Reinforcement Learning with History-Dependent Dynamic Contexts

Guy Tennenholtz [*1] **Nadav Merlis** [*2] **Lior Shani** [1] **Martin Mladenov** [1] **Craig Boutilier** [1]

Abstract

We introduce *Dynamic Contextual Markov Decision Processes (DCMDPs)*, a novel reinforcement learning framework for history-dependent environments that generalizes the contextual MDP framework to handle non-Markov environments, where contexts change over time. We consider special cases of the model, with a focus on *logistic DCMDPs*, which break the exponential dependence on history length by leveraging aggregation functions to determine context transitions. This special structure allows us to derive an upper-confidence-bound style algorithm for which we establish regret bounds. Motivated by our theoretical results, we introduce a practical model-based algorithm for logistic DCMDPs that plans in a latent space and uses optimism over history-dependent features. We demonstrate the efficacy of our approach on a recommendation task (using MovieLens data) where user behavior dynamics evolve in response to recommendations.

1. Introduction

Reinforcement learning (RL) is a paradigm in which an agent learns to act in an environment to maximize long-term reward. RL has been applied to numerous domains, including recommender systems, robot control, video games, and autonomous vehicles (Afsar et al., 2022; Tessler et al., 2019; Mnih et al., 2015; Fayjie et al., 2018). While typical RL approaches rely on a Markov property of both the reward process and environment dynamics, many scenarios are inherently history-dependent (Bacchus et al., 1996; Ronca and Giacomo, 2021), particularly, when humans are involved. As one example, the behavior of users in recommender systems often exhibits non-Markovian characteristics reflective of a user's latent state, including: user preference elicitation sessions, where users respond to a sequence of feedback-gathering interventions (e.g., ratings, comparisons, annotations) (Chen and Pu, 2012; Zhao et al., 2013); user *ad blindness* (i.e., the tendency to gradually ignore ads) (Hohnhold et al., 2015); and the long-term evolution of user satisfaction (Wilhelm et al., 2018; Mladenov et al., 2019). Many aspects of a user's latent state determine their disposition towards specific actions. For example, a user's level of frustration, trust, receptivity, and overall satisfaction, may affect their tendency toward accepting recommendations, providing feedback, or abandoning a session. Notably, such features are cumulatively impacted by the user's long-term history, which makes RL especially challenging due to difficult credit assignment, where the impact of any individual action is usually small and noisy. [1]

In this paper, we introduce *Dynamic Contextual Markov Decision Processes (DCMDPs)* to model such environment dynamics in a ***history-dependent*** *contextual* fashion. DCMDPs decompose the state space to include dynamic history-dependent contexts, where each context represents a different MDP, e.g., preferences of a human interacting with an agent, being affected by previous interactions. Particularly, we introduce a special class of *logistic DCMDPs*, in which context dynamics are determined by the aggregation of a set of feature vectors—functions of the immediate context, state and action—over time. This model is inspired by various psychological studies of human learning and conditioning; in particular, the Rescorla-Wagner (RW) model (Rescorla, 1972), a neuroscience model which describes the diminishing impact of repeated exposure to a stimulus due to historical conditioning. Critically, this structure allows us to develop tractable, UCB-style algorithms (Auer et al., 2008) for logistic DCMDPs that break the exponential dependence on history length in general DCMDPs.

Our contributions are as follows: (1) We introduce DCMDPs, a model that captures non-Markov context dynamics. (2) We introduce a subclass of DCMDPs for which state-action-context features are aggregated over time to determine context dynamics. We show how such problems can

[*]Equal contribution [1]Google Research [2]CREST, ENSAE. Correspondence to: Guy Tennenholtz <guytenn@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]A similar problem occurs in medical settings, where a patient's previous reactions to certain treatments could implicitly affect the physician's receptivity for treatment recommendations over long horizons. Another example includes human driver interventions in autonomous vehicles, where humans may take control of a vehicle for short periods of time.

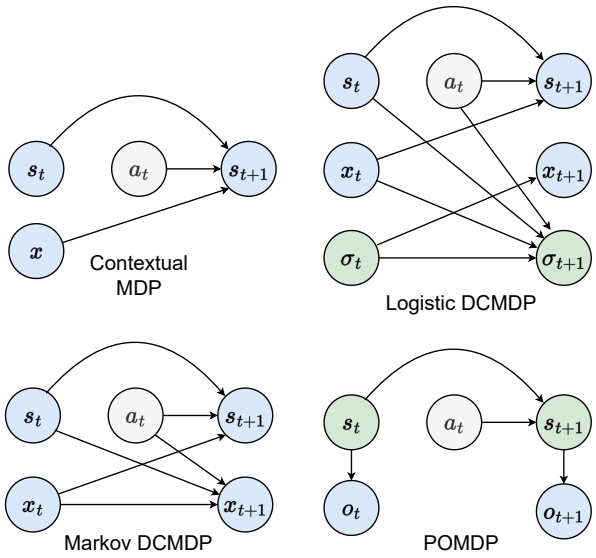

Figure 1: Causal diagrams comparing Contextual MDPs (Hallak et al., 2015), Markov DCMDPs (Section 2.1) and Logistic DCMDPs (Section 3). Logistic DCMDPs are history dependent, where $\sigma_t = \sum_{l=0}^{t-1} \alpha^{t-l-1} \boldsymbol{f}_l^*(s_l, a_l, x_l)$, and $\boldsymbol{f}_l^* : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto \mathbb{R}^M$ are *unknown*, non-stationary, vector valued feature mappings. Green circles represent unobserved variables.

be solved by devising sample efficient and computationally tractable solutions, for which we establish regret bounds. (3) Inspired by our theoretical results, we construct a practical algorithm, based on MuZero (Schrittwieser et al., 2020), and demonstrate its effectiveness on a recommendation system benchmark with long history-dependent contexts.

2. Dynamic Contextual MDPs

We begin by defining *Dynamic Contextual MDPs (DCMDPs)*, a general framework for modeling history-dependent contexts[2]. Let $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{X}$ be state, action, and context spaces, with cardinalities S, A, X, respectively. For any time $t \geq 1$, let $\mathcal{H}_t = \{(s_1, a_1, x_1, \ldots, s_t, a_{t-1}, x_{t-1})\}$ be the set of histories up to time t; and let $\mathcal{H} = \bigcup_t \mathcal{H}_t$. We denote $(s_0, a_0, x_0) = \emptyset$.

A DCMDP is given by the tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, r, P, H)$, where, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto [0, 1]$ is a reward function, $P : \mathcal{H} \times \mathcal{S} \times \mathcal{A} \mapsto \Delta_{\mathcal{S}}$ is a history-dependent transition function, and H is the horizon. DCMDP dynamics proceeds in discrete episodes $k = 1, 2, \ldots, K$. At the beginning of episode k, the agent is initialized at state s_1^k. At any time h, the agent is in state s_h^k, has observed a history $\tau_h^k = (s_1^k, a_1^k, x_1^k, \ldots, s_{h-1}^k, a_{h-1}^k, x_{h-1}^k) \in \mathcal{H}_h$, and selects an action $a_h^k \in \mathcal{A}$. Then, the next context x_h^k occurs with (history-dependent) probability $P(x_h^k | \tau_h^k)$, the agent receives reward $r(s_h^k, a_h^k, x_h^k)$, and the environment transitions to state s_{h+1}^k with probability $P_h(s_{h+1}^k | s_h^k, a_h^k, x_h^k)$.

A policy $\pi : \mathcal{S} \times \mathcal{H} \mapsto \Delta_{\mathcal{A}}$ maps states and histories to distributions over actions. The value of π at time h is defined as $V_h^\pi(s, \tau) = \mathbb{E}_\pi \Big[\sum_{t=h}^{H} r(s_t, a_t, x_t) \,\Big|\, s_h = s, \tau_h = \tau \Big]$, where $a_t \sim \pi(s_t, \tau_t)$, and $x_t \sim P(\cdot | \tau_t)$. An optimal policy π^* maximizes the value over all states and histories ; we denote its value function by V^*. We measure the performance of an RL agent by its *regret* – the difference between its value and that of an optimal policy: $\mathrm{Reg}(K) = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$.

Figure 1 depicts causal diagrams comparing general POMDPs to different types of DCMDPs, including three special cases: Contextual MDPs (Hallak et al., 2015) , Markov DCMDPs, and logistic DCMDPs (defined in the next two sections). DCMDPs are closely related to POMDPs, yet their causal structure allows us to devise more tractable solution (characterized by an aggregation function, as we'll see in Section 3) which can efficiently and tractably capture very long histories. In the next section, we describe a simple instance of DCMDPs, for which contexts are Markov, and show that standard MDP solutions can be applied. Then, in Section 3, we describe a more general DCMDP model, which uses aggregated features to represent histories, for which we provide sample efficient solutions and strong regret guarantees.

2.1. Markov DCMDPs

As a warm-up, we consider a simple version of DCMDPs in which context distributions are Markov w.r.t. the state and previous context. Specifically, we define a *Markov DCMDP* as a DCMDP which satisfies for all $h \in [H]$, $\tau_h = (x_1, s_1, a_1, \ldots, x_{h-1}, s_{h-1}, a_{h-1}) \in \mathcal{H}_h$ $P(x_h | \tau_h) = P(x_h | s_{h-1}, a_{h-1}, x_{h-1})$. A Markov DCMDP $\mathcal{M} = (\mathcal{X}, \mathcal{S}, \mathcal{A}, r, P, H)$ can be reduced to an MDP by augmenting the state space to include the context. To see this, we define the augmented MDP $\overline{\mathcal{M}} = (\bar{\mathcal{S}}, \mathcal{A}, \bar{r}, \bar{P}, H)$, where $\bar{\mathcal{S}} = \mathcal{S} \times \mathcal{X}$ and $\bar{r}(\bar{s}_t, a_t) = r(s_t, a_t, x_t)$, $\bar{P}(\bar{s}_{t+1} | \bar{s}_t, a_t) = P(s_{t+1} | s_t, a_t, x_t) P(x_{t+1} | s_t, a_t, x_t)$. As a consequence, the Markov DCMDP $\mathcal{M}$ and the MDP $\overline{\mathcal{M}}$ have the "same" optimal policy and value, and $\mathcal{M}$ can be solved using standard RL methods, e.g., using UCBVI (Azar et al., 2017) one can obtain $\mathrm{Reg}(K) \leq \tilde{\mathcal{O}}\Big(\sqrt{H^3 SAXK}\Big)$. Markov DCMDPs also generalize contextual MDPs in an especially simple way; but they fail to capture the history dependence of contexts embodied by general DCMDPs. In the next section, we turn to a special case of DCMDPs that does so, but also admits tractable solution methods.

[2]The term "context", as opposed to "state", differentiates between the Markov part of the state and the history dependent part of the state. Additionally, contexts often quantify characteristics of the environment (e.g., types of humans-in-the-loop), which can evolve in a distinct fashion, in contrast to the rest of the state.

3. Logistic DCMDPs

We introduce a general class of DCMDPs, called *logistic DCMDPs*, where history dependence is structured using an aggregation of state-action-context-dependent features. Unlike Markov DCMDPs, logistic DCMDPs allow for context transitions to depend on history.

We define the softmax function $z_i : \mathbb{R}^M \mapsto [0,1]$, with temperature $\eta > 0$ as

$$z_i(\boldsymbol{u}) = \frac{\exp(\eta u_i)}{1 + \sum_{m=1}^{M} \exp(\eta u_m)} \tag{1}$$

for $i \in [M]$, $\boldsymbol{u} \in \mathbb{R}^M$, and $z_{M+1}(\boldsymbol{u}) = 1 - \sum_{i=1}^{M} z_i(\boldsymbol{u})$.

Definition 3.1 (Logistic DCMDP)**.** A *logistic DCMDP* with latent feature maps $\left\{\boldsymbol{f}_h^* : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto \mathbb{R}^M\right\}_{h=0}^{H-1}$ is a DCMDP with context space $\mathcal{X} = \left\{x^{(i)}\right\}_{i=1}^{M+1}$, which satisfies, for all $h \in [H]$, $\tau_h = (s_1, a_1, x_1, \ldots, s_{h-1}, a_{h-1}, x_{h-1}) \in \mathcal{H}_h$, and $i \in [M+1]$:

$$P_{\boldsymbol{f}^*}(x_h^{(i)} | \tau_h) = z_i\left(\sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}_t^*(s_t, a_t, x_t))\right),$$

where $\alpha \in [0,1]$ is a *history discount factor*.

Note that the latent functions $\boldsymbol{f}_h^*$ are vector-valued and *unknown*. In a recommender system, $\boldsymbol{f}_h^*$ may represent a user's unknown degree of trust in the system, or the effect of a sequence of recommendations on their satisfaction. The discount α allows for immediate effects to diminish over time (if less than 1).

A logistic DCMDP is denoted by $(\mathcal{X}, \mathcal{S}, \mathcal{A}, r, P, H, \boldsymbol{f}^*, \alpha)$. We assume $\boldsymbol{f}^*$ is ℓ_2-bounded with $\sqrt{\sum f_{h,i}^{*2}(s,a,x)} \leq L$, and we denote

$$\mathcal{F} = \{\boldsymbol{f} : |f_{h,i}(s,a,x)| \leq b_{h,i}(s,a,x)\} \tag{2}$$

the (rectangular) set where $b_{h,i}(s,a,x)$ are upper bounds on $\boldsymbol{f}^*$. Throughout our analysis we denote the effective history horizon $H_\alpha = \frac{\alpha^{2H}-1}{\alpha-1}$, and without loss of generality scale transitions in z_i (Equation (1)) with temperature $\eta = H_\alpha^{-1}$.[3] For clarity, we write $r(s,a,x^{(i)}) = r_i(s,a)$, $P(s'|s,a,x^{(i)}) = P_i(s'|s,a)$, and $\boldsymbol{r}(s,a) = (r_1(s,a), \ldots, r_{M+1}(s,a))^T$, $\boldsymbol{P}(s'|s,a) = (P_1(s'|s,a), \ldots, P_{M+1}(s'|s,a))^T$. We also denote by $n_h^k(s,a,x)$ the number of visits to s,a,x at time step h of episode $k-1$.

Next, we define a sufficient statistic for logistic DCMDPs that will prove valuable in our solution methods that follow.

[3]We set $\eta = H_\alpha^{-1/2}$ for convenience. Different choices of η are equivalent to varying the bounds on $\mathcal{F}$ in Equation (2).

Definition 3.2 (Sufficient Statistic)**.** Given a logistic DCMDP with feature maps $\boldsymbol{f}$, define $\boldsymbol{\sigma} : \mathcal{H} \mapsto R^M$ as $\boldsymbol{\sigma}(\tau_h; \boldsymbol{f}) := \sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}_t(s_t, a_t, x_t)$, and the set of sufficient statistics by $\boldsymbol{\Sigma}(\boldsymbol{f}) := \{\boldsymbol{\sigma}(\tau; \boldsymbol{f})\}_{\tau \in \mathcal{H}}$.

In Appendix B.1, we prove that $\boldsymbol{\sigma}(\tau_h; \boldsymbol{f})$ is a sufficient statistic of the history for purposes of computing the optimal policy at time h. We do so by defining an equivalent MDP with state space $\mathcal{S} \times \boldsymbol{\Sigma}(\boldsymbol{f})$ with well-defined dynamics and reward, and an equivalent optimal policy, which achieves the same optimal value.

Finally, similar to previous work on logistic and multinomial bandits (Abeille et al., 2021; Amani and Thrampoulidis, 2021), we define a problem-dependent constant for logistic DCMDPs which plays a key role in characterizing the behavior of $M \geq 1$ mulitnomial logit bandit algorithms. For $\boldsymbol{x} \in \mathbb{R}^{M+1}$ and $\tau \in \mathcal{H}$, let $\boldsymbol{z}(\boldsymbol{x}) = (z_0(\boldsymbol{x}), \ldots, z_{M+1}(\boldsymbol{x}))^T$, $\boldsymbol{A}(\tau; \boldsymbol{f}) = \mathrm{diag}(\boldsymbol{z}(\boldsymbol{\sigma}(\tau; \boldsymbol{f}))) - \boldsymbol{z}(\boldsymbol{\sigma}(\tau; \boldsymbol{f}))\boldsymbol{z}(\boldsymbol{\sigma}(\tau; \boldsymbol{f}))^T$, and $1/\kappa = \inf_{\tau \in \mathcal{H}} \lambda_{\min}\{\boldsymbol{A}(\tau; \boldsymbol{f}^*)\}$. Informally, κ is related to saturation of the softmax z_i. For logistic DCMDPs, it is related to a worst-case context distribution w.r.t. $\boldsymbol{f}^*$ and $\tau \in \mathcal{H}$. We refer to Abeille et al. (2021); Amani and Thrampoulidis (2021) for details, as well as lower bounds using this constant in logistic bandits.

The Rescorla-Wagner Model in Recommenders. Before continuing to provide sample efficient methods for solving logistic DCMDPs, we turn to motivate the aggregated model of history through the lens of the Rescola-Wagner (RW) model (Rescorla, 1972) in a recommendation setting.

Logistic DCMDPs generate context transitions based on the sum of specific features of prior states, actions, and contexts, as captured by $\boldsymbol{f}^*$, with backward discounting to diminish the effect of past features or experiences, as captured by α. Such a model can be used to capture a (very simple) RW formulation of user behavior in an interactive recommender system. Let $I = \{i_1, \ldots, i_n\}$ be a set of items. A user may like, dislike, or be unfamiliar with any of these items, represented by $u \in \{1, 0, -1\}^n$. Let g_t be the user's (latent) current degree of satisfaction or engagement with the system. At each time t, the system asks the user for their disposition (e.g., rating) of an item $i_t \in I$. The user decides to answer the question with probability $z_1(g_t)$ (Equation (1)), which is strictly increasing with higher degrees of engagement level. The engagement level then evolves as $g_{t+1} = \alpha g_t + \beta u_{i_t}$, where $\alpha \in [0,1]$, and β is a user-specific sensitivity factor. This model gives rise to a logistic DCMDP, whose solution gives the optimal recommender system policy. Specifically, actions $a_t := i_t \in I$ are the questions asked by the system, $f^*(s_t, a_t, x_t) = \beta u_{a_t}$ depends only on a_t, user engagement is $g_h = \sum_{t=0}^{h-1} \alpha^{h-t-1} f^*(s_t, a_t, x_t) = \sum_{t=0}^{h-1} \alpha^{h-t-1} \beta u_{i_t}$, x_t is the decision whether to answer, and s_t is the observation of the answer.

Algorithm 1 LDC-UCB

1: **for** $k = 1, \ldots, K$ **do**
2: $\quad \bar{r}^k_{i,h}(s,a) = \hat{r}^k_{i,h}(s,a) + b^k_{i,h}(s,a), \forall i, h, s, a$
3: $\quad \bar{\pi}^k \leftarrow$ Optimistic Planner$\big(\bar{\mathcal{M}}_k(\delta)\big)$ // Eq. 6
4: $\quad$ Rollout a trajectory by acting $\bar{\pi}^k$
5: $\quad \hat{\boldsymbol{f}}^k \in \arg\max_{\boldsymbol{f} \in \mathcal{C}_k(\delta)} \mathcal{L}^k_\lambda(\boldsymbol{f})$ // Eq. 3
6: $\quad$ Update $\hat{P}^{k+1}_i(s,a), \hat{r}^{k+1}_i(s,a), n^{k+1}(s,a,x)$ over rollout trajectory
7: **end for**

4. Optimistic Methods for Logistic DCMDPs

Logistic DCMDPs' aggregation of features allow us to obtain sample efficient and computationally tractable solutions; namely, solutions which do not depend exponentially on history. In this section, we describe an optimistic algorithm for solving logistic DCMDPs and provide regret bounds. We focus on theoretical motivations here, and address computational tractability in the next section.

We first develop *Logistic Dynamic Context Upper Confidence Bound (LDC-UCB)*, a general RL method for logistic DCMDPs with unknown latent features (see Algorithm 1). At each episode k, LDC-UCB uses estimates of rewards $\hat{r}^k_{x,h}(s,a) = \frac{\sum_{k'=1}^k \mathbb{1}\{x^{k'}_h = x, s^{k'}_h = s, a^{k'}_h = a\} r^{k'}_h}{n^k_h(s,a,x)}$, transitions $\hat{P}^k_{x,h}(s'|s,a) = \frac{\sum_{k'=1}^k \mathbb{1}\{x^{k'}_h = x, s^{k'}_h = s, a^{k'}_h = a, s^{k'}_{h+1} = s'\}}{n^k_h(s,a,x)}$, and a projected estimate of $\hat{\boldsymbol{f}}$, calculated by maximizing the regularized log likelihood:

$$\mathcal{L}^k_\lambda(\boldsymbol{f}) = \sum_{k'=1}^{k} \sum_{h=1}^{H-1} \sum_{i=1}^{M+1} \mathbb{1}\{x^k_h = i\} \ell^k_{i,h}(\boldsymbol{f}) - \lambda \|\boldsymbol{f}\|_2^2, \tag{3}$$

where $\ell^k_{i,h}(\boldsymbol{f}) = \log\big(z_i(\boldsymbol{\sigma}(\tau^k_h; \boldsymbol{f}))\big)$, $\lambda > 0$, and recall that $\boldsymbol{\sigma}(\tau^k_h; \boldsymbol{f}) = \sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}_t(s^k_t, a^k_t, x^k_t)$.

We account for uncertainty in these estimates by incorporating optimism. For rewards and transitions, we add a bonus term $b^k_{i,h}$ (see Appendix C for explicit definitions) to the estimated reward (line 2). To incorporate optimism in the latent features $\hat{\boldsymbol{f}}$, we build on results from multinomial logistic bandits (Amani and Thrampoulidis, 2021). Specifically, we derive a confidence bound over $\hat{\boldsymbol{f}}$, for which with probability at least $1 - \delta$

$$\left\| g_k(\boldsymbol{f}^*) - g_k(\hat{\boldsymbol{f}}_t) \right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)} \leq \beta_k(\delta), \tag{4}$$

where $H_k(\boldsymbol{f}) = -\nabla^2_{\boldsymbol{f}} \mathcal{L}^k_\lambda(\boldsymbol{f})$, $g_k(\boldsymbol{f}) = -\nabla_{\boldsymbol{f}} \mathcal{L}^k_\lambda(\boldsymbol{f}) + D_k$, $\beta_k(\delta) = \frac{M^{5/2}SAH}{\sqrt{\lambda}}\big(\log\big(1 + \frac{k}{d\lambda}\big) + 2\log\big(\frac{2}{\delta}\big)\big) + \sqrt{\frac{\lambda}{4M}} + \sqrt{\lambda}L$. See Appendix G for exact expressions and a proof of the bound in Equation (4).

Next, we leverage the bound in Equation (4) to construct a feasible set of logistic DCMDPs. Specifically, we define the confidence set

$$\mathcal{C}_k(\delta) = \left\{ \boldsymbol{f} \in \mathcal{F} : \left\| g_k(\boldsymbol{f}) - g_k(\hat{\boldsymbol{f}}_t) \right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f})} \leq \beta_k(\delta) \right\}. \tag{5}$$

and the following set of logistic DCMDPs:

$$\bar{\mathcal{M}}_k(\delta) = \left\{ \Big(\mathcal{X}, \mathcal{S}, \mathcal{A}, \bar{r}, \hat{P}, H, \boldsymbol{f}, \alpha\Big) : \boldsymbol{f} \in \mathcal{C}_k(\delta) \right\}. \tag{6}$$

The optimistic policy $\bar{\pi}^k$ (line 3) is that with greatest value over all DCMDPs in $\bar{\mathcal{M}}_k(\delta)$, i.e., $\bar{\pi}^k$ corresponding to $\max_{\bar{m} \in \bar{\mathcal{M}}_k(\delta)} V^*(s_1; \bar{m})$.

Combining the above, we prove the following regret guarantee for Algorithm 1.

Theorem 4.1. *Let $\lambda = \Theta(\frac{HM^{2.5}SA}{L})$. With probability at least $1 - \delta$, the regret of Algorithm 1 is*

$$\mathrm{Reg}(K) \leq \tilde{\mathcal{O}}(\sqrt{H^6 M^{4.5} S^2 A^2 L^2 \kappa K}).$$

The proof of Theorem 4.1 can be found in Appendix C. We note that computing the optimistic policy over $\bar{\mathcal{M}}_k(\delta)$ (line 3) is computationally difficult, especially due the history dependence of π on the accumulated latent features $\sum_{t=1}^h \alpha^{h-t} \boldsymbol{f}(s_t, a_t, x_t)$. We address this challenge next.

5. Mitigating Computational Complexity

In this section we show how to relax LDC-UCB (Algorithm 1) to mitigate its high computational complexity. Importantly, we maintain regret guarantees similar to those of Theorem 4.1 while obtaining an exponential improvement to computational cost. We later use these results to construct a practical model-based algorithm in Section 6.

To address the computational challenges of Algorithm 1, we focus on two problems. The first involves the set $\mathcal{C}_k(\delta)$ (Equation (5) and line 5 of Algorithm 1) – where computation of the maximum likelihood constrained to set $\mathcal{C}_k(\delta)$ is intractable. To address this, we prove that the constraint on the maximum likelihood estimator can be replaced by a simpler, rectangular set, enabling efficient calculation of the projected maximum likelihood. The second challenge is the complexity of the optimistic planner (Equation (6) and line 3 of Algorithm 1). To overcome this, we develop a *local* confidence bound, for every state-action-context triple (s, a, x), and show it can be leveraged to design an optimistic planner, using a novel thresholding mechanism for optimism in logistic DCMDPs. Pseudocode for this tractable variant of LDC-UCB is presented in Algorithm 2.

5.1. A Tractable Estimator

We begin by constructing a tractable estimator for the latent feature maps $\boldsymbol{f}^*$ which, instead of projecting to the set $\mathcal{C}_k(\delta)$, solves for projected maximum likelihood on the rectangular set $\mathcal{F}$ (Equation (2)). Let $\gamma_k(\delta) = \left(2 + 2L\sqrt{MH} + \sqrt{2(1+L)}\right)\beta_k + \sqrt{\frac{2(1+L)HM}{\lambda}}\beta_k^2(\delta)$. We define the tractable maximum likelihood estimator $\hat{\boldsymbol{f}}_T^k \in \arg\max_{\boldsymbol{f}\in\mathcal{F}} \mathcal{L}_\lambda^k(\boldsymbol{f})$, and have the following bound.

