Appendix to:

Distilled Thompson Sampling: Practical and Efficient Thompson Sampling via Imitation Learning

A Imitation learning with Wasserstein distances

When actions can be naturally embedded in a continuous space, we may want to measure closeness between the imitation and TS policy by incorporating the geometry of the actions taken by the respective policies. In this section, we provide an alternative instantiation of the abstract form of Algorithm 1 that uses Wasserstein distances as the notion of discrepancy $D\left(\cdot, \cdot \mid s\right)$ instead of the KL divergence. Our previous theoretical development for KL divergences has its analogues here, which we now briefly outline.

Given a metric $d(\cdot, \cdot)$ on $\mathcal{A}$, the Wasserstein distance between two distributions q^1 and q^2 on $\mathcal{A}$ is defined by the optimal transport problem

$$D_{\mathrm{w}}\left(q^1, q^2\right) = \inf_{\gamma \in \Gamma(q^1, q^2)} \mathbb{E}_{\gamma}[d(A, A')]$$

where $\Gamma(q^1, q^2)$ denotes the collection of all probabilities on $\mathcal{A} \times \mathcal{A}$ with marginals q^1 and q^2 (i.e., couplings). Intuitively, $D_{\mathrm{w}}\left(q^1, q^2\right)$ measures how much cost $d(A, A')$ is incurred by moving mass away from $A \sim q^1$ to $A' \sim q^2$ in an optimal fashion. Wasserstein distances encode the geometry of the underlying space $\mathcal{A}$ via the distance d. Unlike the KL divergence $D_{\mathrm{kl}}\left(q^1 \| q^2\right)$ that take value ∞ whenever q^1 has support not contained in q^2, the Wasserstein distance allows the imitation policy to have slightly different support than the Thompson sampling policy. For a discrete action space, $D_{\mathrm{w}}\left(\cdot, \cdot\right)$ can be defined with any symmetric matrix $d(a_i, a_j)$ satisfying $d(a_i, a_j) \geq 0$ with 0 iff $a_i = a_j$, and $d(a_i, a_j) \leq d(a_i, a_k) + d(a_k, a_j)$ for any $a_i, a_j, a_k \in \mathcal{A}$. As before, to simplify notation, we let

$$D_{\mathrm{w}}\left(\pi^1, \pi^2 \mid S\right) := D_{\mathrm{w}}\left(\pi^1(\cdot \mid S), \pi^2(\cdot \mid S)\right)$$

for two policies π^1 and π^2.

When Algorithm 1 is instantiated with the Wasserstein distance as its notion of discrepancy $D\left(\cdot, \cdot \mid S\right) = D_{\mathrm{w}}\left(\cdot, \cdot \mid S\right)$, the imitation learning problem (1) becomes

$$\underset{m \in \mathcal{M}}{\text{minimize}}\, \mathbb{E}_{S \sim \mathbb{P}_S}\left[D_{\mathrm{w}}\left(\pi^{\mathrm{TS}}, \pi^m \mid S\right)\right]. \tag{4}$$

To solve the above stochastic optimization problem, we can again use stochastic gradient descent methods, where the stochastic gradient $\nabla_m D_{\mathrm{w}}\left(\pi_t^{\mathrm{TS}}, \pi^m \mid S\right)$ can be computed by solving an optimal transport problem. From Kantorovich-Rubinstein duality (see, for example, [57]), we have

$$\begin{aligned} &D_{\mathrm{w}}\left(\pi_t^{\mathrm{TS}}, \pi^m \mid s\right) \\ &= \sup_{g: \mathcal{A} \to \mathbb{R}} \left\{\mathbb{E}_{A \sim \pi^{\mathrm{TS}}(\cdot \mid s)} g(a) - \mathbb{E}_{A \sim \pi^m(\cdot \mid s)} g(a) :\ g(a) - g(a') \leq d(a, a') \text{ for all } a, a' \in \mathcal{A}\right\}, \end{aligned} \tag{5}$$

where $d(\cdot, \cdot)$ is the metric on $\mathcal{A}$ used to define $D_{\mathrm{w}}\left(\cdot, \cdot\right)$. For discrete action spaces, the maximization problem (5) is a linear program with $O(|\mathcal{A}|)$ variables and constraints; for continuous action spaces, we can solve the problem over empirical distributions to approximate the optimal transport problem. We refer the interested reader to Peyré et al. [43] for a comprehensive introduction to computational methods for solving optimal transport problems.

Letting $g^\star$ denote the optimal solution to the dual problem (5), the envelope theorem (or Danskin's theorem)—see Bonnans and Shapiro [12, Theorem 4.13]—implies that under simple regularity conditions

$$\nabla_m D_{\mathrm{w}}\left(\pi_t^{\mathrm{TS}}, \pi^m \mid s\right) = -\nabla_m \mathbb{E}_{A \sim \pi^m(\cdot \mid s)}[g^\star(a)].$$

Assuming that an appropriate change of gradient and expectation is justified, we can use the policy gradient trick to arrive at

$$-\nabla_m \mathbb{E}_{A\sim\pi^m(\cdot|s)}[g^\star(A)] = -\mathbb{E}_{A\sim\pi^m(\cdot|s)}[g^\star(A)\nabla_m \log \pi^m(A \mid s)].$$

We conclude that for $A \sim \pi^m(\cdot \mid S_i)$,

$$-g^\star(A)\nabla_m \log \pi^m(A \mid S_i) \tag{6}$$

is a stochastic gradient for the imitation problem (4). As before, we can get lower variance estimates by average the above estimator over many actions $A \sim \pi^m(\cdot \mid S_i)$.

Using stochastic gradients (6), we can solve the imitation problem (4) efficiently. We now show that the resulting imitation policy admits a regret decomposition similar to Lemma 1 for KL divergences. As a direct consequence of this decomposition, the regret bounds in Section 4 have their natural analogues with Wasserstein distances replacing KL divergences as the notion of discrepancy, though we omit them for brevity.

Lemma 2. *Let $\pi = \{\pi_t\}_{t\in\mathbb{N}}$ be any set of policies, and let $U_t(\cdot; H_t, S_t) : \mathcal{A} \to \mathbb{R}$ be any upper confidence bound sequence that is measurable with respect to $\sigma(H_t, S_t, A_t)$. For some sequence $M_t(H_t, S_t)$ and a constant $L > 0$, let U_t satisfy*

$$|U_t(a; H_t, S_t) - U_t(a'; H_t, S_t)| \le L d(a, a') \text{ for all } a, a' \in \mathcal{A} \text{ almost surely.} \tag{7}$$

Then for all $T \in \mathbb{N}$,

$$\begin{aligned}\text{BayesRegret}\left(T, \{\pi_t\}_{t\in\mathbb{N}}\right) \le &\sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t)] + \sum_{t=1}^{T} \mathbb{E}[U_t(A_t; H_t, S_t) - f_\theta(A_t, S_t)] \\ &+ L\sum_{t=1}^{T} \mathbb{E}\left[D_{\text{w}}\left(\pi_t^{\text{TS}}, \pi_t \mid S_t\right)\right].\end{aligned} \tag{8}$$

where $D_{\text{w}}(\cdot, \cdot \mid \cdot)$ is the Wasserstein distance defined with the metric d in the condition (7)*.*

Proof Proof. The proof mirrors that of Lemma 2, but bound the differences (13) by $D_{\text{w}}\left(\pi_t^{\text{TS}}, \pi_t \mid S_t\right)$. By the Kantorovich dual representation (5), we have

$$\mathbb{E}[|U_t(A_t^{\text{TS}}; H_t, S_t) - U_t(A_t; H_t, S_t)| \mid H_t, S_t] \le M_t(H_t, S_t) D_{\text{w}}\left(\pi_t^{\text{TS}}, \pi_t \mid S_t\right).$$

Applying this bound in the decomposition (12), and taking expectation over (H_t, S_t) on both sides and summing $t = 1, \ldots, T$, we get the desired bound. □ □

B Contextual Gaussian processes

In this section, we consider the setting where the mean reward function is nonparametric and model it as a sample path of a Gaussian process. Formally, we assume that $(a, s) \mapsto f_\theta(a, s)$ is sampled from a Gaussian process on $\mathcal{A} \times \mathcal{S}$ with mean function $\mu(a, s)$ and covariance function (kernel)

$$k((a, s), (a', s')) := \mathbb{E}[(f_\theta(a, s) - \mu(a, s))(f_\theta(a', s') - \mu(a', s'))].$$

We assume that the decision maker observes rewards $R_t = f_\theta(A_t, S_t) + \epsilon_t$, where the noise $\epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are independent of everything else. Given these rewards, we are interested in optimizing the function $a \mapsto f_\theta(a, S_t)$ for each observed context S_t at time t. Modeling mean rewards as a Gaussian process is advantageous since we can utilize analytic formulae to update the posterior at each step. For large-scale applications, we can parameterize our kernels by a neural network and leverage the recently developed interpolations techniques to perform efficient posterior updates [61, 62, 63].

As before, we measure performance by using the Bayes regret, averaging outcomes over the prior P. We build on the UCB regret bound due to Srinivas et al. [51] and bound the first two terms in the Bayes regret decomposition (Lemma 1). In particular, we show that they can be controlled by the

maximal amount of information on the optimal action that can be gained after T time steps. Recall the definition of mutual information between two random vectors: $I(X,Y) := D_{\mathrm{kl}}\left(P_{X,Y} \| P_X \times P_Y\right)$. We define the maximal possible information gain after T time steps as

$$\gamma_T := \sup_{X \subseteq \mathcal{A}\times\mathcal{S}:|X|=T} I(y_X, f_X)$$

where $y_X = \{f_\theta(x)\}_{x\in X}$ and $f_X = \{f_\theta(x)\}_{x\in X}$. For popular Gaussian and Matern kernels, Srinivas et al. [51] has shown that the maximal information gain can be bounded explicitly; we summarize these bounds shortly.

