Reinforcement Learning with Thompson Sampling: No-Regret Performance over Finite Horizons

Jasmine Bayrooti
University of Cambridge
jgb52@cam.ac.uk

Sattar Vakili
MediaTek Research
sattar.vakili@mtkresearch

Amanda Prorok
University of Cambridge
asp45@cam.ac.uk

Carl Henrik Ek
University of Cambridge
Karolinska Institutet
che29@cam.ac.uk

Abstract

How should agents explore efficiently over extended horizons in sequential decision-making problems? While Thompson sampling (TS) is a widely used exploration strategy, its theoretical performance in reinforcement learning (RL) in the presence of complex temporal structure remains poorly understood. We take a step towards filling this gap by studying TS in episodic RL using models with Gaussian marginal distributions, namely joint Gaussian process (GP) priors over rewards and transitions. To characterize the impact of temporal structure on performance, we derive a regret bound of $\tilde{\mathcal{O}}(\sqrt{KH\Gamma(KH)})$ over K episodes of horizon H, where $\Gamma(\cdot)$ captures the complexity of the GP model. Our analysis addresses the way that uncertainty compounds through recursive updates and offers insights into how uncertainty and temporal structure influences exploration.

1 Introduction

Sequential decision-making under uncertainty lies at the core of reinforcement learning (RL) systems, from robotics [1] and chip design [2] to large language models [3]. In these settings, an agent must make a series of decisions where each choice influences future states and rewards, creating intricate temporal dependencies. Solving complex tasks often requires planning and acting over extended decision horizons. As this horizon grows, exploration becomes more challenging due to uncertainty compounding over time. This raises a central question: how does the length of the decision horizon affect the efficiency of exploration strategies in RL?

We focus on Thompson sampling (TS) which offers a simple and widely used strategy for exploration: sample from a posterior over models and act optimally under the sample [4]. This naturally balances exploration and exploitation and aligns exploration with the agent's uncertainty. TS underpins a range of applications including bandits [5, 6, 7], Bayesian optimization [8, 9], and reinforcement learning (RL) [10, 11, 12]. Empirically, TS has demonstrated strong performance across domains and is widely adopted in practice. Theoretical guarantees for TS have been well-developed in multi-armed bandit settings [13, 14] and have also been extended to RL settings [15, 11, 16, 17]. Existing RL analyses typically rely on discrete state-action spaces, assume linear or kernelized dynamics, or yield regret bounds that scale poorly with state dimensionality [15, 11, 16, 17]. This leaves an open question on how temporal structure, particularly extended horizons, affects TS performance in more general settings. Developing regret bounds for TS in continuous-state, finite-horizon MDPs without strong structural assumptions remains an important problem.

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).

In this work, we study TS-driven exploration in episodic RL with continuous state-action spaces and sequential decision-making under general models where each decision is described by a Gaussian distribution. This setting captures a wide range of problems, including those modeled with Gaussian processes (GPs), a standard tool in Bayesian optimization and RL. Specifically, rewards and transitions are jointly modeled using a multi-output GP, as proposed in [12]. This enables the agent to model correlations across different components of the environment in a flexible and data-efficient manner. We consider an episodic MDP with K episodes of horizon H, where at the start of each episode, the agent samples a realization from the GP posterior and computes an optimal policy with respect to this sample. We use regret, defined as the cumulative loss in value relative to the optimal policy, as a tool to characterize performance over extended horizons. Specifically, we provide a stylized analytical upper bound on the regret, showing its dependence on the horizon, number of episodes, and the complexity of the GP kernel. Our analysis highlights the key theoretical challenges of applying TS in sequential settings, particularly due to the recursive and compositional nature of value functions.

1.1 Contributions

We establish a sublinear regret bound for TS in model-based RL under GP models, an approach we refer to as *Reinforcement Learning with GP Sampling (RL-GPS)*. We prove a regret bound of $\tilde{\mathcal{O}}(\sqrt{KH\Gamma(KH)})$ over K episodes of horizon H, where $\Gamma(\cdot)$ captures the complexity of the GP.

Our theoretical analysis introduces intermediate contributions for deriving the final regret bound. Extending TS regret guarantees to RL presents two main challenges: (i) the optimal value function is a recursive composition of GPs which is not a GP itself [18]; and (ii) TS operates on proxy models induced by Bellman updates, rather than sampling directly from the posterior over the optimal value function. To address (i), we develop high-probability confidence bounds for compositional functions of GPs, formalized in Theorem 1, which allow us to control value estimation errors across the decision horizon. To address (ii), we apply these bounds to derive high-probability confidence intervals for value functions in episodic MDPs, which inherently take a recursive and compositional form (Corollary 1). We then bound the regret in terms of cumulative posterior uncertainty. Here, we derive a new multi-output *elliptical potential lemma* (Lemma 1) to bound this quantity. This result improves upon naively applying the standard versions [8, 19, 20] independently to each output dimension. Instead, our bound jointly tracks uncertainty across multiple outputs, leveraging their correlation structure to obtain tighter regret guarantees.

Finally, we conduct controlled experiments that mirror our theoretical assumptions and validate our regret bounds. The results confirm sublinear cumulative regret on GP-sampled MDPs and sparse navigation tasks. Furthermore, we empirically illustrate how the choice of GP kernel affects learning efficiency, with smoother kernels such as Radial Basis Function (RBF) leading to faster regret decay in smooth environments, and rougher Matérn kernels outperforming in sparse settings. These findings are consistent with the theoretical dependence on model complexity and temporal structure.