Lemma 5.1. *With probability at least $1-\delta$, for all $k \in [K]$,*

$$\left\|\hat{\boldsymbol{f}}_T^k - \boldsymbol{f}^*\right\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \leq \gamma_k(\delta). \tag{7}$$

The proof (see Appendix G.2) uses a convex relaxation of the set $\mathcal{C}_k(\delta)$. Notice that the confidence region for $\hat{\boldsymbol{f}}_T^k$ is looser than that for $\hat{\boldsymbol{f}}^k$ (see Equation (4)), as $\beta_k(\delta) < \gamma_k(\delta)$. Nevertheless, its computation is tractable.

Next we can exploit the confidence bound in Equation (7) to construct a *local* bound for every state-action-context triple (s, a, x) using the number of visits to (s, a, x), i.e., $n_h^k(s, a, x)$. The following result uses structural properties of logistic DCMDPs to achieve a local bound for $\hat{\boldsymbol{f}}_T^k$. Its proof generalizes the local confidence bound in Tennenholtz et al. (2022), and can be found in Appendix G.3.

Lemma 5.2 (Local Estimation Confidence Bound)**.** *For any $\delta > 0$, with probability of at least $1-\delta$, for all $k \in [K], h \in [H], i \in [M]$ and $s, a, x \in \mathcal{S} \times \mathcal{A} \times \mathcal{X}$, it holds that*

$$\left|\left(\hat{\boldsymbol{f}}_T^k(s,a,x)\right)_{i,h} - \left(\boldsymbol{f}^*(s,a,x)\right)_{i,h}\right| \leq \frac{2\sqrt{\kappa}\gamma_k(\delta)}{\sqrt{n_h^k(s,a,x)+4\lambda}}.$$

Lemma 5.2 allows one to reason about the unknown features locally for any visited (s, a, x), a vital step toward an efficient optimistic planner. Indeed, as we see in the next section, the cost of planning in logistic DCMDPs can be reduced significantly using this bound.

5.2. Threshold Optimistic Planning

We now address the major computational challenge of Algorithm 1 – the complexity of optimistic planning (line 3 of Algorithm 1). To do this, we leverage the local bound in Lemma 5.2 and construct an optimistic planner using a novel threshold mechanism, as we describe next.

Recall the set of sufficient statistics $\boldsymbol{\Sigma}(\boldsymbol{f}) = \{\boldsymbol{\sigma}(\tau; \boldsymbol{f})\}_{\tau\in\mathcal{H}}$ (Definition 3.2), which is a finite, vector-valued set with cardinality $|\boldsymbol{\Sigma}(\boldsymbol{f})| = \mathcal{O}\left((SAMH)^{MH}\right)$, making planning in state space $S \times \boldsymbol{\Sigma}$ exponentially hard. Consequently, searching for the optimistic DCMDP in the space of feature maps satisfying $\boldsymbol{f} \in \mathcal{C}_k(\delta)$ (Equation (6)) requires searching over an exponentially large space.

We mitigate this problem exponentially, by leveraging the local confidence bound in Lemma 5.2. Let $\mathcal{B}_k(\delta) \subset \mathbb{R}^M \times \mathbb{R}^M$ be the rectangular cuboid of all candidate confidence intervals satisfying the bound in Lemma 5.2. That is, $\mathcal{B}_k(\delta)$ is the set of all M dimensional intervals $\left[\boldsymbol{l}_h^k(s,a,x), \boldsymbol{u}_h^k(s,a,x)\right]$, such that for all h, s, a, x, $\boldsymbol{f}_h^*(s,a,x) \in \left[\boldsymbol{l}_h^k(s,a,x), \boldsymbol{u}_h^k(s,a,x)\right]$, where, $\boldsymbol{u}_h^k(s,a,x), \boldsymbol{l}_h^k(s,a,x) = \hat{\boldsymbol{f}}_T^k \pm \left(\frac{2\sqrt{\kappa}\gamma_k(\delta)}{\sqrt{n_h^k(s,a,x^{(1)})+4\lambda}}, \ldots, \frac{2\sqrt{\kappa}\gamma_k(\delta)}{\sqrt{n_h^k(s,a,x^{(M)})+4\lambda}}\right)^T$. In what follows, we identify key characteristics of the optimistic value when optimized over $\mathcal{B}_k(\delta)$. Specifically, we show that an optimistic solution lies on the extreme points of $\mathcal{B}_k(\delta)$, but more importantly, at one of M *specific extreme points*. This limits the search required by optimistic planning to a much smaller set, which can be approximated effectively in practice.

Optimism in intervals. Instead of augmenting the state space with $\boldsymbol{\Sigma}(\boldsymbol{f})$, we use the set of confidence intervals defined by $\mathcal{B}_k(\delta)$. We denote by $\boldsymbol{C}_h^k : \boldsymbol{\Sigma}(\hat{\boldsymbol{f}}^k) \mapsto \mathbb{R}^M \times \mathbb{R}^M$ the confidence interval of the sufficient statistic $\boldsymbol{\sigma}(\tau_h^k, \hat{\boldsymbol{f}}^k)$. That is,

$$\begin{aligned}\boldsymbol{C}_h^k &= \boldsymbol{C}(\boldsymbol{\sigma}(\tau_h^k; \hat{\boldsymbol{f}}^k)) \\ &= \left[\sum_{t=0}^{h-1} \alpha^{h-t-1}\boldsymbol{l}_h^k(s_t^k, a_t^k, x_t^k), \sum_{t=0}^{h-1} \alpha^{h-t-1}\boldsymbol{u}_h^k(s_t^k, a_t^k, x_t^k)\right].\end{aligned}$$

We also denote by $\boldsymbol{\mathcal{I}}^k = \left\{\boldsymbol{C}(\boldsymbol{\sigma}(\tau, \hat{\boldsymbol{f}}_T^k))\right\}_{\tau\in\mathcal{H}}$ the set of possible confidence intervals over $\mathcal{B}_k(\delta)$ in episode k.

Next, we augment the state space $\mathcal{S}$ at every episode k by $\mathcal{S} \times \boldsymbol{\mathcal{I}}^k$, and define the augmented state-action optimistic value for context $i \in [M+1]$ and confidence interval $\boldsymbol{C}_h^k = \boldsymbol{C}(\boldsymbol{\sigma}(\tau_h^k, \boldsymbol{f}^k))$ at time step $h \in [H]$ by

$$\bar{Q}_i(s, a, \boldsymbol{C}_h^k)) = \bar{r}_i(s,a) + \mathbb{E}_{s'\sim\hat{P}_i(\cdot|s,a)}\left[\bar{V}_{h+1}(s', \boldsymbol{C}_{h+1}^k)\right],$$

where, with slight abuse of notation, we used $\boldsymbol{C}_{h+1}^k = \boldsymbol{C}\left(\boldsymbol{\sigma}\left(\tau_h^k \cup \{s, a, x^{(i)}\}, \hat{\boldsymbol{f}}_T^k\right)\right)$ to denote the next aggregated confidence interval. The optimistic value $\bar{V}_h$ is defined by maximizing over sufficient statistics in the confidence set $\boldsymbol{C}_h^k$ and $a \in \mathcal{A}$. That is,

$$\bar{V}_h(s, \boldsymbol{C}_h^k) = \max_{a\in\mathcal{A}} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}_h^k} \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})Q_i(s, a, \boldsymbol{C}_h^k) \tag{8}$$

Indeed, $\bar{V}_h$ is an optimistic value, as shown by the following proposition. Its proof is provided in Appendix D.3.

Proposition 5.3 (Optimistic Value)**.** *Let $\bar{V}_h$ as defined in Equation* (8)*. Then, w.h.p. $\bar{V}_1(s_1^k, \boldsymbol{C}_1^k) \geq V_1^*(s_1^k)$.*

Algorithm 2 Tractable LDC-UCB

for $k = 1, \ldots, K$ **do**
  $\bar{r}^k_{i,h}(s,a) = \hat{r}^k_{i,h}(s,a) + b^k_{i,h}(s,a), \forall i, h, s, a$
  $\bar{\pi}^k \leftarrow$ `Optimistic DP`$\left(\bar{r}^k, \hat{P}^k, \mathcal{B}_k(\delta)\right)$ // Eq. 8
  Rollout a trajectory by acting $\bar{\pi}^k$
  $\hat{\boldsymbol{f}}^{k+1}_T \in \arg\max_{\boldsymbol{f} \in \mathcal{F}} \mathcal{L}^k_\lambda(\boldsymbol{f})$ // Eq. 3
  Update $\hat{P}^{k+1}_i(s,a), \hat{r}^{k+1}_i(s,a), n^{k+1}(s,a,x)$ over rollout trajectory
end for

Next, we turn to show that the maximization problem in Equation (8) can be solved efficiently, though $\boldsymbol{C}^k_h$ is an exponentially large set. Notice that the inner term $\sum_{i=0}^M z_i(\bar{\boldsymbol{\sigma}}) Q_i(s,a,\boldsymbol{C}^k_h)$ in Equation (8) is not convex. Still, our analysis shows that a solution to the inner maximization problem lies in the set of extreme points of $\boldsymbol{C}^k_h$. That said, these 2^M extreme points make exhaustive search intractable. Fortunately, we can also show that the optimal solution lies in a space of exactly M solutions – a linearly sized, tractable search space.

To this end, we define the threshold set, which we will use to construct the (linear) set of feasible extreme points.

Definition 5.4. For a rectangular cuboid defined by the interval $\boldsymbol{C} = [\boldsymbol{l}, \boldsymbol{u}] \subseteq \mathbb{R}^{M+1} \times \mathbb{R}^{M+1}$, vector $\boldsymbol{y} \in \mathbb{R}^{M+1}$ and real number $t \in \mathbb{R}$ we define $\boldsymbol{th}_t(\boldsymbol{y}, \boldsymbol{C}) \in \mathbb{R}^{M+1}$ by

$$[\boldsymbol{th}_t(\boldsymbol{y}, \boldsymbol{C})]_i = \begin{cases} l_i & y_i < t \\ u_i & \text{o.w.} \end{cases}$$

Definition 5.5. For a vector $\boldsymbol{Q} \in \mathbb{R}^{M+1}$, we define the threshold set $\mathcal{T}(\boldsymbol{Q}) = \left\{ \frac{Q_i + Q_{i+1}}{2} \right\}_{i=1}^M$.

We use these definitions to show that the optimal solution to Equation (8) lies in the threshold set of Q-values (see proof in Appendix F.1).

Lemma 5.6 (Threshold Optimism)**.** *Let $\boldsymbol{Q} \in \mathbb{R}^{M+1}$. For any $\boldsymbol{x} \in \mathbb{R}^{M+1}$ such that $x_i = 0$ define $f(\boldsymbol{x}) = \sum_{i=1}^{M+1} z_i(\boldsymbol{x}) Q_i$. Let $\boldsymbol{C} = [\boldsymbol{l}, \boldsymbol{u}] \subseteq \mathbb{R}^{M+1} \times \mathbb{R}^{M+1}$ and assume that $\boldsymbol{l} < \boldsymbol{u}$. Then, there exists $t \in \mathcal{T}(\boldsymbol{Q})$ such that $\boldsymbol{th}_t(\boldsymbol{Q}, \boldsymbol{C}) \in \arg\max_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x})$.*

We can now leverage Lemma 5.6 to solve the inner maximization in Equation (8). For notational convenience, we write $\bar{Q}_i = \bar{Q}_i(s,a,\boldsymbol{C}^k_h)$ and $\boldsymbol{Q} = (Q_1, \ldots, Q_{M+1})^T$. Applying Lemma 5.6, we get that

$$\max_{\bar{\boldsymbol{\sigma}} \in \boldsymbol{C}^k_h} \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}}) \bar{Q}_i = \max_{t \in \mathcal{T}(\boldsymbol{Q})} \sum_{i=1}^{M+1} z_i\Big(\boldsymbol{th}_t\Big(\boldsymbol{Q}, \boldsymbol{C}^k_h\Big)\Big) \bar{Q}_i. \tag{9}$$

As a result, the non-convex maximization problem in Equation (9) reduces the search space to M optimistic candidates.

5.3. Putting It All Together

Using Lemma 5.6 and particularly its derived corollary in Equation (9), we construct an optimistic planner, denoted by `Optimistic DP`, which plans via dynamic programming using Equation (9); we refer to Appendix F for an explicit formulation of the optimistic planner. Finally, using the tractable estimator $\hat{\boldsymbol{f}}^k_T$, and the threshold optimistic planner, we present a tractable variant of LDC-UCB in Algorithm 2, for which we have the following regret guarantee.

Theorem 5.7. *Let $\lambda = \Theta(\frac{HM^{2.5}SA}{L})$. With probability at least $1-\delta$, the regret of Algorithm 2 is*

$$R(K) \leq \tilde{\mathcal{O}}\Big(\sqrt{H^8 M^{6.5} S^2 A^2 L^4 \kappa K}\Big).$$

The proof of the theorem can be found in Appendix D. As expected, the regret upper bound in Algorithm 2 is worse than that of Algorithm 1 by a factor of $\tilde{\mathcal{O}}(HML)$. This result is strongly affected by the looser bound for the tractable feature maps in Lemma 5.2. Nevertheless, the intractability of Algorithm 1 compared to the tractability of Algorithm 2 suggests this is a more-than-reasonable tradeoff. Moreover, our tractable variant of LDC-UCB gives rise to practical optimistic algorithms, as we demonstrate next.

6. DCZero

Motivated by our theoretical results, we present a practical model-based optimistic algorithm for solving DCMDPs. We build on MuZero (Schrittwieser et al., 2020), a recent model-based algorithm which constructs a model in latent space and acts using Monte Carlo Tree Search (MCTS, Coulom (2007)). MuZero uses representation, transition, and prediction networks for training and acting. The representation network first embeds observations in a latent space, after which planning takes place using the transition and prediction networks through a variant of MCTS. Importantly, instead of predicting the next state (e.g., using world models (Hafner et al., 2023)), MuZero trains its latent space by predicting three quantities—the reward, value, and current policy—by rolling out trajectories in latent space (see Schrittwieser et al. (2020) for further details).

We develop DCZero, an algorithm based on MuZero for DCMDPs (see Algorithm 3). Like MuZero, DCZero uses representation, transition, and prediction networks to learn and act in the environment. In contrast to MuZero, DCZero trains an additional ensemble of networks to estimate the unknown features $\boldsymbol{f}^*$ using cross-entropy. Estimated quantities of the ensemble are used to construct confidence intervals for the sufficient statistics, which are used to augment the state. DCZero uses $M+1$ transition networks (one for each context), and predicts $M+1$ reward functions. To incorporate optimism, the value function is trained optimistically using the thresholding technique in the previous

section, where rewards for unseen actions are sampled from the trained reward models r_i and next states are sampled from the trained models P_i.

Movie Recommendation Environment. To evaluate the effectiveness of DCZero, we develop a movie recommendation environment based on the MovieLens dataset (Harper and Konstan, 2015). Users and items are represented in embedding space computed using SVD of the MovieLens ratings matrix. Each of n users is assigned a set of M possible user embeddings; i.e., each user $u \in \{u^{(i)}\}_{i=1}^{n}$ is assigned a set of preference vectors $\boldsymbol{x} = \{\boldsymbol{x}^{(j)}\}_{j=1}^{M+1}, \boldsymbol{x}^{(j)} \in \mathbb{R}^d$. Intuitively, these vectors reflect distinct user preferences corresponding to some aspect of the user's latent state (e.g., mood or current interest (Cen et al., 2020); location, companions, or activity; level of trust or satisfaction with the system) and hence influence u's behavior.

The recommendation agent interacting with a user selects an item x from a random set of A movies, $\{\boldsymbol{v}^{(a)}\}_{a=1}^{A}, \boldsymbol{v}^{(a)} \in \mathbb{R}^d$, and recommends it. The user context then evolves according to some history-dependent dynamics represented by a logistic DCMDP. Specifically, we assume unknown latent features $\boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{v})$ with the user's aggregated features (at time $h \in [H]$, episode k) being: $\boldsymbol{\sigma}_{k,h} = \sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}^*(\boldsymbol{x}_k^{(j_t)}, \boldsymbol{v}^{(a_t)})$. The agent recommends movie $\boldsymbol{v}^{(a)}$ to the user, while the user preference vector is sampled as $\boldsymbol{x}_k^{j_h} \sim z(\boldsymbol{\sigma}_{k,h})$. The agent then receives a reward $r_j(\boldsymbol{x}, a) = (\boldsymbol{x}_k^{(j_h)})^T \boldsymbol{\Sigma} \boldsymbol{v}^{(a)}$ reflecting the user's (current) preference for the movie, and the user's latent state transitions given the unknown function $\boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{v})$ and discount α; that is, $\boldsymbol{\sigma}_{k,h+1} = \alpha \boldsymbol{\sigma}_{k,h} + \boldsymbol{f}^*(\boldsymbol{x}_k^{(j_h)}, \boldsymbol{v}^{(a)})$.

We test our methods in two variants of this environment. In the first, "AttractionEnv", user latent features $\boldsymbol{f}^*$ are correlated with the user's degree of preference for the recommended movie:

$$\boldsymbol{f}^*(\boldsymbol{x}^{(j)}, \boldsymbol{v}) = \mu\big((\boldsymbol{x}^{(j)})^T \boldsymbol{\Sigma} \boldsymbol{v}\big), \qquad \text{(Attraction)}$$

where μ is a component-wise monotonically increasing function. AttractionEnv reflects users with a tendency to desire content similar to those they most recently consumed. This may reflect the positive influence of exposure to new types of content, increased familiarity increasing preference, or content domains (such as music) where some mild consistency of experience is preferred to jarring shifts in style or genre. The second environment, "NoveltyEnv", reflects a contrasting dynamics in which user latent features evolve such that $\boldsymbol{f}^*$ is anti-correlated with the user's preference for the recommended movie:

$$\boldsymbol{f}_i^*(\boldsymbol{x}^{(j)}, \boldsymbol{v}) = \begin{cases} -\mu\big((\boldsymbol{x}^{(j)})^T \boldsymbol{\Sigma} \boldsymbol{v}\big) & , j = i \\ \mu\big((\boldsymbol{x}^{(j)})^T \boldsymbol{\Sigma} \boldsymbol{v}\big) & , \text{o.w.} \end{cases} \qquad \text{(Novelty)}$$

Algorithm 3 DCZero

require: Size of ensemble B
init: Replay buffer $\mathcal{R} \leftarrow \emptyset$
for $k = 1, 2, \ldots$ **do**
  Train bootstrap ensemble of B feature maps $\left\{\hat{\boldsymbol{f}}_{\boldsymbol{\theta}_b} : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto \mathbb{R}^M\right\}_{b=1}^{B}$ over $\mathcal{R}$ using cross-entropy loss.
  Augment state s with aggregated feature confidence $\boldsymbol{C}_h \leftarrow \sum_{t=0}^{h-1} \alpha^{h-t-1} \text{std}_b\left(\left\{\hat{\boldsymbol{f}}_{\boldsymbol{\theta}_b}(s_t, a_t, x_t)\right\}_{b=1}^{B}\right)$.
  Train threshold optimistic value estimator over $\mathcal{R}$

$$\bar{Q}_{i,\boldsymbol{\psi}}(s, a, \boldsymbol{C}_h)) = \hat{r}_i(s, a) + \gamma \mathbb{E}_{s' \sim \hat{P}_i(s,a)} \bar{V}_{\boldsymbol{\psi}}(s', \boldsymbol{C}_{h+1}),$$

$$\bar{V}_{\boldsymbol{\psi}}(s, \boldsymbol{C}_h) = \max_{t \in \mathcal{T}(\boldsymbol{Q}_{\boldsymbol{\psi}})} \sum_{i=1}^{M+1} z_i\Big(\boldsymbol{th}_t\Big(\boldsymbol{Q}_{\boldsymbol{\psi}}, \boldsymbol{C}_h^k\Big)\Big) \bar{Q}_{i,\boldsymbol{\psi}}^{\pi_{\boldsymbol{\phi}}(s)}.$$

  Act and train representation network, M transition networks (for each $\hat{P}_i$), and M prediction networks (for each $\hat{r}_i$) using `MuZero-ALG` with the optimistic value $\bar{V}_{\boldsymbol{\psi}}$ in MCTS. Return replay buffer $\mathcal{R}$.
end for

As a result, movies that previously appealed to the user become less preferred, reflecting a desire for novelty over short time periods.

Experiments. All experiments used a horizon of $H = 300$, $M = 6$ user classes, $A = 6$ slate items (changing every reset), and a user embedding dimension of $d = 20$. We used default parameters for MuZero and applied the same parameters to DCZero. We compared DCZero and MuZero on the AttractionEnv and NoveltyEnv environments. We also tested a history-dependent variant of MuZero, which uses the sequence of past movies and contexts to densely represent history. More specifically, Hist-MuZero uses a stack of 30 previous observations as its state. We implemented both MLP and Transformer-based model architectures, but present results for the Transformer, as both had similar performance.

Figure 2 shows these comparisons. The plots compare the return of DCZero with the two baselines on AttractionEnv and NoveltyEnv with $\alpha = 0.99$; we also vary the values of α on the AttractionEnv. We see that DCZero is able to outperform both baselines, with significant increases in performance for larger values of α (i.e., longer history dependence). This suggests that DCZero can be especially beneficial in problems that exhibit long history dependence. Interestingly, we note that using a dense history-dependent Transformer hurts performance, except for very small values of α (indeed, only for $\alpha = 0.1$ does the sequence model outperform the other methods).

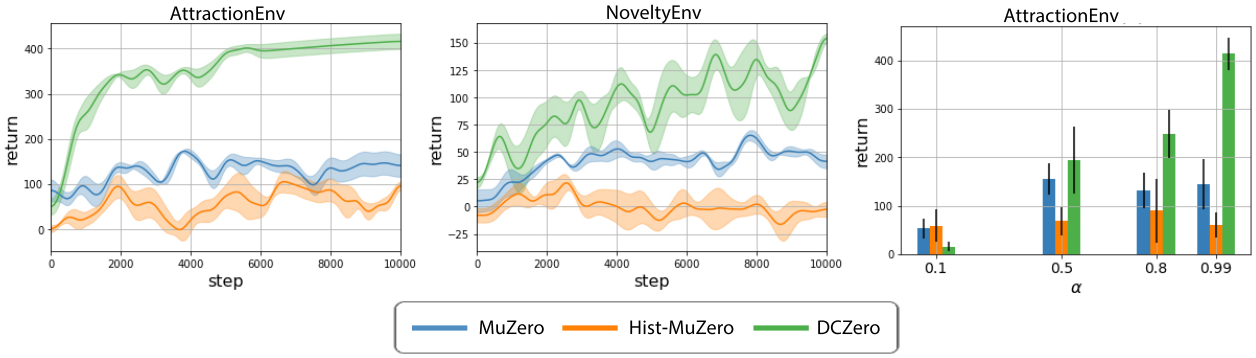

Figure 2: Plots comparing MuZero, Hist-Muzero, and DCZero on the AttractionEnv(left) and NoveltyEnv (middle). We also compare results for different values of α (right). All experiments show mean scores with 95% confidence intervals.

7. Related Work

Contextual MDPs (Hallak et al., 2015) have proven useful in a numerous studies (Jiang et al., 2017; Zintgraf et al., 2019; Kwon et al., 2021). Contexts are sampled once and are fixed throughout the episode. DCMDPs can be seen as a generalization of contextual MDPs, where contexts can change over time in a realistic, history-dependent fashion. Other forms of DCMDPs, are interesting directions for future work, including DCMDPs for which contexts change slowly in time. In Chen et al. (2022) a latent context variable changes abruptly at discrete points in time. Our logistic DCMDP considers history-dependent dynamics of contexts, which can depend on previous states and actions. Moreover, our model can capture smoother behavior which changes very slowly over time (over long histories). Finally, in contrast to Chen et al. (2022), our work provides theoretical guarantees, showing statistical and computational efficiency of our approach. In Mao et al. (2018), a non-stationary contextual environment is considered, yet the contexts are not allowed to depend on previous states and actions. Ren et al. (2022) propose a Bayesian approach for learning contextual MDPs for which contexts can change dynamically. Nevertheless, their model assumes dynamics that are not state-action dependent, and not history dependent.

Partially observable MDPs are widely studied (Papadimitriou and Tsitsiklis, 1987; Vlassis et al., 2012; Krishnamurthy et al., 2016; Tennenholtz et al., 2020; Xiong et al., 2022). As POMDPs are inherently history dependent, recent work has identified models and assumptions for which sample-efficient algorithms can be derived (Xiong et al., 2022; Liu et al., 2022a;b). Nevertheless, such solutions are often computationally intractable, impeding their practical implementation. With DCMPDs, we focus on specific forms of history-dependence, and show them to be computationally tractable, as well as effectively deployable.

Tennenholtz et al. (2022) define TerMDPs, a framework which models exogenous, non-Markovian termination in the environment. Once terminated, the agent stops acting and accrues no further rewards. TerMDPs capture various scenarios in which exogenous actors disengage with the agent (e.g., passengers in autonomous vehicles or users abandoning a recommender), and can be shown to be a special case of logistic DCMDPs (see Appendix B.2). As such, logistic DCMDPs support reasoning about optimizing more general contextual behavior, including: those involving notions of trust (e.g., where users become more or less receptive to agent recommendations); situations where humans override an agent for short periods; and modeling the effects of user satisfaction, moods, etc.

8. Discussion and Future Work

In this work we presented DCMDPs, and logistic DCMDPs in particular—a general history-dependent contextual framework which admits sample and computationally efficient solutions. The aggregation structure of logistic DCMDPs gives rise to efficient estimation of the unknown feature maps. We provided regret guarantees and developed a tractable realization of LDC-UCB using a computational estimator and a novel planning procedure. Finally, we tested DCZero, a model-based implementation of LDC-UCB, demonstrating its efficacy on a recommendation benchmark.