Letting $\mathcal{A} \subseteq [0,r]^d$ for some $r > 0$, we show that the first two terms in the decomposition in Lemma 1 can be bounded by $O(d\gamma_t T \log T)$, thus bounding the Bayes regret up to the sum of imitation error terms. In the following, we use L_f to denote the (random) Lipschitz constant of the map $a \mapsto f_\theta(a,s)$

$$L_f := \sup_{s\in\mathcal{S}} \sup_{a,a'\in\mathcal{A}} \frac{|f_\theta(a,s) - f_\theta(a',s)|}{\|a-a'\|_1}.$$

Theorem 2. *Let $\mathcal{A} \subseteq [0,r]^d$ for some $r > 0$. Assume that*

$$c_1 := \sup_{a\in\mathcal{A},s\in\mathcal{S}} |\mu(a,s)| < \infty, \quad c_2 := \sup_{a,a'\in\mathcal{A},s,s'\in\mathcal{S}} k(a,a') < \infty,$$

and let $c_3 := \left\|\sup_{a\in\mathcal{A},s\in\mathcal{S}} |f_\theta(a,s)|\right\|_{2,P}$. If $\mathbb{E}[L_f^2] < \infty$, then there exists a universal constant $C > 1$ such that

$$\begin{aligned}\mathrm{BayesRegret}\,(T,\pi) \le C\mathbb{E}[L_f] + Cc_2 + Cd\log(rd)\left(c_1\sqrt{\mathbb{E}[L_f]} + c_3\sqrt{\mathbb{E}[L_f^2]}\right) \\ + \left(T\gamma_T \frac{d\log T + d\log rd}{\log(1+\sigma^{-2})}\right)^{1/2} + \sum_{t=1}^{T}(c_3 + Cc_2 d\log rdt)\sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]}.\end{aligned}$$

See Section D.4 for the proof.

To instantiate Theorem 2, it remains to bound smoothness of the reward function $\mathbb{E}[L_f^2]$, and the maximal information gain γ_T. Standard arguments from Gaussian process theory show $\mathbb{E}[L_f^2] < \infty$ holds whenever the mean and covariance function (kernel) is smooth, which holds for commonly used kernels.

Lemma 3 (Theorem 5, Ghosal et al. [24])**.** *If $\mu(\cdot)$ and $k(\cdot,\cdot)$ are 4 times continuously differentiable, then $(a,s) \mapsto f_\theta(a,s)$ is continuously differentiable and follows a Gaussian process again. In particular, $\mathbb{E}[L_f^2] < \infty$.*

To obtain concrete bounds on the maximal information gain γ_T, we use the results of Srinivas et al. [51], focusing on the popular Gaussian and Matern kernels

$$\begin{aligned}k_g(x,x') &:= \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right), \\ k_m(x,x') &:= \frac{2^{1-\nu}}{\Gamma(\nu)} r^\nu B_\nu(r) \;\text{ where } r = \frac{\sqrt{2\nu}}{l}\|x-x'\|,\end{aligned}$$

where we used $B(\cdot)$ and $\Gamma(\cdot)$ to denote the Besel and Gamma functions respectively. To ease notation, we let κ denote the dimension of the underlying space, and define

$$\mathfrak{M}(k_g,T) := (\log T)^{\kappa+1} \;\text{ and }\; \mathfrak{M}(k_m,T) := T^{\frac{\kappa^2+\kappa}{\kappa^2+\kappa+2\nu}}\log T.$$

We have the following bound on γ_T for Gaussian and Matern kernels; the bound is a direct consequence of Theorem 2, [32] and Theorem 5, [51].

Lemma 4. *Let $\mathcal{A} \subseteq \mathbb{R}^d$ and $\mathcal{S} \subseteq \mathbb{R}^{d'}$ be convex and compact. Let the kernel k be given by the sum of two kernels k_A and k_S on $\mathcal{A}$ and $\mathcal{S}$ respectively*

$$k((a,s),(a',s')) = k_A(a,a') + k_S(s,s').$$

If k_A and k_S are either the Gaussian kernel k_g or the Matern kernel k_m with $\nu > 1$, then

$$\gamma_T = O\left(\mathfrak{M}(k_A,T) + \mathfrak{M}(k_S,T) + \log T\right).$$

For example, taking $k_A = k_g$ and $k_S = k_g$, we conclude

$$\text{BayesRegret}\left(T, \{\pi_t\}_{t\in\mathbb{N}}\right) = O\left(\sqrt{dT(\log T)^{\max\{d,d'\}+1}} + d\sum_{t=1}^{T}\log(rdt)\sqrt{\mathbb{E}\left[D_{\text{kl}}\left(\pi_t^{\text{TS}}, \pi_t \mid S_t\right)\right]}\right).$$

C Generalization guarantees for imitation learning

So far, we showed that in order to achieve good regret, it suffices to control the KL-divergence between the imitation and (off-policy) Thompson sampling policy in order. We now show that each of these terms can be optimized efficiently using finite-sample approximations; we are interested in how well the model learned from an empirical approximation of the imitation problem (1) performs with respect to the true imitation objective (KL divergence). Since we consider the problem for any fixed time step t, we omit the subscript t and denote $\pi^{\text{TS}} = \pi_t^{\text{TS}}$. Recalling that N denotes the number of observed "unlabeled" contexts $S_1, \ldots, S_N$, we simulate N_a number of actions from the (off-policy) Thompson sampler $A_{ij}^{\text{TS}} \sim \pi_t^{TS}(\cdot \mid S_i)\ \ j = 1, \ldots, N_a$ for each context S_i.

Since the imitation learning objective $\mathbb{E}[D_{\text{kl}}\left(\pi^{\text{TS}}, \pi^m \mid S\right)]$ is proportional to $-\mathbb{E}_{S, A^{\text{TS}}\sim\pi^{\text{TS}}(\cdot|S)}[\log \pi^m(A^{\text{TS}} \mid S)]$, we are interested in solving the following empirical approximation to the population problem (1)

$$\widehat{m}_{N,N_a} \in \operatorname*{argmax}_{m\in\mathcal{M}} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\log \pi^m(A_{ij}^{\text{TS}} \mid S_i). \tag{9}$$

"Unlabeled" contexts without corresponding action-reward information are often cheap and abundant in internet applications, and we can take N to be very large. For any observed context S_i, the actions $A^{\text{TS}} \sim \pi^{\text{TS}}(\cdot \mid S_i)$ can be generated by posterior sampling. Since this can be done *offline*, and is trivial to parallelize per S_i, we can generate many actions; hence we usually have very large N_a as well.

To make our results concrete, we rely on standard notions of complexity to measure the size of the imitation model class $\mathcal{M}$, using familiar notions based on Rademacher averages. For a sample $\xi_1, \ldots, \xi_n$ and i.i.d. random signs (Rademacher variables) $\varepsilon_i \in \{-1, 1\}$ that are independent of the ξ_i's, the empirical Rademacher complexity of the class of functions $\mathcal{G} \subseteq \{g : \Xi \to \mathbb{R}\}$ is

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_\epsilon\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(\xi_i)\right].$$

For example, when $g \in \mathcal{G}$ takes values in $[-M, M]$ with VC-subgraph dimension d, a standard bound is $\mathbb{E}[\mathfrak{R}_n(\mathcal{G})] \lesssim M\sqrt{\frac{d}{n}}$; see Chapter 2 of van der Vaart and Wellner [56] and Bartlett and Mendelson [9] for a comprehensive treatment.

In what follows, we show that the empirical minimizer $\widehat{m}_{N,N_a}$ achieves good performance optimum with respect to the population problem (1). More concretely, we show that

$$\mathbb{E}\left[D_{\text{kl}}\left(\pi^{\text{TS}}, \pi^{\widehat{m}_{N,N_a}} \mid S\right)\right] \le \inf_{m\in\mathcal{M}}\mathbb{E}\left[D_{\text{kl}}\left(\pi^{\text{TS}}, \pi^m \mid S\right)\right] + N^{-1/2}\left(\mathfrak{T}_1(\mathcal{M}) + N_a^{-1/2}\mathfrak{T}_2(\mathcal{M})\right)$$

where $\mathfrak{T}_1(\mathcal{M})$ and $\mathfrak{T}_1(\mathcal{M})$ are problem-dependent constants that measure the complexity of the imitation model class $\mathcal{M}$. We show that the dominating dimension-dependent constant $\mathfrak{T}_1(\mathcal{M})$ term is in a sense the best one can hope for, matching the generalization guarantee available for the idealized scenario where KL-divergence $D_{\text{kl}}\left(\pi^{\text{TS}}, \pi^m \mid S_i\right)$ can be computed and optimized exactly for each S_i, $i = 1, \ldots, N$.

We begin by first illustrating this "best-case scenario", where we can generate an infinite number of actions (i.e. $N_a = \infty$). We consider the solution $\widehat{m}_{N,\infty}$ to the idealized empirical imitation learning problem where the KL divergence between the imitation policy and the Thompson sampler can be computed (and optimized) exactly. Formally, we let

$$\widehat{m}_{N,\infty} \in \operatorname*{argmin}_{m\in\mathcal{M}}\frac{1}{N}\sum_{i=1}^{N}D_{\text{kl}}\left(\pi^{\text{TS}}, \pi^m \mid S_i\right) = \operatorname*{argmax}_{m\in\mathcal{M}}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{A^{\text{TS}}\sim\pi^{\text{TS}}(\cdot|S_i)}[\log \pi^m(A^{\text{TS}} \mid S)].$$

The Rademacher complexity of the following set of functions controls generalization performance of $\widehat{m}_{N,\infty}$

$$\mathcal{G}_1 := \left\{ s \mapsto \mathbb{E}_{A^{\mathrm{TS}} \sim \pi^{\mathrm{TS}}(\cdot|s)}[\log \pi^m(A^{\mathrm{TS}} \mid s)] : m \in \mathcal{M} \right\}.$$

Lemma 5. *Let $\widehat{m}_{N,\infty}$ be defined as above. If $|\log \pi^m(a \mid s)| \leq M$ for all $a \in \mathcal{A}, s \in \mathcal{S}, m \in \mathcal{M}$, then with probability at least $1 - 2e^{-t}$*

$$\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{\widehat{m}_{N,\infty}} \mid S\right)\right] \leq \inf_{m \in \mathcal{M}} \mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^m \mid S\right)\right] + 4\mathbb{E}[\mathfrak{R}_N(\mathcal{G}_1)] + 2M\sqrt{\frac{2t}{N}}.$$

This lemma follows from a standard concentration argument, which we present in Section E.1 for completeness.