1.2 Related work

RL with TS. The theoretical performance of Thompson sampling (TS) has been extensively studied in the bandit setting, where it achieves near-optimal regret bounds and strong empirical performance [13, 21, 6, 22, 23, 7, 24]. In the GP bandit setting, regret bounds were established for TS under kernel-based assumptions on the target functions, using techniques that our analysis builds on [9]. Extending TS to RL introduces new challenges due to the recursive structure of value functions and the dependence on both states and actions. Several works have established foundational regret guarantees for TS in finite MDPs [15, 25, 26, 27]. In particular, regret analyses for the Posterior Sampling for Reinforcement Learning (PSRL) approach [15] showed that PSRL achieves $\tilde{\mathcal{O}}(H\sqrt{SAT})$ Bayesian regret, where H is the horizon, S is the number of states, A is the number of actions, and T is the total number of steps [11]. Beyond the tabular setting, a TS method with tighter guarantees under linearity assumptions was introduced, although their bounds scale poorly with state dimensionality [17]. Regret bounds have been derived in kernelized RL settings where transition dynamics are modeled as functions in a reproducing kernel Hilbert space (RKHS), with the bounds depending on the maximum information gain of the kernel [16]. Our work differs fundamentally in modeling assumptions as we model both the reward and transition functions jointly as a multi-output GP. As a result, much of our analysis is novel.

Episodic MDP. The episodic MDP framework is a central setting for RL. Sublinear regret bounds have been established for Upper Confidence Bound (UCB) methods in tabular finite-horizon MDPs [28, 29, 30]. Subsequent works have developed regret analyses for UCB-style methods with structural assumptions, including linear [31] and kernelized MDPs [16, 32, 33], as well as under standard assumptions for GP models [34]. These approaches all rely on optimism via constructed confidence sets to guide exploration. In contrast, TS offers an exploration approach that avoids explicit construction of confidence sets and has been less theoretically studied in complex RL settings.

Broader RL settings. Beyond episodic MDPs, it is common to study performance in infinite-horizon discounted [35] and average-reward settings [36]. In the discounted setting, the contraction properties of the Bellman operator enable provably efficient learning [37]. In the infinite-horizon average-reward setting, regret bounds have been established for tabular communicating MDPs with finite diameter using UCB-based strategies [36, 38, 39, 40]. More recent works established regret guarantees in structured infinite-horizon settings for linear mixture MDPs with known feature mappings [41], linear function approximation in weakly communicating MDPs [42], and for nonepisodic RL under continuity and bounded energy assumptions [43]. These approaches rely on structural assumptions with resulting regret bounds depending explicitly on their properties. By focusing on the episodic setting, we sidestep these complications and exploit the finite-horizon structure to derive regret bounds using GP-based recursive analysis.

2 Problem formulation

Reinforcement learning. An episodic MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, f_{\mathrm{R}}, f_{\mathrm{S}}, H)$, where $\mathcal{S} \subset \mathbb{R}^{d_{\mathrm{S}}}$ is the state space, $\mathcal{A} \subset \mathbb{R}^{d_{\mathrm{A}}}$ is the action space, and H is the episode length. The reward function is $f_{\mathrm{R}} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, and the state transition function is $f_{\mathrm{S}} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$. The policy $\pi = \{\pi_h : \mathcal{S} \mapsto \mathcal{A}\}_{h=1}^{H}$ specifies the action $\pi_h(s)$ the agent takes in state s at step h. At the start of each episode k, the environment provides an initial state $s_{1,k}$ and the agent executes a policy $\pi_k = \{\pi_{h,k}\}_{h=1}^{H}$. At each step h, the agent observes the state $s_{h,k}$, selects action $a_{h,k} = \pi_{h,k}(s_{h,k})$, receives reward $f_{\mathrm{R}}(s_{h,k}, a_{h,k})$, and transitions to the new state $s_{h+1,k} = f_{\mathrm{S}}(s_{h,k}, a_{h,k})$.

In an episodic MDP, the agent aims to maximize the cumulative reward collected over an episode. To formalize this, we define the *value function* of a policy π as the expected total reward obtained when starting at state s at step h and following π thereafter:

$$V_h^{\pi}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} f_{\mathrm{R}}(s_{h'}, a_{h'}) \mid s_h = s\right], \ \forall s \in \mathcal{S}, \ h \in [H].$$

The associated *state-action value function* is defined as:

$$Q_h^{\pi}(s, a) = \mathbb{E}\left[\sum_{h'=h}^{H} f_{\mathrm{R}}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right].$$

We assume the existence of an optimal policy $\pi^\star$ that maximizes the expected total reward from any state and time step. The optimal value and optimal state-action value functions are defined as: $V_h^\star(s) = \max_\pi V_h^\pi(s)$, $Q_h^\star(s,a) = \max_\pi Q_h^\pi(s,a)$. The optimal value function satisfies the Bellman optimality equation: $Q_h^\star(s,a) = f_{\mathrm{R}}(s,a) + V_{h+1}^\star(f_{\mathrm{S}}(s,a))$, $V_h^\star(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a)$, with $V_{H+1}^\star(s) = 0$ for all $s \in \mathcal{S}$. An RL algorithm aims to find a near-optimal policy while interacting with the environment. The *regret* over T timesteps is defined as:

$$\mathrm{Regret}(T) = \sum_{k=1}^{K} (V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k})), \tag{1}$$

where π_k is the policy executed by the agent in episode k, and $s_{1,k}$ is the initial state of that episode, and we use $T = KH$ for the total number of steps.