While logistic DCMDPs assume linear aggregations of past features, other variants with more complex parametric function classes over history are possible. Nevertheless, such complex function classes often require sample-inefficient techniques, suggesting that logistic DCMDPs may be especially well-suited to capturing extended, long history dependence. In particular, they admit sample and computationally efficient solutions, which can be implemented in practice. As future work, a hybrid approach which considers combining dense models (such as Transformers) for short-history dependence, and aggregated models (such as logistic DCMDPs) for very long history dependence, may offer the “best of both worlds" in practice.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034255.

Nadav Merlis is partially supported by the Viterbi Fellowship, Technion.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Sanae Amani and Christos Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34: 2913–2924, 2021.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Fahiem Bacchus, Craig Boutilier, and Adam Grove. Rewarding behaviors. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 1160–1167, Portland, OR, 1996.

KV Bhagwat and R Subramanian. Inequalities between means of positive operators. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 83, pages 393–401. Cambridge University Press, 1978.

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-20)*, pages 2942–2951, 2020.

Li Chen and Pearl Pu. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1):125–150, 2012.

Xiaoyu Chen, Xiangming Zhu, Yufeng Zheng, Pushi Zhang, Li Zhao, Wenxue Cheng, Peng CHENG, Yongqiang Xiong, Tao Qin, Jianyu Chen, et al. An adaptive deep rl method for non-stationary environments with piecewise stable context. In *Advances in Neural Information Processing Systems*, 2022.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2007.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. *arXiv preprint arXiv:2008.06036*, 2020.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In *International Conference on Machine Learning*, pages 2937–2947. PMLR, 2021.

Abdur R Fayjie, Sabir Hossain, Doukhi Oualid, and Deok-Jin Lee. Driverless car: Autonomous driving using deep reinforcement learning in urban environment. In *2018 15th international conference on ubiquitous robots (ur)*, pages 896–901. IEEE, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the Twenty-first ACM International Conference on Knowledge Discovery and Data Mining (KDD-15)*, pages 1849–1858, Sydney, 2015.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34:24523–24534, 2021.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022a.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022b.

Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations*, 2018.

Martin Mladenov, Ofer Meshi, Jayden Ooi, Dale Schuurmans, and Craig Boutilier. Advantage amplification in slowly evolving latent-state environments. In *Proceedings of the Twenty-eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 3165–3172, Macau, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Hang Ren, Aivar Sootla, Taher Jafferjee, Junxiao Shen, Jun Wang, and Haitham Bou Ammar. Reinforcement learning in presence of discrete markovian context evolution. In *International Conference on Learning Representations*, 2022.

Robert A Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Current research and theory*, pages 64–99, 1972.

Alessandro Ronca and Giuseppe De Giacomo. Efficient pac reinforcement learning in regular decision processes. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 2026–2032, Montreal, 2021.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

Guy Tennenholtz, Nadav Merlis, Lior Shani, Shie Mannor, Uri Shalit, Gal Chechik, Assaf Hallak, and Gal Dalal. Reinforcement learning with a terminator. In *Advances in Neural Information Processing Systems*, 2022.

Chen Tessler, Guy Tennenholtz, and Shie Mannor. Distributional policy optimization: An alternative approach for continuous control. *Advances in Neural Information Processing Systems*, 32, 2019.

Nikos Vlassis, Michael L Littman, and David Barber. On the computational complexity of stochastic controller optimization in pomdps. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–8, 2012.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on YouTube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM18)*, pages 2165–2173, Torino, Italy, 2018.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps. *Production and Operations Management*, 31(9):3491–3504, 2022.

Xiaoxue Zhao, Weinan Zhang, and Jun Wang. Interactive collaborative filtering. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 1411–1420, 2013.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

Appendix

The appendix is organized as follows. First, i Appendix A, we define additional notations that are used throughout this work. We then show that $\boldsymbol{\sigma}$ is indeed a sufficient statistic for calculating the optimal policy in a logistic DCMDP (Appendix B.1). Next, we provide regret guarantees for our algorithms for solving logistic DCMDPs: in Appendices C and D, we bound the regret of LDC-UCB (Theorem 4.1) and its tractable variant (Theorem 5.7), respectively. Finally, Appendices E to G contain technical lemmas which are crucial for deriving the above regret guarantees. Specifically, Appendix E is dedicated for optimism lemmas and decomposing the regret for logisitic DCMDPs; Appendix F deals with the threshold optimistic planning required for the tractable version of LDC-UCB; and Appendix G provides confidence sets for the regularized log likelihood procedure, following Abeille et al. (2021); Amani and Thrampoulidis (2021).

A. Additional Notations

In this section, we define additional notation that will be of use throughout the proof. We work with the natural filtration

$$\mathcal{F}_k = \sigma\left(\left\{\tau_{H+1}^{k'}\right\}_{k'\in[k]}, s_1^{k+1}\right) = \sigma\Big(\big\{(s_h^1, a_h^1, x_h^1, R_h^1)\big\}_{h=1}^H, \ldots, \big\{(s_h^k, a_h^k, x_h^1, R_h^k)\big\}_{h=1}^H, s_1^{k+1}\Big),$$

and notice that the policy π^k, which might depend on s_1^k, is $\mathcal{F}_{k-1}$-measurable. For brevity, for any episode $k \in [K]$ and time step $h \in [H]$, we define the probability distribution over the contexts by $z_h^k = \boldsymbol{z}(\boldsymbol{\sigma}(\tau_h^k; \boldsymbol{f}^*)$, with $z_{i,h}^k = z_i(\boldsymbol{\sigma}(\tau_h^k; \boldsymbol{f}^*)$ for any $i \in \mathcal{X}$.

With slight abuse of notation, we treat the latent features $\boldsymbol{f}$ as vectors in $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{X}\times[M]}$ instead of a mapping $\boldsymbol{f}_h^* : \mathcal{S}\times\mathcal{A}\times\mathcal{X} \mapsto \mathbb{R}^M, \forall h \in [H]$ and use the notations $\boldsymbol{f}_i(s,a,h,x) = \boldsymbol{f}(s,a,x,h,i)$. We also let $\boldsymbol{d}_h^k$ be the empirical discounted visitation vector at episode k up to time step h, i.e.,

$$\boldsymbol{d}_h^k(s,a,t,x) = H_\alpha^{-1/2}\alpha^{h-t-1}\mathbb{1}\big\{s_t^k = s, a_t^k = a, x_t^k = x\big\},$$

where $H_\alpha \triangleq \frac{1-\alpha^{2H}}{1-\alpha} \leq \min\Big\{H, \frac{1}{1-\alpha}\Big\}$ is a normalization factor describing the effective historical horizon. Then, one can write $\boldsymbol{\sigma}_i(\tau_h^k; \boldsymbol{f}) = \langle \boldsymbol{f}_i, \boldsymbol{d}_h^k(s,a,t,x)\rangle$. Notice that $\boldsymbol{d}_h^k$ is a vector containing zeros except for h elements with the values $\{H_\alpha^{-1/2}, H_\alpha^{-1/2}\alpha, \ldots, H_\alpha^{-1/2}\alpha^{h-1}\}$, where each value appears exactly once. We denote the set of all possible vectors of such form for any $h \in [H]$ by $\mathcal{D}$, and notice that for all $\boldsymbol{d} \in \mathcal{D}$,

$$\|\boldsymbol{d}\|_2^2 \leq \sum_{h=1}^{H-1} H_\alpha^{-1}\alpha^{2h} = \frac{1-\alpha^{2H}}{1-\alpha}H_\alpha^{-1} = 1.$$

Next, we define the following summation operators:

- For any fixed $h \in [H]$ and $i \in \mathcal{X}$, if $P_{i,h} : \mathcal{S}\times\mathcal{A} \mapsto \Delta_{\mathcal{S}}$ is a transition kernel and $V : \mathcal{S}\times\mathcal{H}_h \mapsto \mathbb{R}$ is a value function, the expected value is denoted by

  $$[P_{i,h}V](s,a,\tau_h) = P_{i,h}(\cdot|s,a)^T V(\cdot,\tau_h) = \sum_{s'\in\mathcal{S}} P_i(s_{h+1} = s'|s_h = s, a_h = a)V(s',\tau_h).$$

  and, in general, use similar notations for any transition kernel $P : \mathcal{Y} \mapsto \Delta_{\mathcal{S}}$ from arbitrary space $\mathcal{Y}$.

- We denote the vectorized version of $P_{i,h}V$ by

  $$P_hV = \big(P_{1,h}V, \ldots, P_{M,h}V\big)^T.$$

- If $Z : \mathcal{Y} \mapsto \Delta_{\mathcal{X}}$ is a mapping to the probability simplex over $\mathcal{X}$ and $U : \mathcal{Z}\times\mathcal{X} \mapsto \mathbb{R}$, where $\mathcal{Y}, \mathcal{W}$ are some arbitrary spaces, we let

  $$[ZU](y,w) = \sum_{i=1}^{M+1} Z(y,i)U(w,i),$$

  and in particular, we use $Z_h^{\boldsymbol{f}}(\tau_h) = \boldsymbol{z}(\boldsymbol{\sigma}(\tau_h; \boldsymbol{f})$ and $Z_h^k = \boldsymbol{z}_h^k$.

- Finally, given a transition kernel P and latent feature $\boldsymbol{f}$, we denote the transition operator over a value V by

  $$[T_h^{P,\boldsymbol{f}}V](s,a,\tau_h) = [Z_h^{\boldsymbol{f}}P_hV](s,a,\tau_h) = \sum_{s'\in\mathcal{S}}\sum_{i=1}^{M} P_i(s'|s,a)z_i(\boldsymbol{\sigma}(\tau_h; \boldsymbol{f})V(s',\tau_h).$$

  we similarly use the notation $T_h^{P,Z}$ when for general context distribution that are not necessarily by latent features $\boldsymbol{f}$.

B. Logistic DCMDPs

B.1. Sufficient Statistic

We prove that the $\boldsymbol{\sigma}$ is a sufficient statistic for calculating the optimal policy. We begin by defining an augmented MDP $\mathcal{M}_{\text{aug}} = (\mathcal{S}_{\text{aug}}, \mathcal{A}_{\text{aug}}, P_{\text{aug}}, r_{\text{aug}}, H)$, where $\mathcal{S}_{\text{aug}} = \mathcal{S} \times \boldsymbol{\Sigma}(\boldsymbol{f}^*)$ is the augmented state space, and $\mathcal{A}_{\text{aug}} = \mathcal{A}$ is the (unchanged) action space. The augmented transition function is defined for $s, \boldsymbol{\sigma} \in \mathcal{S} \times \boldsymbol{\Sigma}(\boldsymbol{f}^*), a \in \mathcal{A}, s', \boldsymbol{\sigma}' \in \mathcal{S} \times \boldsymbol{\Sigma}(\boldsymbol{f}^*)$

$$P_{\text{aug}}(s', \boldsymbol{\sigma}'|s, \boldsymbol{\sigma}, a) = \mathbb{1}\{\boldsymbol{\sigma}' = \alpha\boldsymbol{\sigma} + \boldsymbol{f}^*(s, a, x)\} \sum_{i=1}^{M+1} z_i(\boldsymbol{\sigma}) P_i(s'|s, a).$$

Finally, the augmented reward function r_{aug} satisfies

$$r_{\text{aug}}(s, \boldsymbol{\sigma}, a) = \sum_{i=1}^{M+1} z_i(\boldsymbol{\sigma}) r_i(s, a).$$

The augmented MDP $\mathcal{M}_{\text{aug}}$ is closely related to the logistic DCMDP $(\mathcal{X}, \mathcal{S}, \mathcal{A}, r, P, H, \boldsymbol{f}^*, \alpha)$. In fact, as we will show next, they both achieve the same optimal value. To see this, consider an MDP defined by the tuple $(\mathcal{S}_1 \times \mathcal{S}_2, \mathcal{A}, P, r, H)$, and let $\phi : \mathcal{S}_2 \mapsto D$, where D is some known domain. Define the following set of deterministic policies

$$\Pi_{\text{aug}} = \{\pi : \mathcal{S}_1 \times \mathcal{S}_2 \mapsto \mathcal{A} : \exists \eta : \mathcal{S}_1 \times D \mapsto [0, 1], \pi(s_1, s_2) = \eta(s_1, \phi(s_2)\}.$$

Define the augmented optimal value for some $s \in \mathcal{S}_1 \times \mathcal{S}_2$

$$V^*_{\text{aug},1}(s_1, s_2) = \max_{\pi \in \Pi_{\text{aug}}} \mathbb{E}\left[\sum_{t=1}^{H} r_t(s_t, a_t) \,\middle|\, s_1 = s_1, s_2 = s_2, a_t \sim \pi_t(s_1, s_2) \right].$$

We apply the following proposition using the decomposition $\mathcal{S}_1 = \mathcal{S}$, and $\mathcal{S}_2 = \mathcal{H}$ as the set of possible trajectories in the known logistic DCMDP, where

$$\phi(\tau_h) := \boldsymbol{\sigma}(\tau_h; \boldsymbol{f}) = \sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}_t(s_t, a_t, x_t).$$

Proposition B.1 (Tennenholtz et al. (2022)). *Let $\mathcal{M} = (\mathcal{S}_1 \times \mathcal{S}_2, \mathcal{A}, P, r, H)$. Assume for any $s_1, s_2 \in \mathcal{S}_1 \times \mathcal{S}_2$, $a \in \mathcal{A}$, $P(s_1', \phi(s_2')|s_1, s_2, a) = P(s_1', \phi(s_2')|s_1, \phi(s_2), a)$ and $r(s_1, s_2, a) = g(s_1, a)$, for some deterministic function $g : \mathcal{S}_1 \times \mathcal{A} \mapsto [0, 1]$. Then, for any $s_1, s_2 \in \mathcal{S}_1 \times \mathcal{S}_2$,*

$$V^*_{aug,1}(s_1, s_2) = V^*_1(s_1, s_2).$$

This concludes our claim, proving that σ is indeed sufficient, as playing any policy in Π_{aug} achieves the same value.

B.2. Relation to TerMDPs

A special case of logistic DCMDPs are TerMDPs (Tennenholtz et al., 2022), which model exogenous, non-Markov termination in the environment. When terminated, the agent stops interacting with the environment and cannot collect additional rewards. This setup describes various real-world scenarios, such as passengers in autonomous vehicles or users abandoning a recommender systems. To model a TerMDP as a logistic DCMDP we let $\mathcal{X} = \{0, 1\} = \{\text{term}, \text{no term}\}$, and define $r_i(s, a), P_i(s'|s, a)$ such that s_{term} is a sink state for which $r_i(s_{\text{term}}, a) = 0$. The reward in all other states is defined by $r_1(s, a)$. The transition probabilities are defined by $P_i(s'|s, a) = \begin{cases} s_{\text{term}}, & s = s_{\text{term}} \vee i = 0 \\ P_1(s'|s, a), & \text{o.w.} \end{cases}$.

TerMDPs use a cost functions $c_h(s_t, a_t)$ to define the probability of transitioning to the termination state, as $P(x_h = \text{term}|\tau_h) = z_0\left(\sum_{t=1}^{h-1} c_t(s_t, a_t)\right)$ – a special case of logistic DCMDPs with a two-class, context-independent feature map, and a choice of $\alpha = 1$. Indeed, this choice of parameters defines a TerMDP as proposed in Tennenholtz et al. (2022).

Logistic DCMDPs let us consider generalized notions of such models, for which classes can reflect notions of trust, where humans become less susceptible to following recommendations from an agent, through situations where humans override an agent for short periods, to modeling the effects of changing moods.

C. Regret Analysis of LDC-UCB

In this section, we prove the regret bounds of Theorem 4.1. We start by defining the good event, which holds uniformly for all episode with probability $1-\delta$. Then, we show that LDC-UCB is optimistic under the good event. Next, we decompose the regret to error terms of the reward, transition and latent features, and analyzing each of these terms result with the desired regret bounds.

We start by stating the bonuses which the algorithm uses:

$$b^{r,k}_{x,h}(s,a) = \min\left\{\sqrt{\frac{\log\frac{8SAMHK}{\delta}}{n^k_h(s,a,x)\vee 1}}, 1\right\}$$
$$b^{p,k}_{x,h}(s,a) = \min\left\{H\sqrt{\frac{4S\log\frac{8SAMHK}{\delta}}{n^k_h(s,a,x)\vee 1}}, 2H\right\}$$

C.1. Failure Events

We define the following failure events.

$$F^r_k = \left\{\exists s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, h\in[H]:\ |r_{x,h}(s,a) - \hat{r}^k_{x,h}(s,a)| > \min\left\{\sqrt{\frac{\log\frac{2SAMHK}{\delta'}}{n^k_h(s,a,x)\vee 1}}, 1\right\}\right\}$$
$$F^p_k = \left\{\exists s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, h\in[H]:\ \left\|P_{x,h}(\cdot\mid s,a) - \hat{P}^k_{x,h}(\cdot\mid s,a)\right\|_1 > \min\left\{\sqrt{\frac{4S\log\frac{2SAMHK}{\delta'}}{n^k_h(s,a,x)\vee 1}}, 2\right\}\right\}$$
$$F^n = \left\{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\frac{z^k_{i,h}}{\sqrt{n^k_h(s^k_h,a^k_h,i)\vee 1}}\,\middle|\, F_{k-1}\right] > 18H^2\log\left(\frac{1}{\delta'}\right) + 2HS(M+1)A + 4\sqrt{H^2S(M+1)AK}\right\}$$
$$F^{\boldsymbol{f},\text{global}}_k = \{\boldsymbol{f}^*\notin\mathcal{C}_k(\delta')\},$$

where the definition of $\mathcal{C}_k(\delta)$ can be found at Appendix G.

Then, we define the good event, where none of the aforementioned failure events ever occur, i.e.,

$$\mathcal{G} = \left(\cup_{k\in[K]}\bar{F}^r_k\right)\cap\left(\cup_{k\in[K]}\bar{F}^p_k\right)\cap\left(\cup_{k\in[K]}\bar{F}^{\boldsymbol{f},\text{global}}_k\right)\cap\bar{F}^n$$

Lemma C.1. *Letting $\delta' = \delta/4$, the event $\mathcal{G}$ holds with probability at least $1-\delta$.*

Proof. We show that the probability that the events do not hold for all $k\in[K]$ is smaller than $\delta' = \delta/4$.

- **Reward concentration.** First observe that both the empirical and real rewards are bounded in $[0,1]$, so if the minimizer in F^r_k is 1, the event never holds. Otherwise, for any fixed episode k, number of plays n, state s, action a, context x and timestep h, by Hoeffding's inequality, the estimation error is bounded w.p. $1-\delta'$ by $\sqrt{\frac{\log\frac{2}{\delta'}}{n}}$. Taking the union bound over all possible values of $k, n\geq 1, s, a, x$ and h, w.p. at least $1-\delta'$, for all $k\in[K], s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, h\in[H]$, the estimation error is bounded by

$$|r_{x,h}(s,a) - \hat{r}^k_{x,h}(s,a)| \leq \sqrt{\frac{\log\frac{2SAMHK^2}{\delta'}}{2n^k_h(s,a,x)\vee 1}} \leq \sqrt{\frac{\log\frac{2SAMHK}{\delta'}}{n^k_h(s,a,x)\vee 1}}.$$

  Finally, we remark that since $\delta'\leq 1/4$, the event F^r_k never holds when $n^k_h(s,a,x) = 0$ since the bound is larger than 1. In other words, $\Pr\left\{\cup_{k\in[K]}\bar{F}^r_k\right\}\leq\delta'$.

- **Transition concentration.** By the exact same arguments as the reward concentration, while replacing Hoeffding's inequality by the concentration of the L_1 error of a probability estimator (Weissman et al., 2003), we also get $\Pr\{\cup_{k\in[K]}\bar{F}_k^p\} \le \delta'$. Notice that the L_1 distance between any two probability distributions is bounded by 2, which justifies the minimization in the event.

- **Global feature estimation.** By Lemma G.5, we have that

$$\Pr\Big\{\cup_{k\in[K]}\bar{F}_k^{\boldsymbol{f},\text{global}}\Big\} \le \Pr\{\exists k \ge 1 : \boldsymbol{f}^* \notin \mathcal{C}_k(\delta')\} \le \delta'.$$

- **Expected counts concentration.** By Lemma E.7, we have that $\Pr\{\bar{F}^n\} \le \delta'$.

Fixing $\delta' = \delta/4$ and taking the union bound concludes the proof. □

C.2. Regret Analysis – Proof of Theorem 4.1

Theorem 4.1. *Let $\lambda = \Theta(\frac{HM^{2.5}SA}{L})$. With probability at least $1-\delta$, the regret of Algorithm 1 is*

$$\mathrm{Reg}(K) \le \tilde{\mathcal{O}}(\sqrt{H^6M^{4.5}S^2A^2L^2\kappa K}).$$

Proof. Under the good event, the conditions of the regret decomposition lemma (Lemma E.6) hold with $\bar{r}, \hat{P}, Z^{\bar{\boldsymbol{f}}_k}$ and $c = 4$ due to the truncated value iteration, truncated bonuses and value optimism lemma (Proposition C.5). Therefore, the regret can be decomposed in the following way

$$\begin{aligned}\mathrm{Reg}(K) \le &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\, z_{i,h}^k \left|\bar{r}_{i,h}^k(s_h^k,a_h^k) - r_{i,h}(s_h^k,a_h^k)\right| \,\middle|\, F_{k-1}\right]}_{(i)}\\ &+\underbrace{H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\, z_{i,h}^k \left\|\left(\hat{P}_h^k - P_h\right)(\cdot|s_h^k,a_h^k)\right\|_1 \,\middle|\, F_{k-1}\right]}_{(ii)}\\ &+\underbrace{5H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\left\|Z_h^{\bar{\boldsymbol{f}}_k} - Z_h^{\boldsymbol{f}^*}\right\|_1 \,\middle|\, F_{k-1}\right]}_{(iii)}.\end{aligned}$$

By plugging in Lemma C.2, Lemma C.3 and Lemma C.4, which bound terms (i),(ii) and (iii), respectively, we get,

$$\begin{aligned}\mathrm{Reg}(K) \le &\,\mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)\\ &+\mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)\\ &+\tilde{\mathcal{O}}(\sqrt{S^2A^2H^6M^{4.5}L^2\kappa K})\end{aligned}$$

Noticing that the last term is the dominant, we get

$$\mathrm{Reg}(K) \le \tilde{\mathcal{O}}(\sqrt{S^2A^2H^6M^{4.5}L^2\kappa K}),$$

which concludes the proof. □

Now, we prove the lemmas that bounds the three terms in the regret decomposition in Theorem 4.1.

Lemma C.2 (Reward Concentration)**.** *Under the good event, we have that:*

$$
\begin{aligned}
(i) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left|\bar{r}_{i,h}^k(s_h^k, a_h^k) - r_{i,h}(s_h^k, a_h^k)\right| \,\middle|\, F_{k-1} \,\right] \\
&\leq 2H\sqrt{S\log\frac{8SAMHK}{\delta}} \cdot \left(18H^2\log\left(\frac{4}{\delta}\right) + 2HS(M+1)A + 4\sqrt{H^2S(M+1)AK}\right) && \text{(Under } \mathcal{G}\text{)} \\
&= \mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left|\bar{r}_{i,h}^k(s_h^k, a_h^k) - r_{i,h}(s_h^k, a_h^k)\right| \,\middle|\, F_{k-1} \,\right] \\
&\qquad \leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left(b_{i,h}^{r,k}(s_h^k, a_h^k) + b_{i,h}^{p,k}(s_h^k, a_h^k) + \left|\hat{r}_{i,h}^k(s_h^k, a_h^k) - r_{i,h}(s_h^k, a_h^k)\right|\right) \,\middle|\, F_{k-1} \,\right] \\
&\qquad \leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left(2b_{i,h}^{r,k}(s_h^k, a_h^k) + b_{i,h}^{p,k}(s_h^k, a_h^k)\right) \,\middle|\, F_{k-1} \,\right] && \text{(Under } \mathcal{G}\text{)} \\
&\qquad \leq 2H\sqrt{S\log\frac{8SAMHK}{\delta}} \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, \frac{z_{i,h}^k}{\sqrt{n_h^k(s_h^k, a_h^k, i) \vee 1}} \,\middle|\, F_{k-1} \,\right] \\
&\qquad \leq 2H\sqrt{S\log\frac{8SAMHK}{\delta}} \cdot \left(18H^2\log\left(\frac{4}{\delta}\right) + 2HS(M+1)A + 4\sqrt{H^2S(M+1)AK}\right) && \text{(Under } \mathcal{G}\text{)} \\
&\qquad = \mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)
\end{aligned}
$$

□

Lemma C.3 (Transition Concentration)**.** *Under the good event, we have that:*

$$
\begin{aligned}
(ii) &= H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left\|\left(\hat{P}_h^k - P_h\right)(\cdot|s_h^k, a_h^k)\right\|_1 \,\middle|\, F_{k-1} \,\right] \\
&\leq \mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z_{i,h}^k \left\|\left(\hat{P}_h^k - P_h\right)(\cdot|s_h^k, a_h^k)\right\|_1 \,\middle|\, F_{k-1} \,\right] \\
&\qquad \leq H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M} \mathbb{E}\left[\, z_{i,h}^k \cdot \frac{1}{H} b_{i,h}^{p,k}(s_h^k, a_h^k) \,\middle|\, F_{k-1} \,\right] && \text{(Under } \mathcal{G}\text{)} \\
&\qquad \leq H\sqrt{4S\log\frac{8SAMHK}{\delta}} \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M} \mathbb{E}\left[\, \frac{z_{i,h}^k}{\sqrt{n_h^k(s_h^k, a_h^k, i) \vee 1}} \,\middle|\, F_{k-1} \,\right] \\
&\qquad \leq H\sqrt{4S\log\frac{8SAMHK}{\delta}} \cdot \left(18H^2\log\left(\frac{4}{\delta}\right) + 2HS(M+1)A + 4\sqrt{H^2S(M+1)AK}\right) && \text{(Under } \mathcal{G}\text{)} \\
&\qquad = \mathcal{O}\left(H^2S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right)
\end{aligned}
$$

□

Lemma C.4 (Latent Features Concentration)**.** *Under the good event, if $L = \Omega(1)$ and $\lambda = \Theta(\frac{SAHM^{2.5}}{L})$, we have that:*

$$(iii) = 5H \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\Big[\Big\| Z_h^{\bar{\boldsymbol{f}}_k} - Z_h^{\boldsymbol{f}^*} \Big\|_1 \Big| F_{k-1} \Big] \leq \tilde{\mathcal{O}}(\sqrt{S^2A^2H^6M^{4.5}L^2\kappa K})$$

Proof.

$$\begin{aligned}
5H \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\Big[\Big\| Z_h^{\bar{\boldsymbol{f}}_k} - Z_h^{\boldsymbol{f}^*} \Big\|_1 \Big| F_{k-1} \Big] & \\
&\leq 5H\sqrt{M+1} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\Big[\Big\| Z_h^{\bar{\boldsymbol{f}}_k} - Z_h^{\boldsymbol{f}^*} \Big\|_2 \Big| F_{k-1} \Big] \\
&\leq 10H\sqrt{(1+2L)(M+1)\kappa} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\Big[\beta_k(4\delta) \Big\| \boldsymbol{d}_h^k \Big\|_{\boldsymbol{V}_k^{-1}} \Big| F_{k-1} \Big] && \text{(Lemma G.8)} \\
&\leq 10H\beta_K(4\delta)\sqrt{(1+2L)(M+1)\kappa} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\Big[\Big\| \boldsymbol{d}_h^k \Big\|_{\boldsymbol{V}_k^{-1}} \Big| F_{k-1} \Big] \\
&\leq 10H\beta_K(4\delta)\sqrt{(1+2L)(M+1)\kappa} \frac{\sqrt{2KH^3MSA \log \frac{\kappa\lambda HSMA + k}{\kappa\lambda HSMA}}}{\max\{1, 1/\sqrt{\lambda}\}} && \text{(Corollary E.9)} \\
&= \mathcal{O}\left(\frac{\beta_K(4\delta)}{\max\{1, 1/\sqrt{\lambda}\}} \sqrt{H^5M^2SAL\kappa K \log \frac{\kappa\lambda HSMA + k}{\kappa\lambda HSMA}} \right) \\
&= \tilde{\mathcal{O}}(\sqrt{S^2A^2H^6M^{4.5}L^2\kappa K})
\end{aligned}$$

For both the lemma and the corollary, we remind that $\boldsymbol{d}_h^k \leq 1$. For the last relation, recall that

$$\beta_k(\delta) = \frac{M^{3/2}(M+1)SAH}{\sqrt{\lambda}} \left(\log\left(1 + \frac{k}{(M+1)SA\lambda}\right) + 2\log\left(\frac{2}{\delta}\right) \right) + \sqrt{\frac{\lambda}{4M}} + \sqrt{\lambda}L,$$

and assuming that $L = \Omega(1)$, we take $\lambda = \Theta(\frac{SAHM^{2.5}}{L})$, so $\beta_K(4\delta) = \tilde{\mathcal{O}}(\sqrt{SAHM^{2.5}L})$. □

C.3. Optimism in Logistic DCMDPs

In this section, we prove Proposition C.5, which allows us to apply the regret decomposition (Lemma E.6) necessary for proving Theorem 4.1.