We now show that the empirical approximation (9) enjoys a similar generalization performance as $\widehat{m}_{N,\infty}$, so long as N_a is moderately large. To give our result, we define two additional sets of functions

$$\mathcal{G}_2(s) := \{a \mapsto \log \pi^m(a \mid s) : m \in \mathcal{M}\}$$
$$\mathcal{G}_3 := \{(a, s) \mapsto \log \pi^m(a \mid s) : m \in \mathcal{M}\}.$$

For $\mathcal{G}_3$, we abuse notation slightly and write

$$\mathfrak{R}_{NN_a}\mathcal{G}_3 := \mathbb{E}_\epsilon \left[\sup_{m \in \mathcal{M}} \frac{1}{N} \sum_{i=1}^{n} \frac{1}{N_a} \sum_{j=1}^{N_a} \varepsilon_{ij} \log \pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) \right]$$

for i.i.d. random signs ε_{ij}. The following lemma, whose proof we give in Section E.2, shows that the empirical solution (9) generalizes at a rate comparable to the idealized model $\widehat{m}_{N,\infty}$.

Theorem 3. *Let $|\log \pi^m(a \mid s)| \in M$ for all $a \in \mathcal{A}, s \in \mathcal{S}, m \in \mathcal{M}$. Then, with probability at least $1 - 3e^{-t}$,*

$$\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{\widehat{m}_{N,N_a}} \mid S\right)\right] \leq \inf_{m \in \mathcal{M}} \mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^m \mid S\right)\right] + 4\mathbb{E}[\mathfrak{R}_N(\mathcal{G}_1)] + 2M\sqrt{\frac{2t}{N}}$$
$$+ \sqrt{\frac{32t}{N}} \sup_{s \in \mathcal{S}} \mathbb{E}_{A_j^{\mathrm{TS}} \overset{\mathrm{iid}}{\sim} \pi^{\mathrm{TS}}(\cdot|s)}[\mathfrak{R}_{N_a}(\mathcal{G}_2(s))] + 8\mathbb{E}[\mathfrak{R}_{NN_a}(\mathcal{G}_3)]$$

Recalling the standard scaling $\mathbb{E}[\mathfrak{R}_n(\mathcal{G})] \lesssim M\sqrt{\frac{d}{n}}$, we see that $\widehat{m}_{N,N_a}$ achieves performance comparable to the idealized solution $\widehat{m}_{N,\infty}$, up to an $O(N^{-1/2}N_a^{-1/2})$-error term. Although we omit it for brevity, boundedness of $\log \pi^m(a \mid s)$ can be relaxed to sub-Gaussianity by using standard arguments (see, for example, Chapter 2.14 [56]).

We now provide an application of the theorem. **Example 1:** Let $\mathcal{V}$ be a vector space, and $V \subset \mathcal{V}$ be any collection of vectors in $\mathcal{V}$. Let $\|\cdot\|$ be a (semi)norm on $\mathcal{V}$. A collection $v_1, \ldots, v_N \subset \mathcal{V}$ is an *ϵ-cover* of V if for each $v \in V$, there exists v_i such that $\|v - v_i\| \leq \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is then

$$N(V, \epsilon, \|\cdot\|) := \inf\{N \in \mathbb{N} : \exists \text{ an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\|\}.$$

Letting $\mathcal{G}$ be a collection of functions $g : \mathcal{X} \to \mathbb{R}$, a standard argument due to Pollard [44] yields

$$\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \varepsilon_i g(\xi_i) \right] \lesssim \inf_{\delta \geq 0} \left\{ \delta + \frac{1}{\sqrt{n}} \sqrt{\log N(\mathcal{G}, \delta, \|\cdot\|_{L^2(P_n)})} \right\} \tag{10}$$

where P_n denotes the point masses on $\xi_1, \ldots, \xi_n$ and $\|\cdot\|_{L^2(P_n)}$ is the empirical L^2-norm on functions $g : \Xi \to [-M, M]$.

Let $\mathcal{M} \subset \mathbb{R}^{d_m}$ and assume that $m \mapsto \log \pi^m(a \mid s)$ is C-Lipschitz with respect to the ℓ_2-norm for all $a \in \mathcal{A}, s \in \mathcal{S}$ so that

$$|\log \pi^m(a \mid s) - \log \pi^{m'}(a \mid s)| \leq C \|m - m'\|_2.$$

Any ϵ-covering $\{m_1, \dots, m_N\}$ of $\mathcal{M}$ in ℓ_2-norm yields $\min_i |\log \pi^m(a \mid s) - \log \pi^{m_i}(a \mid s)| \le C\epsilon$ for all $m \in \mathcal{M}, a \in \mathcal{A}, s \in \mathcal{S}$. This implies that ℓ_2-covering numbers of $\mathcal{M}$ control L^∞-covering numbers of the set of functions $\mathcal{G}_1, \mathcal{G}_2(s), \mathcal{G}_3$:

$$\max\left\{N(\mathcal{G}_1, \epsilon, L^\infty), \sup_{s\in\mathcal{S}} N(\mathcal{G}_2(s), \epsilon, L^\infty), N(\mathcal{G}_3, \epsilon, L^\infty)\right\} \le N(\mathcal{M}, \epsilon/C, \|\cdot\|_2) \le \left(1 + \frac{\operatorname{diam}(\mathcal{M})C}{\epsilon}\right)^{d_m},$$

where $\operatorname{diam}(\mathcal{M}) = \sup_{m,m\in\mathcal{M}} \|m - m'\|_2$. In Pollard's discretization-based bound (10), setting $\delta = \operatorname{diam}(\mathcal{M})CN^{-1}$ yields

$$\mathbb{E}[\mathfrak{R}_N(\mathcal{G}_1)] \lesssim \sqrt{\frac{d_m}{N}} + \frac{\operatorname{diam}(\mathcal{M})C}{N}.$$

Plugging this bound in Lemma 5, the idealized empirical model $\widehat{m}_{N,\infty}$ achieves

$$\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{\widehat{m}_{N,\infty}} \mid S\right)\right] \lesssim \inf_{m\in\mathcal{M}} \mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{m} \mid S\right)\right] + \sqrt{\frac{d_m}{N}} + \frac{\operatorname{diam}(\mathcal{M})C}{N} \tag{11}$$

with probability at least $1 - 2e^{-t}$, where $\lesssim$ denotes an inequality up to some universal constant.

We now show that $\widehat{m}_{N,N_a}$ achieves a similar generalization guarantee as the idealized model $\widehat{m}_{N,\infty}$. Using the bound (10), we again get

$$\sup_{s\in\mathcal{S}} \mathbb{E}_{A_j^{\mathrm{TS}} \overset{\mathrm{iid}}{\sim} \pi^{\mathrm{TS}}(\cdot|s)}[\mathfrak{R}_{N_a}(\mathcal{G}_2(s))] \lesssim \sqrt{\frac{d_m}{N_a}} + \frac{\operatorname{diam}(\mathcal{M})C}{N_a},$$
$$\mathbb{E}[\mathfrak{R}_{NN_a}(\mathcal{G}_3)] \lesssim \sqrt{\frac{d_m}{NN_a}} + \frac{\operatorname{diam}(\mathcal{M})C}{NN_a}.$$

Applying these bounds to Theorem 3, we see that $\widehat{m}_{N,N_a}$ enjoys the same guarantee (11) as the "best-case" idealized empirical solution $\widehat{m}_{N,\infty}$ (up to constants). $\diamond$

D Proof of regret bounds

D.1 Proof of regret decomposition (Lemma 1)

Conditional on (H_t, S_t), A_t^{TS} has the same distribution as $A_t^\star$. Since $U_t(a; H_t, S_t)$ is a deterministic function conditional on (H_t, S_t), we have

$$\mathbb{E}[U_t(A_t^{\mathrm{TS}}; H_t, S_t) \mid H_t, S_t] = \mathbb{E}[U_t(A_t^\star; H_t, S_t) \mid H_t, S_t].$$

We can rewrite the (conditional) instantenous regret as

$$\begin{aligned}
&\mathbb{E}[f_\theta(A_t^\star, S_t) - f_\theta(A_t, S_t) \mid H_t, S_t] \\
&= \mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t) \mid H_t, S_t] + \mathbb{E}[U_t(A_t^{\mathrm{TS}}; H_t, S_t) - f_\theta(A_t, S_t) \mid H_t, S_t] \\
&= \mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t) \mid H_t, S_t] + \mathbb{E}[U_t(A_t; H_t, S_t) - f_\theta(A_t, S_t) \mid H_t, S_t] \\
&\quad + \mathbb{E}[U_t(A_t^{\mathrm{TS}}; H_t, S_t) - U_t(A_t; H_t, S_t) \mid H_t, S_t].
\end{aligned} \tag{12}$$

We proceed by bounding the gap

$$\mathbb{E}[U_t(A_t^{\mathrm{TS}}; H_t, S_t) - U_t(A_t; H_t, S_t) \mid H_t, S_t] \tag{13}$$

by the KL divergence between π_t^{TS} and π_t. Recall Pinsker's inequality [55]

$$\|P - Q\|_{\mathrm{TV}} := \frac{1}{2} \sup_{g:\mathcal{A}\to[-1,1]} |\mathbb{E}_P[g(A)] - \mathbb{E}_Q[g(A)]| \le \sqrt{\frac{1}{2} D_{\mathrm{kl}}\left(P \| Q\right)}.$$

From the hypothesis, Pinsker's inequality implies

$$\begin{aligned}
\mathbb{E}[|U_t(A_t^{\mathrm{TS}}; H_t, S_t) - U_t(A_t; H_t, S_t)| \mid H_t, S_t] &\le 2M_t(H_t, S_t) \left\|\pi_t^{\mathrm{TS}}(\cdot \mid S_t) - \pi_t(\cdot \mid S_t)\right\|_{\mathrm{TV}} \\
&\le M_t(H_t, S_t)\sqrt{2D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)}.
\end{aligned}$$

Applying this bound in the decomposition (12), and taking expectation over (H_t, S_t) on both sides and summing $t = 1, \ldots, T$, we get

$$\begin{aligned}\mathrm{BayesRegret}\,(T,\pi) \le \sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t)] + \sum_{t=1}^{T} \mathbb{E}[U_t(A_t; H_t, S_t) - f_\theta(A_t, S_t)] \\ + \sum_{t=1}^{T} \mathbb{E}\left[M_t(H_t, S_t)\sqrt{2D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)}\right].\end{aligned}$$

Applying Cauchy-Schwarz inequality and noting that $\sqrt{\mathbb{E}[M_t(H_t, S_t)^2]} \le L$, we obtain the final decomposition.

