Preference-Based Reward Learning under Partial Observability with Inexact Dynamics

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we study how partial observability and inexact latent-state inference affect reward learning from preferences. To that end, we study preference-based reward learning under partial observability, where the learner forms latent-state estimates using an inexact learned POMDP model, so model error can accumulate over time. For finite log-linear POMDPs, we characterize this error term by establishing the stability of the belief filter to parametric model error under certain mixing conditions, yielding bounds on the belief mismatch in expectation and in high probability. We further extend this stability mechanism beyond the log-linear setting to neural-softmax POMDP models with overparameterized neural networks. We then propagate these errors into trajectory-level feature perturbations and derive finite-sample guarantees for constrained Bradley–Terry reward estimation from preferences. Our results decouple statistical error from an irreducible model-mismatch bias, and clarify when preference-based reward learning remains feasible under partial observability with imperfect dynamics.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) combines *preference-based reward learning* and *policy optimization*. A common pipeline collects pairwise (or K-wise) comparisons between trajectory segments, fits a parametric preference model, e.g., Bradley–Terry model (Bradley & Terry, 1952), to cumulative trajectory features, and then optimizes a policy using the learned reward (Christiano et al., 2017; Wirth et al., 2017; Kaufmann et al., 2025). Recent theory for this mechanism in fully observed MDPs analyzes regularized preference maximum likelihood estimation (MLE) under self-normalized concentration (Zhu et al., 2023; Du et al., 2024). Consider an interactive agent such as an LLM: the user-facing observations are text, clicks, or other surface signals, while the *real* state includes latent human factors (e.g., beliefs, goals, interpretations, accuracy and affective state) that are not accessible to the system. The learner must act through an inferred posterior belief rather than the true environment state. In particular, we argue that RLHF must recover an *unknown* reward from preference feedback while simultaneously operating under latent-state uncertainty.

This motivates modeling RLHF interaction as a Partially Observable Markov Decision Process (POMDP), where an unobserved latent state evolves Markovianly and the agent receives noisy observations. A standard way to obtain a Markovian control representation is via the induced belief-state process (the belief-MDP) (Åström, 1965; Kaelbling et al., 1998; Uesato et al., 2022). Accordingly, we lift the preference dataset (trajectory pairs with preference labels) into the belief-MDP representation by constructing beliefs along each recorded trajectory under the learned model and forming belief-based comparison features. However, this reduction introduces an additional, orthogonal difficulty: belief states are not directly observed and must be computed via Bayesian filtering. In practice, filtering uses an *inexact learned* POMDP model estimated from data, so preference features are evaluated on *approximate* belief trajectories, creating a structured, history-dependent model-mismatch channel. Therefore, approximation error propagates from learned dynamics to belief error, then to trajectory-feature perturbations, and finally to bias in the preference MLE (Figure 1).

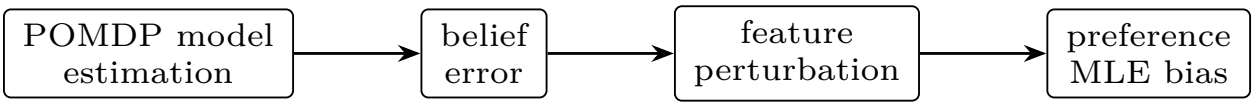


Figure 1: Error-propagation pipeline

The central technical obstacle is that Bayesian filtering is generally *not* contractive in total variation and may amplify errors over time; stability requires conditions that jointly control transition mixing and informativeness of the observation channel. Building on a recent stochastic filter stability theorem Mcdonald & Yüksel (2024), we establish expectation-level stability and quantify the bias introduced by parametric model mismatch at each filtering step. This analysis yields an explicit Lipschitz stability bound for the parameter-to-belief mapping, along with high-probability control over the time-averaged belief error using a martingale argument. We first develop this mechanism for finite log-linear POMDPs and then extend it to neural-softmax POMDP models in the kernel regime, where finite-width linearization errors enter additively into the belief perturbation bound. These results can then be extended to define a uniform belief-accuracy event applicable across all comparisons within an RLHF dataset.

We position our analysis at the interface of controlled filter stability and preference-based reward estimation, to quantify how learned-dynamics error in a POMDP propagates through belief updates into reward learning in the belief-MDP. This belief mismatch perturbs the trajectory-level feature differences used in preference comparisons and the associated empirical feature covariance. We propagate these perturbations through a Bradley–Terry MLE analysis, obtaining a finite-sample reward recovery bound that separates (i) a statistical term vanishing with the number of human comparisons N_{HF}, (ii) an irreducible model-mismatch bias controlled by filtering stability, and (iii) the regularization bias introduced to ensure well-posedness when the feature covariance is ill-conditioned or singular. This decomposition clarifies when preference-based RLHF remains statistically feasible under partial observability with an *inexact* learned POMDP model, and which stability and modeling choices govern the achievable accuracy.

Contributions.

- **A theoretical framework for preference-based reward learning under partial observability.** We establish, to the best of our knowledge, the first finite-sample reward-learning guarantees for preference-based RLHF under partial observability with inexact learned dynamics. We lift the preference dataset into the belief-MDP representation and provide an explicit error-propagation analysis from POMDP model mismatch to reward-estimation bias. This setting is not covered by existing fully observed RLHF theory, where trajectory features are assumed to be directly observed and exactly specified.

- **Belief stability under learned log-linear dynamics and conditional neural-softmax extensions.** Under partial observability, preference features are built from inferred beliefs rather than observed states, so learned-model error can propagate through the Bayesian filter into reward learning. Existing filter-stability results mainly address sensitivity to initialization error and do not directly quantify this recurring model-mismatch effect. We close this gap by proving that, for log-linear POMDP dynamics, the belief filter is Lipschitz-stable in expectation under parametric model error and a Dobrushin-type mixing condition (Theorem 4.1). The proof decomposes each one-step filtering error into observation-model perturbation, transition-model perturbation, and propagated belief error, yielding a stronger sufficient contraction condition because fresh model error is injected at every step. We further give a conditional neural-softmax extension under a local NTK-style linearization event and use martingale concentration to obtain a high-probability time-averaged belief-error event for the preference dataset.

- **Robust Bradley–Terry reward learning from approximate beliefs.** We derive a finite-sample guarantee for the constrained Bradley–Terry MLE (Theorem 5.8) that decomposes into a statistical error decaying as $O(N_{\mathrm{HF}}^{-1/2})$, an irreducible model-mismatch bias controlled by belief error, and a regularization bias. A key consequence is that, unlike in the fully observed setting, more human feedback cannot reduce the estimation error below the bias floor induced by imperfect dynamics, clarifying when and why reward learning degrades under partial observability.

Organization. Section 3 introduces the POMDP model, log-linear dynamics, Bayesian filtering, and the induced belief-MDP representation. Section 4 derives expectation-level and high-probability stability guarantees for Bayesian filtering under learned log-linear dynamics. Appendix C extends the belief-stability mechanism to neural-softmax models through an NTK-style linearization. Section 5 formalizes the preference-learning setup and establishes our main reward estimation guarantee under belief approximation error. Section 6 discusses implications, limitations, and future directions. Technical proofs are deferred to the appendix.

2 Related Work

Our setting draws on three adjacent literatures: learning under partial observability, stability of Bayesian filtering, and preference-based RLHF. POMDPs formalize decision-making with latent states and noisy observations (Åström, 1965), but are statistically hard to learn in full generality (Krishnamurthy et al., 2016; Jin et al., 2020), motivating work under structural assumptions and algorithmic frameworks for planning and learning under partial observability (Kaelbling et al., 1998; Du et al., 2019; Liu et al., 2022; Uehara et al., 2022; Guo et al., 2023; Cayci & Eryilmaz, 2025).

A closely related technical thread studies *filter stability* under initialization error. Classical analyses typically require strong mixing assumptions (Van Handel, 2009), while recent Dobrushin-type results provide explicit exponential stability by combining transition and observation contractions (Mcdonald & Yüksel, 2024). However, these results do not directly address the learned-model regime considered here, where parametric errors in both the transition and observation models introduce fresh perturbations at every Bayesian update rather than only an initial filtering discrepancy. This repeated error-injection mechanism prevents a direct application of existing initialization-stability bounds. Our stability theorem fills this gap by showing that, under a stronger sufficient Dobrushin condition, learned-model error induces a controlled and explicit belief-mismatch bound.

RLHF learns rewards from human feedback and then optimizes policies using standard RL machinery (Christiano et al., 2017; Wirth et al., 2017; Ji et al., 2023). Reward learning is often instantiated via Bradley–Terry model (Bradley & Terry, 1952; Plackett, 1975) and combined with policy optimization methods such as PPO (Schulman et al., 2017), underpinning successes in robotics and interactive systems (Brown et al., 2020) and LLM alignment (Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022; Achiam et al., 2023). Recent theory analyzes preference-based MLE and RLHF-style learning in fully observed MDPs, from tabular regimes to function approximation and exploration-driven policy optimization (Novoseller et al., 2020; Chen et al., 2022b; Zhu et al., 2023; Du et al., 2024; Kaufmann et al., 2025; Pacchiano et al., 2021).

A smaller line of work connects RLHF to latent-state structure and non-Markovian rewards. Most closely, Kausik et al. (2024) model RLHF via *partially-observed reward-states* (PORRL/PORMDP), where the environment state is observed but feedback depends on an additional latent "internal" reward-state. In contrast, we consider latent environment states with no access to rewards, and the agent interacts only through noisy observations and preference feedback. We quantify how belief computation under an *inexact learned* dynamics model perturbs trajectory features and induces a principled bias in Bradley–Terry reward estimation.

3 Setup and Preliminaries

3.1 POMDP model and trajectories

We work with a finite POMDP $\big(\mathbb{S},\ \mathbb{A},\ \hat{\mathbb{S}},\ P_\theta,\ \Phi_w,\ r,\ \nu_0\big)$, where $\mathbb{S}$ is the finite latent state space, $\hat{\mathbb{S}}$ is the finite observation space, $\mathbb{A}$ is the finite action set, $P_\theta(\cdot \mid s, a) \in \Delta_{\mathbb{S}}$ is the transition kernel, and $\Phi_w(\cdot \mid s) \in \Delta_{\hat{\mathbb{S}}}$ is the observation model. The reward function is $r : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$, and the initial state distribution $\nu_0 \in \Delta_{\mathbb{S}}$ is assumed known. The stability analysis below depends on the transition and observation channels only through their Dobrushin coefficients and the resulting explicit cardinality-dependent constants.

Assumption 3.1 (Model structure and bounded features)**.** Model dynamics admit the following log-linear structure

$$P_\theta(s'|s,a) = \frac{\exp\{\theta^\top \phi_p(s,a,s')\}}{\sum_{\bar{s}\in\mathbb{S}} \exp\{\theta^\top \phi_p(s,a,\bar{s})\}}, \qquad \Phi_w(\hat{s}\,|\,s) = \frac{\exp\{w^\top \phi_\Phi(s,\hat{s})\}}{\sum_{\bar{s}\in\hat{\mathbb{S}}} \exp\{w^\top \phi_\Phi(s,\bar{s})\}}.$$

Here ϕ_p is a feature map for the transition kernel and ϕ_Φ is a feature map for the observation model, and we assume

$$\sup_{s,a,s'} \|\phi_p(s,a,s')\|_2 \le B, \qquad \sup_{s,\hat{s}} \|\phi_\Phi(s,\hat{s})\|_2 \le B.$$

For some $B < \infty$. The initialization also satisfies

$$1 > \nu_{\max} \ge \nu_0(s) \ge \nu_{\min} > 0, \qquad \forall s \in \mathbb{S}. \tag{1}$$

Remark 3.2 (Exponential-family modeling and model-based RL). The log-linear parametrizations in Assumption 3.1 are finite exponential-family models, standard in statistical modeling (e.g., Bishop (2007); Wainwright & Jordan (2008); van Handel (2008)). One can use the assumed learned models for belief filtering and downstream policy optimization. In this work we focus on the preference-based reward learning guarantees, treating the policy-optimization stage as modular and leaving its integration to future work.

Assumption 3.3 (Realizability and local parameter error)**.** There exists a true parameter $\Theta^\star = (\theta^\star, w^\star) \in \mathbb{R}^{d_\theta + d_w}$ such that the data-generating POMDP has transition kernel $P_{\theta^\star}$ and observation model $\Phi_{w^\star}$.

Fix $r_\Theta > 0$, and suppose the learned parameter $\Theta = (\theta, w)$ satisfies

$$\Theta \in \mathbb{B}_2(\Theta^\star, r_\Theta) := \{\Theta' \in \mathbb{R}^{d_\theta + d_w} : \|\Theta' - \Theta^\star\|_2 \le r_\Theta\}.$$

Equivalently, $\|\Theta - \Theta^\star\|_2^2 = \|\theta - \theta^\star\|_2^2 + \|w - w^\star\|_2^2 \le r_\Theta^2$. We write $\delta(\theta) := \|\theta - \theta^\star\|_2$, $\delta(w) := \|w - w^\star\|_2$, so that $\delta(\theta)^2 + \delta(w)^2 \le r_\Theta^2$.

Given trajectories of observations and actions, a natural way to estimate the model parameter $\Theta = (\theta, w)$ in this class is via maximum-likelihood or an EM-type procedure. Fix a horizon $T \in \mathbb{N}$. For a single trajectory, let the latent state sequence be $Z := (s_{0:T}) \in \mathbb{S}^{T+1}$ and let the observation sequence be $Y := (\hat{s}_{1:T},\, a_{0:T-1}) \in \hat{\mathbb{S}}^T \times \mathbb{A}^T$. We treat the action sequence $a_{0:T-1}$ as observed and work with the likelihood conditional on actions. Under $\Theta = (\theta, w)$, the joint conditional trajectory likelihood factorizes as

$$P_\Theta(Y, Z) = \nu_0(s_0) \prod_{k=0}^{T-1} P_\theta(s_{k+1} \mid s_k, a_k) \prod_{k=1}^{T} \Phi_w(\hat{s}_k \mid s_k). \tag{2}$$

Given a collection of trajectories, one may estimate Θ by maximizing the marginal likelihood $P_\Theta(Y) = \sum_Z P_\Theta(Y, Z)$, for example, via EM. Under standard local identifiability and regularity conditions (e.g., strong concavity/smoothness of the standard EM auxiliary objective, i.e., the expected complete-data log-likelihood in a neighborhood of $\Theta^\star$), EM-type procedures admit local linear convergence to $\Theta^\star$ at the population level, with corresponding finite-sample perturbation bounds (Cappé et al., 2005; Wang & Zhao, 2006; van Handel, 2008; Balakrishnan et al., 2017).

Given the observation–action history up to time t, $Y_t := (\hat{s}_{1:t},\, a_{0:t-1})$, the corresponding belief is defined as

$$b_t^\Theta(s) \;:=\; \mathbb{P}_\Theta(s_t = s \mid Y_t) \in \Delta_{\mathbb{S}}, \qquad b_0^\Theta(s) = \nu_0(s).$$

For parameters $\Theta = (\theta, w)$, the Bayesian filter update is

$$b_{t+1}^\Theta(s') := \frac{\Phi_w(\hat{s}_{t+1} \mid s') \sum_{s\in\mathbb{S}} P_\theta(s' \mid s, a_t)\, b_t^\Theta(s)}{\sum_{\bar{s}\in\mathbb{S}} \Phi_w(\hat{s}_{t+1} \mid \bar{s}) \sum_{s\in\mathbb{S}} P_\theta(\bar{s} \mid s, a_t)\, b_t^\Theta(s)}. \tag{3}$$

Moreover, the induced belief-MDP reward can be written as

$$r_b(b_t^\Theta, a_t) := \mathbb{E}_{s_t \sim b_t^\Theta}\big[r(s_t, a_t)\big] = \sum_{s\in\mathbb{S}} b_t^\Theta(s)\, r(s, a_t).$$

The recursion Eq. 3 defines a nonlinear operator on $\Delta_{\mathbb{S}}$ obtained by composing a prediction step (via P_θ) with a Bayes reweighting step (via Φ_w). In general, this operator can amplify small perturbations in the model parameters or in the current belief over time. To obtain quantitative *stability* guarantees for the belief process, we impose an explicit *mixing* criterion that controls how strongly transitions forget the previous latent states. This is captured by the following Dobrushin coefficients, which will serve as the key constants governing one-step contraction and the resulting stability rate.

Definition 3.4 (Dobrushin coefficient)**.** Let $P_\theta(\cdot \mid s,a) \in \Delta_{\mathbb{S}}$ be the transition kernel and $\Phi_w \in \Delta_{\hat{\mathbb{S}}}$ the observation model. Define the uniform Dobrushin coefficient of the models over the feasible choice of parameter set by

$$\kappa_P := \inf_\theta \inf_{s,s',a} \sum_{x\in\mathbb{S}} \min\{P_\theta(x \mid s,a), P_\theta(x \mid s',a)\}, \qquad \kappa_\Phi := \inf_w \inf_{s,s'\in\mathbb{S}} \sum_{\hat{s}\in\hat{\mathbb{S}}} \min\{\Phi_w(\hat{s} \mid s), \Phi_w(\hat{s} \mid s')\}.$$

Equivalently,

$$\kappa_P = 1 - \sup_\theta \sup_{a,s,s'} \frac{1}{2}\|P_\theta(\cdot \mid s,a) - P_\theta(\cdot \mid s',a)\|_1.$$

Assumption 3.5 (Dobrushin stability)**.** Let $\kappa_P, \kappa_\Phi \in (0,1]$ be the uniform Dobrushin coefficients in Definition 3.4. Assume that

$$\alpha := (1-\kappa_P)(4-3\kappa_\Phi) < 1.$$

Remark 3.6 (Role of α). In the log-linear model, Lemma A.2 yields the explicit lower bounds

$$\kappa_P \geq \frac{|\mathbb{S}|}{1+(|\mathbb{S}|-1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)}, \qquad \kappa_\Phi \geq \frac{|\hat{\mathbb{S}}|}{1+(|\hat{\mathbb{S}}|-1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}.$$

Therefore, α can be controlled by the feature bound B and the feasible radii; for example, by introducing temperature parameters to control the magnitude of B, one can effectively reduce α. This condition is a non-trivial stability requirement since the Bayesian filtering update need not be a contraction in general (see Example 3.3, Mcdonald & Yüksel (2024)).

Practical note. Although κ_P and κ_Φ are worst-case quantities, in practice rather than plugging in analytic lower bounds Eq. 30 and Eq. 31 that must hold uniformly over all (s,s',a) in the feasible parameter set, we can estimate κ_P and κ_Φ directly from the realized transition and observation kernels. This yields a substantially less pessimistic stability certificate and better matches the observed contraction.

To show exponential filter stability in expectation for an *incorrectly initialized* POMDP, Mcdonald & Yüksel (2024) assumed $(1-\kappa_P)(2-\kappa_\Phi) < 1$. Building on this result, we show filter stability in expectation under POMDP model parameter perturbation. We set the assumption $(1-\kappa_P)(4-3\kappa_\Phi) < 1$ in Theorem 4.1. The stronger $(4-3\kappa_\Phi)$ factor arises because model mismatch injects fresh error at every step, as shown by terms (I)-(II) in Appendix B, and not just an initial discrepancy. This assumption is a *sufficient* technical bridge from POMDP model-mismatch to robust preference-based reward learning, as formalized in Theorem 5.8. We emphasize that this assumption is not claimed to be a *necessary* condition for the relation to hold; empirical performance may remain robust even when these worst-case mixing bounds are not strictly satisfied.

The next section quantifies this effect by proving that, under $\alpha < 1$, the belief mapping $\Theta \mapsto b_t^\Theta$ is Lipschitz in expectation around $\Theta^\star$ (Theorem 4.1). This stability bound is the key technical input for transferring POMDP parameter error into controlled feature perturbations in the downstream RLHF analysis (Theorem 5.8).

4 Belief Stability Under Model Mismatch

Let the true parameters be $\Theta^\star = (\theta^\star, w^\star)$ and the estimate $\Theta = (\theta, w)$, living in a compact feasible parameter set as in Assumption 3.3.

Theorem 4.1. *Consider the POMDP with log-linear dynamics, under Assumptions 3.1, 3.3, 3.5. Fix an action sequence $(a_0, \dots, a_{t-1})$, estimated model parameters $\Theta = (\theta, w)$, true parameters $\Theta^\star = (\theta^\star, w^\star)$ and corresponding belief processes $(b_k^\Theta)_{k=0}^t$ and $(b_k^{\Theta^\star})_{k=0}^t$ with priors $b_0^\Theta = b_0^{\Theta^\star} = \nu_0 \in \Delta_{\mathbb{S}}$. Then, the mapping from parameters $\Theta = (\theta, w) \to b_t^\Theta$ is Lipschitz at $\Theta^\star$ in expectation. In particular, for every $t \geq 1$,*

$$\mathbb{E}\Big[\|b_t^\Theta - b_t^{\Theta^\star}\|_{\mathrm{TV}}\Big] \leq \frac{(1-\alpha^t)}{1-\alpha} c_b\left(\delta(\theta), \delta(w)\right), \tag{4}$$

where $c_b\left(\delta(\theta), \delta(w)\right) := B\big(\,\delta(w) + \frac{(3-2\kappa_\Phi)}{2}\,\delta(\theta)\big)$, and the expectation is over $\hat{s}_{1:t} \sim P_{\Theta^\star}(\,\cdot\,|a_{0:t-1})$.

Proof sketch. The proof decomposes one-step belief error into (I) observation-model perturbation, (II) transition-model perturbation, and (III) propagation under filter mixing, then unrolls the resulting recursion. The detailed proof is provided in Appendix B.

Proof. The full proof is given in Appendix B. □

Remark 4.2 (Error in belief stability). The bound in Eq. 4 splits into a *time-propagation* factor and a *one-step mismatch* term:

$$\mathbb{E}\Big[\|b_t^\Theta - b_t^{\Theta^\star}\|_{\mathrm{TV}}\Big] \;\leq\; \underbrace{\frac{1-\alpha^t}{1-\alpha}}_{\text{propagation over time}} \times \underbrace{\Big(B\,\delta(w) + \tfrac{B}{2}(3-2\kappa_\Phi)\,\delta(\theta)\Big)}_{\text{model mismatch}}.$$

When $\alpha < 1$, the propagation factor is uniformly bounded by $(1-\alpha)^{-1}$, so the belief error scales linearly with the parameter deviations $\delta(w)$ and $\delta(\theta)$.

Remark 4.3 (Filter stability challenges). Analyzing filter stability over long horizons is notoriously difficult because the Bayesian update operator ψ is generally *not* a contraction and can expand the total variation distance between beliefs (see Mcdonald & Yüksel (2024), Example 3.3). Standard results often impose restrictive strong mixing conditions on the transition kernel $P(\cdot|\cdot,\cdot)$ and assume that it is sufficiently ergodic Van Handel (2009). Using the Dobrushin coefficient κ_Φ of the observation model can account for the *joint* contraction properties of the transition and measurement steps, allowing for stability even when the transition kernel alone is not sufficiently mixing. The price for this generality is that the contraction holds in *expectation* rather than almost surely. Moreover, while Mcdonald & Yüksel (2024) establishes stability with respect to incorrect *initialization*, our Theorem 4.1 extends this machinery to the learning setting. We establish stability with respect to *parametric model mismatch* (Θ vs. $\Theta^\star$), deriving explicit perturbation bounds specific to the log-linear family. This quantifies how parameter error injects bias at every step (terms (I) and (II) in Eq. 45 in Appendix B), a distinct challenge from the pure initialization decay studied in classical filtering literature.

Theorem 4.1 controls the belief error in *expectation* at a fixed time t under model mismatch. For the RLHF reduction we will need a *uniform-in-time* control along a trajectory, since downstream feature and preference observations depend on the entire history. We therefore convert the per-step expected contraction into a *high-probability bound* on the *time-average* belief error.

Definition 4.4 (Neural-softmax transition and observation models)**.** Let the transition and observation kernels be parameterized by neural-network scores

$$P_{W_p}(s' \mid s, a) = \frac{\exp\{F_p(s, a, s'; W_p)\}}{\sum_{\bar{s}\in\mathbb{S}} \exp\{F_p(s, a, \bar{s}; W_p)\}}\,, \qquad \Phi_{W_\Phi}(\hat{s} \mid s) = \frac{\exp\{F_\Phi(s, \hat{s}; W_\Phi)\}}{\sum_{\bar{s}\in\hat{\mathbb{S}}} \exp\{F_\Phi(s, \bar{s}; W_\Phi)\}} \tag{5}$$

We take both score functions to be two-layer ReLU networks with width m.