We start by clearly stating the output value of the planning algorithm. For any $\boldsymbol{f} \in \mathcal{F}$, we define the truncated optimistic value under $\boldsymbol{f}$ as the solution to the following value iteration problem:

$$\begin{aligned}
&\bar{V}_{H+1}^{k,\boldsymbol{f}}(s, \tau_{H+1}) = 0, && \forall s \in \mathcal{S}, \tau_{H+1} \in \mathcal{H}_{H+1} \\
&\bar{V}_h^{k,\boldsymbol{f}}(s, \tau_h) = \min\Big\{ H, \max_a \Big\{ \Big[Z_h^{\boldsymbol{f}} \bar{r}_h^k \Big](s, a, \tau_h) + \Big[T_h^{\hat{P}_h^k, \boldsymbol{f}} \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a, \tau_h) \Big\} \Big\}, && \forall h \in [H], s \in \mathcal{S}, \tau_h \in \mathcal{H}_h.
\end{aligned}$$

Then, given an initial state s, we define the optimistic value of the DCMDP by $\bar{V}_1^k(s, \tau_h) = \max_{\boldsymbol{f} \in \mathcal{C}_k(\delta)} \bar{V}_1^{k,\boldsymbol{f}}(s, \tau_h)$. For this value, the following holds:

Proposition C.5. *Under the good event $\mathcal{G}$, for any $k \geq 1$ and any initial state $s \in \mathcal{S}$, it holds that $\bar{V}_1^k(s) \geq V_1^*(s)$.*

Proof. Assume that $\mathcal{G}$ holds, and let $\bar{V}_h^{k,\boldsymbol{f}}(s, \tau_h)$, $\bar{V}_1^k(s, \tau_h)$ as defined by the beginning of the section. In the proof, we will show that for any $k \in [K]$, $s \in \mathcal{S}$, $h \in [H]$ and $\tau_h \in \mathcal{H}_h$ and $\boldsymbol{f} \in \mathcal{F}$, it holds that $\bar{V}_h^{k,\boldsymbol{f}}(s, \tau_h) \geq V^{*,\boldsymbol{f}}(s, \tau_h)$. Since under

$\mathcal{G}$, we know that $\boldsymbol{f} \in \mathcal{C}_k(\delta)$, we then have that

$$\bar{V}_1^k(s) = \max_{\boldsymbol{f} \in \mathcal{C}_k(\delta)} \bar{V}_1^{k,\boldsymbol{f}}(s, \tau_h) \geq \max_{\boldsymbol{f} \in \mathcal{C}_k(\delta)} V_1^{*,\boldsymbol{f}}(s, \tau_h) \geq V_1^*(s),$$

which would conclude the prove. Throughout this proof, we assume w.l.o.g. that all optimistic values are smaller than H; otherwise, they will be truncated to H, which still always optimistic since the rewards are in $[0, 1]$ and the horizon is H.

We prove that $\bar{V}_h^{k,\boldsymbol{f}}(s, \tau_h) \geq V^{*,\boldsymbol{f}}(s, \tau_h)$ by backward-induction. First notice that the claim holds for $h = H$, since

$$\begin{aligned}
\bar{V}_H^{k,\boldsymbol{f}}(s, \tau_H; \boldsymbol{f}) - V_H^{*,\boldsymbol{f}}(s, \tau_H) &= \max_a \Big\{ \Big[Z_H^{\boldsymbol{f}} \bar{r}_H^k \Big](s, a, \tau_H) \Big\} - \max_a \Big\{ \Big[Z_H^{\boldsymbol{f}} r_H \Big](s, a, \tau_H) \Big\} \\
&\overset{(1)}{\geq} \Big[Z_H^{\boldsymbol{f}} \bar{r}_H^k \Big](s, a^*, \tau_H) - \Big[Z_H^{\boldsymbol{f}} r_H \Big](s, a^*, \tau_H) \\
&\qquad\qquad\qquad\qquad \text{(for } a^* \in \arg\max_a \Big\{ \Big[Z_H^{\boldsymbol{f}} r_H \Big](s, a, \tau_H) \Big\}) \\
&\geq \Big[Z_H^{\boldsymbol{f}} (\hat{r}_H^k + b_H^{r,k} - r) \Big](s, a^*, \tau_H) \\
&\geq 0 \qquad\qquad\qquad\qquad \text{(Lemma E.1)}
\end{aligned}$$

Now let $h \in [H-1]$ and assume that the claim holds for $h+1$. Then, for

$$a^* \in \arg\max_a \Big\{ \Big[Z_h^{\boldsymbol{f}} r_h \Big](s, a, \tau_h) + \Big[T_h V_{h+1}^{*,\boldsymbol{f}} \Big](s, a, \tau_h) \Big\},$$

we have

$$\begin{aligned}
&\bar{V}_h^{k,\boldsymbol{f}}(s, \tau_h) - V_h^{*,\boldsymbol{f}}(s, \tau_h) \\
&\quad = \max_a \Big\{ \Big[Z_h^{\boldsymbol{f}} \bar{r}_h^k \Big](s, a, \tau_h) + \Big[T_h^{\hat{P}_h^k, \boldsymbol{f}} \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a, \tau_h) \Big\} - \max_a \Big\{ \Big[Z_h^{\boldsymbol{f}} r_h \Big](s, a, \tau_h) + \Big[T_h V_{h+1}^{*,\boldsymbol{f}} \Big](s, a, \tau_h) \Big\} \\
&\quad \overset{(1)}{\geq} \Big[Z_h^{\boldsymbol{f}} \bar{r}_h^k \Big](s, a^*, \tau_h) + \Big[T_h^{\hat{P}_h^k, \boldsymbol{f}} \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a^*, \tau_h) - \Big[Z_h^{\boldsymbol{f}} r_h \Big](s, a^*, \tau_h) - \Big[T_h V_{h+1}^{*,\boldsymbol{f}} \Big](s, a^*, \tau_h) \\
&\quad = \Big[Z_h^{\boldsymbol{f}} \bar{r}_h^k \Big](s, a^*, \tau_h) - \Big[Z_h^{\boldsymbol{f}} r_h \Big](s, a^*, \tau_h) + \Big[\Big(T_h^{\hat{P}_h^k, \boldsymbol{f}} - T_h \Big) \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a^*, \tau_h) + \Big[T_h \Big(\bar{V}_{h+1}^{k,\boldsymbol{f}} - V_{h+1}^{*,\boldsymbol{f}} \Big) \Big](s, a^*, \tau_h) \\
&\quad \overset{(2)}{\geq} \Big[Z_h^{\boldsymbol{f}} \bar{r}_h^k \Big](s, a^*, \tau_h) - \Big[Z_h^{\boldsymbol{f}} r_h \Big](s, a^*, \tau_h) + \Big[\Big(T_h^{\hat{P}_h^k, \boldsymbol{f}} - T_h \Big) \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a^*, \tau_h) \\
&\quad = \Big[Z_h^{\boldsymbol{f}} \big(\bar{r}_h^k - r_h \big) \Big](s, a^*, \tau_h) + \Big[Z_h^{\boldsymbol{f}} \Big(\hat{P}_h^k - P_h \Big) \bar{V}_{h+1}^{k,\boldsymbol{f}} \Big](s, a^*, \tau_h),
\end{aligned}$$

where in (1) we used the definition of the max operator, and in (2) the induction step. Overall, replacing $\bar{r}_h^k$ with its definition, we get that

$$\begin{aligned}
\bar{V}_h^k(s, \tau_h) - V_h^*(s, \tau_h) &\geq \Big[Z_h^{\boldsymbol{f}} \Big(\hat{r}_h^k + b_h^{r,k} - r_h \Big) \Big](s, a^*, \tau_h) + \Big[Z_h^{\boldsymbol{f}} \Big(\Big(\hat{P}_h^k - P_h \Big) \bar{V}_{h+1}^{k,\boldsymbol{f}} + b_h^{p,k} \Big) \Big](s, a^*, \tau_h) \\
&\geq 0,
\end{aligned}$$

where the second inequality is by Lemma E.1 and Lemma E.2, which hold under $\mathcal{G}$.

□

D. Regret Analysis for Tractable LDC-UCB

In this section, we prove the regret bounds of Theorem 5.7. We start by defining the good event, which holds uniformly for all episode with probability $1-\delta$. Then, we show that the Tractable LDC-UCB is optimistic under the good event. Next, we decompose the regret to error terms of the reward, transition and latent features, and analyzing each of these terms result with the desired regret bounds.

We start by stating the bonuses which the algorithm uses:

$$b^{r,k}_{x,h}(s,a) = \min\left\{\sqrt{\frac{\log\frac{8SAMHK}{\delta}}{n^k_h(s,a,x)\vee 1}}, 1\right\}$$
$$b^{p,k}_{x,h}(s,a) = \min\left\{H\sqrt{\frac{4S\log\frac{8SAMHK}{\delta}}{n^k_h(s,a,x)\vee 1}}, 2H\right\}$$
$$b^{\boldsymbol{f},k}_{x,h}(s,a) = \frac{2\sqrt{\kappa}\gamma_k(4\delta)}{\sqrt{n^k_h(s,a,x)+4\lambda}}$$

The confidence intervals can then be written as

$$\boldsymbol{C}^k_h = \left[\boldsymbol{\sigma}(\tau^k_h;\hat{\boldsymbol{f}}^k) - \sum_{t=0}^{h-1}\alpha^{h-t-1}b^{\boldsymbol{f},k}_{x^k_t,t}(s^k_t,a^k_t), \quad \boldsymbol{\sigma}(\tau^k_h;\hat{\boldsymbol{f}}^k) + \sum_{t=0}^{h-1}\alpha^{h-t-1}b^{\boldsymbol{f},k}_{x^k_t,t}(s^k_t,a^k_t)\right]$$

D.1. Failure Events

We define the following failure events.

$$F^r_k = \left\{\exists s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, h\in[H]: \ |r_{x,h}(s,a) - \hat{r}^k_{x,h}(s,a)| > \min\left\{\sqrt{\frac{\log\frac{2SAMHK}{\delta'}}{n^k_h(s,a,x)\vee 1}}, 1\right\}\right\}$$
$$F^p_k = \left\{\exists s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, h\in[H]: \ \left\|P_{x,h}(\cdot\mid s,a) - \hat{P}^k_{x,h}(\cdot\mid s,a)\right\|_1 > \min\left\{\sqrt{\frac{4S\log\frac{2SAMHK}{\delta'}}{n^k_h(s,a,x)\vee 1}}, 2\right\}\right\}$$
$$F^n = \left\{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\frac{z^k_{i,h}}{\sqrt{n^k_h(s^k_h,a^k_h,i)\vee 1}}\,\middle|\, F_{k-1}\right] > 18H^2\log\left(\frac{1}{\delta'}\right) + 2HS(M+1)A + 4\sqrt{H^2S(M+1)AK}\right\}$$
$$F^{\boldsymbol{f},\text{local}}_k = \left\{\exists s\in\mathcal{S}, a\in\mathcal{A}, x\in\mathcal{X}, i\in[M], h\in[H]: \left|\hat{f}^k_{i,h}(s,a,x) - f^*_{i,h}(s,a,x)\right| > \frac{2\sqrt{\kappa}\gamma_k(\delta')}{\sqrt{n^k_h(s,a,x)+4\lambda}}\right\}$$

where $\gamma_k(\delta)$ is defined in Proposition G.10.

Then, we define the good event, where none of the aforementioned failure events ever occur, i.e.,

$$\mathcal{G} = \left(\cup_{k\in[K]}\bar{F}^r_k\right)\cap\left(\cup_{k\in[K]}\bar{F}^p_k\right)\cap\left(\cup_{k\in[K]}\bar{F}^{\boldsymbol{f},\text{local}}_k\right)\cap\bar{F}^n$$

Lemma D.1. *Letting $\delta' = \delta/4$, the event $\mathcal{G}$ holds with probability at least $1-\delta$.*

Proof. The proof is almost identical to the one of Lemma C.1. We only need to prove that $\Pr\left\{\cup_{k\in[K]}\bar{F}^{\boldsymbol{f},\text{local}}_k\right\} \le \delta'$, which directly follows by Lemma 5.2. □

D.2. Regret Analysis – Proof of Theorem 5.7

Theorem 5.7. *Let $\lambda = \Theta(\frac{HM^{2.5}SA}{L})$. With probability at least $1-\delta$, the regret of Algorithm 2 is*

$$R(K) \leq \tilde{\mathcal{O}}\Big(\sqrt{H^8 M^{6.5} S^2 A^2 L^4 \kappa K}\Big).$$

Proof. Let

$$\bar{\boldsymbol{\sigma}}_h^k(s,\tau_h) \in \arg\max_{\bar{\boldsymbol{\sigma}} \in \boldsymbol{C}_h^k} \max_a \left\{ \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}}) \bar{r}_{i,h}^k(s,a) + \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}}) \hat{P}_{i,h}^k(\cdot|s,a)^T \bar{V}_{h+1}^k(\cdot,\tau_{h+1}) \right\},$$

and denote $\bar{Z}_h^k(s,\tau_h) = \boldsymbol{z}(\bar{\boldsymbol{\sigma}}_h^k(s,\tau_h))$.

Under the good event, the conditions of the regret decomposition lemma (Lemma E.6) hold with $\bar{r}, \hat{P}, \bar{Z}$ and $c=4$ due to the truncated value iteration, truncated bonuses and value optimism lemma (Proposition C.5). Therefore, the regret can be decomposed in the following way,

$$\begin{aligned}
\mathrm{Reg}(K) \leq & \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\big[\, z_{i,h}^k \left|\bar{r}_{i,h}^k(s_h^k,a_h^k) - r_{i,h}(s_h^k,a_h^k)\right| \,\big|\, F_{k-1} \,\big]}_{(i)} \\
& + \underbrace{H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\Big[\, z_{i,h}^k \left\|\Big(\hat{P}_h^k - P_h\Big)(\cdot|s_h^k,a_h^k)\right\|_1 \,\Big|\, F_{k-1} \,\Big]}_{(ii)} \\
& + \underbrace{5H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\, \left\|\bar{Z}_h^k - Z_h^{\boldsymbol{f}^*}\right\|_1 \,\Big|\, F_{k-1} \,\Big]}_{(iii)}.
\end{aligned}$$

The terms (i) and (iii) are identical to the ones in the proof of Theorem 4.1 (as the reward bonuses are identical), and thus can be bounded by Lemma C.2 and Lemma C.3. Term (iii) can be bounded by Lemma D.2.

Thus, we obtain,

$$\begin{aligned}
\mathrm{Reg}(K) \leq & \, \mathcal{O}\left(H^2 S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right) \\
& + \mathcal{O}\left(H^2 S\sqrt{MAK\log\frac{SAMHK}{\delta}}\right) \\
& + \tilde{\mathcal{O}}\Big(\sqrt{H^8 S^2 A^2 M^{6.5} L^4 \kappa K}\Big) \\
\leq & \, \tilde{\mathcal{O}}\Big(\sqrt{H^8 S^2 A^2 M^{6.5} L^4 \kappa K}\Big)
\end{aligned}$$

□

Lemma D.2 (Latent Features Concentration)**.** *Under the good event, it holds that*

$$\begin{aligned}
5H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\, \left\|Z_h^{\bar{\boldsymbol{f}}_k} - Z_h^{\boldsymbol{f}^*}\right\|_1 \,\Big|\, F_{k-1} \,\Big] & \leq \mathcal{O}\left(\left(\frac{1}{\sqrt{\lambda}} \vee 1\right)\gamma_K(4\delta)\sqrt{\kappa H^6 S M^3 A K}\right) \\
& \leq \tilde{\mathcal{O}}\Big(\sqrt{H^8 S^2 A^2 M^{6.5} L^4 \kappa K}\Big)
\end{aligned}$$

Proof.

$$
\begin{aligned}
&5H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\Big\|Z_h^{\bar{\boldsymbol{f}}_k}-Z_h^{\boldsymbol{f}^*}\Big\|_1\;\Big|\;F_{k-1}\Big]\\
&\qquad=5H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\big[\,\big\|\boldsymbol{z}(\bar{\boldsymbol{\sigma}}_h^k(s,\tau_h^k))-\boldsymbol{z}(\boldsymbol{\sigma}(\tau_h^k;\boldsymbol{f}^*)\big\|_1\;\big|\;F_{k-1}\big]\\
&\qquad\overset{(1)}{\le}10H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\big[\,\big|z_i(\bar{\boldsymbol{\sigma}}_h^k(s,\tau_h^k))-z_i(\boldsymbol{\sigma}(\tau_h^k;\boldsymbol{f}^*)\big|\;\big|\;F_{k-1}\big]\\
&\qquad\overset{(2)}{\le}2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\big[\,\big|\bar{\boldsymbol{\sigma}}_{i,h}^k(s,\tau_h^k)-\boldsymbol{\sigma}_i(\tau_h^k;\boldsymbol{f}^*)\big|\;\big|\;F_{k-1}\big]
\end{aligned}
$$

where relation (1) holds by substituting $z_{M+1}(\boldsymbol{x})=1-\sum_{i=1}^{M}z_i(\boldsymbol{x})$ and applying the triangle inequality, and relation (2) is since the function $f(x)=e^x/(a+e^x)$ is $\frac{1}{4}$-Lipschitz, and $z_i(\boldsymbol{x})$ can be represented as such a function of x_i. Then,

$$
\begin{aligned}
&5H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\Big\|Z_h^{\bar{\boldsymbol{f}}_k}-Z_h^{\boldsymbol{f}^*}\Big\|_1\;\Big|\;F_{k-1}\Big]\\
&\qquad\le 2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\big[\,\big|\bar{\boldsymbol{\sigma}}_{i,h}^k(s,\tau_h^k)-\boldsymbol{\sigma}_i(\tau_h^k;\boldsymbol{f}^*\big|\;\big|\;F_{k-1}\big]\\
&\qquad\le 2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\Big[\,\Big|\boldsymbol{\sigma}_i(\tau_h;\hat{\boldsymbol{f}}^k)-\boldsymbol{\sigma}_i(\tau_h^k;\boldsymbol{f}^*\Big|\;\Big|\;F_{k-1}\Big]\\
&\qquad\quad+2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\left[\sum_{t=0}^{h-1}\alpha^{h-t-1}b_{x_t^k,t}^{\boldsymbol{f},k}(s_t^k,a_t^k)\;\middle|\;F_{k-1}\right]\\
&\qquad\le 2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\left[\sum_{t=0}^{h-1}\alpha^{h-t-1}b_{x_t^k,t}^{\boldsymbol{f},k}(s_t^k,a_t^k)\;\middle|\;F_{k-1}\right]\\
&\qquad\quad+2.5H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M}\mathbb{E}\left[\sum_{t=0}^{h-1}\alpha^{h-t-1}b_{x_t^k,t}^{\boldsymbol{f},k}(s_t^k,a_t^k)\;\middle|\;F_{k-1}\right] && \text{(Under }\mathcal{G}\text{)}\\
&\qquad=5HM\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\sum_{t=0}^{h-1}\alpha^{h-t-1}b_{x_t^k,t}^{\boldsymbol{f},k}(s_t^k,a_t^k)\;\middle|\;F_{k-1}\right]\\
&\qquad\le 5H^2M\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\,b_{x_h^k,h}^{\boldsymbol{f},k}(s_h^k,a_h^k)\;\Big|\;F_{k-1}\Big]\\
&\qquad\le 10H^2M\sqrt{\kappa}\gamma_K(4\delta)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)+4\lambda}}\;\middle|\;F_{k-1}\right]\\
&\qquad\le 10H^2M\left(\frac{1}{2\sqrt{\lambda}}\vee 1\right)\sqrt{\kappa}\gamma_K(4\delta)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\;\middle|\;F_{k-1}\right]\\
&\qquad\le 10H^2M\left(\frac{1}{2\sqrt{\lambda}}\vee 1\right)\sqrt{\kappa}\gamma_K(4\delta)\left(18H^2\log\left(\frac{1}{\delta}\right)+2HS(M+1)A+4\sqrt{H^2S(M+1)AK}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Lemma E.7)}\\
&\qquad=\mathcal{O}\left(\left(\frac{1}{\sqrt{\lambda}}\vee 1\right)\gamma_K(4\delta)\sqrt{\kappa H^6SM^3AK}\right)
\end{aligned}
$$

Now, note that by the definition in Proposition G.10, $\gamma_k(\delta) := \Big(2+2L+\sqrt{2(1+L)}\Big)\beta_k(\delta) + \sqrt{\frac{2(1+L)HM}{\lambda}}\beta_k^2(\delta)$. Also, by Equation (12), $\beta_k(\delta) = \frac{M^{3/2}(M+1)SAH}{\sqrt{\lambda}}\Big(\log\Big(1+\frac{k}{(M+1)SA\lambda}\Big)+2\log\big(\frac{2}{\delta}\big)\Big)+\sqrt{\frac{\lambda}{4M}}+\sqrt{\lambda}L$. Plugging in these definitions we have that,

$$
\begin{aligned}
\gamma_k(\delta) &= \Big(2+2L+\sqrt{2(1+L)}\Big)\left(\frac{M^{3/2}(M+1)SAH}{\sqrt{\lambda}}\left(\log\left(1+\frac{k}{(M+1)SA\lambda}\right)+2\log\left(\frac{2}{\delta}\right)\right)+\sqrt{\frac{\lambda}{4M}}+\sqrt{\lambda}L\right)\\
&+\sqrt{\frac{2(1+L)HM}{\lambda}}\left(\frac{M^{3/2}(M+1)SAH}{\sqrt{\lambda}}\left(\log\left(1+\frac{k}{(M+1)SA\lambda}\right)+2\log\left(\frac{2}{\delta}\right)\right)+\sqrt{\frac{\lambda}{4M}}+\sqrt{\lambda}L\right)^2\\
&\le \mathcal{O}\left(\frac{LM^{5/2}SAH}{\lambda}\log\left(\frac{1}{\delta}+\frac{k}{MSA\lambda\delta}\right)+L\sqrt{\frac{\lambda}{M}}+\sqrt{\lambda}L^2\right)\\
&+\sqrt{\frac{2(1+L)HM}{\lambda}}\mathcal{O}\left(\frac{M^5S^2AH^2}{\lambda}\log^2\left(\frac{1}{\delta}+\frac{k}{MSA\lambda\delta}\right)+\frac{\lambda}{M}+\lambda L^2\right)\\
&=\tilde{\mathcal{O}}\Big(\lambda^{-1}LM^{5/2}SAH+\lambda^{1/2}LM^{-1/2}+\lambda^{1/2}L^2\Big)\\
&+\tilde{\mathcal{O}}\Big(\lambda^{-3/2}L^{1/2}M^{11/2}S^2A^2H^{5/2}+\lambda^{1/2}L^{1/2}M^{-1/2}H^{1/2}+\lambda^{1/2}L^{5/2}M^{1/2}H^{1/2}\Big)\\
&\le\tilde{\mathcal{O}}\Big(\lambda^{-1}LM^{5/2}SAH+\lambda^{1/2}LM^{-1/2}+\lambda^{-3/2}L^{1/2}M^{11/2}S^2A^2H^{5/2}\\
&\qquad+\lambda^{1/2}L^{1/2}M^{-1/2}H^{1/2}+\lambda^{1/2}L^{5/2}M^{1/2}H^{1/2}\Big)\\
&\le\tilde{\mathcal{O}}\Big(L^2M^{7/4}S^{1/2}A^{1/2}H\Big),
\end{aligned}
$$

where we used $\lambda = \frac{M^{2.5}SAH}{L}$ to minimize the above term.