D.2 Proof of Theorem 1

We begin by defining a few requisite concepts. Recall that a collection $v_1, \ldots, v_N$ is an *ϵ-cover* of a set V in norm $\|\cdot\|$ if for each $v \in \mathcal{V}$, there exists v_i such that $\|v - v_i\| \le \epsilon$. The *covering number* is

$$N(V, \epsilon, \|\cdot\|) := \inf\{N \in \mathbb{N} \mid \text{there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\|\}.$$

For a class of functions $\mathcal{H} \subset \{f : \mathcal{A} \times \mathcal{S} \to \mathbb{R}\}$, we consider the sup-norm $\|h\|_{L^\infty(\mathcal{X})} := \sup_{a\in\mathcal{A}, s\in\mathcal{S}} |h(a,s)|$.

We use the notion of *eluder dimension* proposed by Russo and Van Roy [47], which quantifies the size of the function class $\mathcal{F} = \{f_\theta(\cdot,\cdot) : \theta \in \Theta\}$ for sequential decision making problems.

Definition 1. *An action-state pair $(a, s) \in (\mathcal{A}, \mathcal{S})$ is ϵ-dependent on $\{(a_1, s_1), \ldots, (a_n, s_n)\} \subset \mathcal{A} \times \mathcal{S}$ with respect to $\mathcal{F}$ if for any $f, f' \in \mathcal{F}$*

$$\left(\sum_{i=1}^{n}(f(a_i, s_i) - f'(a_i, s_i))^2\right)^{\frac{1}{2}} \le \epsilon \;\; \textit{implies} \;\; f(a,s) - f'(a,s) \le \epsilon.$$

We say that $(a, s) \in \mathcal{A} \times \mathcal{S}$ is ϵ-independent of $\{(a_1, s_1), \ldots, (a_n, s_n)\}$ with respect to $\mathcal{F}$ if (a, s) is not ϵ-dependent on $\{(a_1, s_1), \ldots, (a_n, s_n)\}$.

Definition 2. *The* eluder dimension $d_{\mathrm{E}}(\mathcal{F}, \epsilon)$ *of $\mathcal{F}$ is the length of the longest sequence in $\mathcal{A} \times \mathcal{S}$ such that for some $\epsilon' \ge \epsilon$, every element in the sequence is ϵ'-independent of its predecessors.*

The eluder dimension bounds the Bayes regret decomposition given in Lemma 1.

Lemma 6 (Russo and Van Roy [47])**.** *Let $\pi = \{\pi_t\}_{t\ge1}$ be any policy, and $\mathcal{F} = \{(a,s) \mapsto f_\theta(a,s) : \theta \in \Theta\}$. Assume $f_\theta(a,s) \in [-M, M]$ for all $\theta \in \Theta, a \in \mathcal{A}, s \in \mathcal{S}$, and $R_t - f_\theta(A_t, S_t)$ is σ sub-Gaussian conditional on (θ, H_t, S_t, A_t). When $\sup_{a\in\mathcal{A}} |U_t(a; H_t, S_t)| \le M_t(H_t, S_t)$ holds, we have*

$$\begin{aligned}\mathrm{BayesRegret}\,(T,\pi) \le & C d_{\mathrm{E}}(\mathcal{F}, T^{-1}) + \sigma\sqrt{T d_{\mathrm{E}}(\mathcal{F}, T^{-1})(\log T + \log N(\mathcal{F}, T^{-1}, \|\cdot\|_{L^\infty(\mathcal{X})}))} \\ & + L\sum_{t=1}^{T}\sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]},\end{aligned}$$

and when condition (7) *holds, we have*

$$\begin{aligned}\mathrm{BayesRegret}\,(T,\pi) \le & C d_{\mathrm{E}}(\mathcal{F}, T^{-1}) + \sigma\sqrt{T d_{\mathrm{E}}(\mathcal{F}, T^{-1})(\log T + \log N(\mathcal{F}, T^{-1}, \|\cdot\|_{L^\infty(\mathcal{X})}))} \\ & + L\sum_{t=1}^{T}\mathbb{E}\left[D_{\mathrm{w}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right].\end{aligned}$$

for some constant $C > 0$ that only depends on M.

From Lemma 6, it suffices to bound the covering number and the eluder dimension of the linear model class

$$\mathcal{F} = \{(a,s) \mapsto g(\langle\phi(a,s), \theta\rangle) : \theta \in \Theta\}.$$

Since $\theta \mapsto g(\langle \phi(a,s), \theta \rangle)$ is c_2-Lipschitz with respect to $\|\cdot\|_2$, a standard covering argument (e.g. see Chapter 2.7.4 of van der Vaart and Wellner [56]) gives

$$N\left(\mathcal{H}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}\right) \leq N\left(\Theta, \frac{\epsilon}{c_2}, \|\cdot\|\right) \leq \left(1 + \frac{2c_1c_2}{\epsilon}\right)^d.$$

Proposition 11, Russo and Van Roy [47] shows that

$$d_{\mathrm{E}}(\mathcal{F}, T^{-1}) \leq Cdr^2 \log rT$$

for some constant C that depends only on c_1 and c_2. Using these bounds in Lemma 6, we obtain the result.

D.3 Explicit regret bounds for linear bandits

In the case of linear bandits, we can use a more direct argument that leverage the rich analysis of UCB algorithms provided by previous authors [17, 1, 2], instead of the eluder dimension argument used to show Theorem 1.

Instead of bounding the eluder dimension, we can directly bound the upper confidence bounds in the decomposition in Lemma 1. By using the regret analysis of Dani et al. [17], Abbasi-Yadkori et al. [1, 2] for UCB algorithms, we obtaint he following result for linear contextual bandits.

Lemma 7. *Let $\phi : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ such that $f_\theta(a,s) = \phi(a,s)^\top \theta$ for all $\theta \in \Theta$. Let $c_1, c_2, \sigma > 0$ be such that*

$$\sup_{\theta \in \Theta} \|\theta\|_2 \leq c_1, \quad \sup_{a \in \mathcal{A}, s \in \mathcal{S}} \|\phi(a,s)\|_2 \leq c_2,$$

and assume that $R_t - f_\theta(A_t, S_t)$ is σ-sub-Gaussian conditional on (θ, H_t, S_t, A_t). Then, there exists a constant C that depends on c_1, c_2, σ such that

$$\begin{aligned}\mathrm{BayesRegret}\left(T, \{\pi_t\}_{t\in\mathbb{N}}\right) \leq & 2\left((c_1+1)c_1 + \sigma\sqrt{d + \log\sqrt{T}\left(1 + \frac{c_2^2 T}{\lambda}\right)}\right)\sqrt{2Td\log\left(\lambda + \frac{Tc_2^2}{d}\right)} \\ & + 4c_1c_2\sqrt{T} + c_1c_2\sum_{t=1}^T \sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]} \end{aligned} \tag{14}$$

Furthermore, if $a \mapsto \phi(a,s)$ is L-Lipschitz with respect to a metric d, then the same bound holds with $L\sum_{t=1}^T \mathbb{E}\left[D_{\mathrm{w}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]$ replacing the last sum, where $D_{\mathrm{w}}(\cdot,\cdot \mid \cdot)$ is the Wasserstein distance defined with the metric d.

Although we omit it for brevity, the above $O(\sqrt{dT}\log T)$ regret bound can be improved to $\tilde{O}(\mathbb{E}[\sqrt{\|\theta\|_0 \, dT}])$ by using a similar argument as below (see Proposition 3, [47] and [2]).

Proof

Lemma 7 follows from a direct consequence of Lemma 1, and Dani et al. [17], Abbasi-Yadkori et al. [1]; we detail it below for completeness. We first show the bound (14). Letting $L_t(a; H_t, S_t)$ be an arbitrary sequence of measurable functions denoting lower confidence bounds, the Bayes regret decomposition in Lemma 1 implies

$$\begin{aligned}\mathrm{BayesRegret}\left(T, \pi\right) \leq & \sum_{t=1}^T \mathbb{E}[U_t(A_t; H_t, S_t) - L_t(A_t; H_t, S_t)] \\ & + 2c_1c_2\sum_{t=1}^T \left\{\mathbb{P}(f_\theta(A_t, S_t) \leq L_t(A_t; H_t, S_t)) + \mathbb{P}(f_\theta(A_t^\star, S_t) \geq U_t(A_t^\star; H_t, S_t))\right\} \\ & + L\sum_{t=1}^T \sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]}. \end{aligned} \tag{15}$$

We proceed by bounding the first and second sum in the above inequality.

To ease notation, for a fixed $\lambda \geq 1 \vee c_2^2$ define

$$X_t := \begin{bmatrix} \phi(A_1, S_1)^\top \\ \vdots \\ \phi(A_t, S_t)^\top \end{bmatrix}, \quad Y_t := \begin{bmatrix} R_1 \\ \vdots \\ R_t \end{bmatrix}, \quad V_t := \lambda I + \sum_{k=1}^{t} \phi(A_k, S_k)\phi(A_k, S_k)^\top$$

for all $t \in \mathbb{N}$, and we let $V_0 := \lambda I$. We use the following key result due to Dani et al. [17], Abbasi-Yadkori et al. [1].

Lemma 8 (Theorem 2, Abbasi-Yadkori et al. [1])**.** *Under the conditions of the proposition, for any* $\delta > 0$

$$\mathbb{P}\left(\left\| \theta - \widehat{\theta}_t \right\|_{V_t} \leq \sqrt{\lambda} c_1 + \sigma \sqrt{d \left(\log \frac{1}{\delta} + \log\left(1 + \frac{c_2^2 t}{\lambda}\right)\right)} =: \beta_t(\delta) \text{ for all } t \geq 0 \;\middle|\; \theta \right) \geq 1 - \delta$$

where we used $\|\theta\|_A := \sqrt{\theta^\top A \theta}$.