Corollary 4.5 (Belief perturbation bound for neural-softmax models)**.** *Consider the POMDP introduced earlier with neural-softmax transition and observation models in Definition 4.4, under Assumption C.3. Let $W = (W_p, W_\Phi) \in \mathcal{D}$ be an estimated neural parameter and let $W^\star = (W_p^\star, W_\Phi^\star) \in \mathcal{D}$ be the true parameter generating the POMDP. Consider an action sequence $(a_0, \dots, a_{t-1})$ and the corresponding belief processes*

$(b_k^W)_{k=0}^t$ and $(b_k^{W^\star})_{k=0}^t$, initialized from the same prior $b_0^W = b_0^{W^\star} = \nu_0 \in \Delta_{\mathbb{S}}$. Let κ_P^{NN} and $\kappa_\Phi^{\mathrm{NN}}$ be the uniform neural Dobrushin coefficients from Lemma C.6, and assume $\alpha_{\mathrm{NN}} := (1-\kappa_P^{\mathrm{NN}})(4-3\kappa_\Phi^{\mathrm{NN}}) < 1$. For any $\delta_{\mathrm{NN}} \in (0,1)$, on the event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$, and for every $t \geq 1$ it holds

$$\mathbb{E}\Big[\|b_t^W - b_t^{W^\star}\|_{\mathrm{TV}}\Big] \leq \frac{1-\alpha_{\mathrm{NN}}^t}{1-\alpha_{\mathrm{NN}}}\, c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}}), \tag{6}$$

where the expectation is taken with respect to $\hat{s}_{1:t} \sim P_{W^\star}(\,\cdot\,|a_{0:t-1})$, and

$$c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}}) := B^{\mathrm{NN}}\|W_\Phi - W_\Phi^\star\|_F + 2\varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}) \tag{7}$$

$$+ (3 - 2\kappa_\Phi^{\mathrm{NN}})\left(\frac{B^{\mathrm{NN}}}{2}\|W_p - W_p^\star\|_F + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}})\right). \tag{8}$$

In particular, as $m \to \infty$, this recovers a Lipschitz-type belief perturbation bound in the neural parameters.

Proof. The full proof is given in Appendix D. □

Proof sketch. The proof follows the same one-step decomposition as Theorem 4.1. The only changes are that the log-linear row-wise kernel perturbation bounds are replaced by Lemma C.5, and the Dobrushin constants are replaced by the neural constants κ_P^{NN} and $\kappa_\Phi^{\mathrm{NN}}$ from Lemma C.6. The resulting recursion has additive term $c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}})$ and contraction coefficient α_{NN}.

4.1 High-probability time-average belief error

The expectation bounds in Theorems 4.1 and Corollary 4.5 control the belief mismatch at each fixed time. For preference-based reward learning, however, the relevant objects are trajectory-level feature sums, and hence the error depends on the accumulated belief mismatch along the whole rollout. Therefore, it suffices to control the *time-average* belief error. The following corollary converts the one-step stability recursion into a high-probability time-average bound via a martingale argument. This bound will later be union-bounded over all trajectory pairs in the preference dataset to define a belief-accuracy event for the reward learning analysis.

Corollary 4.6. *Consider the setting of Theorem 4.1 and fix $\Theta = (\theta, w) \in \mathbb{B}_2(\Theta^\star, r_\Theta)$ such that $\|\theta - \theta^\star\|_2 \leq \delta(\theta)$ and $\|w - w^\star\|_2 \leq \delta(w)$. For $t = 0, 1, \ldots, T-1$ consider the random variables $X_t := \left\|b_t^\Theta - b_t^{\Theta^\star}\right\|_{\mathrm{TV}}$. For the filtrations $\{\mathcal{F}_t^-\}_{t\geq 0}$ and $\{\mathcal{F}_t\}_{t\geq 0}$ defined by $\mathcal{F}_t^- := \sigma(\hat{s}_{1:t}, a_{0:t-1}), \mathcal{F}_t := \sigma(\hat{s}_{1:t}, a_{0:t})$, with $\mathcal{F}_{-1}^- := \sigma(\varnothing, \Omega)$, the random variable X_t is $\mathcal{F}_t^-$-measurable. For any $\delta_b \in (0,1)$, with probability at least $1-\delta_b$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t \leq \frac{2c_b\left(\delta(\theta), \delta(w)\right)}{1-\alpha} + \frac{2}{1-\alpha}\sqrt{\frac{2c_b\left(\delta(\theta), \delta(w)\right)\log(1/\delta_b)}{T}}$$
$$+ \frac{\log(1/\delta_b)}{T(1-\alpha)}\left(\frac{4}{3} + \frac{2\alpha}{1-\alpha}\right) =: \epsilon_b(\delta_b). \tag{9}$$

Moreover, the same statement holds for the neural-softmax model of Theorem 4.5. In particular, for any $\delta_{\mathrm{NN}} \in (0,1)$, on the linearization event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$ from Definition C.2, if $X_t^{\mathrm{NN}} := \left\|b_t^W - b_t^{W^\star}\right\|_{\mathrm{TV}}$, and if $\alpha_{\mathrm{NN}} < 1$, then with conditional probability at least $1-\delta_b$, equivalently with joint probability at least $1-(\delta_b + \delta_{\mathrm{NN}})$,

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t^{\mathrm{NN}} \leq \frac{2c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}})}{1-\alpha_{\mathrm{NN}}} + \frac{2}{1-\alpha_{\mathrm{NN}}}\sqrt{\frac{2c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}})\log(1/\delta_b)}{T}}$$
$$+ \frac{\log(1/\delta_b)}{T(1-\alpha_{\mathrm{NN}})}\left(\frac{4}{3} + \frac{2\alpha_{\mathrm{NN}}}{1-\alpha_{\mathrm{NN}}}\right) \tag{10}$$
$$=: \epsilon_b^{\mathrm{NN}}(\delta_b, \delta_{\mathrm{NN}}), \tag{11}$$

where $c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}})$ is defined in Eq. 7.

Proof sketch. Starting from a one-step conditional drift inequality for X_t, define a martingale difference sequence and apply Freedman's inequality to control deviations of $\sum_t X_t$ from its conditional expectation, yielding the stated time-average bound.

Remark 4.7 (Learning Beliefs Directly). Our analysis is modular with respect to the source of the approximate beliefs. Although Sections 3–4 instantiate $\tilde{b}_t = b_t^{\Theta}$ through Bayesian filtering under a learned POMDP model, one may instead use a direct belief-inference module that maps histories $(\hat{s}_{1:t}, a_{0:t-1})$ to approximate beliefs. Recent examples include Deep Belief Markov Models Arcieri et al. (2026), which learn belief-transition and belief-inference operators via variational inference, and flow-based recurrent belief models, which use normalizing flows to represent flexible belief distributions (Chen et al., 2022a). This provides an alternative approach to bypass explicit POMDP model estimation when convenient. In our reward-learning analysis in the following section, such a module can be substituted for the Bayesian filter whenever it provides a belief-accuracy guarantee of the form $\frac{1}{T}\sum_{t=0}^{T-1} \|\tilde{b}_t - b_t^{\Theta^\star}\|_{\mathrm{TV}} \le \epsilon_b^{\mathrm{approx}}$.

for some approximation error $\epsilon_b^{\mathrm{approx}}$. Then the belief-approximation error enters the next section only through the trajectory-feature perturbation term, and we can replace $\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$ by $\epsilon_b^{\mathrm{approx}}$ to obtain the same bound on accumulated feature difference $\|\tilde{\phi}_i - \phi_i\|_2$, and the subsequent Bradley–Terry estimation analysis proceeds unchanged with this substituted error level.

5 Preference-Based Reward Learning Under Belief Error

Having established quantitative belief stability under model mismatch, we turn to learning a reward model from pairwise preferences when trajectories are represented in the belief-MDP. Note that the model mismatch affects reward learning through the induced belief features, since preference labels come from rollouts, but the learner's feature construction relies on b_t^{Θ} rather than the oracle beliefs $b_t^{\Theta^\star}$. Accordingly, this section formalizes the Bradley–Terry preference model (Bradley & Terry, 1952; Wirth et al., 2017) over belief-based trajectory features, and quantifies how the belief error bounds from Section 4 translate into an explicit perturbation (bias) term in reward estimation, alongside the usual finite-sample statistical error.

5.1 Reward model and preference data

Assumption 5.1 (Reward model structure and preference data for belief-MDP)**.**

1. **Linear latent reward.** For all $(s, a) \in \mathbb{S} \times \mathbb{A}$, the reward is linear: $r(s, a) = \phi_r(s, a)^\top \mu^\star$ with the corresponding feature map satisfying $\|\phi_r(s, a)\|_2 \le B_r < \infty$, and the unknown reward parameter $\mu^\star \in \mathbb{B}_2(r_\mu) \subset \mathbb{R}^d$.
2. **Belief feature map.** Define $\phi_b(b, a) := \sum_{s\in\mathbb{S}} b(s)\, \phi_r(s, a)$ for $(b, a) \in \Delta_{\mathbb{S}} \times \mathbb{A}$. Then $r_b(b, a) := \mathbb{E}_{s\sim b}\big[r(s, a)\big] = \sum_{s\in\mathbb{S}} b(s)\, \phi_r(s, a)^\top \mu^\star = \phi_b(b, a)^\top \mu^\star$.
3. **Preference realizability.** Let $\{(\tau_i^{(1)}, \tau_i^{(2)}, y_i)\}_{i=1}^{N_{\mathrm{HF}}}$ be N_{HF} independent preference comparisons with fixed trajectory horizon $T \in \mathbb{N}$. Here $y_i \in \{0, 1\}$ and $\sigma(z) = 1/(1 + e^{-z})$. Let $\phi_i \in \mathbb{R}^d$ be the clean trajectory-level feature difference defined in Eq. 13. Then, for every i, $\mathbb{P}(y_i = 1 \mid \phi_i) = \sigma(\phi_i^\top \mu^\star)$.

Remark 5.2. Since ϕ_b is linear in b, it inherits boundedness:

$$\|\phi_b(b, a)\|_2 \le \sum_{s\in\mathbb{S}} b(s)\, \|\phi_r(s, a)\|_2 \le B_r, \forall (b, a) \in \Delta_{\mathbb{S}} \times \mathbb{A}.$$

Moreover, for any $b, b' \in \Delta_{\mathbb{S}}$ and all $a \in \mathbb{A}$,

$$\|\phi_b(b, a) - \phi_b(b', a)\|_2 = \Big\| \sum_{s\in\mathbb{S}} \big(b(s) - b'(s)\big)\phi_r(s, a)\Big\|_2 \le B_r\|b - b'\|_1 = 2B_r\|b - b'\|_{\mathrm{TV}}. \tag{12}$$

Remark 5.3 (Generalization to non-linear rewards). Linear reward parameterizations are standard in preference-based RLHF and are widely used as a tractable baseline (see, e.g., Zhu et al. (2023); Du et al.

(2024); Cen et al. (2025)). Our analysis isolates how model-mismatch propagates through belief errors into perturbed trajectory features, which in turn affects reward learning in Theorem 5.8. More general reward function classes can be accommodated by introducing an approximation error term (relative to the linear class) and carrying it through the final bound. In particular, over-parameterized neural networks are a plausible choice as discussed in Appendix C for POMDP model approximation.

5.2 Clean vs. perturbed trajectory features and MLE

Throughout the following, we work on the belief state space $\Delta_{\mathbb{S}}$ and use the belief feature map ϕ_b. Let N_{HF} denote the number of *independent* preference comparisons, and consider a dataset $\{(\tau_i^{(1)}, \tau_i^{(2)}, y_i)\}_{i=1}^{N_{\mathrm{HF}}}$, where $y_i \in \{0,1\}$ is the observed preference label and $y_i = 1$ indicates a preference for $\tau_i^{(1)}$ over $\tau_i^{(2)}$. For each $i \in \{1, \dots, N_{\mathrm{HF}}\}$ and $j \in \{1,2\}$, let the observed history be $Y_i^{(j)} := \big(\hat{s}_{i,1:T}^{(j)}, a_{i,0:T-1}^{(j)}\big)$. and define the corresponding belief sequences $(b_{i,h}^{(j),\Theta})_{h=0}^{T-1}$ and $(b_{i,h}^{(j),\Theta^\star})_{h=0}^{T-1}$ as the filtering distributions computed under Θ and $\Theta^\star$, respectively, on the *same* history $Y_i^{(j)}$ via the Bayesian update Eq. 3. We then write the history-indexed belief trajectories as $\tau_i^{(j),\Theta} := \big(b_{i,h}^{(j),\Theta}, a_{i,h}^{(j)}\big)_{h=0}^{T-1}, \tau_i^{(j),\Theta^\star} := \big(b_{i,h}^{(j),\Theta^\star}, a_{i,h}^{(j)}\big)_{h=0}^{T-1}$, and the action sequence $(a_{i,h}^{(j)})_{h=0}^{T-1}$ is treated as observed. Throughout this work, we assume that all trajectories have a fixed horizon $T < \infty$.

Remark 5.4 (Single realized action sequence vs. fixed-history analysis). In an RLHF interaction there is only one physically realized history $(\hat{s}_{1:T}, a_{0:T-1})$. Our stability guarantees compare b_t^{Θ} and $b_t^{\Theta^\star}$ computed on this *same realized history* (i.e., conditioning on the realized action sequence and coupling on the realized observations). This fixed-history coupling is exactly what controls feature perturbations in Eq. 13–Eq. 14. Analyzing additional divergence caused by two models inducing different closed-loop action sequences is beyond scope.

Reward parameter learning from human preferences. We consider the estimation of the reward parameter $\mu^\star \in \mathbb{R}^d$ from human preference data under a trajectory-level Bradley–Terry model Bradley & Terry (1952). Human feedback is provided as pairwise trajectory comparisons $\{(\tau_i^{(1)}, \tau_i^{(2)}, y_i)\}_{i=1}^{N_{\mathrm{HF}}}$. For each comparison, we define the trajectory-level exact and perturbed accumulated feature differences as

$$\phi_i := \sum_{h=0}^{T-1} \phi_b\big(b_{i,h}^{(1),\Theta^\star}, a_{i,h}^{(1)}\big) - \sum_{h=0}^{T-1} \phi_b\big(b_{i,h}^{(2),\Theta^\star}, a_{i,h}^{(2)}\big), \tag{13}$$

$$\tilde{\phi}_i := \sum_{h=0}^{T-1} \phi_b\big(b_{i,h}^{(1),\Theta}, a_{i,h}^{(1)}\big) - \sum_{h=0}^{T-1} \phi_b\big(b_{i,h}^{(2),\Theta}, a_{i,h}^{(2)}\big). \tag{14}$$

The feature vector ϕ_i corresponds to the trajectory-level reward feature difference evaluated along the true belief process $(b_{i,h}^{(j),\Theta^\star})_h$. The perturbed feature $\tilde{\phi}_i$ is computed using approximate beliefs $(b_{i,h}^{(j),\Theta})_h$, where the differences arise from model mismatch. Lemma E.2 bounds $\|\phi_i - \tilde{\phi}_i\|_2$ on the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ in the Appendix. These two features induce conditional preference models; for example, $\tilde{\phi}_i$ induces $\mathbb{P}(y_i = 1 \mid \tilde{\phi}_i, \mu) = \sigma(\tilde{\phi}_i^\top \mu)$, with $\sigma(z) = 1/(1+e^{-z})$. We estimate the reward parameter using the constrained maximum likelihood estimator

$$\tilde{\mu} \;:=\; \operatorname*{arg\,min}_{\|\mu\|_2 \le r_\mu} \left\{ -\sum_{i=1}^{N_{\mathrm{HF}}} \Big[y_i \log \sigma(\tilde{\phi}_i^\top \mu) + (1-y_i) \log\big(1 - \sigma(\tilde{\phi}_i^\top \mu)\big) \Big] \right\}. \tag{15}$$

In the following analysis, we study the statistical behavior of $\tilde{\mu}$ and its robustness to trajectory-level perturbations arising from belief approximation. For clarity, the complete computational pipeline from recursive belief filtering to the optimization of the regularized estimator in Eq. 15 is summarized in Algorithm 1.

Remark 5.5 (Belief-based features). Classical Bradley–Terry reward learning typically assumes access to true states and uses state-based trajectory features (Wirth et al., 2017; Zhu et al., 2023; Du et al., 2019; Cen et al., 2025). Under partial observability, we instead construct ϕ_i and $\tilde{\phi}_i$ from belief–action pairs; hence model mismatch affects reward learning through belief-induced feature perturbations.

Algorithm 1 Preference-Based Reward Learning via Learned Beliefs

Input: Horizon T; preference data $\mathcal{D} = \{(Y_i^{(1)}, Y_i^{(2)}, y_i)\}_{i=1}^{N_{\mathrm{HF}}}$ with $Y_i^{(j)} = (\hat{s}_{i,1:T}^{(j)}, a_{i,0:T-1}^{(j)})$; labels $y_i \in \{0,1\}$; learned POMDP model $\Theta = (\theta, w)$; belief feature map ϕ_b; constraint radius r_μ.
Phase 1: Reconstruct Beliefs and Features
for $i = 1, \dots, N_{\mathrm{HF}}$ **do**
  for $j \in \{1,2\}$ **do**
    $b_{i,0}^{(j),\Theta} \leftarrow \nu_0$
    for $h = 0, \dots, T-1$ **do**
      $b_{i,h+1}^{(j),\Theta} \leftarrow \mathrm{BeliefUpdate}\Big(b_{i,h}^{(j),\Theta}, a_{i,h}^{(j)}, \hat{s}_{i,h+1}^{(j)}; \Theta\Big)$ (Eq. 3)
    end for
  end for
  Compute perturbed feature difference $\tilde{\phi}_i$ (Eq. 14):
  $\tilde{\phi}_i \leftarrow \sum_{h=0}^{T-1} \phi_b\Big(b_{i,h}^{(1),\Theta}, a_{i,h}^{(1)}\Big) \; - \; \sum_{h=0}^{T-1} \phi_b\Big(b_{i,h}^{(2),\Theta}, a_{i,h}^{(2)}\Big)$
end for
Phase 2: Maximum Likelihood Estimation
Solve for $\tilde{\mu}$ (Eq. 15):
$\tilde{\mu} \leftarrow \arg\min_{\|\mu\|_2 \le r_\mu} \; -\sum_{i=1}^{N_{\mathrm{HF}}} \Big[y_i \log \sigma(\tilde{\phi}_i^\top \mu) + (1-y_i) \log\big(1 - \sigma(\tilde{\phi}_i^\top \mu)\big)\Big]$
Output: Estimated reward parameter $\tilde{\mu}$

5.3 Belief accuracy event for preference learning

In this subsection we lift the time-average belief error guarantee from Corollary 4.6 to the full preference dataset. Note that we do *not* need a uniform-in-time control such as $\sup_{0\le h\le T-1} \|b_h^{\Theta} - b_h^{\Theta^\star}\|_{\mathrm{TV}}$. Since Bradley–Terry comparisons are driven by accumulated trajectory-level feature differences (13 and 14), the relevant quantity is the *per-trajectory time-average* belief mismatch, which then translates into a bound on the induced feature perturbation. We therefore define a high-probability event on which every rollout used to form the comparisons satisfies a small average belief error.

Belief accuracy event. Consider preference comparisons $\{(\tau_i^{(1)}, \tau_i^{(2)}, y_i)\}_{i=1}^{N_{\mathrm{HF}}}$. For any confidence level $\delta_b \in (0,1)$, define the belief-accuracy events for $i \in \{1, \dots, N_{\mathrm{HF}}\},\ j \in \{1,2\}$ as

$$\mathcal{E}_{b,i}^{(j)}(\delta_b) \;:=\; \left\{ \frac{1}{T} \sum_{h=0}^{T-1} \big\| b_{i,h}^{(j),\Theta} - b_{i,h}^{(j),\Theta^\star} \big\|_{\mathrm{TV}} \;\le\; \epsilon_b(\delta_b) \right\}, \quad \mathcal{E}_b^{\mathrm{RLHF}}(\delta_b) \;:=\; \bigcap_{i=1}^{N_{\mathrm{HF}}} \bigcap_{j\in\{1,2\}} \mathcal{E}_{b,i}^{(j)}\left(\frac{\delta_b}{2N_{\mathrm{HF}}}\right). \tag{16}$$

By Corollary 4.6 and a union bound, $\mathbb{P}\big(\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)\big) \;\ge\; 1-\delta_b$. Throughout the following analysis, we work on the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ for log-linear POMDP model and use the shorthand

$$\epsilon_b(\delta_b, 2N_{\mathrm{HF}}) := \; \epsilon_b\left(\frac{\delta_b}{2N_{\mathrm{HF}}}\right) = \frac{2c_b\left(\delta(\theta), \delta(w)\right)}{1-\alpha} + \frac{2}{1-\alpha}\sqrt{\frac{2c_b\left(\delta(\theta), \delta(w)\right)\log(2N_{\mathrm{HF}}/\delta_b)}{T}} + \frac{\log(2N_{\mathrm{HF}}/\delta_b)}{T(1-\alpha)}\left(\frac{4}{3} + \frac{2\alpha}{1-\alpha}\right). \tag{17}$$

That is, $\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$ is given by Eq. 9 with the substitution $\log\big(2N_{\mathrm{HF}}/\delta_b\big)$. Analogously, for the neural-softmax POMDP model under the setting of Theorem 4.5, fix $\delta_{\mathrm{NN}} \in (0,1)$. On the event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$, define $\mathcal{E}_b^{\mathrm{RLHF,NN}}(\delta_b, \delta_{\mathrm{NN}})$ similarly to Eq. 16, with $b_{i,h}^{(j),\Theta}$ and $b_{i,h}^{(j),\Theta^\star}$ replaced by $b_{i,h}^{(j),W}$ and $b_{i,h}^{(j),W^\star}$, respectively, and with ϵ_b replaced by the neural bound ϵ_b^{NN}. More explicitly, we use the shorthand

$$\epsilon_b^{\mathrm{NN}}(\delta_b, 2N_{\mathrm{HF}}, \delta_{\mathrm{NN}}) := \epsilon_b^{\mathrm{NN}}\left(\frac{\delta_b}{2N_{\mathrm{HF}}}, \delta_{\mathrm{NN}}\right),$$

where $\epsilon_b^{\mathrm{NN}}(\cdot, \delta_{\mathrm{NN}})$ is defined in Eq. 10 using the linearization-error levels $\varepsilon_p^{\mathrm{NN}}(m, \delta_{\mathrm{NN}})$ and $\varepsilon_\Phi^{\mathrm{NN}}(m, \delta_{\mathrm{NN}})$. By the neural high-probability belief bound and a union bound over the $2N_{\mathrm{HF}}$ trajectories, together with

$\mathcal{E}_{\text{lin}}^{\text{NN}}(\delta_{\text{NN}})$, we have

$$\mathbb{P}\Big(\mathcal{E}_b^{\text{RLHF,NN}}(\delta_b, \delta_{\text{NN}})\Big) \geq 1 - (\delta_b + \delta_{\text{NN}}).$$

Remark 5.6 (Preference realizability and conditional independence). Assumption 5.1 (item 3) assumes a correctly specified Bradley–Terry preference model, where each label y_i is generated from an independent trajectory pair and is conditionally independent given the corresponding clean feature difference ϕ_i, with $\mathbb{P}(y_i = 1 \mid \phi_i) = \sigma(\phi_i^\top \mu^\star)$. This provides the link from trajectory-level features to observable preference feedback. (See, also Proof E.1)

5.4 Reward estimation guarantee

We now state our main theorem for estimating the Bradley–Terry reward parameter $\mu^\star$. Our analysis is carried out in the *local geometry* induced by the empirical design covariance, which captures the curvature of the empirical MLE objective. Accordingly, we measure error in the covariance-weighted norm $\|\cdot\|_{\tilde{\Sigma}+\zeta I}$, (See Definition 5.7), which is a natural scale for evaluating the parameter optimality gap. However, our learner features are *perturbed* through belief mismatch, so the usual self-normalized arguments Zhu et al. (2023); Abbasi-yadkori et al. (2011) cannot be applied directly to the gradient formed with $\tilde{\phi}_i$. This introduces an additional technical difficulty; see Remark F.1.

