vector space of real symmetric $r \times r$ matrices.

**Proposition 25** (Rank Bound for the Context Design Matrix)**.** *The context design matrix satisfies*

$$\mathrm{rank}(Z_{\mathrm{ctx}}(W)) \le r \ \dim(\mathbb{S}^r) = \tfrac{r^2(r+1)}{2}. \tag{28}$$

*Consequently, full row rank requires*

$$m \le \frac{r^2(r+1)}{2} \le \frac{\min\{d, D\}^2(\min\{d, D\} + 1)}{2}. \tag{29}$$

**Assumption 26** (Full Row Rank of the Shared Context Design Matrix)**.** *For the fixed shared embedding matrix $W$, there is a constant $\zeta > 0$ such that*

$$\sigma_{\min}(Z_{\mathrm{ctx}}(W)) \ge \zeta. \tag{30}$$

The condition is computable from the serializer and the shared embedding matrix. It introduces no trainable parameter indexed explicitly by a history, although different contexts may have different serialized index sequences $(\xi_{i,1}, \ldots, \xi_{i,\ell_i})$. The condition depends only on the representation at initialization and not on the neural tangent kernel along the training trajectory.

## C.2 EXISTENCE AND GENERIC FULL ROW RANK

Use the same serializer with the one-hot shared embedding $x_{i,j}^{\mathrm{oh}} = e_{\xi_{i,j}} \in \mathbb{R}^D$. Let $u_i^{\mathrm{oh}}$, $C_i^{\mathrm{oh}}$, and $Z_{\mathrm{oh}} \in \mathbb{R}^{m \times D^3}$ be defined by the preceding formulas with these one-hot vectors.

**Assumption 27** (Full Row Rank of the One-Hot Context Design Matrix)**.** *The one-hot context design matrix satisfies* $\mathrm{rank}(Z_{\mathrm{oh}}) = m$.

**Proposition 28** (Existence and Generic Full Row Rank for Shared Embeddings)**.** *Suppose Assumption 27 holds and $d \ge D$. Then a shared embedding matrix $W_0 \in \mathbb{R}^{d \times D}$ exists for which $Z_{\mathrm{ctx}}(W_0)$ has full row rank. More generally, fix any embedding dimension $d$. If at least one matrix $W_0 \in \mathbb{R}^{d \times D}$ makes $Z_{\mathrm{ctx}}(W_0)$ full row rank, then the set $\{W \in \mathbb{R}^{d \times D} : \mathrm{rank}(Z_{\mathrm{ctx}}(W)) < m\}$ has Lebesgue measure zero. Any distribution on $W$ that is absolutely continuous with respect to Lebesgue measure therefore gives full row rank almost surely.*

The genericity statement alone does not quantify $\zeta$. The following sufficient condition, based on a triple of indices in $\Xi$ that occurs in only one context, gives an explicit witness for the padded one-hot embedding and nearby context matrices. It is not required by the main convergence theorems, which assume only the singular-value condition in Assumption 26.

**Corollary 29** (A Sufficient Condition for a Singular-Value Lower Bound)**.** *Assume every context $c_i$ has two distinct positions $p_i$ and $q_i$ such that the triple $(\xi_{i,\ell_i}, \xi_{i,p_i}, \xi_{i,q_i})$ does not occur in any other*

*context. Here occurrence in another context means that its final index equals $\xi_{i,\ell_i}$ and its serialized index sequence contains both $\xi_{i,p_i}$ and $\xi_{i,q_i}$. Then*

$$\sigma_{\min}(Z_{\mathrm{oh}}) \geq \min_{1\leq i\leq m} \ell_i^{-2} \geq L^{-2}. \tag{31}$$

*In particular, Assumption 27 holds. If $d \geq D$ and $W_0$ is the padded one-hot embedding used in Proposition 28, then*

$$\sigma_{\min}(Z_{\mathrm{ctx}}(W_0)) = \sigma_{\min}(Z_{\mathrm{oh}}) \geq L^{-2}. \tag{32}$$

*For this witness, Assumption 26 holds with $\zeta = L^{-2}$, and Proposition 30 permits $\gamma = \sigma_v^2 L^{-4}$. Moreover, if*

$$\|Z_{\mathrm{ctx}}(W) - Z_{\mathrm{ctx}}(W_0)\|_{\mathrm{op}} \leq \frac{1}{2L^2}, \tag{33}$$

*then $\sigma_{\min}(Z_{\mathrm{ctx}}(W)) \geq 1/(2L^2)$, so one may take $\gamma = \sigma_v^2/(4L^4)$. By continuity of $W \mapsto Z_{\mathrm{ctx}}(W)$, this bound holds on an open neighborhood of $W_0$.*

Corollary 29 gives an explicit sufficient condition showing that Assumption 26 is satisfiable; it is not a structural assumption of Theorems 32–35. Those theorems apply whenever $Z_{\mathrm{ctx}}$ has the stated singular-value margin, whether or not any context has a triple $(\xi_{i,\ell_i}, \xi_{i,p_i}, \xi_{i,q_i})$ that occurs in no other context.

For one head, define the score-parameter Jacobian matrix

$$\boldsymbol{\Phi}_A(A, v) \in \mathbb{R}^{m\times d^2}, \quad [\boldsymbol{\Phi}_A(A, v)]_{i,:} = \mathrm{vec}\left(\nabla_A \varphi_i(A, v)\right)^\top. \tag{34}$$

We also write $\boldsymbol{\Phi}_A(\omega) = \boldsymbol{\Phi}_A(A, v)$ when $\omega = (A, v)$. At initialization, let

$$\boldsymbol{G}_A = \mathbb{E}_{v^0}\left[\boldsymbol{\Phi}_A(0, v^0)\boldsymbol{\Phi}_A(0, v^0)^\top\right]. \tag{35}$$

**Proposition 30** (Jacobian Gram Matrix for the Attention Score Matrices)**.** *Under the initialization in equation* (20)*,*

$$\nabla_A \varphi_i(0, v) = u_i(C_i v)^\top, \tag{36}$$

*and*

$$\boldsymbol{G}_A = \sigma_v^2 Z_{\mathrm{ctx}} Z_{\mathrm{ctx}}^\top. \tag{37}$$

*Under Assumption 26,*

$$\boldsymbol{G}_A \succeq \gamma \boldsymbol{I}_m, \gamma := \sigma_v^2 \zeta^2. \tag{38}$$

The lower bound is therefore obtained from the last-token queries, the within-context token covariances, and the variance of the trainable value-vector initialization. No fixed feature branch supplies context-dependent logits.

### C.3 Derivative bounds

The following bounds are used to control neural tangent kernel drift. They depend only on the embedding radius and a local bound on the value vectors. For a bilinear map $T : \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \to \mathbb{R}^d$, define $\|T\|_{\mathrm{op}} = \sup_{\|U\|_F\leq 1, \|V\|_F\leq 1} \|T[U, V]\|_2$.

**Lemma 31** (Smoothness of one attention head)**.** *For every context $c_i$ and score matrices $A, A'$, the attention summary satisfies*

$$\|s_i(A)\|_2 \leq R, \tag{39}$$

$$\|s_i(A) - s_i(A')\|_2 \leq L_1 \|A - A'\|_F, L_1 = 2R^3. \tag{40}$$

*The second derivative satisfies*

$$\left\|D_A^2 s_i(A)\right\|_{\mathrm{op}} \leq L_2, L_2 = 6R^5. \tag{41}$$

*Fix a radius $r \leq 1$. On the set $\|v\|_2 \leq C_{\mathrm{init}} + r$, define*

$$C_{\text{score}} = L_1(C_{\text{init}} + r),\ C_{\text{head}} = \sqrt{R^2 + C_{\text{score}}^2}, \tag{42}$$

$$L_{\text{score}} = L_1 + (C_{\text{init}} + r)L_2. \tag{43}$$

*Then*

$$\|\nabla_\omega \varphi_i(\omega)\|_2 \le C_{\text{head}}, \tag{44}$$

$$\|\nabla_A \varphi_i(\omega)\|_F \le C_{\text{score}}, \tag{45}$$

$$\|\text{vec}(\nabla_A \varphi_i(\omega)) - \text{vec}(\nabla_A \varphi_i(\omega'))\|_2 \le L_{\text{score}} \|\omega - \omega'\|_2 \tag{46}$$

*whenever $\omega$ and $\omega'$ lie in that set.*

### C.4 Initial Neural Tangent Kernel Concentration

Let $\mathbf{\Theta}_A(0)$ be the neural tangent operator contributed only by the score matrices $A_{r,+}$ and $A_{r,-}$ at initialization. Because the two copies enter the logits with opposite signs and have identical parameters,

$$\mathbf{\Theta}_A(0) = \frac{1}{M}\sum_{r=1}^{M} \left[\mathbf{\Phi}_A(\omega_r^0)\mathbf{\Phi}_A(\omega_r^0)^\top\right] \otimes \left(b_r b_r^\top\right). \tag{47}$$

Since $\mathbb{E}[b_r b_r^\top] = \boldsymbol{I}_K$,

$$\mathbb{E}[\mathbf{\Theta}_A(0)] = \boldsymbol{G}_A \otimes \boldsymbol{I}_K. \tag{48}$$

**Theorem 32** (Initial Lower Bound on the Neural Tangent Kernel Contribution from the Attention Score Matrices)**.** *Assume Assumption 26, set $\gamma = \sigma_v^2\zeta^2$ as in Proposition 30, and use the bounds of Lemma 31 with $r = 1$. If*

$$M \ge \frac{8KmC_{\text{score}}^2}{\gamma}\log\left(\frac{mK}{\delta}\right), \tag{49}$$

*then, with probability at least $1-\delta$ over the random initialization $\{(\epsilon_r, b_r)\}_{r=1}^M$,*

$$\mathbf{\Theta}_A(0) \succeq \frac{\gamma}{2}\boldsymbol{I}_{mK}. \tag{50}$$

The proof is a direct matrix Chernoff argument (Tropp, 2012). Each summand in equation (47) is positive semidefinite and has operator norm at most $KmC_{\text{score}}^2$.

### C.5 Centered Logits and the Full Neural Tangent Kernel Matrix

Softmax is invariant to adding a constant to all logits at one context. Define

$$P_K = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \boldsymbol{P} = \boldsymbol{I}_m \otimes P_K. \tag{51}$$

The gradient of cross-entropy belongs to the range of $\boldsymbol{P}$. Let $\mathbf{\Theta}(t)$ denote the full neural tangent kernel matrix for all trainable score and value parameters. Since neural tangent kernel contributions are positive semidefinite,

$$\mathbf{\Theta}(t) \succeq \mathbf{\Theta}_A(t). \tag{52}$$

The main result of the next section proves a lower bound for $\mathbf{\Theta}_A(t)$ for all $t \ge 0$.

## D Uniform Neural Tangent Kernel Lower Bound and Loss Convergence

### D.1 Cross-entropy geometry on a loss sublevel set

Let

$$E(t) = \mathcal{L}(f(\theta_t)) - \mathcal{L}^*. \tag{53}$$

Because $f(\theta_0) = 0$,

$$\mathcal{L}(f(\theta_0)) = \log K,\ E_0 = \log K - \sum_{i=1}^{m} d_i \operatorname{Ent}(q_i). \tag{54}$$

Gradient flow makes $\mathcal{L}$ nonincreasing. For every context and outcome,

$$\operatorname{softmax}(f_i(t))_y \geq p_* = \exp\left(-\frac{\log K}{\eta d_{\min}}\right). \tag{55}$$

Indeed, each weighted context loss is at most $\log K$, and each target probability is at least $\eta$.

**Lemma 33** (Strong convexity on centered logits)**.** *On the loss sublevel set $\mathcal{L}(f) \leq \log K$, the population loss is strongly convex on the range of $\boldsymbol{P}$ with constant*

$$\mu = d_{\min} K p_*^2. \tag{56}$$

*Consequently,*

$$\|\nabla_f \mathcal{L}(f)\|_2^2 \geq 2\mu\big(\mathcal{L}(f) - \mathcal{L}^*\big). \tag{57}$$

The bound follows from the softmax Hessian identity

$$u^\top\big(\operatorname{diag}(p) - pp^\top\big)u = \frac{1}{2}\sum_{a,b=1}^{K} p_a p_b (u_a - u_b)^2 \tag{58}$$

for $u^\top \mathbf{1} = 0$.

### D.2 Functional gradient-flow equation

Let $J(\theta)$ be the Jacobian of the stacked logits with respect to all trainable parameters. Then

$$\boldsymbol{\Theta}(\theta) = J(\theta)J(\theta)^\top. \tag{59}$$

The logit trajectory satisfies the exact identity

$$\dot{f}_t = -\boldsymbol{\Theta}(\theta_t)\nabla_f \mathcal{L}(f_t). \tag{60}$$

Therefore,

$$\dot{E}(t) = -\nabla_f \mathcal{L}(f_t)^\top \boldsymbol{\Theta}(\theta_t)\nabla_f \mathcal{L}(f_t). \tag{61}$$

### D.3 Bootstrap for the trainable attention scores

Set

$$c = \frac{\gamma\mu}{2} = \frac{\sigma_v^2 \zeta^2 d_{\min} K p_*^2}{2}. \tag{62}$$

Using the constants from Lemma 31 with radius $r = 1$, define

$$r_M = \frac{2\sqrt{K} C_{\text{head}} \sqrt{E_0}}{c\sqrt{M}} \tag{63}$$

and

$$r_* = \min\left\{1, \frac{\gamma}{8KmC_{\text{score}}L_{\text{score}}}\right\}. \tag{64}$$

**Theorem 34** (Uniform Lower Bound on the Neural Tangent Kernel Contribution from the Attention Score Matrices)**.** *Assume Assumption 26 and set $\gamma = \sigma_v^2\zeta^2$. Suppose equation* (49) *holds and*

$$r_M \leq r_*/2. \tag{65}$$

*Then, with probability at least $1 - \delta$, the gradient-flow solution exists for every $t \geq 0$ and satisfies*

$$\boldsymbol{P}\boldsymbol{\Theta}_A(t)\boldsymbol{P} \succeq \frac{\gamma}{4}\boldsymbol{P} \tag{66}$$

*for every $t \geq 0$. Moreover, for every pair $r$ and sign $s \in \{+, -\}$,*

$$\sup_{t \geq 0} \left\| \omega_{r,s}(t) - \omega_r^0 \right\|_2 \leq r_M. \tag{67}$$

Theorem 34 is the architecture-dependent step. The lower bound on the neural tangent kernel contribution from the attention score matrices is a conclusion derived from the corresponding Jacobian and finite width. The proof does not assume a uniform-in-time lower bound on the neural tangent kernel.

**Theorem 35** (Exponential Decay of the Excess Population Loss)**.** *Under the conditions of Theorem 34,*

$$E(t) \leq E_0 e^{-ct} \tag{68}$$

*for every $t \geq 0$, where $c = \gamma\mu/2$. In addition,*

$$\int_0^\infty \left\| \nabla_f \mathcal{L}(f_t) \right\|_2 \, \mathrm{d}t \leq \frac{2\sqrt{2E_0}}{c}. \tag{69}$$

The result is finite-width. It does not first pass to an infinite-width mean-field equation. Width is used to concentrate the neural tangent kernel contribution from the attention score matrices and to control its drift. Because the targets $q_i$ are stochastic, the limiting objective value is $\mathcal{L}^* = \sum_i d_i \operatorname{Ent}(q_i)$; the theorem proves convergence of conditional distributions in population cross-entropy, not zero-loss interpolation of deterministic labels.

# E Model-Based Planning and Bayes Value Suboptimality

The simulation argument in this section is used by both optimization regimes. Corollaries 8 and 24 apply it to population and empirical prediction error on unseen binary feature vectors, while Theorem 38 applies it to the general environment model on the finite reachable history–action set.

The Bayes predictive kernel $P_\nu$ is supported on the history tree by construction. In the one-sparse logistic regression setting, both $P_\nu(\cdot \mid h, a)$ and $\widehat{P}_B(\cdot \mid h, a)$ are supported on $\{y_u^+, y_u^-\}$. In the finite reachable-set setting, the full-support condition $q_i(y) \geq \eta$ places every outcome extension in the next-stage history set, and the softmax predictor is therefore also supported on the same history tree.

## E.1 Bayes Control by Dynamic Programming on Histories

Let $\mathsf{P}$ be any history-dependent stochastic kernel on $\mathcal{C}$ that is supported on the history tree, and use the initial observation distribution $P_\nu^0$. For a history policy $\pi$, set $V_{\mathsf{P},H+1}^\pi(h) = 0$ and, for $k = H, \ldots, 1$, define

$$V_{\mathsf{P},k}^\pi(h) = \sum_{a \in \mathcal{A}} \pi(a \mid h) \sum_{y=(r,o') \in \mathcal{Y}} \mathsf{P}(y \mid h, a) \left[ r + V_{\mathsf{P},k+1}^\pi(h, a, y) \right]. \tag{70}$$

The root value and optimal value are

$$V_{\mathsf{P}}^\pi := \sum_{(o) \in \mathcal{H}_1} P_\nu^0(o) V_{\mathsf{P},1}^\pi((o)), \; V_{\mathsf{P}}^* := \max_\pi V_{\mathsf{P}}^\pi. \tag{71}$$

Here $V_{\mathsf{P},k+1}^\pi(h, a, y)$ denotes the continuation value at the history obtained by appending action $a$ and outcome $y$ to $h$.

**Proposition 36** (Equivalence of Bayes control and dynamic programming under $P_\nu$)**.** *For every task-independent history policy $\pi$,*

$$V_\nu^\pi = V_{P_\nu}^\pi. \tag{72}$$

*Consequently,*

$$V_\nu^* = V_{P_\nu}^*, \tag{73}$$

*and every policy obtained by exact backward induction under $P_\nu$ is Bayes-optimal for the original prior over tasks.*

This equivalence is exact. It does not approximate the task posterior by a Markov state or assume that the latent task is identifiable.

### E.2 Learned Predictive Kernel

At training time $t$, define, for every $(h_i, a_i) \in \mathcal{C}$,

$$\widehat{P}_t(y \mid h_i, a_i) = \mathrm{softmax}(f_i(\theta_t))_y. \tag{74}$$

Since the population excess loss is an average conditional KL divergence,

$$E(t) = \sum_{i=1}^{m} d_i D_{\mathrm{KL}}\big(q_i \,\|\, \widehat{P}_t(\cdot \mid c_i)\big). \tag{75}$$

Pinsker's inequality and the convention $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$ give

$$\epsilon_t := \max_{1 \le i \le m} \mathrm{TV}\left(q_i, \widehat{P}_t(\cdot \mid c_i)\right) \le \sqrt{\frac{E(t)}{2d_{\min}}}. \tag{76}$$

### E.3 Exact planning in the learned predictive kernel

Let $\widehat{V}_{H+1,t}(h) = 0$. For $k = H, \ldots, 1$, define

$$\widehat{Q}_{k,t}(h, a) = \sum_{y=(r,o')\in\mathcal{Y}} \widehat{P}_t(y \mid h, a)\left[r + \widehat{V}_{k+1,t}(h, a, y)\right], \tag{77}$$

$$\widehat{V}_{k,t}(h) = \max_{a\in\mathcal{A}} \widehat{Q}_{k,t}(h, a). \tag{78}$$

Choose any maximizing policy

$$\widehat{\pi}_t(h) \in \arg\max_{a\in\mathcal{A}} \widehat{Q}_{k,t}(h, a). \tag{79}$$

The condition $q_i(y) \ge \eta$ ensures that every outcome under the learned predictive kernel leads to a history in the next-stage reachable history set. The planner therefore never queries $\widehat{P}_t$ outside $\mathcal{C}$. This planner is external to the causal self-attention model. The theorem concerns world-model learning followed by exact planning, not the formation of an internal planning algorithm.

**Lemma 37** (Finite-Horizon Simulation Bound)**.** *Let $\mathsf{P}$ and $\widehat{P}$ be two history-dependent stochastic kernels supported on the same history tree, use the same initial distribution $P^0_\nu$, have rewards in $[0, 1]$, and satisfy*

$$\max_{(h,a)\in\mathcal{C}} \mathrm{TV}\left(\mathsf{P}(\cdot \mid h, a), \widehat{P}(\cdot \mid h, a)\right) \le \epsilon. \tag{80}$$

*For every history policy $\pi$,*

$$\left|V^\pi_{\mathsf{P}} - V^\pi_{\widehat{P}}\right| \le \frac{H(H+1)}{2}\epsilon. \tag{81}$$

*If $\widehat{\pi}$ is optimal under $\widehat{P}$, then*

$$V^*_{\mathsf{P}} - V^{\widehat{\pi}}_{\mathsf{P}} \le H(H+1)\epsilon. \tag{82}$$

**Theorem 38** (Causal self-attention training to Bayes value suboptimality)**.** *Under the conditions of Theorem 34, with probability at least $1 - \delta$ over the network initialization, the policy obtained by exact planning in $\widehat{P}_t$ satisfies*

$$\mathrm{SubOpt}_H(\widehat{\pi}_t) := V^*_\nu - V^{\widehat{\pi}_t}_\nu \le H(H+1)\sqrt{\frac{E_0}{2d_{\min}}}e^{-ct/2}. \tag{83}$$

**Corollary 39** (Population training time for a prescribed value error)**.** *Under the conditions of Theorem 38, fix $\varepsilon_{\mathrm{val}} > 0$. The inequality $\mathrm{SubOpt}_H(\widehat{\pi}_t) \le \varepsilon_{\mathrm{val}}$ holds whenever*

$$t \geq \frac{2}{c} \max \left\{ 0, \log \left( \frac{H(H+1)}{\varepsilon_{\mathrm{val}}} \sqrt{\frac{E_0}{2d_{\min}}} \right) \right\}. \tag{84}$$

Theorem 38 is not an imitation-learning statement and its left-hand side is a value gap for one planned policy, not cumulative regret. Proposition 36 identifies this value gap with the corresponding gap in the Bayes predictive history process. The targets $q_i$ are transition-and-reward conditionals under that process. The optimal policy is computed after training from the learned model, and its value is evaluated on the fresh test episode defined in Section 2.

# F Proofs for the One-Sparse Logistic Regression Model, Finite Samples, and Unseen Binary Feature Vectors

*Proof of Proposition 18.* Let $d_0 = n_{\mathrm{q}} + P + 1$. For feature coordinate $j$ and sign $z_j$, use the fixed token vector $x_j(z_j) = (\mathbf{0}_{n_{\mathrm{q}}}, e_j, z_j) \in \mathbb{R}^{d_0}$. For query $u$, use the fixed final-token vector $x_{\mathrm{q}}(u) = (e_u, \mathbf{0}_P, 0) \in \mathbb{R}^{d_0}$. Apply a strict causal mask under which the final query token attends only to the preceding $P$ feature tokens.

Let $A_B \in \mathbb{R}^{d_0 \times d_0}$ be zero outside its query-by-feature block, set that block equal to $B$, and keep all remaining entries fixed at zero. Let $v = (\mathbf{0}_{n_{\mathrm{q}}}, \mathbf{0}_P, 1)$. Then $x_{\mathrm{q}}(u)^\top A_B x_j(z_j) = B_{u,j}$ and $v^\top x_j(z_j) = z_j$. The attention weights are therefore $\alpha_j(B; u)$, and the scalar attention output is $\sum_j \alpha_j(B; u) z_j = r_B(u, z)$. We use the unnormalized bilinear score above. If a fixed nonzero scalar score factor $\kappa$ is retained in the raw parameterization, the effective-score dynamics are accelerated by the factor $\kappa^2$. Equivalently, the resulting effective-score trajectory is obtained by replacing $t$ with $\kappa^2 t$ under gradient flow, while a raw gradient-descent step size $\eta_{\mathrm{raw}}$ induces the effective step size $\kappa^2 \eta_{\mathrm{raw}}$.

The linear map $B \mapsto A_B$ inserts $B$ into the only trainable block and preserves Frobenius inner products. Hence the pullback of the Euclidean gradient with respect to the trainable entries of $A_B$ is exactly the Euclidean gradient with respect to $B$. Therefore gradient flow and full-batch gradient descent in this constrained attention-score parameterization induce exactly the score trajectories analyzed in the sparse model. □

*Proof of Proposition 17.* Let $\mathcal{Z}_{\mathrm{des}} = \{z^{j,s} : 1 \leq j \leq P,\ s \in \{-1, +1\}\}$ be the set of distinct binary feature vectors. For every index $(j, s)$, introduce a distinct initial observation $o_{j,s}$. For every binary feature vector $z \in \{-1, +1\}^P \setminus \mathcal{Z}_{\mathrm{des}}$, introduce one additional initial observation $o_z$, and introduce one common continuation observation $o_{\mathrm{c}}$. Let

$$\mathcal{O}_{\mathrm{init}} = \{o_{j,s} : 1 \leq j \leq P, s \in \{-1,+1\}\} \cup \{o_z : z \in \{-1,+1\}^P \setminus \mathcal{Z}_{\mathrm{des}}\}, \mathcal{O} = \mathcal{O}_{\mathrm{init}} \cup \{o_{\mathrm{c}}\}. \tag{85}$$

Define the feature vector from the initial observation and keep it fixed throughout the episode. If the first observation is $o_{j,s}$, set $z(h_\ell) = z^{j,s}$. If it is $o_z$, set $z(h_\ell) = z$. The serializer exposes the $P$ feature values and the stage–action query, but it does not expose the raw identity of the first observation. Thus two raw initial observations with the same feature vector have the same serialized input.