To instantiate the decomposition (15), we let

$$U_t(a; H_t, S_t) := \sup_{\theta' : \left\|\theta' - \widehat{\theta}_{t-1}\right\|_{V_{t-1}} \leq \beta_{t-1}(\delta)} \phi(a, S_t)^\top \theta',$$
$$L_t(a; H_t, S_t) := \inf_{\theta' : \left\|\theta' - \widehat{\Theta}_{t-1}\right\|_{V_{t-1}} \leq \beta_{t-1}(\delta)} \phi(a, S_t)^\top \theta'.$$

We are now ready to bound the second term in the decomposition (15). On the event

$$\mathcal{E} := \left\{ \left\| \theta - \widehat{\theta}_t \right\|_{V_t} \leq \beta_t(\delta) \text{ for all } t \geq 0 \right\},$$

we have $f_\theta(A_t, S_t) \geq L_t(A_t; H_t, S_t)$ and $f_\theta(A_t^\star, S_t) \leq U_t(A_t^\star; H_t, S_t)$ by definition. Since Lemma 8 states $\mathbb{P}(\mathcal{E} \mid \theta) \geq 1 - \delta$, we conclude that the second sum in the decomposition (15) is bounded by $4 c_1 c_2 T \delta$.

To bound the first sum in the decomposition (15), we use the following bound on the norm of feature vectors.

Lemma 9 (Lemma 11, Abbasi-Yadkori et al. [1])**.** *If* $\lambda \geq c_2^2 \vee 1$*, for any sequence of* a_t, s_t *for* $t \geq 1$*, and corresponding* $A_t := \lambda I + \sum_{k=1}^{t} \phi(a_k, s_k)\phi(a_k, s_k)^\top$*, we have*

$$\sum_{t=1}^{T} \|\phi(A_t, S_t)\|^2_{V_{t-1}^{-1}} \leq 2d \log\left(\lambda + \frac{T c_2^2}{d}\right).$$

Noting that by definition

$$U_t(A_t; H_t, S_t) - L_t(A; H_t, S_t) \leq 2 \|\phi(A_t, S_t)\|_{V_{t-1}^{-1}} \beta_{t-1}(\delta),$$

we obtain

$$\begin{aligned}
\sum_{t=1}^{T} U_t(A_t; H_t, S_t) - L_t(A; H_t, S_t) &\leq 2 \sum_{t=1}^{T} \|\phi(A_t, S_t)\|_{V_{t-1}^{-1}} \beta_{t-1}(\delta) \\
&\overset{(a)}{\leq} 2\beta_T(\delta) \sum_{t=1}^{T} \|\phi(A_t, S_t)\|_{V_{t-1}^{-1}} \\
&\overset{(b)}{\leq} 2\beta_T(\delta) \sqrt{T \sum_{t=1}^{T} \|\phi(A_t, S_t)\|^2_{V_{t-1}^{-1}}} \\
&\overset{(c)}{\leq} 2\beta_T(\delta) \sqrt{2Td \log\left(\lambda + \frac{Tc_2^2}{d}\right)}
\end{aligned}$$

where we used monotonicity of $t \mapsto \beta_t(\delta)$ in step (a), Cauchy-Schwarz inequality in step (b), and Lemma 9 in step (c).

Collecting these bounds, we conclude

$$\begin{aligned}\mathrm{BayesRegret}\left(T, \pi\right) \leq & 2\beta_T(\delta)\sqrt{2Td\log\left(\lambda + \frac{Tc_2^2}{d}\right)} + 4c_1c_2T\delta \\ & + L\sum_{t=1}^{T}\sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]}.\end{aligned}$$

Setting $\delta = 1/\sqrt{T}$, we obtain the first result. The second result is immediate by starting with the decomposition (8) and using an identical argument.

D.4 Proof of Theorem 2

In what follows, we abuse notation and let C be a universal constant that changes line by line. Since $f_\theta(a, s)$ follows a Gaussian process, its posterior mean and variance is given by

$$\begin{aligned}\mu_t(a, s) &:= \mathbb{E}[f_\theta(a, s) \mid H_t] = k_t(a, s)^\top (K_t + \sigma^2 I)^{-1} y_t, \\ \sigma_t^2(a, s) &:= \mathrm{Var}(f_\theta(a, s) \mid H_t) = k((a, s), (a, s)) - k_t(a, s)^\top (K_t + \sigma^2 I)^{-1} k_t(a, s)\end{aligned}$$

where $k_t(a, s) := [k((A_j, S_j), (a, s))]_{1\leq j\leq t}$, $K_t := [k((A_i, S_i), (A_j, S_j))]_{1\leq i,j\leq t}$ and $y_t = [r_j]_{1\leq j\leq t}$. Define the upper confidence bound

$$U_t(a; H_t, s) := \mu_t(a, s) + \sqrt{\beta_t}\sigma_t(a, s)$$

where $\beta_t = 2\log((t^4 rd)^d t^2)$. Noting that

$$|U_t(a; H_t, s)| \leq \sup_{a\in\mathcal{A}, s\in\mathcal{S}} |\mathbb{E}\left[f_\theta(a, s) \mid H_t\right]| + \sqrt{\beta_t}k((a, s), (a, s)) \leq \mathbb{E}\left[\sup_{a\in\mathcal{A}, s\in\mathcal{S}} |f_\theta(a, s)| \;\middle|\; H_t\right] + \sqrt{\beta_t}c_2,$$

a minor modification to the proof of Lemma 1 yields

$$\begin{aligned}\mathrm{BayesRegret}\left(T, \pi\right) \leq & \sum_{t=1}^{T}\mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t)] + \sum_{t=1}^{T}\mathbb{E}[U_t(A_t; H_t, S_t) - f_\theta(A_t, S_t)] \\ & + \sum_{t=1}^{T}\left(\left\|\mathbb{E}\left[\sup_{a\in\mathcal{A}, s\in\mathcal{S}} |f_\theta(a, s)| \;\middle|\; H_t\right]\right\|_{2,P} + \sqrt{\beta}_t c_2\right)\sqrt{2\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi_t^{\mathrm{TS}}, \pi_t \mid S_t\right)\right]}.\end{aligned} \tag{16}$$

From Jensen's inequality and the tower property,

$$\left\|\mathbb{E}\left[\sup_{a\in\mathcal{A}, s\in\mathcal{S}} |f_\theta(a, s)| \;\middle|\; H_t\right]\right\|_{2,P} \leq \left\|\sup_{a\in\mathcal{A}, s\in\mathcal{S}} |f_\theta(a, s)|\right\|_{2,P} = c_3.$$

From Borell-TIS inequality (e.g., see [4]), we have $c_3 < \infty$.

We now proceed by bounding the first two terms in the regret decomposition (16). Let $\mathcal{A}_t$ be a $(1/t^4)$-cover of $\mathcal{A}$, so that for any $a \in \mathcal{A}$, there exists $[a]_t \in \mathcal{A}_t$ such that $\|a - [a]_t\|_1 \leq 1/t^4$. Since $|\mathcal{A}_t| \leq (t^4 rd)^d$, we have $2\log(|\mathcal{A}_t|t^2) \leq \beta_t$. We begin by decomposing the first term in the decomposition.

$$\begin{aligned}\sum_{t=1}^{T} f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t) = & \underbrace{\sum_{t=1}^{T}\mathbb{E}[f_\theta(A_t^\star, S_t) - f_\theta([A_t^\star]_t, S_t)]}_{(a)} + \underbrace{\sum_{t=1}^{T}\mathbb{E}[f_\theta([A_t^\star]_t, S_t) - U_t([A_t^\star]_t; H_t, S_t)]}_{(b)} \\ & + \underbrace{\sum_{t=1}^{T}\mathbb{E}[U_t([A_t^\star]_t; H_t, S_t) - U_t(A_t^\star; H_t, S_t)]}_{(c)}.\end{aligned}$$

Using the definition of L_f, the first term (a) in the above equality is bounded by

$$\sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^\star, S_t) - f_\theta([A_t^\star]_t, S_t)] \leq \mathbb{E}[L_f] \sum_{t=1}^{T} \|A_t^\star - [A_t^\star]_t\|_1 \leq \mathbb{E}[L_f] \sum_{t=1}^{\infty} \frac{1}{t^4} \leq C\mathbb{E}[L_f]$$

where we used the fact that $\mathcal{A}_t$ is a $1/t^4$-cover of $\mathcal{A}$. To bound the second term (b), note that since $f_\theta(a, s) \mid H_t \sim N(\mu_t(a, s), \sigma_t^2(a, s))$, we have

$$\mathbb{E}[f_\theta(a, s) - U_t(a; H_t, s) \mid H_t] \leq \mathbb{E}[(f_\theta(a, s) - U_t(a; H_t, s))_+ \mid H_t] = \frac{\sigma_t(a, s)}{\sqrt{2\pi}} e^{-\frac{\beta_t}{2}} \leq \frac{c_2}{\sqrt{2\pi} t^2 |\mathcal{A}_t|}. \tag{17}$$

Hence, we obtain the bound

$$\sum_{t=1}^{T} \mathbb{E}[f_\theta([A_t^\star]_t, S_t) - U_t([A_t^\star]_t; H_t, S_t)] \leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \mathbb{E}[f_\theta(a, S_t) - U_t(a; H_t, S_t)] \leq \sum_{t=1}^{\infty} \frac{c_2}{\sqrt{2\pi} t^2} \leq C c_2$$

where we used the independence of S_t and H_t, and the bound (17).