Gaussian process modeling. GPs specify distributions over the space of functions, offering calibrated uncertainty estimates that can be leveraged for exploration and decision making. In the single-output case, we model an unknown function $f : \mathcal{Z} \to \mathbb{R}$ as a Gaussian process $f \sim \mathrm{GP}(0, k)$

Algorithm 1 RL with GP Sampling (RL-GPS)

Require: number of episodes K, episode length H, GP kernel k
Initialize: reward-dynamics model $p(f)$, reward and transition buffer $\mathcal{D}$
for episode $k = 1, \ldots, K$ **do**
    // Create the proxy value functions $Q_{h,k}$
    Sample functions $[\hat{f}_{\mathrm{R},k}, \hat{f}_{\mathrm{S},k}]$ from GP posterior $p(f \mid \mathcal{D})$
    Initialize $V_{H+1,k}(\cdot) = 0$
    for $h = H, \ldots, 1$ **do**
        $Q_{h,k}(s,a) = \hat{f}_{\mathrm{R},k}(s,a) + V_{h+1}(\hat{f}_{\mathrm{S},k}(s,a))$
        $V_{h,k}(s) = \max_{a \in \mathcal{A}} Q_{h,k}(s,a)$
    // Follow the greedy policy with respect to $Q_{h,k}$
    Observe initial state $s_{1,k}$
    for $h = 1, \ldots, H$ **do**
        Select action $a_{h,k} = \operatorname{argmax}_{a \in \mathcal{A}} Q_{h,k}(s_{h,k}, a)$
        Observe next state $s_{h+1,k} = f_{\mathrm{S}}(s_{h,k}, a_{h,k})$ and reward $r_{h,k} = f_{\mathrm{R}}(s_{h,k}, a_{h,k})$
        Store reward and transition in buffer $(s_{h,k}, a_{h,k}, r_{h,k}, s_{h+1,k}) \to \mathcal{D}$
    Update GP posterior $p(f \mid \mathcal{D})$ using new transitions

with a scalar-valued kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Given n noisy observations $\{(z_i, y_i)\}_{i=1}^n$ with $y_i = f(z_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \lambda^2)$, the posterior mean and variance at any test point $z \in \mathcal{Z}$ are:

$$\mu_n(z) = \mathbf{k}_n^\top (\mathbf{K}_n + \lambda^2 \mathbf{I}_n)^{-1} \mathbf{y}_n, \quad \sigma_n^2(z) = k(z,z) - \mathbf{k}_n^\top (\mathbf{K}_n + \lambda^2 \mathbf{I}_n)^{-1} \mathbf{k}_n,$$

where $\mathbf{K}_n \in \mathbb{R}^{n \times n}$ is the kernel matrix with $[\mathbf{K}_n]_{ij} = k(z_i, z_j)$, $\mathbf{k}_n(z) \in \mathbb{R}^n$ has entries $k(z_i, z)$, and $\mathbf{y}_n \in \mathbb{R}^n$ is the vector of observed outputs.

Multi-output Gaussian processes. In many applications, we wish to model a vector-valued function $f : \mathcal{Z} \to \mathbb{R}^d$ jointly across multiple correlated outputs. In this setting, f is modeled as a multi-output GP where $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^{d \times d}$ is a matrix-valued positive semidefinite kernel that encodes both input similarity and output correlations. Given n input points $z_1, \ldots, z_n$, let the observed outputs be collected into a vector $\mathbf{y}_n \in \mathbb{R}^{nd}$ by stacking all d outputs at each input. The full joint prior over $\mathbf{y}_n$ is a multivariate Gaussian with zero mean and block kernel matrix $\mathbf{K}_n \in \mathbb{R}^{nd \times nd}$ defined by: $\mathbf{K}_n[(i-1)d + r, (j-1)d + s] = [k(z_i, z_j)]_{rs}$ for $i, j \in [n],\ r, s \in [d]$.

The posterior at test point z is Gaussian with mean $\mu_n(z) \in \mathbb{R}^d$ and covariance $\Sigma_n(z) \in \mathbb{R}^{d \times d}$ given by:

$$\mu_n(z) = \mathbf{k}_n(z)^\top (\mathbf{K}_n + \lambda^2 \mathbf{I}_{nd})^{-1} \mathbf{y}_n, \quad \Sigma_n(z) = k(z,z) - \mathbf{k}_n(z)^\top (\mathbf{K}_n + \lambda^2 \mathbf{I}_{nd})^{-1} \mathbf{k}_n(z),$$

where $\mathbf{k}_n(z) \in \mathbb{R}^{nd \times d}$ is the cross-covariance between $f(z)$ and the training outputs, defined via $[\mathbf{k}_n(z)]_{(i-1)d+r,s} = [k(z_i, z)]_{rs}$. We define $\sigma_n^2(z) := \operatorname{diag}(\Sigma_n(z)) \in \mathbb{R}^d$ as the marginal predictive variances. With slight abuse of notation, we use $\sigma_n(z)$ to denote the vector of marginal standard deviations, where $\sigma_{n,i}(z) = (\sigma_{n,i}^2(z))^{1/2}$ for $i = 1, \ldots, d$. The posterior mean $\mu_n(z)$ and covariance $\Sigma_n(z)$ enable multi-output GPs to jointly model transitions and rewards with uncertainty, making them well-suited for RL where both functions must be estimated simultaneously.

3 Reinforcement learning with GP sampling

In this section, we present RL-GPS for learning episodic MDPs with joint GP modeling of the reward and transition functions, following a recently proposed model [12].

Assumption 1. *Let $f = [f_R, f_S]$ denote the joint reward and transition function. We assume f is distributed as a multi-output GP: $f \sim GP(\mathbf{0}, k)$, for a known kernel $k : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}^{d \times d}$.*

The RL-GPS algorithm follows a value-iteration form of TS, where at the start of each episode, the agent samples a realization of the reward and transition functions from the GP posterior and computes proxy value functions $Q_{h,k} : \mathcal{Z} \mapsto \mathbb{R}$ via backward induction. The agent then executes the greedy policy under the function during the episode. This approach encourages exploration by introducing

structured randomness into value estimates, naturally balancing exploitation of high-reward regions with exploration of uncertain areas of the state-action space. Pseudocode is given in Algorithm 1.

4 Analysis

In this section, we derive a regret bound for RL-GPS (Algorithm 1) in episodic MDPs with a multi-output GP model. We introduce intermediate results and provide the regret bound in Theorem 2.

4.1 Confidence intervals

To analyze the regret, we require high-probability bounds on the accuracy of GP predictions. For a single-output GP f with posterior mean μ_n and standard deviation σ_n, the tail decay of the Gaussian distribution implies that, with probability at least $1-\delta$, the following holds uniformly over z:

$$|f(z) - \mu_n(z)| \leq \beta_n(\delta)\sigma_n(z) \tag{2}$$

where $\beta_n(\delta) = \mathcal{O}(\sqrt{\log(\frac{n}{\delta})})$ [e.g., see 8].