Definition 5.7 (Clean and perturbed empirical covariances)**.** Given trajectory-level feature differences $\{\phi_i\}_{i=1}^{N_{\text{HF}}}$ and their perturbed counterparts $\{\tilde{\phi}_i\}_{i=1}^{N_{\text{HF}}}$, define the corresponding empirical covariance matrices by

$$\Sigma := \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top, \qquad \tilde{\Sigma} := \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \tilde{\phi}_i \tilde{\phi}_i^\top. \tag{18}$$

Theorem 5.8 (Reward parameter estimation under belief error)**.** *Assume the conditions of Theorem 4.1 and Assumption 5.1. Let $\Theta = (\theta, w) \in \mathbb{B}_2(\Theta^\star, r_\Theta)$ and let $\{(Y_i^{(1)}, Y_i^{(2)}, y_i)\}_{i=1}^{N_{\text{HF}}}$ be N_{HF} independent preference comparisons. Form perturbed trajectory feature differences $\{\tilde{\phi}_i\}_{i=1}^{N_{\text{HF}}}$ via Eq. 14 using beliefs $(b_{i,h}^{(j),\Theta})_{h=0}^{T}$ computed from Eq. 3, and let $\tilde{\mu}$ be the constrained Bradley–Terry MLE in Eq. 15. Fix $\delta_b, \delta_c \in (0,1)$. Then, for any factor $c_\zeta > 1$, with probability at least $1 - (\delta_b + \delta_c)$, the estimation error satisfies*

$$\begin{aligned}
\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I} \leq & \frac{2\sqrt{c_\zeta}}{\rho\sqrt{N_{\text{HF}}(c_\zeta - 1)}} \sqrt{d \log\left(1 + \frac{4T^2 B_r^2}{\zeta d}\right) + 2\log\left(\frac{1}{\delta_c}\right)} \\
& + \frac{8TB_r\, \epsilon_b(\delta_b, 2N_{\text{HF}})}{\rho\sqrt{\zeta}} \left(1 + \tfrac{1}{2} TB_r\, r_\mu\right) + 2r_\mu\sqrt{\zeta} && (19) \\
:= & \ \epsilon(\delta_b, \delta_c, \zeta), && (20)
\end{aligned}$$

where $\rho = 1/\big(2 + \exp(2TB_r r_\mu) + \exp(-2TB_r r_\mu)\big)$, and $\zeta := c_\zeta 16T^2 B_r^2 \epsilon_b(\delta_b, 2N_{\text{HF}})(1 + \epsilon_b(\delta_b, 2N_{\text{HF}}))$.

The same conclusion holds for the neural-softmax POMDP model of Theorem 4.5. In that case, additionally fix $\delta_{\text{NN}} \in (0,1)$. With probability at least $1-(\delta_b+\delta_c+\delta_{\text{NN}})$, the bound Eq. 20 holds after replacing $b_{i,h}^{(j),\Theta}, b_{i,h}^{(j),\Theta^\star}$ by $b_{i,h}^{(j),W}, b_{i,h}^{(j),W^\star}$, respectively, and replacing $\epsilon_b(\delta_b, 2N_{\text{HF}})$ everywhere, including in the definition of ζ, by $\epsilon_b^{\text{NN}}(\delta_b, 2N_{\text{HF}}, \delta_{\text{NN}})$.

Proof sketch. We show $\tilde{L}$ is ρ-strongly convex on $\mathbb{B}_2(r_\mu)$ in the $(\tilde{\Sigma} + \zeta I)$-geometry, and then control the estimation error by bounding the gradient at $\mu^\star$ $\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I} \leq \frac{2}{\rho} \|\nabla \tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}} + 2r_\mu\sqrt{\zeta}$. Then, we decompose $\nabla \tilde{L}(\mu^\star)$ into (i) a clean term handled by an elliptical-potential bound, and (ii) a perturbation term controlled on $\mathcal{E}_b^{\text{RLHF}}(\delta_b)$. Finally, we transfer the bound from the clean norm $(\Sigma + \zeta I)$ to the perturbed norm $(\tilde{\Sigma} + \zeta I)$ by an inverse-comparison argument. More details on ζ is provided in Remarks F.3-F.4. □

Proof. The full proof is given in Appendix F. □

Remark 5.9 (The cost of partial observability and error decomposition). Compare Theorem 5.8 to the oracle fully observed setting, where the learner observes the true state and thus eliminates the belief-induced model-mismatch bias terms. In that case one recovers the standard $\tilde{O}\left(\sqrt{\frac{d}{\rho^2 N_{\mathrm{HF}}}}\right)$ decay rate (e.g., Lemma 5.1 in Zhu et al. (2023)). In contrast, under partial observability with a fixed learned dynamics model, the bound contains additional terms governed by the belief-approximation level rather than by the preference-sample size, apart from the logarithmic dependence introduced by the union bound over the $2N_{\mathrm{HF}}$ trajectories. Therefore, increasing N_{HF} reduces the statistical term but cannot remove the error floor induced by a fixed inaccurate dynamics model; this floor is controlled by the belief-error level, up to logarithmic confidence factors. This separates the benefit of more feedback from the price of operating with approximate beliefs. More precisely, the bound in Theorem 5.8 separates three contributions to the parameter error. Up to logarithmic factors,

$$\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I} \lesssim \underbrace{\mathcal{O}\left(\frac{1}{\sqrt{N_{\mathrm{HF}}}}\right)}_{\text{Statistical Noise}} + \underbrace{\mathcal{O}\left(\frac{TB_r}{\sqrt{\zeta}}\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})\right)}_{\text{Model-Mismatch Bias}} + \underbrace{\mathcal{O}(r_\mu\sqrt{\zeta})}_{\text{Regularization Bias}} .$$

6 Conclusion

We analyzed preference-based reward learning in belief-MDPs under partial observability, where the true environment state is latent, and the agent instead acts on belief states obtained by Bayesian filtering from observation histories under an inexact learned POMDP. For finite log-linear POMDPs, we proved explicit stability of the filtering recursion to parametric model mismatch under a Dobrushin-type mixing condition, yielding bounds on $\mathbb{E}[\|b_t^{\Theta} - b_t^{\Theta^\star}\|_{\mathrm{TV}}]$ and high-probability time-average control. We further extended the same belief-stability mechanism to neural-softmax POMDP models through an NTK-style linearization, where finite-width linearization errors enter additively into the belief perturbation bound. We then propagated belief mismatch into trajectory-level feature perturbations and derived finite-sample guarantees for Bradley–Terry reward estimation from preferences, decomposing the error upper bound into an $N_{\mathrm{HF}}^{-1/2}$ statistical term, an irreducible model-mismatch bias controlled by the belief error level, and a regularization bias governed by ζ.

A limitation is that our filter comparison treats the action sequence as given and studies stability under a common logged action–observation history, which is the relevant regime for offline RLHF on a fixed dataset Kaufmann et al. (2025). Also, the Dobrushin contraction assumption can be conservative in weakly mixing or nearly deterministic regimes and may only be valid in a local neighborhood of the true parameters. For the neural-softmax extension, the guarantee additionally relies on a lazy-training regime, so the resulting constants depend on the width-dependent approximation errors.

As future directions, one can consider end-to-end closed-loop guarantees that couple belief-filter error with policy-induced trajectory drift, develop weaker stability notions beyond uniform Dobrushin contraction, and extend the analysis to broader learned-dynamics function classes with explicit approximation terms. A natural complementary module is a POMDP parameter-estimation stage, for either log-linear or neural transition and observation models, with finite-sample error control, enabling a fully modular pipeline in which estimation, filtering stability, preference-based reward learning, and policy optimization in the continuous belief-MDP are analyzed and improved componentwise.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alekh Agarwal, Sham Kakade, Mikael Henaff, and Wen Sun. Pc-pg: policy cover directed exploration for provable policy gradient learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Giacomo Arcieri, Konstantinos G. Papakonstantinou, Daniel Straub, and Eleni Chatzi. Deep belief markov models for pomdp inference. *Neural Networks*, 196:108386, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108386. URL https://www.sciencedirect.com/science/article/pii/S0893608025012675.

K. J. Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(65)90154-X. URL https://www.sciencedirect.com/science/article/pii/0022247X6590154X.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435. URL https://doi.org/10.1214/16-AOS1435.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117:30063 – 30070, 2019. URL https://api.semanticscholar.org/CorpusID:195700154.

Mikhail Belkin, Daniel J. Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116: 15849 – 15854, 2018. URL https://api.semanticscholar.org/CorpusID:198496504.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pp. 330–359. PMLR, 2020.

O. Cappé, E. Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, Germany, 2005.

Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for POMDPs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=6G01e0vgIf.

Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=BkEqk7pS1I.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SQnitDuow6.

Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, and Jianyu Chen. Flow-based recurrent belief state learning for POMDPs. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3444–3468. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/chen22q.html.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022b.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Exploration-driven policy optimization in RLHF: Theoretical insights on efficient data utilization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11830–11887. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/du24i.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2959268.

Jiacheng Guo, Zihao Li, Huazheng Wang, Mengdi Wang, Zhuoran Yang, and Xuezhou Zhang. Provably efficient representation learning with tractable planning in low-rank pomdp. In *International Conference on Machine Learning*, pp. 11967–11997. PMLR, 2023.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Ziwei Ji, Justin D. Li, and Matus Telgarsky. Early-stopped neural networks are consistent. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=rMKTq-ca0qu.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=f7OkIurx4b. Survey Certification.

Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. A theoretical framework for partially observed reward-states in rlhf. *arXiv preprint arXiv:2402.03282*, 2024.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2387337ba1e0b0249ba90f55b2ba2521-Paper.pdf.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf.

Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *Advances in Neural Information Processing Systems*, 35:18296–18308, 2022.

Curtis Mcdonald and Serdar Yüksel. Stochastic Observability and Filter Stability Under Several Criteria. *IEEE Transactions on Automatic Control*, 69(5), 2024. doi: 10.1214/20-ECP333.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1029–1038. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/novoseller20a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1:84–105, 2019. URL https://api.semanticscholar.org/CorpusID:61153563.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Siddhartha Satpathi, Harsh Gupta, Shiyu Liang, and R. Srikant. The role of regularization in overparameterized neural networks. In *CDC*, pp. 4683–4688, 2020. URL https://doi.org/10.1109/CDC42340.2020.9304386.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *Advances in Neural Information Processing Systems*, 35:578–592, 2022.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Ramon van Handel. Hidden markov models. *Unpublished lecture notes*, 2008.

Ramon Van Handel. Observability and nonlinear filtering. *Probability theory and related fields*, 145(1):35–74, 2009.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL `http://dx.doi.org/10.1561/2200000001`.

Shaojun Wang and Yunxin Zhao. Almost sure convergence of titterington's recursive estimator for mixture models. *Statistics and Probability Letters*, 76(18):2001–2006, 2006. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2006.05.017. URL `https://www.sciencedirect.com/science/article/pii/S0167715206001842`.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64, 11 2016. doi: 10.1145/3446776.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/zhu23f.html`.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Appendix

Lemma A.1. *Given model dynamic assumption as in 3.1, and finite feature bounds $B < \infty$, for all feasible parameter pairs (θ, w), (θ', w') with $\|\theta - \theta'\|_2 \le \delta(\theta)$, $\|w - w'\|_2 \le \delta(w)$:*

$$\|P_\theta(\cdot|s,a) - P_{\theta'}(\cdot|s,a)\|_1 \le B\delta(\theta), \qquad \forall (s,a) \in \mathbb{S} \times \mathbb{A}, \tag{21}$$

$$\|\Phi_w(\cdot \mid s) - \Phi_{w'}(\cdot \mid s)\|_1 \le B\delta(w), \qquad \forall s \in \mathbb{S}, \tag{22}$$

$$\|P_\theta(s'|\cdot,a) - P_{\theta'}(s'|\cdot,a)\|_1 \le 2B \frac{|\mathbb{S}|}{1 + (|\mathbb{S}|-1)\exp\big(-2B(\|\theta^\star\|_2 + r_\Theta)\big)}\,\delta(\theta), \qquad \forall (a,s') \in \mathbb{A} \times \mathbb{S}, \tag{23}$$

$$\big\|\Phi_w(\hat{s} \mid \cdot) - \Phi_{w'}(\hat{s} \mid \cdot)\big\|_1 \le 2B \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}|-1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)}\,\delta(w), \qquad \forall \hat{s} \in \hat{\mathbb{S}}. \tag{24}$$

Proof. Fix $(s,a) \in \mathbb{S} \times \mathbb{A}$ and consider the probability vector $P_\theta(\cdot \mid s,a) \in \Delta_{\mathbb{S}}$. Denote the Jacobian of $\theta \to P_\theta(\cdot|s,a)$ by $J_{a,s}(\theta) \in \mathbb{R}^{|\mathbb{S}| \times d_\theta}$ where d_θ is the dimension of θ, and the s'-th row is $\nabla_\theta P_\theta(s'|s,a)^\top$. We first establish the row-wise operator bound in 21. For $v \in \mathbb{R}^{d_\theta}$ with $\|v\|_2 = 1$, we have

$$\begin{aligned}
\|J_{a,s}(\theta)v\|_1 &= \sum_{s' \in \mathbb{S}} \Big| \big(J_{a,s}(\theta)v\big)_{s'} \Big| \\
&= \sum_{s' \in \mathbb{S}} \left| P_\theta(s' \mid s,a) \left(\phi_p(s,a,s')^\top v - \sum_{s'' \in \mathbb{S}} P_\theta(s'' \mid s,a) \phi_p(s,a,s'')^\top v \right) \right| \\
&= \sum_{s' \in \mathbb{S}} P_\theta(s' \mid s,a) \left| \phi_p(s,a,s')^\top v - \sum_{s'' \in \mathbb{S}} P_\theta(s'' \mid s,a) \phi_p(s,a,s'')^\top v \right| \\
&= \mathbb{E}_{S' \sim P_\theta(\cdot|s,a)} \Big[\Big| \phi_p(s,a,S')^\top v - \mathbb{E}_{\bar{S} \sim P_\theta(\cdot|s,a)} \big[\phi_p(s,a,\bar{S})^\top v \big] \Big| \Big] \\
&\le B,
\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz and the fact that

$$\mathbb{E}\Big[\big| \phi_p(s,a,S')^\top v - \mathbb{E}[\phi_p(s,a,S')^\top v] \big| \Big] \le \sqrt{\mathrm{Var}(\phi_p(s,a,S')^\top v)} \le B.$$

Therefore

$$\|J_{a,s}(\theta)\|_{2 \to 1} = \sup_{\|v\|_2 = 1} \|J_{a,s}(\theta)v\|_1 \le B,$$

and by the mean-value theorem, we have

$$\begin{aligned}
\big\| P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a) \big\|_1 &\le \sup_{\hat{\theta} \text{ between } \theta \text{ and } \theta'} \|J_{a,s}(\hat{\theta})\|_{2 \to 1} \, \|\theta - \theta'\|_2 \\
&\le B\delta(\theta).
\end{aligned}$$

The proof for $\Phi_w(\cdot \mid s)$ is identical, yielding $\|\Phi_w(\cdot \mid s) - \Phi_{w'}(\cdot \mid s)\|_1 \le B\delta(w)$.

We next compute the coordinate-wise gradient bound needed for the column-wise estimates. For $s' \in \mathbb{S}$,

$$\begin{aligned}
\Big\| \nabla_\theta P_\theta(s' \mid s,a) \Big\|_2 &= \Big\| \sum_{s'' \in \mathbb{S}} \frac{\partial \, P_\theta(s' \mid s,a)}{\partial(\theta^\top \phi_p(s,a,s''))} \, \nabla_\theta \theta^\top \phi_p(s,a,s'') \Big\|_2 \\
&= \Big\| \sum_{s'' \in \mathbb{S}} \big(\mathrm{Diag}(P_\theta(\cdot \mid s,a)) - P_\theta(\cdot \mid s,a) P_\theta(\cdot \mid s,a)^\top \big)_{s's''} \, \phi_p(s,a,s'') \Big\|_2 \\
&= \Big\| P_\theta(s' \mid s,a) \Big(\phi_p(s,a,s') - \sum_{s'' \in \mathbb{S}} P_\theta(s'' \mid s,a) \, \phi_p(s,a,s'') \Big) \Big\|_2 \\
&\le P_\theta(s' \mid s,a) \Big(\|\phi_p(s,a,s')\|_2 + \sum_{s''} P_\theta(s'' \mid s,a) \, \|\phi_p(s,a,s'')\|_2 \Big) \\
&\le P_\theta(s' \mid s,a) \, (B + B) \\
&= 2B \, P_\theta(s' \mid s,a).
\end{aligned} \tag{25}$$

where the first inequality follows by the triangle inequality and Jensen's inequality.

Then, we establish a column–wise Lipschitz bound for the transition kernel (23). Assume $\theta, \theta' \in \mathbb{B}_2(r_\Theta; \theta^\star)$ and fix $a \in \mathbb{A}$, $s' \in \mathbb{S}$. Let $J_{a,s'}(\theta) \in \mathbb{R}^{|\mathbb{S}| \times d_\theta}$ denote the Jacobian of $\theta \mapsto P_\theta(s' \mid \cdot, a)$, whose s–th row is $\nabla_\theta P_\theta(s' \mid s,a)^\top$. By 25, for every $s \in \mathbb{S}$,

$$\big\| \nabla_\theta P_\theta(s' \mid s,a) \big\|_2 \le 2B \, P_\theta(s' \mid s,a).$$

Therefore,

$$\begin{aligned}\|J_{a,s'}(\theta)\|_{2\to 1} &= \sup_{\|v\|_2=1} \sum_{s\in\mathbb{S}} \big|\langle \nabla_\theta P_\theta(s' \mid s,a), v\rangle\big| \\ &\le \sum_{s\in\mathbb{S}} \big\|\nabla_\theta P_\theta(s' \mid s,a)\big\|_2 \\ &\le 2B \sum_{s\in\mathbb{S}} P_\theta(s' \mid s,a).\end{aligned}$$

By Lemma A.2, the column–sum is uniformly bounded on $\mathbb{B}_2(r_\Theta;\theta^\star)$, hence

$$\sum_{s\in\mathbb{S}} P_\theta(s' \mid s,a) \le \frac{|\mathbb{S}|}{1+(|\mathbb{S}|-1)\exp\big(-2B(\|\theta^\star\|_2+r_\Theta)\big)},$$

and thus

$$\|J_{a,s'}(\theta)\|_{2\to 1} \le 2B\frac{|\mathbb{S}|}{1+(|\mathbb{S}|-1)\exp\big(-2B(\|\theta^\star\|_2+r_\Theta)\big)}.$$

Applying the mean–value theorem gives

$$\begin{aligned}\sum_{s\in\mathbb{S}} \big|P_\theta(s' \mid s,a) - P_{\theta'}(s' \mid s,a)\big| &= \Big\|\big(P_\theta(s' \mid s,a)\big)_{s\in\mathbb{S}} - \big(P_{\theta'}(s' \mid s,a)\big)_{s\in\mathbb{S}}\Big\|_1 \\ &\le \sup_{\hat\theta \text{ between } \theta,\theta'} \|J_{a,s'}(\hat\theta)\|_{2\to 1}\,\|\theta-\theta'\|_2 \\ &\le 2B\frac{|\mathbb{S}|}{1+(|\mathbb{S}|-1)\exp\big(-2B(\|\theta^\star\|_2+r_\Theta)\big)}\,\|\theta-\theta'\|_2.\end{aligned}$$

If, in addition, $\|\theta-\theta'\|_2 \le \delta(\theta)$, this yields

$$\sum_{s\in\mathbb{S}} \big|P_\theta(s' \mid s,a) - P_{\theta'}(s' \mid s,a)\big| \le 2B\frac{|\mathbb{S}|}{1+(|\mathbb{S}|-1)\exp\big(-2B(\|\theta^\star\|_2+r_\Theta)\big)}\,\delta(\theta), \qquad \forall\, s' \in \mathbb{S}.$$

To show the claim in 24, fix $\hat s \in \hat{\mathbb{S}}$ and consider $w \to \Phi_w(\hat s|\cdot)$ from $\mathbb{R}^{d_w}$. Following a similar argument, we apply the mean-value theorem to the aforementioned map and bound the operator norm $\|J_{\hat s}(w)\|_{2\to 1}$ of its Jacobian $J_{\hat s}(w) \in \mathbb{R}^{|\mathbb{S}|\times d_w}$. First, we bound the ℓ_2-norm of $\nabla_w \Phi_w(\hat s \mid s)^\top$ which is the s-th row of $J_{\hat s}(w)$,

$$\begin{aligned}\big\|\nabla_w \Phi_w(\hat s \mid s)\big\|_2 &= \|\Phi_w(\hat s \mid s)\Big(\phi_\Phi(s,\hat s) - \sum_{y\in\hat{\mathbb{S}}} \Phi_w(y \mid s)\,\phi_\Phi(s,y)\Big)\|_2 \\ &\le \Phi_w(\hat s \mid s)\Big(\|\phi_\Phi(s,\hat s)\|_2 + \sum_{y\in\hat{\mathbb{S}}} \Phi_w(y \mid s)\,\|\phi_\Phi(s,y)\|_2\Big) \\ &\le \Phi_w(\hat s \mid s)\,(B+B) = 2B\,\Phi_w(\hat s \mid s).\end{aligned}$$

Then, by Lemma A.2, inequality 28,

$$\begin{aligned}\|J_{\hat s}(w)\|_{2\to 1} &= \sup_{\|v\|_2=1} \sum_{s\in\mathbb{S}} \big|\langle \nabla_w \Phi_w(\hat s \mid s), v\rangle\big| \;\le\; \sum_{s\in\mathbb{S}} \big\|\nabla_w \Phi_w(\hat s \mid s)\big\|_2 \\ &\le 2B \sum_{s\in\mathbb{S}} \Phi_w(\hat s \mid s) \\ &\le 2B\frac{|\mathbb{S}|}{1+(|\hat{\mathbb{S}}|-1)\exp\big(-2B(\|w^\star\|_2+r_\Theta)\big)}.\end{aligned}$$

Now applying the mean-value theorem, for some $\tilde{w}$ on the line segment between w and w',

$$\begin{aligned}\left\|\Phi_w(\hat{s} \mid \cdot) - \Phi_{w'}(\hat{s} \mid \cdot)\right\|_1 &\leq \|J_{\hat{s}}(\tilde{w})\|_{2\to 1} \|w - w'\|_2 \\ &\leq 2B \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)} \delta(w),\end{aligned}$$

and the result follows. □

In the following lemma, for simplicity, we consider the parameter set to be $\mathbb{B}_2(\theta^\star, r_\Theta) \times \mathbb{B}_2(w^\star, r_\Theta)$. This result also holds for the feasible parameter set, $\mathbb{B}_2(\Theta^\star, r_\Theta)$, assumed in Assumption 3.3.

Lemma A.2. *Under Assumption 3.1, in the neighborhoods $\theta \in \mathbb{B}_2(r_\Theta; \theta^\star)$, $w \in \mathbb{B}_2(r_\Theta; w^\star)$, the following properties hold:*

$$\sup_{\theta \in \mathbb{B}_2(r_\Theta;\theta^\star)} \sup_{a \in \mathbb{A}} \sup_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} P_\theta(s' \mid s, a) \;\leq\; \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\big(-2B(\|\theta^\star\|_2 + r_\Theta)\big)}, \tag{26}$$

$$\inf_{\theta \in \mathbb{B}_2(r_\Theta;\theta^\star)} \inf_{a \in \mathbb{A}} \inf_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} P_\theta(s' \mid s, a) \;\geq\; \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)}. \tag{27}$$

$$\sup_{w \in \mathbb{B}_2(r_\Theta;w^\star)} \sup_{\hat{s} \in \hat{\mathbb{S}}} \sum_{s \in \mathbb{S}} \Phi_w(\hat{s} \mid s) \;\leq\; \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)}, \tag{28}$$

$$\inf_{w \in \mathbb{B}_2(r_\Theta;w^\star)} \inf_{\hat{s} \in \hat{\mathbb{S}}} \sum_{s \in \mathbb{S}} \Phi_w(\hat{s} \mid s) \;\geq\; \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}. \tag{29}$$

Furthermore, the uniform Dobrushin coefficients in Definition 3.4 satisfy

$$\kappa_P \;\geq\; \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)}, \tag{30}$$

$$\kappa_\Phi \;\geq\; \frac{|\hat{\mathbb{S}}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}. \tag{31}$$

Proof of Lemma A.2. For any $(s, a, s') \in \mathbb{S} \times \mathbb{A} \times \mathbb{S}$ and any $\theta \in \mathbb{B}_2(r_\Theta; \theta^\star)$, we have

$$|\theta^\top \phi_p(s, a, s')| \leq \|\theta\|_2 \, \|\phi_p(s, a, s')\|_2 \leq B(\|\theta^\star\|_2 + r_\Theta).$$

Hence,

$$\exp\big(-B(\|\theta^\star\|_2 + r_\Theta)\big) \leq \exp(\theta^\top \phi_p(s, a, s')) \leq \exp\big(B(\|\theta^\star\|_2 + r_\Theta)\big). \tag{32}$$

Using the fact that the numerator term also appears in the denominator, we have

$$\begin{aligned}P_\theta(s' \mid s, a) &= \frac{\exp(\theta^\top \phi_p(s, a, s'))}{\sum_{\bar{s} \in \mathbb{S}} \exp(\theta^\top \phi_p(s, a, \bar{s}))} \\ &\leq \frac{\exp\big(B(\|\theta^\star\|_2 + r_\Theta)\big)}{\exp\big(B(\|\theta^\star\|_2 + r_\Theta)\big) + (|\mathbb{S}| - 1)\exp\big(-B(\|\theta^\star\|_2 + r_\Theta)\big)} \\ &= \frac{1}{1 + (|\mathbb{S}| - 1)\exp\big(-2B(\|\theta^\star\|_2 + r_\Theta)\big)}.\end{aligned} \tag{33}$$

Summing over $s \in \mathbb{S}$ for fixed (a, s') yields

$$\sum_{s \in \mathbb{S}} P_\theta(s' \mid s, a) \leq \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\big(-2B(\|\theta^\star\|_2 + r_\Theta)\big)}.$$

Taking the supremum over (θ, a, s') proves 26. Similarly,

$$P_\theta(s' \mid s, a) \geq \frac{\exp\big(-B(\|\theta^\star\|_2 + r_\Theta)\big)}{\exp\big(-B(\|\theta^\star\|_2 + r_\Theta)\big) + (|\mathbb{S}| - 1)\exp\big(B(\|\theta^\star\|_2 + r_\Theta)\big)}$$
$$= \frac{1}{1 + (|\mathbb{S}| - 1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)} =: P_{\min}. \tag{34}$$

Taking the infimum over (θ, a, s') and summing $P_{\min}$ over $s \in \mathbb{S}$ proves 27. For the Dobrushin coefficient of P_θ in 30, every entry of $P_\theta(\cdot \mid s, a)$ is bounded from below by $P_{\min}$. Hence,

$$\sum_{s' \in \mathbb{S}} \min\{P_\theta(s' \mid s, a), P_\theta(s' \mid s'', a)\} \geq |\mathbb{S}|\, P_{\min}$$
$$= \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)}. \tag{35}$$

Taking the infimum over $\theta \in \mathbb{B}_2(r_\Theta; \theta^\star)$, $a \in \mathbb{A}$, and $s, s'' \in \mathbb{S}$ yields 30.