To bound the third term (c), we show the claim

$$\begin{aligned} |U_t(a; H_t, s) - U_t(a'; H_t, s)| \leq \mathbb{E}[L_f \mid H_t] \|a - a'\|_1 \\ + \sqrt{\beta_t} \left(2\mathbb{E}\left[L_f \left(\sup_{a \in \mathcal{A}, s \in \mathcal{S}} \mu(a, s)^2 + \sup_{a \in \mathcal{A}, s \in \mathcal{S}} f_\theta(a, s)^2 \right) \mid H_t \right] \right)^{\frac{1}{2}} \|a - a'\|_1^{\frac{1}{2}}. \end{aligned} \tag{18}$$

From the above claimed bound, it follows that

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[U_t([A_t^\star]_t; H_t, S_t) - U_t(A_t^\star; H_t, S_t)] &\leq \sum_{t=1}^{T} \frac{\mathbb{E}[L_f]}{t^4} + \sum_{t=1}^{T} \sqrt{2\beta_t} \frac{c_1 \sqrt{\mathbb{E}[L_f]} + c_3 \sqrt{\mathbb{E}[L_f^2]}}{t^2} \\ &\leq C\mathbb{E}[L_f] + Cd \log(rd) \left(c_1 \sqrt{\mathbb{E}[L_f]} + c_3 \sqrt{\mathbb{E}[L_f^2]} \right). \end{aligned}$$

To show the bound (18), first note that $a \mapsto \mathbb{E}[f_\theta(a, s) \mid H_t]$ and $a \mapsto \mathbb{E}[f_\theta(a, s)^2 \mid H_t]$ is $\mathbb{E}[L_f \mid H_t]$- and $\mathbb{E}[2L_f \sup_{a \in \mathcal{A}, s \in \mathcal{S}} |f_\theta(a, s)| \mid H_t]$- Lipschitz respectively, for all $s \in \mathcal{S}$. Hence, $a \mapsto \sigma_t^2(a, s)$ is $\mathbb{E}[2L_f(c_1^2 + \sup_{a \in \mathcal{A}, s \in \mathcal{S}} |f_\theta(a, s)|^2) \mid H_t]$-Lipschitz. Noting that

$$|\sigma_t(a, s) - \sigma_t(a', s)| = \left| \frac{\sigma_t^2(a, s) - \sigma_t^2(a', s)}{\sigma_t(a, s) + \sigma_t(a', s)} \right| \leq \frac{1}{c} |\sigma_t^2(a, s) - \sigma_t^2(a', s)| + c$$

for any $c > 0$, taking the infimum over $c > 0$ on the right hand side yields

$$\begin{aligned} |\sigma_t(a, s) - \sigma_t(a', s)| &\leq \sqrt{2|\sigma_t^2(a, s) - \sigma_t^2(a', s)|} \\ &\leq \left(2\mathbb{E}\left[L_f \left(c_1^2 + \sup_{a \in \mathcal{A}, s \in \mathcal{S}} f_\theta(a, s)^2 \right) \mid H_t \right] \right)^{\frac{1}{2}} \|a - a'\|_1^{\frac{1}{2}} \end{aligned}$$

which shows the bound (18).

Collecting these bounds, we have shown that

$$\sum_{t=1}^{T} \mathbb{E}[f_\theta(A_t^\star, S_t) - U_t(A_t^\star; H_t, S_t)] \leq C\mathbb{E}[L_f] + Cc_2 + Cd \log(rd) \left(c_1 \sqrt{\mathbb{E}[L_f]} + c_3 \sqrt{\mathbb{E}[L_f^2]} \right). \tag{19}$$

To bound the second term in the Bayes regret decomposition (16), we use the following lemma due to Srinivas et al. [51].

Lemma 10 (Lemma 5.3 Srinivas et al. [51])**.** *For any sequence of A_t and S_t,*

$$\mathbb{E}\left(\sum_{t=1}^{T} \sigma_t(A_t, S_t)^2 \right)^{\frac{1}{2}} \leq \sqrt{\frac{2\gamma_T}{\log(1 + \sigma^{-2})}}$$

Using the lemma, we have

$$\sum_{t=1}^{T} \mathbb{E}[U_t(A_t; H_t, S_t) - f_\theta(A_t, S_t)] = \sum_{t=1}^{T} \sqrt{\beta_t}\mathbb{E}[\sigma_t(A_t, S_t)] \le \sqrt{T\beta_T}\sqrt{\frac{2\gamma_T}{\log(1+\sigma^{-2})}}.$$

Combining this with the bound (19), we obtain our result.

E Proof of generalization results

E.1 Proof of Lemma 5

We use the following standard concentration result based on the bounded differences inequality and a symmetrization argument; see, for example, [13, 58, 14]. We denote by $\widehat{P}_n$ the empirical distribution constructed from any i.i.d. sample $X_i \sim P$.

Lemma 11. *If $|g| \le M$ for all $g \in \mathcal{G}$, then with probability at least $1 - 2e^{-t}$*

$$\sup_{g\in\mathcal{G}} |\mathbb{E}[g(X)] - \mathbb{E}_{\widehat{P}_n}[g(X)]| \le 2\mathbb{E}[\mathfrak{R}_n(\mathcal{G})] + M\sqrt{\frac{2t}{n}}.$$

Noting that for any $m \in \mathcal{M}$

$$\begin{aligned}
&\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{\widehat{m}_{N,\infty}} \mid S\right)\right] - \mathbb{E}[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{m} \mid S\right)] \\
&= \mathbb{E}[\log \pi^m(A^{\mathrm{TS}} \mid S)] - \mathbb{E}[\log \pi^{\widehat{m}_{N,\infty}}(A^{\mathrm{TS}} \mid S)] \\
&= \mathbb{E}[\log \pi^m(A^{\mathrm{TS}} \mid S)] - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^m(A^{\mathrm{TS}} \mid S_i)] \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^m(A^{\mathrm{TS}} \mid S_i)] - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^{\widehat{m}_{N,\infty}}(A^{\mathrm{TS}} \mid S_i)] \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^{\widehat{m}_{N,\infty}}(A^{\mathrm{TS}} \mid S_i)] - \mathbb{E}[\log \pi^{\widehat{m}_{N,\infty}}(A^{\mathrm{TS}} \mid S)] \\
&\le 2\sup_{m\in\mathcal{M}} \left| \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^m(A^{\mathrm{TS}} \mid S_i)] - \mathbb{E}[\log \pi^m(A^{\mathrm{TS}} \mid S)] \right|
\end{aligned}$$

where we used the fact that $\widehat{m}_{N,\infty}$ maximizes $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot|S_i)}[\log \pi^m(A^{\mathrm{TS}} \mid S_i)]$ in the last inequality. Applying Lemma 11 with $\mathcal{G} = \mathcal{G}_1$ and taking the infimum over $m \in \mathcal{M}$, we obtain the result.

E.2 Proof of Theorem 3

We begin by noting that

$$\begin{aligned}
&\mathbb{E}\left[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{\widehat{m}_{N,N_a}} \mid S\right)\right] - \mathbb{E}[D_{\mathrm{kl}}\left(\pi^{\mathrm{TS}}, \pi^{m} \mid S\right)] = \mathbb{E}[\log \pi^m(A^{\mathrm{TS}} \mid S)] - \mathbb{E}[\log \pi^{\widehat{m}_{N,N_a}}(A^{\mathrm{TS}} \mid S)] \\
&= \mathbb{E}[\log \pi^m(A^{\mathrm{TS}} \mid S)] - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_A} \log \pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_A} \log \pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_A} \log \pi^{\widehat{m}_{N,N_a}}(A_{ij}^{\mathrm{TS}} \mid S_i) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_A} \log \pi^{\widehat{m}_{N,N_a}}(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}[\log \pi^{\widehat{m}_{N,N_a}}(A^{\mathrm{TS}} \mid S)].
\end{aligned}$$

Since $\widehat{m}_{N,N_a}$ maximizes $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_A}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i)$, the preceeding display can be bounded by

$$
\begin{aligned}
&2\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid S)]\right| \\
&\leq 2\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid S_i)}[\log\pi^m(A^{\mathrm{TS}} \mid S_i)] - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid S)]\right| \\
&\quad + 2\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid S_i)}[\log\pi^m(A^{\mathrm{TS}} \mid S_i)]\right)\right|
\end{aligned} \tag{20}
$$

We proceed by separately bounding the two terms in the inequality (20). From Lemma 11, the second term is bounded by

$$4\mathbb{E}[\mathfrak{R}_N(\mathcal{G}_1)] + 2M\sqrt{\frac{2t}{N}}$$

with probability at least $1-2e^{-t}$. To bound the first term

$$Z_{N,N_a} := 2\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid S_i)]\right)\right|,$$

consider the Doob martingale

$$M_k := \mathbb{E}[Z_{N,N_a} \mid S_1,\ldots,S_k] \text{ for } 1\leq k\leq N$$

with $M_0 = \mathbb{E}[Z_{N,N_a}]$, which is martingale adapted to the filtration $\mathcal{F}_k := \sigma(S_1,\ldots,S_k)$. Denote the martingale difference sequence $D_k = M_k - M_{k-1}$ for $k\geq 1$. Let $\bar{S}_k$ be an independent copy of S_k that is independent of all $S_i, A_{ij}^{\mathrm{TS}}$ for $i\neq k$, and let $\bar{A}_{kj}^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid\bar{S}_k)$ independent of everything other $\bar{S}_k$. We can write

$$
\begin{aligned}
\frac{1}{2}|D_k| = \mathbb{E}&\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid S_i)]\right)\right|\,\middle|\, S_1,\ldots,S_k\right] \\
&- \mathbb{E}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i\neq k}\left(\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\mathrm{TS}} \mid S_i) - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid S_i)]\right)\right.\right. \\
&\qquad\left.\left. + \frac{1}{N}\left(\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(\bar{A}_{kj}^{\mathrm{TS}} \mid \bar{S}_k) - \mathbb{E}[\log\pi^m(A^{\mathrm{TS}} \mid \bar{S}_k)]\right)\right|\,\middle|\, S_1,\ldots,S_k\right].
\end{aligned}
$$

Thus, we arrive at the bound independence of S_i's yields

$$
\begin{aligned}
\frac{1}{2}|D_k| &\leq \frac{1}{N}\mathbb{E}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N_a}\sum_{j=1}^{N_a}\Big\{\log\pi^m(A_{kj}^{\mathrm{TS}} \mid S_k) - \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid S_k)}[\log\pi^m(A^{\mathrm{TS}} \mid S_k)]\right.\right. \\
&\qquad\left.\left. - \log\pi^m(\bar{A}_{kj}^{\mathrm{TS}} \mid \bar{S}_k) + \mathbb{E}_{\bar{A}^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid\bar{S}_k)}[\log\pi^m(\bar{A}^{\mathrm{TS}} \mid \bar{S}_k)]\Big\}\right|\,\middle|\, S_k\right] \\
&\leq \frac{2}{N}\sup_{s\in\mathcal{S}}\mathbb{E}_{A_j^{\mathrm{TS}}\overset{\mathrm{iid}}{\sim}\pi^{\mathrm{TS}}(\cdot\mid s)}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_j^{\mathrm{TS}} \mid s) - \mathbb{E}_{A^{\mathrm{TS}}\sim\pi^{\mathrm{TS}}(\cdot\mid s)}[\log\pi^m(A^{\mathrm{TS}} \mid s)]\right|\right]
\end{aligned}
$$

where $\bar{S}_k$ is an independent copy of S_k, and similarly $\bar{A}_{kj}^{\mathrm{TS}}\overset{\mathrm{iid}}{\sim}\pi^{\mathrm{TS}}(\cdot\mid\bar{S}_k)$.