In the RL setting, we must also construct confidence intervals for $v(f_S(\cdot))$ as part of the policy's recursive design where $v : \mathcal{S} \mapsto \mathbb{R}$ is a generic value function. The following theorem addresses this.

Theorem 1 (Confidence bounds for composed GPs)**.** *Assume $f : \mathcal{Z} \mapsto \mathcal{S} \subset \mathbb{R}^{d_S}$ is a multi-output GP with posterior mean μ_n and standard deviation σ_n. Let $v : \mathcal{S} \mapsto \mathbb{R}$ be a twice differentiable value function where for all $s \in \mathcal{S}$, $\|\nabla v(s)\| \leq u_G$ and $\|\nabla^2 v(s)\|_{op} \leq u_H$. Define the composition $g(z) = v(f(z))$. We have, with probability $1-\delta$, for all $z \in \mathcal{Z}$,*

$$|g(z) - v(\mu_n(z))| \leq u_G \beta_n(\delta/d_S)\|\sigma_n(z)\| + \frac{1}{2} u_H \beta_n(\delta/d_S)^2 \|\sigma_n(z)\|^2. \tag{3}$$

The proof uses a Taylor expansion of v to bound $|g(z) - v(\mu_n(z))|$ in terms of the first- and second-order behavior of v, together with the standard GP confidence intervals given in (2). A detailed proof is provided in Appendix A.

4.2 Performance analysis of RL-GPS

We first introduce our assumption regarding the smoothness of the value functions.

Assumption 2 (Smoothness of the value functions)**.** *We assume that for all h, V_h is twice differentiable where for all s, $\|\nabla V_h(s)\| \leq u_G$ and $\|\nabla^2 V_h(s)\|_{op} \leq u_H$.*

Remark 1. *The assumptions on the gradient and Hessian norms are mild compared to those typically imposed on value functions in the literature. For example, in [31] and [44], it is assumed that all proxy value functions belong to a function class defined using linear or kernel-based models, respectively. In contrast, we impose a weaker assumption only on the first and second derivatives.*

Now we present the main theorem bounding the regret of RL-GPS.

Theorem 2. *Consider the episodic MDP setting described in Section 2 and the RL-GPS algorithm given in Algorithm 1.Under Assumptions 1 and 2, with probability $1-\delta$,*

$$\textit{Regret}(T) = \mathcal{O}\left(\log(Td/\delta)\sqrt{T\Gamma(T)}\right),$$

where $\Gamma(T) = \sup_{z_{h,k}, h\in[H], k\in[K]} \mathcal{I}_T$, $\mathcal{I}_T := \frac{1}{2}\log\det(\mathbf{I}_{Td} + \frac{1}{\lambda^2}\mathbf{K}_T)$.

The determinant in $\Gamma(\cdot)$ represents the complexity of the function space described by the GP [45] and serves as an upper bound on the information gain, which is discussed in detail in the next section. The regret bound holds with high probability, where the randomness accounts for both the joint GP distribution of the environment and the randomness in the Thompson samples.

Remark 2. *The Matérn family is both theoretically significant and practically prevalent among kernel choices. By substituting the bounds on $\Gamma(\cdot)$ from [46], we obtain the following regret rates. For a base Matérn kernel with smoothness parameter $\nu > 1$, our regret bound becomes:*

$$\textit{Regret}(T) = \tilde{\mathcal{O}}\left(T^{\frac{\nu+d}{2\nu+d}}\right),$$

while for the radial basis function (RBF) kernel, the regret bound simplifies to $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$.

Remark 3. *When $H = 1$, learning in episodic MDPs reduces to the degenerate special case of Bayesian optimization, also known as GP bandits. In this setting, we have $T = K$ and our regret bound becomes:*

$$\textit{Regret}(T) = \mathcal{O}\left(\log(T/\delta)\sqrt{T\Gamma(T)}\right),$$

which recovers the standard regret bounds in Bayesian optimization [e.g., see 8].

4.3 Proof of Theorem 2

We bound total regret by analyzing the per-episode difference between the optimal value function and that of RL-GPS. The full proof is in Appendix B and proceeds in four steps:

Step 1: Regret decomposition. We decompose the per-step regret into two parts: immediate regret due to TS and a recursive component from propagating value uncertainty through transitions:

$$V_h^\star(s_{h,k}) - V_h^{\pi_k}(s_{h,k}) = \underbrace{Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a_{h,k})}_{\text{Immediate regret}} + \underbrace{V_{h+1}^\star(s_{h+1,k}) - V_{h+1}^{\pi_k}(s_{h+1,k})}_{\text{Recursive term}}. \quad (4)$$

Step 2: Bounding the immediate regret. We build on techniques from the analysis of TS in GP bandits [9], but face two key challenges: (i) the target function $Q_h^\star(s,a) = f_{\mathrm{R}}(s,a) + V_{h+1}^\star(f_{\mathrm{S}}(s,a))$ is recursive and cannot be directly modeled as a GP; (ii) TS samples from the posterior of a proxy $Q_{h,k}$, not the true posterior over $Q_h^\star$. To address these issues, we construct high-probability upper and lower confidence bounds on the proxy and target value functions.