Finally, plugging in this expression we have that

$$
\begin{aligned}
5H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\Big[\Big\|Z_h^{\bar{\boldsymbol{f}}_k}-Z_h^{\boldsymbol{f}^*}\Big\|_1\,\Big|\,F_{k-1}\Big] &\le \mathcal{O}\Big(\Big(\frac{1}{\sqrt{\lambda}}\vee 1\Big)\gamma_K(4\delta)\sqrt{\kappa H^6SM^3AK}\Big)\\
&\le\tilde{\mathcal{O}}\Big(L^2M^{7/4}S^{1/2}A^{1/2}H\sqrt{\kappa H^6SM^3AK}\Big)\\
&=\tilde{\mathcal{O}}\Big(\sqrt{H^8S^2A^2M^{6.5}L^4\kappa K}\Big),
\end{aligned}
$$

where we assumed that $\lambda \ge 1$ □

D.3. Optimism in Tractable Logistic DCMDPs – Proof of Proposition 5.3

In this section, we prove Proposition 5.3, which allows us to apply the regret decomposition (Lemma E.6) necessary for proving Theorem 5.7.

Proposition 5.3 (Optimistic Value)**.** *Let $\bar{V}_h$ as defined in Equation* (8)*. Then, w.h.p. $\bar{V}_1(s_1^k, \boldsymbol{C}_1^k) \ge V_1^*(s_1^k)$.*

Proof. We divide the proof into two steps. Defining $\bar{V}_h(s,\tau_h)$ the optimistic value function which follows the equations

$$
\begin{aligned}
&\bar{V}_{H+1}^k(s,\tau_{H+1}) = 0, \quad \forall s\in\mathcal{S}, \tau_{H+1}\in\mathcal{H}_{H+1}, \text{ and}\\
&\bar{V}_h^k(s,\tau_h) = \min\left\{H, \max_a \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}(\boldsymbol{\sigma}(\tau_h))}\left\{\sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\bar{r}_{i,h}^k(s,a)+\sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\hat{P}_{i,h}^k(\cdot|s,a)^T\bar{V}_{h+1}^k(\cdot,\tau_{h+1})\right\}\right\},\\
&\forall h\in[H], s\in\mathcal{S}, \tau_h\in\mathcal{H}_h,
\end{aligned}
$$

we first show that $\bar{V}_h^k(s,\boldsymbol{C}_h^k(\tau_h)) = \bar{V}_h^k(s,\tau_h)$ for all $k\in[K], h\in[H], s\in\mathcal{S}$ and $\tau_h\in\mathcal{H}_h$. This follows due to a simple induction; first notice that the claim trivially holds when $h = H+1$, where both values are 0. Now fix $h\in[H]$ and assume

that $\bar{V}^k_{h+1}(s, \boldsymbol{C}^k_h(\tau_{h+1})) = \bar{V}^k_{h+1}(s, \tau_{h+1})$ for all $s \in \mathcal{S}$ and $\tau_{h+1} \in \mathcal{H}_{h+1}$. In the following, we prove that this implies $\bar{V}^k_h(s, \boldsymbol{C}^k_h(\tau_h)) = \bar{V}^k_h(s, \tau_h)$ for any $s \in \mathcal{S}$ and $\tau_h \in \mathcal{H}_h$, which prove the claim.

$$
\begin{aligned}
\bar{V}_h(s, \boldsymbol{C}^k_h(\tau_h)) &= \min\left\{\max_{a\in\mathcal{A}} \max_{t\in\mathcal{T}(\bar{\boldsymbol{Q}})} \sum_{i=0}^{M} z_i\Big(\boldsymbol{th}_t\Big(\bar{\boldsymbol{Q}}, \boldsymbol{C}^k_h(\tau_h)\Big)\Big)\bar{Q}_i(s, a, \boldsymbol{C}^k_h(\tau_h)), H\right\} \\
&\overset{(1)}{=} \min\left\{\max_{a\in\mathcal{A}} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h(\tau_h)} \sum_{i=0}^{M} z_i(\bar{\boldsymbol{\sigma}})\bar{Q}_i(s, a, \boldsymbol{C}^k_h(\tau_h)), H\right\} \\
&= \min\left\{\max_{a\in\mathcal{A}} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h(\tau_h)} \sum_{i=0}^{M} z_i(\bar{\boldsymbol{\sigma}})\Big(\bar{r}_i(s,a) + \mathbb{E}_{s'\sim\hat{P}_i(\cdot|s,a)}\big[\bar{V}_{h+1}(s', \boldsymbol{C}_{h+1}(a,i))\big]\Big), H\right\} \\
&\overset{(2)}{=} \min\left\{\max_{a\in\mathcal{A}} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h(\tau_h)} \sum_{i=0}^{M} z_i(\bar{\boldsymbol{\sigma}})\Big(\bar{r}_i(s,a) + \mathbb{E}_{s'\sim\hat{P}_i(\cdot|s,a)}\Big[\bar{V}_{h+1}(s', \boldsymbol{C}^k_{h+1})\Big]\Big), H\right\} \\
&\overset{(3)}{=} \min\left\{\max_{a\in\mathcal{A}} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h(\tau_h)} \sum_{i=0}^{M} z_i(\bar{\boldsymbol{\sigma}})\Big(\bar{r}_i(s,a) + \mathbb{E}_{s'\sim\hat{P}_i(\cdot|s,a)}\big[\bar{V}^k_{h+1}(s, \tau_h)\big]\Big), H\right\} \\
&= \bar{V}^k_h(s, \tau_h)
\end{aligned}
$$

Relation (1) is by Lemma 5.6, which proves that when the confidence interval of a multinomial function is rectangular, one of the maximizers of an linear combination w.r.t. this function is a threshold function; therefore, the maximum over threshold functions achieves the same value at the rectangular set. Relation (2) is by the definition of $\boldsymbol{C}_{h+1}(a_h, x_h)$ at Algorithm 4, which implies that $\boldsymbol{C}_{h+1}(a_h, x_h) = \boldsymbol{C}^k_{h+1}$. Finally, (3) is by the induction hypothesis.

Next, we prove that under the good event, $\bar{V}^k_h(s, \tau_h) \geq V^*_h(s, \tau_h)$ for all $k \in [K], h \in [H], s \in \mathcal{S}$ and $\tau_h \in \mathcal{H}_h$. This claim is also proved by induction and clearly holds when $h = H + 1$, when all values equal zero. Assume that $\bar{V}^k_{h+1}(s, \tau_{h+1}) \geq V^*_{h+1}(s, \tau_{h+1})$ for all $s \in \mathcal{S}$, $\tau_{h+1} \in \mathcal{H}_{h+1}$. Also, assume w.l.o.g. that $\bar{V}^k_h(s, \tau_h) < H$, otherwise the claim trivially holds. Then, denoting

$$
a^* \in \arg\max_a \Big\{\Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a, \tau_h) + \Big[T_h V^{*,\boldsymbol{f}^*}_{h+1}\Big](s, a, \tau_h)\Big\},
$$

and under $\mathcal{G}$,

$$
\begin{aligned}
&\bar{V}^k_h(s, \tau_h) - V^*_h(s, \tau_h) \\
&\quad= \max_a \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h} \left\{\sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\bar{r}^k_{i,h}(s,a) + \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\hat{P}^k_{i,h}(\cdot|s,a)^T \bar{V}^k_{h+1}(\cdot, \tau_{h+1})\right\} \\
&\qquad - \max_a \Big\{\Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a, \tau_h) + \Big[T_h V^{*,\boldsymbol{f}^*}_{h+1}\Big](s, a, \tau_h)\Big\} \\
&\quad\overset{(1)}{\geq} \max_{\bar{\boldsymbol{\sigma}}\in\boldsymbol{C}^k_h} \left\{\sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\bar{r}^k_{i,h}(s,a^*) + \sum_{i=1}^{M+1} z_i(\bar{\boldsymbol{\sigma}})\hat{P}^k_{i,h}(\cdot|s,a^*)^T \bar{V}^k_{h+1}(\cdot, \tau_{h+1})\right\} \\
&\qquad - \Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a^*, \tau_h) - \Big[T_h V^{*,\boldsymbol{f}^*}_{h+1}\Big](s, a^*, \tau_h) \\
&\quad\overset{(2)}{\geq} \Big[Z^{\boldsymbol{f}^*}_h \bar{r}^k_h\Big](s, a^*, \tau_h) - \Big[T^{\hat{P}^k_h, \boldsymbol{f}^*}_h \bar{V}^k_{h+1}\Big](s, a^*, \tau_h) - \Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a^*, \tau_h) - \Big[T_h V^{*,\boldsymbol{f}^*}_{h+1}\Big](s, a^*, \tau_h) \\
&\quad= \Big[Z^{\boldsymbol{f}^*}_h \bar{r}^k_h\Big](s, a^*, \tau_h) - \Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a^*, \tau_h) + \Big[\Big(T^{\hat{P}^k_h, \boldsymbol{f}^*}_h - T_h\Big)\bar{V}^{k,\boldsymbol{f}^*}_{h+1}\Big](s, a^*, \tau_h) + \Big[T_h\Big(\bar{V}^{k,\boldsymbol{f}^*}_{h+1} - V^{*,\boldsymbol{f}^*}_{h+1}\Big)\Big](s, a^*, \tau_h) \\
&\quad\overset{(3)}{\geq} \Big[Z^{\boldsymbol{f}^*}_h \bar{r}^k_h\Big](s, a^*, \tau_h) - \Big[Z^{\boldsymbol{f}^*}_h r_h\Big](s, a^*, \tau_h) + \Big[\Big(T^{\hat{P}^k_h, \boldsymbol{f}^*}_h - T_h\Big)\bar{V}^{k,\boldsymbol{f}^*}_{h+1}\Big](s, a^*, \tau_h) \\
&\quad= \Big[Z^{\boldsymbol{f}^*}_h \big(\bar{r}^k_h - r_h\big)\Big](s, a^*, \tau_h) + \Big[Z^{\boldsymbol{f}^*}_h \Big(\hat{P}^k_h - P_h\Big)\bar{V}^{k,\boldsymbol{f}^*}_{h+1}\Big](s, a^*, \tau_h).
\end{aligned}
$$

In (1) we used the definition of the max operator. Relation (2) holds since under the good event,

$$\begin{aligned}
\boldsymbol{C}(\boldsymbol{\sigma}(\tau_h^k;\boldsymbol{f}^*)) &= \sum_{t=0}^{h-1}\alpha^{h-t-1}\boldsymbol{f}_t(s_t,a_t,x_t) \\
&\in \left[\boldsymbol{\sigma}(\tau_h;\hat{\boldsymbol{f}}^k) - \sum_{t=0}^{h-1}\alpha^{h-t-1}b^{\boldsymbol{f},k}_{x_t^k,t}(s_t^k,a_t^k),\quad \boldsymbol{\sigma}(\tau;\hat{\boldsymbol{f}}^k) + \sum_{t=0}^{h-1}\alpha^{h-t-1}b^{\boldsymbol{f},k}_{x_t^k,t}(s_t^k,a_t^k)\right] \qquad \text{(Under } \mathcal{G}\text{)} \\
&= \boldsymbol{C}_h^k.
\end{aligned}$$

In (3), we used the induction step the induction step. Overall, replacing $\bar{r}_h^k$ with its definition, we get that

$$\begin{aligned}
\bar{V}_h^k(s,\tau_h) - V_h^*(s,\tau_h) &\geq \left[Z_h^{\boldsymbol{f}^*}\left(\hat{r}_h^k + b_h^{r,k} - r_h\right)\right](s,a^*,\tau_h) + \left[Z_h^{\boldsymbol{f}^*}\left(\left(\hat{P}_h^k - P_h\right)\bar{V}_{h+1}^{k,\boldsymbol{f}^*} + b_h^{p,k}\right)\right](s,a^*,\tau_h) \\
&\geq 0,
\end{aligned}$$

where the second inequality is by Lemma E.1 and Lemma E.2, which hold under $\mathcal{G}$. □

E. Useful Lemmas

E.1. Optimism Lemmas

Lemma E.1 (Reward Optimism)**.** *For any $k \geq 1$, define the event*

$$F_k^r = \Big\{\exists s \in \mathcal{S}, a \in \mathcal{A}, i \in [M], h \in [H] : \ |r_{i,h}(s,a) - \hat{r}_{i,h}^{k,r}(s,a)| > b_{i,h}^k(s,a)\Big\}.$$

Then, under $\bar{F}_k^r$, for any $\boldsymbol{f} \in \mathcal{F}$, $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $\tau_h \in \mathcal{H}_h$, it holds that

$$\Big[Z_h^{\boldsymbol{f}}(\hat{r}_h^k + b_h^{r,k} - r)\Big](s,a,\tau_h) \geq 0$$

Proof. The result directly follows by the definition of $\bar{F}_k^r$, since

$$\begin{aligned}
\Big[Z_h^{\boldsymbol{f}}(\hat{r}_h^k + b_h^{r,k} - r)\Big](s,a,\tau_h) &\geq \min_i \Big\{\big(\hat{r}_{i,h}^k(s,a) - r_i(s,a)\big) + b_{i,h}^{r,k}(s,a)\Big\} \\
&\geq \min_i \Big\{-b_{i,h}^{r,k}(s,a) + b_{i,h}^{r,k}(s,a)\Big\} && \text{(Under } \bar{F}_k^r\text{)} \\
&= 0
\end{aligned}$$

□

Lemma E.2 (Transition Optimism)**.** *For any $k \geq 1$, define the event*

$$F_k^p = \Big\{\exists s \in \mathcal{S}, a \in \mathcal{A}, i \in [M], h \in [H] : \ \Big\|P_{i,h}(\cdot \mid s,a) - \hat{P}_{i,h}^k(\cdot \mid s,a)\Big\|_1 > \frac{1}{H} b_{i,h}^{p,k}(s,a)\Big\}.$$

Then, under $\bar{F}_k^p$, for any $\boldsymbol{f} \in \mathcal{F}$, $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, $\tau_h \in \mathcal{H}_h$ and $V \in [0,H]^S$, it holds that

$$\Big[Z_h^{\boldsymbol{f}}\Big(\Big(\hat{P}_h^k - P_h\Big)V + b_h^{p,k}\Big)\Big](s,a,\tau_h) \geq 0$$

Proof. The result directly follows by the definition of $\bar{F}_k^p$ and Cauchy-Schwartz inequality, since

$$\begin{aligned}
\Big[Z_h^{\boldsymbol{f}}\Big(\Big(\hat{P}_h^k - P_h\Big)V + b_h^{p,k}\Big)\Big](s,a,\tau_h) &\geq \min_i \Big\{\Big[\Big(\hat{P}_h^k - P_h\Big)V\Big](s,a) + b_{i,h}^{p,k}(s,a)\Big\} \\
&\geq \min_i \Big\{-\Big\|P_{i,h}(\cdot \mid s,a) - \hat{P}_{i,h}^k(\cdot \mid s,a)\Big\|_1 \|V\|_\infty + b_{i,h}^{p,k}(s,a)\Big\} && \text{(C.S)} \\
&\geq \min_i \Big\{-\frac{1}{H} b_{i,h}^{p,k}(s,a) \cdot H + b_{i,h}^{p,k}(s,a)\Big\} && \text{(Under } \bar{F}_k^p\text{)} \\
&= 0
\end{aligned}$$

□

E.2. Decomposition Lemmas

Lemma E.3.

$$\begin{aligned}
V &: \mathcal{S} \times \mathcal{H} \mapsto \mathbb{R}, \\
Z^{(1)}, Z^{(2)} &: \mathcal{S} \times \mathcal{A} \times \mathcal{H} \mapsto \Delta_{\mathcal{X}}, \\
r^{(1)}, r^{(2)} &: \mathcal{S} \times \mathcal{A} \times \mathcal{H} \times \mathcal{X} \mapsto \mathbb{R}, \text{ and} \\
P^{(1)}, P^{(2)} &: \mathcal{S} \times \mathcal{A} \times \mathcal{H} \times \mathcal{X} \mapsto \Delta_{\mathcal{S}}.
\end{aligned}$$

Then, for any $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], \tau_h \in \mathcal{H}$

$$\begin{aligned}\Big[Z_h^{(1)} r_h^{(1)} + T_h^{P^{(1)},Z^{(1)}} V_{h+1}\Big](s,a,\tau_h) - \Big[Z_h^{(2)} r_h^{(2)} + T_h^{P^{(2)},Z^{(2)}} V_{h+1}\Big](s,a,\tau_h) \\ = \Big[Z_h^{(2)}\Big(r_h^{(1)} - r_h^{(2)}\Big)\Big](s,a,\tau_h) \\ + \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big)\Big(r_h^{(1)} + P_h^{(1)} V_{h+1}\Big)\Big](s,a,\tau_h) \\ + \Big[Z_h^{(2)}\Big(P_h^{(1)} - P_h^{(2)}\Big) V_{h+1}\Big](s,a,\tau_h).\end{aligned}$$

Proof. We have that

$$\begin{aligned}&\Big[Z_h^{(1)} r_h^{(1)} + T_h^{P^{(1)},Z^{(1)}} V_{h+1}\Big](s,a,\tau_h) - \Big[Z_h^{(2)} r_h^{(2)} + T_h^{P^{(2)},Z^{(2)}} V_{h+1}\Big](s,a,\tau_h) \\ &= \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big) r_h^{(1)}\Big](s,a,\tau_h) + \Big[Z^{(2)}\Big(r_h^{(1)} - r_h^{(2)}\Big)\Big](s,a,\tau_h) + \Big[\Big(T_h^{P^{(1)},Z^{(1)}} - T_h^{P^{(2)},Z^{(2)}}\Big) V_{h+1}\Big](s,a,\tau_h) \\ &= \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big) r_h^{(1)}\Big](s,a,\tau_h) + \Big[Z^{(2)}\Big(r_h^{(1)} - r_h^{(2)}\Big)\Big](s,a,\tau_h) + \Big[\Big(Z_h^{(1)} P_h^{(1)} - Z_h^{(2)} P_h^{(2)}\Big) V_{h+1}\Big](s,a,\tau_h) \\ &= \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big) r_h^{(1)}\Big](s,a,\tau_h) + \Big[Z^{(2)}\Big(r_h^{(1)} - r_h^{(2)}\Big)\Big](s,a,\tau_h) \\ &\quad + \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big) P_h^{(1)} V_{h+1}\Big](s,a,\tau_h) + \Big[Z_h^{(2)}\Big(P_h^{(1)} - P_h^{(2)}\Big) V_{h+1}\Big](s,a,\tau_h) \\ &= \Big[Z_h^{(2)}\Big(r_h^{(1)} - r_h^{(2)}\Big)\Big](s,a,\tau_h) + \Big[\Big(Z_h^{(1)} - Z_h^{(2)}\Big)\Big(r_h^{(1)} + P_h^{(1)} V_{h+1}\Big)\Big](s,a,\tau_h) + \Big[Z_h^{(2)}\Big(P_h^{(1)} - P_h^{(2)}\Big) V_{h+1}\Big](s,a,\tau_h).\end{aligned}$$

This completes the proof. □

Next, recall that by embedding the history into the state, every DCMDP can be represented as an MDP. This equivalence will allow us to apply the following lemma on DCMDPs:

Lemma E.4 (Value difference lemma, e.g., Dann et al. (2017), Lemma E.15)**.** *Consider two MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H)$ and $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P', r', H)$. For any policy π and any s, h, the following relation holds:*

$$\begin{aligned}&V_h^\pi(s;\mathcal{M}') - V_h^\pi(s;\mathcal{M}) \\ &\quad = \mathbb{E}\left[\sum_{t=h}^{H} (r_t'(s_t,a_t) - r_t(s_t,a_t)) + (P' - P)(\cdot \mid s_t, a_t)^T V_{t+1}^\pi(\cdot;\mathcal{M}') | s_h = s, \pi, P\right]\end{aligned}$$

Corollary E.5 (Truncated value difference lemma)**.** *Consider two MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H)$ and $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P', r', H)$. Also, for any $C \in \mathbb{R}$, define the truncated value of a policy π under MDP $\mathcal{M}$ by the solution to the truncated dynamic programming problem*

$$\begin{aligned}&V_{H+1}^\pi(s;\mathcal{M},C) = 0, && \forall s \in \mathcal{S} \\ &V_h^\pi(s;\mathcal{M},C) = \mathbb{E}_{a\sim\pi}\big[\min\big\{C, r_h(s,a) + P(\cdot \mid s,a)^T V_{h+1}^\pi(\cdot;\mathcal{M},C)\big\}\big], && \forall h \in [H], s \in \mathcal{S}.\end{aligned}$$

Then, for any policy π, any $s \in \mathcal{S}, h \in [H]$ and any $C \in \mathbb{R}$, the following relation holds:

$$\begin{aligned}&V_h^\pi(s;\mathcal{M}',C) - V_h^\pi(s;\mathcal{M}) \\ &\quad \le \mathbb{E}\left[\sum_{t=h}^{H} (r_t'(s_t,a_t) - r_t(s_t,a_t)) + (P' - P)(\cdot \mid s_t, a_t)^T V_{t+1}^\pi(\cdot;\mathcal{M}',C) | s_h = s, \pi, P\right]\end{aligned}$$

Proof. We build an MDP whose value (without truncation) is $V_h^\pi(s;\mathcal{M}',C)$ and its reward are always smaller than the rewards of $\mathcal{M}'$. In particular, for any $h \in [H], s \in \mathcal{S}$ and $a \in \mathcal{A}$, define the new reward function

$$\bar{r}_h(s,a) = r_h'(s,a) - \max\big\{0, r_h'(s,a) + P'(\cdot \mid s,a)^T V_{h+1}^\pi(s;\mathcal{M}',C) - C\big\} \le r_h'(s,a), \tag{10}$$

and denote $\bar{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, P', \bar{r}, H)$. Clearly, $V^{\pi}_{H+1}(s; \bar{\mathcal{M}}) = V^{\pi}_{H+1}(s; \mathcal{M}', C) = 0$. Now assume by induction the the equality holds for all $t > h$ and all $s \in \mathcal{S}$.

Let $s \in \mathcal{S}$ be some state. If for some $a \in \mathcal{A}$, the maximizer in Equation (10) equals zero, then there was no truncation in the value iteration and so, by the induction hypothesis, we get

$$\begin{aligned}\bar{r}_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \bar{\mathcal{M}}) &= r'_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \mathcal{M}', C) \\ &= \min\big\{C, r_h(s,a) + P(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \mathcal{M}', C)\big\}.\end{aligned}$$

On the other hand, if the maximizer in Equation (10) is not zero, then one can easily verify that

$$\bar{r}_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \bar{\mathcal{M}}) = \min\big\{C, r_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \mathcal{M}', C)\big\} = C$$

Therefore, this equality holds for all $a \in \mathcal{A}$ and thus

$$\begin{aligned}V^{\pi}_h(s; \bar{\mathcal{M}}) &= \mathbb{E}_{a\sim\pi}\big[\bar{r}_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \bar{\mathcal{M}})\big] \\ &= \mathbb{E}_{a\sim\pi}\big[\min\big\{C, r_h(s,a) + P'(\cdot \mid s,a)^T V^{\pi}_{h+1}(\cdot; \mathcal{M};, C)\big\}\big] \\ &= V^{\pi}_h(s; \mathcal{M}', C),\end{aligned}$$

and by induction, this equality holds for all $h \in [H]$ and $s \in \mathcal{S}$. Now, using this fact with Lemma E.4 on $\mathcal{M}$ and $\bar{\mathcal{M}}$, we get:

$$\begin{aligned}&V^{\pi}_h(s; \mathcal{M}', C) - V^{\pi}_h(s; \mathcal{M}) \\ &\quad= \mathbb{E}\left[\sum_{t=h}^{H} (\bar{r}_t(s_t,a_t) - r_t(s_t,a_t)) + (P' - P)(\cdot \mid s_t,a_t)^T V^{\pi}_{t+1}(\cdot; \mathcal{M}', C) | s_h = s, \pi, P\right] \\ &\quad\le \mathbb{E}\left[\sum_{t=h}^{H} (r'_t(s_t,a_t) - r_t(s_t,a_t)) + (P' - P)(\cdot \mid s_t,a_t)^T V^{\pi}_{t+1}(\cdot; \mathcal{M}', C) | s_h = s, \pi, P\right]\end{aligned}$$

where the inequality is since $\bar{r}_h(s,a) \le r'_h(s,a)$ for all h, s, a. □

We are now ready to present the general regret decomposition lemma.