Let the initial observation be uniform on $\mathcal{O}_{\mathrm{init}}$, let the reward alphabet be $\{0, 1\}$, and use a single task with prior mass one. Let $y^+ = (1, o_{\mathrm{c}})$ and $y^- = (0, o_{\mathrm{c}})$, and set $y_u^+ = y^+$ and $y_u^- = y^-$ for every $u \in \mathcal{Q}$.

Fix a relevant-coordinate map $j_\star$. At stage $\ell$, given history $h_\ell$ and action $a$, draw $Y_\ell \in \{-1, +1\}$ according to $\Pr(Y_\ell = +1 \mid h_\ell, a) = \mathrm{sigmoid}\left(\lambda z_{j_\star(\ell,a)}(h_\ell)\right)$. The process emits $y^+$ when $Y_\ell = +1$ and emits $y^-$ when $Y_\ell = -1$. Because both conditional outcome probabilities lie in $(0, 1)$, the set of supported raw histories is independent of $j_\star$.

Let the behavior policy be uniform on $\mathcal{A}$ at every history. Retain an exploratory episode exactly when its initial observation is one of the indexed initial observations $o_{j,s}$. This retention rule is independent of $j_\star$. Conditional on retention, the index $(j, s)$ is uniform on its $2P$ possible values. After retention, select one of the $H$ stages uniformly at random. Then, for every $u = (\ell, a)$ and every index $(j, s)$,

$$\Pr\big(u=(\ell,a),\ (j,s)\mid \text{retained}\big)=\frac{1}{H}\frac{1}{|\mathcal{A}|}\frac{1}{2P}=\frac{1}{2Pn_{\mathrm{q}}}. \tag{86}$$

The previous actions and outcomes may vary across retained episodes, but all histories with the same $u$ and $(j,s)$ have the same serialized input $(u,z^{j,s})$ and the same conditional outcome distribution $P_\nu(\cdot\mid u,z^{j,s})$. Hence the retained examples have exactly the training distribution specified in Assumption 5. □

*Proof of Proposition 23.* For coordinate $j$, collect its values across the fixed training binary feature vectors in the binary column pattern $c_j=(z_j^{(1)},\dots,z_j^{(T)})\in\{-1,1\}^T$. If $c_j=c_k$ for two coordinates $j\neq k$, then Assumption 4 gives the same population conditional probabilities on every training binary feature vector when coordinate $j$ is relevant and when coordinate $k$ is relevant. The two candidate relevant coordinates are therefore not distinguishable from these training binary feature vectors. Distinguishing all $P$ candidate relevant coordinates requires $P$ distinct binary column patterns, while $T$ binary entries admit at most $2^T$ binary patterns. Hence $P\le 2^T$ and $T\ge\lceil\log_2 P\rceil$. □

*Proof of Theorem 6.* The loss separates across the rows of $B$. Fix one query $u$ and write $j_\star=j_\star(u)$. For the two sign-paired inputs associated with candidate $k$, the attention summaries are $r_B(u,z^{k,+1})=2\alpha_k-1$ and $r_B(u,z^{k,-1})=1-2\alpha_k$. For candidate $k$, the two sign contexts have predictions $p_k=\mathrm{sigmoid}(\lambda(2\alpha_k-1))$ and $1-p_k$. Their averaged target is $p_\lambda=\mathrm{sigmoid}(\lambda)$ when $k=j_\star$ and $1-p_\lambda$ when $k\neq j_\star$. The one-row loss is therefore $\ell(\alpha)=P^{-1}[\mathrm{ce}_{\mathrm{bin}}(p_\lambda,p_{j_\star})+\sum_{k\neq j_\star}\mathrm{ce}_{\mathrm{bin}}(1-p_\lambda,p_k)]$.

At initialization every score in the row is zero. The row loss is invariant under every permutation of the irrelevant coordinates. Uniqueness of the gradient-flow solution therefore preserves one relevant score $a(t)$ and one common irrelevant score $b(t)$. Every query has the same row loss after relabeling its relevant coordinate, so the same functions apply to all rows. The vector field is smooth and bounded, which gives a unique global solution.

Let $g=a-b$, $\alpha=e^g/(e^g+P-1)$, $\varepsilon=1-\alpha$, and $w=\varepsilon/(P-1)$. Along the symmetric trajectory, the row loss is $\ell_{\mathrm{sym}}(\alpha)=P^{-1}[\mathrm{ce}_{\mathrm{bin}}(p_\lambda,\mathrm{sigmoid}(\lambda(2\alpha-1)))+(P-1)\,\mathrm{ce}_{\mathrm{bin}}(1-p_\lambda,\mathrm{sigmoid}(\lambda(2w-1)))]$. Differentiation gives $-\ell'_{\mathrm{sym}}(\alpha)=2\lambda P^{-1}[(p_\lambda-\mathrm{sigmoid}(\lambda(2\alpha-1)))+(\mathrm{sigmoid}(\lambda(2w-1))-(1-p_\lambda))]$.

The arguments of every sigmoid in the preceding expression lie in $[-\lambda,\lambda]$, and the sigmoid derivative on this interval is at least $m_\lambda$. The mean-value theorem gives $p_\lambda-\mathrm{sigmoid}(\lambda(2\alpha-1))\ge 2\lambda m_\lambda\varepsilon$ and $\mathrm{sigmoid}(\lambda(2w-1))-(1-p_\lambda)\ge 2\lambda m_\lambda w$. Hence $-\ell'_{\mathrm{sym}}(\alpha)\ge 4\lambda^2 m_\lambda\varepsilon/(P-1)$.

The overall loss is the average of the $n_{\mathrm{q}}$ row losses. Along the symmetric trajectory, changing the relevant score gives $\partial\alpha/\partial a=\alpha\varepsilon$. By symmetry, the gradient with respect to each common irrelevant score is $-(P-1)^{-1}\alpha\varepsilon\ell'_{\mathrm{sym}}(\alpha)$. Euclidean gradient flow in the $P$ row scores gives $\dot g=-P\alpha\varepsilon\ell'_{\mathrm{sym}}(\alpha)/(n_{\mathrm{q}}(P-1))$. Thus $g$ is strictly increasing at every finite point. Since $\dot\varepsilon=-\alpha\varepsilon\dot g$ and $\alpha(t)\ge 1/P$, the preceding derivative bound yields $\dot\varepsilon\le -4\lambda^2 m_\lambda\varepsilon^3/(n_{\mathrm{q}}P(P-1)^2)$. Because $\varepsilon(0)=(P-1)/P$, integration gives $\varepsilon(t)^{-2}\ge P^2/(P-1)^2+8\lambda^2 m_\lambda t/(n_{\mathrm{q}}P(P-1)^2)$. This proves the stated rate. Its right-hand side diverges, so $\varepsilon(t)\to 0$, $\alpha(t)\to 1$, and $g(t)\to\infty$.

Now fix any query $u$ and binary vector $z\in\{-1,1\}^P$, and write $z_\star=z_{j_\star(u)}$. The total attention mass on irrelevant coordinates is $\varepsilon_{\mathrm{sel}}(t)$, so $|r_{B_t}(u,z)-z_\star|=|\sum_{j\neq j_\star(u)}\alpha_j(B_t;u)(z_j-z_\star)|\le 2\varepsilon_{\mathrm{sel}}(t)$. The sigmoid is $1/4$-Lipschitz, which proves the total-variation bound.

For the Kullback–Leibler bound, let $A_{\log}(x)=\log(1+e^x)$. Bernoulli Kullback–Leibler divergence in natural parameters is the Bregman divergence generated by $A_{\log}$, and $A''_{\log}(x)\le 1/4$. The natural-parameter difference is at most $2\lambda\varepsilon_{\mathrm{sel}}(t)$, so the divergence is at most $\lambda^2\varepsilon_{\mathrm{sel}}(t)^2/2$. Since $\mathcal{L}_{\mathrm{sel}}(B_t)-\mathcal{L}^*_{\mathrm{sel}}$ is the uniform average of the corresponding conditional Kullback–Leibler divergences over all $(u,j,s)$, the same bound yields $\mathcal{L}_{\mathrm{sel}}(B_t)-\mathcal{L}^*_{\mathrm{sel}}\le\lambda^2\varepsilon_{\mathrm{sel}}(t)^2/2$. □

*Proof of Theorem 19.* The population loss separates across rows and is invariant under permutations of the irrelevant coordinates. Starting from zero, full-batch gradient descent therefore preserves one relevant score $a_n$ and one common irrelevant score $b_n$ in every row. Use the notation $g_n$, $\alpha_n$, and $\varepsilon_n$

for one row, and let $\ell_{\mathrm{sym}}$ be the symmetric one-row loss from the proof of Theorem 6. The exact update is

$$g_{n+1} - g_n = -\frac{\eta_{\mathrm{gd}} P \alpha_n \varepsilon_n}{n_{\mathrm{q}}(P-1)} \ell'_{\mathrm{sym}}(\alpha_n). \tag{87}$$

The derivative bound in the population proof gives $-\ell'_{\mathrm{sym}}(\alpha_n) \geq 4\lambda^2 m_\lambda \varepsilon_n/(P-1)$. Hence the increment $\Delta_n = g_{n+1} - g_n$ is positive and satisfies $\Delta_n \geq 4\eta_{\mathrm{gd}}\lambda^2 m_\lambda P \alpha_n \varepsilon_n^2/[n_{\mathrm{q}}(P-1)^2]$.

Since $\varepsilon(g) = (P-1)/(e^g + P - 1)$, the exact relation between two successive iterates is $\varepsilon_{n+1} = \varepsilon_n/[\varepsilon_n + \alpha_n e^{\Delta_n}]$. Using $e^x \geq 1 + x$ gives $\varepsilon_{n+1} \leq \varepsilon_n/(1 + \alpha_n \Delta_n)$. Therefore

$$\varepsilon_{n+1}^{-2} - \varepsilon_n^{-2} \geq 2\alpha_n \Delta_n \varepsilon_n^{-2} \geq \frac{8\eta_{\mathrm{gd}}\lambda^2 m_\lambda}{n_{\mathrm{q}} P (P-1)^2}, \tag{88}$$

where the last step uses $\alpha_n \geq 1/P$. Summation from 0 to $n-1$ and $\varepsilon_0 = (P-1)/P$ proves equation (2). The right-hand side diverges, so $\varepsilon_n \to 0$ and $g_n \to \infty$. The distributional and loss bounds depend only on the total attention mass on irrelevant coordinates, so the final part of the proof of Theorem 6 applies without change. □

*Proof of Corollary 8.* For every planning context $c = (h, a) \in \mathcal{C}_k$, let $u = (k, a)$. Assumption 4 gives $P_\nu(\cdot \mid h, a) = P^\star_{\mathrm{feat}}(\cdot \mid u, z(h))$. Theorems 6 and 19 therefore give uniform total-variation errors of at most $\lambda\varepsilon_{\mathrm{sel}}(t)/2$ and $\lambda\varepsilon^{(n)}_{\mathrm{sel}}/2$ on every reachable history–action pair. Proposition 36 identifies Bayes control under the task prior with dynamic programming in the Bayes predictive kernel. Applying Lemma 37 with $\widehat{P} = \widehat{P}_{B_t}$ and with $\widehat{P} = \widehat{P}_{B^{(n)}}$ gives the two stated suboptimality bounds.

□

*Proof of Corollary 21.* Fix a relevant-coordinate map $j_\star$, a query $u = (\ell, a)$, and an index $(j, s)$. In the construction of Proposition 17, the initial observation $o_{j,s}$ has positive probability and determines the feature vector $z^{j,s}$ throughout the episode.

Fix any action $a_0 \in \mathcal{A}$. Consider the stage-$\ell$ history that starts from $o_{j,s}$, takes action $a_0$ at every stage before $\ell$, and observes outcome label $+1$ after each of these actions. The behavior policy assigns positive probability to every selected action. Moreover, because $\lambda > 0$ is finite, $\mathrm{sigmoid}(\lambda z^{j,s}_{j_\star(k,a_0)}) \in (0,1)$ for every preceding stage $k$. Hence every selected outcome has positive conditional probability, and the resulting stage-$\ell$ history, denoted by $h_{u,j,s}$, is supported.

By construction, $z(h_{u,j,s}) = z^{j,s}$. Assumption 4 therefore gives $P_\nu(\cdot \mid h_{u,j,s}, a) = P_\nu(\cdot \mid u, z^{j,s})$. A generative model queried independently at the supported history–action pair $(h_{u,j,s}, a)$ consequently produces independent outcomes with this common distribution. □

*Proof of Theorem 22.* For each pair $(u, k)$, Assumption 4 implies that the $2N_{\mathrm{sel}}$ variables $\mathbf{1}\{Y^{(b)}_{u,k,+1} = +1\}$ and $\mathbf{1}\{Y^{(b)}_{u,k,-1} = -1\}$ are independent Bernoulli variables with common mean $t_{u,k}$. Hence $\widehat{t}_{u,k}$ is their average. Hoeffding's inequality (Hoeffding, 1963) and a union bound over the $Pn_{\mathrm{q}}$ pairs give, with probability at least $1 - \delta$ over the conditional outcome samples, $\max_{u,k} |\widehat{t}_{u,k} - t_{u,k}| \leq \varepsilon_{\mathrm{stat}}$. Work on this event. The gradient field of $\widehat{\mathcal{L}}_{\mathrm{sel}}$ is smooth and bounded, so the empirical gradient flow has a unique global solution.

Fix an arbitrary query $u$ and suppress the query index. Let $\alpha_\star$ be the attention weight on the relevant coordinate $j_\star$, let $\alpha_j$ denote the weight on an irrelevant coordinate, and set $\varepsilon = 1 - \alpha_\star = \sum_{j \neq j_\star} \alpha_j$. Define $p(x) = \mathrm{sigmoid}(\lambda(2x-1))$. For each candidate coordinate $k$, sign pairing reduces its population contribution to $\mathrm{ce}_{\mathrm{bin}}(t_k, p(\alpha_k))$ and its empirical contribution to $\mathrm{ce}_{\mathrm{bin}}(\widehat{t}_k, p(\alpha_k))$. Hence the corresponding population derivative is $r_k = 2\lambda(p(\alpha_k) - t_k)$. In particular, $r_\star = 2\lambda(p(\alpha_\star) - p_\lambda)$ and, for $j \neq j_\star$, $r_j = 2\lambda(p(\alpha_j) - (1 - p_\lambda))$. With $c_\lambda = 4\lambda^2 m_\lambda$, the mean-value theorem gives $r_\star \leq -c_\lambda \varepsilon$ and $r_j \geq c_\lambda \alpha_j$.

The empirical derivative is $\widehat{r}_k = 2\lambda(p(\alpha_k) - \widehat{t}_k) = r_k + e_k$, where $e_k = 2\lambda(t_k - \widehat{t}_k)$. On the concentration event, $|e_k| \leq d_\lambda := 2\lambda\varepsilon_{\mathrm{stat}}$. Let $\overline{r} = \sum_k \alpha_k r_k$ and $\overline{e} = \sum_k \alpha_k e_k$. Gradient flow for

one row is $\dot{B}_k = -\alpha_k(\widehat{r}_k - \sum_\ell \alpha_\ell \widehat{r}_\ell)/(n_{\mathrm{q}}P)$. For an irrelevant coordinate, define the score gap $g_j = B_{j_\star} - B_j$. The exact-target part of its derivative is $\dot{g}_j^{(0)} = [\alpha_j(r_j - \overline{r}) - \alpha_\star(r_\star - \overline{r})]/(n_{\mathrm{q}}P)$. Consider any time interval on which $\alpha_\star \geq \alpha_j$ for every irrelevant coordinate. Since $r_\star \leq 0$, $r_j \geq 0$, and $\overline{r} \geq \alpha_\star r_\star$, direct rearrangement gives $\dot{g}_j^{(0)} \geq [\alpha_j r_j + \alpha_\star(-r_\star)(\varepsilon + \alpha_j)]/(n_{\mathrm{q}}P) \geq c_\lambda \alpha_\star \varepsilon^2/(n_{\mathrm{q}}P)$.

The perturbation of this gap derivative is $[\alpha_j(e_j - \overline{e}) - \alpha_\star(e_\star - \overline{e})]/(n_{\mathrm{q}}P)$. The identities $e_\star - \overline{e} = \sum_{i \neq j_\star} \alpha_i(e_\star - e_i)$ and $e_j - \overline{e} = \alpha_\star(e_j - e_\star) + \sum_{i \neq j_\star} \alpha_i(e_j - e_i)$ imply $|e_\star - \overline{e}| \leq 2d_\lambda\varepsilon$ and $|e_j - \overline{e}| \leq 2d_\lambda$. Since $\alpha_j \leq \varepsilon$, the absolute perturbation is at most $4d_\lambda\varepsilon/(n_{\mathrm{q}}P)$.

At initialization all attention weights equal $1/P$. As long as $\varepsilon \geq r_{\mathrm{stat}} = 8Pd_\lambda/c_\lambda$, the preceding bounds and $\alpha_\star \geq 1/P$ give $\dot{g}_j \geq c_\lambda\varepsilon^2/(2n_{\mathrm{q}}P^2) > 0$. At initialization all relevant–irrelevant score gaps are zero. Therefore, by a first-crossing argument, the strict positivity of $\dot{g}_j$ whenever $\varepsilon \geq r_{\mathrm{stat}}$ prevents any gap from becoming negative in this region. Hence $\alpha_\star \geq \alpha_j$ is preserved for every irrelevant coordinate.

The softmax identity gives $\dot{\alpha}_\star = \alpha_\star \sum_{j \neq j_\star} \alpha_j \dot{g}_j$, so whenever $\varepsilon \geq r_{\mathrm{stat}}$ we have $\dot{\varepsilon} \leq -c_\lambda\varepsilon^3/(2n_{\mathrm{q}}P^3) = -2\lambda^2 m_\lambda \varepsilon^3/(n_{\mathrm{q}}P^3)$. Integration from $\varepsilon(0) = (P-1)/P$ proves the inverse-square-root term in the theorem until the trajectory first enters $\{\varepsilon \leq r_{\mathrm{stat}}\}$.

Because $r_{\mathrm{stat}} \leq 1/2$, every point on the boundary $\varepsilon = r_{\mathrm{stat}}$ satisfies $\alpha_\star = 1 - r_{\mathrm{stat}} \geq \alpha_j$ for all irrelevant coordinates. The same gap estimate is strictly positive on that boundary, so $\dot{\varepsilon} < 0$ there. The set $\{\varepsilon \leq r_{\mathrm{stat}}\}$ is therefore forward invariant. The score ordering also persists inside this set. Indeed, if a relevant–irrelevant gap first reached zero, then $\alpha_j = \alpha_\star$ and $\varepsilon \geq \alpha_j = 1 - \varepsilon$, so $\varepsilon \geq 1/2$. The only possible equality case has $\varepsilon = r_{\mathrm{stat}} = 1/2$, where the gap derivative is strictly positive. Thus no gap can cross zero. This proves the all-time bound, the score ordering, and the limiting statistical radius for every row.

The total-variation and Kullback–Leibler estimates use only the total attention mass on irrelevant coordinates. The final part of the population proof therefore applies row by row with $\varepsilon_{\mathrm{sel}}$ replaced by $\varepsilon_u$. Since binary cross-entropy excess over the conditional entropy equals the Bernoulli Kullback–Leibler divergence, averaging this bound uniformly over all $(u, j, s)$ proves the population loss estimate along the empirical trajectory. □

*Proof of Corollary 24.* For every planning context $c = (h, a) \in \mathcal{C}_k$, let $u = (k, a)$. Assumption 4 gives $P_\nu(\cdot \mid h, a) = P^\star_{\mathrm{feat}}(\cdot \mid u, z(h))$. Theorem 22 therefore gives a uniform predictive-kernel error of at most $\lambda \max_u \varepsilon_u(t)/2$ on every reachable history–action pair. Proposition 36 and Lemma 37 prove the finite-time bound. Taking the upper limit and substituting $r_{\mathrm{stat}} = 4P\varepsilon_{\mathrm{stat}}/(\lambda m_\lambda)$ gives the asymptotic bound. □

*Proof of Corollary 20.* Substitute the inverse-square-root bounds of Theorems 6 and 19 into Corollary 8 and solve for $t$ and $n$. For the empirical result, Corollary 24 gives the asymptotic bound $PH(H+1)m_\lambda^{-1}\sqrt{L_\delta/N_{\mathrm{sel}}}$. Requiring this bound to be at most $\varepsilon_{\mathrm{val}}$ and solving for $N_{\mathrm{sel}}$ gives the second term in equation (4). The condition $r_{\mathrm{stat}} \leq 1/2$ gives the first term. □

## G Trajectory-Level Finite-Sample Training for the Finite Reachable-Set Setting

The population theorem uses exact normalized visitation weights and exact conditional distributions. We now train on the empirical distribution of next-outcome prediction examples produced by independent exploratory episodes.

Observe $N$ independent training episodes of length $H$ under the behavior policy $\beta$. For context $c_i$, let $I_{n,i}$ indicate whether episode $n$ visits $c_i$, and define

$$N_i = \sum_{n=1}^{N} I_{n,i}, \ \rho_i = \mathbb{E}[I_{n,i}], \ \rho_{\min} = \min_i \rho_i > 0. \tag{89}$$

A particular stage-specific history–action pair can occur at most once in one episode. Let $i(n,k)$ be the index of the context observed at stage $k$ of episode $n$, and let $Y_{n,k}$ be its following outcome. The empirical loss is defined on every data realization as the average negative log-likelihood over all $NH$ next-outcome prediction examples:

$$\widehat{\mathcal{L}}(f) := \frac{1}{NH}\sum_{n=1}^{N}\sum_{k=1}^{H} -\log \operatorname{softmax}(f_{i(n,k)})_{Y_{n,k}}. \tag{90}$$

In this section, $\theta_t$ follows empirical gradient flow from the paired initialization,

$$\dot{\theta}_t = -\nabla_\theta \widehat{\mathcal{L}}(f(\theta_t)). \tag{91}$$

The learned predictive kernel $\widehat{P}_t$ is defined from this empirical trajectory by equation (74).

The empirical normalized visitation weights are $\widehat{d}_i = N_i/(NH)$. Whenever $N_i > 0$, define

$$\widehat{q}_i(y) = \frac{1}{N_i}\sum_{n:I_{n,i}=1} \mathbf{1}\{Y_{n,i} = y\}, \tag{92}$$

where $Y_{n,i}$ is the outcome following $c_i$ in a visited episode. Since every episode contributes $H$ training examples, $\sum_i N_i = NH$ and $\sum_i \widehat{d}_i = 1$. On the event that every reachable history–action pair is observed, the empirical loss can be grouped as

$$\widehat{\mathcal{L}}(f) = \sum_{i=1}^{m} \widehat{d}_i \operatorname{CE}\left(\widehat{q}_i, \operatorname{softmax}(f_i)\right). \tag{93}$$

On the full-coverage event, define

$$\widehat{E}(t) = \widehat{\mathcal{L}}(f_t) - \sum_{i=1}^{m} \widehat{d}_i \operatorname{Ent}(\widehat{q}_i),\ \widehat{E}_0 = \widehat{E}(0). \tag{94}$$

For $\delta \in (0, 1/2)$, set

$$\overline{\epsilon}_{\infty,N}(\delta) = \sqrt{\frac{\log(4mK/\delta)}{N\rho_{\min}}},\ \overline{\epsilon}_{1,N}(\delta) = K\overline{\epsilon}_{\infty,N}(\delta),\ \underline{d} = \frac{\rho_{\min}}{2H}. \tag{95}$$

Assume

$$N \geq \frac{8}{\rho_{\min}} \log\left(\frac{2m}{\delta}\right),\ \overline{\epsilon}_{\infty,N}(\delta) \leq \frac{\eta}{2}. \tag{96}$$

Define conservative empirical lower bounds and the resulting exponential convergence-rate constant by

$$\underline{p}_* = \exp\left(-\frac{2\log K}{\eta \underline{d}}\right),\ \underline{\mu} = \underline{d}K\underline{p}_*^2,\ \underline{c} = \frac{\gamma\underline{\mu}}{2}. \tag{97}$$

Use the constants of Lemma 31 with radius one, and define

$$\overline{r}_M = \frac{2\sqrt{K}C_{\text{head}}\sqrt{\log K}}{\underline{c}\sqrt{M}}. \tag{98}$$

**Theorem 40** (Trajectory-Level Prediction and Value-Suboptimality Bound)**.** *Suppose Assumption 26 holds. Assume the sample-size conditions above, the initial-width condition of Theorem 32, and $\overline{r}_M \leq r_*/2$. Then, with probability at least $1 - 2\delta$ over the $N$ training episodes and the network initialization, every reachable history–action pair is observed, $\widehat{d}_i \geq \underline{d}$, and*

$$\widehat{E}(t) \leq \widehat{E}_0 e^{-\underline{c}t} \tag{99}$$

*for every $t \geq 0$. The learned predictive kernel satisfies*

$$\max_i \operatorname{TV}\left(q_i, \widehat{P}_t(\cdot \mid c_i)\right) \leq \frac{1}{2}\overline{\epsilon}_{1,N}(\delta) + \sqrt{\frac{\widehat{E}_0}{2\underline{d}}} e^{-\underline{c}t/2}. \tag{100}$$

*The policy obtained by exact planning in $\widehat{P}_t$ satisfies*

$$\mathrm{SubOpt}_H(\widehat{\pi}_t) \le H(H+1)\left[\frac{1}{2}\bar{\epsilon}_{1,N}(\delta) + \sqrt{\frac{\widehat{E}_0}{2\underline{d}}}e^{-\underline{c}t/2}\right]. \tag{101}$$

The first term is conditional-distribution estimation error and the second is optimization error. The theorem also accounts for random visitation counts through the lower bound $N_i \ge N\rho_{\min}/2$. The suboptimality is evaluated on a fresh test episode independent of the training episodes and network initialization. The theorem does not replace a general mixing analysis for overlapping truncated contexts or a single continuing trajectory.