Then, we prove 29 and 28, which provide uniform column-sum bounds for the observation model Φ_w. The proof is similar to the derivation of 26 and 27 for P_θ. Fix $(s, \hat{s}) \in \mathbb{S} \times \hat{\mathbb{S}}$ and let $w \in \mathbb{B}_2(r_\Theta; w^\star)$. By the feature bound $\|\phi_\Phi(s, y)\|_2 \leq B$ and the radius assumption,

$$|w^\top \phi_\Phi(s, y)| \leq \|w\|_2\, \|\phi_\Phi(s, y)\|_2 \leq B(\|w^\star\|_2 + r_\Theta), \qquad \forall y \in \hat{\mathbb{S}}.$$

Hence, for all $y \in \hat{\mathbb{S}}$,

$$\exp\big(-B(\|w^\star\|_2 + r_\Theta)\big) \leq \exp\big(w^\top \phi_\Phi(s, y)\big) \leq \exp\big(B(\|w^\star\|_2 + r_\Theta)\big).$$

Recalling the log-linear observation model in Assumption 3.1, we obtain the uniform pointwise bounds

$$\Phi_w(\hat{s} \mid s) \leq \frac{\exp\big(B(\|w^\star\|_2 + r_\Theta)\big)}{\exp\big(B(\|w^\star\|_2 + r_\Theta)\big) + (|\hat{\mathbb{S}}| - 1)\exp\big(-B(\|w^\star\|_2 + r_\Theta)\big)}$$
$$= \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)} =: \Phi_{\max}, \tag{36}$$

$$\Phi_w(\hat{s} \mid s) \geq \frac{\exp\big(-B(\|w^\star\|_2 + r_\Theta)\big)}{\exp\big(-B(\|w^\star\|_2 + r_\Theta)\big) + (|\hat{\mathbb{S}}| - 1)\exp\big(B(\|w^\star\|_2 + r_\Theta)\big)}$$
$$= \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)} =: \Phi_{\min}. \tag{37}$$

Summing 36 and 37 over $s \in \mathbb{S}$ yields, for every fixed $\hat{s} \in \hat{\mathbb{S}}$,

$$\sum_{s \in \mathbb{S}} \Phi_w(\hat{s} \mid s) \leq \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)},$$

$$\sum_{s \in \mathbb{S}} \Phi_w(\hat{s} \mid s) \geq \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}.$$

Taking the supremum and infimum over $(w, \hat{s})$ gives 28 and 29, respectively. To obtain the Dobrushin coefficient in 31, using $\Phi_{\min}$ by the exact same argument used to obtain κ_P, we have

$$\sum_{\hat{s} \in \hat{\mathbb{S}}} \min\{\Phi_w(\hat{s} \mid s), \Phi_w(\hat{s} \mid s'')\} \geq |\hat{\mathbb{S}}|\, \Phi_{\min}$$
$$= \frac{|\hat{\mathbb{S}}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}. \tag{38}$$

Taking the infimum over $w \in \mathbb{B}_2(r_\Theta; w^\star)$ and $s, s'' \in \mathbb{S}$ proves 31. □

B Proof of Theorem 4.1

Remark B.1. First, recall 1 where we assumed $1 > \nu_{\max} \geq \nu_0(s) \geq \nu_{\min} > 0$. Plus, similar to the argument in Lemma A.2 that we have uniformly bounded $P_\theta(s'|s,a)$ in Eq. 33 and 34, and given the similar log-linear structure of models P_θ and Φ_w assumed in 3.1 and bounded feature maps in 3.1, we can establish uniform bounds over Φ_w and P_θ. Denote

$$P_{\min} := \frac{1}{1 + (|\mathbb{S}| - 1)\exp\big(2B(\|\theta^\star\|_2 + r_\Theta)\big)}, \qquad P_{\max} := \frac{1}{1 + (|\mathbb{S}| - 1)\exp\big(-2B(\|\theta^\star\|_2 + r_\Theta)\big)}, \tag{39}$$

$$\Phi_{\min} := \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(2B(\|w^\star\|_2 + r_\Theta)\big)}, \qquad \Phi_{\max} := \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\big(-2B(\|w^\star\|_2 + r_\Theta)\big)}. \tag{40}$$

where for all $(s, s', a, \hat{s})$ and $\Theta \in \mathbb{B}_2(\Theta^\star, r_\Theta)$ we have

$$0 < \nu_{\min} \;\leq\; \nu_0(s) \;\leq\; \nu_{\max}, \quad 0 < P_{\min} \;\leq\; P_\theta(s' \mid s, a) \;\leq\; P_{\max}, \quad 0 < \Phi_{\min} \;\leq\; \Phi_w(\hat{s} \mid s) \;\leq\; \Phi_{\max}. \tag{41}$$

Besides, define the prediction and observation operators

$$(P_\theta^{a_t} q)(s') = \sum_{s \in \mathbb{S}} P_\theta(s' \mid s, a_t)\, q(s), \quad \Delta_{\mathbb{S}} \to \Delta_{\mathbb{S}}, \tag{42}$$

$$(Q_w q)(\hat{s}) = \sum_{s' \in \mathbb{S}} \Phi_w(\hat{s} \mid s')\, q(s'), \quad \Delta_{\mathbb{S}} \to \Delta_{\hat{\mathbb{S}}}, \tag{43}$$

and the Bayesian update operator

$$\psi_{\Phi_w} : \Delta_{\mathbb{S}} \times \hat{\mathbb{S}} \to \Delta_{\mathbb{S}} \cup \{0\},$$
$$\psi_{\Phi_w}(b, \hat{s})(s') = \frac{\Phi_w(\hat{s}|s')b(s')}{\sum_{\bar{s} \in \mathbb{S}} \Phi_w(\hat{s}|\bar{s})b(\bar{s})} = \frac{\Phi_w(\hat{s}|s')b(s')}{(Q_w b)(\hat{s})}.$$

Then, the belief update 3 can be expressed as

$$b_{t+1}^\Theta = \psi_{\Phi_w}(P_\theta^{a_t} b_t^\Theta,\, \hat{s}_{t+1}).$$

Now, we can prove Theorem 4.1.

Proof of Theorem 4.1. For two parameter sets Θ and $\Theta^\star$, we have

$$b_{k+1}^\Theta = \psi_{\Phi_w}\big(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}\big), \qquad b_{k+1}^{\Theta^\star} = \psi_{\Phi_{w^\star}}\big(P_{\theta^\star}^{a_k} b_k^{\Theta^\star}, \hat{s}_{k+1}\big), \qquad b_0^\Theta = b_0^{\Theta^\star} = \nu_0.$$

We work under the conditional law $\mathbb{P}_{\Theta^\star}(\,\cdot\,|a_{0:t-1})$, i.e., the action sequence $(a_0, \ldots, a_{t-1})$ is treated as *fixed*. We use the filtrations

$$\mathcal{F}_k^- := \sigma(\hat{s}_{1:k}, a_{0:k-1}), \qquad \mathcal{F}_k := \sigma(\hat{s}_{1:k}, a_{0:k}),$$

so that $b_k^{\Theta^\star}$ and b_k^Θ are $\mathcal{F}_k^-$-measurable and a_k is $\mathcal{F}_k$-measurable. In particular, the conditional law of $\hat{s}_{k+1}$ given $\mathcal{F}_k$ is

$$\hat{s}_{k+1} \mid \mathcal{F}_k \sim Q_{w^\star}\Big(P_{\theta^\star}^{a_k} b_k^{\Theta^\star}\Big).$$

Fix $k \in \{0, \ldots, t-1\}$. Using the filter recursions $b_{k+1}^\Theta = \psi_{\Phi_w}\big(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}\big)$ and $b_{k+1}^{\Theta^\star} = \psi_{\Phi_{w^\star}}\big(P_{\theta^\star}^{a_k} b_k^{\Theta^\star}, \hat{s}_{k+1}\big)$, add and subtract the intermediate terms $\psi_{\Phi_{w^\star}}\big(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}\big)$ and $\psi_{\Phi_{w^\star}}\big(P_{\theta^\star}^{a_k} b_k^\Theta, \hat{s}_{k+1}\big)$:

$$\|b_{k+1}^\Theta - b_{k+1}^{\Theta^\star}\|_{\mathrm{TV}} = \big\|\psi_{\Phi_w}(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}) - \psi_{\Phi_{w^\star}}(P_{\theta^\star}^{a_k} b_k^{\Theta^\star}, \hat{s}_{k+1})\big\|_{\mathrm{TV}} \tag{44}$$

$$\leq \underbrace{\big\|\psi_{\Phi_w}(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}) - \psi_{\Phi_{w^\star}}(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1})\big\|_{\mathrm{TV}}}_{\text{(I)}} \tag{45}$$

$$+ \underbrace{\big\|\psi_{\Phi_{w^\star}}(P_\theta^{a_k} b_k^\Theta, \hat{s}_{k+1}) - \psi_{\Phi_{w^\star}}(P_{\theta^\star}^{a_k} b_k^\Theta, \hat{s}_{k+1})\big\|_{\mathrm{TV}}}_{\text{(II)}} \tag{46}$$

$$+ \underbrace{\big\|\psi_{\Phi_{w^\star}}(P_{\theta^\star}^{a_k} b_k^\Theta, \hat{s}_{k+1}) - \psi_{\Phi_{w^\star}}(P_{\theta^\star}^{a_k} b_k^{\Theta^\star}, \hat{s}_{k+1})\big\|_{\mathrm{TV}}}_{\text{(III)}}. \tag{47}$$

Term (I). This term corresponds to the perturbation of the observation model. First, consider the following observation for any generic $b \in \Delta_{|\mathbb{S}|}$:

$$
\begin{aligned}
\|\psi_{\Phi_w}(b,\hat{s}) - \psi_{\Phi_{w^\star}}(b,\hat{s})\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{s'\in\mathbb{S}} \left| \frac{\Phi_w(\hat{s} \mid s')\, b(s')}{\sum_{\bar{s}} \Phi_w(\hat{s} \mid \bar{s})\, b(\bar{s})} - \frac{\Phi_{w^\star}(\hat{s} \mid s')\, b(s')}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \right| \\
&\leq \frac{1}{2}\sum_{s'\in\mathbb{S}} \left| \frac{\big(\Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s')\big)\, b(s')}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \right| \\
&\qquad + \frac{1}{2}\sum_{s'\in\mathbb{S}} \left| \Phi_w(\hat{s} \mid s')\, b(s') \Big(\frac{1}{\sum_{\bar{s}} \Phi_w(\hat{s} \mid \bar{s})\, b(\bar{s})} - \frac{1}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \Big) \right| \\
&\leq \frac{1}{2\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right| \\
&\qquad + \frac{1}{2}\Big(\sum_{s'\in\mathbb{S}} \Phi_w(\hat{s} \mid s')\, b(s') \Big) \left| \frac{1}{\sum_{\bar{s}} \Phi_w(\hat{s} \mid \bar{s})\, b(\bar{s})} - \frac{1}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \right| \\
&= \frac{1}{2\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right| \\
&\qquad + \frac{1}{2} \frac{\left| \sum_{\bar{s}} \Phi_w(\hat{s} \mid \bar{s})\, b(\bar{s}) - \sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s}) \right|}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \\
&\leq \frac{1}{2\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right| \\
&\qquad + \frac{1}{2\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})} \sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right| \\
&= \frac{\sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right|}{\sum_{\bar{s}} \Phi_{w^\star}(\hat{s} \mid \bar{s})\, b(\bar{s})}. \qquad (48)
\end{aligned}
$$

Now we take the expectation under $\hat{s} \sim (Q_{w^\star} b)(\cdot)$. Since $(Q_{w^\star} b)(\cdot)$ is a probability distribution on $\hat{\mathbb{S}}$ with non-zero support, we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\hat{s}\sim(Q_{w^\star} b)} \left[\frac{\sum_{s'\in\mathbb{S}} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right|}{(Q_{w^\star} b)(\hat{s})} \right] &= \sum_{\hat{s}\in\hat{\mathbb{S}}} (Q_{w^\star} b)(\hat{s}) \, \frac{\sum_{s'} b(s') \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right|}{(Q_{w^\star} b)(\hat{s})} \\
&= \sum_{s'\in\mathbb{S}} b(s') \sum_{\hat{s}\in\hat{\mathbb{S}}} \left| \Phi_w(\hat{s} \mid s') - \Phi_{w^\star}(\hat{s} \mid s') \right| \\
&= \sum_{s'\in\mathbb{S}} b(s') \left\| \Phi_w(\cdot \mid s') - \Phi_{w^\star}(\cdot \mid s') \right\|_1 \\
&\leq \sup_{s'\in\mathbb{S}} \left\| \Phi_w(\cdot \mid s') - \Phi_{w^\star}(\cdot \mid s') \right\|_1 \\
&\leq B\, \delta(w), \qquad (49)
\end{aligned}
$$

where in the last step we used Lemma A.1. Now, we observe that for any measurable $f : \hat{\mathbb{S}} \to [0,1]$ and distributions ν, ν' on $\hat{\mathbb{S}}$,

$$
\mathbb{E}_{\hat{s}\sim\nu}[f(\hat{s})] \leq \mathbb{E}_{\hat{s}\sim\nu'}[f(\hat{s})] + \|\nu - \nu'\|_{\mathrm{TV}}.
$$

Applying this with

$$
f(\hat{s}) = \|\psi_{\Phi_w}(P_\theta^{a_k} b_k^{\Theta}, \hat{s}) - \psi_{\Phi_{w^\star}}(P_\theta^{a_k} b_k^{\Theta}, \hat{s})\|_{\mathrm{TV}} \in [0,1],
$$

$\nu = Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k)$, and $\nu' = Q_{w^\star}(P^{a_k}_{\theta} b^{\Theta}_k)$, yields

$$\begin{aligned}\mathbb{E}\big[(\mathrm{I}) \mid \mathcal{F}_k\big] &= \mathop{\mathbb{E}}_{\hat{s}\sim Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k)}\Big[\|\psi_{\Phi_w}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s}) - \psi_{\Phi_{w^\star}}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s})\|_{\mathrm{TV}}\Big] \\ &\le \mathop{\mathbb{E}}_{\hat{s}\sim Q_{w^\star}(P^{a_k}_{\theta} b^{\Theta}_k)}\Big[\|\psi_{\Phi_w}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s}) - \psi_{\Phi_{w^\star}}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s})\|_{\mathrm{TV}}\Big] \\ &\qquad + \Big\|Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k) - Q_{w^\star}(P^{a_k}_{\theta} b^{\Theta}_k)\Big\|_{\mathrm{TV}}. \end{aligned} \tag{50}$$

The first term is controlled by Eq. 48–Eq. 49 with $b = P^{a_k}_{\theta} b^{\Theta}_k$.

For the mismatch term in Eq. 50, since $Q_{w^\star}$ is a kernel, its Dobrushin contraction gives

$$\begin{aligned}\Big\|Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k) - Q_{w^\star}(P^{a_k}_{\theta} b^{\Theta}_k)\Big\|_{\mathrm{TV}} &\le (1-\kappa_\Phi)\,\Big\|P^{a_k}_{\theta^\star} b^{\Theta^\star}_k - P^{a_k}_{\theta} b^{\Theta}_k\Big\|_{\mathrm{TV}} \\ &\le (1-\kappa_\Phi)\Big(\Big\|P^{a_k}_{\theta^\star} b^{\Theta^\star}_k - P^{a_k}_{\theta^\star} b^{\Theta}_k\Big\|_{\mathrm{TV}} + \Big\|P^{a_k}_{\theta^\star} b^{\Theta}_k - P^{a_k}_{\theta} b^{\Theta}_k\Big\|_{\mathrm{TV}}\Big) \\ &\le (1-\kappa_\Phi)\Big((1-\kappa_P)\,\|b^{\Theta^\star}_k - b^{\Theta}_k\|_{\mathrm{TV}} + \frac{B}{2}\,\delta(\theta)\Big). \end{aligned} \tag{51}$$

Combining Eq. 50, Eq. 49, and Eq. 51, we obtain

$$\mathbb{E}\big[(\mathrm{I}) \mid \mathcal{F}_k\big] \le B\,\delta(w) + \frac{B}{2}(1-\kappa_\Phi)\,\delta(\theta) + (1-\kappa_P)(1-\kappa_\Phi)\,\|b^{\Theta^\star}_k - b^{\Theta}_k\|_{\mathrm{TV}}. \tag{52}$$

Term (II). To control the perturbation of the transition kernel, we keep the sampling law of $\hat{s}_{k+1}$ under the true model and add a change-of-measure step, as in Term (I):

$$\begin{aligned}\mathbb{E}\big[(\mathrm{II}) \mid \mathcal{F}_k\big] &= \mathbb{E}_{\hat{s}\sim Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k)}\Big[\big\|\psi_{\Phi_{w^\star}}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s}) - \psi_{\Phi_{w^\star}}(P^{a_k}_{\theta^\star} b^{\Theta}_k, \hat{s})\big\|_{\mathrm{TV}}\Big] \\ &\le \mathbb{E}_{\hat{s}\sim Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta}_k)}\Big[\big\|\psi_{\Phi_{w^\star}}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s}) - \psi_{\Phi_{w^\star}}(P^{a_k}_{\theta^\star} b^{\Theta}_k, \hat{s})\big\|_{\mathrm{TV}}\Big] \\ &\qquad + \Big\|Q_{w^\star}\big(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k\big) - Q_{w^\star}\big(P^{a_k}_{\theta^\star} b^{\Theta}_k\big)\Big\|_{\mathrm{TV}}. \end{aligned} \tag{53}$$

We now bound the first term in Eq. 53 using Lemma 3.2 Mcdonald & Yüksel (2024), which gives

$$\begin{aligned}\mathbb{E}_{\hat{s}\sim Q_{w^\star}(P^{a_k}_{\theta^\star} b^{\Theta}_k)}\Big[\big\|\psi_{\Phi_{w^\star}}(P^{a_k}_{\theta} b^{\Theta}_k, \hat{s}) - \psi_{\Phi_{w^\star}}(P^{a_k}_{\theta^\star} b^{\Theta}_k, \hat{s})\big\|_{\mathrm{TV}}\Big] &\le (2-\kappa_\Phi)\,\|P^{a_k}_{\theta} b^{\Theta}_k - P^{a_k}_{\theta^\star} b^{\Theta}_k\|_{\mathrm{TV}} \\ &= (2-\kappa_\Phi)\,\big\|P^{a_k}_{\theta} b^{\Theta}_k - P^{a_k}_{\theta^\star} b^{\Theta}_k\big\|_{\mathrm{TV}} \\ &\le \frac{B}{2}(2-\kappa_\Phi)\delta(\theta), \end{aligned} \tag{54}$$

where we have used

$$\begin{aligned}\big\|P^{a_k}_{\theta} b - P^{a_k}_{\theta^\star} b\big\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{s'\in\mathbb{S}}\Big|\sum_{s\in\mathbb{S}} b(s)\,\Big(P_\theta(s' \mid s, a_k) - P_{\theta^\star}(s' \mid s, a_k)\Big)\Big| \\ &\le \frac{1}{2}\sum_{s'\in\mathbb{S}}\sum_{s\in\mathbb{S}} b(s)\,\Big|P_\theta(s' \mid s, a_k) - P_{\theta^\star}(s' \mid s, a_k)\Big| \\ &= \frac{1}{2}\sum_{s\in\mathbb{S}} b(s)\,\sum_{s'\in\mathbb{S}}\Big|P_\theta(s' \mid s, a_k) - P_{\theta^\star}(s' \mid s, a_k)\Big| \\ &\le \sup_{s\in\mathbb{S}}\frac{1}{2}\sum_{s'\in\mathbb{S}}\Big|P_\theta(s' \mid s, a_k) - P_{\theta^\star}(s' \mid s, a_k)\Big| \\ &= \sup_{s\in\mathbb{S}}\frac{1}{2}\,\big\|P_\theta(\cdot \mid s, a_k) - P_{\theta^\star}(\cdot \mid s, a_k)\big\|_1 \\ &\le \frac{B}{2}\delta(\theta). \end{aligned}$$

It remains to bound the second term in Eq. 53. Using the Dobrushin coefficient of $Q_{w^\star}$ and then of $P^{a_k}_{\theta^\star}$ gives

$$\begin{aligned}\left\|Q_{w^\star}\big(P^{a_k}_{\theta^\star} b^{\Theta^\star}_k\big) - Q_{w^\star}\big(P^{a_k}_{\theta^\star} b^{\Theta}_k\big)\right\|_{\mathrm{TV}} &\le (1-\kappa_\Phi)\left\|P^{a_k}_{\theta^\star} b^{\Theta^\star}_k - P^{a_k}_{\theta^\star} b^{\Theta}_k\right\|_{\mathrm{TV}} \\ &\le (1-\kappa_\Phi)(1-\kappa_P)\left\|b^{\Theta^\star}_k - b^{\Theta}_k\right\|_{\mathrm{TV}}. \end{aligned} \tag{55}$$

Combining Eq. 54–Eq. 55 yields

$$\mathbb{E}\big[(\mathrm{II}) \mid \mathcal{F}_k\big] \le \frac{B}{2}(2-\kappa_\Phi)\,\delta(\theta) + (1-\kappa_\Phi)(1-\kappa_P)\left\|b^{\Theta^\star}_k - b^{\Theta}_k\right\|_{\mathrm{TV}}. \tag{56}$$

Remark B.2 (Dominance condition). We note that the dominance condition $P^{a_k}_{\theta^\star} b^{\Theta}_k \ll P^{a_k}_{\theta} b^{\Theta}_k$ required for the application of filter stability results (e.g., Lemma 3.2 in Mcdonald & Yüksel (2024)) is automatically satisfied for all k in our setting. Under the log-linear parameterization in Assumption 3.1, the transition kernel P_θ and observation model Φ_w, due to the softmax structure, are strictly positive everywhere, ensuring that $P^{a_k}_{\theta} b^{\Theta}_k$ and $P^{a_k}_{\theta^\star} b^{\Theta}_k$ remain mutually absolutely continuous.