Definition 1 (Upper and lower confidence bounds for value functions)**.** *We define the upper confidence bounds recursively as:*

$$Q_{h,k}^u(s,a) = \mu_{R,k}(s,a) + V_{h+1,k}^u(\mu_{S,k}(s,a)) + \xi_k(s,a), \quad V_{h,k}^u(s) = \max_{a\in\mathcal{A}} Q_{h,k}^u(s,a),$$

with $V_{H+1,k}^u(s) = 0$. Similarly, the lower confidence bounds are defined as:

$$Q_{h,k}^l(s,a) = \mu_{R,k}(s,a) + V_{h+1,k}^l(\mu_{S,k}(s,a)) - \xi_k(s,a), \quad V_{h,k}^l(s) = \max_{a\in\mathcal{A}} Q_{h,k}^l(s,a),$$

with $V_{H+1,k}^l(s) = 0$. The confidence width $\xi_k(s,a)$ is given by:

$$\xi_k(s,a) = \beta_k(\delta/Td)\sigma_{R,k}(s,a) + u_G\beta_k(\delta/Td)\|\sigma_{S,k}(s,a)\| + \frac{1}{2}u_H\beta_k(\delta/Td)^2\|\sigma_{S,k}(s,a)\|^2,$$

based on Theorem 1 where $\mu_{R,k}$, $\sigma_{R,k}$ and $\mu_{S,k}$, $\sigma_{S,k}$ are the posterior mean and standard deviation of f_{R} and f_{S}, respectively.

Corollary 1. *With probability at least $1-\delta$, we have:*

$$Q_{h,k}^l(s,a) \le Q_{h,k}(s,a),\ Q_h^\star(s,a) \le Q_{h,k}^u(s,a), \quad \forall(h,k,s,a).$$

Step 3: Accumulating episode regret. Unrolling the decomposition over $h = 1,\dots,H$ yields a per-episode bound in terms of the confidence widths: for a constant c,

$$\text{Regret}(T) \le c\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_k(s_{h,k},a_{h,k}).$$

Step 4: Bounding cumulative regret. Summing over K episodes leads to total regret bounds involving posterior variances. These are controlled using a new elliptical potential lemma for multi-output GPs (Lemma 1) and its delayed-update extension (Lemma 2). The latter accounts for the batched nature of episode-level GP updates, introducing dependence on H. We apply techniques from delayed feedback GP analysis [47, 48] to obtain the final bound.

4.4 Elliptical potential lemma for multi-output GPs

A key term in bounding cumulative regret is the sum of posterior variances along a sequence of inputs. For scalar-output GPs, this is bounded by a log-determinant term (sometimes referred to as the *elliptical potential lemma* [8, 20]): $\sum_{t=1}^{T}\sigma_{t-1}^2(z_t) \le C\log\det\left(\mathbf{I}_{Td} + (1/\lambda^2)\mathbf{K}_T\right)$.

In our setting, the transition function is modeled as a multi-output GP. Applying the scalar result separately to each dimension leads to suboptimal dependence on the state dimension d_{S}. To avoid this, we derive a version tailored for vector-valued GPs:

Lemma 1 (Elliptical potential lemma for multi-output GPs)**.** *Let $f : \mathcal{Z} \to \mathbb{R}^d$ be modeled by a multi-output GP with kernel k. For inputs $z_1, \ldots, z_T$, let $\sigma_{t-1}(z_t) \in \mathbb{R}^d$ denote the posterior standard deviations. Then,*

$$\sum_{t=1}^{T} \|\sigma_{t-1}(z_t)\|^2 \leq C\mathcal{I}_T, \quad \textit{where} \quad \mathcal{I}_T = \frac{1}{2}\log\det\left(\mathbf{I}_{Td} + \lambda^{-2}\mathbf{K}_T\right), \quad \textit{and} \quad C = \frac{2}{\log(1+\lambda^{-2})}.$$

The proof, given in Appendix C.2, extends the classical scalar result to the multi-output setting.

5 Experiments

To empirically validate the regret scaling in Theorem 2, we study how kernel complexity impacts regret in synthetic MDPs generated from GP-sampled environments. The state and action spaces are continuous, $\mathcal{S} = [0,1]^2$ and $\mathcal{A} = [0,1]$, but each dimension is discretized into 25 equally sized bins to enable tractable value function approximation. We compare three common kernel functions: the Radial Basis Function (RBF) kernel and Matérn kernels with smoothness parameters $\nu = 2.5$ and $\nu = 1.5$. For the multi-output GP model, we use the popular *linear model of coregionalization* (LMC) [49, 50], which predicts the vector-valued output as a linear combination of independent latent GPs (see Appendix D for details). For each kernel, a sparse LMC multi-output GP with zero mean and fixed linear correlations is used to sample the ground-truth reward and transition functions, f_{R} and f_{S} with rewards normalized to $r \in [0,1]$ per step, thereby defining the MDP. The optimal value function $V^\star$ is computed using finite-horizon value iteration with $H = 20$. Algorithm 1 is run for $K = 1000$ episodes and cumulative regret relative to the optimal value function is quantified. The results are averaged over 200 randomly sampled environments and shown in Figure 1. Across all kernels, cumulative regret grows sublinearly, which is consistent with our theoretical analysis. Performance varies with the complexity of the kernel: the RBF kernel yields the lowest regret, followed by Matérn $\nu = 2.5$ and then Matérn $\nu = 1.5$, as predicted by Remark 2. This reflects that rougher kernels correspond to more complex function classes and require more data to accurately estimate value functions. These results empirically confirm the theoretical link between kernel smoothness and learning efficiency that our analysis elucidates through the regret bound's dependence on information gain $\Gamma(\cdot)$. Additional experiments and information are given in Appendix E.

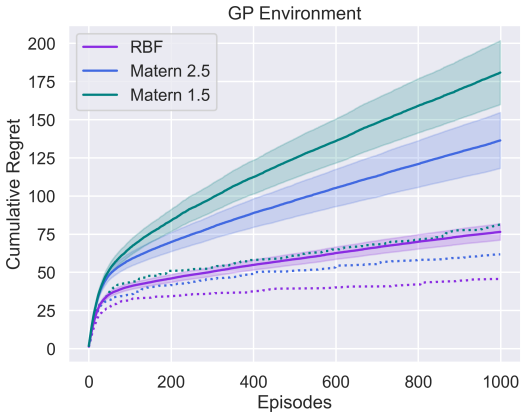

Figure 1: Cumulative regret over different kernels on GP-sampled environments over 200 trials. The shaded region around each curve represents ± 1 standard error of the mean across trials. Dotted lines represent median regrets.