**Corollary 41** (Explicit finite-sample value requirement)**.** *Under the conditions of Theorem 40, fix $\varepsilon_{\mathrm{val}} > 0$. In addition to the sample-size and width conditions of that theorem, suppose*

$$N \ge \frac{K^2H^2(H+1)^2}{\rho_{\min}\varepsilon_{\mathrm{val}}^2}\log\left(\frac{4mK}{\delta}\right). \tag{102}$$

*If*

$$t \ge \frac{2}{\underline{c}}\max\left\{0, \log\left(\frac{2H(H+1)}{\varepsilon_{\mathrm{val}}}\sqrt{\frac{\widehat{E}_0}{2\underline{d}}}\right)\right\}, \tag{103}$$

*then $\mathrm{SubOpt}_H(\widehat{\pi}_t) \le \varepsilon_{\mathrm{val}}$ on the event of Theorem 40.*

## H Proof of Passive-Data Impossibility

*Proof of Proposition 3.* Consider a single observation $o$ and two actions 0 and 1. In task $e = 0$, action 0 gives reward one and action 1 gives reward zero. In task $e = 1$, action 1 gives reward one and action 0 gives reward zero. Let the passive corpus contain only the observation token $o$, with no action or action-conditioned reward. Its distribution is identical in the two tasks.

Any learner that receives data only from this passive distribution has the same distribution over output policies in the two tasks. Let $p$ be its marginal probability of selecting action 0. Its expected value suboptimality is $1 - p$ in task $e = 0$ and $p$ in task $e = 1$. Therefore

$$\max\{p, 1-p\} \ge 1/2. \tag{104}$$

□

## I Behavior-Policy and Shared-Embedding Proofs

*Proof of Proposition 1.* For a history $h_k$ generated by a task-independent policy $\pi$, define

$$\Psi_\pi(h_k) := \prod_{t=1}^{k-1}\pi(a_t \mid h_t). \tag{105}$$

Conditional on task $e$, the history probability factorizes as

$$\Pr_\pi(h_k \mid e) = \Psi_\pi(h_k)\Lambda_e(h_k). \tag{106}$$

The policy factor is the same for every task. Bayes' rule therefore gives

$$\begin{aligned}\Pr_\pi(e \mid h_k) &= \frac{\nu(e)\Psi_\pi(h_k)\Lambda_e(h_k)}{\sum_{e'\in\mathcal{E}}\nu(e')\Psi_\pi(h_k)\Lambda_{e'}(h_k)} \\ &= \nu(e \mid h_k).\end{aligned} \tag{107}$$

If $\pi(a \mid h_k) > 0$, the factor $\pi(a \mid h_k)$ also cancels from the posterior. Thus $\Pr_\pi(e \mid h_k, a) = \nu(e \mid h_k)$. Conditioning on the next outcome and summing over tasks gives

$$\Pr_\pi(y \mid h_k, a) = \sum_{e\in\mathcal{E}}\nu(e \mid h_k)P_e(y \mid h_k, a) = P_\nu(y \mid h_k, a). \tag{108}$$

□

*Proof of Lemma 2.* By the definition $\mathcal{H}_{k+1} = \{(h, a, r, o') : (h, a) \in \mathcal{C}_k,\ y = (r, o') \in \mathcal{Y},\ P_\nu(y \mid h, a) > 0\}$, the stage-indexed history sets are closed under positive-probability one-step extensions. We next prove by induction on the stage that every history generated by a task-independent policy belongs to the corresponding history set almost surely.

The initial observation belongs to $\mathcal{H}_1$ almost surely because its marginal distribution is $P^0_\nu$. Assume $h_k \in \mathcal{H}_k$ almost surely. Any task-independent policy selects an action in $\mathcal{A}$, and Proposition 1 shows that the conditional outcome distribution is $P_\nu(\cdot \mid h_k, a_k)$. Every realized outcome therefore has positive probability under the Bayes predictive kernel and produces a history in $\mathcal{H}_{k+1}$.

Under $\beta$, the full-support condition gives $\beta(a \mid h) > 0$ for every $(h, a) \in \mathcal{C}$. Fix any context $c_i = (h_k, a_k) \in \mathcal{C}_k$, where $h_k$ contains the actions $a_1, \ldots, a_{k-1}$ and outcomes $y_1, \ldots, y_{k-1}$. By the recursive definition of the reachable history sets, $P^0_\nu(o_1) > 0$ and $P_\nu(y_t \mid h_t, a_t) > 0$ for every $t \in \{1, \ldots, k-1\}$. Therefore,

$$\Pr_\beta(\text{the episode visits } c_i) = P^0_\nu(o_1) \left[\prod_{t=1}^{k-1} \beta(a_t \mid h_t) P_\nu(y_t \mid h_t, a_t)\right] \beta(a_k \mid h_k) > 0. \quad (109)$$

Hence every context in $\mathcal{C}$ has positive episode-level visitation probability. □

*Proof of Proposition 25.* Let $S = \mathrm{col}(W)$, so $\dim S = r$. Every embedding $x_{i,j} = We_{\xi_{i,j}}$ and every last-token query $u_i = x_{i,\ell_i}$ belongs to $S$. Since $\bar{x}_i \in S$, each centered embedding $x_{i,j} - \bar{x}_i$ also belongs to $S$. Hence $C_i$ is symmetric, its range is contained in $S$, and it vanishes on $S^\perp$.

Let $Q \in \mathbb{R}^{d \times r}$ have orthonormal columns spanning $S$. Then every such $C_i$ can be written as $C_i = QM_iQ^\top$ for some symmetric matrix $M_i \in \mathbb{R}^{r \times r}$. Therefore the linear space containing the matrices $C_i$ has dimension at most $r(r+1)/2$.

Since $u_i \in S$, every vector $u_i \otimes \mathrm{vec}(C_i)$ lies in a tensor-product space of dimension at most $r \cdot r(r+1)/2 = r^2(r+1)/2$. The rows of $Z_{\mathrm{ctx}}$ therefore span a space of dimension at most $r^2(r+1)/2$, which proves equation (28). Finally, $r = \mathrm{rank}(W) \le \min\{d, D\}$ gives the second bound. □

*Proof of Proposition 28.* For the first claim, let $W_0 \in \mathbb{R}^{d \times D}$ have its first $D$ rows equal to $\boldsymbol{I}_D$ and all remaining rows equal to zero. Then $W_0 e_\xi$ is the one-hot embedding padded with zeros. Consequently, $u_i(W_0)$ and $C_i(W_0)$ are padded versions of $u_i^{\mathrm{oh}}$ and $C_i^{\mathrm{oh}}$, and $Z_{\mathrm{ctx}}(W_0)$ is obtained from $Z_{\mathrm{oh}}$ by an injective coordinate embedding. Assumption 27 gives $\mathrm{rank}(Z_{\mathrm{ctx}}(W_0)) = m$.

For the genericity claim, fix any dimension $d$ and any $W_0 \in \mathbb{R}^{d \times D}$ for which $Z_{\mathrm{ctx}}(W_0)$ has full row rank. Choose an $m \times m$ minor of $Z_{\mathrm{ctx}}(W)$ that is nonzero at $W_0$. Every entry of $u_i(W) \otimes \mathrm{vec}(C_i(W))$ is a polynomial of degree three in the entries of $W$, so the determinant of this minor is a polynomial in $W$. It is not the zero polynomial because it is nonzero at $W_0$. The zero set of a nonzero real polynomial has Lebesgue measure zero. Every rank-deficient $W$ makes the selected minor vanish, so the rank-deficient set also has measure zero. An absolutely continuous distribution assigns probability zero to that set. □

*Proof of Corollary 29.* Under the one-hot embedding, $\xi_{i,p} \neq \xi_{i,q}$ whenever $p \neq q$, because $\xi_{i,j} = (\tau_{i,j}, j)$. For such $p$ and $q$, the covariance coordinate indexed by $(\xi_{i,p}, \xi_{i,q})$ is $-1/\ell_i^2$. Hence the coordinate of $u_i^{\mathrm{oh}} \otimes \mathrm{vec}(C_i^{\mathrm{oh}})$ indexed by $(\xi_{i,\ell_i}, \xi_{i,p_i}, \xi_{i,q_i})$ equals $-1/\ell_i^2$. The unique-triple condition makes this coordinate zero in every other row. The selected coordinates are distinct across contexts, so the corresponding $m \times m$ column submatrix is

$$B_{\mathrm{uniq}} = \mathrm{diag}(-\ell_1^{-2}, \ldots, -\ell_m^{-2}). \quad (110)$$

After a column permutation, write $Z_{\mathrm{oh}} = [B_{\mathrm{uniq}}\ R_{\mathrm{uniq}}]$. Then

$$Z_{\mathrm{oh}} Z_{\mathrm{oh}}^\top = B_{\mathrm{uniq}} B_{\mathrm{uniq}}^\top + R_{\mathrm{uniq}} R_{\mathrm{uniq}}^\top \succeq B_{\mathrm{uniq}} B_{\mathrm{uniq}}^\top. \quad (111)$$

Therefore $\sigma_{\min}(Z_{\mathrm{oh}}) \geq \min_i \ell_i^{-2} \geq L^{-2}$. For the padded one-hot embedding, $Z_{\mathrm{ctx}}(W_0)$ is obtained by an isometric coordinate embedding of $Z_{\mathrm{oh}}$, so the two matrices have the same nonzero singular values. Finally, Weyl's inequality gives

$$\sigma_{\min}(Z_{\mathrm{ctx}}(W)) \geq \sigma_{\min}(Z_{\mathrm{ctx}}(W_0)) - \|Z_{\mathrm{ctx}}(W) - Z_{\mathrm{ctx}}(W_0)\|_{\mathrm{op}}\,, \tag{112}$$

which proves the neighborhood bound. Continuity follows because every entry of $Z_{\mathrm{ctx}}(W)$ is a polynomial in the entries of $W$. □

# J Attention Derivatives and Jacobian Rank

## J.1 Proof of Proposition 30

*Proof.* At $A = 0$, $\alpha_{i,j}(0) = 1/\ell_i$ for every $j$, and therefore $s_i(0) = \bar{x}_i$. Equation (18) gives

$$\nabla_A \varphi_i(0, v) = \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} \left[v^\top (x_{i,j} - \bar{x}_i)\right] u_i x_{i,j}^\top. \tag{113}$$

By the definition of $C_i$ and the identity $\sum_{j=1}^{\ell_i}(x_{i,j} - \bar{x}_i) = 0$,

$$\begin{aligned} C_i &= \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^\top \\ &= \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} x_{i,j}(x_{i,j} - \bar{x}_i)^\top. \end{aligned} \tag{114}$$

Hence

$$\nabla_A \varphi_i(0, v) = u_i (C_i v)^\top. \tag{115}$$

This proves equation (36).

For contexts $i$ and $j$, the Frobenius inner product of the two score gradients satisfies

$$\begin{aligned} [\boldsymbol{G}_A]_{ij} &= \mathbb{E}_{v^0} \left\langle u_i (C_i v^0)^\top, u_j (C_j v^0)^\top \right\rangle_F \\ &= (u_i^\top u_j) \mathbb{E}_{v^0} \left[(v^0)^\top C_i C_j v^0\right] \\ &= \sigma_v^2 (u_i^\top u_j) \operatorname{tr}(C_i C_j), \end{aligned} \tag{116}$$

where the last equality uses $\mathbb{E}[v^0 (v^0)^\top] = \sigma_v^2 \boldsymbol{I}_d$.

Since $C_i$ and $C_j$ are symmetric, $\operatorname{vec}(C_i)^\top \operatorname{vec}(C_j) = \operatorname{tr}(C_i C_j)$. Together with $\langle a \otimes b, c \otimes d \rangle = (a^\top c)(b^\top d)$, this gives

$$[\boldsymbol{G}_A]_{ij} = \sigma_v^2 \left\langle u_i \otimes \operatorname{vec}(C_i), u_j \otimes \operatorname{vec}(C_j) \right\rangle. \tag{117}$$

Therefore

$$\boldsymbol{G}_A = \sigma_v^2 Z_{\mathrm{ctx}} Z_{\mathrm{ctx}}^\top. \tag{118}$$

Assumption 26 and the definition of $\sigma_{\min}$ give

$$Z_{\mathrm{ctx}} Z_{\mathrm{ctx}}^\top \succeq \zeta^2 \boldsymbol{I}_m. \tag{119}$$

Consequently,

$$\boldsymbol{G}_A \succeq \sigma_v^2 \zeta^2 \boldsymbol{I}_m. \tag{120}$$

This proves equation (38). □

## J.2 Proof of Lemma 31

*Proof.* Equation (39) follows because $s_i(A)$ is a convex combination of vectors with norm at most $R$.

For a perturbation $U$, equation (17) and Cauchy–Schwarz give

$$\begin{aligned}\|D_A s_i(A)[U]\|_2 &\le \sum_j \alpha_{i,j}(A)\,\|x_{i,j} - s_i(A)\|_2\,\left|x_{i,\ell_i}^\top U x_{i,j}\right| \\ &\le 2R^3\,\|U\|_F\,.\end{aligned} \tag{121}$$

The mean-value theorem proves equation (40).

For directions $U$ and $V$ with unit Frobenius norm, write

$$\zeta_U(j) = x_{i,\ell_i}^\top U x_{i,j},\ \zeta_V(j) = x_{i,\ell_i}^\top V x_{i,j}. \tag{122}$$

Under the attention distribution, $|\zeta_U(j)|$ and $|\zeta_V(j)|$ are at most $R^2$. Using $D_A\mathbb{E}_\alpha[g][V] = \mathrm{Cov}_\alpha(g, \zeta_V)$, a second differentiation gives

$$\begin{aligned}D_A^2 s_i(A)[U,V] = \mathrm{Cov}_\alpha(x\zeta_U, \zeta_V) - \mathrm{Cov}_\alpha(x, \zeta_V)\mathbb{E}_\alpha[\zeta_U] \\ - s_i(A)\,\mathrm{Cov}_\alpha(\zeta_U, \zeta_V).\end{aligned} \tag{123}$$

Each of the three terms has norm at most $2R^5$. Hence

$$\left\|D_A^2 s_i(A)\right\|_{\mathrm{op}} \le 6R^5 = L_2. \tag{124}$$

The gradient of $\varphi_i(A, v) = v^\top s_i(A)$ has blocks

$$\nabla_v \varphi_i = s_i(A),\ \nabla_A \varphi_i = D_A s_i(A)^*[v]. \tag{125}$$

Equations (44) and (45) follow from the previous bounds. Finally,

$$\begin{aligned}&\|D_A s_i(A)^*[v] - D_A s_i(A')^*[v']\|_F \\ &\le L_1\,\|v - v'\|_2 + (C_{\mathrm{init}} + r)L_2\,\|A - A'\|_F\,,\end{aligned} \tag{126}$$

which implies equation (46) with $L_{\mathrm{score}} = L_1 + (C_{\mathrm{init}} + r)L_2$. □

## K Initial Neural Tangent Kernel Concentration

*Proof of Theorem 32.* For one base pair, define

$$Z_r = \left[\boldsymbol{\Phi}_A(\omega_r^0)\boldsymbol{\Phi}_A(\omega_r^0)^\top\right] \otimes (b_r b_r^\top). \tag{127}$$

Then equation (47) gives

$$\boldsymbol{\Theta}_A(0) = \frac{1}{M}\sum_{r=1}^{M} Z_r. \tag{128}$$

The matrices $Z_r$ are independent and positive semidefinite. Since $\|b_r\|_2^2 = K$ and every row of $\boldsymbol{\Phi}_A(\omega_r^0)$ has norm at most $C_{\mathrm{score}}$,

$$\lambda_{\max}(Z_r) \le KmC_{\mathrm{score}}^2. \tag{129}$$

Independence of $b_r$ and $\omega_r^0$ gives

$$\mathbb{E}[Z_r] = \boldsymbol{G}_A \otimes \boldsymbol{I}_K \succeq \gamma \boldsymbol{I}_{mK}. \tag{130}$$

The lower-tail matrix Chernoff (Tropp, 2012) bound with relative deviation $1/2$ implies

$$\Pr\left[\lambda_{\min}(\boldsymbol{\Theta}_A(0)) \le \gamma/2\right] \le mK\exp\left(-\frac{M\gamma}{8KmC_{\mathrm{score}}^2}\right). \tag{131}$$

Equation (49) makes the right-hand side at most $\delta$. □

## L Cross-Entropy Geometry

*Proof of Lemma 33.* Let $p_i = \mathrm{softmax}(f_i)$. Since $\mathcal{L}(f) \le \log K$ and every term is nonnegative,

$$\mathrm{CE}(q_i, p_i) \le \frac{\log K}{d_i} \le \frac{\log K}{d_{\min}}. \tag{132}$$

For every outcome $y$,

$$-\eta \log p_i(y) \le -q_i(y) \log p_i(y) \le \mathrm{CE}(q_i, p_i). \tag{133}$$

This proves equation (55).

The Hessian of $\mathrm{CE}(q_i, \mathrm{softmax}(f_i))$ is

$$H_i = \mathrm{diag}(p_i) - p_i p_i^\top. \tag{134}$$

For $u^\top \mathbf{1} = 0$, equation (58) gives

$$\begin{aligned} u^\top H_i u &\ge \frac{p_*^2}{2} \sum_{a,b=1}^{K} (u_a - u_b)^2 \\ &= K p_*^2 \left\| u \right\|_2^2 . \end{aligned} \tag{135}$$

After multiplying by $d_i$ and summing over contexts, the population Hessian is bounded below by $\mu \boldsymbol{P}$ with $\mu = d_{\min} K p_*^2$.

Let $f^{\mathrm{c}} = \boldsymbol{P} f$ and let $f^\star$ be the centered minimizer with block $f_i^\star = \log q_i - K^{-1} \mathbf{1}\mathbf{1}^\top \log q_i$. Softmax invariance gives $\mathcal{L}(f^{\mathrm{c}}) = \mathcal{L}(f)$, and convexity keeps the segment between $f^{\mathrm{c}}$ and $f^\star$ inside the same loss sublevel set. Strong convexity on this centered segment implies

$$\mathcal{L}(f) - \mathcal{L}^* \le \frac{1}{2\mu} \left\| \nabla_f \mathcal{L}(f) \right\|_2^2 , \tag{136}$$

which is equation (57). □

# M Uniform Neural Tangent Kernel Control

## M.1 Parameter-gradient bound

Let

$$g_i(t) = \mathrm{softmax}(f_i(t)) - q_i, \ \left\| g(t) \right\|_d^2 = \sum_{i=1}^{m} d_i \left\| g_i(t) \right\|_2^2 . \tag{137}$$

Pinsker's inequality gives

$$\left\| g(t) \right\|_d^2 \le 2E(t). \tag{138}$$

For one signed copy, differentiating equation (23) yields

$$\dot{\omega}_{r,\pm} = \mp \tfrac{1}{\sqrt{2M}} \textstyle\sum_{i=1}^{m} d_i \big(b_r^\top g_i(t)\big) \nabla_\omega \varphi_i(\omega_{r,\pm}), \tag{139}$$

Weighted Cauchy–Schwarz and Lemma 31 give

$$\left\| \dot{\omega}_{r,s}(t) \right\|_2 \le \frac{\sqrt{K} C_{\mathrm{head}}}{\sqrt{2M}} \left\| g(t) \right\|_d \tag{140}$$

while the parameter remains in the unit initialization ball.

## M.2 Neural Tangent Kernel Drift

Let $J_{A,r,s}(t)$ be the Jacobian of the stacked logits with respect to the entries of $A_{r,s}$. Within the unit initialization ball,

$$\left\| J_{A,r,s}(t) \right\|_{\mathrm{op}} \le \frac{\sqrt{Km} C_{\mathrm{score}}}{\sqrt{2M}}, \tag{141}$$

$$\left\| J_{A,r,s}(t) - J_{A,r,s}(0) \right\|_{\mathrm{op}} \le \frac{\sqrt{Km} L_{\mathrm{score}}}{\sqrt{2M}} \left\| \omega_{r,s}(t) - \omega_r^0 \right\|_2 . \tag{142}$$

Using

$$\left\|UU^\top - VV^\top\right\|_{\mathrm{op}} \le \left(\|U\|_{\mathrm{op}} + \|V\|_{\mathrm{op}}\right)\|U - V\|_{\mathrm{op}}, \tag{143}$$

we obtain, whenever all heads move by at most $r \le 1$,

$$\|\boldsymbol{\Theta}_A(t) - \boldsymbol{\Theta}_A(0)\|_{\mathrm{op}} \le 2KmC_{\mathrm{score}}L_{\mathrm{score}}r. \tag{144}$$

### M.3 Proof of Theorems 34 and 35

*Proof.* Work on the event in Theorem 32. Define the first exit time

$$T_* = \inf\left\{t \ge 0 : \max_{r,s}\left\|\omega_{r,s}(t) - \omega_r^0\right\|_2 \ge r_*\right\}. \tag{145}$$

For $t < T_*$, equation (144) and the definition of $r_*$ imply

$$\|\boldsymbol{\Theta}_A(t) - \boldsymbol{\Theta}_A(0)\|_{\mathrm{op}} \le \gamma/4. \tag{146}$$

Together with equation (50),

$$\boldsymbol{P}\boldsymbol{\Theta}_A(t)\boldsymbol{P} \succeq \gamma\boldsymbol{P}/4. \tag{147}$$

The full neural tangent kernel matrix dominates the contribution from the attention score matrices. Equations (61), (57), and (147) give

$$\dot{E}(t) \le -\frac{\gamma}{4}\|\nabla_f\mathcal{L}(f_t)\|_2^2 \le -\frac{\gamma\mu}{2}E(t) = -cE(t). \tag{148}$$

Therefore

$$E(t) \le E_0 e^{-ct} \tag{149}$$

for $t < T_*$. Equations (138) and (140) imply

$$\begin{aligned}\left\|\omega_{r,s}(t) - \omega_r^0\right\|_2 &\le \frac{\sqrt{K}C_{\mathrm{head}}}{\sqrt{2M}}\int_0^t \sqrt{2E_0}e^{-cu/2}\,\mathrm{d}u \\ &\le \frac{2\sqrt{K}C_{\mathrm{head}}\sqrt{E_0}}{c\sqrt{M}} \\ &= r_M.\end{aligned} \tag{150}$$

Since $r_M \le r_*/2$, continuity prevents the trajectory from reaching the exit boundary. The gradient-flow vector field is smooth in the finite-dimensional parameters. On every bounded initialization ball it is locally Lipschitz and bounded, so the standard continuation theorem rules out a finite maximal existence time while the trajectory remains in that ball. Hence $T_* = \infty$, and equations (66), (67), and (68) follow.