Next, we use a standard symmetrization result to bound the preceding display; see, for example, Chapter 2.3, van der Vaart and Wellner [56] for a comprehensive treatment.

Lemma 12. *If $X_i \overset{\text{iid}}{\sim} P$, we have*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}(g(X_i)-\mathbb{E}[g(X)])\right|\right] \le 4\mathbb{E}[\mathfrak{R}_n(\mathcal{G})]$$

Applying Lemma 12 to the bound on $|D_k|$. we conclude $|D_k| \le \frac{8}{N}\sup_{s\in\mathcal{S}}\mathbb{E}_{A_j^{\text{TS}}\overset{\text{iid}}{\sim}\pi^{\text{TS}}(\cdot|s)}[\mathfrak{R}_{N_a}(\mathcal{G}_2(s))]$. Then, Azuma-Hoeffding bound (Corollary 2.1, Wainwright [58]) yields

$$Z_{N,N_a} \le \mathbb{E}[Z_{N,N_a}] + \sqrt{\frac{32t}{N}}\sup_{s\in\mathcal{S}}\mathbb{E}_{A_j^{\text{TS}}\overset{\text{iid}}{\sim}\pi^{\text{TS}}(\cdot|s)}[\mathfrak{R}_{N_a}(\mathcal{G}_2(s))]$$

with probability at least $1-e^{-t}$.

It now remains to bound $\mathbb{E}[Z_{N,N_a}]$, for which we use a symmetrization argument. Although $(S_i, A_{ij}^{\text{TS}})$ are not i.i.d., a standard argument still applies, which we outline for completeness. Denoting by $(\bar{S}_i, \bar{A}_{ij}^{\text{TS}})$ independent copies of $(S_i, A_{ij}^{\text{TS}})$, note that

$$\begin{aligned}
\mathbb{E}[Z_{N,N_a}] &= 2\mathbb{E}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\text{TS}}\mid S_i) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(\bar{A}_{ij}^{\text{TS}}\mid \bar{S}_i)\right]\right|\right] \\
&\le 2\mathbb{E}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\log\pi^m(A_{ij}^{\text{TS}}\mid S_i) - \log\pi^m(\bar{A}_{ij}^{\text{TS}}\mid \bar{S}_i)\right|\right] \\
&= 2\mathbb{E}\left[\sup_{m\in\mathcal{M}}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_a}\sum_{j=1}^{N_a}\epsilon_{ij}(\log\pi^m(A_{ij}^{\text{TS}}\mid S_i) - \log\pi^m(\bar{A}_{ij}^{\text{TS}}\mid \bar{S}_i))\right|\right] \\
&\le 8\mathbb{E}[\mathfrak{R}_{NN_a}(\mathcal{G}_3)].
\end{aligned}$$

Collecting these bounds, we conclude that with probability $1-3e^{-t}$, the right hand side of the inequality (20) is bounded by

$$4\mathbb{E}[\mathfrak{R}_N(\mathcal{G}_1)] + 2M\sqrt{\frac{2t}{n}} + \sqrt{\frac{32t}{N}}\sup_{s\in\mathcal{S}}\mathbb{E}_{A_j^{\text{TS}}\overset{\text{iid}}{\sim}\pi^{\text{TS}}(\cdot|s)}[\mathfrak{R}_{N_a}(\mathcal{G}_2(s))] + 8\mathbb{E}[\mathfrak{R}_{NN_a}(\mathcal{G}_3)].$$

F Experiment Details

F.1 Hyperparameters

We use hyperparameters from Riquelme et al. [45] as follows. The NEURALGREEDY, NEURALLINEARTS methods use a fully-connected neural network with two hidden layers of containing 100 rectified linear units. The networks are multi-output, where each output corresponds for predicted reward under each action. The networks are trained using 100 mini-batch updates at each period to minimize the mean-squared error via RMSProp with an initial learning rate of 0.01. The learning rate is decayed after each mini-batch update according to an inverse time decay schedule with a decay rate of 0.55 and the learning rate is reset the initial learning rate each update period. For BOOTSTRAP-NN-TS, we use 10 replicates and train each replicate with all observations as in Riquelme et al. [45].

The Bayesian linear regression models used on the last linear layer for NEURALLINEAR-TS use the normal inverse gamma prior NIG($\mu_a = \mathbf{0}, \alpha_a = 3, \beta_a = 3, \Lambda_a = 0.25I_d$). LINEAR-TS uses a NIG($\mu_a = \mathbf{0}, \alpha_a = 6, \beta_a = 6, \Lambda_a = 0.25I_d$) prior distribution.

The imitation models used by the IL methods are fully-connected neural networks with two hidden layers of 100 units and hyperbolic tangent activations. The networks use a Softmax function on the outputs to predict the probability of selecting each action. The networks are trained using 2000 mini-batch updates via RMSProp to minimize the KL-divergence between the predicted probabilities and the approximate propensity scores of the Thompson sampling policy π^{TS}. For each observed

context S_i, we approximate the propensity scores of the Thompson sampling policy $\pi^{TS}(\cdot|S_i)$ using $N_a = 2048$ Monte Carlo samples: $\hat{\pi}^{TS}(a|S_i) = \frac{1}{N_a}\sum_{j=1}^{N_a} \mathbb{1}(A_{ij} = a)$ where $A_{ij} \sim \pi^{TS}(\cdot|S_i)$. We use an initial learning rate of 0.001. learning rate is decayed every 100 mini-batches according to an inverse time decay schedule with a decay rate of 0.05. In practice, the hyperparameters of the imitation model can be optimized or adjusted at each update period by minimizing the KL-divergence on a held-out subset of the observed data, which may lead to better regret performance. We do not use inverse propensity-weighting on the observations, but we suspect that may it may further improve performance.

Data preprocessing We normalize all numeric features to be in [0,1] and one-hot encode all categorical features. For the Warfarin dataset, we also normalize the rewards to be in [0,1].

F.2 Posterior Inference for Bayesian Linear Regression

LINEAR-TS: For each action, We assume the data for action a were generated from the linear function: $r_a = \boldsymbol{s}^T\boldsymbol{\theta}_a + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_a^2)$.

$$\sigma_a^2 \sim \text{IG}(\alpha_a, \beta_a), \;\; \boldsymbol{\theta_a}|\sigma_a^2 \sim \mathcal{N}(\boldsymbol{\mu}_a, \sigma_a^2\Sigma_a),$$

where the prior distribution is given by $\text{NIG}(\boldsymbol{\mu}_a, \Lambda_a, \alpha_a, \beta_a)$ and $\Lambda_a = \Sigma_a^{-1}$ is the precision matrix. After n_a observations of contexts $X_a \in \mathbb{R}^{n_a \times (d+1)}$ and rewards $\boldsymbol{y}_a \in \mathbb{R}^{n_a \times 1}$, we denote the joint posterior by $P(\boldsymbol{\theta}_a, \sigma_a^2) \sim \text{NIG}(\bar{\boldsymbol{\mu}}_a, \bar{\Lambda}_a, \bar{\alpha}_a, \bar{\beta}_a)$, where

$$\bar{\Lambda} = X_a^T X_a + \Lambda_a, \;\; \bar{\boldsymbol{\mu}}_a = \bar{\Lambda}_a^{-1}(\Lambda_a\boldsymbol{\mu}_a + X_a^T\boldsymbol{y}_a)$$
$$\bar{\alpha}_a = \alpha + \frac{n_a}{2}, \;\; \bar{\beta}_a = \beta + \frac{1}{2}(\boldsymbol{y}_a^T\boldsymbol{y}_a + \boldsymbol{\mu}_a^T\Lambda_a\boldsymbol{\mu}_a - \bar{\boldsymbol{\mu}}_a^T\bar{\Lambda}_a\bar{\boldsymbol{\mu}}_a).$$

F.3 Benchmark Problem Datasets

Mushroom UCI Dataset: This real dataset contains 8,124 examples with 22 categorical valued features containing descriptive features about the mushroom and labels indicating if the mushroom is poisonous or not. With equal probability, a poisonous mushroom may be unsafe and hurt the consumer or it may be safe and harmless. At each time step, the policy must choose whether to eat the new mushroom or abstain. The policy receives a small positive reward (+5) for eating a safe mushroom, a large negative reward (-35) for eating an unsafe mushroom, and zero reward for abstaining. We one-hot encode all categorical features, which results in 117-dimesional contexts.

Pharamcological Dosage Optimization Warfarin is common anticoagulant (blood thinner) that is prescribed to patients with atrial fibrillation to prevent strokes [65]. The optimal dosage varies from person to person and prescribing the incorrect dosage can have severe consequences. The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) includes a dataset with a 17-dimensional feature set containing numeric features including age, weight, and height, along with one-hot encoded categorical features indicating demographics and the presence of genetic markers. The dataset also includes the optimal dosage for each patient refined by physicians over time. We use this supervised dataset as a contextual bandit benchmark using the dosage as the action and defining the reward function to be the distance between the selected dosage and the optimal dosage. We discretize the action space into 20 (or 50) equally spaced dosage levels.