6 Conclusion

This work investigated the performance of Thompson sampling in episodic reinforcement learning with extended decision horizons. Using a multi-output Gaussian Process model over rewards and transitions, we provided theoretical evidence that learning efficiency depends on the horizon length, number of episodes, and complexity of the underlying model. The sublinear regret bound, $\tilde{\mathcal{O}}(\sqrt{KH\Gamma(KH)})$, serves as a characterization of these relationships. To derive this bound, we introduced new tools for bounding uncertainty in recursive value functions, including confidence intervals for compositional functions of GPs and a multi-output elliptical potential lemma that captures correlations across components. Our experiments on synthetic tasks validated our theoretical predictions and highlighted how GP kernel structure influences learning dynamics.

Overall, this work underscores the role of temporal structure and posterior uncertainty in exploration efficiency and provides a foundation for further analysis of TS in sequential settings. Future directions for research include extending these results to non-Gaussian or learned model classes as well as

infinite-horizon problems. More broadly, understanding how exploration performance scales with horizon length remains a key challenge in reinforcement learning with implications for designing agents that can reason effectively over long time scales in complex environments.

Acknowledgments

J. Bayrooti is supported by a DeepMind scholarship. A. Prorok is supported in part by European Research Council (ERC) Project 949940 (gAIa).

References

[1] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[2] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Shen Song, Eric Wang, Young-Joon Lee, James Johnson, Arvind Pathak, Tom Duerig, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[4] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[5] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pages 2249–2257, 2011.

[6] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[7] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[8] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, page 1015–1022, 2010.

[9] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

[10] Remo Sasso, Michelangelo Conserva, and Paulo Rauber. Posterior sampling for deep reinforcement learning. In *International Conference on Machine Learning*, pages 30042–30061. PMLR, 2023.

[11] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017.

[12] Jasmine Bayrooti, Carl Henrik Ek, and Amanda Prorok. Efficient model-based reinforcement learning through optimistic thompson sampling. *International Conference on Learning Representations*, 2025.

[13] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[14] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[15] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[16] Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3197–3205. PMLR, 2019.

[17] Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 12040–12051, 2021.

[18] Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 207–215. PMLR, 2013.

[19] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[20] Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

[21] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[22] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013.

[23] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[24] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

[25] Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[26] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

[27] Daniel Russo. Worst-case regret bounds for exploration via thompson sampling. In *Advances in Neural Information Processing Systems*, pages 14196–14206, 2019.

[28] 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.

[29] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[30] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

[31] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[32] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[33] Sattar Vakili and Julia Olkhovskaya. Kernelized reinforcement learning with order optimal regret bounds. In *Advances in Neural Information Processing Systems*, volume 36, pages 4225–4247, 2023.

[34] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *Advances in Neural Information Processing Systems*, volume 33, pages 14156–14170, 2020.

[35] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[36] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[37] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[38] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, volume 11, pages 1563–1600, 2010.

[39] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[40] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

[41] Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3883–3913. PMLR, 2022.

[42] Arnob Ghosh and Xingyu Zhou. Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. *International Conference on Learning Representations*, 2023.

[43] Bhavya Sukhija, Lenart Treven, Florian Dörfler, Stelian Coros, and Andreas Krause. NeoRL: Efficient exploration for nonepisodic RL. In *Advances in Neural Information Processing Systems*, volume 37, pages 74966–74998, 2024.

[44] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. In *Advances in Neural Information Processing Systems*, volume 33, pages 13903–13916, 2020.

[45] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[46] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

[47] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Near-linear time gaussian process optimization with adaptive batching and resparsification. In *International Conference on Machine Learning*, pages 1295–1305. PMLR, 2020.

[48] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Scaling Gaussian process optimization by evaluating a few unique candidates multiple times. In *International Conference on Machine Learning*, pages 2523–2541. PMLR, 2022.

[49] Michel Grzebyk and Hans Wackernagel. Multivariate analysis and spatial/temporal scales: real and complex models. In *International Biometrics Conference*, volume 1, pages 19–33, 1994.

[50] Hans Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2003.

[51] Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of GP-UCB. In *Advances in Neural Information Processing Systems*, volume 36, pages 35266–35276, 2023.

[52] Kingma Diederik and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[53] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A Proof of Theorem 1

In this section, we provide a detailed proof of Theorem 1. Recall from (2) that for a single-output GP f with posterior mean and standard deviation μ_n and σ_n, we have, with probability $1-\delta$, uniformly in z,

$$|f(z) - \mu_n(z)| \leq \beta_n(\delta)\sigma_n(z) \tag{5}$$

where $\beta_n(\delta) = \mathcal{O}(\sqrt{\log(\frac{n}{\delta})})$.

To extend this to a composition $v(f(z))$, where $f(z) \in \mathbb{R}^d$ is drawn from a multi-output GP and $v : \mathbb{R}^d \to \mathbb{R}$ is a smooth function, we apply Taylor's theorem. For any $z \in \mathcal{Z}$, there exists a point ζ on the line segment connecting $f(z)$ and $\mu_n(z)$ such that:

$$v(f(z)) = v(\mu_n(z)) + \nabla v(\mu_n(z))^\top (f(z) - \mu_n(z)) + \frac{1}{2}(f(z) - \mu_n(z))^\top \nabla^2 v(\zeta)\,(f(z) - \mu_n(z)).$$

Taking absolute values and using the bounds on gradient and Hessian

$$\begin{aligned}|v(f(z)) - v(\mu_n(z))| &\leq \|\nabla^\top v(\mu_n(z))\|\|f(z) - \mu_n(z)\| + \frac{1}{2}\|\nabla^2 v(\zeta)\|_{\text{op}}\|f(z) - \mu_n(z)\|^2 \\ &\leq u_G\|f(z) - \mu_n(z)\| + \frac{1}{2}u_H^2\|f(z) - \mu_n(z)\|^2\end{aligned}$$

By the standard single GP confidence bound (2), with probability $1 - \delta/d_S$, we have $|f_j(z) - \mu_{n,j}(z)| \leq \beta_n(\delta/d_S)\sigma_{n,j}(z)$ for all j, and hence, applying a probability union bound, with probability $1-\delta$,

$$\|f(z) - \mu_n(z)\| \leq \beta_n(\delta/d_s)\|\sigma_n(z)\|.$$

Substituting into the bound above, we obtain

$$|v(f(z)) - v(\mu_n(z))| \leq u_G\beta_n(\delta/d_s)\|\sigma_n(z)\| + \frac{1}{2}u_H^2\beta_n(\delta/d_s)^2\|\sigma_n(z)\|^2,$$

that completes the proof of Theorem 1.