Finally, because $\nabla_{f_i}\mathcal{L}(f_t) = d_i g_i(t)$ and $0 < d_i \le 1$ for every $i$,

$$\begin{aligned}\|\nabla_f\mathcal{L}(f_t)\|_2^2 &= \sum_{i=1}^m d_i^2\|g_i(t)\|_2^2 \\ &\le \sum_{i=1}^m d_i\|g_i(t)\|_2^2 \\ &= \|g(t)\|_d^2.\end{aligned} \tag{151}$$

Therefore,

$$\|\nabla_f\mathcal{L}(f_t)\|_2 \le \|g(t)\|_d \le \sqrt{2E_0}e^{-ct/2}. \tag{152}$$

Integrating the preceding bound gives equation (69). □

## N PLANNING PROOFS

*Proof of Proposition 36.* Under the task mixture, the initial observation has distribution $P_\nu^0$. Proposition 1 shows that, for every task-independent policy and every supported history–action pair, the conditional distribution of the next outcome is $P_\nu(\cdot \mid h, a)$. Hence, for every complete trajectory $\tau = (o_1, a_1, y_1, \ldots, a_H, y_H)$ with positive probability,

$$\Pr_\nu^\pi(\tau) = P_\nu^0(o_1) \prod_{k=1}^{H} \pi(a_k \mid h_k) P_\nu(y_k \mid h_k, a_k). \tag{153}$$

The right-hand side is exactly the trajectory probability under the history process with initial law $P_\nu^0$ and Bayes predictive kernel $P_\nu$. The two processes therefore have the same trajectory distribution and the same expected return for every policy, which proves equation (72).

The sets are finite and the horizon is finite, so backward induction on the history tree attains $V_{P_\nu}^*$ with a deterministic history policy. Equality of the value of every policy under the two processes implies $V_\nu^* = V_{P_\nu}^*$ and shows that every optimizer under $P_\nu$ is Bayes-optimal under the original task prior. □

*Proof of Lemma 37.* Let $\Delta_n$ be the largest conditional value difference over supported histories for a fixed policy when $n$ decision steps remain. Since rewards lie in $[0, 1]$, a one-step reward plus a continuation value with $n - 1$ remaining steps lies in $[0, n]$. Total variation therefore gives

$$\Delta_n \le n\epsilon + \Delta_{n-1}, \ \Delta_0 = 0. \tag{154}$$

Thus

$$\Delta_H \le \sum_{n=1}^{H} n\epsilon = \frac{H(H+1)}{2}\epsilon. \tag{155}$$

This proves equation (81).

Let $\pi^*$ be optimal under $\mathsf{P}$ and let $\widehat{\pi}$ be optimal under $\widehat{P}$. Then

$$\begin{aligned} V_{\mathsf{P}}^* - V_{\mathsf{P}}^{\widehat{\pi}} &= V_{\mathsf{P}}^{\pi^*} - V_{\widehat{P}}^{\pi^*} + V_{\widehat{P}}^{\pi^*} - V_{\widehat{P}}^{\widehat{\pi}} + V_{\widehat{P}}^{\widehat{\pi}} - V_{\mathsf{P}}^{\widehat{\pi}} \\ &\le 2\Delta_H, \end{aligned} \tag{156}$$

which proves equation (82). □

*Proof of Theorem 38.* Equation (75) implies

$$D_{\mathrm{KL}}\big(q_i \,\|\, \widehat{P}_t(\cdot \mid c_i)\big) \le E(t)/d_{\min} \tag{157}$$

for every context. Pinsker's inequality proves equation (76). Proposition 36 gives $V_\nu^* = V_{P_\nu}^*$ and $V_\nu^{\widehat{\pi}_t} = V_{P_\nu}^{\widehat{\pi}_t}$. Apply Lemma 37 with $\mathsf{P} = P_\nu$ and $\widehat{P} = \widehat{P}_t$, then use equation (68). □

*Proof of Corollary 39.* Substitute equation (68) into equation (83) and solve the resulting exponential inequality for $t$. □

## O TRAJECTORY-LEVEL FINITE-SAMPLE PROOF

*Proof of Theorem 40.* For fixed $i$, the indicators $I_{1,i}, \ldots, I_{N,i}$ are independent Bernoulli variables with mean $\rho_i$. A multiplicative Chernoff bound gives

$$\Pr\left(N_i < \frac{N\rho_i}{2}\right) \le \exp\left(-\frac{N\rho_i}{8}\right). \tag{158}$$

The first sample-size condition and a union bound imply, with probability at least $1 - \delta/2$, that $N_i \ge N\rho_{\min}/2$ for every $i$. On this event, $\widehat{d}_i = N_i/(NH) \ge \underline{d}$.

For each fixed context $i$, condition only on the vector $(I_{1,i}, \dots, I_{N,i})$. The event $I_{n,i} = 1$ is measurable before the outcome following $c_i$ is drawn. By Proposition 1 and independence across episodes, conditional on this vector, the outcomes $\{Y_{n,i} : I_{n,i} = 1\}$ are independent and have common distribution $q_i$.

Let $\mathcal{E}_{\mathrm{count},i} = \{N_i \geq N\rho_{\min}/2\}$ and $\mathcal{E}_{\mathrm{count}} = \bigcap_{i=1}^{m} \mathcal{E}_{\mathrm{count},i}$. For every realization of $(I_{1,i}, \dots, I_{N,i})$ in $\mathcal{E}_{\mathrm{count},i}$ and every $y \in \mathcal{Y}$, conditional Hoeffding's inequality gives

$$\Pr\left(|\widehat{q}_i(y) - q_i(y)| > \overline{\epsilon}_{\infty,N}(\delta) \mid I_{1,i}, \dots, I_{N,i}\right) \leq \frac{\delta}{2mK}. \tag{159}$$

Indeed, on $\mathcal{E}_{\mathrm{count},i}$, $2N_i\overline{\epsilon}_{\infty,N}(\delta)^2 \geq \log(4mK/\delta)$. Let $\mathcal{E}_{i,y} = \{|\widehat{q}_i(y) - q_i(y)| \leq \overline{\epsilon}_{\infty,N}(\delta)\}$. Integrating the preceding conditional inequality over the fixed-context indicator vector yields

$$\Pr\left(\mathcal{E}_{i,y}^{\mathrm{c}} \cap \mathcal{E}_{\mathrm{count}}\right) \leq \Pr\left(\mathcal{E}_{i,y}^{\mathrm{c}} \cap \mathcal{E}_{\mathrm{count},i}\right) \leq \frac{\delta}{2mK}. \tag{160}$$

A union bound over $i$ and $y$, followed by combination with the count-event bound, therefore gives, with probability at least $1 - \delta$,

$$\max_{i,y} |\widehat{q}_i(y) - q_i(y)| \leq \overline{\epsilon}_{\infty,N}(\delta), \ \max_{i} \|\widehat{q}_i - q_i\|_1 \leq \overline{\epsilon}_{1,N}(\delta). \tag{161}$$

The positivity condition implies $\widehat{q}_i(y) \geq \eta/2$ on this event.

On the full-coverage event, equation (93) has the same form as the population loss, with weights $\widehat{d}_i$ and targets $\widehat{q}_i$. Its initial value is $\log K$, and $\widehat{E}_0 \leq \log K$. Since $\widehat{d}_i \geq \underline{d}$ and $\widehat{q}_i(y) \geq \eta/2$, the sublevel-set probability floor is at least $\underline{p}_*$, the centered-logit curvature is at least $\underline{\mu}$, and the exponential rate is at least $\underline{c}$. The initial tangent concentration event has probability at least $1 - \delta$. The condition $\overline{r}_M \leq r_*/2$ then repeats the uniform-kernel bootstrap with the conservative constants and gives $\widehat{E}(t) \leq \widehat{E}_0 e^{-\underline{c}t}$ for all $t \geq 0$. A union bound over the data and initialization events gives probability at least $1 - 2\delta$.

For every context,

$$\begin{aligned} \mathrm{TV}\left(q_i, \widehat{P}_t(\cdot \mid c_i)\right) &\leq \mathrm{TV}(q_i, \widehat{q}_i) + \mathrm{TV}\left(\widehat{q}_i, \widehat{P}_t(\cdot \mid c_i)\right) \\ &\leq \frac{1}{2}\overline{\epsilon}_{1,N}(\delta) + \sqrt{\frac{D_{\mathrm{KL}}(\widehat{q}_i \,\|\, \widehat{P}_t(\cdot \mid c_i))}{2}} \\ &\leq \frac{1}{2}\overline{\epsilon}_{1,N}(\delta) + \sqrt{\frac{\widehat{E}(t)}{2\underline{d}}}. \end{aligned} \tag{162}$$

Use the empirical convergence bound, Proposition 36, and then Lemma 37. □

*Proof of Corollary 41.* Under equation (102), the conditional-distribution estimation term in equation (101) is at most $\varepsilon_{\mathrm{val}}/2$. Under equation (103), the optimization term is at most $\varepsilon_{\mathrm{val}}/2$. Adding the two terms proves the claim. □

## P Additional Related Work

Sparse and structured dependencies have been studied through relevant-token selection, input-location recovery, and causal attention (Wang et al., 2024b; Nichani et al., 2024; Marion et al., 2025; Dohmatob, 2025; Duranthon et al., 2026; Wu et al., 2025; Golowich et al., 2025; Chen et al., 2025; Wen et al., 2025). Complementary analyses address optimization and training dynamics for specific attention parameterizations (Wu et al., 2023; Gao et al., 2024; Kim & Suzuki, 2024; Huang et al., 2024; Li et al., 2024; Huang et al., 2025; Saponati et al., 2025; Varre et al., 2025). Transition estimation and recovery, next-token prediction, and planning have been considered in Ildiz et al. (2024); Shi & Cao (2025); Wang et al. (2024a); He et al. (2024). Generalization of next-token-prediction pretraining is studied by Li et al. (2025); model-based reinforcement-learning analyses connect model error to decision quality (Ayoub et al., 2020; Fan & Ramadge, 2021; Ávila Pires & Szepesvári, 2016). Predictive-state and minimal-realization theories likewise characterize sufficient or minimal internal states from controlled

input–output behavior (Littman et al., 2001; Sakcak et al., 2024). Our contribution is not the general existence of such a state, but an explicit weight-derived realization for autoregressive Transformers together with a fixed-token-pair finite inverse, cache-level reconstruction, and intervention-specific dimension and identifiability results.

The initialization results address a different issue from architectural interaction order. Ustaomeroglu & Qu (2025) analyze representation and learning of pairwise interactions by linear attention, with higher-order results for modified attention architectures. Parity has also been used to study Transformer expressivity: Kozachinskiy et al. (2026) establish a lower bound for one-layer Transformers and give deeper softmax constructions compatible with causal masking. Our softmax architecture is unchanged between the negative and positive results. The obstruction instead concerns the Jacobian at uniform-attention initialization: zero scores restrict its interaction degree, whereas target-independent nonzero score initialization yields an explicit higher-order spectrum, nonasymptotic eigenvalue bounds, and finite-width population and empirical learning guarantees. This initialization effect is related to broader work on gradient-based learning failures, tangent-kernel approximations, and attention kernels (Shalev-Shwartz et al., 2017; Malach et al., 2021; Chizat et al., 2019; Hron et al., 2020; Song et al., 2024). The empirical result treats unprojected nonlinear training under bounded perturbations of paired initialization.

**Scope of the one-sparse comparison.** Recent analyses of single-location regression study localization and optimization in attention models, including population dynamics and finite-sample generalization (Marion et al., 2025; Dohmatob, 2025; Duranthon et al., 2026). Our one-sparse model instead fixes the feature map and the relevant coordinate for each query, while allowing a general binary input distribution characterized by feature-disagreement probabilities. This yields an exact identification condition, a matching fixed-design bound, and population, finite-sample, and exploratory-episode optimization guarantees under the same condition. The assumptions and conclusions are therefore different from those of single-location regression rather than being a strict generalization of that setting.

## Q Constants for the Finite Reachable-Set Guarantee

To state the finite-width result, we first collect the quantities that control the cross-entropy decay and the stability of the attention-score Jacobian. Let $E(t) = \mathcal{L}(f(\theta_t)) - \mathcal{L}^*$ denote the excess population cross-entropy at time $t$, and let $E_0 = \log K - \sum_i d_i \operatorname{Ent}(q_i) = E(0)$. The full-support condition yields the probability lower bound $p_* = \exp(-\log K/(\eta d_{\min}))$ on the relevant loss sublevel set. Set $\mu = d_{\min} K p_*^2$, which gives the corresponding lower curvature bound for the cross-entropy. The initialization condition $\sigma_{\min}(Z_{\mathrm{ctx}}) \geq \zeta$ gives the Jacobian Gram lower-bound constant $\gamma = \sigma_v^2 \zeta^2$. We then define $c = \gamma\mu/2$, which will be the exponential decay rate of the excess loss.

The remaining quantities control Jacobian variation and parameter movement away from initialization. From Lemma 31, let $L_1 = 2R^3$ and $L_2 = 6R^5$ denote the first- and second-order bounds for the attention summary, and let $C_{\mathrm{init}} = \sigma_v\sqrt{d}$ be the initialization norm of each value vector. With radius fixed to one, set $C_{\mathrm{score}} = L_1(C_{\mathrm{init}} + 1)$, $C_{\mathrm{head}} = \sqrt{R^2 + C_{\mathrm{score}}^2}$, and $L_{\mathrm{score}} = L_1 + (C_{\mathrm{init}} + 1)L_2$. Here $C_{\mathrm{score}}$ bounds the gradient with respect to the attention score matrix, $C_{\mathrm{head}}$ bounds the full single-head gradient, and $L_{\mathrm{score}}$ bounds the variation of the attention-score Jacobian. Finally, define $r_* = \min\{1, \gamma/(8KmC_{\mathrm{score}}L_{\mathrm{score}})\}$ as the radius within which the kernel lower bound remains stable, and $r_M = 2\sqrt{K}C_{\mathrm{head}}\sqrt{E_0}/(c\sqrt{M})$ as the resulting upper bound on the movement of each head from initialization.

## R General Training Distributions and Independent Episodes

This section retains Assumption 4 and the attention parameterization of Section 3, but replaces the symmetric input distribution by arbitrary distributions $D_u$ on $\{-1, 1\}^P$. Write $\Delta_P = \{\alpha \in [0, 1]^P : \sum_j \alpha_j = 1\}$ and $c_\lambda = 4\lambda^2 m_\lambda$. The weights $w_u$ are positive and sum to one. A common input distribution means that, when comparing candidate relevant coordinates, the law of the input is held fixed. This requirement is used for the necessity statement about identification. The sufficient optimization and sampling results hold for the input law of each fixed environment.

### R.1 IDENTIFICATION AND POPULATION OPTIMIZATION

**Lemma 42** (Loss and score differences)**.** *Fix a query, suppress its index, and write $\star = j_\star(u)$. Let $F_D(\alpha) = \mathbb{E}_D[\mathrm{ce}_{\mathrm{bin}}(\mathrm{sigmoid}(\lambda Z_\star), \mathrm{sigmoid}(\lambda \alpha^\top Z))]$ and $\kappa_j = \Pr_D\{Z_j \neq Z_\star\}$. Define $\varepsilon = 1 - \alpha_\star$, $M(Z) = \sum_{j \neq \star} \alpha_j \mathbf{1}\{Z_j \neq Z_\star\}$, and $q_j = \partial_j F_D - \partial_\star F_D$ for $j \neq \star$. Then*

$$\frac{c_\lambda}{2}\mathbb{E}_D[M(Z)^2] \leq F_D(\alpha) - F_D(e_\star) \leq \frac{\lambda^2}{2}\mathbb{E}_D[M(Z)^2], \tag{163}$$

*and $q_j \geq c_\lambda \mathbb{E}_D[\mathbf{1}\{Z_j \neq Z_\star\} M(Z)] \geq c_\lambda \kappa_j \alpha_j$. If $\kappa = \min_{j \neq \star} \kappa_j > 0$, the unique minimizer on $\Delta_P$ is $e_\star$.*

*For gradient flow on $wF_D(\mathrm{softmax}(B))$, put $g_j = B_\star - B_j$ and $\overline{q} = \sum_{i \neq \star} \alpha_i q_i$. Then*

$$\dot{g}_j = w[(\alpha_\star - \alpha_j)\overline{q} + \alpha_j q_j]. \tag{164}$$

*On the region $\alpha_\star \geq \alpha_j$ for every $j \neq \star$, if $\kappa > 0$, every gap satisfies $\dot{g}_j \geq w A_D \varepsilon^2$, where $A_D = 3c_\lambda \kappa / [4P(P-1)]$.*

*Proof.* For binary coordinates and $\alpha \in \Delta_P$, the identity $\alpha^\top Z - Z_\star = -2Z_\star M(Z)$ holds pointwise. Both natural parameters $\lambda \alpha^\top Z$ and $\lambda Z_\star$ lie in $[-\lambda, \lambda]$. The second derivative of $A_{\log}(x) = \log(1 + e^x)$ lies between $m_\lambda$ and $1/4$ on this interval. Binary cross-entropy excess equals the Bernoulli Kullback–Leibler divergence, which is the Bregman divergence of $A_{\log}$ in natural parameters. Its integral remainder is therefore between $m_\lambda/2$ and $1/8$ times their squared difference. Substituting the difference $2\lambda M(Z)$ proves the two loss bounds.

Differentiation gives

$$q_j = \lambda \mathbb{E}_D\left[\left(\mathrm{sigmoid}(\lambda \alpha^\top Z) - \mathrm{sigmoid}(\lambda Z_\star)\right)(Z_j - Z_\star)\right]. \tag{165}$$

On $\{Z_j \neq Z_\star\}$, the second factor is $-2Z_\star$ and the first has the same sign, with magnitude at least $2\lambda m_\lambda M(Z)$. Their product is nonnegative elsewhere as well, and the integrand vanishes when $Z_j = Z_\star$. This proves the lower bound on $q_j$. Because $M(Z) \geq \alpha_j \mathbf{1}\{Z_j \neq Z_\star\}$, the second inequality follows. Moreover, since all summands of $M$ are nonnegative,

$$\mathbb{E}_D[M(Z)^2] \geq \sum_{j \neq \star} \kappa_j \alpha_j^2 \geq \frac{\kappa \varepsilon^2}{P-1}. \tag{166}$$

When $\kappa > 0$, the loss bound is strictly positive for every $\alpha \neq e_\star$, proving uniqueness.

The softmax Jacobian is $\partial \alpha_i / \partial B_j = \alpha_i(\mathbf{1}\{i = j\} - \alpha_j)$. Subtracting the two score derivatives gives the stated gap equation. To bound it, let $a = \alpha_\star$, $b = \alpha_j$, and $S_2 = \sum_{i \neq \star} \alpha_i^2$. On the indicated region, $a \geq b$, $a \geq 1/P$, and $S_2 \leq a\varepsilon \leq a$. Since $\overline{q} \geq c_\lambda \kappa S_2$ and $q_j \geq c_\lambda \kappa b$,

$$\frac{\dot{g}_j}{w} \geq c_\lambda \kappa[(a-b)S_2 + b^2] \geq c_\lambda \kappa \left(aS_2 - \frac{S_2^2}{4}\right) \geq \frac{3c_\lambda \kappa a S_2}{4} \geq A_D \varepsilon^2. \tag{167}$$

Here the second inequality uses $b^2 - bS_2 \geq -S_2^2/4$, and the last uses $S_2 \geq \varepsilon^2/(P-1)$ and $a \geq 1/P$. □

*Proof of Theorem 7.* First fix one query. If $\kappa_j = 0$, then $Z_j = Z_\star$ almost surely under $D$, so the two candidate relevant coordinates give identical conditional outcome probabilities on the training distribution. They cannot be identified under a common input law. Conversely, if every $\kappa_j > 0$, each competing coordinate disagrees with $Z_\star$ on a set of positive probability; strict monotonicity of the sigmoid makes the corresponding conditional probabilities different. Thus the true coordinate is identifiable. Lemma 42 also shows that the simplex minimizer is unique.

Each score row evolves independently. The gradient fields are smooth with bounded derivatives because the input space is finite and all coordinates of $Z$ are bounded; in particular, the vector fields are globally Lipschitz and bounded. Gradient flow therefore has a unique global solution. At zero initialization all gaps vanish. On any boundary $g_j = 0$ of the region where all gaps are nonnegative, Lemma 42 gives $\dot{g}_j > 0$. The region is consequently positively invariant. For gradient descent, its exact gap update is the gap derivative in that lemma multiplied by $\eta_{\mathrm{gd}}$, evaluated at the current iterate.

Hence every nonnegative gap strictly increases at each finite iterate, for every positive step size. This proves score ordering for both algorithms without any symmetry assumption on $D$.

Under gradient flow, the softmax identity gives $\dot{\alpha}_\star = \alpha_\star \sum_{j \neq \star} \alpha_j \dot{g}_j$. With $A_D$ from Lemma 42,

$$\dot{\varepsilon} \leq -\frac{wA_D}{P}\varepsilon^3, \tag{168}$$

so $\varepsilon(t)^{-2} \geq \varepsilon(0)^{-2} + 2wA_D t/P$. To analyze gradient descent, write $R = \sum_{j \neq \star} e^{-g_j} = \varepsilon/\alpha_\star$ and let primes denote the next iterate. Every gap increment is at least $\eta_{\mathrm{gd}} w A_D \varepsilon^2$, so

$$R' \leq R e^{-\eta_{\mathrm{gd}} w A_D \varepsilon^2}, \ \varepsilon' \leq \frac{\varepsilon}{\varepsilon + \alpha_\star e^{\eta_{\mathrm{gd}} w A_D \varepsilon^2}} \leq \frac{\varepsilon}{1 + \alpha_\star \eta_{\mathrm{gd}} w A_D \varepsilon^2}. \tag{169}$$

The last inequality uses $e^x \geq 1 + x$. Consequently $(\varepsilon')^{-2} - \varepsilon^{-2} \geq 2\eta_{\mathrm{gd}} w A_D \alpha_\star \geq 2\eta_{\mathrm{gd}} w A_D/P$. Summing the increments, substituting $\varepsilon(0) = (P-1)/P$ and $c_\lambda = 4\lambda^2 m_\lambda$, and restoring the query index proves the rate in the theorem. In particular $\varepsilon_u \to 0$ for every query.

For every test vector $z$, the bound $|\alpha^\top z - z_\star| \leq 2\varepsilon$ yields $\mathrm{TV} \leq \lambda\varepsilon/2$ by the $1/4$-Lipschitz property of the sigmoid and $D_{\mathrm{KL}} \leq \lambda^2\varepsilon^2/2$ by the upper Bregman bound used above. Averaging also gives $\mathcal{L}_D(B) - \inf \mathcal{L}_D \leq (\lambda^2/2) \sum_u w_u \varepsilon_u^2$. Assumption 4 identifies these predictions with the conditional outcome kernels at every supported planning context. Proposition 36 and Lemma 37 therefore give $\mathrm{SubOpt}_H \leq \lambda H(H+1) \max_u \varepsilon_u/2$. □

**Corollary 43** (A logarithmic fixed input design)**.** *For every $P \geq 2$, there is a collection of $T = \lceil \log_2 P \rceil$ binary input vectors, chosen independently of the relevant-coordinate map, such that its uniform distribution satisfies $\kappa_u \geq 1/T$ for every query and every relevant coordinate. Hence Theorem 7 applies. Together with Proposition 23, this matches the minimum number of fixed input vectors needed for relevant-coordinate identification from population conditional probabilities.*

*Proof.* Choose $P$ distinct codewords in $\{-1, 1\}^T$ and place them as columns of a $T$-by-$P$ matrix. Use its rows as the training vectors. Every two columns disagree in at least one row, so their disagreement probability under the uniform indexed-row distribution is at least $1/T$. The construction uses no relevant-coordinate information. If rows coincide, their multiplicities can be retained, as in Assumption 5; in fact the lower bound in Proposition 23 prevents a smaller set of distinct rows from distinguishing all $P$ columns. The theorem applies for all query weights. The lower bound concerns identification only and does not assert sample or optimization optimality. □

### R.2 Finite samples at arbitrary inputs

For every query $u$, fix $N_u \geq 1$ inputs $z_{u,1}, \ldots, z_{u,N_u}$ and draw independent labels with $\Pr(Y_{u,i} = +1) = p_{u,i} := P_\nu(+1 \mid u, z_{u,i})$. Labels are conditionally independent within each query; dependence between different queries is allowed. Define $\widetilde{Y}_{u,i} = \mathbf{1}\{Y_{u,i} = +1\}$ and

$$\widehat{F}_u(\alpha) = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathrm{ce}_{\mathrm{bin}}\Big(\widetilde{Y}_{u,i}, \mathrm{sigmoid}(\lambda \alpha^\top z_{u,i})\Big), \tag{170}$$

with empirical loss $\widehat{\mathcal{L}}_D(B) = \sum_u w_u \widehat{F}_u(\alpha(B;u))$. Write $\widehat{\kappa}_u = \min_{j \neq j_\star(u)} N_u^{-1} \sum_i \mathbf{1}\{z_{u,i,j} \neq z_{u,i,j_\star(u)}\}$.