Lemma E.6 (Regret Decomposition)**.** *Assume that there exist an optimistic value function $\bar{V}^k_h$ such that the following hold:*

1. ***Value representation.*** *For all $k \in [K]$ and $h \in [H]$, there exist $\bar{Z}^k_h : \mathcal{S} \times \mathcal{A} \times \mathcal{H}_h \mapsto \Delta_{\mathcal{X}}$, $\bar{r}^k_h : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto \mathbb{R}$ and $\bar{P}^k_h : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \mapsto \Delta_{\mathcal{S}}$ such that for all $k \ge 1$, $h \in [H]$, $s, \in \mathcal{S}$ and $\tau_h \in \mathcal{H}_h$, it holds that*

$$\bar{V}^k_h(s, \tau_h) \le \bar{Z}^k_h \bar{r}^k_h + T_h^{\bar{P}^k_h, \bar{\boldsymbol{f}}_k}(\cdot | s^k_h, a^k_h, \tau^k_h)^T \bar{V}^k_{h+1}(\cdot, \tau^k_{h+1}).$$

2. ***Boundedness.*** *For all $k \ge 1$, $h \in [H]$, $s, \in \mathcal{S}$, $a \in \mathcal{A}$ $i \in [M+1]$ and $\tau_h \in \mathcal{H}_h$, it holds that $0 \le \bar{V}^k_h(s, \tau_h) \le H$ and $0 \le \bar{r}^k_{i,h}(s, a, \tau_h) \le cH$ for some $c > 0$.*

3. ***Optimism.*** *For all $k \ge 1$, it holds that $\bar{V}^k_h(s^k_1) \ge V^*_1(s^k_1)$.*

Then, the regret can be bounded by

$$\begin{aligned}\mathrm{Reg}(K) \le &\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z^k_{i,h} \left|\bar{r}^k_{i,h}(s^k_h, a^k_h) - r_{i,h}(s^k_h, a^k_h)\right| \,\middle|\, F_{k-1}\,\right] \\ &+ H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\left[\, z^k_{i,h} \left\|\big(\bar{P}^k_h - P_h\big)(\cdot | s^k_h, a^k_h)\right\|_1 \,\middle|\, F_{k-1}\,\right] \\ &+ (c+1)H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\left[\, \left\|\bar{Z}^k_h - Z^{\boldsymbol{f}^*}_h\right\|_1 \,\middle|\, F_{k-1}\,\right]\end{aligned}$$

Proof.

$$
\begin{aligned}
\mathrm{Reg}(K) &= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&\le \sum_{k=1}^{K} \bar{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) && \text{(Optimism)} \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[(\bar{Z}_h^k \bar{r}_h^k - Z_h^{\boldsymbol{f}^*} r_h)(s_h^k, a_h^k, \tau_h^k) + (T_h^{\bar{P}_h^k, \bar{\boldsymbol{f}}_k} - T_h)(\cdot|s_h^k, a_h^k, \tau_h^k)^T \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \,\Big|\, \mathcal{F}_{k-1} \Big] \\
& && \text{(Corollary E.5)} \\
&\le \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\Big[Z_h^{\boldsymbol{f}^*} \big(\bar{r}_h^k - r_h\big) \Big](s_h^k, a_h^k, \tau_h^k) \,\Big|\, F_{k-1} \Big]}_{(i)} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\Big(\bar{Z}_h^k - Z_h^{\boldsymbol{f}^*} \Big) \big(\bar{r}_h^k(s_h^k, a_h^k) + \bar{P}_h^k(\cdot|s_h^k, a_h^k) \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \big) \,\Big|\, F_{k-1} \Big]}_{(ii)} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[Z_h^{\boldsymbol{f}^*} \big[\bar{P}_h^k - P_h \big](\cdot|s_h^k, a_h^k, \tau_h^k)^T \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \,\Big|\, F_{k-1} \Big]}_{(iii)} && \text{(Lemma E.3)}
\end{aligned}
$$

Notice that in the application of Lemma E.4, which was applied w.r.t. π^k, we used the fact that any DCMDP can be represented as an MDP whose history was embedded into the state. We now bound each of the terms of the decomposition.

Reward error

$$
\begin{aligned}
(i) &= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\left[\sum_{i=1}^{M+1} z_{i,h}^k \big(\bar{r}_{i,h}^k(s_h^k, a_h^k) - r_{i,h}(s_h^k, a_h^k) \big) \,\middle|\, F_{k-1} \right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\big[z_{i,h}^k \big| \bar{r}_{i,h}^k(s_h^k, a_h^k) - r_{i,h}(s_h^k, a_h^k) \big| \,\big|\, F_{k-1} \big]
\end{aligned}
$$

Latent features error

$$
\begin{aligned}
(ii) &\le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\Big\| \bar{Z}_h^k - Z_h^{\boldsymbol{f}^*} \Big\|_1 \Big\| \bar{r}_h^k(s_h^k, a_h^k) + \hat{P}_h^k(\cdot|s_h^k, a_h^k)^T \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \Big\|_\infty \,\Big|\, F_{k-1} \Big] && \text{(Hölder)} \\
&\le (c+1) H \sqrt{M+1} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\Big[\Big\| \bar{Z}_h^k - Z_h^{\boldsymbol{f}^*} \Big\|_1 \,\Big|\, F_{k-1} \Big]
\end{aligned}
$$

where the last inequality is since the optimistic value is bounded in $[0, H]$ and the reward is in $[0, cH]$.

Transition error

$$
\begin{aligned}
(iii) &= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\left[\sum_{i=1}^{M+1} z_{i,h}^k \Big(\Big(\hat{P}_h^k - P_h \Big)(\cdot|s_h^k, a_h^k)^T \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \Big) \,\middle|\, F_{k-1} \right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\Big[z_{i,h}^k \Big\| \Big(\hat{P}_h^k - P_h \Big)(\cdot|s_h^k, a_h^k) \Big\|_1 \big\| \bar{V}_{h+1}^k(\cdot, \tau_{h+1}^k) \big\|_\infty \,\Big|\, F_{k-1} \Big] && \text{(Hölder)} \\
&\le H \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1} \mathbb{E}\Big[z_{i,h}^k \Big\| \Big(\hat{P}_h^k - P_h \Big)(\cdot|s_h^k, a_h^k) \Big\|_1 \,\Big|\, F_{k-1} \Big] && \text{(Boundedness)}
\end{aligned}
$$

Combining all bounds concludes the proof.

□

E.3. Visitation-Summation Lemmas

Lemma E.7 (Expected Cumulative Visitation Bound, Lemma 22, Efroni et al. (2020), adapted to DCMDPs)**.** *Let* $\{\mathcal{F}_k\}_{k=1}^K$ *be the natural filtration. Then, with probability greater than* $1-\delta$ *it holds that*

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\frac{z_{i,h}^k}{\sqrt{n_h^k(s_h^k,a_h^k,i)\vee 1}}\,\middle|\,F_{k-1}\right] &= \sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\,\middle|\,F_{k-1}\right]\\ &\le 18H^2\log\left(\frac{1}{\delta}\right)+2HS(M+1)A+4\sqrt{H^2S(M+1)AK}\\ &= \mathcal{O}\left(H\left(SMA+H\log\left(\frac{1}{\delta}\right)\right)+\sqrt{H^2SMAK}\right)\\ &= \tilde{\mathcal{O}}\left(\sqrt{H^2SMAK}\right)\end{aligned}$$

Proof. We start by rewriting the sum as follows:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\frac{z_{i,h}^k}{\sqrt{n_h^k(s_h^k,a_h^k,i)\vee 1}}\,\middle|\,F_{k-1}\right] &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\sum_{i=1}^{M+1}z_{i,h}^k\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,i)\vee 1}}\,\middle|\,F_{k-1}\right]\\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\mathbb{E}_{x_h^k\sim z_{i,h}^k}\left[\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\right]\,\middle|\,F_{k-1}\right]\\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\,\middle|\,F_{k-1}\right]\\ &= \sum_{k=1}^{K}\mathbb{E}\left[\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\,\middle|\,F_{k-1}\right],\end{aligned}$$

which proves the first equality. Now, defining $Y_k=\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}$, which is $\mathcal{F}_k$-measurable and bounded almost surely in $[0,H]$, we can apply Lemma 27 of (Efroni et al., 2021) and get that for any $\delta>0$, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{M+1}\mathbb{E}\left[\frac{z_{i,h}^k}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k))\vee 1}}\,\middle|\,F_{k-1}\right] &\le \left(1+\frac{1}{2H}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}+2(2H+1)^2\log\frac{1}{\delta}\\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}+18H^2\log\frac{1}{\delta}.\end{aligned}$$

Finally, observing that every time a context-state-action is visited, its count increases, we can bound the sum by

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}} &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{\mathbb{1}\{x_h^k=x,s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s_h^k,a_h^k,x_h^k)\vee 1}}\\
&\le \sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(1+\sum_{n=1}^{n_h^K(s,a,x)}\frac{1}{\sqrt{n}}\right)\\
&\le HS(M+1)A+\sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}2\sqrt{n_h^K(s,a,x)}\\
&\le HS(M+1)A+2\sqrt{HS(M+1)A\underbrace{\sum_{h=1}^{H}\sum_{x\in\mathcal{X}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}n_h^K(s,a,x)}_{=HK}} \quad \text{(Cauchy Schwartz)}\\
&= HS(M+1)A+2\sqrt{H^2S(M+1)AK}.
\end{aligned}$$

Substituting this bound concludes the proof. □

Lemma E.8 (Elliptical potential lemma, Abbasi-Yadkori et al. (2011))**.** *Let $\{x_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ such that $\|x_t\|_2 \le L$ for all $t\ge 1$ and let $\boldsymbol{V}_t=\lambda_I+\sum_{s=1}^{t-1}x_sx_s^T$. Then,*

$$\sum_{t=1}^{n}\min\left\{\|x_t\|^2_{\boldsymbol{V}_t^{-1}},1\right\}\le 2d\log\frac{\lambda d+nL^2}{\lambda d}$$

Corollary E.9. *Let $\{x_h^k\}_{k\ge1,h\in[H]}$ be a sequence in $\mathbb{R}^d$ such that $\|x_t\|_2 \le L$ for all k,h and let $\boldsymbol{V}_k=\lambda_I+\sum_{k'=1}^{k-1}\sum_{t=1}^{H}x_t^{k'}{x_t^{k'}}^T$. Then,*

$$\sum_{k'=1}^{k}\sum_{h=1}^{H}\left\|x_h^{k'}\right\|_{\boldsymbol{V}_{k'}^{-1}}\le\frac{\sqrt{2KH^2d\log\frac{\lambda d+kL^2}{\lambda d}}}{\max\{1,L/\sqrt{\lambda}\}}.$$

Proof. Define the matrices $\boldsymbol{V}_{k,h}=\sum_{k'=1}^{k-1}x_h^k{x_h^k}^T$; clearly, it holds that $\boldsymbol{V}_k\preceq\boldsymbol{V}_{k,h}$ for all k,h, and thus, by applying Lemma E.8 for each of these matrices, we get

$$\sum_{k'=1}^{k}\sum_{h=1}^{H}\min\left\{\left\|x_h^{k'}\right\|^2_{\boldsymbol{V}_{k'}^{-1}},1\right\}\le\left(\sum_{k'=1}^{k}\min\left\{\left\|x_h^{k'}\right\|^2_{\boldsymbol{V}_{k',h}^{-1}},1\right\}\right)\le\sum_{h=1}^{H}2d\log\frac{\lambda d+kL^2}{\lambda d}=2dH\log\frac{\lambda d+kL^2}{\lambda d}.$$

Also, notice that if $\|x\|_2\le L$, then $\|x\|^2_{\boldsymbol{V}_k^{-1}}\le\frac{L^2}{\lambda_{\min}(\boldsymbol{V}_k)}\le\frac{L^2}{\lambda}$, and thus

$$\left\|x_h^{k'}\right\|^2_{\boldsymbol{V}_{k'}^{-1}}\le\frac{\min\left\{\left\|x_h^{k'}\right\|^2_{\boldsymbol{V}_{k'}^{-1}},1\right\}}{\max\{1,L^2/\lambda\}}.$$

Finally, the desired result is achieved by the Cauchy-Schwartz inequality:

$$\begin{aligned}
\sum_{k'=1}^{k}\sum_{h=1}^{H}\left\|x_h^{k'}\right\|^2_{\boldsymbol{V}_{k'}^{-1}} &\le\sqrt{KH\sum_{k'=1}^{k}\sum_{h=1}^{H}\|x_h^{k'}\|^2_{\boldsymbol{V}_{k'}^{-1}}}\\
&\le\frac{\sqrt{KH\sum_{k'=1}^{k}\sum_{h=1}^{H}\min\left\{\|x_h^{k'}\|^2_{\boldsymbol{V}_{k'}^{-1}},1\right\}}}{\max\{1,L/\sqrt{\lambda}\}}\\
&\le\frac{\sqrt{2KH^2d\log\frac{\lambda d+kL^2}{\lambda d}}}{\max\{1,L/\sqrt{\lambda}\}}
\end{aligned}$$

Algorithm 4 Optimistic Threshold Planner for Logistic DCMDPs

require: Optimistic reward $\bar{r}$, estimated transition $\hat{P}$, and rectangular confidence set $\mathcal{B}_k(\delta)$ for $\hat{\boldsymbol{f}}_T$.
init: $\bar{V}_H(s, \boldsymbol{C}) \leftarrow 0$, for all $s, \boldsymbol{C} \in \mathcal{S} \times \boldsymbol{\mathcal{I}}^k$
for $h = H-1, \dots, 1$ **do**
  for each $s \in \mathcal{S}, \boldsymbol{C}_h \in \left\{\boldsymbol{C}(\boldsymbol{\sigma}(\tau_h; \hat{\boldsymbol{f}}_T)) : \tau_h \in \mathcal{H}_h\right\}$ **do**
    $\boldsymbol{C}_{h+1} := \alpha \boldsymbol{C}_h + \left[\boldsymbol{l}_h^k, \boldsymbol{u}_h^k\right]$
    $\bar{Q}_i(s, a, \boldsymbol{C}_h) = \bar{r}_i(s, a) + \mathbb{E}_{s' \sim \hat{P}_i(\cdot|s,a)}\left[\bar{V}_{h+1}(s', \boldsymbol{C}_{h+1})\right]$ // State-action optimistic value
    $\bar{V}_h(s, \boldsymbol{C}_h) = \min\left\{\max_{a \in \mathcal{A}, t \in \mathcal{T}(\bar{\boldsymbol{Q}})} \sum_{i=0}^{M} z_i\big(\boldsymbol{th}_t(\bar{\boldsymbol{Q}}, \boldsymbol{C}_h)\big)\bar{Q}_i(s, a, \boldsymbol{C}_h), H\right\}$ // Lemma 5.6
    $\bar{\pi}(s, \boldsymbol{C}_h) \in \arg\max_{a \in \mathcal{A}} \max_{t \in \mathcal{T}(\bar{\boldsymbol{Q}})} \sum_{i=0}^{M} z_i\big(\boldsymbol{th}_t(\bar{\boldsymbol{Q}}, \boldsymbol{C}_h)\big)\bar{Q}_i(s, a, \boldsymbol{C}_h)$
  end for
end for
Output $\bar{\pi}(s, \tau) = \bar{\pi}(s, \boldsymbol{C}(\boldsymbol{\sigma}(\tau)))$

□

F. Threshold Optimistic Planning

F.1. Proof of Threshold Optimism – Lemma 5.6

Lemma 5.6 (Threshold Optimism)**.** *Let $\boldsymbol{Q} \in \mathbb{R}^{M+1}$. For any $\boldsymbol{x} \in \mathbb{R}^{M+1}$ such that $x_i = 0$ define $f(\boldsymbol{x}) = \sum_{i=1}^{M+1} z_i(\boldsymbol{x}) Q_i$. Let $\boldsymbol{C} = [\boldsymbol{l}, \boldsymbol{u}] \subseteq \mathbb{R}^{M+1} \times \mathbb{R}^{M+1}$ and assume that $\boldsymbol{l} < \boldsymbol{u}$. Then, there exists $t \in \mathcal{T}(\boldsymbol{Q})$ such that $\boldsymbol{th}_t(\boldsymbol{Q}, \boldsymbol{C}) \in \arg\max_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x})$.*

Proof. For brevity, throughout the proof, we assume that $\eta = 1$, namely, $z_i(\boldsymbol{x}) = \frac{\exp(x_i)}{1+\sum_{m=1}^{M}\exp(x_m)}$. This has no impact on the proof, since one can always denote $\left[\tilde{\boldsymbol{l}}, \tilde{\boldsymbol{u}}\right] = [\eta \boldsymbol{l}, \eta \boldsymbol{u}]$ and follow the rest of the proof with the modified intervals.

Let $X^* = \arg\max_{\boldsymbol{x} \in \mathcal{R}} f(\boldsymbol{x})$. We start by showing that there exists at least one solution at the extreme points of $\mathcal{R}$. We then show that solutions at the extreme points have a thresholding behavior.

Part 1. We first show that $X^* \cap \text{ext}(\mathcal{R}) \neq \emptyset$, i.e., there exists $\boldsymbol{x}^* \in X^*$ that is an extreme point of the set $\mathcal{R}$. Note that f is continuous and $\mathcal{R}$ is a compact set, therefore X^* is nonempty.

Let $\boldsymbol{x}^* \in X^*$ and choose some $k \in [M]$. We show that by replacing x_k^* by either l_k or u_k, we get another solution at X^*. Repeatedly doing so for all $k \in [M]$ will lead to $\boldsymbol{x}^* \in \text{ext}(\mathcal{R})$ and conclude this part of the proof.

We now fix $x_1^*, \dots, x_{k-1}^*, x_{k+1}^*, \dots, x_M^*$ and study $f(\boldsymbol{x}^*)$ only as a function of x_k^*. We also use the convention, $x_0^* = 0$. Then,

$$
\begin{aligned}
f(\boldsymbol{x}^*) &= \sum_{i=0}^{M} z_i(\boldsymbol{x}^*) v_i \\
&= \sum_{i=0}^{M} \frac{\exp(x_i^*)}{\sum_{j=0}^{M}\exp\big(x_j^*\big)} v_i \\
&= \frac{\exp(x_k^*) v_k}{\sum_{j=0}^{M}\exp\big(x_j^*\big)} + \frac{\sum_{i \neq k}\exp(x_i^*) v_i}{\sum_{j=0}^{M}\exp\big(x_j^*\big)} \\
&= \frac{\exp(x_k^*) v_k}{\sum_{j=0}^{M}\exp\big(x_j^*\big)} + \left(1 - \frac{\exp(x_k^*)}{\sum_{j=0}^{M}\exp\big(x_j^*\big)}\right) \frac{\sum_{i \neq k}\exp(x_i^*) v_i}{\sum_{j \neq k}\exp\big(x_j^*\big)}.
\end{aligned}
$$

Denote $\lambda(x_k^*) = \frac{\exp(x_k^*)}{\sum_{j=0}^{M}\exp\left(x_j^*\right)}$, and $v_{\text{ref}} = \frac{\sum_{i\neq k}\exp(x_i^*)v_i}{\sum_{j\neq k}\exp\left(x_j^*\right)}$. Then,

$$f(\boldsymbol{x}^*) = \lambda(x_k^*)v_k + (1-\lambda(x_k^*))v_{\text{ref}}.$$

Note that, since we fixed $x_1^*, \ldots, x_{k-1}^*, x_{k+1}^*, \ldots, x_M^*$, then v_{ref} is constant (does not depend on x_k^*). Also, $\lambda(x_k^*)$ is a strictly monotonically increasing function in x_k^* and $f(\boldsymbol{x}^*)$ is linear in $\lambda(x_k^*)$. Hence $\max_{x_k^*} f(\boldsymbol{x}^*)$ is achieved either for $\tilde{x}_k^* = \arg\min_{x_k\in[l_k,u_k]}\lambda(x_k) = l_k$ or $\tilde{x}_k^* = \arg\max_{x_k\in[l_k,u_k]}\lambda(x_k) = u_k$. Denoting the solution that replaces x_k^* with the maximizer $\tilde{x}_k^*$ by $\tilde{\boldsymbol{x}}^*$, we get that $f(\tilde{\boldsymbol{x}}^*) \geq f(\boldsymbol{x}^*)$, but since $\boldsymbol{x}^* \in X^*$, so does $\tilde{\boldsymbol{x}}^* \in X^*$. Following this process for all $k \in [M]$ leads to an optimal $\tilde{\boldsymbol{x}}^* \in \text{ext}(\mathcal{R})$ and thus $X^* \cap \text{ext}(\mathcal{R}) \neq \emptyset$.

Part 2. For the next part of the proof, we show that there exists an optimal solution that is a threshold function. Without loss of generality, assume that $(v_1, \ldots, v_M)$ are sorted in ascending order, such that $v_1 \leq v_2 \leq \ldots \leq v_M$. Let $\boldsymbol{x}^* \in X^* \cap \text{ext}(\mathcal{R})$, and assume by contradiction there exists $i, j \in [M], i < j$, such that $x_i^* = u_i, x_j^* = l_j$ and $v_i < v_j$. Denote

$$\begin{aligned}
\epsilon_i &= \min\left\{x_i^* - \log\left(\exp(x_i^*) + \exp\left(x_j^*\right) - \exp(u_j)\right), x_i^* - l_i\right\} \\
\epsilon_j &= \log\left(\exp(x_i^*) + \exp\left(x_j^*\right) - \exp(x_i^* - \epsilon_i)\right) - x_j^*
\end{aligned}$$

and let $\tilde{\boldsymbol{x}} = \boldsymbol{x}^* - \epsilon_i \boldsymbol{e}_i + \epsilon_j \boldsymbol{e}_j$. Then, ϵ_i, ϵ_j enjoy the folllowing properties.

1. $\epsilon_i, \epsilon_j > 0$, since

$$\begin{aligned}
\epsilon_i &\geq x_i^* - \log\left(\exp(x_i^*) + \underbrace{\exp\left(x_j^*\right) - \exp(u_j)}_{>0}\right) > 0, \\
\epsilon_j &= \log\left(\exp(x_i^*) + \underbrace{\exp\left(x_j^*\right) - \exp(x_i^* - \epsilon_i)}_{>0}\right) - x_j^* > 0.
\end{aligned}$$

2. By definition, $\epsilon_i \leq x_i^* - l_i$ by definition, and $\epsilon_j \leq u_j - x_j^*$, since by substituting ϵ_i, we get

$$\begin{aligned}
\epsilon_j &\leq \log\left(\exp(x_i^*) + \exp\left(x_j^*\right) - \exp\left(x_i^* - \left[x_i^* - \log\left(\exp(x_i^*) + \exp\left(x_j^*\right) - \exp(u_j)\right)\right]\right)\right) - x_j^* \\
&= \log\left(\exp(x_i^*) + \exp\left(x_j^*\right) - \left[\exp(x_i^*) + \exp\left(x_j^*\right) - \exp(u_j)\right]\right) - x_j^* \\
&= u_j - x_j^*.
\end{aligned}$$

   In particular, given that $\epsilon_i, \epsilon_j > 0$, it implies that $l_i \leq x_i^* - \epsilon_i \leq u_i$ and $l_j \leq x_j^* + \epsilon_j \leq u_j$.

3. The total weight of i, j is preserved

$$\begin{aligned}
\exp(x_i^* - \epsilon_i) + \exp\left(x_j^* + \epsilon_j\right) &= \exp(x_i^* - \epsilon_i) + \left[\exp(x_i^*) + \exp\left(x_j^*\right) - \exp(x_i^* - \epsilon_i)\right] \\
&= \exp(x_i^*) + \exp\left(x_j^*\right)
\end{aligned}$$

Given these properties, $\tilde{\boldsymbol{x}}$ is a valid solution for which we have that

$$\begin{aligned}
f(\tilde{\boldsymbol{x}}) &= \frac{\exp(\tilde{x}_i)v_i}{1+\sum_{k=0}^{M}\exp(\tilde{x}_k)} + \frac{\exp(\tilde{x}_j)v_j}{1+\sum_{k=0}^{M}\exp(\tilde{x}_k)} + \frac{\sum_{k\neq i,j}\exp(\tilde{x}_k)v_k}{1+\sum_{k=1}^{M}\exp(\tilde{x}_k)} \\
&= \frac{\exp(\tilde{x}_i)v_i}{1+\sum_{k=0}^{M}\exp(x_k^*)} + \frac{\exp(\tilde{x}_j)v_j}{1+\sum_{k=0}^{M}\exp(x_k^*)} + \frac{\sum_{k\neq i,j}\exp(x_k^*)v_k}{1+\sum_{k=1}^{M}\exp(x_k^*)}. \qquad \text{(By property (3))}
\end{aligned}$$

Therefore,

$$f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*) = \frac{\exp(x_i^* - \epsilon_i)v_i + \exp\left(x_j^* + \epsilon_j\right)v_j - \exp(x_i^*)v_i - \exp\left(x_j^*\right)v_j}{1+\sum_{k=0}^{M}\exp(x_k^*)}.$$

Considering the numerator, we have that

$$\begin{aligned}
&\exp(x_i^* - \epsilon_i)v_i + \exp\big(x_j^* + \epsilon_j\big)v_j - \exp(x_i^*)v_i - \exp\big(x_j^*\big)v_j \\
&\quad = \exp(x_i^* - \epsilon_i)v_i + \big(\exp(x_i^*) + \exp\big(x_j^*\big) - \exp(x_i^* - \epsilon_i)\big)v_j - \exp(x_i^*)v_i - \exp\big(x_j^*\big)v_j \qquad \text{(By property (3))} \\
&\quad = \exp(x_i^* - \epsilon_i)v_i + \exp(x_i^*)v_j - \exp(x_i^* - \epsilon_i)v_j - \exp(x_i^*)v_i \\
&\quad = (\exp(x_i^*) - \exp(x_i^* - \epsilon_i))(v_j - v_i) \\
&\quad > 0,
\end{aligned}$$

where the inequality is since $\epsilon_i > 0$ and $v_i < v_j$. That is, $f(\tilde{\boldsymbol{x}}) > f(\boldsymbol{x}^*)$, in contradiction to $\boldsymbol{x} \in X^*$. To summarize, we prove that for any $\boldsymbol{x}^* \in X^* \cap \text{ext}(\mathcal{R})$, if $v_i < v_j$, then $x_i^* \leq x_j^*$, which corresponds to a thresholding function. All that is left is to prove that if $v_i = v_j = v$, there exists a solution $\boldsymbol{x}^* \in X^* \cap \text{ext}(\mathcal{R})$ such that either $x_i^* = u_i, x_j^* = u_j$ or $x_i^* = l_i, x_j^* = l_j$. To show this, we follow a similar path to the first part of the proof and write

$$\begin{aligned}
f(\boldsymbol{x}^*) &= \sum_{k=0}^{M} z_i(\boldsymbol{x}^*)v_i \\
&= \frac{\exp(x_i^*)v_i + \exp\big(x_j^*\big)v_j}{\sum_{k=0}^{M}\exp(x_k^*)} + \left(1 - \frac{\exp(x_i^*) + \exp\big(x_j^*\big)}{\sum_{k=0}^{M}\exp(x_k^*)}\right)\frac{\sum_{k\neq i,j}\exp(x_k^*)v_k}{\sum_{k\neq i,j}\exp(x_k^*)} \\
&= \frac{\exp(x_i^*) + \exp\big(x_j^*\big)}{\sum_{k=0}^{M}\exp(x_k^*)}v + \left(1 - \frac{\exp(x_i^*) + \exp\big(x_j^*\big)}{\sum_{k=0}^{M}\exp(x_k^*)}\right)\frac{\sum_{k\neq i,j}\exp(x_k^*)v_k}{\sum_{k\neq i,j}\exp(x_k^*)}.. \qquad (v_i = v_j = v)
\end{aligned}$$

Now, denoting $\lambda(x_i^*, x_j^*) = \frac{\exp(x_i^*)+\exp\big(x_j^*\big)}{\sum_{k=0}^{M}\exp\big(x_k^*\big)}$, and $v_{\text{ref}} = \frac{\sum_{k\neq i,j}\exp(x_k^*)v_k}{\sum_{k\neq i,j}\exp\big(x_k^*\big)}$, we can follow the exact same line of the proof as the first part, and conclude that there exist another solution $\tilde{\boldsymbol{x}}$. in which $\lambda(x_i^*, x_j^*)$ is either maximized or minimized – either $x_i^* = u_i, x_j^* = u_j$ or $x_i^* = l_i, x_j^* = l_j$.