B Proof of Theorem 2

In this section, we provide a detailed proof of Theorem 2 on the regret performance of RL-GPS. We bound the total regret by analyzing the per-episode difference between the optimal value function and the value function of the GP-TS-RL algorithm. We structure the proof in four main steps.

First, we decompose the per-step regret into two components: an immediate regret term arising from TS, and a recursive term capturing uncertainty in value propagation through the transition model.

Second, we bound the immediate regret using techniques inspired by those used in the analysis of TS in GP bandits. There are however certain challenges which are discussed below.

Third, we unroll the recursion and accumulate these bounds over all steps within an episode.

Fourth, we bound the cumulative regret by bounding the sum of posterior standard deviations in GPs, which appear in the third step, to complete the regret bound.

First step: Decomposing the per-step regret.

We begin by analyzing the regret incurred at step h of episode k. Let $a_{h,k} = \pi_k(s_{h,k})$ denote the action taken by the algorithm, and let $a^\star_{h,k} = \arg\max_{a\in\mathcal{A}} Q^\star_h(s_{h,k}, a)$ denote the optimal action at

that state. The per-step regret is defined as the difference between the optimal and executed value functions:

$$V_h^\star(s_{h,k}) - V_h^{\pi_k}(s_{h,k}).$$

By the Bellman equation, this can be written as:

$$\begin{aligned}
&V_h^\star(s_{h,k}) - V_h^{\pi_k}(s_{h,k}) \\
&= \left(f_{\mathrm{R}}(s_{h,k}, a_{h,k}^\star) + V_{h+1}^\star(f_{\mathrm{S}}(s_{h,k}, a_{h,k}^\star)) - \left(f_{\mathrm{R}}(s_{h,k}, a_{h,k}) + V_{h+1}^{\pi_k}(f_{\mathrm{S}}(s_{h,k}, a_{h,k}))\right)\right) \\
&= \left(f_{\mathrm{R}}(s_{h,k}, a_{h,k}^\star) + V_{h+1}^\star(f_{\mathrm{S}}(s_{h,k}, a_{h,k}^\star)) - \left(f_{\mathrm{R}}(s_{h,k}, a_{h,k}) + V_{h+1}^\star(f_{\mathrm{S}}(s_{h,k}, a_{h,k}))\right)\right) \\
&\quad + \left(V_{h+1}^\star(f_{\mathrm{S}}(s_{h,k}, a_{h,k})) - V_{h+1}^{\pi_k}(f_{\mathrm{S}}(s_{h,k}, a_{h,k}))\right) \\
&= Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a_{h,k}) + \left(V_{h+1}^\star(s_{h+1,k}) - V_{h+1}^{\pi_k}(s_{h+1,k})\right), \qquad (6)
\end{aligned}$$

where the first equality follows from the definition of the value function, the second adds and subtracts the term $V_{h+1}^\star(f_{\mathrm{S}}(s_{h,k}, a_{h,k}))$, and the third rewrites the expression using the definition of $Q_h^\star$ and noting $s_{h+1,k} = f_{\mathrm{S}}(s_{h,k}, a_{h,k})$.

We split this expression into two terms:

$$V_h^\star(s_{h,k}) - V_h^{\pi_k}(s_{h,k}) = \underbrace{Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a_{h,k})}_{\text{(immediate regret)}} + \underbrace{\left(V_{h+1}^\star(s_{h+1,k}) - V_{h+1}^{\pi_k}(s_{h+1,k})\right)}_{\text{(recursive part of regret)}}. \qquad (7)$$

The first term captures the immediate regret incurred by TS. The second term reflects the recursive component of regret, which arises due to uncertainty in the transition model and its impact on value propagation. We next proceed to bound the immediate regret term.

Second step: Bounding the immediate regret (TS).

This term captures the suboptimality of the action $a_{h,k}$ chosen by TS, relative to the optimal action $a_{h,k}^\star$, under the target function $Q_h^\star(\cdot, \cdot)$. The analysis presents two key challenges compared to the standard Thompson sampling analysis in kernel bandits [9]:

1. The target function $Q_h^\star(\cdot, \cdot) = f_{\mathrm{R}}(\cdot, \cdot) + V_{h+1}^\star(f_{\mathrm{S}}(\cdot, \cdot))$ is more complex, as it has a recursive and compositional structure involving both the reward and value functions and cannot be directly modeled as a GP.
2. The TS algorithm does not sample directly from the posterior of this target function, but instead from the posterior of a proxy Q_h defined in the algorithm.

To address both challenges, we create upper and lower confidence bounds Q^u and Q^l for both Q_h and $Q_h^\star$.

Recall the following upper and lower confidence bounds from Definition 1. We define the upper confidence bounds recursively as:

$$Q_{h,k}^u(s, a) = \mu_{R,k}(s, a) + V_{h+1,k}^u(\mu_{S,k}(s, a)) + \xi_k(s, a), \;\; V_{h,k}^u(s) = \max_{a \in \mathcal{A}} Q_{h,k}^u(s, a),$$

initialized with $V_{H+1,k}^u(s) = 0$. Similarly, the lower confidence bounds are defined as:

$$Q_{h,k}^l(s, a) = \mu_{R,k}(s, a) + V_{h+1,k}^l(\mu_{S,k}(s, a)) - \xi_k(s, a), \;\; V_{h,k}^l(s) = \max_{a \in \mathcal{A}} Q_{h,k}^l(s, a).$$

initialized with $V_{H+1,k}^l(s) = 0$. The confidence width $\xi_k(s, a)$ is given by:

$$\xi_k(s, a) = \beta_k(\delta/Td)\sigma_{R,k}(s, a) + G\beta_k(\delta/Td)\|\sigma_{S,k}(s, a)\| + \frac{1}{2}H\beta_k(\delta/Td)^2\|\sigma_{S,k}(s, a)\|^2, \quad (8)$$

which is based on the confidence bound from Theorem 1.