Term (III). This term measures the propagation of error in belief for one time step. Under Assumption 3.5, and by Theorems 3.3 and 4.1 of Mcdonald & Yüksel (2024), for one step we have

$$\mathbb{E}[\,(\mathrm{III})\,|\,\mathcal{F}_k] \le (1-\kappa_P)(2-\kappa_\Phi)\|b^{\Theta}_k - b^{\Theta^\star}_k\|_{\mathrm{TV}}. \tag{57}$$

Since $b^{\Theta}_0 = b^{\Theta^\star}_0 = \nu_0$, we have

$$\mathbb{E}\Big[\|b^{\Theta}_0 - b^{\Theta^\star}_0\|_{\mathrm{TV}}\Big] = \|\nu_0 - \nu_0\|_{\mathrm{TV}} = 0.$$

Putting all together, and unrolling over time yields

$$\begin{aligned}\mathbb{E}\left[\|b^{\Theta}_{k+1} - b^{\Theta^\star}_{k+1}\|_{\mathrm{TV}}\right] &= \mathbb{E}\Big[\mathbb{E}\left[\|b^{\Theta}_{k+1} - b^{\Theta^\star}_{k+1}\|_{\mathrm{TV}} \,\middle|\, \mathcal{F}_k\right]\Big] \\ &\le \mathbb{E}\Big[\mathbb{E}\big[(\mathrm{I}) \mid \mathcal{F}_k\big] + \mathbb{E}\big[(\mathrm{II}) \mid \mathcal{F}_k\big] + \mathbb{E}\big[(\mathrm{III}) \mid \mathcal{F}_k\big]\Big] \\ &\le \Big((1-\kappa_P)(2-\kappa_\Phi) + 2(1-\kappa_P)(1-\kappa_\Phi)\Big)\,\mathbb{E}\Big[\|b^{\Theta}_k - b^{\Theta^\star}_k\|_{\mathrm{TV}}\Big] \\ &\qquad + B\,\delta(w) + \frac{B}{2}\big((2-\kappa_\Phi) + (1-\kappa_\Phi)\big)\delta(\theta) \\ &= \alpha\,\mathbb{E}\Big[\|b^{\Theta}_k - b^{\Theta^\star}_k\|_{\mathrm{TV}}\Big] + B\,\delta(w) + \frac{B}{2}(3-2\kappa_\Phi)\,\delta(\theta), \end{aligned} \tag{58}$$

with $\alpha = (1-\kappa_P)(4-3\kappa_\Phi)$. Unrolling Eq. 58 and using $\mathbb{E}[\|b^{\Theta}_0 - b^{\Theta^\star}_0\|_{\mathrm{TV}}] = 0$ gives, for all $k \ge 1$,

$$\begin{aligned}\mathbb{E}\left[\|b^{\Theta}_k - b^{\Theta^\star}_k\|_{\mathrm{TV}}\right] &\le \left(B\,\delta(w) + \frac{B}{2}(3-2\kappa_\Phi)\,\delta(\theta)\right)\sum_{j=0}^{k-1}\alpha^j \\ &= B\left(\delta(w) + \frac{3-2\kappa_\Phi}{2}\,\delta(\theta)\right)\frac{1-\alpha^k}{1-\alpha}. \end{aligned} \tag{59}$$

This concludes the proof. □

C Neural Network Extension

The linear approximation of the POMDP model studied earlier provides a setting in which the dependence of the POMDP kernels on the learned parameters can be controlled explicitly. However, log-linear scores may be too restrictive for complex models. We therefore consider a neural-softmax extension, where the scores defining $P(\cdot \mid s, a)$ and $\Phi(\cdot \mid s)$ are represented by neural networks.

This extension is motivated by standard results on over-parameterized neural networks. Such networks can interpolate training data under first-order optimization methods while still exhibiting strong generalization behavior (Belkin et al., 2018; Bartlett et al., 2019; Zhang et al., 2016). A central theoretical explanation is the lazy-training or neural tangent kernel (NTK) regime, where sufficiently wide networks trained near random initialization behave approximately like their first-order linearization around initialization (Jacot et al., 2018; Li & Liang, 2018; Du et al., 2019; Oymak & Soltanolkotabi, 2019; Chizat et al., 2019). In this regime,

$$F(x;W) \approx F(x;W_0) + \big\langle \nabla_W F(x;W_0),\, W - W_0 \big\rangle,$$

so the tangent feature map $\nabla_W F(x;W_0)$ plays the role of an effective feature representation. In the infinite-width limit, these tangent features induce an RKHS description, and finite-width networks approximate the corresponding kernel functions with errors of order $O(m^{-1/2})$ under standard assumptions.

In this section, we extend our belief-stability mechanism beyond the log-linear class. The neural-softmax model can be viewed as an approximate log-linear model in its NTK tangent features, with additional finite-width linearization errors $\varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}})$ and $\varepsilon_\Phi^{\mathrm{NN}}(m, \delta_{\mathrm{NN}})$. These errors additively contribute to the belief perturbation bound and vanish as the lazy-training approximation improves.

Definition C.1 (Symmetric random initialization)**.** Assume the fixed width m is even. For each model component $g \in \{p, \Phi\}$, let $d_{x,g}$ denote the input dimension. For $i = 1, \ldots, m/2$, sample independently $c_{g,i} \sim \mathrm{Rad}$, $\omega_{g,i,0} \sim \mathcal{N}(0, I_{d_{x,g}})$, and define the second half of the initialization by

$$c_{g,i+m/2} := -c_{g,i}, \qquad \omega_{g,i+m/2,0} := \omega_{g,i,0}, \qquad i = 1, \ldots, m/2.$$

Then, $W_{g,0} := (\omega_{g,1,0}, \ldots, \omega_{g,m,0})$ is called a symmetric random initialization for the two-layer ReLU network.

Definition C.2 (NTK-linearized scores and tangent feature maps)**.** Let $W_{p,0}$ and $W_{\Phi,0}$ be reference initialization points for the transition and observation networks with parameters W_p and W_Φ, respectively. Let $x_p(s,a,s') \in \mathbb{R}^{d_{x,p}}, x_\Phi(s,\hat{s}) \in \mathbb{R}^{d_{x,\Phi}}$ denote fixed input encodings of the transition and observation tuples, respectively. We write the trainable first-layer weights as

$$W_p = (\omega_{p,1}, \ldots, \omega_{p,m}) \in \mathbb{R}^{m \times d_{x,p}}, \qquad W_\Phi = (\omega_{\Phi,1}, \ldots, \omega_{\Phi,m}) \in \mathbb{R}^{m \times d_{x,\Phi}},$$

and fix coefficients $c_{p,i}, c_{\Phi,i} \in \{-1, +1\}$. The neural scores are defined as

$$F_p(s,a,s';W_p) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_{p,i}\ \mathrm{ReLU}\big(\omega_{p,i}^\top x_p(s,a,s')\big)\,, \tag{60}$$

$$F_\Phi(s,\hat{s};W_\Phi) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_{\Phi,i}\ \mathrm{ReLU}\big(\omega_{\Phi,i}^\top x_\Phi(s,\hat{s})\big)\,, \tag{61}$$

Throughout this section, norms on W_p and W_Φ are Frobenius norms. Equivalently, after vectorization, one may regard $W_p \in \mathbb{R}^{d_p}$ and $W_\Phi \in \mathbb{R}^{d_\Phi}$, where $d_p = m d_{x,p}$ and $d_\Phi = m d_{x,\Phi}$.

Assumption C.3 (Over-parameterized neural-softmax model)**.** The neural-softmax model in Definition 4.4 satisfies the following conditions.

- **Realizability.** There exists $W^\star = (W_p^\star, W_\Phi^\star) \in \mathcal{D}$ generating the true POMDP.
- **Initialization.** The reference points $W_{p,0}$ and $W_{\Phi,0}$ are symmetric initializations as in Definition C.1, for the transition and observation networks, respectively. Under symmetric construction we have

$$F_p(s,a,s';W_{p,0}) = 0, \qquad F_\Phi(s,\hat{s};W_{\Phi,0}) = 0, \tag{62}$$

  for all $(s,a,s') \in \mathbb{S} \times \mathbb{A} \times \mathbb{S}$ and $(s,\hat{s}) \in \mathbb{S} \times \hat{\mathbb{S}}$.

- **Uniformly bounded tangent features.** For some $B^{\mathrm{NN}} < \infty$, the tangent feature maps in Definition C.2 satisfy

$$\sup_{s\in\mathbb{S},\ a\in\mathbb{A},\ s'\in\mathbb{S}} \|\phi_p^{\mathrm{NN}}(s,a,s')\|_F \le B^{\mathrm{NN}}, \tag{63}$$

$$\sup_{s\in\mathbb{S},\ \hat{s}\in\hat{\mathbb{S}}} \|\phi_\Phi^{\mathrm{NN}}(s,\hat{s})\|_F \le B^{\mathrm{NN}}. \tag{64}$$

Linearization. Define the first-order NTK-linearized scores around initialization by

$$F_{p,W_p}^{\mathrm{lin}}(s,a,s') := F_p(s,a,s';W_{p,0}) + \left\langle \nabla_{W_p} F_p(s,a,s';W_{p,0}), W_p - W_{p,0}\right\rangle_F, \tag{65}$$

$$F_{\Phi,W_\Phi}^{\mathrm{lin}}(s,\hat{s}) := F_\Phi(s,\hat{s};W_{\Phi,0}) + \left\langle \nabla_{W_\Phi} F_\Phi(s,\hat{s};W_{\Phi,0}), W_\Phi - W_{\Phi,0}\right\rangle_F. \tag{66}$$

The corresponding finite-width tangent feature maps are defined as

$$\phi_p^{\mathrm{NN}}(s,a,s') := \nabla_{W_p} F_p(s,a,s';W_{p,0}), \tag{67}$$

$$\phi_\Phi^{\mathrm{NN}}(s,\hat{s}) := \nabla_{W_\Phi} F_\Phi(s,\hat{s};W_{\Phi,0}). \tag{68}$$

Linearization event. Let the bounded convex set $\mathcal{D} = \mathcal{D}_p \times \mathcal{D}_\Phi$ denote a local lazy-training neighborhood around initialization, chosen so that the high-probability linearization bounds below hold. For $\delta_{\mathrm{NN}} \in (0,1)$, define the high-probability linearization event

$$\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}}) := \Bigg\{ \sup_{W_p\in\mathcal{D}_p} \sup_{s,a,s'} \left| F_p(s,a,s';W_p) - F_{p,W_p}^{\mathrm{lin}}(s,a,s') \right| \le \varepsilon_p^{\mathrm{NN}}(m,\delta_{\mathrm{NN}}),$$
$$\sup_{W_\Phi\in\mathcal{D}_\Phi} \sup_{s,\hat{s}} \left| F_\Phi(s,\hat{s};W_\Phi) - F_{\Phi,W_\Phi}^{\mathrm{lin}}(s,\hat{s}) \right| \le \varepsilon_\Phi^{\mathrm{NN}}(m,\delta_{\mathrm{NN}}) \Bigg\}, \tag{69}$$

where $\mathbb{P}(\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})) \ge 1 - \delta_{\mathrm{NN}}$.

Remark C.4 (Neural linearization event). Under Assumption C.3, the event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$ is used to replace the neural-softmax scores in Definition 4.4 by their NTK-linearized counterparts in Definition C.2. This event follows from standard two-layer ReLU arguments around symmetric initialization (see Lemma 2 of Cayci et al. (2024)) that gives the corresponding local linearization control, while Lemma 4.1(7) of Satpathi et al. (2020) provides the required uniform control over bounded inputs. Applying these bounds separately to the transition and observation score networks, and taking a union bound over the two initializations, yields $\mathbb{P}(\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})) \ge 1 - \delta_{\mathrm{NN}}$ with, for universal constants $C_p, C_\Phi > 0$,

$$\varepsilon_p^{\mathrm{NN}}(m_p,\delta_{\mathrm{NN}}) \le C_p \frac{R_p}{\sqrt{m_p}} \left(R_p + \sqrt{\log\frac{2}{\delta_{\mathrm{NN}}}} + \sqrt{d_p \log m_p} \right)$$

$$\varepsilon_\Phi^{\mathrm{NN}}(m_\Phi,\delta_{\mathrm{NN}}) \le C_\Phi \frac{R_\Phi}{\sqrt{m_\Phi}} \left(R_\Phi + \sqrt{\log\frac{2}{\delta_{\mathrm{NN}}}} + \sqrt{d_\Phi \log m_\Phi} \right),$$

with

$$R_p := \sqrt{m_p} \sup_{W_p\in\mathcal{D}_p} \max_{i\in[m_p]} \|W_{p,i} - W_{p,i,0}\|_2, \qquad R_\Phi := \sqrt{m_\Phi} \sup_{W_\Phi\in\mathcal{D}_\Phi} \max_{i\in[m_\Phi]} \|W_{\Phi,i} - W_{\Phi,i,0}\|_2.$$

Throughout the neural-softmax analysis, all bounds involving Lemma C.5 and Corollary 4.5 are understood on this event.

Lemma C.5 (Neural-softmax kernel perturbation bounds)**.** *Fix $\delta_{\mathrm{NN}} \in (0,1)$ and work on the event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$. Assume Assumption C.3. For $W_p, W_p' \in \mathcal{D}_p$ and $W_\Phi, W_\Phi' \in \mathcal{D}_\Phi$, let, for $W \in \mathcal{D}_p$ and $W' \in \mathcal{D}_\Phi$,*

$$\delta^{\mathrm{NN}}(W) := \|W - W_{p,0}\|_F, \qquad \delta^{\mathrm{NN}}(W') := \|W' - W_{\Phi,0}\|_F. \tag{70}$$

Then,

$$\left\|P_{W_p}(\cdot \mid s,a) - P_{W'_p}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \le \frac{1}{2}B^{\mathrm{NN}}\|W_p - W'_p\|_F + \varepsilon_p^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}), \qquad \forall (s,a) \in \mathbb{S}\times\mathbb{A}. \tag{71}$$

$$\left\|\Phi_{W_\Phi}(\cdot \mid s) - \Phi_{W'_\Phi}(\cdot \mid s)\right\|_{\mathrm{TV}} \le \frac{1}{2}B^{\mathrm{NN}}\|W_\Phi - W'_\Phi\|_F + \varepsilon_\Phi^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}), \qquad \forall s \in \mathbb{S}. \tag{72}$$

$$\left\|P_{W_p}(s' \mid \cdot,a) - P_{W'_p}(s' \mid \cdot,a)\right\|_1 \le 2B^{\mathrm{NN}}c_p\|W_p - W'_p\|_F + 4c_p\varepsilon_p^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}), \qquad \forall (s',a) \in \mathbb{S}\times\mathbb{A}. \tag{73}$$

$$\left\|\Phi_{W_\Phi}(\hat{s} \mid \cdot) - \Phi_{W'_\Phi}(\hat{s} \mid \cdot)\right\|_1 \le 2B^{\mathrm{NN}}c_\Phi\|W_\Phi - W'_\Phi\|_F + 4c_\Phi\varepsilon_\Phi^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}), \qquad \forall \hat{s} \in \hat{\mathbb{S}}, \tag{74}$$

where

$$c_p := \frac{|\mathbb{S}|}{1 + (|\mathbb{S}|-1)\exp\bigl(-2(B^{\mathrm{NN}}r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}))\bigr)}, \qquad c_\Phi := \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}|-1)\exp\bigl(-2(B^{\mathrm{NN}}r_\Phi^{\mathrm{NN}} + \varepsilon_\Phi^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}))\bigr)},$$

and

$$r_p^{\mathrm{NN}} := \sup_{\bar{W}_p \in \mathcal{D}_p} \|\bar{W}_p - W_{p,0}\|_F, \qquad r_\Phi^{\mathrm{NN}} := \sup_{\bar{W}_\Phi \in \mathcal{D}_\Phi} \|\bar{W}_\Phi - W_{\Phi,0}\|_F. \tag{75}$$

Proof. Define the linearized score mismatches

$$\Delta_p^{\mathrm{lin}}(W_p, W'_p) := \sup_{s\in\mathbb{S},\ a\in\mathbb{A},\ s'\in\mathbb{S}} \left|F_{p,W_p}^{\mathrm{lin}}(s,a,s') - F_{p,W'_p}^{\mathrm{lin}}(s,a,s')\right|, \tag{76}$$

$$\Delta_\Phi^{\mathrm{lin}}(W_\Phi, W'_\Phi) := \sup_{s\in\mathbb{S},\ \hat{s}\in\hat{\mathbb{S}}} \left|F_{\Phi,W_\Phi}^{\mathrm{lin}}(s,\hat{s}) - F_{\Phi,W'_\Phi}^{\mathrm{lin}}(s,\hat{s})\right|. \tag{77}$$

By the symmetric Xavier initialization in Assumption C.3, Eq. 65 becomes

$$F_{p,W_p}^{\mathrm{lin}}(s,a,s') = \left\langle \phi_p^{\mathrm{NN}}(s,a,s'),\, W_p - W_{p,0}\right\rangle_F.$$

Thus, given initialization, the transition kernel $P_{W_p}^{\mathrm{lin}}(\cdot \mid s,a)$, obtained by using F_{p,W_p}^{lin} in Eq. 5, is a log-linear model with feature map ϕ_p^{NN} and parameter $W_p - W_{p,0}$. Since $P_{W_p}^{\mathrm{lin}}(\cdot \mid s,a)$ is a log-linear softmax model with feature map ϕ_p^{NN}, parameter $W_p - W_{p,0}$, and feature bound B^{NN}, the row-wise part of Lemma A.1 applies with θ replaced by $W_p - W_{p,0}$. Therefore,

$$\left\|P_{W_p}^{\mathrm{lin}}(\cdot \mid s,a) - P_{W'_p}^{\mathrm{lin}}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \le \frac{B^{\mathrm{NN}}}{2}\|W_p - W'_p\|_F. \tag{78}$$

Moreover, by the same row-wise argument used in Lemma A.1, applied directly to the score vectors, we also have

$$\left\|P_{W_p}^{\mathrm{lin}}(\cdot \mid s,a) - P_{W'_p}^{\mathrm{lin}}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \le \frac{1}{2}\Delta_p^{\mathrm{lin}}(W_p, W'_p). \tag{79}$$

Moreover, the definition of F_{p,W_p}^{lin} and the uniform tangent-feature bound imply

$$\Delta_p^{\mathrm{lin}}(W_p, W'_p) \le B^{\mathrm{NN}}\|W_p - W'_p\|_F, \tag{80}$$

then,

$$\left|F_{p,W_p}^{\mathrm{lin}}(s,a,s') - F_{p,W'_p}^{\mathrm{lin}}(s,a,s')\right| = \left|\left\langle \phi_p^{\mathrm{NN}}(s,a,s'),\, W_p - W'_p\right\rangle_F\right| \le B^{\mathrm{NN}}\|W_p - W'_p\|_F.$$

Next, by the definition of the score-linearization error,

$$\sup_{s,a,s'} \left|F_p(s,a,s';W_p) - F_{p,W_p}^{\mathrm{lin}}(s,a,s')\right| \le \varepsilon_p^{\mathrm{NN}}(m,\,\delta_{\mathrm{NN}}),$$

and the same bound holds with W_p' in place of W_p. Applying the same row-wise score argument used in Lemma A.1 to the nonlinear and linearized score vectors gives

$$\left\|P_{W_p}(\cdot \mid s,a) - P^{\mathrm{lin}}_{W_p}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \le \frac{1}{2}\varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}), \tag{81}$$

$$\left\|P_{W_p'}(\cdot \mid s,a) - P^{\mathrm{lin}}_{W_p'}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \le \frac{1}{2}\varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}). \tag{82}$$

Combining Eq. 79, Eq. 81, and Eq. 82 by the triangle inequality yields

$$\begin{aligned}\left\|P_{W_p}(\cdot \mid s,a) - P_{W_p'}(\cdot \mid s,a)\right\|_{\mathrm{TV}} &\le \left\|P_{W_p}(\cdot \mid s,a) - P^{\mathrm{lin}}_{W_p}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \\ &\quad + \left\|P^{\mathrm{lin}}_{W_p}(\cdot \mid s,a) - P^{\mathrm{lin}}_{W_p'}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \\ &\quad + \left\|P^{\mathrm{lin}}_{W_p'}(\cdot \mid s,a) - P_{W_p'}(\cdot \mid s,a)\right\|_{\mathrm{TV}} \\ &\le \frac{1}{2}\Delta_p^{\mathrm{lin}}(W_p, W_p') + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}).\end{aligned}$$

Using Eq. 80 proves the claim.

It remains to prove the column-wise transition bound. For any $W_p \in \mathcal{D}_p$, the initialization condition and tangent-feature bound imply

$$\left|F^{\mathrm{lin}}_{p,W_p}(s,a,s')\right| = \left|\left\langle \phi_p^{\mathrm{NN}}(s,a,s'),\, W_p - W_{p,0}\right\rangle_F\right| \le B^{\mathrm{NN}}\delta^{\mathrm{NN}}(W_p) \le B^{\mathrm{NN}} r_p^{\mathrm{NN}}.$$

Hence, for every $s, s' \in \mathbb{S}$,

$$P^{\mathrm{lin}}_{W_p}(s' \mid s,a) \le \frac{1}{1 + (|\mathbb{S}|-1)\exp\!\big(-2B^{\mathrm{NN}} r_p^{\mathrm{NN}}\big)} \le \frac{1}{1 + (|\mathbb{S}|-1)\exp\!\big(-2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\big)}.$$

Therefore,

$$\sum_{s\in\mathbb{S}} P^{\mathrm{lin}}_{W_p}(s' \mid s,a) \le c_p.$$

Repeating the column-wise Jacobian argument in Lemma A.1, now applied to the linearized model and using score mismatch $\Delta_p^{\mathrm{lin}}(W_p, W_p')$, gives

$$\left\|P^{\mathrm{lin}}_{W_p}(s' \mid \cdot,a) - P^{\mathrm{lin}}_{W_p'}(s' \mid \cdot,a)\right\|_1 \le 2c_p\, \Delta_p^{\mathrm{lin}}(W_p, W_p').$$

Moreover, the score-linearization error implies

$$\sup_{s,a,s'} |F_p(s,a,s'; W_p)| \le B^{\mathrm{NN}}\delta^{\mathrm{NN}}(W_p) + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}) \le B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}).$$

Thus, for the same coordinate-wise softmax-Jacobian estimate,

$$\left\|P_{W_p}(s' \mid \cdot,a) - P^{\mathrm{lin}}_{W_p}(s' \mid \cdot,a)\right\|_1 \le 2c_p\, \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}),$$

and the same bound holds with W_p' in place of W_p. By the triangle inequality,

$$\begin{aligned}\left\|P_{W_p}(s' \mid \cdot,a) - P_{W_p'}(s' \mid \cdot,a)\right\|_1 &\le \left\|P_{W_p}(s' \mid \cdot,a) - P^{\mathrm{lin}}_{W_p}(s' \mid \cdot,a)\right\|_1 \\ &\quad + \left\|P^{\mathrm{lin}}_{W_p}(s' \mid \cdot,a) - P^{\mathrm{lin}}_{W_p'}(s' \mid \cdot,a)\right\|_1 \\ &\quad + \left\|P^{\mathrm{lin}}_{W_p'}(s' \mid \cdot,a) - P_{W_p'}(s' \mid \cdot,a)\right\|_1 \\ &\le 2c_p\, \Delta_p^{\mathrm{lin}}(W_p, W_p') + 4c_p\, \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}).\end{aligned}$$

Using Eq. 80 gives the parameter-distance version.