This completes the proof.

□

G. Confidence Sets

We require the following quantities, used by Amani and Thrampoulidis (2021) or adapted from Abeille et al. (2021). First, recall that for any $\boldsymbol{f} \in \mathbb{R}^{M(M+1)SAH}$ and $\boldsymbol{d} \in \mathbb{R}^{MSAH}$, we have

$$\boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}) := \text{diag}(\boldsymbol{z}(\boldsymbol{d}, \boldsymbol{f})) - \boldsymbol{z}(\boldsymbol{d}, \boldsymbol{f})\boldsymbol{z}(\boldsymbol{d}, \boldsymbol{f})^T$$

Also, for any $\boldsymbol{f} \in \mathbb{R}^{M(M+1)SAH}$, define

$$\boldsymbol{g}_k(\boldsymbol{f}) := \lambda \boldsymbol{f} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} z_i(\boldsymbol{d}_h^{k'}, \boldsymbol{f}) \otimes \boldsymbol{d}_h^{k'}, \qquad \text{and} \qquad \boldsymbol{H}_k(\boldsymbol{f}) := \lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{A}(\boldsymbol{d}_h^{k'}, \boldsymbol{f}) \otimes \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'^T}.$$

Note that by definition

$$\nabla_{\boldsymbol{f}} \mathcal{L}_\lambda^k(\boldsymbol{f}) = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{m}_h^{k'} \otimes \boldsymbol{d}_h^{k'} - \boldsymbol{g}_k(\boldsymbol{f}) \qquad \text{and} \qquad \nabla_{\boldsymbol{f}}^2 \mathcal{L}_\lambda^k(\boldsymbol{f}) = -\boldsymbol{H}_k(\boldsymbol{f}) \tag{11}$$

Confidence Set

$$\mathcal{C}_k(\delta) := \left\{ \boldsymbol{f} \in \mathcal{F} : \left\| g_k(\boldsymbol{f}) - g_k(\hat{\boldsymbol{f}}_t) \right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f})} \leq \beta_k(\delta) \right\}, \tag{12}$$

where $\beta_k(\delta) = \frac{M^{3/2}(M+1)SAH}{\sqrt{\lambda}} \left(\log\left(1 + \frac{k}{(M+1)SA\lambda}\right) + 2\log\left(\frac{2}{\delta}\right) \right) + \sqrt{\frac{\lambda}{4M}} + \sqrt{\lambda} L.$

Other Notations For any $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M(M+1)SAH}$ and $\boldsymbol{d} \in \mathbb{R}^{MSAH}$, define

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) &:= \int_0^1 \boldsymbol{A}(\boldsymbol{d}, v\boldsymbol{f}_1 + (1-v)\boldsymbol{f}_2) dv, \\
\tilde{\boldsymbol{B}}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) &:= \int_0^1 (1-v)\boldsymbol{A}(\boldsymbol{d}, v\boldsymbol{f}_1 + (1-v)\boldsymbol{f}_2) dv, \\
\boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2) &:= \lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{B}(\boldsymbol{d}_h^k, \boldsymbol{f}_1, \boldsymbol{f}_2) \otimes \boldsymbol{d}_h^k \boldsymbol{d}_h^{k^T}, \\
\tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1, \boldsymbol{f}_2) &:= \lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \tilde{\boldsymbol{B}}(\boldsymbol{d}_h^k, \boldsymbol{f}_1, \boldsymbol{f}_2) \otimes \boldsymbol{d}_h^k \boldsymbol{d}_h^{k^T}, \\
\boldsymbol{V}_k &:= \lambda \boldsymbol{I}_{(M+1)SAH} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'^T}.
\end{aligned}$$

Note that we stray from the notation of $\boldsymbol{V}_k$ in (Amani and Thrampoulidis, 2021), by removing the factor of κ from the regularization term.

G.1. Useful Lemmas

We now provide a list of lemmas required for providing confidence intervals for the logistic history dependent transition model in Appendices C and D. In what follows we will use the following expression:

$$d_2(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) \triangleq \left\| (\boldsymbol{f}_1 - \boldsymbol{f}_2)^T \boldsymbol{d} \right\|_2 \leq \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_2 \|\boldsymbol{d}\|_2 \leq L, \tag{13}$$

To this end, the following properties hold:

Lemma G.1 (Amani and Thrampoulidis (2021), Lemma 2)**.** *For any* $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M(M+1)SAH}$ *and* $\boldsymbol{d} \in \mathbb{R}^{MSAH}$

$$z(\boldsymbol{d}, \boldsymbol{f}_1) - z(\boldsymbol{d}, \boldsymbol{f}_2) = [\boldsymbol{B}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) \otimes \boldsymbol{d}](\boldsymbol{f}_1 - \boldsymbol{f}_2)$$

Lemma G.2 (Amani and Thrampoulidis (2021), Lemma 3)**.** *For any* $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M(M+1)SAH}$,

$$\boldsymbol{g}_k(\boldsymbol{f}_1) - \boldsymbol{g}_k(\boldsymbol{f}_2) = \boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)(\boldsymbol{f}_1 - \boldsymbol{f}_2)$$

Lemma G.3 (Amani and Thrampoulidis (2021), Lemma 4)**.** *For any* $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathcal{F}$*, it holds that* $(1+2L)^{-1}\boldsymbol{H}_k(\boldsymbol{f}_1) \preceq \boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)$ *and* $(1+2L)^{-1}\boldsymbol{H}_k(\boldsymbol{f}_2) \preceq \boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)$.

Lemma G.4 (Amani and Thrampoulidis (2021), Lemma 5)**.** *For any* $\boldsymbol{f} \in \mathbb{R}^{M(M+1)SAH}$ *and* $\boldsymbol{d} \in \mathbb{R}^{MSAH}$*, the matrix* $\boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f})$ *is strictly diagonally dominant and thus positive definite.*

Lemma G.5 (Amani and Thrampoulidis (2021), Theorem 1)**.** *Let* $\delta \in (0,1)$*. With probability at least* $1-\delta$*, for all* $k \geq 1$*, it holds that* $\boldsymbol{f}^* \in \mathcal{C}_k(\delta)$.

Remark G.6. Notice that Lemma G.4 also implies that all matrices $\boldsymbol{B}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2), \tilde{\boldsymbol{B}}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2), \boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2), \tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)$ are positive definite. More over, it implies that $\boldsymbol{B}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) \succeq \tilde{\boldsymbol{B}}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2)$ (since $1 - v \in [0,1]$), and therefore, $\boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2) \succeq \tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)$.

Relying on our different definition for V_k, the next lemma allows us to gain dependence on the derivative at the *real* latent features $\boldsymbol{f}^*$, instead of the worst-case derivative as in (Amani and Thrampoulidis, 2021).

Lemma G.7 (Connection between Local and Global Design Matrices)**.** *For any* $\boldsymbol{f} \in \mathcal{F}$*, denote* $\kappa(\boldsymbol{f}) = \frac{1}{\inf_{\boldsymbol{d} \in \mathcal{D}} \lambda_{\min}\{\boldsymbol{A}(\boldsymbol{d};\boldsymbol{f})\}}$*. It holds that*

$$\kappa(\boldsymbol{f})\boldsymbol{H}_k(\boldsymbol{f}) \succeq \boldsymbol{I_M} \otimes \boldsymbol{V}_k \ ,$$

and specifically for $f^* \in \mathcal{F}$*,*

$$\kappa \boldsymbol{H}_k(\boldsymbol{f}^*) \succeq \boldsymbol{I_M} \otimes \boldsymbol{V}_k \ .$$

Proof. For clarity, we state the dimension of the identity matrices throughout the proof. Throughout the analysis, recall that if $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \succeq 0$ and $\boldsymbol{A} \succeq \boldsymbol{B}$ then $\boldsymbol{A} \otimes \boldsymbol{C} \succeq \boldsymbol{B} \otimes \boldsymbol{C}$. For any $\boldsymbol{f} \in \mathcal{F}$,

$$\begin{aligned}
\boldsymbol{H}_k(\boldsymbol{f}) &= \lambda \boldsymbol{I}_{M(M+1)SAH} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \boldsymbol{A}(\boldsymbol{d}_h^{k'}, \boldsymbol{f}) \otimes \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T} \\
&\succeq \lambda \boldsymbol{I}_{M(M+1)SAH} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \lambda_{\min}\Big\{\boldsymbol{A}(\boldsymbol{d}_h^{k'}; \boldsymbol{f})\Big\} \boldsymbol{I}_{M+1} \otimes \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T} \\
&\succeq \lambda \boldsymbol{I}_{M(M+1)SAH} + \left(\inf_{\boldsymbol{d} \in \mathcal{D}} \lambda_{\min}\{\boldsymbol{A}(\boldsymbol{d}; \boldsymbol{f})\}\right) \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \boldsymbol{I}_{M+1} \otimes \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T} \\
&= \lambda \boldsymbol{I}_{M(M+1)SAH} + \frac{1}{\kappa(\boldsymbol{f})} \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \boldsymbol{I}_{M+1} \otimes \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T} \\
&= \frac{1}{\kappa(\boldsymbol{f})} \boldsymbol{I}_M \otimes \left(\kappa(\boldsymbol{f})\lambda \boldsymbol{I}_{(M+1)SAH} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T}\right) \\
&\overset{(*)}{\succeq} \frac{1}{\kappa(\boldsymbol{f})} \boldsymbol{I}_M \otimes \left(\lambda \boldsymbol{I}_{(M+1)SAH} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H} \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'T}\right) \\
&= \frac{1}{\kappa(\boldsymbol{f})} \boldsymbol{I_M} \otimes \boldsymbol{V}_k \ ,
\end{aligned}$$

where $(*)$ holds since $\kappa(\boldsymbol{f}) \geq 1$ (see, e.g., eq. (29) of Amani and Thrampoulidis (2021) when fixing the set of possible parameters $\mathcal{F}$ to be a singleton $\mathcal{F} = \{\boldsymbol{f}\}$).

Finally, we conclude the proof by noting that for $\boldsymbol{f}^* \in \mathcal{F}$, it holds by definition that $\kappa = \kappa(\boldsymbol{f}^*)$. □

Lemma G.8. *For all* $\boldsymbol{d} \in \mathbb{R}^{MSAH}$ *such that* $\|\boldsymbol{d}\|_2 \leq 1$*,* $k \geq 1$ *and* $\boldsymbol{f} \in \mathcal{C}_k(\delta)$*, if* $\boldsymbol{f}^* \in \mathcal{C}_k(\delta)$ *then*

$$\|z(\boldsymbol{d}, \boldsymbol{f}) - z(\boldsymbol{d}, \boldsymbol{f}^*)\|_2 \leq 2\beta_k(\delta)(1+2L)\sqrt{\kappa}\, \|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}$$

Proof. Here, we closely follow the proof of Lemma 1 in (Amani and Thrampoulidis, 2021), with the exception that we apply Lemma G.7 to achieve dependence on κ. Specifically, we let $\kappa_{\max} := \sup_{\boldsymbol{d}t\in\mathcal{D},\boldsymbol{f}\in\mathcal{F}} \lambda_{\max}\{\boldsymbol{A}(\boldsymbol{d};\boldsymbol{f})\}$. Notice that following (Amani and Thrampoulidis, 2021, Section 3), it holds that $\kappa_{\max} \leq 1$.

$$
\begin{aligned}
&\|z(\boldsymbol{d},\boldsymbol{f}) - z(\boldsymbol{d},\boldsymbol{f}^*)\|_2 \\
&= \|[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}](\boldsymbol{f}^*-\boldsymbol{f})\| && \text{(Lemma G.1)}\\
&= \left\|[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}]\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\boldsymbol{G}_k^{1/2}(\boldsymbol{f}^*,\boldsymbol{f})(\boldsymbol{f}^*-\boldsymbol{f})\right\| \\
&\leq \left\|[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}]\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\right\| \|\boldsymbol{f}^*-\boldsymbol{f}\|_{\boldsymbol{G}_k(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(Cauchy-Schwartz)}\\
&= \left\|[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}]\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\right\| \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(Lemma G.2)}\\
&= \sqrt{\lambda_{\max}\left([\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}]\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\\
&= \sqrt{\lambda_{\max}\left(\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\left[\boldsymbol{B}^T(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}\right]\left[\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}^T\right]\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\\
&&& \text{(cyclic property of } \lambda_{\max})\\
&= \sqrt{\lambda_{\max}\left(\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\left[\boldsymbol{B}^T(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\boldsymbol{B}(\boldsymbol{d},\boldsymbol{f}^*,\boldsymbol{f})\otimes\boldsymbol{d}\boldsymbol{d}^T\right]\boldsymbol{G}_k^{-1/2}(\boldsymbol{f}^*,\boldsymbol{f})\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\\
&&& \text{(mixed-product property)}\\
&\leq \kappa_{\max}\sqrt{\lambda_{\max}\left(\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})\left[\boldsymbol{I_M}\otimes\boldsymbol{d}\boldsymbol{d}^T\right]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(definition of } \kappa_{\max})\\
&= \sqrt{\lambda_{\max}\left(\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})[\boldsymbol{I_M}\otimes\boldsymbol{d}]\left[\boldsymbol{I_M}\otimes\boldsymbol{d}^T\right]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(cyclic property of } \lambda_{\max}\text{, and } \kappa_{\max}\leq 1)\\
&= \sqrt{\lambda_{\max}\left([\boldsymbol{I_M}\otimes\boldsymbol{d}]\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})\left[\boldsymbol{I_M}\otimes\boldsymbol{d}^T\right]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(mixed-product property)}\\
&\leq \sqrt{1+2L}\sqrt{\lambda_{\max}\left([\boldsymbol{I_M}\otimes\boldsymbol{d}]\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)\left[\boldsymbol{I_M}\otimes\boldsymbol{d}^T\right]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(Lemma G.3)}\\
&\leq \sqrt{\kappa(1+2L)}\sqrt{\lambda_{\max}\left([\boldsymbol{I_M}\otimes\boldsymbol{d}]\left[\boldsymbol{I_M}\otimes\boldsymbol{V}_k^{-1}(\boldsymbol{f}^*)\right]\left[\boldsymbol{I_M}\otimes\boldsymbol{d}^T\right]\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(Lemma G.7)}\\
&= \sqrt{\kappa(1+2L)}\sqrt{\lambda_{\max}\left(\boldsymbol{I_M}\otimes\|\boldsymbol{d}\|^2_{\boldsymbol{V}_k^{-1}}\right)}\, \|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})} && \text{(mixed-product property)}\\
&= \sqrt{\kappa(1+2L)}\,\|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}\|g_k(\boldsymbol{f}^*)-g_k(\boldsymbol{f})\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\\
&= \sqrt{\kappa(1+2L)}\,\|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}\left\|g_k(\boldsymbol{f}^*)-g_k(\hat{\boldsymbol{f}})+g_k(\hat{\boldsymbol{f}})-g_k(\boldsymbol{f})\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\\
&= \sqrt{\kappa(1+2L)}\,\|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}\left[\left\|g_k(\boldsymbol{f}^*)-g_k(\hat{\boldsymbol{f}})\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}+\left\|g_k(\hat{\boldsymbol{f}})-g_k(\boldsymbol{f})\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}^*,\boldsymbol{f})}\right]\\
&\leq (1+2L)\sqrt{\kappa}\,\|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}\left[\left\|g_k(\boldsymbol{f}^*)-g_k(\hat{\boldsymbol{f}})\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)}+\left\|g_k(\hat{\boldsymbol{f}})-g_k(\boldsymbol{f})\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)}\right] && \text{(Lemma G.3)}\\
&\leq 2\beta_k(\delta)(1+2L)\sqrt{\kappa}\,\|\boldsymbol{d}\|_{\boldsymbol{V}_k^{-1}}. && (\boldsymbol{f},\boldsymbol{f}^*\in\mathcal{C}_k(\delta))
\end{aligned}
$$

The *cyclic property* of $\lambda_{\max}$ refers to the fact that for two matrices $\boldsymbol{M}_1, \boldsymbol{M}_2$, the eigenvalues of $\boldsymbol{M}_1\boldsymbol{M}_2$ are the same as the eigenvalues of $\boldsymbol{M}_2\boldsymbol{M}_1$, and thus the same hold for the maximal eigenvalue.

□

Lemma G.9 (Adaptation of (Abeille et al., 2021), Lemma 8, to the multinomial case in Amani and Thrampoulidis (2021), Lemma 13)**.** *For any $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathcal{F}$, it holds that $(2+2L)^{-1}\boldsymbol{H}_k(\boldsymbol{f}_1) \preceq \tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)$ and $(2+2L)^{-1}\boldsymbol{H}_k(\boldsymbol{f}_2) \preceq \tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)$.*

Proof. According to Sun and Tran-Dinh (2019)[Eq. 16], for any $\boldsymbol{d} \in \mathbb{R}^{MSAH}$, $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M(M+1)SAH}$, and for any $v \in [0,1]$, we have that $\nabla^2 f(vx+(1-v)y) \succeq e^{-vd_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)}\nabla^2 f(y)$, where $d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)$ is defined in Equation (13).

Thus,

$$\int_0^1 (1-v)\nabla^2 f(vx+(1-v)y)dv \succeq \nabla^2 f(y)\int_0^1 (1-v)e^{-vd_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)}dv,$$

and replacing with the notation $\nabla^2 f(\boldsymbol{x}) \triangleq \boldsymbol{A}(\boldsymbol{d},\boldsymbol{x})$, we get

$$\tilde{\boldsymbol{B}}(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2) \triangleq \int_0^1 (1-v)\boldsymbol{A}(\boldsymbol{d},v\boldsymbol{f}_1+(1-v)\boldsymbol{f}_2)dv \succeq \boldsymbol{A}(\boldsymbol{d},\boldsymbol{f}_2)\int_0^1 (1-v)e^{-vd_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)}dv,$$

Integrating the RHS by parts,

$$\tilde{\boldsymbol{B}}(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2) \succeq \left(\frac{1}{d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)} + \frac{e^{-d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)}-1}{(d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2))^2}\right)\nabla^2 f(y) = g(d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2))\boldsymbol{A}(\boldsymbol{d},\boldsymbol{f}_2),$$

where we defined $g(z) \triangleq \frac{1}{z}\left(1+\frac{e^{-z}-1}{z}\right)$.

Next, by Abeille et al. (2021)[Lemma 10], for all $z \geq 0$, it holds that $g(z) \geq \frac{1}{2+z}$, and therefore,

$$\tilde{\boldsymbol{B}}(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2) \succeq (2+d_2(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2))^{-1}\boldsymbol{A}(\boldsymbol{d},\boldsymbol{f}_2)) \succeq (2+2L)^{-1}\boldsymbol{A}(\boldsymbol{d},\boldsymbol{f}_2), \tag{14}$$

where the last inequality follows is due to Equation (13).

Plugging in Equation (14) with the definition of $\tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)$,

$$\begin{aligned}
\tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1,\boldsymbol{f}_2) &= \lambda\boldsymbol{I} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\tilde{\boldsymbol{B}}(\boldsymbol{d}_h^k,\boldsymbol{f}_1,\boldsymbol{f}_2)\otimes \boldsymbol{d}_h^k{\boldsymbol{d}_h^k}^T \\
&\succeq (2+2L)^{-1}\left(\lambda\boldsymbol{I} + \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\boldsymbol{A}(\boldsymbol{d}_h^{k'},\boldsymbol{f}_2)\otimes \boldsymbol{d}_h^{k'}{\boldsymbol{d}_h^{k'}}^T\right) \\
&= (2+2L)^{-1}H_k(\boldsymbol{f}_2).
\end{aligned}$$

By the symmetry in $\boldsymbol{f}_1,\boldsymbol{f}_2$ in the definition of $\tilde{B}(\boldsymbol{d},\boldsymbol{f}_1,\boldsymbol{f}_2)$, we can similarly prove that $(2+2L)^{-1}H_k(\boldsymbol{f}_1) \succeq \tilde{\boldsymbol{G}}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)$

□

G.2. Convex Relaxation

Similar to Abeille et al. (2021), we define the convex relaxation of the set $\mathcal{C}_k(\delta)$ by

$$\mathcal{E}_k(\delta) = \left\{\boldsymbol{f}\in\mathcal{F} : \mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f}) \leq \xi^2(\delta)\right\} \qquad \text{where } \xi(\delta) = \beta_k(\delta) + \sqrt{\frac{HM}{\lambda}}\beta_k^2(\delta). \tag{15}$$

The next proposition is an adaptation of (Abeille et al., 2021)[Lemma 1] to the multinomial setting of (Amani and Thrampoulidis, 2021). Importantly, this proposition provides a confidence interval for the convex relaxation set Equation (15), which serves as the basis for the tractable estimator in Section 5.1.

Proposition G.10. *Let $\delta \in (0,1)$.*

1. *$\mathcal{C}_k(\delta) \subseteq \mathcal{E}_k(\delta)$ for all $k \geq 1$ and therefore, w.p. $1-\delta$, $\boldsymbol{f}^* \in \mathcal{E}_k(\delta)$ for all $k \geq 1$.*
2. *With probability $1-\delta$, it holds that*

$$\forall \boldsymbol{f}\in\mathcal{E}_k(\delta), \quad \|\boldsymbol{f}-\boldsymbol{f}^*\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \leq (2+2L)\beta_k(\delta) + \sqrt{2(1+L)}\xi_k(\delta).$$

In particular, for $\bar{\boldsymbol{f}} \in \arg\max_{\boldsymbol{f}\in\mathcal{F}}\mathcal{L}_\lambda^k(\boldsymbol{f})$, with probability $1-\delta$,

$$\left\|\bar{\boldsymbol{f}}-\boldsymbol{f}^*\right\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \leq (2+2L)\beta_k(\delta) + \sqrt{2(1+L)}\xi_k(\delta) \triangleq \gamma_k(\delta),$$

where $\gamma_k(\delta) := \left(2+2L+\sqrt{2(1+L)}\right)\beta_k(\delta) + \sqrt{\frac{2(1+L)HM}{\lambda}}\beta_k^2(\delta)$.

As in (Abeille et al., 2021), in order to prove Proposition G.10, we first require the following side-lemma:

Lemma G.11 (Counterpart of Abeille et al. (2021), Lemma 2)**.** *Let $\delta \in (0,1)$. For all $\boldsymbol{f} \in \mathcal{C}_k(\delta)$, it holds that*

$$\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\boldsymbol{f}^*)\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}})} \leq \xi_k(\delta).$$

Proof. First notice that by Remark G.6, the norm w.r.t. the inverse is well-defined and we further have $\lambda_{\min}(\boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2)) \geq \lambda$. Next, we utilise Amani and Thrampoulidis (2021), (eq. 61), which states that for any $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathcal{F}$, $\boldsymbol{d} \in \mathcal{D}$ (Recall that $\boldsymbol{B}$ is symmetric between $\boldsymbol{f}_1, \boldsymbol{f}_2$):

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{d}, \boldsymbol{f}_1, \boldsymbol{f}_2) &\succeq \left(1 + \left\| \left[\boldsymbol{f}_{11}^T \boldsymbol{d} - \boldsymbol{f}_{21}^T \boldsymbol{d}, \ldots, \boldsymbol{f}_{1M}^T \boldsymbol{d} - \boldsymbol{f}_{2M}^T \boldsymbol{d}\right] \right\| \right)^{-1} \boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}_1) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Amani and Thrampoulidis (2021), (eq. 61))} \\
&\succeq \left(1 + \|\mathbf{1}_M \otimes \boldsymbol{d}\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}_1,\boldsymbol{f}_2)} \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_{\boldsymbol{G}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)}\right)^{-1} \boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}_1) \qquad \text{(Cauchy-Schwartz)} \\
&\succeq \left(1 + \sqrt{\frac{HM}{\lambda}} \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_{\boldsymbol{G}_k(\boldsymbol{f}_1,\boldsymbol{f}_2)}\right)^{-1} \boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}_1), \qquad (\boldsymbol{G}_k(\boldsymbol{f}_1, \boldsymbol{f}_2) \succeq \lambda \boldsymbol{I})
\end{aligned}$$

where $j \in [1, ..., M]$, $\boldsymbol{f}_{ij}$ is the j-th coordinate of $\boldsymbol{f}_i$ and $\mathbf{1}_M \in \mathbb{R}^M$ is a vector of ones.