Wheel Bandit Problem The wheel bandit problem is a synthetic problem specifically designed to require exploration [45]. 2-dimensional contexts are sampled from inside the unit circle with uniform random probability. There are 5 actions where one action always has a mean reward of $\mathbb{E}[r(\boldsymbol{s}, a_1)] = 1.2$ independent of the context, and the mean rewards of the other actions depend on the context. If $||\boldsymbol{s}||_2 \leq \delta$, then the other 4 actions are non-optimal with a mean reward of 1. If $||\boldsymbol{s}||_2 > \delta$, then 1 of the 4 remaining actions is optimal—and determined by the sign of the two dimensions of $\boldsymbol{s}$ —with a mean reward of 50. The remaining 3 actions all have a mean reward of 1. All rewards are observed with zero-mean additive Gaussian noise with standard deviation $\sigma = 0.01$. We set $\delta = 0.95$, which means the probability of a sampling a context on the perimeter ($||\boldsymbol{s}||_2 \geq \delta$) where one action yields a large reward is $1 - (0.95)^2 = 0.0975$.

Real World Video Upload Transcoding Optimization We demonstrate performance of the imitation learning algorithm on a real world video upload transcoding application. At each time step, the policy receives a request to upload a video along with contextual features and the policy is tasked

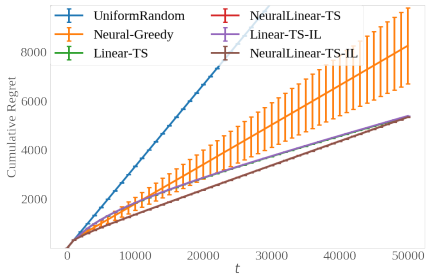

Figure 2: Cumulative regret on the Warfarin problem with 50 actions

with deciding how the video should be transcoded (e.g. what quality level to use) when uploading the video to the service. It is preferable to upload videos at a high quality because it can lead to a better viewer experience (if the viewer has a sufficiently good network connection). However, higher quality videos have larger file sizes. Uploading a large file is more likely to to fail than uploading a small file; uploading a larger file takes more time, which increases the likelihood that the network connection will drop or that person uploading the video will grow frustrated and cancel.

The contextual information accompanying each video includes dense and sparse features about: the video file (e.g. the raw bitrate, resolution, and file size) and the network connection (e.g. connection type, download bandwidth, country). There are 7 actions corresponding to a unique (resolution, bitrate) pairs. The actions are ranked ordered in terms of quality: action i yields a video with higher quality than action j if and only if $i \geq j$. The reward for a successful upload is a positive and monotonically increasing function of the action. The reward for a failed upload is 0.

We evaluate the performance of different contextual bandit algorithms using the unbiased, offline, policy evaluation technique proposed by Li et al. [34]. The method evaluates a CB algorithm by performing rejection sampling on a stream of logged observation tuples of the form $(\mathbf{x}_t, a_t, r_t)$ collected under a uniform random policy. Specifically, the observation tuple t is rejected if the logged action does not match the action selected by the CB algorithm being evaluated. For this demonstration we leverage a real video upload transcoding dataset containing 8M observations logged under a uniform random policy. We evaluate each algorithm using the stream of logged data until each algorithm has "observed" $T = 50,000$ *valid* time steps.

F.4 Additional Results

Warfarin - 50 Actions Figure 2 shows the cumulative regret on Warfarin using 50 actions. The imitation learning methods match the cumulative regret of the vanilla Thompson sampling methods.

G Time and Space Complexity

G.1 Complexity of Evaluated Methods

Table 2 shows the decision-making time complexity for the methods used in our empirical analysis. The time complexity is equivalent to the space complexity for all evaluated methods.

NEURALGREEDY The time complexity of NEURALGREEDY is the sum of matrix-vector multiplications involved in a forward pass.

LINEAR-TS The time complexity of LINEAR-TS is dominated by sampling from the joint posterior, which requires sampling from a multivariate normal with dimension d. To draw a sample from the joint posterior $P(\boldsymbol{\theta}, \sigma)$ at decision time, we first sample the noise level $\tilde{\sigma}^2 \sim \text{IG}(\alpha, \beta)$ and then sample $\tilde{\boldsymbol{\theta}} | \tilde{\sigma}^2 \sim \mathcal{N}\left(\boldsymbol{\mu}, \tilde{\sigma}^2 \Lambda^{-1}\right)$. Rather than inverting the precision matrix $\tilde{\Sigma} = \tilde{\sigma}^2 \Lambda^{-1}$, we compute root decomposition (e.g. a Cholesky decomposition) of the $d \times d$ precision matrix $\Lambda = LL^T$. The root decomposition can be computed once, with cost $O(d^3)$, after an offline batch update and cached until the next batch update. Given L^T, we sample directly by computing $\tilde{\boldsymbol{\theta}} = \boldsymbol{\mu} + \boldsymbol{z}$, where

$$\frac{1}{\tilde{\sigma}} L^T \boldsymbol{z} = \boldsymbol{\zeta} \tag{21}$$

and $\boldsymbol{\zeta} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Since L^T is upper triangular, Eqn. (21) can be solved using a backward substitution in quadratic time: $O(d^2)$.[2]

NEURALLINEAR-TS The time complexity of NEURALLINEAR-TS is the sum of a forward pass up to the last hidden layer and sampling from a multivariate normal with dimension h_M, where h_M is the size of the last hidden layer.

IL The IL methods have the same time complexity as NEURALGREEDY, ignoring the cost of sampling from multinomial with k categories.

G.2 Complexity Using Embedded Actions

An alternative modeling approach for the non-imitation methods is to embed the action with the context as input to the reward model.

NEURALGREEDY Using an embedded action, the time complexity for a forward pass up to the last layer is $O_{\text{last-layer}} = O\big(kd_a h_1 + k\sum_{m=1}^{M-1} h_m h_{m+1}\big)$ because the input at decision time is a $k \times d_a$ matrix where the context is embedded with each of the k actions and the each context-action vector has dimension d_a. The time complexity of computing the output layer remains $O(kh_M)$. The space complexity remains linear in the number of parameters, but it also requires computing temporary intermediate tensors of size $k \times h_m$ for $m = 1...M$: $O\big(d_a h_1 + \sum_{m=1}^{M-1} h_m h_{m+1} + \sum_{m=1}^{M} k h_m\big)$.

LINEAR-TS Linear-TS with an embedded action only requires using a single sample of the parameters, which yields a complexity of to $O(d_a^2 + kd_a)$ for LINEAR-TS. The space complexity is also $O(d_a^2 + kd_a)$.

NEURALLINEAR-TS For NEURALLINEAR-TS the time complexity of computing the outputs given the last hidden layer is $O(h_M^2 + kh_M)$, since only a single sample of h_M parameters is required for computed the reward for all actions. The space complexity for NEURALLINEAR-TS the sum the space complexities of NEURALGREEDY and LINEAR-TS.

IL The computatiuonal cost of the IL methods would be unchanged.

We choose to empirically evaluate models *without* embedded actions because linear methods using embedded actions cannot model reward functions that involve non-linear interactions between the contexts and actions, whereas modeling each action independently allows for more flexibility. Riquelme et al. [45] find that Thompson sampling using disjoint, exact linear bayesian regressions are a strong baseline in many applications. Furthermore, Riquelme et al. [45] observe that it is important to model the noise levels independently for each action.

G.3 Complexity of Alternative Methods

Alternative Thompson sampling methods including mean-field approaches, the low-rank approximations of the covariance matrix, and bootstrapping can also decrease the computational cost of posterior sampling. Mean-field approaches can reduce time complexity of sampling parameters from the posterior from quadratic $O(n^2)$ to linear $O(n)$ in the number of parameters n.[3] However, assuming independence among parameters has been observed to result in worse performance in some settings [45]. Low-rank approximations of the covariance matrix allow for sampling parameters in $O((n + 1)\rho)$, where ρ is the rank of the approximate covariance, but such methods have a space complexity of $O(\rho n)$ since they require storing ρ copies of the parameters [66, 36]. Bootstrapping also requires storing multiple copies of the parameters, so the space is $O(bn)$ where b is the number of bootstrap replicates. However, bootstrapping simply requires a multinomial draw to select one set of bootstrapped parameters. All these methods require a forward pass using the sampled parameters, and the time complexity is the sum of the time complexities of sampling parameters and the forward pass.

[2]The alternative approach of inverting the precision matrix to compute the covariance matrix $\Sigma = \Lambda^{-1}$, computing and caching its root decomposition $\Sigma = L_\Sigma L_\Sigma^T$, and sampling $\tilde{\boldsymbol{\theta}}$ as $\tilde{\boldsymbol{\theta}} = \boldsymbol{\mu} + L_\Sigma \boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ also has a time complexity of $O(d^2)$ from the matrix-vector multiplication $L_\Sigma \boldsymbol{\zeta}$.

[3]We describe space complexity in terms of the number of parameters n, so that we do not make assumptions about the underlying model.

Table 2. Decision-making time complexity and space complexity for each method . For methods relying on fully-connected neural networks, the time complexity of a forward pass to the last hidden layer is $C_{\text{last-layer}} = dh_1 + \sum_{m=1}^{M-1} h_m h_{m+1}$, where d is the dimension of the context and h_m is the number of units in hidden layer m. For BOOTSTRAP-NN-TS, B denotes the number of bootstrap replicates.

METHOD	TIME COMPLEXITY	SPACE COMPLEXITY
NEURALGREEDY	$O(C_{\text{LAST-LAYER}}) + O(kh_M)$	$O(C_{\text{LAST-LAYER}}) + O(kh_M)$
LINEAR-TS	$O(kd^2)$	$O(kd^2)$
NEURALLINEAR-TS	$O(C_{\text{LAST-LAYER}}) + O\big(kh_M^2\big)$	$O(C_{\text{LAST-LAYER}}) + O\big(kh_M^2\big)$
BOOTSTRAP-NN-TS	$O(C_{\text{LAST-LAYER}}) + O(kh_M)$	$O(C_{\text{LAST-LAYER}} \cdot B) + O(kh_M B))$
IL	$O(C_{\text{LAST-LAYER}}) + O(kh_M)$	