The proof for the observation kernel is the same, replacing the transition score family F_p by F_Φ, the parameter W_p by W_Φ, the tangent feature map ϕ_p^{NN} by ϕ_Φ^{NN}, and the softmax output space $\mathbb{S}$ by $\hat{\mathbb{S}}$. The resulting fixed-$\hat{s}$ column is summed over $s \in \mathbb{S}$, which gives the constant c_Φ. This proves Eq. 72–Eq. 74. □

Lemma C.6 (Neural-softmax column sums and Dobrushin bounds)**.** *Fix $\delta_{\mathrm{NN}} \in (0,1)$ and work on the event $\mathcal{E}_{\mathrm{lin}}^{\mathrm{NN}}(\delta_{\mathrm{NN}})$. Assume Assumption C.3. Then, the following properties hold.*

$$\sup_{W_p \in \mathcal{D}_p} \sup_{a \in \mathbb{A}} \sup_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} P_{W_p}(s' \mid s, a) \leq \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\bigl(-2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}. \tag{83}$$

$$\inf_{W_p \in \mathcal{D}_p} \inf_{a \in \mathbb{A}} \inf_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} P_{W_p}(s' \mid s, a) \geq \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}. \tag{84}$$

$$\sup_{W_\Phi \in \mathcal{D}_\Phi} \sup_{\hat{s} \in \hat{\mathbb{S}}} \sum_{s \in \mathbb{S}} \Phi_{W_\Phi}(\hat{s} \mid s) \leq \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(-2(B^{\mathrm{NN}} r_\Phi^{\mathrm{NN}} + \varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}. \tag{85}$$

$$\inf_{W_\Phi \in \mathcal{D}_\Phi} \inf_{\hat{s} \in \hat{\mathbb{S}}} \sum_{s \in \mathbb{S}} \Phi_{W_\Phi}(\hat{s} \mid s) \geq \frac{|\mathbb{S}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r_\Phi^{\mathrm{NN}} + \varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}. \tag{86}$$

Furthermore, the uniform neural Dobrushin coefficients defined as

$$\kappa_P^{\mathrm{NN}} := \inf_{W_p \in \mathcal{D}_p} \inf_{a \in \mathbb{A}} \inf_{s, s'' \in \mathbb{S}} \sum_{x \in \mathbb{S}} \min\{P_{W_p}(x \mid s, a), P_{W_p}(x \mid s'', a)\},$$

$$\kappa_\Phi^{\mathrm{NN}} := \inf_{W_\Phi \in \mathcal{D}_\Phi} \inf_{s, s'' \in \mathbb{S}} \sum_{\hat{s} \in \hat{\mathbb{S}}} \min\{\Phi_{W_\Phi}(\hat{s} \mid s), \Phi_{W_\Phi}(\hat{s} \mid s'')\},$$

satisfy

$$\kappa_P^{\mathrm{NN}} \geq \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}, \tag{87}$$

$$\kappa_\Phi^{\mathrm{NN}} \geq \frac{|\hat{\mathbb{S}}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r_\Phi^{\mathrm{NN}} + \varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}. \tag{88}$$

Proof of Lemma C.6. We first prove the claims for the transition kernel. Fix $(s, a, s') \in \mathbb{S} \times \mathbb{A} \times \mathbb{S}$ and $W_p \in \mathcal{D}_p$. By the symmetric initialization in Assumption C.3,

$$F_{p,W_p}^{\mathrm{lin}}(s, a, s') = \bigl\langle \phi_p^{\mathrm{NN}}(s, a, s'), W_p - W_{p,0} \bigr\rangle_F.$$

Hence, by the tangent-feature bound and the definition of r_p^{NN},

$$\bigl|F_{p,W_p}^{\mathrm{lin}}(s, a, s')\bigr| \leq \|\phi_p^{\mathrm{NN}}(s, a, s')\|_F \, \|W_p - W_{p,0}\|_F \leq B^{\mathrm{NN}} r_p^{\mathrm{NN}}.$$

Using the score-linearization error in Definition C.2, we obtain the full nonlinear score bound

$$|F_p(s, a, s'; W_p)| \leq \bigl|F_{p,W_p}^{\mathrm{lin}}(s, a, s')\bigr| + \bigl|F_p(s, a, s'; W_p) - F_{p,W_p}^{\mathrm{lin}}(s, a, s')\bigr| \leq B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}).$$

Thus the neural-softmax transition kernel satisfies the same uniform score bound as the log-linear kernel in Lemma A.2, with $B(\|\theta^\star\|_2 + r_\Theta)$ replaced by $B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}})$. Repeating the pointwise softmax upper-bound argument in Eq. 33 gives, for all (s, a, s'),

$$P_{W_p}(s' \mid s, a) \leq \frac{1}{1 + (|\mathbb{S}| - 1)\exp\bigl(-2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)}.$$

Summing over $s \in \mathbb{S}$ and taking the supremum over (W_p, a, s') proves Eq. 83. Similarly, repeating the lower-bound argument in Eq. 34 gives

$$P_{W_p}(s' \mid s, a) \geq \frac{1}{1 + (|\mathbb{S}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r_p^{\mathrm{NN}} + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}))\bigr)} =: P_{\min}^{\mathrm{NN}}.$$

Summing $P^{\mathrm{NN}}_{\min}$ over $s \in \mathbb{S}$ and taking the infimum over (W_p, a, s') proves Eq. 84. Moreover, since every entry of $P_{W_p}(\cdot \mid s, a)$ is bounded from below by $P^{\mathrm{NN}}_{\min}$, the same Dobrushin argument as in Eq. 35 yields

$$\sum_{x\in\mathbb{S}} \min\{P_{W_p}(x \mid s, a), P_{W_p}(x \mid s'', a)\} \geq |\mathbb{S}| P^{\mathrm{NN}}_{\min} = \frac{|\mathbb{S}|}{1 + (|\mathbb{S}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r^{\mathrm{NN}}_p + \varepsilon^{\mathrm{NN}}_p(m,\, \delta_{\mathrm{NN}}))\bigr)}.$$

Taking the infimum over $W_p \in \mathcal{D}_p$, $a \in \mathbb{A}$, and $s, s'' \in \mathbb{S}$ proves Eq. 87.

We now prove the observation bounds. Fix $(s, \hat{s}) \in \mathbb{S} \times \hat{\mathbb{S}}$ and $W_\Phi \in \mathcal{D}_\Phi$. By the symmetric initialization,

$$F^{\mathrm{lin}}_{\Phi, W_\Phi}(s, \hat{s}) = \bigl\langle \phi^{\mathrm{NN}}_\Phi(s, \hat{s}), W_\Phi - W_{\Phi,0} \bigr\rangle_F.$$

The tangent-feature bound and the definition of r^{NN}_Φ imply

$$\bigl|F^{\mathrm{lin}}_{\Phi, W_\Phi}(s, \hat{s})\bigr| \leq \|\phi^{\mathrm{NN}}_\Phi(s, \hat{s})\|_F \, \|W_\Phi - W_{\Phi,0}\|_F \leq B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi.$$

Together with the score-linearization error, this gives

$$|F_\Phi(s, \hat{s}; W_\Phi)| \leq B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi + \varepsilon^{\mathrm{NN}}_\Phi(m,\, \delta_{\mathrm{NN}}).$$

Thus the neural-softmax observation kernel satisfies the same uniform score bound as the log-linear observation kernel in Lemma A.2, with $B(\|w^\star\|_2 + r_\Theta)$ replaced by $B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi + \varepsilon^{\mathrm{NN}}_\Phi(m,\, \delta_{\mathrm{NN}})$. Repeating the pointwise upper- and lower-bound arguments in Eq. 36–Eq. 37 gives, for all $(s, \hat{s})$,

$$\Phi_{W_\Phi}(\hat{s} \mid s) \leq \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(-2(B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi + \varepsilon^{\mathrm{NN}}_\Phi(m,\, \delta_{\mathrm{NN}}))\bigr)} =: \Phi^{\mathrm{NN}}_{\max},$$

and

$$\Phi_{W_\Phi}(\hat{s} \mid s) \geq \frac{1}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi + \varepsilon^{\mathrm{NN}}_\Phi(m,\, \delta_{\mathrm{NN}}))\bigr)} =: \Phi^{\mathrm{NN}}_{\min}.$$

Summing these pointwise bounds over $s \in \mathbb{S}$ and taking the supremum and infimum over $(W_\Phi, \hat{s})$ proves Eq. 85 and Eq. 86, respectively. Finally, since every entry of $\Phi_{W_\Phi}(\cdot \mid s)$ is bounded from below by $\Phi^{\mathrm{NN}}_{\min}$, the same Dobrushin argument as in Eq. 38 gives

$$\sum_{\hat{s}\in\hat{\mathbb{S}}} \min\{\Phi_{W_\Phi}(\hat{s} \mid s), \Phi_{W_\Phi}(\hat{s} \mid s'')\} \geq |\hat{\mathbb{S}}| \Phi^{\mathrm{NN}}_{\min} = \frac{|\hat{\mathbb{S}}|}{1 + (|\hat{\mathbb{S}}| - 1)\exp\bigl(2(B^{\mathrm{NN}} r^{\mathrm{NN}}_\Phi + \varepsilon^{\mathrm{NN}}_\Phi(m,\, \delta_{\mathrm{NN}}))\bigr)}.$$

Taking the infimum over $W_\Phi \in \mathcal{D}_\Phi$ and $s, s'' \in \mathbb{S}$ proves Eq. 88. □

D Proof of Corollary 4.5

Proof. For $a \in \mathbb{A}$, write

$$(P^a_{W_p} q)(s') := \sum_{s\in\mathbb{S}} P_{W_p}(s' \mid s, a) q(s), \qquad (Q_{W_\Phi} q)(\hat{s}) := \sum_{s'\in\mathbb{S}} \Phi_{W_\Phi}(\hat{s} \mid s') q(s').$$

The belief recursions are

$$b^W_{k+1} = \psi_{\Phi_{W_\Phi}}\bigl(P^{a_k}_{W_p} b^W_k, \hat{s}_{k+1}\bigr), \qquad b^{W^\star}_{k+1} = \psi_{\Phi_{W^\star_\Phi}}\bigl(P^{a_k}_{W^\star_p} b^{W^\star}_k, \hat{s}_{k+1}\bigr), \qquad b^W_0 = b^{W^\star}_0 = \nu_0.$$

We work under $\mathbb{P}_{W^\star}(\cdot \mid a_{0:t-1})$, with the same filtrations $\mathcal{F}^-_k$ and $\mathcal{F}_k$ as in the proof of Theorem 4.1. Thus

$$\hat{s}_{k+1} \mid \mathcal{F}_k \sim Q_{W^\star_\Phi}\Bigl(P^{a_k}_{W^\star_p} b^{W^\star}_k\Bigr).$$

Define

$$\ell_p^{\mathrm{NN}} := \frac{B^{\mathrm{NN}}}{2}\|W_p - W_p^\star\|_F + \varepsilon_p^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}), \tag{89}$$

$$\ell_\Phi^{\mathrm{NN}} := B^{\mathrm{NN}}\|W_\Phi - W_\Phi^\star\|_F + 2\varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}). \tag{90}$$

By Lemma C.5, for every $b \in \Delta_{\mathbb{S}}$ and $a \in \mathbb{A}$,

$$\left\|P^a_{W_p} b - P^a_{W_p^\star} b\right\|_{\mathrm{TV}} \le \ell_p^{\mathrm{NN}}, \tag{91}$$

$$\sup_{s\in\mathbb{S}}\left\|\Phi_{W_\Phi}(\cdot \mid s) - \Phi_{W_\Phi^\star}(\cdot \mid s)\right\|_1 \le \ell_\Phi^{\mathrm{NN}}. \tag{92}$$

Fix $k \in \{0, \dots, t-1\}$. As in the proof of Theorem 4.1, add and subtract

$$\psi_{\Phi_{W_\Phi^\star}}(P^{a_k}_{W_p} b_k^W, \hat{s}_{k+1}) \quad \text{and} \quad \psi_{\Phi_{W_\Phi^\star}}(P^{a_k}_{W_p^\star} b_k^W, \hat{s}_{k+1}),$$

and denote the resulting three terms by (I), (II), (III), corresponding respectively to observation-model perturbation, transition-model perturbation, and propagation of the previous belief error. The algebraic decomposition is identical to the decomposition in the proof of Theorem 4.1.

For Term (I), the Bayes-operator perturbation argument in Eq. 48–Eq. 49, with $\Phi_w, \Phi_{w^\star}$ replaced by $\Phi_{W_\Phi}, \Phi_{W_\Phi^\star}$, gives

$$\mathbb{E}_{\hat{s}\sim Q_{W_\Phi^\star} b}\Big[\|\psi_{\Phi_{W_\Phi}}(b, \hat{s}) - \psi_{\Phi_{W_\Phi^\star}}(b, \hat{s})\|_{\mathrm{TV}}\Big] \le \ell_\Phi^{\mathrm{NN}}.$$

The same change-of-measure step as in Eq. 50, together with the Dobrushin contraction of $Q_{W_\Phi^\star}$, the contraction of $P^{a_k}_{W_p^\star}$, and Eq. 91, yields

$$\mathbb{E}[(\mathrm{I}) \mid \mathcal{F}_k] \le \ell_\Phi^{\mathrm{NN}} + (1-\kappa_\Phi^{\mathrm{NN}})\ell_p^{\mathrm{NN}} + (1-\kappa_P^{\mathrm{NN}})(1-\kappa_\Phi^{\mathrm{NN}})\|b_k^{W^\star} - b_k^W\|_{\mathrm{TV}}. \tag{93}$$

For Term (II), applying Lemma 3.2 of Mcdonald & Yüksel (2024) as in Eq. 54, and using Eq. 91, gives

$$\mathbb{E}_{\hat{s}\sim Q_{W_\Phi^\star}(P^{a_k}_{W_p^\star} b_k^W)}\Big[\|\psi_{\Phi_{W_\Phi^\star}}(P^{a_k}_{W_p} b_k^W, \hat{s}) - \psi_{\Phi_{W_\Phi^\star}}(P^{a_k}_{W_p^\star} b_k^W, \hat{s})\|_{\mathrm{TV}}\Big] \le (2-\kappa_\Phi^{\mathrm{NN}})\ell_p^{\mathrm{NN}}.$$

The change-of-measure term is the same as in Eq. 55, with the neural Dobrushin coefficients:

$$\Big\|Q_{W_\Phi^\star}(P^{a_k}_{W_p^\star} b_k^{W^\star}) - Q_{W_\Phi^\star}(P^{a_k}_{W_p^\star} b_k^W)\Big\|_{\mathrm{TV}} \le (1-\kappa_\Phi^{\mathrm{NN}})(1-\kappa_P^{\mathrm{NN}})\|b_k^{W^\star} - b_k^W\|_{\mathrm{TV}}.$$

Therefore,

$$\mathbb{E}[(\mathrm{II}) \mid \mathcal{F}_k] \le (2-\kappa_\Phi^{\mathrm{NN}})\ell_p^{\mathrm{NN}} + (1-\kappa_\Phi^{\mathrm{NN}})(1-\kappa_P^{\mathrm{NN}})\|b_k^{W^\star} - b_k^W\|_{\mathrm{TV}}. \tag{94}$$

For Term (III), Theorems 3.3 and 4.1 of Mcdonald & Yüksel (2024), applied with the neural Dobrushin coefficients from Lemma C.6, give

$$\mathbb{E}[(\mathrm{III}) \mid \mathcal{F}_k] \le (1-\kappa_P^{\mathrm{NN}})(2-\kappa_\Phi^{\mathrm{NN}})\|b_k^W - b_k^{W^\star}\|_{\mathrm{TV}}. \tag{95}$$

The required dominance condition clearly holds, since the neural-softmax transition and observation kernels are strictly positive on finite spaces.

Combining Eq. 93–Eq. 95 and taking outer expectation yields

$$\mathbb{E}\Big[\|b_{k+1}^W - b_{k+1}^{W^\star}\|_{\mathrm{TV}}\Big] \le \alpha_{\mathrm{NN}}\,\mathbb{E}\Big[\|b_k^W - b_k^{W^\star}\|_{\mathrm{TV}}\Big] + \ell_\Phi^{\mathrm{NN}} + (3-2\kappa_\Phi^{\mathrm{NN}})\ell_p^{\mathrm{NN}}, \tag{96}$$

where $\alpha_{\mathrm{NN}} = (1-\kappa_P^{\mathrm{NN}})(4-3\kappa_\Phi^{\mathrm{NN}})$. Since $b_0^W = b_0^{W^\star}$, unrolling Eq. 96 gives

$$\mathbb{E}\Big[\|b_t^W - b_t^{W^\star}\|_{\mathrm{TV}}\Big] \le \Big(\ell_\Phi^{\mathrm{NN}} + (3-2\kappa_\Phi^{\mathrm{NN}})\ell_p^{\mathrm{NN}}\Big)\frac{1-\alpha_{\mathrm{NN}}^t}{1-\alpha_{\mathrm{NN}}}.$$

Substituting Eq. 89 and Eq. 90 proves Eq. 6. □

Remark D.1 (Finite-width neural-network approximation). The quantities $\varepsilon_p^{\mathrm{NN}}(m, \delta_{\mathrm{NN}})$ and $\varepsilon_\Phi^{\mathrm{NN}}(m, \delta_{\mathrm{NN}})$ measure the finite-width error incurred by replacing the nonlinear ReLU scores with their first-order NTK linearizations around initialization. In the NTK regime, sufficiently over-parameterized networks remain close to initialization and their outputs are well approximated by this linearized model (Jacot et al., 2018; Chizat et al., 2019; Ji et al., 2021). For two-layer ReLU networks with symmetric random initialization, high-probability local linearization bounds imply that, on a neighborhood of radius $O(m^{-1/2})$ around initialization, the approximation error decays as

$$\varepsilon_p^{\mathrm{NN}}(m, \delta_{\mathrm{NN}}) = \tilde{O}(m^{-1/2}), \qquad \varepsilon_\Phi^{\mathrm{NN}}(m,\, \delta_{\mathrm{NN}}) = \tilde{O}(m^{-1/2}),$$

Consequently, under the stated local linearization event, the additive neural approximation terms in Corollary 4.5 vanish as $m \to \infty$, recovering the same belief-stability mechanism as in the log-linear model, with the NTK tangent features playing the role of the finite-dimensional feature maps.

E Proof of Corollary 4.6

Proof. Consider $X_t := \|b_t^\Theta - b_t^{\Theta^\star}\|_{\mathrm{TV}} \in [0,1]$ and use the filtration $\{\mathcal{F}_t\}_{t\geq -1}$ defined by $\mathcal{F}_{-1} := \sigma(\varnothing, \Omega)$ and $\mathcal{F}_t := \sigma(\hat{s}_{1:t}, a_{0:t})$ for $t \geq 0$. As $\mathcal{F}_t^- \subseteq \mathcal{F}_t$, the random variable X_t is also $\mathcal{F}_t$-measurable, and in particular, under any non-anticipative policy, a_t is $\mathcal{F}_t^-$-measurable and $\mathcal{F}_t = \sigma(\mathcal{F}_t^-, a_t)$. From one-step conditional bound established in the proof of Theorem 4.1 (the bounds for terms (I)–(III) up to 58 without taking the outer expectation), for all $t \geq 1$, we proved

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] \leq \alpha X_{t-1} + c_b\left(\delta(\theta), \delta(w)\right). \tag{97}$$

Define the martingale difference sequence $D_t := \; X_t \; - \; \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right]$ and the partial sums $S_n := \sum_{t=0}^{n-1} D_t$ for $n = 1, \ldots, T$. Note that $D_0 = X_0 - \mathbb{E}[X_0|\mathcal{F}_{-1}] = 0$ since $X_0 = 0$ as $b_0^\Theta = b_0^{\Theta^\star} = \nu_0$. Then $\left\{S_n,\, \mathcal{F}_{n-1}\right\}_{n=1}^T$ is a martingale. Summing 97 over $t = 1, \ldots, T-1$ yields

$$\begin{aligned}\sum_{t=1}^{T-1} \mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] &\leq \alpha \sum_{t=1}^{T-1} X_{t-1} + (T-1) c_b\left(\delta(\theta), \delta(w)\right) \\ &\leq \; \alpha \sum_{t=0}^{T-1} X_t + (T-1) c_b\left(\delta(\theta), \delta(w)\right).\end{aligned} \tag{98}$$

where the last inequality follows from adjusting the indices and knowing that $X_t \in [0,1]$. Moreover, observe that

$$\begin{aligned}\sum_{t=0}^{T-1} X_t &= \sum_{t=0}^{T-1} D_t + \sum_{t=0}^{T-1} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \\ &= S_T + \sum_{t=0}^{T-1} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]\end{aligned}$$

together with the fact that $X_0 = 0$ and upper-bounding the last summation by 98

$$\sum_{t=0}^{T-1} X_t \leq \; \alpha \sum_{t=0}^{T-1} X_t + (T-1) c_b\left(\delta(\theta), \delta(w)\right) \; + \; S_T \tag{99}$$

$$\Rightarrow (1-\alpha) \sum_{t=0}^{T-1} X_t \leq \; (T-1) c_b\left(\delta(\theta), \delta(w)\right) \; + \; S_T \tag{100}$$

Now, for $n = 1, \ldots, T$ define $V_n := \sum_{t=0}^{n-1} \mathbb{E}\left[D_t^2 \mid \mathcal{F}_{t-1}\right]$. Since $X_t \in [0,1]$, we have $\mathbb{E}[D_t^2 \mid \mathcal{F}_{t-1}] = \mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) \leq \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}] \leq \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$, then for $n = T$, using the drift summation bound 98

$$\begin{aligned} V_T &\leq \sum_{t=0}^{T-1} \mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] \\ &\leq \alpha \sum_{t=0}^{T-1} X_t + (T-1) c_b\left(\delta(\theta), \delta(w)\right) \\ &\leq \alpha \sum_{t=0}^{T-1} X_t + T c_b\left(\delta(\theta), \delta(w)\right) \end{aligned} \tag{101}$$

Moreover, $X_t \in [0,\, 1]$ yields $|D_t| \leq 1$, then by Freedman's inequality Freedman (1975), with probability at least $1 - \delta_b$,

$$S_T \leq \sqrt{2 V_T \log(1/\delta_b)} + \frac{2}{3} \log(1/\delta_b). \tag{102}$$

Substitute 101 into 102, and then into 99:

$$\begin{aligned} (1-\alpha) \sum_{t=0}^{T-1} X_t &\leq T\, c_b\left(\delta(\theta), \delta(w)\right) + \sqrt{2 \log(1/\delta_b)(\alpha \sum_{t=0}^{T-1} X_t + T c_b\left(\delta(\theta), \delta(w)\right))} + \frac{2}{3} \log(1/\delta_b) \\ &\leq T c_b\left(\delta(\theta), \delta(w)\right) + \sqrt{2\alpha \log(1/\delta_b) \sum_{t=0}^{T-1} X_t} + \sqrt{2 T c_b\left(\delta(\theta), \delta(w)\right) \log(1/\delta_b)} + \frac{2}{3} \log(1/\delta_b) \end{aligned}$$

where last inequality follows from $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. Also, by applying the arithmetic mean-geometric mean inequality

$$\begin{aligned} \sqrt{2\alpha \log(1/\delta_b) \sum_{t=0}^{T-1} X_t} &= \Big(\sqrt{(1-\alpha) \sum_{t=0}^{T-1} X_t}\Big) . \Big(\sqrt{\frac{2\alpha \log(1/\delta_b)}{1-\alpha}}\Big) \\ &\leq \frac{(1-\alpha) \sum_{t=0}^{T-1} X_t}{2} + \frac{\alpha \log(1/\delta_b)}{1-\alpha} \end{aligned}$$

and rearranging terms we have

$$\begin{aligned} &\frac{1-\alpha}{2} \sum_{t=0}^{T-1} X_t \leq T c_b\left(\delta(\theta), \delta(w)\right) + \sqrt{2 T c_b\left(\delta(\theta), \delta(w)\right) \log(1/\delta_b)} + \log(1/\delta_b) \left(\frac{2}{3} + \frac{\alpha}{1-\alpha}\right) \\ &\Rightarrow \sum_{t=0}^{T-1} X_t \leq \frac{2 T c_b\left(\delta(\theta), \delta(w)\right)}{1-\alpha} + \frac{2\sqrt{2 T c_b\left(\delta(\theta), \delta(w)\right) \log(1/\delta_b)}}{1-\alpha} + \log(1/\delta_b) \frac{2\alpha + 4}{3(1-\alpha)^2} \\ &\Rightarrow \frac{1}{T} \sum_{t=0}^{T-1} X_t \leq \frac{2 c_b\left(\delta(\theta), \delta(w)\right)}{1-\alpha} + \frac{2}{1-\alpha} \sqrt{\frac{2 c_b\left(\delta(\theta), \delta(w)\right) \log(1/\delta_b)}{T}} + \frac{\log(1/\delta_b)}{T(1-\alpha)} \left(\frac{4}{3} + \frac{2\alpha}{1-\alpha}\right). \end{aligned}$$

The neural-softmax case follows by the same argument. One replaces $X_t = \|b_t^{\Theta} - b_t^{\Theta^\star}\|_{\mathrm{TV}}$ by $X_t^{\mathrm{NN}} = \|b_t^{W} - b_t^{W^\star}\|_{\mathrm{TV}}$, and uses the one-step recursion from Theorem 4.5,

$$\mathbb{E}[X_t^{\mathrm{NN}} \mid \mathcal{F}_{t-1}] \leq \alpha_{\mathrm{NN}} X_{t-1}^{\mathrm{NN}} + c_b^{\mathrm{NN}}(W, W^\star; \delta_{\mathrm{NN}}).$$

Repeating the preceding Freedman argument with (α, c_b) replaced by $(\alpha_{\mathrm{NN}}, c_b^{\mathrm{NN}})$ gives the identical high-probability time-average bound for the neural-softmax model. □