**Theorem 44** (Empirical gradient flow at arbitrary inputs)**.** *Let $\delta \in (0,1)$ and define*

$$d_u = \lambda \sqrt{\frac{\log(2Pn_{\mathrm{q}}/\delta)}{2N_u}}, \ r_u = \frac{32P(P-1)d_u}{3c_\lambda \widehat{\kappa}_u}. \tag{171}$$

*Suppose $\widehat{\kappa}_u > 0$ and $r_u \leq 1/2$ for every query. With probability at least $1 - \delta$, empirical gradient flow on $\widehat{\mathcal{L}}_D$ from zero exists globally, preserves $\alpha_{j_\star(u)} \geq \alpha_j$ for every $u, j$, and satisfies*

$$\varepsilon_u(t) \leq \max\left\{ \left[ \frac{P^2}{(P-1)^2} + \frac{3w_u c_\lambda \widehat{\kappa}_u t}{4P^2(P-1)} \right]^{-1/2}, r_u \right\}. \tag{172}$$

*The uniform total-variation, Kullback–Leibler, and planning bounds of Theorem 7 hold on this event. Its deterministic conclusion holds for every positive choice of query weights simultaneously.*

*Proof.* Let $F_u^0$ be the loss $\widehat{F}_u$ with each label $\widetilde{Y}_{u,i}$ replaced by its mean $p_{u,i}$. Direct differentiation shows that

$$\nabla_\alpha \widehat{F}_u(\alpha) - \nabla_\alpha F_u^0(\alpha) = -\frac{\lambda}{N_u} \sum_{i=1}^{N_u} (\widetilde{Y}_{u,i} - p_{u,i}) z_{u,i} =: e_u. \tag{173}$$

In particular, this perturbation is independent of $\alpha$. For each coordinate, the summands before multiplication by $\lambda/N_u$ are independent centered variables with intervals of length one. Hoeffding's inequality and a union bound over queries and coordinates give $\max_j |e_{u,j}| \le d_u$ for every $u$ with probability at least $1-\delta$. No independence between queries is needed. Work on this event and fix one row. The smooth bounded gradient field gives a unique global solution.

Use the exact-target gap equation of Lemma 42 for the empirical input distribution, with $\widehat{\kappa} = \widehat{\kappa}_u$ and $A = 3c_\lambda \widehat{\kappa}/[4P(P-1)]$. Write $a = \alpha_\star$, $b = \alpha_j$, and $\overline{e} = \sum_i \alpha_i e_i$. The error in the gap derivative, divided by $w = w_u$, is $b(e_j - \overline{e}) - a(e_\star - \overline{e})$. Since $|e_\star - \overline{e}| \le 2d\varepsilon$, $|e_j - \overline{e}| \le 2d$, and $b \le \varepsilon$, its absolute value is at most $4d\varepsilon$. On the region where $a \ge \alpha_j$ for all irrelevant coordinates, Lemma 42 therefore gives

$$\dot{g}_j \ge w(A\varepsilon^2 - 4d\varepsilon) \ge \frac{wA}{2}\varepsilon^2 \text{ whenever } \varepsilon \ge r := \frac{8d}{A}. \tag{174}$$

Starting from zero, a first-crossing argument preserves nonnegative gaps while $\varepsilon \ge r$. In this region $a \ge 1/P$, so $\dot{\varepsilon} \le -wA\varepsilon^3/(2P)$. Integrating until the first entry into $\{\varepsilon \le r\}$ gives the stated inverse-square-root bound.

Since $r \le 1/2$, every point on the boundary $\varepsilon = r$ has $a = 1 - r \ge \alpha_j$ for all irrelevant coordinates. The displayed gap bound yields $\dot{\varepsilon} < 0$ on that boundary, so this set is positively invariant. Within it $a \ge 1/2$ and $\alpha_j \le \varepsilon \le 1/2$, which preserves score ordering. If equality is possible at $r = 1/2$, the gap derivative at equality is strictly positive. Thus the trajectory cannot cross into the opposite ordering. The resulting global bound is the maximum of the transient bound and $r$. Restoring indices proves the assertion. The prediction and planning statements follow from the last paragraph of the proof of Theorem 7. The event controlling $e_u$ contains no query weights, and the row inequalities merely multiply by $w_u > 0$, proving the last assertion. □

### R.3 Independent exploratory episodes

Now collect $N_{\mathrm{ep}}$ independent length-$H$ episodes under the protocol of Section 2: each episode independently draws a task $e \sim \nu$ and follows the same fixed task-independent behavior policy $\beta$. The policy may depend on the history within an episode. For $u = (k, a)$, define $\rho_u = \Pr\{A_k = a\} > 0$ and $D_u = \mathrm{Law}(z(H_k) \mid A_k = a)$, and use the corresponding $\kappa_u$ from Theorem 7. Here $H_k$ denotes the random decision history. Pool the samples having query $u$, with count $N_u$. Each episode contributes at most one sample to this query. Let $\widehat{\mathcal{L}}_D$ use these samples with any positive query weights. In particular, the empirical loss that weights every observed stage equally has $w_u = N_u/(HN_{\mathrm{ep}})$ whenever all counts are nonzero.

**Corollary 45** (Prediction and planning from independent episodes)**.** *Assume $\kappa_u > 0$ for every query. Fix $\delta \in (0,1)$, write $L_{\mathrm{ep}} = \log(3Pn_{\mathrm{q}}/\delta)$, and suppose*

$$N_{\mathrm{ep}}\rho_u \ge \frac{16L_{\mathrm{ep}}}{\kappa_u} \text{ for every } u. \tag{175}$$

*Define $R_{\mathrm{ep},u} = 16P(P-1)(3\lambda m_\lambda \kappa_u)^{-1}\sqrt{L_{\mathrm{ep}}/(N_{\mathrm{ep}}\rho_u)}$. If also $R_{\mathrm{ep},u} \le 1/2$ for every $u$, then with probability at least $1-\delta$, all counts satisfy $N_u \ge N_{\mathrm{ep}}\rho_u/2$, and empirical gradient flow from zero satisfies*

$$\varepsilon_u(t) \le \max\left\{ \left[ \frac{P^2}{(P-1)^2} + \frac{3w_u c_\lambda \kappa_u t}{8P^2(P-1)} \right]^{-1/2}, R_{\mathrm{ep},u} \right\}. \tag{176}$$

*The same event gives the uniform prediction and planning bounds of Theorem 7. For the ordinary pooled empirical loss, $w_u \ge \rho_u/(2H)$. In particular, the hypotheses hold and $\limsup_{t\to\infty} \mathrm{SubOpt}_H(\widehat{\pi}_t) \le \varepsilon_{\mathrm{val}}$ whenever, for every query,*

$$N_{\mathrm{ep}}\rho_u \ge \max\left\{ \frac{16L_{\mathrm{ep}}}{\kappa_u}, \frac{1024P^2(P-1)^2L_{\mathrm{ep}}}{9\lambda^2 m_\lambda^2 \kappa_u^2}, \frac{64P^2(P-1)^2H^2(H+1)^2L_{\mathrm{ep}}}{9m_\lambda^2\kappa_u^2\varepsilon_{\mathrm{val}}^2} \right\}. \tag{177}$$

*Here $\widehat{\pi}_t$ is any exact dynamic-programming policy for the resulting predictive kernel.*

*Proof.* For each fixed query, $N_u$ is binomial with mean $N_{\mathrm{ep}}\rho_u$. The multiplicative Chernoff inequality gives $\Pr\{N_u < N_{\mathrm{ep}}\rho_u/2\} \le \exp(-N_{\mathrm{ep}}\rho_u/8)$. Conditional on the set of episode indices selected by $A_k = a$, their stage-$k$ feature vectors are independent with law $D_u$. Conditional on $N_u = n$, the disagreement count for coordinate $j$ is binomial with mean $n\kappa_{u,j}$, where $\kappa_{u,j} = \Pr_{D_u}\{Z_j \neq Z_{j_\star(u)}\} \ge \kappa_u$. Applying the same inequality, on counts at least $N_{\mathrm{ep}}\rho_u/2$, bounds the probability of an empirical disagreement frequency below $\kappa_{u,j}/2$ by $\exp(-N_{\mathrm{ep}}\rho_u\kappa_{u,j}/16)$. The stated sample condition makes both bounds at most $e^{-L_{\mathrm{ep}}}$. A union bound over counts and irrelevant coordinates thus has failure probability at most $Pn_{\mathrm{q}}e^{-L_{\mathrm{ep}}} = \delta/3$. On its complement, $N_u \ge N_{\mathrm{ep}}\rho_u/2$ and $\widehat{\kappa}_u \ge \kappa_u/2$ for every query.

For label concentration, fix one query and condition only on its selected histories and actions before their outcomes are produced. The selection event $A_k = a$ is determined before the current outcome, and episodes are independent. Proposition 1 and Assumption 4 therefore make the selected outcome indicators conditionally independent Bernoulli variables with means $P_\nu(+1 \mid u, z(H_k))$. The perturbation formula in the proof of Theorem 44 applies to this query. With $d_u = \lambda\sqrt{L_{\mathrm{ep}}/(2N_u)}$, its coordinate-wise Hoeffding bound has failure probability at most $2Pe^{-L_{\mathrm{ep}}}$ whenever $N_u \ge 1$. Applying these conditional bounds separately and then taking a union bound over queries gives failure probability at most $2Pn_{\mathrm{q}}e^{-L_{\mathrm{ep}}} = 2\delta/3$. This argument does not condition simultaneously on later-stage histories, which can contain earlier outcomes; it requires no independence between query datasets.

On the intersection of the two events, whose probability is at least $1-\delta$, we have $d_u \le \lambda\sqrt{L_{\mathrm{ep}}/(N_{\mathrm{ep}}\rho_u)}$ and $\widehat{\kappa}_u \ge \kappa_u/2$. Substitution into the deterministic proof of Theorem 44 gives $r_u \le R_{\mathrm{ep},u}$ and the transient rate in the corollary. That proof holds for all positive weights, including the weights computed from the observed counts. The count bound yields $w_u \ge \rho_u/(2H)$ for the pooled empirical loss. The uniform prediction and planning conclusions follow as before.

Finally, $R_{\mathrm{ep},u} \le 1/2$ is equivalent to the second sample requirement in the maximum. Multiplying $R_{\mathrm{ep},u}$ by $\lambda H(H+1)/2$ gives $8P(P-1)H(H+1)(3m_\lambda\kappa_u)^{-1}\sqrt{L_{\mathrm{ep}}/(N_{\mathrm{ep}}\rho_u)}$. Requiring this quantity to be at most $\varepsilon_{\mathrm{val}}$ gives the third requirement. This proves the explicit asymptotic value guarantee. □

The episode result concerns gradient flow on the fixed empirical loss formed after data collection. It allows nonuniform query visitation and nonuniform feature distributions without conditional generative-model queries. It does not assume independence of successive observations within an episode, and it does not analyze stochastic-gradient updates or behavior policies adapted across episodes.

# S Initialization, Walsh Expansions, and the Attention Jacobian

This section proves Proposition 13 and Theorem 14. It concerns a fixed query and binary input tokens, and uses the direct score-matrix parameterization of the main model. We first derive the local expansion, then prove nonasymptotic lower bounds. Finite-width population and empirical training are treated separately in Appendix T.

## S.1 Input representation and kernel normalization

Fix an integer $P \ge 1$, put $\ell = P+1$ and $d = 2P+1$, and let

$$\{e_0\} \cup \{e_{j,b} : 1 \le j \le P,\ b \in \{-1,1\}\} \tag{178}$$

be an orthonormal basis of $\mathbb{R}^d$. For $z \in \{-1,1\}^P$, the sequence consists of $x_j(z) = e_{j,z_j}$ for $1 \le j \le P$, followed by $x_0(z) = u = e_0$. The subscript zero denotes the fixed final query, rather than its sequence position. We include this query among the attended tokens. Thus

$$\alpha_j(A,z) = \frac{\exp(u^\top A x_j(z))}{\sum_{i=0}^{P}\exp(u^\top A x_i(z))},\ s(A,z) = \sum_{j=0}^{P}\alpha_j(A,z)x_j(z), \tag{179}$$

and $\varphi_z(A, v) = v^\top s(A, z)$. The parameter norm is $\|A\|_F^2 + \|v\|_2^2$. Define the attention-weighted covariance

$$C(A, z) = \sum_{j=0}^{P} \alpha_j(A, z) x_j(z) x_j(z)^\top - s(A, z) s(A, z)^\top. \tag{180}$$

The softmax derivative gives, for every $A$ and $v$,

$$\nabla_A \varphi_z(A, v) = u(C(A, z)v)^\top, \ \nabla_v \varphi_z(A, v) = s(A, z). \tag{181}$$

Indeed, differentiating the summary in direction $U$ gives $D_A s(A, z)[U] = C(A, z)U^\top u$, and pairing with $v$ yields the first identity.

For $\xi_{\mathrm{sc}} \in \{-1, 1\}^P$ and real $\tau$, set

$$A^\tau(\xi_{\mathrm{sc}}) = \tau e_0 \left( \sum_{j=1}^{P} \xi_{\mathrm{sc},j} (e_{j,1} - e_{j,-1}) \right)^\top. \tag{182}$$

Write $s_\tau(z; \xi_{\mathrm{sc}}) = s(A^\tau(\xi_{\mathrm{sc}}), z)$ and $C_\tau(z; \xi_{\mathrm{sc}}) = C(A^\tau(\xi_{\mathrm{sc}}), z)$. At this initialization,

$$D_\tau(z; \xi_{\mathrm{sc}}) = 1 + \sum_{j=1}^{P} e^{\tau \xi_{\mathrm{sc},j} z_j}, \ \alpha_0 = \frac{1}{D_\tau(z; \xi_{\mathrm{sc}})}, \ \alpha_j = \frac{e^{\tau \xi_{\mathrm{sc},j} z_j}}{D_\tau(z; \xi_{\mathrm{sc}})}. \tag{183}$$

Throughout this section, $\xi_{\mathrm{sc}}$ is uniform on the binary cube, $v$ is independent of $\xi_{\mathrm{sc}}$, and $\mathbb{E}[vv^\top] = \sigma_v^2 I_d$ with $\sigma_v > 0$. In particular, the value initialization in the main model has this property. Define the expected scalar-head kernels

$$\begin{aligned} k_A^\tau(z, z') &= \mathbb{E}_{\xi_{\mathrm{sc}}, v} \langle \nabla_A \varphi_z(A^\tau(\xi_{\mathrm{sc}}), v), \nabla_A \varphi_{z'}(A^\tau(\xi_{\mathrm{sc}}), v) \rangle_F \\ &= \sigma_v^2 \mathbb{E}_{\xi_{\mathrm{sc}}} \langle C_\tau(z; \xi_{\mathrm{sc}}), C_\tau(z'; \xi_{\mathrm{sc}}) \rangle_F, \\ k_v^\tau(z, z') &= \mathbb{E}_{\xi_{\mathrm{sc}}} \langle s_\tau(z; \xi_{\mathrm{sc}}), s_\tau(z'; \xi_{\mathrm{sc}}) \rangle_2, \\ k_{\mathrm{full}}^\tau(z, z') &= k_A^\tau(z, z') + k_v^\tau(z, z'). \end{aligned} \tag{184}$$

The covariance identity uses $\|u\|_2 = 1$ and the isotropic second moment of $v$.

For real-valued functions on the cube, use the inner product $\langle g, h \rangle = 2^{-P} \sum_z g(z) h(z)$. For $B \in \{A, v, \mathrm{full}\}$, the corresponding operator is

$$(\mathcal{T}_B^\tau g)(z) = 2^{-P} \sum_{z'} k_B^\tau(z, z') g(z'). \tag{185}$$

These operator eigenvalues equal the eigenvalues of the finite Gram matrix divided by $2^P$. For a subset $S \subseteq \{1, \dots, P\}$, put $\chi_S(z) = \prod_{j \in S} z_j$, with $\chi_\varnothing = 1$. The $2^P$ Walsh functions are an orthonormal basis: $\langle \chi_S, \chi_T \rangle$ is one when $S = T$ and zero otherwise.

The relation to the main model's binary logits is as follows. Each pair has logits scaled by $1/\sqrt{2M}$, shares its initialization between the two copies, and has an independent Rademacher output vector $b_r \in \{-1, 1\}^2$. The two copies contribute

$$\frac{1}{M} b_r b_r^\top \langle \nabla \varphi_z(\omega_r^0), \nabla \varphi_{z'}(\omega_r^0) \rangle \tag{186}$$

to the full logit kernel block. Since $\mathbb{E}[b_r b_r^\top] = I_2$, the expected block is $k_{\mathrm{full}}^\tau(z, z') I_2$. For the scalar logit difference $F(z) = f_1(z) - f_2(z)$, the expected kernel is therefore $2k_{\mathrm{full}}^\tau(z, z')$. Its normalized operator has eigenvalues $2(\lambda_s^A(\tau) + \lambda_s^v(\tau))$. The same factor of two applies to the score contribution alone. Pairing guarantees $F(z) = 0$ at initialization for every $\tau$.

### S.2 Symmetries and Fourier coefficients

**Lemma 46** (Walsh diagonalization)**.** *For every real $\tau$, both $\mathcal{T}_A^\tau$ and $\mathcal{T}_v^\tau$ are diagonal in the Walsh basis. Their eigenvalues depend only on $s = |S|$. If*

$$\widehat{C}_\tau(S; \xi_{\mathrm{sc}}) = 2^{-P} \sum_z \chi_S(z) C_\tau(z; \xi_{\mathrm{sc}}), \ \widehat{s}_\tau(S; \xi_{\mathrm{sc}}) = 2^{-P} \sum_z \chi_S(z) s_\tau(z; \xi_{\mathrm{sc}}), \tag{187}$$

*then*

$$\lambda_s^A(\tau) = \sigma_v^2 \mathbb{E}_{\xi_{\rm sc}} \|\widehat{C}_\tau(S;\xi_{\rm sc})\|_F^2,\ \lambda_s^v(\tau) = \mathbb{E}_{\xi_{\rm sc}} \|\widehat{s}_\tau(S;\xi_{\rm sc})\|_2^2. \tag{188}$$

*All these eigenvalues are nonnegative, real analytic in $\tau$, and even in $\tau$.*

*Proof.* For $\epsilon \in \{-1,1\}^P$, let $U_\epsilon$ be the orthogonal permutation that fixes $e_0$ and maps $e_{j,b}$ to $e_{j,\epsilon_j b}$. Write $\epsilon z$ and $\epsilon\xi_{\rm sc}$ for coordinatewise products. The formulas for the attention weights imply

$$s_\tau(\epsilon z;\xi_{\rm sc}) = U_\epsilon s_\tau(z;\epsilon\xi_{\rm sc}),$$
$$C_\tau(\epsilon z;\xi_{\rm sc}) = U_\epsilon C_\tau(z;\epsilon\xi_{\rm sc}) U_\epsilon^\top.$$

Orthogonal invariance of the inner products and invariance of the law of $\xi_{\rm sc}$ under $\xi_{\rm sc} \mapsto \epsilon\xi_{\rm sc}$ give $k_B^\tau(\epsilon z, \epsilon z') = k_B^\tau(z,z')$ for $B = A, v$. Hence the kernel depends only on $zz'$. Substituting $z' = zw$ into the operator shows that each $\chi_S$ is an eigenfunction. Permuting the $P$ positions simultaneously in $z, z', \xi_{\rm sc}$ and in the embedding basis preserves the kernel as well. Any two subsets of equal size are related by a position permutation, so their eigenvalues agree.

Because $\|\chi_S\|_2 = 1$, its eigenvalue is $\langle \chi_S, \mathcal{T}_B^\tau \chi_S\rangle$. Substituting the kernel formulas and interchanging the finite sums gives the two squared-norm identities, and also proves nonnegativity. Each denominator $D_\tau(z;\xi_{\rm sc})$ is strictly positive for real $\tau$. All entries are therefore real analytic, and finite sums preserve real analyticity. Finally, replacing $\tau$ by $-\tau$ is equivalent to replacing $\xi_{\rm sc}$ by $-\xi_{\rm sc}$, which preserves its distribution. The eigenvalues are even. □

### S.3 Zero scores and exact stationarity

*Proof of Proposition 13.* First allow any fixed linear transformation of the shared one-hot embeddings. The final query remains a fixed vector $u$, and each variable token has the form $x_j(z) = p_j + z_j r_j$ for fixed vectors $p_j, r_j$. At $A = 0$, the summary is

$$\overline{x}(z) = \frac{1}{\ell}\left(u + \sum_{j=1}^P (p_j + z_j r_j)\right). \tag{189}$$

Each coordinate belongs to the span of Walsh functions of degree at most one. Moreover,

$$C(0,z) = \frac{1}{\ell}\left(uu^\top + \sum_{j=1}^P x_j(z)x_j(z)^\top\right) - \overline{x}(z)\overline{x}(z)^\top \tag{190}$$

has entries of Walsh degree at most two. The gradient identities above show that every coordinate of the score Jacobian has degree at most two, and every coordinate of the value Jacobian has degree at most one. Fixed output weights, finite sums of heads, and the two signs in a pair preserve these inclusions.

For the expected-rank assertion, return to the orthonormal one-hot representation. Define

$$c_j = \frac{e_{j,1} + e_{j,-1}}{\sqrt{2}},\ b_j = \frac{e_{j,1} - e_{j,-1}}{\sqrt{2}}. \tag{191}$$

Together with $e_0$, these vectors form an orthonormal basis. At zero scores, every attention weight is $1/\ell$. For the constant Fourier coefficient, the $(e_0, e_0)$ entry of the covariance is $P/\ell^2$, which is nonzero. For $S = \{j\}$, the $(e_0, b_j)$ entry is $-z_j/(\sqrt{2}\ell^2)$, so its Fourier coefficient is nonzero. For $S = \{j,k\}$ with $j \neq k$, the $(b_j, b_k)$ entry is $-z_j z_k/(2\ell^2)$, again giving a nonzero Fourier coefficient. Lemma 46 therefore gives strictly positive score eigenvalues in degrees zero, one, and, when $P \geq 2$, two. All higher-degree coefficients vanish by the degree bound. Thus the score operator has range exactly $\mathcal{V}_2$ and rank $1 + P + \binom{P}{2}$. The value operator has range contained in $\mathcal{V}_2$ and is positive semidefinite, so adding it leaves the range and rank unchanged.

Finally, suppose $|S| \geq 3$ and $q(1 \mid z) = 1/2 + \delta_{\rm p}\chi_S(z)$. Paired initialization gives $f_2(z) = f_1(z) = 0$, hence both predicted probabilities are $1/2$. For any trainable scalar parameter $\theta_j$, the derivative of

the population cross-entropy at initialization is

$$\begin{aligned}\partial_{\theta_j}\mathcal{L}(\theta_0) &= 2^{-P}\sum_{z}\sum_{y\in\{1,2\}}\left(\frac{1}{2}-q(y\mid z)\right)\partial_{\theta_j}f_y(z;\theta_0)\\ &= \delta_{\mathrm{p}}2^{-P}\sum_{z}\chi_S(z)\left(\partial_{\theta_j}f_2(z;\theta_0)-\partial_{\theta_j}f_1(z;\theta_0)\right)=0.\end{aligned} \tag{192}$$

The last equality is Walsh orthogonality to $\mathcal{V}_2$. This argument holds for every realization of the finite collection of initial value and output weights. The loss is smooth, so its gradient vector field is locally Lipschitz. The constant trajectory $\theta_t = \theta_0$ is a solution and, by uniqueness, is the population gradient-flow trajectory for all times. Exact population gradient descent also remains at $\theta_0$ by induction, independently of its step-size sequence. Uniform prediction has total variation distance $\delta_{\mathrm{p}}$ from the target at each input. In particular, its excess cross-entropy remains the strictly positive quantity

$$\log 2 - \mathrm{Ent}\left(\left(\frac{1}{2}+\delta_{\mathrm{p}},\frac{1}{2}-\delta_{\mathrm{p}}\right)\right). \tag{193}$$

This stationarity claim uses the exact population objective; random empirical frequencies or stochastic gradients need not preserve the required orthogonality. □

### S.4 The first nonzero eigenvalue at each interaction order

**Lemma 47** (Fourier coefficients of analytic functions)**.** *Let $F$ be real analytic near zero in $\mathbb{R}^P$, fix $\xi_{\mathrm{sc}} \in \{-1,1\}^P$, and let $B, S \subseteq \{1,\dots,P\}$. The Fourier coefficient at $S$ of $\chi_B(z)F(\tau\xi_{\mathrm{sc},1}z_1,\dots,\tau\xi_{\mathrm{sc},P}z_P)$ has a convergent power series near $\tau = 0$ whose powers have the form $|S\triangle B| + 2j$, $j \geq 0$. In particular, that coefficient is $O(\tau^{|S\triangle B|})$.*

*Proof.* Expand $F(t) = \sum_{m\in\mathbb{N}_0^P} c_m \prod_i t_i^{m_i}$ in a convergent multivariate power series. The Fourier coefficient of the term indexed by $m$ vanishes unless $m_i$ is odd exactly at the coordinates in $S \triangle B$. A surviving term therefore has total degree $|S\triangle B| + 2j$ for an integer $j \geq 0$. Since the cube is finite, the series can be averaged term by term in a sufficiently small neighborhood of zero. □

*Proof of Theorem 14.* Walsh diagonalization and the squared-norm formulas follow from Lemma 46. It remains to identify their leading terms. Use the orthonormal basis $e_0, c_1, b_1, \dots, c_P, b_P$ defined above, and set $t_i = \tau\xi_{\mathrm{sc},i}z_i$. The summary coordinates are

$$\langle e_0, s_\tau\rangle = \alpha_0,\ \langle c_j, s_\tau\rangle = \frac{\alpha_j}{\sqrt{2}},\ \langle b_j, s_\tau\rangle = \frac{z_j\alpha_j}{\sqrt{2}}. \tag{194}$$

Here $\alpha_0 = (1+\sum_i e^{t_i})^{-1}$ and $\alpha_j = e^{t_j}(1+\sum_i e^{t_i})^{-1}$ are analytic functions of $t$ near zero.