Thus, we can write for any $\boldsymbol{f} \in \mathcal{C}_k(\delta)$

$$\begin{aligned}
\boldsymbol{G}_k(\boldsymbol{f}, \hat{\boldsymbol{f}}_k) &= \lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{B}(\boldsymbol{d}_h^k, \boldsymbol{f}, \hat{\boldsymbol{f}}_k) \otimes \boldsymbol{d}_h^k {\boldsymbol{d}_h^k}^T \\
&\succeq \lambda \boldsymbol{I} + \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k\right\|_{\boldsymbol{G}_k(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right)^{-1} \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}) \otimes \boldsymbol{d}_h^k {\boldsymbol{d}_h^k}^T \qquad (\forall A, B \succ 0 \Rightarrow A \otimes B \succ 0) \\
&\succeq \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k\right\|_{\boldsymbol{G}_k(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right)^{-1} \left(\lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \boldsymbol{A}(\boldsymbol{d}, \boldsymbol{f}) \otimes \boldsymbol{d}_h^k {\boldsymbol{d}_h^k}^T\right) \\
&= \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k\right\|_{\boldsymbol{G}_k(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right)^{-1} \boldsymbol{H}_k(\boldsymbol{f}) \\
&= \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right)^{-1} \boldsymbol{H}_k(\boldsymbol{f}), \qquad \text{(Lemma G.2)}
\end{aligned}$$

Using this inequality, we get

$$\begin{aligned}
\left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|^2_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}}_k)} &\leq \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right) \left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|^2_{\boldsymbol{H}_k^{-1}(\boldsymbol{f})} \\
&\leq \left(1 + \sqrt{\frac{HM}{\lambda}} \left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}}_k)}\right) \beta_k^2(\delta). \quad (\boldsymbol{f} \in \mathcal{C}_k(\delta), \text{see Equation (12)})
\end{aligned}$$

Solving this inequality finally yields the desired result (see, e.g. Abeille et al., 2021, Proposition 7):

$$\left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|_{\boldsymbol{G}_k^{-1}(\boldsymbol{f},\hat{\boldsymbol{f}}_k)} \leq \beta_k(\delta) + \sqrt{\frac{HM}{\lambda}} \beta_k^2(\delta)$$

□

We are now ready to prove Proposition G.10.

Proof of Proposition G.10. **Part 1.** We start by writing the exact second-order Taylor expansion of the likelihood $\mathcal{L}_\lambda^k(\boldsymbol{f})$, which holds for any $\boldsymbol{f} \in \mathbb{R}^{M(M+1)SAH}$

$$\mathcal{L}_\lambda^k(\boldsymbol{f}) = \mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) + \nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k)^T(\boldsymbol{f} - \hat{\boldsymbol{f}}_k) + (\boldsymbol{f} - \hat{\boldsymbol{f}}_k)^T\left(\int_{v=0}^1 (1-v)\nabla_{\boldsymbol{f}}^2\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k + v(\boldsymbol{f} - \hat{\boldsymbol{f}}_k))dv\right)(\boldsymbol{f} - \hat{\boldsymbol{f}}_k).$$

Since $\hat{\boldsymbol{f}}_k$ is the solution to the unconstrained minimization of the concave likelihood $\mathcal{L}_\lambda^k(\boldsymbol{f})$, we have that $\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) = 0$. Recalling that $\nabla_{\boldsymbol{f}}^2\mathcal{L}_\lambda^k(f) = -H_k(f)$, we get

$$\begin{aligned}
\mathcal{L}_\lambda^k(\boldsymbol{f}) - \mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) &= \nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k)^T(\boldsymbol{f}^* - \hat{\boldsymbol{f}}_k) + (\boldsymbol{f} - \hat{\boldsymbol{f}}_k)^T\left(\int_{v=0}^1 (1-v)\nabla_{\boldsymbol{f}}^2\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k + v(\boldsymbol{f} - \hat{\boldsymbol{f}}_k))dv\right)(\boldsymbol{f} - \hat{\boldsymbol{f}}_k) \\
&= -(\boldsymbol{f} - \hat{\boldsymbol{f}}_k)^T\left(\int_{v=1}^1 (1-v)\boldsymbol{H}_k(\hat{\boldsymbol{f}}_k + v(\boldsymbol{f} - \hat{\boldsymbol{f}}_k))dv\right)(\boldsymbol{f} - \hat{\boldsymbol{f}}_k) \\
&= -\left\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k\right\|^2_{\tilde{\boldsymbol{G}}_k(\hat{\boldsymbol{f}}_k, \boldsymbol{f})} && \text{(Def. of } \tilde{\boldsymbol{G}}_k(\hat{\boldsymbol{f}}_k, \boldsymbol{f})) \\
&\geq -\left\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k)\right\|^2_{\boldsymbol{G}_k(\hat{\boldsymbol{f}}_k, \boldsymbol{f})} && (\tilde{\boldsymbol{G}}_k \preceq \boldsymbol{G}_k) \\
&= -\left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k))\right\|^2_{\boldsymbol{G}_k^{-1}(\hat{\boldsymbol{f}}_k, \boldsymbol{f})} && \text{(Lemma G.2)} \\
&= -\left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k))\right\|^2_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}, \hat{\boldsymbol{f}}_k)}. && (\boldsymbol{G}_k(\hat{\boldsymbol{f}}_k, \boldsymbol{f}) = \boldsymbol{G}_k(\boldsymbol{f}, \hat{\boldsymbol{f}}_k))
\end{aligned}$$

Rearranging, we get that for any $\boldsymbol{f} \in \mathbb{R}^{M(M+1)SAH}$,

$$\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f}) \leq \left\|\boldsymbol{g}_k(\boldsymbol{f}) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k))\right\|^2_{\boldsymbol{G}_k^{-1}(\boldsymbol{f}, \hat{\boldsymbol{f}}_k)},$$

and thus, the above inequality holds for any $\boldsymbol{f} \in \mathcal{C}_k(\delta) \subseteq \mathbb{R}^{M(M+1)SAH}$.

Finally, by Lemma G.11, for any $\boldsymbol{f} \in \mathcal{C}_k(\delta)$, we have that

$$\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f}) \leq \xi_k^2(\delta),$$

which implies that $\mathcal{C}_k(\delta) \subseteq \mathcal{E}_k(\delta)$ by the definition of Equation (15). In particular, by Lemma G.5, $\boldsymbol{f}^* \in \mathcal{C}_k(\delta)$ with probability at least $1 - \delta$, and thus the same holds for $\mathcal{E}_k(\delta)$.

Part 2. For this part, assume that $\boldsymbol{f}^* \in \mathcal{C}_k(\delta) \subseteq \mathcal{E}_k(\delta)$ for all $k \geq 1$, an event that holds with probability at least $1 - \delta$. Also, let $\boldsymbol{f} \in \mathcal{E}_k(\delta)$. Writing the Taylor expansion of $\mathcal{L}_\lambda^k(\boldsymbol{f})$, following the same derivation as the last part, we get

$$\begin{aligned}
\mathcal{L}_\lambda^k(\boldsymbol{f}) &= \mathcal{L}_\lambda^k(\boldsymbol{f}^*) + \nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) + (\boldsymbol{f} - \boldsymbol{f}^*)^T\left(\int_{v=1}^1 (1-v)\nabla_{\boldsymbol{f}}^2\mathcal{L}_\lambda^k(\boldsymbol{f}^* + v(\boldsymbol{f} - \boldsymbol{f}^*))dv\right)(\boldsymbol{f} - \boldsymbol{f}^*) \\
&= \mathcal{L}_\lambda^k(\boldsymbol{f}^*) + \nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) - \|\boldsymbol{f} - \boldsymbol{f}^*\|^2_{\tilde{\boldsymbol{G}}_k(\boldsymbol{f}^*, \boldsymbol{f})}. \\
&\leq \mathcal{L}_\lambda^k(\boldsymbol{f}^*) + \nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) - (2+2L)^{-1}\|\boldsymbol{f} - \boldsymbol{f}^*\|^2_{\boldsymbol{H}_k(\boldsymbol{f}^*)}. && \text{(Lemma G.9)}
\end{aligned}$$

Rearranging this inequality, we get,

$$\begin{aligned}
\|\boldsymbol{f} - \boldsymbol{f}^*\|^2_{\boldsymbol{H}_k(\boldsymbol{f}^*)} &\leq (2+2L)\big(\mathcal{L}_\lambda^k(\boldsymbol{f}^*) - \mathcal{L}_\lambda^k(\boldsymbol{f})\big) + (2+2L)\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) \\
&\leq (2+2L)\Big(\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f})\Big) + (2+2L)\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) && \text{(Def. of } \hat{\boldsymbol{f}}_k) \\
&\leq (2+2L)\xi_k^2(\delta) + (2+2L)\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) && (\boldsymbol{f} \in \mathcal{E}_k(\delta)) \\
&\leq (2+2L)\xi_k^2(\delta) + (2+2L)\|\boldsymbol{f} - \boldsymbol{f}^*\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)}\left\|\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)} && \text{(Cauchy-Schwartz)} \\
&\leq (2+2L)\xi_k^2(\delta) + (2+2L)\beta_k(\delta)\|\boldsymbol{f} - \boldsymbol{f}^*\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)}
\end{aligned}$$

where the last inequality is since

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*)\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)} &= \left\|\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\boldsymbol{f}^*) - \underbrace{\nabla_{\boldsymbol{f}}\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k)}_{=0}\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)} \\
&= \left\|\boldsymbol{g}_k(\boldsymbol{f}^*) - \boldsymbol{g}_k(\hat{\boldsymbol{f}}_k)\right\|_{\boldsymbol{H}_k^{-1}(\boldsymbol{f}^*)} && \text{(see Equation (11))} \\
&\le \beta_k(\delta). && (\boldsymbol{f}^* \in \mathcal{C}_k(\delta))
\end{aligned}$$

Thus, we have the inequality

$$\|\boldsymbol{f} - \boldsymbol{f}^*\|^2_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \le (2+2L)\xi_k^2(\delta) + (2+2L)\beta_k(\delta)\,\|\boldsymbol{f} - \boldsymbol{f}^*\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)},$$

which implies that (see, e.g. Abeille et al., 2021, Proposition 7)

$$\|\boldsymbol{f} - \boldsymbol{f}^*\|^2_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \le \sqrt{(2+2L)}\xi_k(\delta) + (2+2L)\beta_k(\delta).$$

To conclude the proof, notice that under the event that $\boldsymbol{f}^* \in \mathcal{C}_k(\delta)$ for all $k \ge 1$, we also have that $\boldsymbol{f}^* \in \mathcal{E}_k(\delta)$, and therefore, $\mathcal{E}_k(\delta)$ is not empty for all $k \ge 1$. Specifically, when the set is nonempty, by the definition of the set $\mathcal{E}_k(\delta)$ (Equation (15)), there exists a $f \in \mathcal{F}$ for which $\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f}) \le \xi_k^2(\delta)$.

Now, by definition, it holds for the constrained maximizer $\bar{\boldsymbol{f}}_k \in \arg\max_{\boldsymbol{f}\in\mathcal{F}} \mathcal{L}_\lambda^k(\boldsymbol{f})$ that for any $\boldsymbol{f} \in \mathcal{F}$, $\mathcal{L}_\lambda^k(\bar{\boldsymbol{f}}) \ge \mathcal{L}_\lambda^k(\boldsymbol{f})$. Consequently, for any $\boldsymbol{f} \in \mathcal{F}$, $\mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\bar{\boldsymbol{f}}_k) \le \mathcal{L}_\lambda^k(\hat{\boldsymbol{f}}_k) - \mathcal{L}_\lambda^k(\boldsymbol{f})$. Thus, when the set is nonempty, it must contain the constrained maximizer $\bar{\boldsymbol{f}}_k$. A direct conclusion of the previous inequality is that w.p. at least $1-\delta$, for all $k \ge 1$,

$$\left\|\bar{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \le (2+2L)\beta_k(\delta) + \sqrt{2(1+L)}\xi_k(\delta).$$

□

G.3. Local Confidence Bound

To prove the local confidence bound, we adapt the proofs of (Tennenholtz et al., 2022, Appendix K) to the multinomial case, while also taking into account the discounting of the latent features.

Next, we prove that the inverse of the Gram matrix of each episode is well behaved – its diagonal is bounded at any visited state-action-context.

Lemma G.12 (Inverse Eigenvalues Bound)**.** *Let $\boldsymbol{D}_k = \sum_{h=1}^H \boldsymbol{d}_h^{k'} \boldsymbol{d}_h^{k'^T}$ be the Gram matrix that corresponds to the discounted visitations during episode k. If $(s,a,x) \in \tau_h^k$ and $\boldsymbol{e}_{x,s,a,h} \in \mathbb{R}^{(M+1)SAH}$ is a unit vector in the coordinate (x,s,a,h), then $\boldsymbol{e}_{x,s,a,h}^T(\lambda\boldsymbol{I} + \boldsymbol{D}_k)^{-1}\boldsymbol{e}_{x,s,a,h} \le \frac{1}{\frac{1}{4H_\alpha}+\lambda}$.*

Proof. We closely follow the proof of (Tennenholtz et al., 2022, Lemma 7), while incorporating discount to the visitation vector. For brevity, and with some abuse of notations, we use $\boldsymbol{e}_n \in \mathbb{R}^{(M+1)SAH}$ to denote the unit vector in the n-th coordinate. In the following, we assume w.l.o.g. that the t-th coordinate of the vector $\boldsymbol{d}_h^k$ represents the state that was visited on the t-th time step (while unvisited states can be arbitrarily ordered). As done by (Tennenholtz et al., 2022, Lemma 7), this can be done using any permutation matrix $\boldsymbol{P}_k$ such that $\boldsymbol{e}_{x_t^k,,s_t^k,a_t^k,t} = \boldsymbol{P}_k\boldsymbol{e}_t$ for all $t \in [H]$. Then, denoting $\bar{\boldsymbol{e}}_t = H_\alpha^{-1/2}\sum_{n=1}^t \alpha^{t-n}\boldsymbol{e}_n = \big(\underbrace{\alpha^{t-1}, \alpha^{t-2}, \ldots, 1}_{t-\text{elements}}, 0, \ldots, 0\big)^T$, we can write $\boldsymbol{d}_t^k = H_\alpha^{-1/2}\sum_{n=1}^t \alpha^{t-n}\boldsymbol{e}_{x_n^k,s_n^k,a_n^k,n} = H_\alpha^{-1/2}\sum_{n=1}^t \alpha^{t-n}\boldsymbol{P}_k\boldsymbol{e}_n = \boldsymbol{P}_k\bar{\boldsymbol{e}}_t$.

Now, recalling that permutation matrices are orthogonal ($\boldsymbol{P}_k^{-1} = \boldsymbol{P}_k^T$) we can write

$$\begin{aligned}
\boldsymbol{e}_{x,s,a,h}^T(\lambda \boldsymbol{I} + \boldsymbol{D}_k)^{-1}\boldsymbol{e}_{x,s,a,h} &= \boldsymbol{e}_{x,s,a,h}^T\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \boldsymbol{d}_t^{k'}\boldsymbol{d}_t^{k'\,T}\right)^{-1}\boldsymbol{e}_{x,s,a,h} \\
&= \boldsymbol{e}_{x,s,a,h}^T\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \boldsymbol{P}_k\bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T\boldsymbol{P}_k^T\right)^{-1}\boldsymbol{e}_{x,s,a,h} \\
&= \boldsymbol{e}_{x,s,a,h}^T\left(\boldsymbol{P}_k\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T\right)\boldsymbol{P}_k^T\right)^{-1}\boldsymbol{e}_{x,s,a,h} \\
&= \boldsymbol{e}_{x,s,a,h}^T\boldsymbol{P}_k\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T\right)^{-1}\boldsymbol{P}_k^T\boldsymbol{e}_{x,s,a,h} \\
&= \boldsymbol{e}_h^T\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T\right)^{-1}\boldsymbol{e}_h \ .
\end{aligned}$$

Next, notice that $\lambda \boldsymbol{I} + \sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T$ is a block-diagonal matrix, whose first block is of size $H \times H$ (and the rest of the matrix is fully diagonal). We denote this first block of $\sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T$ by $\boldsymbol{C}$. For block-diagonal matrices, each block can be inverted independently of the other blocks, and for any coordinate $h \in [H]$, if $\boldsymbol{u}_h \in \mathbb{R}^H$ is the unit vector at coordinate h, we thus have

$$\begin{aligned}
\boldsymbol{e}_{x,s,a,h}^T(\lambda \boldsymbol{I} + \boldsymbol{D}_k)^{-1}\boldsymbol{e}_{x,s,a,h} &= \boldsymbol{e}_h^T\left(\lambda \boldsymbol{I} + \sum_{t=1}^{H} \bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T\right)^{-1}\boldsymbol{e}_h \\
&= \boldsymbol{u}_h^T(\lambda \boldsymbol{I} + \boldsymbol{C})^{-1}\boldsymbol{u}_h \\
&\leq \frac{\|\boldsymbol{u}_h\|_2^2}{\lambda_{\min}(\lambda \boldsymbol{I} + \boldsymbol{C})} \\
&= \frac{1}{\lambda + \lambda_{\min}(\boldsymbol{C})}
\end{aligned} \tag{16}$$

In the rest of the proof, we focus on bounding $\lambda_{\min}(\boldsymbol{C})$. First, observe that for any $t \in [H]$, we have

$$(\bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T)(i,j) = \begin{cases} H_\alpha^{-1}\alpha^{2t-i-j} & i \leq t, j \leq t \\ 0 & else \end{cases}$$

and thus

$$\boldsymbol{C}(i,j) = \sum_{t=1}^{H}(\bar{\boldsymbol{e}}_t\bar{\boldsymbol{e}}_t^T)(i,j) = H_\alpha^{-1}\sum_{t=\max\{i,j\}}^{H}\alpha^{2t-i-j} \ .$$

In particular, notice that for any $i < j$ (above diagonal), we have $\boldsymbol{C}(i,j) = \alpha\boldsymbol{C}(i+1,j)$, while for $i \geq j$ (below and on

diagonal), we have $\boldsymbol{C}(i,j) = \alpha \boldsymbol{C}(i+1,j) + \alpha^{i-j}$. Using this structure, we can calculate its inverse using diagonalization:

$$\left(\begin{array}{ccccc|ccccc} \sum_{t=1}^{H}\alpha^{2t-2} & \sum_{t=2}^{H}\alpha^{2t-3} & \sum_{t=3}^{H}\alpha^{2t-4} & \dots & \alpha^{H-1} & 1 & 0 & 0 & \dots & 0 \\ \sum_{t=2}^{H}\alpha^{2t-3} & \sum_{t=2}^{H}\alpha^{2t-4} & \sum_{t=3}^{H}\alpha^{2t-5} & \dots & \alpha^{H-2} & 0 & 1 & 0 & \dots & 0 \\ \sum_{t=3}^{H}\alpha^{2t-4} & \sum_{t=3}^{H}\alpha^{2t-5} & \sum_{t=3}^{H}\alpha^{2t-6} & \dots & \alpha^{H-3} & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{H-1} & \alpha^{H-2} & \alpha^{H-3} & \dots & 1 & 0 & 0 & 0 & \dots & 1 \end{array}\right)$$

$$= \left(\begin{array}{ccccc|ccccc} 1 & 0 & 0 & \dots & 0 & 1 & -\alpha & 0 & \dots & 0 \\ \alpha & 1 & 0 & \dots & 0 & 0 & 1 & -\alpha & \dots & 0 \\ \alpha^2 & \alpha & 1 & \dots & 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha^{H-1} & \alpha^{H-2} & \alpha^{H-3} & \dots & 1 & 0 & 0 & 0 & \dots & 1 \end{array}\right)$$

$$= \left(\begin{array}{ccccc|ccccc} 1 & 0 & 0 & \dots & 0 & 1 & -\alpha & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & -\alpha & 1+\alpha^2 & -\alpha & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & -\alpha & 1+\alpha^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 1+\alpha^2 \end{array}\right)$$

In the first relation, we subtracted α-times the $i+1$ rows from the i rows, while in the second one, we subtracted α-times the $i-1$ rows from the i rows. Thus, the inverse can be explicitly written as:

$$\boldsymbol{C}^{-1}_{i,j} = H_\alpha \begin{cases} 1 & i=j=1 \\ 1+\alpha^2 & i=j>1 \\ -\alpha & i=j-1 \text{ or } i=j+1 \\ 0 & \text{o.w.} \end{cases}$$

Notice that the absolute values of all rows is smaller than $H_\alpha(1+\alpha)^2 \le 4H_\alpha$. Then (e.g., by Gershgorin circle theorem), $\lambda_{\max}(\boldsymbol{C}^{-1}) \le 4H_\alpha$, and since B is PSD, $\lambda_{\min}(\boldsymbol{C}) \ge \frac{1}{4H_\alpha}$. The proof is concluded by substituting this result back into Equation (16). □

We are now ready to prove the local concentration results for the latent features of a DCMDP:

Lemma G.13 (Local Estimation Confidence Bound)**.** *Let $\hat{\boldsymbol{f}}^k \in \arg\max_{\boldsymbol{f}\in\mathcal{F}} \mathcal{L}^k_\lambda(\boldsymbol{f})$ be the maximum likelihood estimate of the features. Then, for any $\delta > 0$, with probability of at least $1-\delta$, for all $k \in [K], h \in [H], i \in [M]$ and $s,a,x \in \mathcal{S}\times\mathcal{A}\times\mathcal{X}$, it holds that*

$$\left|\hat{f}^k_{i,h}(s,a,x) - f^*_{i,h}(s,a,x)\right| \le \frac{2\gamma_k(\delta)\sqrt{\kappa H_\alpha}}{\sqrt{n^k_h(s,a,x) + 4\lambda H_\alpha}},$$

where $\gamma_k(\delta)$ is defined in Proposition G.10.

Proof. The proof follows Lemma 6 of (Tennenholtz et al., 2022).

For any $k \in [K], h \in [H], i \in [M]$ and $s,a,x \in \mathcal{S}\times\mathcal{A}\times\mathcal{X}$, let $\boldsymbol{e}_{x,s,a,h} \in \mathbb{R}^{(M+1)SAH}$ be a unit vector in the (x,s,a,h) coordinate and denote $\boldsymbol{f}_i \in \mathbb{R}^{(M+1)SAH}$ the latent features that correspond to a next latent state i. We start by bounding

$$\begin{aligned} \left|\hat{f}^k_{i,h}(s,a,x) - f^*_{i,h}(s,a,x)\right| &= \left|\left\langle \boldsymbol{e}_{x,s,a,h}, \hat{\boldsymbol{f}}_i - \boldsymbol{f}^*_i \right\rangle\right| \\ &\le \|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}} \left\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}^*_i\right\|_{\boldsymbol{V}_k} && \text{(Cauchy-Schwartz)} \\ &\le \|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}} \sqrt{\sum_{i'=1}^{M} \left\|\hat{\boldsymbol{f}}_{i'} - \boldsymbol{f}^*_{i'}\right\|^2_{\boldsymbol{V}_k}} \\ &= \|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}} \left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{I}_M\otimes\boldsymbol{V}_k}, && (17) \end{aligned}$$

We now turn our focus to bound $\|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}}$. Using the notation $\boldsymbol{D}_k$, as defined in Lemma G.12, we have

$$\begin{aligned}
\boldsymbol{V}_k^{-1} &= \left(\lambda \boldsymbol{I} + \sum_{k'=1}^{k-1} \boldsymbol{D}_{k'} \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\}\right)^{-1} \\
&\preceq \left(\sum_{k'=1}^{k-1} \left(\frac{\lambda}{n_h(s,a,x)} \boldsymbol{I} + \boldsymbol{D}_{k'}\right) \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\}\right)^{-1} \\
&\preceq \frac{1}{(n_h(s,a,x))^2} \sum_{k'=1}^{k-1} \left(\frac{\lambda}{n_h(s,a,x)} \boldsymbol{I} + \boldsymbol{D}_{k'}\right)^{-1} \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\} \ ,
\end{aligned}$$

where $n_h^k(s,a,x) = \sum_{k'=1}^{k-1} \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\}$ and the third transition is due to HM-AM inequality for positive matrices (Bhagwat and Subramanian, 1978). Next, we combine this result with Lemma G.12 and get

$$\begin{aligned}
\|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}}^2 &= \boldsymbol{e}_{x,s,a,h}^T \boldsymbol{V}_k^{-1} \boldsymbol{e}_{x,s,a,h} \\
&\leq \frac{1}{\left(n_h^k(s,a,x)\right)^2} \sum_{k'=1}^{k-1} \boldsymbol{e}_{x,s,a,h}^T \left(\frac{\lambda}{n_h(s,a,x)} \boldsymbol{I} + \boldsymbol{D}_{k'}\right)^{-1} \boldsymbol{e}_{x,s,a,h} \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\} \\
&\leq \frac{1}{(n_h(s,a,x))^2} \sum_{k'=1}^{k-1} \frac{1}{\frac{1}{4H_\alpha} + \frac{\lambda}{n_h(s,a,x)}} \mathbb{1}\left\{(s,a,x) \in \tau_h^{k'}\right\} && \text{(Lemma G.12)} \\
&= \frac{n_h(s,a,x)}{(n_h(s,a,x))^2} \frac{1}{\frac{1}{4H_\alpha} + \frac{\lambda}{n_h(s,a,x)}} \\
&= \frac{1}{\frac{n_h(s,a,x)}{4H_\alpha} + \lambda} \\
&= \frac{4H_\alpha}{n_h(s,a,x) + 4\lambda H_\alpha}.
\end{aligned}$$

By plugging into Equation (17), we obtain that for any k and any h, s, a

$$\begin{aligned}
\left|\hat{f}_{i,h}^k(s,a,x) - f_{i,h}^*(s,a,x)\right| &\leq \|\boldsymbol{e}_{x,s,a,h}\|_{\boldsymbol{V}_k^{-1}} \left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{I}_M \otimes \boldsymbol{V}_k} \\
&\leq \left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{I}_M \otimes \boldsymbol{V}_k} \frac{2\sqrt{H_\alpha}}{\sqrt{n_h(s,a,x) + 4\lambda H_\alpha}} \\
&\leq \left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \frac{2\sqrt{\kappa H_\alpha}}{\sqrt{n_h(s,a,x) + 4\lambda H_\alpha}}.
\end{aligned}$$

Finally by Proposition G.10, with probability $1-\delta$, for all $k \geq 1$, it holds that $\left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_{\boldsymbol{H}_k(\boldsymbol{f}^*)} \leq \gamma_k(\delta)$, and substituting this bound concludes the proof. □