Lemma E.1 (Clean negative log-likelihood gradient bound)**.** *Consider the* clean *(5.2) negative log-likelihood gradient*

$$\nabla L(\mu^\star) = \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \big(\sigma(\phi_i^\top \mu^\star) - y_i\big)\phi_i.$$

where σ is the sigmoid function and $\phi_i \in \{\phi_j\}_{j=1}^{N_{\text{HF}}} \subset \mathbb{R}^d$ is the given trajectory-level exact (clean) accumulated features differences for $i-th$ trajectory, defined in 13. Given $\zeta > 0$ define the clean empirical covariance $\Sigma := \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top$ and regularized empirical covariance by $\Sigma + \zeta I$. Then, for any $\delta_c \in (0,1)$ with probability at least $1-\delta_c$ it holds

$$\big\|\nabla L(\mu^\star)\big\|_{(\Sigma+\zeta I)^{-1}} \;\le\; \frac{1}{\sqrt{N_{\text{HF}}}} \sqrt{d \log\Big(1 + \frac{4T^2 B_r^2}{\zeta d}\Big) + 2\log\Big(\frac{1}{\delta_c}\Big)}. \tag{103}$$

Proof of lemma E.1. Define the shorthand $\xi_i := \sigma(\phi_i^\top \mu^\star) - y_i \in [-1,1]$, $S := \sum_{i=1}^{N_{\text{HF}}} \xi_i \phi_i$, and $V := N_{\text{HF}}\, \zeta I + \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top$. By Assumption 5.1, item 3 (Preference Realizability) and Remark 5.6, for $0 \le i \le N_{\text{HF}} - 1$ we have $\mathbb{E}[\xi_i \mid \phi_i] = \sigma(\phi_i^\top \mu^\star) - \mathbb{E}[y_i \mid \phi_i] = \sigma(\phi_i^\top \mu^\star) - \sigma(\phi_i^\top \mu^\star) = 0$. Also, since we have $\xi_i \in [-1,1]$, by Hoeffding's lemma ξ_i is conditionally 1-sub-Gaussian given ϕ_i, i.e.,

$$\forall t \in \mathbb{R}, \qquad \mathbb{E}\big[\exp(t\xi_i) \,|\, \phi_i\big] \le \exp(\frac{t^2}{2})$$

Then, by Theorem 1 of Abbasi-yadkori et al. (2011) (elliptical potential), for any $\delta_c \in (0,1)$, with probability at least $1-\delta_c$

$$\begin{aligned}
\|S\|_{V^{-1}} &\le \sqrt{2\log\left(\frac{\det(V)^{1/2}}{\det(N_{\text{HF}}\zeta I)^{1/2}} \cdot \frac{1}{\delta_c}\right)} \\
&= \sqrt{\log\det\left(\frac{1}{N_{\text{HF}}\zeta} I \cdot \left(N_{\text{HF}}\zeta I + \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top\right)\right) + 2\log(\frac{1}{\delta_c})} \\
&= \sqrt{\log\det\left(I + \frac{1}{N_{\text{HF}}\zeta} \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top\right) + 2\log(\frac{1}{\delta_c})} \\
&\le \sqrt{d\log\Big(1 + \frac{(2TB_r)^2}{\zeta d}\Big) + 2\log\Big(\frac{1}{\delta_c}\Big)}
\end{aligned} \tag{104}$$

Where the last inequality follows from Lemma 10 in Abbasi-yadkori et al. (2011) (determinant-trace inequality). Finally, note that $\nabla L(\mu^\star) = \frac{1}{N_{\text{HF}}} S$ and $\Sigma + \zeta I = \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \phi_i \phi_i^\top + \zeta I = \frac{1}{N_{\text{HF}}} V$, we have

$$\begin{aligned}
\big\|\nabla L(\mu^\star)\big\|_{(\Sigma+\zeta I)^{-1}} &= \Big\|\frac{1}{N_{\text{HF}}} S\Big\|_{(\frac{1}{N_{\text{HF}}} V)^{-1}} = \sqrt{\Big(\frac{1}{N_{\text{HF}}} S\Big)^\top \big(N_{\text{HF}} V^{-1}\big)\Big(\frac{1}{N_{\text{HF}}} S\Big)} \\
&= \frac{1}{\sqrt{N_{\text{HF}}}} \|S\|_{V^{-1}}.
\end{aligned} \tag{105}$$

This result combined with 104 completes the proof.

□

Lemma E.2 (Trajectory-level feature perturbation bound)**.** *Fix $i \in \{1, \dots, N_{\text{HF}}\}$ and define the clean and perturbed trajectory-level feature differences ϕ_i and $\tilde{\phi}_i$ as in 13–14. Let $\Delta_{\phi,i} := \tilde{\phi}_i - \phi_i$. Suppose the event $\mathcal{E}_b^{\text{RLHF}}(\delta_b)$ in 16 holds. Then, it holds*

$$\|\Delta_{\phi,i}\|_2 \le\; 4TB_r\, \epsilon_b(\delta_b, 2N_{\text{HF}}) \tag{106}$$

where $\epsilon_b(\delta_b, 2N_{\text{HF}})$ is defined in 17.

Proof of lemma E.2. By the triangle inequality and 12 we have

$$\begin{aligned}
\|\Delta_{\phi,i}\|_2 &= \left\| \sum_{h=0}^{T-1} \big[\phi_b(b_{i,h}^{(1),\Theta}, a_{i,h}^{(1)}) - \phi_b(b_{i,h}^{(1),\Theta^\star}, a_{i,h}^{(1)})\big] - \sum_{h=0}^{T-1} \big[\phi_b(b_{i,h}^{(2),\Theta}, a_{i,h}^{(2)}) - \phi_b(b_{i,h}^{(2),\Theta^\star}, a_{i,h}^{(2)})\big] \right\|_2 \\
&\le \sum_{h=0}^{T-1} \big\|\phi_b(b_{i,h}^{(1),\Theta}, a_{i,h}^{(1)}) - \phi_b(b_{i,h}^{(1),\Theta^\star}, a_{i,h}^{(1)})\big\|_2 + \sum_{h=0}^{T-1} \big\|\phi_b(b_{i,h}^{(2),\Theta}, a_{i,h}^{(2)}) - \phi_b(b_{i,h}^{(2),\Theta^\star}, a_{i,h}^{(2)})\big\|_2 \\
&\le \sum_{h=0}^{T-1} B_r \|b_{i,h}^{(1),\Theta} - b_{i,h}^{(1),\Theta^\star}\|_1 + \sum_{h=0}^{T-1} B_r \|b_{i,h}^{(2),\Theta} - b_{i,h}^{(2),\Theta^\star}\|_1 \\
&= 2B_r \sum_{h=0}^{T-1} \|b_{i,h}^{(1),\Theta} - b_{i,h}^{(1),\Theta^\star}\|_{\mathrm{TV}} + 2B_r \sum_{h=0}^{T-1} \|b_{i,h}^{(2),\Theta} - b_{i,h}^{(2),\Theta^\star}\|_{\mathrm{TV}}. \qquad (107)
\end{aligned}$$

On the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$, for each $j \in \{1, 2\}$,

$$\frac{1}{T} \sum_{h=0}^{T-1} \|b_{i,h}^{(j),\Theta} - b_{i,h}^{(j),\Theta^\star}\|_{\mathrm{TV}} \le \epsilon_b(\delta_b, 2N_{\mathrm{HF}}),$$

hence $\sum_{h=0}^{T-1} \|b_{i,h}^{(j),\Theta} - b_{i,h}^{(j),\Theta^\star}\|_{\mathrm{TV}} \le T\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$. Substituting into 107 yields

$$\|\Delta_{\phi,i}\|_2 \le 2B_r \cdot T\epsilon_b(\delta_b, 2N_{\mathrm{HF}}) + 2B_r \cdot T\epsilon_b(\delta_b, 2N_{\mathrm{HF}}) = 4TB_r\, \epsilon_b(\delta_b, 2N_{\mathrm{HF}}),$$

which proves Eq. 106.

□

F Proof of Theorem 5.8

Proof. Assume the true reward model parameter $\mu^\star$ exists and consider the set $\mathbb{B}_2(r_\mu)$ (Assumption 5.1). Define the maximum likelihood estimation objective for perturbed features (14)

$$\begin{aligned}
\tilde{L}(\mu) &:= -\frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \log\left(\frac{\mathbf{1}\{y_i = 1\}}{1 + \exp\big(-(\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}})^\top \mu\big)} + \frac{\mathbf{1}\{y_i = 0\}}{1 + \exp\big((\tilde{\phi}^{\tau_i^{(1)},\tau_i^{(2)}})^\top \mu\big)} \right) \\
&= -\frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \Big[y_i \ln \sigma(\langle \tilde{\phi}_i, \mu \rangle) + (1 - y_i) \ln\big(1 - \sigma(\langle \tilde{\phi}_i, \mu \rangle)\big) \Big] \\
&= -\frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \Big[y_i \ln \sigma(\langle \phi_i + \Delta_{\phi,i}, \mu \rangle) + (1 - y_i) \ln\big(1 - \sigma(\langle \phi_i + \Delta_{\phi,i}, \mu \rangle)\big) \Big] \qquad (108)
\end{aligned}$$

where $\sigma(\cdot)$ is the sigmoid function. The gradient and Hessian of $\tilde{L}(\mu)$ are

$$\begin{aligned}
\nabla \tilde{L}(\mu) &= \frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \big(\sigma(\tilde{\phi}_i^\top \mu) - y_i\big) \tilde{\phi}_i \\
&= \frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \big(\sigma((\phi_i + \Delta_{\phi,i})^\top \mu) - y_i\big)(\phi_i + \Delta_{\phi,i}) \qquad (109)
\end{aligned}$$

$$\nabla^2 \tilde{L}(\mu) = \frac{1}{N_{\mathrm{HF}}} \sum_{i=1}^{N_{\mathrm{HF}}} \sigma(\tilde{\phi}_i^\top \mu)\big(1 - \sigma(\tilde{\phi}_i^\top \mu)\big)\, \tilde{\phi}_i \tilde{\phi}_i^\top. \qquad (110)$$

The function $\sigma(z)(1-\sigma(z))$ is even and strictly decreasing on $[0,\infty)$. For any $\mu \in \mathbb{B}_2(r_\mu)$, we have $|\tilde{\phi}_i^\top \mu| \leq \|\tilde{\phi}_i\|_2 \|\mu\|_2 \leq 2TB_r r_\mu$, then for all i,

$$\sigma(\tilde{\phi}_i^\top \mu)(1-\sigma(\tilde{\phi}_i^\top \mu)) \geq \frac{1}{2+\exp(2TB_r r_\mu)+\exp(-2TB_r r_\mu)} = \rho$$

substituting into 110 yields

$$\nabla^2 \tilde{L}(\mu) = \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \sigma(\tilde{\phi}_i^\top \mu)\big(1-\sigma(\tilde{\phi}_i^\top \mu)\big)\, \tilde{\phi}_i \tilde{\phi}_i^\top \tag{111}$$

$$\succeq \rho\big(\frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \tilde{\phi}_i \tilde{\phi}_i^\top\big) = \rho \tilde{\Sigma} \tag{112}$$

where $\tilde{\Sigma}$ is the feature covariance matrix defined in 18. Now, for any μ_1 and μ_2 in $\mathbb{B}_2(r_\mu)$ and line segment $\mu_\lambda = \lambda\mu_1 + (1-\lambda)\mu_2$ with $\lambda \in [0,1]$, we apply the second order Taylor expansion

$$\begin{aligned}
&\tilde{L}(\mu_1) - \tilde{L}(\mu_2) - \langle \nabla \tilde{L}(\mu_2), \mu_1 - \mu_2 \rangle \\
&\quad = \int_0^1 (1-\lambda)\,(\mu_1-\mu_2)^\top \nabla^2 \tilde{L}(\mu_\lambda)\,(\mu_1-\mu_2)\, d\lambda \\
&\quad \geq \frac{\rho}{2}\,(\mu_1-\mu_2)^\top (\tilde{\Sigma})(\mu_1-\mu_2).
\end{aligned} \tag{113}$$

Therefore,

$$\tilde{L}(\mu_1) - \tilde{L}(\mu_2) - \langle \nabla \tilde{L}(\mu_2), \mu_1 - \mu_2 \rangle \geq \frac{\rho}{2} \|\mu_1 - \mu_2\|_{\tilde{\Sigma}}^2. \tag{114}$$

This proves the $\rho-$strong convexity of $\tilde{L}(\mu)$ with $\mu \in \mathbb{B}_2(r_\mu)$ with respect to the semi-norm $\|\cdot\|_{\tilde{\Sigma}}$. Next, writing 113 for $\mu^\star$ and $\tilde{\mu}$ yields

$$\begin{aligned}
\frac{\rho}{2}\,\|\tilde{\mu}-\mu^\star\|_{\tilde{\Sigma}}^2 \leq\ & \tilde{L}(\tilde{\mu}) - \tilde{L}(\mu^\star) - \langle \nabla \tilde{L}(\mu^\star), \tilde{\mu}-\mu^\star \rangle \\
\overset{(a)}{\leq}\ & -\langle \nabla \tilde{L}(\mu^\star), \tilde{\mu}-\mu^\star \rangle \\
\overset{(b)}{\leq}\ & \|\nabla \tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}}\, \|\tilde{\mu}-\mu^\star\|_{(\tilde{\Sigma}+\zeta I)}
\end{aligned} \tag{115}$$

where (a) follows from the fact that $\tilde{L}(\tilde{\mu}) - \tilde{L}(\mu^\star) \leq 0$ as $\tilde{\mu} \in \arg\min_{\mu \in \mathbb{B}_2(r_\mu)} \tilde{L}(\mu)$, and (b) follows from applying Cauchy–Schwarz in the dual pair of norms induced by $\tilde{\Sigma} + \zeta I$. Using the identity $\|\tilde{\mu}-\mu^\star\|_{\tilde{\Sigma}}^2 = \|\tilde{\mu}-\mu^\star\|_{(\tilde{\Sigma}+\zeta I)}^2 - \zeta\,\|\tilde{\mu}-\mu^\star\|_2^2$, we have

$$\frac{\rho}{2}\,\|\tilde{\mu}-\mu^\star\|_{(\tilde{\Sigma}+\zeta I)}^2 \leq \|\nabla \tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}}\, \|\tilde{\mu}-\mu^\star\|_{(\tilde{\Sigma}+\zeta I)} + 2\rho\,\zeta\, r_\mu^2. \tag{116}$$

Solving this quadratic inequality and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ yields

$$\|\tilde{\mu}-\mu^\star\|_{(\tilde{\Sigma}+\zeta I)} \leq\ \frac{2}{\rho}\,\|\nabla \tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}}\ +\ 2r_\mu\sqrt{\zeta}. \tag{117}$$

Now we proceed with bounding $\|\nabla \tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}}$ with reference to 109. we define per-sample gradients at $\mu^\star$ as $g_i^\star := \big(\sigma(\phi_i^\top \mu^\star) - y_i\big)\phi_i$ and $\tilde{g}_i^\star := \big(\sigma(\tilde{\phi}_i^\top \mu^\star) - y_i\big)\tilde{\phi}_i$, with respect to exact and perturbed features. Hence, $\nabla \tilde{L}(\mu^\star) = \frac{1}{N_{\text{HF}}} \sum_i \tilde{g}_i(\mu^\star)$, and we can write

$$\nabla \tilde{L}(\mu^\star) = \nabla L(\mu^\star) + \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \big(\tilde{g}_i^\star - g_i^\star\big). \tag{118}$$

Consider the following decomposition

$$\tilde{g}_i^\star - g_i^\star = \underbrace{\Big(\sigma(\phi_i^\top \mu^\star + \Delta_{\phi,i}^\top \mu^\star) - \sigma(\phi_i^\top \mu^\star)\Big)\phi_i}_{(I)} + \underbrace{\Big(\sigma(\phi_i^\top \mu^\star + \Delta_{\phi,i}^\top \mu^\star) - y_i\Big)\Delta_{\phi,i}}_{(II)}.$$

Since $\sigma(\cdot)$ is $\frac{1}{4}$-Lipschitz, $\sigma(\cdot) \in (0,1)$ and $y_i \in \{0,1\}$, we have

$$\|(I)\|_2 \le \frac{1}{4}\,|\Delta_{\phi,i}^\top \mu^\star|\,\|\phi_i\|_2 \le \frac{2TB_r}{4}\|\Delta_{\phi,i}\|_2\|\mu^\star\|_2 \le \frac{TB_r}{2} r_\mu \|\Delta_{\phi,i}\|_2$$

$$\|(II)\|_2 \le \|\Delta_{\phi,i}\|_2.$$

where we used $\mu^\star \in \mathbb{B}_2(r_\mu)$, $|\Delta_{\phi,i}^\top \mu^\star| \le \|\Delta_{\phi,i}\|_2\|\mu^\star\|_2 \le r_\mu\|\Delta_{\phi,i}\|_2$, together with $\|\phi_i\|_2 \le 2TB_r$. Thus, for $i = 1, \dots, N_{\mathrm{HF}}$, we obtain

$$\|\tilde{g}_i^\star - g_i^\star\|_2 \le \Big(1 + \tfrac{1}{2}\,TB_r\,r_\mu\Big)\,\|\Delta_{\phi,i}\|_2, \tag{119}$$

which yields

$$\begin{aligned}\big\|\nabla\tilde{L}(\mu^\star) - \nabla L(\mu^\star)\big\|_2 = \Big\|\frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}(\tilde{g}_i^\star - g_i^\star)\Big\|_2 \;&\le\; \frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\|\tilde{g}_i^\star - g_i^\star\|_2 \\ &\le \Big(1 + \tfrac{1}{2}\,TB_r\,r_\mu\Big)\cdot\frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\|\Delta_{\phi,i}\|_2.\end{aligned} \tag{120}$$

Since $\tilde{\Sigma} + \zeta I \succeq \zeta I$, the inequality

$$\begin{aligned}\big\|\nabla\tilde{L}(\mu^\star) - \nabla L(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} &\le \frac{1}{\sqrt{\zeta}}\big\|\nabla\tilde{L}(\mu^\star) - \nabla L(\mu^\star)\big\|_2 \\ &\le \frac{1}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}\,TB_r\,r_\mu\Big)\cdot\frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\|\Delta_{\phi,i}\|_2\end{aligned} \tag{121}$$

on the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ Eq. 16 where $\|\Delta_{\phi,i}\|_2 \le 4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$ for all i with the probability at least $1 - \delta_b$, implies

$$\begin{aligned}\big\|\nabla\tilde{L}(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} &\le \big\|\nabla L(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} + \frac{1}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}\,TB_r\,r_\mu\Big)\cdot\frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\|\Delta_{\phi,i}\|_2 \\ &\le \big\|\nabla L(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} + \frac{4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}\,TB_r\,r_\mu\Big).\end{aligned} \tag{122}$$

By lemma E.1, for any $\delta_c \in (0,1)$ with probability at least $1 - \delta_c$ we have

$$\big\|\nabla L(\mu^\star)\big\|_{(\Sigma+\zeta I)^{-1}} \le \frac{1}{\sqrt{N_{\mathrm{HF}}}}\sqrt{d\log\Big(1 + \frac{4T^2B_r^2}{\zeta d}\Big) + 2\log\Big(\frac{1}{\delta_c}\Big)}. \tag{123}$$

Now, we study the covariance perturbation. Define

$$E := \tilde{\Sigma} - \Sigma = \frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\Big(\phi_i\Delta_{\phi,i}^\top + \Delta_{\phi,i}\phi_i^\top + \Delta_{\phi,i}\Delta_{\phi,i}^\top\Big)$$

which is a symmetric. On $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$, we have $\|\Delta_{\phi,i}\|_2 \le 4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$, for all i. Moreover, we have $\|\phi_i\|_2 \le 2TB_r$, hence, using $\|uv^\top\|_{op} \le \|u\|_2\|v\|_2$ and the triangle inequality,

$$\begin{aligned}
\|E\|_{op} &\le \frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\Big(\|\phi_i\Delta_{\phi,i}^\top\|_{op} + \|\Delta_{\phi,i}\phi_i^\top\|_{op} + \|\Delta_{\phi,i}\Delta_{\phi,i}^\top\|_{op}\Big)\\
&\le \frac{1}{N_{\mathrm{HF}}}\sum_{i=1}^{N_{\mathrm{HF}}}\Big(2\|\phi_i\|_2\|\Delta_{\phi,i}\|_2 + \|\Delta_{\phi,i}\|_2^2\Big)\\
&\le 16T^2B_r^2\epsilon_b(\delta_b, 2N_{\mathrm{HF}}) + (4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}}))^2 \qquad (124)\\
&= 16T^2B_r^2\epsilon_b(\delta_b, 2N_{\mathrm{HF}})(1+\epsilon_b(\delta_b, 2N_{\mathrm{HF}}))\\
&:= \epsilon_\Sigma \qquad (125)
\end{aligned}$$

Therefore, a sufficient level of regularization on $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ is ϵ_Σ. Then, for any $1 < c_\zeta$, choosing $\zeta = c_\zeta\epsilon_\Sigma$ on $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ yields

$$\frac{\|E\|_{op}}{\zeta} \le \frac{1}{c_\zeta} < 1 \qquad \Longrightarrow \qquad \frac{1}{\sqrt{1-\|E\|_{op}/\zeta}} \le \sqrt{\frac{c_\zeta}{c_\zeta - 1}}. \qquad (126)$$

Since E is symmetric, $-\|E\|_{op}I \preceq E \preceq \|E\|_{op}I$, and hence

$$\begin{aligned}
&\Sigma + \zeta I - \|E\|_{op}I \;\preceq\; \tilde{\Sigma} + \zeta I \;\preceq\; \Sigma + \zeta I + \|E\|_{op}I\\
&\Longrightarrow \Sigma + \zeta I - \|E\|_{op}I \succeq \Big(1 - \frac{\|E\|_{op}}{\zeta}\Big)(\Sigma + \zeta I)\\
&\Longrightarrow (\tilde{\Sigma} + \zeta I)^{-1} \;\preceq\; \frac{1}{1-\|E\|_{op}/\zeta}(\Sigma + \zeta I)^{-1} \qquad (127)
\end{aligned}$$

as $\|E\|_{op}/\zeta < 1$. This implies that for any $v \in \mathbb{R}^d$,

$$\|v\|_{(\tilde{\Sigma}+\zeta I)^{-1}} \;\le\; \frac{1}{\sqrt{1-\|E\|_{op}/\zeta}}\,\|v\|_{(\Sigma+\zeta I)^{-1}} \le \sqrt{\frac{c_\zeta}{c_\zeta - 1}}\|v\|_{(\Sigma+\zeta I)^{-1}}. \qquad (128)$$

Finally, returning to Eq. 122 and combining Eq. 128 with Lemma E.1 (Eq. Eq. 123), on the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ we obtain

$$\begin{aligned}
\big\|\nabla\tilde{L}(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} &\le \big\|\nabla L(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} + \frac{4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}TB_r\,r_\mu\Big)\\
&\le \frac{1}{\sqrt{1-\|E\|_{op}/\zeta}}\,\big\|\nabla L(\mu^\star)\big\|_{(\Sigma+\zeta I)^{-1}} + \frac{4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}TB_r\,r_\mu\Big)\\
&\le \sqrt{\frac{c_\zeta}{c_\zeta - 1}}\,\big\|\nabla L(\mu^\star)\big\|_{(\Sigma+\zeta I)^{-1}} + \frac{4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}TB_r\,r_\mu\Big)\\
&\le \sqrt{\frac{c_\zeta}{N_{\mathrm{HF}}(c_\zeta - 1)}}\sqrt{d\log\Big(1 + \frac{4T^2B_r^2}{\zeta d}\Big) + 2\log\Big(\frac{1}{\delta_c}\Big)} + \frac{4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}TB_r\,r_\mu\Big).
\end{aligned}$$

Substituting the above bound into Eq. 117, on the event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ and the event of Lemma E.1 (with probability at least $1-\delta_c$), we obtain

$$\begin{aligned}
\|\tilde{\mu} - \mu^\star\|_{(\tilde{\Sigma}+\zeta I)} &\le \frac{2}{\rho}\big\|\nabla\tilde{L}(\mu^\star)\big\|_{(\tilde{\Sigma}+\zeta I)^{-1}} \;+\; 2r_\mu\sqrt{\zeta}\\
&\le \frac{2\sqrt{c_\zeta}}{\rho\sqrt{N_{\mathrm{HF}}(c_\zeta - 1)}}\sqrt{d\log\Big(1 + \frac{4T^2B_r^2}{\zeta d}\Big) + 2\log\Big(\frac{1}{\delta_c}\Big)}\\
&\qquad + \frac{8TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})}{\rho\sqrt{\zeta}}\Big(1 + \tfrac{1}{2}TB_r\,r_\mu\Big) \;+\; 2r_\mu\sqrt{\zeta}. \qquad (129)
\end{aligned}$$

Therefore, taking a union bound over $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$ and the event of Lemma E.1, the claim of Theorem 5.8 holds with probability at least $1 - \delta_b - \delta_c$.