For completeness, every covariance entry in this basis is listed below. Symmetric entries agree, and the entries with $j \neq k$ are specified separately:

$$\begin{aligned}
C_{00} &= \alpha_0 - \alpha_0^2,\\
C_{0,c_j} &= -\alpha_0\alpha_j/\sqrt{2}, & C_{0,b_j} &= -z_j\alpha_0\alpha_j/\sqrt{2},\\
C_{c_j,c_j} &= (\alpha_j-\alpha_j^2)/2, & C_{b_j,b_j} &= (\alpha_j-\alpha_j^2)/2,\\
C_{c_j,b_j} &= z_j(\alpha_j-\alpha_j^2)/2,\\
C_{c_j,c_k} &= -\alpha_j\alpha_k/2, & C_{c_j,b_k} &= -z_k\alpha_j\alpha_k/2,\ j\neq k,\\
C_{b_j,b_k} &= -z_jz_k\alpha_j\alpha_k/2,\ j\neq k.
\end{aligned} \tag{195}$$

Each entry is an analytic function of $t$ multiplied by a Walsh function of degree at most two. The only entries with a degree-two multiplier are $C_{b_j,b_k}$ for distinct $j, k$.

Fix $S$ of size $s \geq 2$. Lemma 47 shows that a Fourier coefficient of order $\tau^{s-2}$ can occur only in $C_{b_j,b_k}$ with $j,k \in S$, $j \neq k$. All other entries have Fourier coefficients $O(\tau^{s-1})$. For such a pair, set $R = S\setminus\{j,k\}$ and $h = |R| = s-2$. After canceling $z_jz_k$ against the corresponding factors in $\chi_S$,

the coefficient at order $\tau^h$ is determined by the mixed derivative, once in each variable indexed by $R$, of

$$-\frac{1}{2}\frac{e^{t_j+t_k}}{(1+\sum_i e^{t_i})^2} \tag{196}$$

at $t=0$. Because $j,k \notin R$, the numerator is constant for this derivative. Each differentiated variable occurs once, so the chain rule gives

$$\left(\prod_{i\in R}\partial_{t_i}\right)(1+\sum_i e^{t_i})^{-2}\Big|_{t=0} = \frac{(-1)^h(h+1)!}{\ell^{h+2}}. \tag{197}$$

Consequently,

$$[\widehat{C}_\tau(S;\xi_{\mathrm{sc}})]_{b_j,b_k} = \frac{(-1)^{s-1}(s-1)!}{2\ell^s}\left(\prod_{i\in R}\xi_{\mathrm{sc},i}\right)\tau^{s-2} + O_{P,s}(\tau^s). \tag{198}$$

The remainder starts at power $s$, rather than $s-1$, by the parity assertion in Lemma 47. The constants can be taken uniformly in $\xi_{\mathrm{sc}}$ and the indices because these sets are finite at fixed $P$.

The Frobenius norm counts both symmetric entries for each of the $\binom{s}{2}$ unordered pairs. Their leading coefficients have the same squared magnitude, since $\xi_{\mathrm{sc},i}^2 = 1$. All other entries contribute $O_{P,s}(\tau^{2s-2})$ after squaring. It follows that

$$\|\widehat{C}_\tau(S;\xi_{\mathrm{sc}})\|_F^2 = \frac{\binom{s}{2}((s-1)!)^2}{2\ell^{2s}}\tau^{2s-4} + O_{P,s}(\tau^{2s-2}). \tag{199}$$

Averaging in $\xi_{\mathrm{sc}}$ and multiplying by $\sigma_v^2$ proves the stated score-kernel expansion.

For the value kernel, the same calculation also applies at degree one. For $s \geq 1$, only coordinates $b_j$ with $j \in S$ can contribute at order $\tau^{s-1}$. In each such coordinate, differentiating $(1+\sum_i e^{t_i})^{-1}$ once in the $s-1$ variables in $S \setminus \{j\}$ gives

$$[\widehat{s}_\tau(S;\xi_{\mathrm{sc}})]_{b_j} = \frac{(-1)^{s-1}(s-1)!}{\sqrt{2}\ell^s}\left(\prod_{i\in S\setminus\{j\}}\xi_{\mathrm{sc},i}\right)\tau^{s-1} + O_{P,s}(\tau^{s+1}). \tag{200}$$

The remaining summary coordinates have Fourier coefficients $O_{P,s}(\tau^s)$. Summing the squared coordinates and averaging gives

$$\lambda_s^v(\tau) = \frac{s((s-1)!)^2}{2\ell^{2s}}\tau^{2s-2} + O_{P,s}(\tau^{2s}),\ 1 \leq s \leq P. \tag{201}$$

In particular, the full scalar kernel has the same leading term as the score kernel for every $s \geq 2$.

The proof of Proposition 13 already gives $\lambda_0^A(0) > 0$ and $\lambda_1^A(0) > 0$. They remain positive in a neighborhood of zero by continuity. For each $2 \leq s \leq P$, the leading coefficient in the score expansion is strictly positive. Hence there exists a positive radius on which that eigenvalue is positive at every nonzero $\tau$. Taking the minimum of these finitely many radii proves the existence of $\tau_0(P) > 0$. The radius can be chosen independently of $\sigma_v > 0$ because every score eigenvalue has the same multiplicative factor $\sigma_v^2$. When $P = 1$, positivity follows directly from the degree-zero and degree-one eigenvalues. This establishes positivity near zero. Proposition 49 below proves the stronger claim in the theorem for every nonzero score scale, for both kernels. □

*Remark* 48 (Scope of the spectral expansion). The expansions hold for fixed $P, s$, and $\sigma_v$ as $\tau \to 0$ and are not uniform in growing $P$; the following nonasymptotic bounds remove this restriction. Expected-kernel positive definiteness alone does not imply full rank for a finite collection of heads, so the finite-width results below instead use a target-dependent comparison point and control nonlinear parameter movement. Exact zero-score stationarity continues to require paired initialization and the population objective.

### S.5 NONASYMPTOTIC EIGENVALUE BOUNDS

The power series above identifies the first nonzero term. A Laplace integral representation gives a lower bound without a small-initialization restriction. This distinction matters when the dimension or interaction order varies.

**Proposition 49** (Eigenvalue bounds at every nonzero score scale)**.** *Use the representation and normalization of Appendix S. For every $\tau \neq 0$ and $2 \leq s \leq P$,*

$$\lambda_s^A(\tau) \geq \frac{\sigma_v^2 \binom{s}{2}((s-1)!)^2 e^{-4|\tau|} \sinh^{2s-4}(|\tau|)}{2(1+Pe^{|\tau|})^{2s}}. \tag{202}$$

*For every $1 \leq s \leq P$,*

$$\lambda_s^v(\tau) \geq \frac{s((s-1)!)^2 e^{-2|\tau|} \sinh^{2s-2}(|\tau|)}{2(1+Pe^{|\tau|})^{2s}}. \tag{203}$$

*The degree-zero eigenvalues and the degree-one score eigenvalue are strictly positive at every real $\tau$. Consequently, each expected kernel is positive definite for every $\tau \neq 0$. The bounds hold simultaneously for every finite $P$ and every indicated order $s$.*

*Proof.* Evenness permits us to take $\tau > 0$. The sign-change identity in Lemma 46 gives

$$\begin{aligned} \widehat{C}_\tau(S;\xi_{\mathrm{sc}}) &= \chi_S(\xi_{\mathrm{sc}}) U_{\xi_{\mathrm{sc}}} \widehat{C}_\tau(S;\mathbf{1}) U_{\xi_{\mathrm{sc}}}^\top, \\ \widehat{s}_\tau(S;\xi_{\mathrm{sc}}) &= \chi_S(\xi_{\mathrm{sc}}) U_{\xi_{\mathrm{sc}}} \widehat{s}_\tau(S;\mathbf{1}). \end{aligned} \tag{204}$$

Their squared norms are therefore independent of $\xi_{\mathrm{sc}}$. It suffices to calculate at $\xi_{\mathrm{sc}} = \mathbf{1}$. For an integration variable $x \geq 0$, define

$$\begin{aligned} m_0(x) &= \tfrac{1}{2}(e^{-xe^{\tau}} + e^{-xe^{-\tau}}), \\ m_1(x) &= \tfrac{1}{2}(e^{-xe^{\tau}} - e^{-xe^{-\tau}}), \\ m_e(x) &= \tfrac{1}{2}(e^{\tau}e^{-xe^{\tau}} + e^{-\tau}e^{-xe^{-\tau}}). \end{aligned} \tag{205}$$

Here $m_0, m_e$ are positive and $m_1(x) < 0$ for $x > 0$. For distinct $j, k \in S$, the covariance entry $C_{b_j,b_k} = -z_j z_k \alpha_j \alpha_k / 2$ and the identity $D^{-2} = \int_0^\infty x e^{-xD}\,\mathrm{d}x$ give

$$[\widehat{C}_\tau(S;\mathbf{1})]_{b_j,b_k} = -\frac{1}{2}\int_0^\infty x e^{-x} m_e(x)^2 m_1(x)^{s-2} m_0(x)^{P-s}\,\mathrm{d}x. \tag{206}$$

Indeed, the two factors $z_j, z_k$ cancel the corresponding factors of $\chi_S$. The numerator contributes $e^{\tau z_j} e^{\tau z_k}$; the remaining coordinates in $S$ contribute $m_1$, and the coordinates outside $S$ contribute $m_0$. The input sum is finite, and all its integrals are absolutely convergent, justifying the interchange. The integrand has a constant sign, so there is no cancellation in its absolute integral.

For every $x \geq 0$,

$$m_e(x) \geq e^{-\tau} e^{-xe^{\tau}},\ m_0(x) \geq e^{-xe^{\tau}},\ |m_1(x)| \geq x \sinh(\tau) e^{-xe^{\tau}}. \tag{207}$$

The last inequality follows by integrating $xe^{-xu}$ over $u \in [e^{-\tau}, e^{\tau}]$ and dividing by two. Hence

$$\begin{aligned} |[\widehat{C}_\tau(S;\mathbf{1})]_{b_j,b_k}| &\geq \frac{e^{-2\tau}\sinh^{s-2}(\tau)}{2}\int_0^\infty x^{s-1} e^{-x(1+Pe^{\tau})}\,\mathrm{d}x \\ &= \frac{e^{-2\tau}(s-1)!\sinh^{s-2}(\tau)}{2(1+Pe^{\tau})^s}. \end{aligned} \tag{208}$$

There are $2\binom{s}{2}$ such ordered covariance entries. Squaring, summing them, and multiplying by $\sigma_v^2$ proves the score bound.

For $j \in S$, the summary entry is $\langle b_j, s_\tau \rangle = z_j \alpha_j / \sqrt{2}$. Using $D^{-1} = \int_0^\infty e^{-xD}\,\mathrm{d}x$ instead yields

$$[\widehat{s}_\tau(S;\mathbf{1})]_{b_j} = \frac{1}{\sqrt{2}}\int_0^\infty e^{-x} m_e(x) m_1(x)^{s-1} m_0(x)^{P-s}\,\mathrm{d}x. \tag{209}$$

The same sign argument and pointwise bounds give

$$|[\widehat{s}_\tau(S;\mathbf{1})]_{b_j}| \geq \frac{e^{-\tau}(s-1)!\sinh^{s-1}(\tau)}{\sqrt{2}(1+Pe^{\tau})^s}. \tag{210}$$

Summing the squares over the $s$ choices of $j$ proves the value bound.

Finally, $C_{00} = \alpha_0(1-\alpha_0) > 0$ gives a nonzero degree-zero covariance coefficient. For $S = \{j\}$, the $(e_0, b_j)$ covariance coefficient equals $-\mathbb{E}_z[\alpha_0\alpha_j]/\sqrt{2} < 0$. The degree-zero summary has positive $e_0$ coordinate. These observations prove the remaining positivity claims. The bounds used only finite sums and pointwise inequalities; none uses an asymptotic remainder or a restriction on growing $P$. □

### S.6 Interaction order and the cost of a kernel guarantee

Define the explicit lower bound on the binary-margin eigenvalue by

$$\underline{\kappa}_s(\tau) = \frac{\sigma_v^2\binom{s}{2}((s-1)!)^2 e^{-4|\tau|}\sinh^{2s-4}(|\tau|)}{(1+Pe^{|\tau|})^{2s}},\ 2 \leq s \leq P. \tag{211}$$

Thus $\kappa_s \geq \underline{\kappa}_s(\tau) > 0$ for every $\tau \neq 0$. Since $1 + Pe^u \leq (P+1)e^u$ and $\sinh u \geq u$ for $u \geq 0$, this further implies

$$\kappa_s \geq \frac{\sigma_v^2\binom{s}{2}((s-1)!)^2}{(P+1)^{2s}} e^{-(2s+4)|\tau|}|\tau|^{2s-4}. \tag{212}$$

In particular, if $0 < |\tau| \leq (2s+4)^{-1}$, the exponential factor is at least $e^{-1}$. Unlike the Taylor remainder, this lower bound is explicit and simultaneous in dimension and order.

**Proposition 50** (Trace constraint)**.** *For every real score scale $\tau$, the normalized expected binary-margin kernel satisfies*

$$\sum_{j=0}^{P}\binom{P}{j}\kappa_j = \mathbb{E}_z K_\infty(z,z) \leq 2(\sigma_v^2+1),\ \kappa_s \leq \frac{2(\sigma_v^2+1)}{\binom{P}{s}}. \tag{213}$$

*Whenever $\kappa_s > 0$, the squared reproducing-kernel Hilbert space norm of $a_{\mathrm{p}}\chi_S$ is $a_{\mathrm{p}}^2/\kappa_s$.*

*Proof.* The weighted covariance is positive semidefinite and has trace $\sum_j \alpha_j\|x_j\|^2 - \|s\|^2 \leq 1$. Thus $\|C\|_F^2 \leq (\operatorname{tr} C)^2 \leq 1$, while $\|s\|^2 \leq 1$. The diagonal kernel identity in Appendix S gives the upper bound. The Walsh basis has $\binom{P}{j}$ eigenfunctions of order $j$; taking the operator trace gives the equality. Nonnegativity of every eigenvalue gives the individual bound. The Hilbert space norm follows by expanding in the normalized eigenbasis. □

For $2 \leq s \leq P$ and $\tau \neq 0$, every sufficient condition in Theorem 15 and Theorem 51 remains valid when $\kappa_s$ is replaced throughout by $\underline{\kappa}_s(\tau)$. This gives bounds involving only the displayed initialization scale, dimension, order, target amplitude, and accuracy parameters. For fixed $P, s, \sigma_v, \rho_{\mathrm{p}}, \delta, \varepsilon$ and a bounded initialization perturbation, these sufficient choices have training time and sample size $O(|\tau|^{4-2s})$ and width $O(|\tau|^{8-4s})$ as $\tau \to 0$, for $s \geq 3$. This scaling describes the sufficient guarantees only and does not give a lower bound on nonlinear training time. The dimension dependence remains explicit in equation (212) and in the derivative constants.

A limitation independent of $\tau$ follows from the trace constraint. At $s = \lfloor P/2 \rfloor$, the multiplicity $\binom{P}{s}$ is of order $2^P/\sqrt{P}$. For fixed $\sigma_v$ and nonzero $a_{\mathrm{p}}$, Proposition 50 therefore forces an exponentially large target kernel norm. Hence the kernel-based sufficient guarantees can become exponential at growing interaction order, although this does not constitute a computational lower bound for the nonlinear network. Finally, the exact leading coefficients give

$$\frac{\lambda_s^A(\tau)}{\lambda_s^v(\tau)} = \frac{\sigma_v^2(s-1)}{2\tau^2}\big(1 + O_{P,s,\sigma_v}(\tau^2)\big),\ 2 \leq s \leq P. \tag{214}$$

Thus the score and value contributions have different small-scale dependence within the same architecture; the comparison does not require changing its interaction mechanism.

# T FINITE-WIDTH LEARNING FROM A TARGET EIGENVALUE

This section proves Theorem 15. All expectations over inputs use the uniform distribution on $\{-1,1\}^P$. For a real-valued function $g$, write $\|g\|_2^2 = \mathbb{E}_z[g(z)^2]$ and $\langle g,h\rangle_2 = \mathbb{E}_z[g(z)h(z)]$. In particular, $\|\chi_S\|_2 = 1$. These normalized norms do not introduce a factor equal to the number of inputs.

## T.1 CONSTANTS FOR THE POPULATION GUARANTEE

For Theorem 15, $d = 2P+1$ and the unit-norm token representation give

$$\begin{aligned} C_v &= \sigma_v\sqrt{d}, & C_0 &= 2\sqrt{1+4C_v^2}, \\ C_1 &= \sqrt{1+4(C_v+1)^2}, & C_2 &= 4+6(C_v+1). \end{aligned} \tag{215}$$

With $R_s^2 = 2a_{\mathrm{p}}^2/(\kappa_s\delta)$ and $T = 2R_s^2/\varepsilon$, the sufficient width is

$$M \geq \max\left\{\frac{4a_{\mathrm{p}}^2C_0^2}{\kappa_s\delta\varepsilon}, 8C_1^2T^2, \frac{2a_{\mathrm{p}}^2C_0^2}{\kappa_s^2}, 32C_2^2T^2\right\}. \tag{216}$$

The argument below derives these constants and proves the main-text statement. The same derivative constants are used for empirical training.

## T.2 THE BINARY MARGIN AND THE EXPECTED KERNEL

Write $\varphi_z(\omega) = v^\top s_z(A)$ for a scalar attention head, where $\omega = (A,v)$. Let $b_r = (b_{r,1}, b_{r,2})^\top$ be the fixed Rademacher output vector and set $\beta_r = b_{r,1} - b_{r,2}$. Thus $|\beta_r| \leq 2$ and $\mathbb{E}\beta_r^2 = 2$. The binary margin of the paired network is

$$F_\theta(z) = f_{\theta,1}(z) - f_{\theta,2}(z) = \frac{1}{\sqrt{2M}}\sum_{r=1}^{M}\beta_r\left[\varphi_z(\omega_{r,+}) - \varphi_z(\omega_{r,-})\right]. \tag{217}$$

The two copies in each pair start at the same $\omega_r^0 = (A_r^0, v_r^0)$, so $F_{\theta_0} = 0$. The head pairs are independent and identically distributed. Define

$$g_r(z) = \nabla_\omega\varphi_z(\omega_r^0),\ \psi_r(z) = \beta_r g_r(z),\ k_\infty(z,z') = \mathbb{E}_r\langle g_r(z), g_r(z')\rangle. \tag{218}$$

The full binary-margin kernel at initialization and its expectation are

$$K_{M,0}(z,z') = \frac{1}{M}\sum_{r=1}^{M}\langle\psi_r(z),\psi_r(z')\rangle,\ K_\infty(z,z') = 2k_\infty(z,z'). \tag{219}$$

The associated operators integrate against the uniform input distribution. The spectral assumption used below is precisely $K_\infty\chi_S = \kappa_s\chi_S$, with $\kappa_s = 2\lambda_s > 0$. No other eigenvalue is used.

For $q(z) = 1/2 + \rho_{\mathrm{p}}\chi_S(z)$, the optimal binary margin is $F^*(z) = a_{\mathrm{p}}\chi_S(z)$, where $a_{\mathrm{p}} = \log((1/2+\rho_{\mathrm{p}})/(1/2-\rho_{\mathrm{p}}))$. Define

$$\ell_q(x) = \log(1+e^x) - qx,\ \mathcal{L}(\theta) = \mathbb{E}_z\ell_{q(z)}(F_\theta(z)),\ \mathcal{L}^* = \mathbb{E}_z\ell_{q(z)}(F^*(z)). \tag{220}$$

The identities $\ell_q'(x) = \mathrm{sigmoid}(x) - q$ and $0 \leq \ell_q''(x) \leq 1/4$ imply convexity in the margin, $|\ell_q'(x)| \leq 1$, and

$$0 \leq \mathbb{E}_z\ell_{q(z)}(F(z)) - \mathcal{L}^* \leq \frac{1}{8}\|F - F^*\|_2^2 \tag{221}$$

for every real-valued margin $F$.

## T.3 UNIFORM DERIVATIVE BOUNDS

We give the derivative bounds used in the proof. Every key and query embedding has norm at most one. Fix an input and write $u$ for its query, $\alpha_j$ for its attention weights, and $s = \sum_j \alpha_j x_j$. For a score perturbation $U$, set $\ell_U(j) = u^\top U x_j$ and use $\mathbb{E}_\alpha$ for averaging over token positions. Then

$$D_A s[U] = \mathbb{E}_\alpha[x\ell_U] - s\mathbb{E}_\alpha[\ell_U]. \tag{222}$$

Since $\|x_j\| \le 1$ and $|\ell_U(j)| \le \|U\|_F$, we have $\|s\| \le 1$ and $\|D_A s[U]\| \le 2\|U\|_F$. Differentiating once more gives

$$D_A^2 s[U,V] = \mathbb{E}_\alpha[x\ell_U\ell_V] - \mathbb{E}_\alpha[x\ell_U]\mathbb{E}_\alpha[\ell_V] - \mathbb{E}_\alpha[x\ell_V]\mathbb{E}_\alpha[\ell_U] \\ - s\mathbb{E}_\alpha[\ell_U\ell_V] + 2s\mathbb{E}_\alpha[\ell_U]\mathbb{E}_\alpha[\ell_V]. \tag{223}$$

Each term is bounded by its coefficient times $\|U\|_F\|V\|_F$, so $\|D_A^2 s[U,V]\| \le 6\|U\|_F\|V\|_F$.

The value initialization satisfies $\|v_r^0\| = C_v = \sigma_v\sqrt{d}$. On the product Euclidean ball of radius one about any $\omega_r^0$, the value vector has norm at most $C_v + 1$. The gradient of $\varphi = v^\top s(A)$ consequently satisfies

$$\|\nabla_\omega \varphi_z(\omega)\| \le C_1 = \sqrt{1 + 4(C_v+1)^2}. \tag{224}$$

For two parameter directions $(U, w)$ and $(V, t)$,

$$D_\omega^2\varphi[(U,w),(V,t)] = v^\top D_A^2 s[U,V] + w^\top D_A s[V] + t^\top D_A s[U]. \tag{225}$$

It follows that the Hessian operator norm is at most $C_2 = 4 + 6(C_v + 1)$. At initialization, $\|g_r(z)\| \le \sqrt{1 + 4C_v^2}$, and hence $\|\psi_r(z)\| \le C_0 = 2\sqrt{1 + 4C_v^2}$. These bounds hold uniformly over inputs and over the score matrices, including their nonzero initial values.