The proof for the neural-softmax model is identical. By the modular structure of the pipeline, the only POMDP model-specific input used above is the belief-accuracy event $\mathcal{E}_b^{\mathrm{RLHF}}(\delta_b)$, through the bound $\|\Delta_{\phi,i}\|_2 \le 4TB_r\,\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$. Under Theorem 4.5 and its corresponding high-probability event established in the second part of Corollary 4.6, similar event holds with $b_{i,h}^{(j),\Theta}, b_{i,h}^{(j),\Theta^\star}$ replaced by $b_{i,h}^{(j),W}, b_{i,h}^{(j),W^\star}$, and with $\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$ replaced by $\epsilon_b^{\mathrm{NN}}(\delta_b, 2N_{\mathrm{HF}}, \delta_{\mathrm{NN}})$. All subsequent steps, e.g., the feature perturbation bound, covariance perturbation bound, gradient decomposition, and strong-convexity argument are downstream and therefore remain unchanged after this replacement. □

Remark F.1. A common approach to bound $\|\nabla\tilde{L}(\mu^\star)\|_{(\tilde{\Sigma}+\zeta I)^{-1}}$ would be to apply an elliptical potential inequality directly to the summands $\tilde{g}_i^\star = \big(\sigma(\tilde{\phi}_i^\top \mu^\star) - y_i\big)\tilde{\phi}_i, i = 1, \ldots, N_{\mathrm{HF}}$. as in E.1, but this is not valid here because the required conditional mean-zero property fails as $\mathbb{E}[y_i \mid \phi_i] = \sigma(\phi_i^\top \mu^\star)$, hence in general $\mathbb{E}\big[\sigma(\tilde{\phi}_i^\top \mu^\star) - y_i \,\big|\, \tilde{\phi}_i\big] \neq 0$. Therefore, $\{\tilde{g}_i^\star\}$ is not a martingale-difference sequence w.r.t. $\tilde{\phi}_i$, and we instead perform the decompose in 122 and apply the self-normalized bound only to the clean gradient $\nabla L(\mu^\star)$, for which $\mathbb{E}[\sigma(\phi_i^\top \mu^\star) - y_i \mid \phi_i] = 0$ holds.

Remark F.2. The inequality in Eq. 117 is a standard device that makes the analysis well-posed and compatible with self-normalized concentration arguments. Indeed, the strong convexity step yields control only in the semi-norm induced by $\tilde{\Sigma}$,

$$\tilde{L}(\mu_1) - \tilde{L}(\mu_2) - \langle \nabla\tilde{L}(\mu_2), \mu_1 - \mu_2 \rangle \;\ge\; \frac{\rho}{2}\,\|\mu_1 - \mu_2\|_{\tilde{\Sigma}}^2,$$

where $\tilde{\Sigma}$ may be singular (see also Agarwal et al. (2020); Du et al. (2024); Zhu et al. (2023); Abbasi-yadkori et al. (2011)). Introducing the term ζI ensures invertibility and allows us to apply Cauchy–Schwarz in the dual pair $\|\cdot\|_{\tilde{\Sigma}+\zeta I}$ and $\|\cdot\|_{(\tilde{\Sigma}+\zeta I)^{-1}}$. This modification contributes an additive regularization bias of order $\sqrt{\zeta}\|\tilde{\mu} - \mu^\star\|_2 \le 2r_\mu\sqrt{\zeta}$ (e.g., Theorem 5.2 in Zhu et al. (2023)).

Remark F.3 (Making ζ fully explicit). Although we write

$$\zeta \;:=\; c_\zeta\, 16T^2 B_r^2\, \epsilon_b(\delta_b, 2N_{\mathrm{HF}})\Big(1 + \epsilon_b(\delta_b, 2N_{\mathrm{HF}})\Big),$$

the quantity $\epsilon_b(\delta_b, 2N_{\mathrm{HF}})$ is *explicit* in our analysis: it is given in Eq. 17 (see also Corollary 4.6). In particular, ϵ_b depends only on $(T, N_{\mathrm{HF}}, \delta_b)$ and the problem parameters $\alpha = (1-\kappa_P)(4-3\kappa_\Phi)$ and $c_b(\delta(\theta), \delta(w))$. To obtain a deterministic bound that does not depend on the unknown perturbations, $\delta(\theta), \delta(w)$ (or on κ_P, κ_Φ), one may replace them by worst-case bounds over the feasible set, e.g. $\delta(\theta), \delta(w) \le r_\Theta$ and lower bounds on κ_P, κ_Φ from Lemma A.2, thereby obtaining a closed-form upper bound $\bar{\epsilon}_b$ and the explicit choice

$$\zeta \;:=\; c_\zeta\, 16T^2 B_r^2\, \bar{\epsilon}_b\,(1 + \bar{\epsilon}_b).$$

This substitution only affects constants, while preserving the stated guarantees. Moreover, when $\epsilon_b = 0$, the same argument may be run with any $\zeta > 0$.

Remark F.4 (Bias scaling under the prescribed choice of ζ). The bound in Theorem 5.8 depends on ζ in all three terms; the statistical term depends on ζ only through the logarithmic factor, while the explicit belief-mismatch and regularization bias terms have the form $C_1\epsilon_b/\sqrt{\zeta}$ and $C_2\sqrt{\zeta}$. Under the prescribed tuning $\zeta = \bar{c}\,\epsilon_b(1+\epsilon_b)$ (Theorem 5.8 and Remark F.3), the two explicit bias contributions satisfy

$$C_1\frac{\epsilon_b}{\sqrt{\zeta}} + C_2\sqrt{\zeta} = \frac{C_1}{\sqrt{\bar{c}}}\sqrt{\frac{\epsilon_b}{1+\epsilon_b}} + C_2\sqrt{\bar{c}}\sqrt{\epsilon_b(1+\epsilon_b)} = \mathcal{O}(\sqrt{\epsilon_b}),$$

where we used that $\epsilon_b \in (0, 1]$ implies $1 + \epsilon_b \in [1, 2]$. Thus, as the time-averaged belief error ϵ_b decreases, the explicit belief-mismatch/regularization bias floor does not blow up under this tuning and decays at rate $\sqrt{\epsilon_b}$, up to the logarithmic ζ-dependence in the statistical term.

Corollary F.5 (Value perturbation from reward parameter error)**.** *Working under the setting of Theorem 5.8, fix the (possibly randomized) policy $\pi : \Delta_{\mathbb{S}} \to \Delta_{\mathbb{A}}$ used in data generation and any $\gamma \in (0,1)$. For $\mu \in \mathbb{R}^d$, define the discounted value*

$$V_\mu^\pi(b) \;:=\; \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t\, r_\mu(b_t, a_t) \;\middle|\; b_0 = b\right].$$

Then with probability at least $1 - (\delta_b + \delta_c)$,

$$\sup_{b \in \Delta_{\mathbb{S}}} \left|V_{\tilde{\mu}}^\pi(b) - V_{\mu^\star}^\pi(b)\right| \;\le\; \frac{B_r}{(1-\gamma)\sqrt{\zeta}}\; \epsilon(\delta_b, \delta_c, \zeta),$$

where $\epsilon(\delta_b, \delta_c, \zeta)$ is as in Eq. 20.

Proof of Corollary F.5. Fix any $(b, a) \in \Delta_{\mathbb{S}} \times \mathbb{A}$. By boundedness of the belief feature map, $\|\phi_b(b, a)\|_2 \le B_r$, hence for all $\mu, \mu' \in \mathbb{R}^d$,

$$\begin{aligned} |r_\mu(b,a) - r_{\mu'}(b,a)| &= \left|\phi_b(b,a)^\top(\mu - \mu')\right| \\ &\le B_r\, \|\mu - \mu'\|_2. \end{aligned}$$

Therefore, by linearity of expectation and $\sum_{t \ge 0} \gamma^t = (1-\gamma)^{-1}$,

$$\left|V_\mu^\pi(b) - V_{\mu'}^\pi(b)\right| \le \frac{B_r}{1-\gamma}\, \|\mu - \mu'\|_2.$$

Taking $\mu = \tilde{\mu}$, $\mu' = \mu^\star$, and using $\tilde{\Sigma} + \zeta I \succeq \zeta I$ (so $\|\nu\|_2 \le \zeta^{-1/2} \|\nu\|_{\tilde{\Sigma} + \zeta I}$), we obtain

$$\left|V_{\tilde{\mu}}^\pi(b) - V_{\mu^\star}^\pi(b)\right| \le \frac{B_r}{(1-\gamma)\sqrt{\zeta}}\; \|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma} + \zeta I}.$$

On the event of Theorem 5.8, $\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma} + \zeta I} \le \epsilon(\delta_b, \delta_c, \zeta)$. Taking the supremum over b completes the proof.

□

G Experiments

G.1 Belief-stability bound

We first test the belief-stability component of Theorem 4.1. The purpose of this experiment is to verify the qualitative dependences predicted by the theorem once a controlled model perturbation is introduced. In particular, Theorem 4.1 predicts that, for histories generated from the true model,

$$\mathbb{E}\Big[\|b_t^\Theta - b_t^{\Theta^\star}\|_{\mathrm{TV}}\Big] \le \frac{1-\alpha^t}{1-\alpha}\, c_b(\delta(\theta), \delta(w)), \qquad c_b(\delta(\theta), \delta(w)) = B\left(\delta(w) + \frac{3 - 2\kappa_\Phi}{2}\delta(\theta)\right).$$

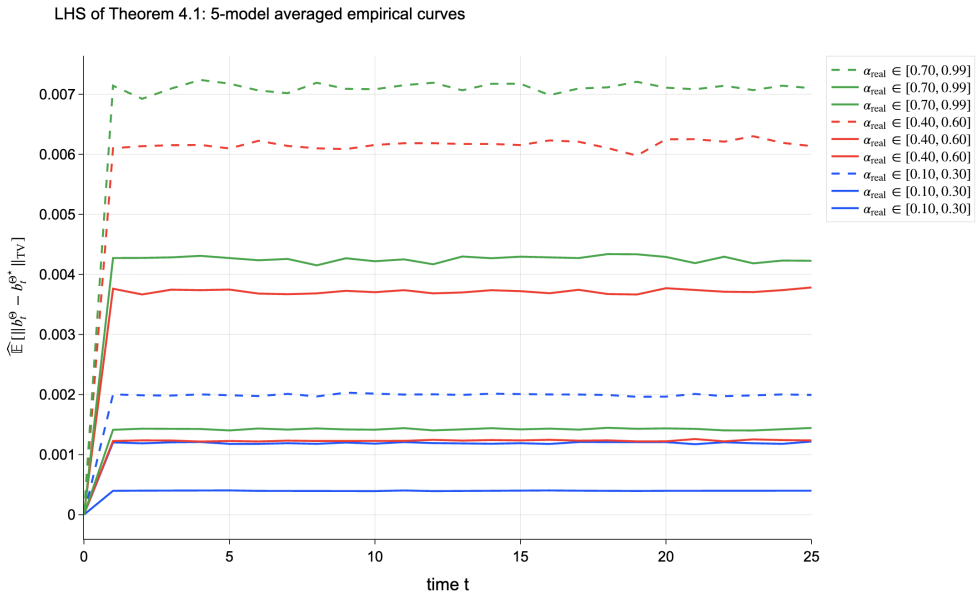
Thus, for $\alpha < 1$, the time-propagation factor saturates at order $(1-\alpha)^{-1}$, while the one-step perturbation term is linear in the parameter deviations. The experiment is designed to separately demonstrates these two effects.

Setting. We generate finite log-linear POMDPs satisfying the bounded-feature condition

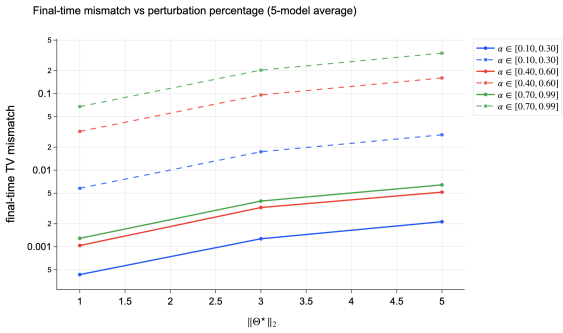
$$\|\phi_p(s, a, s')\|_2 \le B, \qquad \|\phi_\Phi(s, \hat{s})\|_2 \le B,$$

with $|\mathbb{S}| = 20$, $|\mathbb{A}| = 10$, $|\hat{\mathbb{S}}| = 20$, $d_\theta = d_w = 10$, horizon $T = 25$, and $B = 0.1$. True parameters $\Theta^\star = (\theta^\star, w^\star)$ are sampled and retained according to the realized stability coefficient

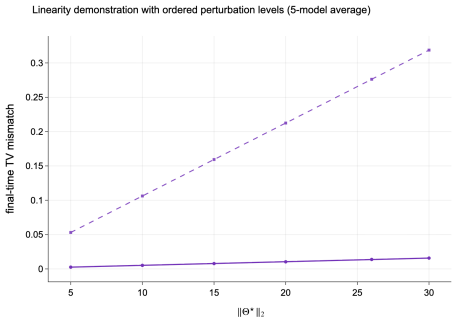
$$\alpha_{\mathrm{real}} := (1 - \kappa_P)(4 - 3\kappa_\Phi),$$



(a) Empirical estimates of the LHS, $\widehat{\mathbb{E}}[\|b_t^{\Theta} - b_t^{\Theta^\star}\|_{\mathrm{TV}}]$, over time for different realized stability bands and perturbation levels.



(b) Final-time empirical mismatch and realized theorem RHS as a function of the perturbation magnitude, grouped by realized α-bands.



(c) Fixed-α linearity check: both the final-time empirical mismatch and the realized theorem RHS scale approximately linearly with the perturbation magnitude.

Figure 2: Synthetic validation of the belief-stability mechanism in Theorem 4.1. The experiment isolates the filtering component by comparing Bayesian filters computed under the true log-linear POMDP model and a perturbed model on common histories generated from the true model. The plots are intended to test the qualitative predictions of the theorem: bounded-in-time propagation under $\alpha < 1$, larger mismatch for weaker stability, and approximately linear dependence on the parameter perturbation size.

so that the accepted models fall into prescribed stability bands. We then form a perturbed model $\Theta = (\theta, w)$ by moving $\Theta^\star$ in a fixed perturbation direction with prescribed magnitude. For each accepted model, observation histories are generated from the true POMDP, and the two Bayesian filters $b_t^{\Theta^\star}$ and b_t^{Θ} are run on the same action–observation histories. This common-history comparison matches the distributional object in Theorem 4.1 and removes policy-induced trajectory drift from the experiment. To stress the least stable case within the fixed-action-sequence setting of the theorem, the action sequence is fixed to the worst realized action for the sampled model.

The plotted empirical quantity is the Monte Carlo estimate

$$\widehat{\mathbb{E}}\Big[\|b_t^{\Theta} - b_t^{\Theta^\star}\|_{\mathrm{TV}}\Big],$$

where the expectation is over observation histories generated under $\Theta^\star$. In the reported plots, curves are averaged over 5 independently accepted models in the corresponding stability regime and over 50 simulated observation histories per model. The displayed RHS is obtained by evaluating the theorem expression with the realized κ_P, κ_Φ and the corresponding perturbation term $c_b(\delta(\theta), \delta(w))$. Since this uses realized stability constants and finite Monte Carlo estimates, the figure should be interpreted as a synthetic stability diagnostic rather than a claim of tightness of the worst-case theorem constants.

Results. Figure 2a shows that the empirical belief mismatch rises from zero and then remains stable over the horizon. This behavior is consistent with the structure of Theorem 4.1; the two filters start from the same prior, model mismatch injects fresh error after filtering begins, and the Dobrushin contraction prevents this error from accumulating linearly in time. The ordering across stability regimes is also consistent with the theorem. Models with larger realized α_{real} exhibit larger belief mismatch, while more stable models have substantially smaller plateaus. Within each stability band, increasing the perturbation magnitude increases the empirical mismatch.

Figure 2b compares the final-time empirical mismatch with the realized theorem RHS. The empirical curves remain below the corresponding proof-level upper bounds and inherit the same monotone structure and demonstrate how larger parameter perturbations lead to larger final-time belief error, and weaker stability, i.e., larger α_{real}, shifts both the empirical mismatch and the bound upward. The gap between the empirical values and the RHS is expected, because the theorem uses total-variation Lipschitz bounds, Dobrushin worst-case contractions, and a uniform propagation argument; these constants are designed for robustness rather than numerical tightness.

Finally, Figure 2c isolates the perturbation dependence by using a separate fixed-stability experiment with $\alpha_{\text{real}} \in [0.25, 0.40]$. In this setting the propagation factor $(1-\alpha^t)/(1-\alpha)$ is approximately fixed across models, so the theorem predicts that the dominant dependence on the perturbation size should enter through the linear term $c_b(\delta(\theta), \delta(w))$. The observed final-time empirical mismatch and the realized RHS both grow approximately linearly with the perturbation magnitude, supporting the Lipschitz interpretation of Theorem 4.1. Overall, under a stable filter, model mismatch induces a controlled belief-error level, but this error increases with both the learned-model perturbation and the weakness of the realized stability coefficient.

G.2 Reward-learning

We next isolate the downstream reward-estimation component of Theorem 5.8. Owing to the modular structure of our error-propagation pipeline, this experiment separates reward estimation from the preceding belief-filtering analysis. In particular, it does not estimate a POMDP model and does not run Bayesian filtering. Instead, it directly generates clean belief-based trajectory features and perturbed belief-based trajectory features whose discrepancy is controlled through a prescribed belief-accuracy level ϵ_b. The goal is therefore to test the second part of the theory and show that once belief approximation induces trajectory-feature perturbations, how does this perturbation affect Bradley–Terry reward estimation.

The theorem controls the perturbed-feature estimator in the adaptive covariance geometry. In particular, ignoring logarithmic factors, Theorem 5.8 has the qualitative form of

$$\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I} \lesssim \underbrace{N_{\text{HF}}^{-1/2}}_{\text{statistical error}} + \underbrace{\frac{TB_r}{\sqrt{\zeta}}\epsilon_b}_{\text{belief-induced bias}} + \underbrace{r_\mu\sqrt{\zeta}}_{\text{regularization bias}} .$$

Thus, for a fixed belief-error level ϵ_b and fixed regularization ζ, increasing N_{HF} should reduce the statistical component, while the belief-induced and regularization terms determine the residual scale that cannot be removed merely by collecting more preference comparisons.

Setting. We use a finite synthetic belief-MDP with $|\mathbb{S}| = 20$, $|\mathbb{A}| = 5$, reward-feature dimension $d = 5$, horizon $T = 10$, feature bound $B_r = 0.1$, and parameter constraint radius $r_\mu = 1$. The true reward parameter

satisfies $\|\mu^\star\|_2 = 0.8 r_\mu$. For each

$$N_{\mathrm{HF}} \in \{10, 100, 500, 1000, 5000, 10000, 20000, 50000\} \quad \text{and} \quad \epsilon_b \in \{0.05, 0.10, 0.15, 0.20, 0.30\},$$

we generate N_{HF} independent pairwise comparisons. Clean belief-based trajectory features are denoted by ϕ_i, while perturbed features are denoted by $\tilde{\phi}_i$. The perturbations are constructed so that each trajectory satisfies the prescribed average total-variation belief-error level ϵ_b, matching the type of belief-accuracy event used in the proof of Theorem 5.8.

Preference labels are sampled from the clean Bradley–Terry model

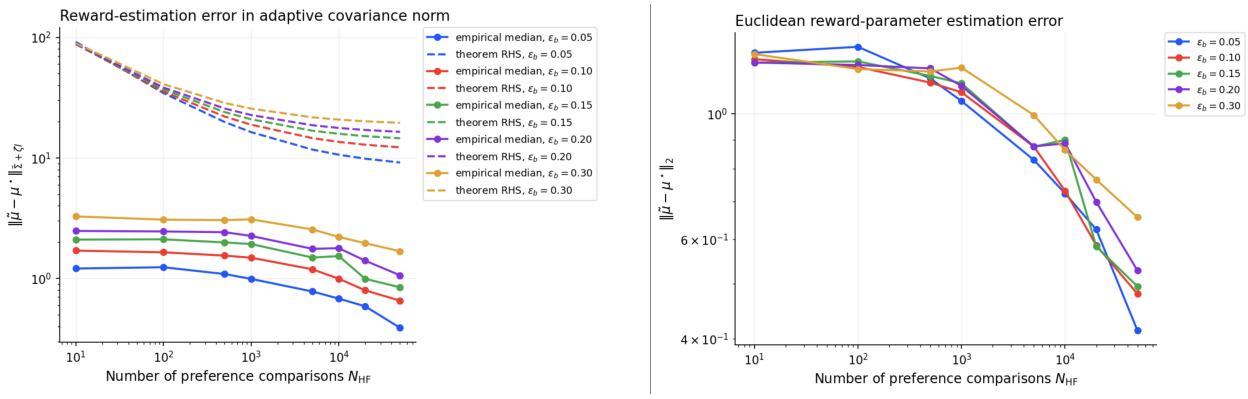
$$\mathbb{P}(y_i = 1 \mid \phi_i) = \sigma(\phi_i^\top \mu^\star),$$

whereas the estimator is fitted using the perturbed features $\tilde{\phi}_i$:

$$\tilde{\mu} \in \arg\min_{\|\mu\|_2 \le r_\mu} -\sum_{i=1}^{N_{\mathrm{HF}}} \Big[y_i \log \sigma(\tilde{\phi}_i^\top \mu) + (1 - y_i) \log\big(1 - \sigma(\tilde{\phi}_i^\top \mu)\big) \Big].$$

This creates the downstream misspecification studied by the theorem, that the preferences are generated according to the oracle belief features, but the learner observes only perturbed belief-induced features.

For each ϵ_b, we set the adaptive-norm regularization level according to the deterministic theorem scale $\zeta = 1.05 \cdot 16 T^2 B_r^2 \epsilon_b (1 + \epsilon_b)$, up to a negligible numerical floor.



(a) The error $\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I}$ and the corresponding proof-level RHS.

(b) Euclidean parameter error $\|\tilde{\mu} - \mu^\star\|_2$, reported only as an auxiliary diagnostic.

Figure 3: Synthetic downstream reward-learning experiment. The left panel reports the adaptive covariance norm controlled by Theorem 5.8; the right panel reports the Euclidean error only for interpretability.

Results. Figure 3 reports medians over 100 independent trials. The left panel shows the theorem-relevant quantity $\|\tilde{\mu} - \mu^\star\|_{\tilde{\Sigma}+\zeta I}$ together with the corresponding theorem RHS. The empirical adaptive-norm error decreases as N_{HF} increases, reflecting the reduction of the statistical component in Theorem 5.8. At the same time, for a fixed sample size, larger prescribed belief error ϵ_b generally leads to larger estimation error. This is consistent with the perturbation term in the theorem, where belief mismatch enters through the accumulated trajectory-feature error.

The empirical curves remain below the proof-level RHS across the tested regimes. The gap is expected and should not be interpreted as a failure of the scaling law, as he theorem is a high-probability guarantee and uses conservative ingredients. The left panel also illustrates the main qualitative implication of the theorem. More preference comparisons reduce the statistical error, but they do not remove the effect of a fixed belief approximation error. In the theorem RHS, once the statistical term becomes small, the remaining scale is governed by the belief-induced bias term and the regularization bias. This matches the conceptual message

of the paper, that under partial observability with imperfect belief construction, reward learning has an additional error channel that is absent from the fully observed Bradley–Terry setting.

The right panel reports $\|\tilde{\mu} - \mu^\star\|_2$ as a diagnostic in the ordinary Euclidean parameter norm. This is not the norm controlled by Theorem 5.8; therefore, crossings and small non-monotonicities across ϵ_b-curves should not be overinterpreted. Converting the adaptive-norm bound into a Euclidean bound would require additional lower-eigenvalue control of $\tilde{\Sigma}+\zeta I$. Nevertheless, the Euclidean plot shows the same broad behavior mentioned earlier that increasing N_{HF} improves reward-parameter recovery, while larger belief perturbations tend to make estimation harder.

Summary of experimental findings. Overall, the experiments provide controlled synthetic checks of the two error-propagation mechanisms analyzed in the paper. The belief-stability experiment shows that larger model perturbations and weaker realized stability produce larger belief mismatch, while the error remains stable over the tested horizon. The reward-learning experiment isolates the effect of a prescribed belief-error level and shows the expected pattern, that increasing N_{HF} reduces the statistical component, whereas fixed belief perturbation leaves a residual error contribution.