### T.4 A PARAMETER COMPARISON POINT

Define the independent random vectors and functions

$$h_r = \mathbb{E}_z[\psi_r(z)\chi_S(z)],\ X_r(z) = \langle \psi_r(z), h_r\rangle. \tag{226}$$

Their expectations satisfy

$$\mathbb{E}_r X_r = \kappa_s\chi_S,\ \mathbb{E}_r\|h_r\|^2 = \langle \chi_S, K_\infty\chi_S\rangle_2 = \kappa_s. \tag{227}$$

Also $\|h_r\| \le C_0$ and $\|X_r\|_2^2 \le C_0^2\|h_r\|^2$. Independence and the Hilbert-space variance identity therefore imply

$$\mathbb{E}\left\|\frac{1}{M}\sum_{r=1}^M X_r - \kappa_s\chi_S\right\|_2^2 = \frac{1}{M}\mathbb{E}\|X_1 - \mathbb{E}X_1\|_2^2 \le \frac{C_0^2\kappa_s}{M}. \tag{228}$$

Construct a parameter displacement by

$$\Delta_{r,+} = \frac{a_{\mathrm{p}}}{\sqrt{2M}\kappa_s}h_r,\ \Delta_{r,-} = -\frac{a_{\mathrm{p}}}{\sqrt{2M}\kappa_s}h_r,\ \theta_c = \theta_0 + \Delta. \tag{229}$$

This point is only a comparison point in the proof; the learning algorithm does not need the target set $S$ or the vectors $h_r$. The derivative of the binary margin at initialization gives

$$D_\theta F_{\theta_0}[\Delta] = \frac{a_{\mathrm{p}}}{M\kappa_s}\sum_{r=1}^M X_r. \tag{230}$$

Furthermore,

$$\mathbb{E}\|\Delta\|^2 = \frac{a_{\mathrm{p}}^2}{\kappa_s},\ \mathbb{E}\,\|D_\theta F_{\theta_0}[\Delta] - F^*\|_2^2 \le \frac{a_{\mathrm{p}}^2 C_0^2}{M\kappa_s}. \tag{231}$$

Applying Markov's inequality to these two nonnegative quantities and taking a union bound shows that, with probability at least $1 - \delta$,

$$\|\Delta\|^2 \le R_s^2 = \frac{2a_{\mathrm{p}}^2}{\kappa_s\delta},\ \|D_\theta F_{\theta_0}[\Delta] - F^*\|_2^2 \le \frac{2a_{\mathrm{p}}^2 C_0^2}{M\kappa_s\delta}. \tag{232}$$

Fix this event for the remainder of the proof. The bound $M \ge 4a_{\mathrm{p}}^2 C_0^2/(\kappa_s\delta\varepsilon)$ makes the second quantity at most $\varepsilon/2$.

The comparison point also obeys a deterministic bound for every head:

$$\|\Delta_{r,\pm}\| \le \frac{a_{\mathrm{p}}C_0}{\kappa_s\sqrt{2M}} \le \frac{1}{2}, \tag{233}$$

where the last inequality follows from $M \geq 2a_{\mathrm{p}}^2 C_0^2/\kappa_s^2$. The Hessian bound and the factor $|\beta_r|/\sqrt{2M} \leq \sqrt{2/M}$ give the Taylor remainder

$$\|F_{\theta_c} - D_\theta F_{\theta_0}[\Delta]\|_\infty \leq \frac{C_2}{\sqrt{2M}}\|\Delta\|^2 \leq \frac{C_2 R_s^2}{\sqrt{2M}}. \tag{234}$$

The width condition $M \geq 32C_2^2T^2$, with $T = 2R_s^2/\varepsilon$, bounds this remainder by $\varepsilon/16$. Since $0 < \varepsilon < 1$, the comparison loss satisfies

$$\mathcal{L}(\theta_c) - \mathcal{L}^* \leq \frac{1}{8}\left(\sqrt{\varepsilon/2} + \varepsilon/16\right)^2 \leq \varepsilon/4. \tag{235}$$

### T.5 The nonlinear gradient flow

Consider the population gradient flow $\dot\theta_t = -\nabla_\theta \mathcal{L}(\theta_t)$ starting from $\theta_0$. Its vector field is smooth, so a unique local solution exists. While every head remains within distance one of its initialization, the gradient bound and $|\ell_q'| \leq 1$ imply

$$\|\dot\omega_{r,\pm}(t)\| \leq \frac{\sqrt{2}C_1}{\sqrt{M}}. \tag{236}$$

The condition $M \geq 8C_1^2T^2$ implies that each head moves by at most $1/2$ before time $T$. To justify this without assuming existence on the entire interval, apply the estimate up to the first exit from a radius-one ball or the maximal existence time, whichever occurs first. The estimate excludes an exit before $T$. All parameters then remain in a bounded set, so the smooth vector field allows continuation through any finite endpoint before $T$. The unique solution therefore exists on $[0, T]$ and satisfies the stated movement bound.

Define

$$E(t) = \mathcal{L}(\theta_t) - \mathcal{L}^*,\ D_{\mathrm{cmp}}(t) = \frac{1}{2}\|\theta_t - \theta_c\|^2,\ c_M = \frac{\sqrt{2}C_2}{\sqrt{M}}. \tag{237}$$

Both endpoints $\theta_t$ and $\theta_c$ remain within radius $1/2$ for every head. Their joining segment therefore lies in the region where the Hessian bound holds. Taylor's theorem, expanded at $\theta_t$, gives

$$\|F_{\theta_t} - F_{\theta_c} - D_\theta F_{\theta_t}[\theta_t - \theta_c]\|_\infty \leq \frac{C_2}{\sqrt{2M}}\|\theta_t - \theta_c\|^2 = c_M D_{\mathrm{cmp}}(t). \tag{238}$$

Convexity of binary cross-entropy in its margin gives

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_c) \leq \mathbb{E}_z\left[\ell'_{q(z)}(F_{\theta_t}(z))\big(F_{\theta_t}(z) - F_{\theta_c}(z)\big)\right]. \tag{239}$$

Using the gradient-flow equation, the last two inequalities, and $|\ell_q'| \leq 1$, we obtain

$$\begin{aligned}\dot D_{\mathrm{cmp}}(t) &= -\langle \nabla_\theta \mathcal{L}(\theta_t), \theta_t - \theta_c\rangle \\ &\leq \mathcal{L}(\theta_c) - \mathcal{L}(\theta_t) + c_M D_{\mathrm{cmp}}(t) \\ &\leq \varepsilon/4 - E(t) + c_M D_{\mathrm{cmp}}(t).\end{aligned} \tag{240}$$

Since $E(t) \geq 0$ and $D_{\mathrm{cmp}}(0) \leq R_s^2/2$, integration yields

$$D_{\mathrm{cmp}}(t) \leq e^{c_M t}\left(R_s^2/2 + \varepsilon t/4\right). \tag{241}$$

Our width condition ensures $c_M T \leq 1/4$. Together with $T = 2R_s^2/\varepsilon$, this gives $D_{\mathrm{cmp}}(t) \leq e^{1/4}R_s^2$ for $0 \leq t \leq T$.

Along exact population gradient flow, $\frac{d}{dt}\mathcal{L}(\theta_t) = -\|\nabla_\theta \mathcal{L}(\theta_t)\|^2 \leq 0$. Thus $E(T)$ is at most the time average of $E(t)$. Integrating the preceding differential inequality for $D_{\mathrm{cmp}}$ and using $D_{\mathrm{cmp}}(T) \geq 0$ gives

$$\begin{aligned}E(T) &\leq \frac{1}{T}\int_0^T E(t)\,dt \\ &\leq \frac{\varepsilon}{4} + \frac{R_s^2}{2T} + c_M e^{1/4} R_s^2 \\ &\leq \frac{\varepsilon}{2} + \frac{e^{1/4}\varepsilon}{8} < \varepsilon.\end{aligned} \tag{242}$$

This proves the loss guarantee without an inverse of $K_\infty$ on the orthogonal complement of the target and without a lower bound for any other kernel eigenvalue.

Finally, excess binary cross-entropy is the average conditional Kullback–Leibler divergence. Pinsker's inequality for Bernoulli distributions gives

$$\mathbb{E}_z\big[(\widehat{p}_T(z) - q(z))^2\big] \le \frac{1}{2}\mathbb{E}_z \,\mathrm{KL}\big(\mathrm{Bern}(q(z)) \,\|\, \mathrm{Bern}(\widehat{p}_T(z))\big) = E(T)/2 \le \varepsilon/2. \tag{243}$$

The probability statement follows from the two Markov bounds used to construct the comparison point. All subsequent estimates are deterministic on that event. This completes the proof.

### T.6 Finite samples and perturbed paired initialization

The population theorem does not by itself control training on sampled outcomes. We now analyze gradient flow on a fixed empirical loss. A uniform deviation bound is used both to control its parameter trajectory and to transfer its terminal loss to the population distribution. The same argument allows a bounded perturbation of the paired initialization, so the initial output need not vanish.

Let $(Z_i, Y_i)_{i=1}^{n_\mathrm{p}}$ be independent, identically distributed examples, with $Z_i$ uniform on $\{-1,1\}^P$ and $Y_i \in \{0,1\}$ satisfying

$$\Pr(Y_i = 1 \mid Z_i = z) = q(z) = 1/2 + \rho_\mathrm{p}\chi_S(z). \tag{244}$$

Thus $Y_i$ is the indicator of the first binary outcome. These examples are independent of the random reference initialization $\theta_0$. Define

$$\widehat{\mathcal{L}}_\mathrm{p}(\theta) = \frac{1}{n_\mathrm{p}}\sum_{i=1}^{n_\mathrm{p}} \ell_{Y_i}(F_\theta(Z_i)),\ \ell_y(x) = \log(1+e^x) - yx. \tag{245}$$

Training starts at $\theta_\mathrm{init} = \theta_0 + U$, where $U$ is independent of these examples and is chosen without using $S$, but may depend on the reference initialization. The two copies of each pair need not agree at $\theta_\mathrm{init}$. The architecture and its fixed output weights are unchanged. Write $\vartheta_t$ for the training trajectory, to distinguish $\vartheta_0$ from the paired reference point $\theta_0$. All score and value parameters follow the unprojected flow $\dot{\vartheta}_t = -\nabla_\theta \widehat{\mathcal{L}}_\mathrm{p}(\vartheta_t)$, with $\vartheta_0 = \theta_\mathrm{init}$.

**Theorem 51** (Finite-sample learning from a target eigenvalue)**.** *Use the reference network, target, and constants $a_\mathrm{p}, C_0, C_1, C_2$ of Theorem 15, and suppose $\kappa_s > 0$. Let $\varepsilon, \delta \in (0,1)$. Assume almost surely that $\|U\| \le r_\mathrm{init}$ and $\max_{r,\pm}\|U_{r,\pm}\| \le 1/4$, where $r_\mathrm{init} \ge 0$ is deterministic. Set*

$$\begin{aligned} r_\mathrm{tar}^2 &= \frac{4a_\mathrm{p}^2}{\kappa_s\delta}, & R_\mathrm{emp} &= \max\{1, r_\mathrm{tar} + r_\mathrm{init}\},\\ B_\mathrm{emp} &= 4R_\mathrm{emp}, & T_\mathrm{emp} &= \frac{4R_\mathrm{emp}^2}{\varepsilon},\\ V_\mathrm{emp} &= 4C_0B_\mathrm{emp} + (C_0B_\mathrm{emp} + a_\mathrm{p})\sqrt{2\log(8/\delta)}. \end{aligned} \tag{246}$$

*Suppose that*

$$n_\mathrm{p} \ge \frac{1024V_\mathrm{emp}^2}{\varepsilon^2}, \tag{247}$$

*and*

$$M \ge \max\left\{\frac{32a_\mathrm{p}^2C_0^2}{\kappa_s\delta\varepsilon}, 32C_1^2T_\mathrm{emp}^2, \frac{2a_\mathrm{p}^2C_0^2}{\kappa_s^2}, \frac{2048C_2^2B_\mathrm{emp}^4}{\varepsilon^2}\right\}. \tag{248}$$

*With probability at least $1-\delta$ over the reference initialization and the examples, empirical gradient flow exists on $[0, T_\mathrm{emp}]$, every head stays within distance $1/2$ of its reference initialization, and*

$$\mathcal{L}(\vartheta_{T_\mathrm{emp}}) - \mathcal{L}^* \le \varepsilon,\ \mathbb{E}_Z[(\mathrm{sigmoid}(F_{\vartheta_{T_\mathrm{emp}}}(Z)) - q(Z))^2] \le \varepsilon/2. \tag{249}$$

*No input-coverage condition or lower bound on any other eigenvalue is required.*

The perturbation condition describes a neighborhood in parameter space, not merely a bound on the initial output. In that neighborhood Taylor's theorem gives

$$\|F_{\theta_{\mathrm{init}}}\|_\infty \le C_0 r_{\mathrm{init}} + \frac{C_2 r_{\mathrm{init}}^2}{\sqrt{2M}}. \tag{250}$$

Thus the theorem covers nonzero initial outputs in a neighborhood of paired initialization, while the target set $S$ enters only through the proof and target specification, not through the features supplied to the learner. The stopping time and resource bounds depend on the target order and eigenvalue, or on the explicit lower bound in equation (211).

*Proof.* Write $T = T_{\mathrm{emp}}$, $B = B_{\mathrm{emp}}$, and $R = R_{\mathrm{emp}}$ within this proof. The comparison construction above, with failure parameter $\delta/2$, gives an event of probability at least $1 - \delta/2$ on which

$$\|\Delta\|^2 \le r_{\mathrm{tar}}^2,\ \|D_\theta F_{\theta_0}[\Delta] - F^*\|_2^2 \le \frac{4a_{\mathrm{p}}^2 C_0^2}{M\kappa_s \delta} \le \varepsilon/8. \tag{251}$$

Its deterministic headwise bound and equation (248) also give $\|\Delta_{r,\pm}\| \le 1/2$. Fix this initialization event, let $\theta_c = \theta_0 + \Delta$, and condition on the reference initialization for the sampling argument.

**Uniform deviation on the parameter neighborhood.** Let $\mathcal{B}$ consist of parameters with $\|\theta - \theta_0\| \le B$ and $\max_{r,\pm} \|\omega_{r,\pm} - \omega_r^0\| \le 1$. From the paired normalization, the full reference margin Jacobian obeys $\|\nabla_\theta F_{\theta_0}(z)\| \le C_0$ uniformly in $z$. For every $\theta \in \mathcal{B}$, the head Hessian bound gives

$$\|F_\theta - D_\theta F_{\theta_0}[\theta - \theta_0]\|_\infty \le Q_{\mathrm{lin}} := \frac{C_2 B^2}{\sqrt{2M}} \le \varepsilon/64. \tag{252}$$

In particular, the $1/4$-smoothness of the population loss in its margin yields

$$E_c := \mathcal{L}(\theta_c) - \mathcal{L}^* \le \frac{1}{8}(\sqrt{\varepsilon/8} + \varepsilon/64)^2 \le \varepsilon/16. \tag{253}$$

For completeness, consider the tangent loss class

$$g_w(z, y) = \ell_y(D_\theta F_{\theta_0}(z)[w]) - \ell_y(0),\ \|w\| \le B. \tag{254}$$

Let $\mathbb{P}$ and $\mathbb{P}_{n_{\mathrm{p}}}$ denote the true and empirical expectations of functions of $(Z, Y)$ in this proof. Each map $x \mapsto \ell_y(x) - \ell_y(0)$ is one-Lipschitz and vanishes at zero. Symmetrization, contraction for Rademacher averages, and Cauchy–Schwarz give

$$\mathbb{E} \sup_{\|w\| \le B} |(\mathbb{P} - \mathbb{P}_{n_{\mathrm{p}}}) g_w| \le \frac{4BC_0}{\sqrt{n_{\mathrm{p}}}}. \tag{255}$$

To see the dimension-free estimate in the last step, for independent Rademacher signs $\epsilon_i$, the tangent linear class satisfies

$$\mathbb{E}_\epsilon \sup_{\|w\| \le B} \left| \frac{1}{n_{\mathrm{p}}} \sum_i \epsilon_i D_\theta F_{\theta_0}(Z_i)[w] \right| \le \frac{B}{n_{\mathrm{p}}} \left( \sum_i \|\nabla_\theta F_{\theta_0}(Z_i)\|^2 \right)^{1/2} \le \frac{BC_0}{\sqrt{n_{\mathrm{p}}}}. \tag{256}$$

The factors two from symmetrization and from the absolute-value contraction inequality give the preceding factor four. Since $|g_w| \le BC_0$, changing one example changes the supremum deviation by at most $2BC_0/n_{\mathrm{p}}$. The bounded-differences inequality therefore bounds that deviation, with failure probability at most $\delta/4$, by

$$\frac{4BC_0}{\sqrt{n_{\mathrm{p}}}} + BC_0 \sqrt{\frac{2\log(4/\delta)}{n_{\mathrm{p}}}}. \tag{257}$$

The fixed function $\ell_Y(0) - \ell_Y(F^*(Z))$ has absolute value at most $a_{\mathrm{p}}$. Hoeffding's inequality bounds its deviation by $a_{\mathrm{p}}\sqrt{2\log(8/\delta)/n_{\mathrm{p}}}$ with failure probability at most $\delta/4$. Combining these two events and the nonlinear remainder, uniformly on $\mathcal{B}$,

$$\begin{aligned} \left| [\mathcal{L}(\theta) - \mathcal{L}^*] - [\widehat{\mathcal{L}}_{\mathrm{p}}(\theta) - \widehat{\mathcal{L}}_{\mathrm{p}}(F^*)] \right| &\le \frac{V_{\mathrm{emp}}}{\sqrt{n_{\mathrm{p}}}} + 2Q_{\mathrm{lin}} \\ &\le \varepsilon/16 =: G_{\mathrm{emp}}. \end{aligned} \tag{258}$$

Here $\widehat{\mathcal{L}}_{\mathrm{p}}(F^*) = n_{\mathrm{p}}^{-1} \sum_i \ell_{Y_i}(F^*(Z_i))$ denotes an empirical comparator loss, not a network parameter. The sampling event has conditional probability at least $1 - \delta/2$ for every reference initialization. Its intersection with the initialization event thus has probability at least $1 - \delta$. All subsequent estimates are deterministic on this intersection.

**Existence and control of the empirical trajectory.** While each head stays within distance one of its reference initialization, $|\ell'_{Y_i}| \le 1$ and the derivative bound imply

$$\|\dot{\omega}_{r,\pm}(t)\| \le \frac{\sqrt{2}C_1}{\sqrt{M}}. \tag{259}$$

The second width condition makes the movement over $[0,T]$ at most $1/4$. Adding the initial displacement of at most $1/4$ gives a distance of at most $1/2$ from the reference point. Applying the estimate up to the first exit from a unit head ball excludes such an exit, and smoothness gives continuation through every finite endpoint before $T$.

It remains to control the global distance needed in the uniform deviation bound. Define $D_{\mathrm{cmp}}(t) = \|\vartheta_t - \theta_c\|^2/2$ and $c_M = \sqrt{2}C_2/\sqrt{M}$. The fourth width condition gives $c_M T \le 1/4$. Both endpoints and their headwise segments satisfy the Hessian bound. Convexity of each empirical loss in its margin and Taylor expansion at $\vartheta_t$ therefore imply, before any exit from $\mathcal{B}$,

$$\begin{aligned} \dot{D_{\mathrm{cmp}}}(t) &\le \widehat{\mathcal{L}}_{\mathrm{p}}(\theta_c) - \widehat{\mathcal{L}}_{\mathrm{p}}(\vartheta_t) + c_M D_{\mathrm{cmp}}(t) \\ &\le E_c - E(t) + 2G_{\mathrm{emp}} + c_M D_{\mathrm{cmp}}(t), \end{aligned} \tag{260}$$

where $E(t) = \mathcal{L}(\vartheta_t) - \mathcal{L}^* \ge 0$. Since $\|\theta_{\mathrm{init}} - \theta_c\| \le r_{\mathrm{init}} + r_{\mathrm{tar}} \le R$, integration gives

$$D_{\mathrm{cmp}}(t) \le e^{c_M t}\left(\frac{R^2}{2} + \frac{3\varepsilon t}{16}\right) \le \frac{5}{4}e^{1/4}R^2 < 2R^2,\ 0 \le t \le T. \tag{261}$$

Consequently $\|\vartheta_t - \theta_0\| \le \|\Delta\| + \sqrt{2D_{\mathrm{cmp}}(t)} < 3R < B$. Continuity excludes an exit through the global-radius boundary as well. Thus the deviation bound applies along the whole trajectory without projecting the gradient updates.

**Terminal population error.** Only the empirical loss is guaranteed to be nonincreasing. Using its monotonicity and equation (258) at the endpoint and along the trajectory gives

$$E(T) \le \frac{1}{T}\int_0^T E(t)\,\mathrm{d}t + 2G_{\mathrm{emp}}. \tag{262}$$

Integrating the differential inequality for $D_{\mathrm{cmp}}$, using $D_{\mathrm{cmp}}(T) \ge 0$, and substituting the preceding bounds, we obtain

$$\begin{aligned} E(T) &\le E_c + 4G_{\mathrm{emp}} + \frac{R^2}{2T} + 2c_M R^2 \\ &\le \frac{\varepsilon}{16} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} + \frac{\varepsilon}{8} < \varepsilon. \end{aligned} \tag{263}$$

Finally, population excess cross-entropy is the expected conditional Bernoulli Kullback–Leibler divergence. Pinsker's inequality gives the asserted mean squared probability error. This completes the proof. □

The result concerns independent noisy outcomes, and its sample and width requirements are sufficient rather than optimized.

### T.7 A PLANNING CONSEQUENCE OF EXACT POPULATION STATIONARITY

**Proposition 52** (A model-only planning obstruction)**.** *Fix $S \subseteq \{1,\dots,P\}$ with $|S| \ge 3$, $0 < \delta_{\mathrm{p}} < 1/2$, and $\delta_{\mathrm{c}} \in (0,1)$. Consider two horizon-one environments, indexed by $b \in \{-1,1\}$. In each environment the initial observation $z$ is uniform on $\{-1,1\}^P$, the action is $a \in \{-1,1\}$, and the reward distribution is*

$$R \mid z, a \sim \mathrm{Bernoulli}\left(\frac{1}{2} + b\delta_{\mathrm{p}} a \chi_S(z)\right). \tag{264}$$

*The next observation is a common deterministic terminal symbol. Each environment has a singleton task prior. Train the paired binary predictor under uniform behavior actions, serializing $z$ as the tokens $e_{j,z_j}$ followed by the action query $e_{0,a}$, with all token types represented by distinct one-hot vectors. At zero score initialization, exact population gradient flow and exact population gradient descent leave the predicted reward kernel uniform, at every finite width.*

*Let $\widehat{\pi}$ be any fixed, possibly randomized planner that uses only this learned kernel and the known observation and action spaces. Its tie-breaking and randomization law must be the same in the two environments. Then*

$$\frac{1}{2}\sum_{b\in\{-1,1\}} \mathrm{SubOpt}_b(\widehat{\pi}) = \delta_{\mathrm{p}},\ \max_b \mathrm{SubOpt}_b(\widehat{\pi}) \geq \delta_{\mathrm{p}}, \tag{265}$$

*where values include the planner's randomization. Nevertheless, if $S$ is known, $N \geq \log(1/\delta_{\mathrm{c}})/(2\delta_{\mathrm{p}}^2)$ independent behavior samples identify $b$ with probability at least $1-\delta_{\mathrm{c}}$.*

*Proof.* Condition on either action. Its query is fixed, so the Jacobian argument in Proposition 13 applies: every parameter derivative has degree at most two in $z$. Paired initialization gives zero logits. The cross-entropy residual is proportional to $ba\chi_S(z)$, which is orthogonal to every such derivative under the uniform cube measure. Every population gradient coordinate therefore vanishes, also after averaging over actions. The initialized parameters are a fixed point of exact population gradient descent and a stationary solution of population gradient flow. The latter solution is unique by local smoothness.

Both learned kernels are identical. Write $m(z) = \mathbb{E}_{\widehat{\pi}}[a \mid z]$. The optimal action in environment $b$ is $b\chi_S(z)$, and hence

$$V_b^* = \frac{1}{2} + \delta_{\mathrm{p}},\ V_b^{\widehat{\pi}} = \frac{1}{2} + b\delta_{\mathrm{p}}\mathbb{E}_z[m(z)\chi_S(z)]. \tag{266}$$

Averaging the two suboptimalities cancels the second term. This average compares two separate singleton-task instances; it is not a hidden mixture prior within either instance.

For the statistical claim, define $U_i = (2R_i - 1)a_i\chi_S(z_i)$. Then $U_i \in \{-1, 1\}$ and $\mathbb{E}_b U_i = 2b\delta_{\mathrm{p}}$. Hoeffding's inequality gives

$$\Pr_b\left\{\mathrm{sign}\left(\frac{1}{N}\sum_{i=1}^{N} U_i\right) \neq b\right\} \leq e^{-2N\delta_{\mathrm{p}}^2}, \tag{267}$$

with any fixed convention at zero. This proves the sample bound. □

The obstruction is specific to the exact population training dynamics considered above.

# U Additional Experimental Details and Numerical Checks

## U.1 Computation and evaluation

The recorded environment uses an NVIDIA GeForce RTX 4090, PyTorch 2.5.1, and CUDA 12.4. Computations use FP64 arithmetic, deterministic algorithms, and disabled TF32. Experiment 1 uses exact conditional probabilities and has no sampling variability. Experiments 2–4 and 6 use master seeds $0, \ldots, 9$, while Experiment 5 uses master seeds $1000, \ldots, 1009$; Experiment 3 additionally evaluates 100 initializations with seeds $0, \ldots, 99$. In Experiment 4, independent data and initialization seeds are deterministically generated from each master seed. Standard deviations are computed across seeds within a setting. Checkpoints from the same run are not treated as independent repetitions.

Total variation and Kullback–Leibler divergence are evaluated against the known target distributions, with the target as the first argument of KL divergence. Sparse-model errors are maxima over all queries and the indicated feature vectors; multi-head errors are maxima over the finite context set. Excess cross-entropy subtracts the corresponding target conditional entropy, using empirical targets and visitation weights when reporting empirical excess loss. Descriptive scaling slopes are obtained by fitting logarithms of group means against logarithms of sample size or width; they are not estimates from independent checkpoint observations.

## U.2 Experiment 1: Integration and discrete optimization

We run 5,000 full-batch gradient-descent updates at each step size $\eta_{\mathrm{gd}} \in \{1, 10, 100\}$. All 96 recorded flow and descent checkpoints satisfy the attention, TV, and KL bounds of Theorems 6

Table 1: Experiment 1 at $t = 50{,}000$. Input counts are per query; errors are maxima over all four queries and the indicated vectors.

| $P$ | Design | Unseen | $\alpha_\star$ | Max. TV all | Max. TV unseen | Max. KL unseen |
|---|---|---|---|---|---|---|
| 4 | 8 | 8 | 0.98910 | 0.00431 | 0.00287 | $2.08 \times 10^{-5}$ |
| 8 | 16 | 240 | 0.98174 | 0.00724 | 0.00620 | $9.68 \times 10^{-5}$ |
| 12 | 24 | 4,072 | 0.97625 | 0.00944 | 0.00857 | $1.84 \times 10^{-4}$ |
| 16 | 32 | 65,504 | 0.97156 | 0.01133 | 0.01056 | $2.79 \times 10^{-4}$ |

and 19, although the bounds are conservative. The descent results show the same feature-selection behavior at all tested step sizes; their fixed-iteration comparison is reported in Appendix U.2.

Population gradient flow is integrated with fourth-order Runge–Kutta, step size 10, and checkpoints at $t \in \{0, 100, 1000, 5000, 20000, 50000\}$. Each query has $2P$ design feature vectors, so the objective contains $2Pn_{\mathrm{q}}$ query–vector inputs. Their exact conditional probabilities are used rather than a finite sample of outcome labels. Enumeration of the full binary cube therefore tests generalization beyond the training input support, not generalization from a stated number of sampled labels.

Full-batch gradient descent uses $\eta_{\mathrm{gd}} \in \{1, 10, 100\}$. Checkpoints are recorded after 0, 10, 50, 200, 1000, and 5000 updates. For $P = 16$, maximum unseen-input TV after 5,000 updates is 0.04615, 0.01056, and 0.00316, respectively. These are fixed-iteration comparisons: $\eta_{\mathrm{gd}} n$ differs across settings, so they are not comparisons at matched gradient-flow time. The result is specific to the symmetric model and does not establish unrestricted step-size stability for general attention architectures.

All 96 recorded population-flow and gradient-descent checkpoints satisfy each of the explicit attention, TV, and KL inequalities. The time-dependent bounds are conservative. At $P = 16$ and $t = 50{,}000$, the attention-error bound is 0.38927, whereas the observed irrelevant attention mass is 0.02844. Satisfaction of these bounds is a comparison with an upper bound, not evidence that the bound tightly describes the observed trajectory.

### U.3 Experiment 2: Sampling protocol and complete sample-size comparison

At $\delta = 0.05$, the concentration event holds in all 120 sampled instances. The additional condition $r_{\mathrm{stat}} \leq 1/2$ holds only for $(P, N_{\mathrm{sel}})$ equal to $(4, 65536)$, $(4, 262144)$, and $(8, 262144)$; all 150 checkpoints from these 30 instances satisfy the attention, TV, and KL bounds of Theorem 22. The remaining settings are not counted as tests under that theorem's assumptions.

For each indexed input $(u, j, s)$, we collect $N_{\mathrm{sel}}$ independent conditional outcomes. The total number of outcomes is $2Pn_{\mathrm{q}}N_{\mathrm{sel}}$, and the empirical frequencies remain fixed throughout training. We retain $\lambda = 1$ and $n_{\mathrm{q}} = 4$. Empirical gradient flow starts from zero scores and is integrated by fourth-order Runge–Kutta with step size 100, with checkpoints at $t \in \{0, 1000, 5000, 20000, 50000\}$. The two dimensions, six sample counts, and ten seeds give 120 sampled instances and 600 checkpoint records. Sampling radii and theorem conditions use $\delta = 0.05$.

Table 2: Experiment 2 at $t = 50{,}000$. Sampling is the mean maximum symmetrized-frequency error; TV is mean $\pm$ standard deviation across ten seeds. A dagger marks a setting satisfying $r_{\mathrm{stat}} \leq 1/2$.

| | $P = 4$ | | $P = 8$ | |
|---|---|---|---|---|
| $N_{\mathrm{sel}}$ | Sampling | Max. TV | Sampling | Max. TV |
| 256 | 0.04344 | $0.02663 \pm 0.01146$ | 0.04430 | $0.03394 \pm 0.01304$ |
| 1,024 | 0.01928 | $0.01129 \pm 0.00348$ | 0.02368 | $0.01301 \pm 0.00566$ |
| 4,096 | 0.01042 | $0.00802 \pm 0.00365$ | 0.01153 | $0.01022 \pm 0.00346$ |
| 16,384 | 0.00480 | $0.00521 \pm 0.00072$ | 0.00538 | $0.00795 \pm 0.00028$ |
| 65,536 | 0.00234 | $(0.00464 \pm 0.00023)^{\dagger}$ | 0.00271 | $0.00755 \pm 0.00019$ |
| 262,144 | 0.00118 | $(0.00448 \pm 0.00013)^{\dagger}$ | 0.00139 | $(0.00738 \pm 0.00009)^{\dagger}$ |

The three applicable settings have $(P, N_{\mathrm{sel}})$ equal to $(4, 65536)$, $(4, 262144)$, and $(8, 262144)$, with deterministic radii $0.4040$, $0.2020$, and $0.4251$, respectively. Only these settings satisfy $r_{\mathrm{stat}} \le 1/2$. The concentration event also holds in the other sampled instances, but this alone does not make Theorem 22 applicable. Accordingly, its attention, TV, and KL inequalities are counted as applicable at 150 checkpoints and left inapplicable at the remaining 450.

Relevant-score ordering is preserved at every recorded checkpoint, including those outside the radius condition. Population prediction error nevertheless need not decrease monotonically along an empirical trajectory. For $P = 8$ and $N_{\mathrm{sel}} = 256$, mean maximum TV increases from $0.03136$ at $t = 20{,}000$ to $0.03394$ at $t = 50{,}000$. Minimizing empirical cross-entropy and reducing prediction error against the population distribution are distinct objectives. Likewise, the nonzero terminal errors at large sample sizes should not be interpreted as measured asymptotic statistical floors, since the population trajectories still have finite-time optimization error.

### U.4 Experiment 3: Kernel diagnostics and theorem conditions

Table 3: Experiment 3. Initialization error and the $\gamma/2$ count use 100 initializations. Maximum single-head movement, relative kernel drift, and maximum TV are measured at $t = 500$ over ten training seeds. Entries are means; TV also includes standard deviation.

| Scale | $M$ | Init. error | $\gamma/2$ count | Movement | Relative drift | Max. TV |
|---|---|---|---|---|---|---|
| 1 | 16 | 0.480 | 29/100 | 1.177 | 0.45702 | $0.0106 \pm 0.0033$ |
| 1 | 64 | 0.244 | 99/100 | 0.592 | 0.14047 | $0.0165 \pm 0.0048$ |
| 1 | 256 | 0.123 | 100/100 | 0.298 | 0.03856 | $0.0190 \pm 0.0024$ |
| 0.25 | 16 | 0.517 | 12/100 | 0.803 | 0.02020 | $0.2042 \pm 0.0064$ |
| 0.25 | 64 | 0.265 | 95/100 | 0.393 | 0.00646 | $0.2057 \pm 0.0048$ |
| 0.25 | 256 | 0.134 | 100/100 | 0.196 | 0.00180 | $0.2069 \pm 0.0020$ |

This experiment uses ten fixed full-support binary target distributions and uniform context weights, rather than the trajectory-dependent weights of Experiment 4. The two representation settings, indexed by context scales 1 and $0.25$, give $\gamma = 6.173 \times 10^{-3}$ and $\gamma = 2.050 \times 10^{-4}$, respectively. Population gradient flow is approximated by explicit Euler with step size $0.05$. The 60 training runs are evaluated at $t \in \{0, 1.25, 5, 10, 20, 50, 100, 250, 500\}$, giving 540 checkpoint records; the initialization evaluation comprises 600 records.

Initialization error is $\|\boldsymbol{\Theta}_A(0) - \mathbb{E}\boldsymbol{\Theta}_A(0)\|_{\mathrm{op}} / \|\mathbb{E}\boldsymbol{\Theta}_A(0)\|_{\mathrm{op}}$. Relative training drift is $\|\boldsymbol{\Theta}_A(t) - \boldsymbol{\Theta}_A(0)\|_{\mathrm{op}} / \|\boldsymbol{\Theta}_A(0)\|_{\mathrm{op}}$, with each trajectory compared against its own initialization. Maximum single-head movement is $\max_{r,s} \|\omega_{r,s}(t) - \omega_r^0\|_2$. The centered kernel minimum eigenvalue is evaluated on the range of $\boldsymbol{P}$, excluding the structural zero eigenvalues of the projected matrix. The $\gamma/2$ counts in Table 3 instead refer to the unprojected initialization inequality $\lambda_{\min}(\boldsymbol{\Theta}_A(0)) \ge \gamma/2$.

Three-point log–log fits across $M \in \{16, 64, 256\}$ give initialization-error slopes $-0.491$ and $-0.487$ for scales 1 and $0.25$, and terminal-movement slopes $-0.495$ and $-0.508$. These finite-range observations are consistent with inverse-square-root width dependence, but do not establish an asymptotic exponent. At scale 1, terminal excess losses are $1.27 \times 10^{-4}$, $3.18 \times 10^{-4}$, and $4.01 \times 10^{-4}$ as width increases. At scale $0.25$, they remain between $0.0439$ and $0.0450$. Thus the representation with the smaller singular-value margin trains more slowly even though both context matrices retain full row rank.

The sufficient initialization-width requirements are approximately $2.48 \times 10^6$ and $7.48 \times 10^7$ for the two representations, far above the tested widths. The separate parameter-movement condition is also not met. Consequently, numerical satisfaction of the centered kernel inequality at the recorded checkpoints is not a test of the full finite-width theorem under all its sufficient conditions.

### U.5 Experiment 4: Coverage, stopping, and error decomposition

We use $M = 64$ head pairs and independent length-two episodes from a history-dependent environment with two actions, binary outcomes, and ten contexts. The behavior policy selects action 1 with

probability $b \in \{0.5, 0.1, 0.02\}$. Training retains the empirical visitation weights $\widehat{d}_i = N_i/(2N)$. All runs reach $t = 500$ and then stop at a prescribed checkpoint when maximum empirical fitting error is at most $0.005$, or at $t = 5000$; budget-exhausted runs are retained.

Table 4: Experiment 4. Covered counts fully observed runs out of ten. $S$ averages only those runs; $F$ and $D$ use all ten. Entries are means, with standard deviation for terminal prediction $D(T)$.

| $b$ | $N$ | Covered | $S$ | $F(T)$ | $D(500)$ | $D(T)$ |
|---|---|---|---|---|---|---|
| 0.5 | 1,024 | 10/10 | 0.0736 | 0.0048 | 0.1100 | $0.0734 \pm 0.0169$ |
| 0.5 | 4,096 | 10/10 | 0.0347 | 0.0049 | 0.0972 | $0.0342 \pm 0.0062$ |
| 0.5 | 16,384 | 10/10 | 0.0185 | 0.0049 | 0.0911 | $0.0189 \pm 0.0047$ |
| 0.5 | 65,536 | 10/10 | 0.0078 | 0.0048 | 0.0916 | $0.0093 \pm 0.0015$ |
| 0.1 | 1,024 | 9/10 | 0.3503 | 0.2153 | 0.3652 | $0.2462 \pm 0.0983$ |
| 0.1 | 4,096 | 10/10 | 0.1051 | 0.2214 | 0.3651 | $0.1897 \pm 0.0504$ |
| 0.1 | 16,384 | 10/10 | 0.0602 | 0.2192 | 0.3682 | $0.2018 \pm 0.0374$ |
| 0.1 | 65,536 | 10/10 | 0.0238 | 0.2079 | 0.3685 | $0.2113 \pm 0.0083$ |
| 0.02 | 1,024 | 0/10 | N/A | 0.1624 | 0.3832 | $0.3615 \pm 0.0548$ |
| 0.02 | 4,096 | 3/10 | 0.4947 | 0.4086 | 0.4056 | $0.3971 \pm 0.0336$ |
| 0.02 | 16,384 | 9/10 | 0.2690 | 0.4864 | 0.4052 | $0.3990 \pm 0.0136$ |
| 0.02 | 65,536 | 10/10 | 0.1183 | 0.3319 | 0.4008 | $0.3928 \pm 0.0071$ |

The model uses embedding dimension 16, $\sigma_v = 0.25$, and a fixed one-hot context-token-and-action serializer. The constant behavior probabilities $b = 0.5$, $0.1$, and $0.02$ give minimum episode-level visitation probabilities $\rho_{\min} = 0.0875$, $0.0035$, and $0.00014$, respectively. Each setting uses $N \in \{1024, 4096, 16384, 65536\}$ independent episodes and ten data/initialization seeds. The empirical objective preserves $\widehat{d}_i = N_i/(2N)$ without context reweighting, probability smoothing, or task-identity input.

Explicit Euler uses step size $0.05$. All runs first reach 10,000 steps, or $t = 500$, with common checkpoints at steps 0, 100, 300, 1000, 3000, and 10000. From step 10,000 onward, the fitting criterion is evaluated every 1,000 steps. A run stops when $F(t) \le 0.005$ or when it reaches 100,000 steps, corresponding to $t = 5000$. Thus $D(500)$ compares a common optimization budget, whereas $D(T)$ uses run-specific stopping times under the same empirical criterion. All runs that fail to meet the fitting criterion within the budget remain in the reported results.

Prediction error is evaluated over all contexts, whereas empirical fitting error uses only visited contexts:

$$D(t) = \max_i \mathrm{TV}(q_i, \widehat{P}_t(\cdot \mid c_i)),$$
$$F(t) = \max_{i:N_i>0} \mathrm{TV}(\widehat{q}_i, \widehat{P}_t(\cdot \mid c_i)).$$

The sampling error $S = \max_i \mathrm{TV}(q_i, \widehat{q}_i)$ is reported only when every context has been observed. Incompletely covered runs still contribute to $D$ and $F$, but not to the full-context sampling statistic. The entry N/A denotes the absence of fully covered runs, not zero sampling error. The bound $D(t) \le S + F(t)$ is checked within each fully covered run, not by combining group means computed on different subsets. It holds at all 101 fully covered terminal records and all 7,770 fully covered training checkpoints.

For $b = 0.02$, full coverage occurs in $0/10$, $3/10$, $9/10$, and $10/10$ runs as $N$ increases. Even at $N = 65{,}536$, the expected count of the least-visited context is only $N\rho_{\min} = 9.175$. At $b = 0.5$, all runs are fully covered, and a four-point log–log fit to mean sampling TV gives slope $-0.531$, consistent with $(N\rho_{\min})^{-1/2}$ dependence at fixed behavior policy.

Coverage and fitting remain distinct at larger sample sizes. For $b = 0.1$, increasing $N$ from $4{,}096$ to $65{,}536$ reduces mean sampling TV from $0.1051$ to $0.0238$, while terminal fitting error remains near $0.21$. At the largest sample size, mean empirical excess losses are only $1.75 \times 10^{-4}$ and $1.17 \times 10^{-4}$ for $b = 0.1$ and $b = 0.02$, respectively, despite their large worst-context errors. These results concern the observed behavior of the visitation-weighted objective. The sufficient initialization-width

requirement is approximately $2.48 \times 10^6$, and neither the width nor the movement condition is met, so the experiment does not verify the full trajectory-level theorem under its sufficient conditions.

### U.6 Experiment 5: Initialization, higher-order learning, and finite-sample diagnostics

We use the binary parity model of Section 4.1 with target amplitude $\rho_{\mathrm{p}} = 0.2$ and value-initialization scale $\sigma_v = 0.25$. Training experiments use ten seeds, 1000–1009, explicit Euler integration with step size $0.1$, and checkpoints at $t \in \{0, 1, 5, 10, 20, 50, 100, 200\}$. Population losses, Walsh coefficients, and prediction errors are evaluated by enumerating the complete Boolean cube.

**Initialization spectrum and population training.** For the spectral calculation, we use $P \in \{4, 6, 8, 12\}$, interaction orders $2 \le s \le \min\{6, P\}$, and score scales $\tau \in \{0, 0.05, 0.1, 0.2, 0.4, 0.8, 2, 4\}$. The small-scale exponents are estimated from $\tau \in \{0.05, 0.1, 0.2\}$ and compared with the predicted powers $2s - 4$ and $2s - 2$ for the score and value kernels. For population training, we fix $P = 6$, use $s \in \{3, 4\}$, $\tau \in \{0, 2, 4\}$, and $M \in \{64, 256, 1024\}$. We record population excess cross-entropy, mean-squared and maximum absolute prediction error, target Walsh coefficients of the margin and predicted probability, maximum single-head movement, relative kernel drift, and the target-direction kernel Rayleigh quotient. The fitted exponent errors are below $0.047$. Zero-score initialization retains the initial excess loss $0.082283$, whereas at $\tau = 4$ and $M = 1024$ the terminal excess losses are $0.072567$ and $0.076896$ for $s = 3$ and $s = 4$, respectively.

**Finite samples and incomplete input support.** Inputs are sampled independently and uniformly from the Boolean cube, and labels are drawn independently from $q(z) = 1/2 + \rho_{\mathrm{p}} \chi_S(z)$. Repeated inputs are aggregated into empirical conditional frequencies and weighted according to their observed counts. We use $P \in \{6, 8\}$, $s \in \{3, 4\}$, $n_{\mathrm{p}} \in \{32, 128, 512, 2048\}$, $\tau = 4$, and $M = 256$. At the same initialization and checkpoints, we evaluate the full score-and-value linearized model defined by the initialization Jacobian. Recorded quantities include empirical and population excess loss, population MSE, seen- and unseen-input MSE, target Walsh coefficients, and parameter movement. Across recorded checkpoints, the mean population excess losses of the nonlinear and linearized models differ by at most $1.61 \times 10^{-3}$. For the incomplete-support evaluation, we use $P = 8$, $n_{\mathrm{p}} = 1024$, and $s \in \{3, 4\}$. Although mean input coverage exceeds $98\%$, terminal unseen-input MSE reaches $0.04583$ and $0.04272$, respectively, compared with $0.04$ at initialization. Thus high input coverage alone does not imply improved prediction on the remaining unseen inputs.

**Perturbed initialization.** To examine the local nonzero-output regime, we fix $P = 6$, $s = 3$, $n_{\mathrm{p}} = 128$, $\tau = 4$, and $M = 256$. Starting from the paired reference initialization, we apply random parameter perturbations with global radii $r_{\mathrm{init}} \in \{0, 0.5, 1, 2\}$, subject to a maximum initial perturbation norm of $0.25$ for each head. Training then follows the same integration and checkpoint schedule. We record the realized perturbation radius, maximum per-head perturbation, distance from the paired reference initialization, empirical and population prediction errors, and target Walsh coefficients. At $t = 200$, the mean target Walsh coefficient of the predicted probability lies between $0.01250$ and $0.01252$ across the four perturbation radii, while the mean excess losses differ by less than $7 \times 10^{-5}$ and remain near $0.1068$. In these tested settings, bounded nonzero-output perturbations preserve the observed target-component learning but do not remove finite-sample overfitting.

### U.7 Experiment 6: Planning protocol

We use a two-stage planning instance with two latent tasks under a uniform prior and binary actions and outcomes. The task-specific Bernoulli reward probabilities are $0.55$ and $0.55$ for action $0$, and $0.9$ and $0.1$ for action $1$, respectively. The immediate-reward greedy root action has value $1.1$, whereas the Bayes-optimal root action has value $1.185$. The predictive model uses $M = 128$ head pairs and is trained from the exact Bayes predictive conditionals and normalized visitation weights for master seeds $0, \ldots, 9$. The main runs use explicit Euler integration with step size $0.05$ through $t = 500$, with policy metrics monitored every $5$ time units. At every recorded checkpoint, we form the learned conditional outcome kernel and solve the resulting finite history tree exactly by backward induction; policy value is then evaluated under the true Bayes predictive kernel. We record the complete policy, its Bayes value suboptimality, and the maximum total-variation prediction error. For master seed $0$,

matched-time integration checks use Euler step sizes $dt \in \{0.05, 0.025, 0.0125\}$ and compare the resulting policies and values at identical training times.

## V NOTATION SUMMARY

| Symbol | Meaning |
|---|---|
| $\mathcal{E}, \nu$ | finite task family and prior |
| $\mathcal{Q}, n_{\mathrm{q}}$ | stage–action query set and its size |
| $P, z(h), j_\star(u)$ | number of fixed history features, feature vector, and relevant coordinate |
| $P^\star_{\mathrm{feat}}(\cdot \mid u, z)$ | target conditional distribution in the one-sparse logistic regression model |
| $\lambda, p_\lambda, m_\lambda$ | logistic regression coefficient, positive-outcome probability at feature value $+1$, and minimum sigmoid derivative |
| $B, \alpha_\star, \varepsilon_{\mathrm{sel}}$ | attention score matrix, relevant attention weight, and population attention mass on irrelevant coordinates |
| $N_{\mathrm{sel}}, \widehat{t}_{u,k}, \varepsilon_{\mathrm{stat}}$ | conditional outcome samples for each $(u, j, s)$, symmetrized empirical target probability, and sampling radius |
| $P^0_e, P^0_\nu$ | task-specific and prior predictive initial observation distributions |
| $h, a, y$ | decision history, action, and next reward–observation outcome |
| $K = \lvert\mathcal{Y}\rvert$ | number of outcome coordinates |
| $\nu(e \mid h), P_\nu(y \mid h, a)$ | Bayes posterior and Bayes predictive kernel |
| $\mathcal{H}_k, \mathcal{C} = \{c_i\}_{i=1}^m$ | reachable histories at each stage and finite reachable history–action set |
| $q_i, d_i$ | Bayes predictive distribution and normalized visitation weight |
| $A, v$ | trainable bilinear attention score matrix and value vector |
| $u_i, \bar{x}_i, C_i$ | last-token query vector, arithmetic mean of the embeddings, and within-context covariance |
| $Z_{\mathrm{ctx}}, Z_{\mathrm{oh}}, \zeta$ | context design matrices and minimum singular-value margin |
| $r = \mathrm{rank}(W)$ | rank of the shared embedding matrix |
| $\sigma_v$ | value-vector initialization scale |
| $\tau, \xi_{\mathrm{sc}}, a_{\mathrm{p}}, \rho_{\mathrm{p}}$ | score scale, random score-sign vector, target margin amplitude, and parity probability amplitude |
| $\lambda^A_s, \lambda^v_s, \kappa_s$ | normalized score, value, and binary-margin kernel eigenvalues at order $s$ |
| $\underline{\kappa}_s(\tau), n_{\mathrm{p}}, r_{\mathrm{init}}$ | explicit target-eigenvalue lower bound, independent sample count, and initialization-perturbation bound |
| $s_i(A), \varphi_i(A, v)$ | causal self-attention summary and scalar head output |
| $M$ | number of independent head pairs |
| $\mathbf{\Phi}_A(A, v)$ | Jacobian of the scalar head outputs across contexts with respect to the score-matrix parameters |
| $\boldsymbol{G}_A$ | expected Jacobian Gram matrix for the attention score matrices |
| $\mathbf{\Theta}_A(t), \mathbf{\Theta}(t)$ | neural tangent kernel contribution from the attention score matrices and full neural tangent kernel matrix |
| $E(t)$ | population excess cross-entropy |
| $\widehat{P}_t$ | learned predictive kernel |
| $\widehat{\pi}_t$ | policy computed by exact dynamic programming under $\widehat{P}_t$ |
| $V^\pi_\nu, V^\pi_{\mathsf{P}}$ | policy values under the task prior and a history-dependent stochastic kernel |
| $\mathrm{SubOpt}_H(\widehat{\pi}_t)$ | Bayes value suboptimality on a fresh test episode |