Also recall Corollary 1. With probability at least $1 - \delta$, the optimal and proxy value functions are bounded by the high-probability confidence intervals:

$$Q_{h,k}^l(s, a) \le Q_{h,k}(s, a), \quad Q_h^\star(s, a) \le Q_{h,k}^u(s, a),$$

for all (h, k, s, a). Let us denote this event as $\mathcal{E}$.

Following the analysis technique used in the kernel bandit setting [9], we aim to show:

$$Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a_{h,k}) \leq c\, \xi_{k-1}(s_{h,k}, a_{h,k}), \tag{9}$$

for some universal constant $c > 0$.

To this end, we define the *saturated set* of actions at step (h, k) as:

$$\mathcal{S}_{h,k} := \left\{ a \in \mathcal{A} \,:\, Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a) > 2\, \xi_{k-1}(s_{h,k}, a) \right\}. \tag{10}$$

Intuitively, $\mathcal{S}_{h,k}$ includes actions that are significantly suboptimal under the true value function $Q_h^\star$, with a suboptimality gap that exceeds twice their confidence width.

We now prove a loose lower bound on the probability of selecting an action from unsaturated set. Specifically:

$$\begin{aligned}
\Pr\left[a_{h,k} \notin \mathcal{S}_{h,k}\right] &\geq \Pr\left[Q_{h,k}(s_{h,k}, a_{h,k}^\star) > Q_{h,k}(s_{h,k}, a) \quad \forall a \in \mathcal{S}_{h,k}\right] \\
&\geq \Pr\left[Q_{h,k}(s_{h,k}, a_{h,k}^\star) > Q_h^\star(s_{h,k}, a_{h,k}^\star) \wedge \right. \\
&\quad \left. Q_h^\star(s_{h,k}, a_{h,k}^\star) > Q_{h,k}(s_{h,k}, a),\ \forall a \in \mathcal{S}_{h,k}\right] \\
&\geq \Pr\left[Q_{h,k}(s_{h,k}, a_{h,k}^\star) > Q_h^\star(s_{h,k}, a_{h,k}^\star)\right] - \Pr\left[\bar{\mathcal{E}}\right] \\
&\geq \frac{1}{2} - \delta.
\end{aligned}$$

The first inequality holds because $a_{h,k}^\star$ is, by definition, the optimal action under $Q_h^\star$ and therefore saturated. The second step follows by decomposing the event into two sufficient conditions. To see the third line, observe that under the event $\mathcal{E}$, for any saturated action $a \in \mathcal{S}_{h,k}$, we have:

$$Q_{h,k}(s_{h,k}, a) \leq Q_h^\star(s_{h,k}, a) + 2\xi_{k-1}(s_{h,k}, a) \leq Q_h^\star(s_{h,k}, a_{h,k}^\star).$$

The fourth inequality follows from the symmetry and probability bound for $\mathcal{E}$ from Corollary 1.

Let $b_{h,k} = \arg\min_{a \in \mathcal{A} \setminus \mathcal{S}_{h,k}} \xi_{k-1}(s_{h,k}, a)$ denote the unsaturated action with the smallest confidence width. Then, using the law of total expectation:

$$\begin{aligned}
\mathbb{E}\left[\xi_{k-1}(s_{h,k}, a_{h,k})\right] &\geq \mathbb{E}\left[\xi_{k-1}(s_{h,k}, a_{h,k}) \mid a_{h,k} \notin \mathcal{S}_{h,k}\right] \Pr\left[a_{h,k} \notin \mathcal{S}_{h,k}\right] \\
&\geq \xi_{k-1}(s_{h,k}, b_{h,k})(\frac{1}{2} - \delta).
\end{aligned} \tag{11}$$

We now upper bound the immediate regret using $b_{h,k}$ as a reference:

$$\begin{aligned}
&Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, a_{h,k}) \\
&= \left(Q_h^\star(s_{h,k}, a_{h,k}^\star) - Q_h^\star(s_{h,k}, b_{h,k})\right) + \left(Q_h^\star(s_{h,k}, b_{h,k}) - Q_h^\star(s_{h,k}, a_{h,k})\right) \\
&\leq 2\, \xi_{k-1}(s_{h,k}, b_{h,k}) + \left(Q_{h,k}(s_{h,k}, b_{h,k}) + \xi_{k-1}(s_{h,k}, b_{h,k})\right) \\
&\quad - \left(Q_{h,k}(s_{h,k}, a_{h,k}) - \xi_{k-1}(s_{h,k}, a_{h,k})\right) \\
&\leq 3\, \xi_{k-1}(s_{h,k}, b_{h,k}) + \xi_{k-1}(s_{h,k}, a_{h,k}) \\
&\leq \left(\frac{3}{\frac{1}{2} - \delta} + 1\right) \xi_{k-1}(s_{h,k}, a_{h,k}),
\end{aligned}$$

where the first inequality adds and subtract the value at $b_{h,k}$, the first inquality uses definitions of the set $\mathcal{S}$, $b_{h,k}$ and event $\mathcal{E}$ and the final step uses the earlier bound on $\xi_{k-1}(s_{h,k}, b_{h,k})$.

This completes the bound on immediate regret in terms of the confidence width at the selected action.

Third step: Bounding the episode regret.

From the per-step regret decomposition (7) and the bound on the immediate regret (9) and using a telescoping sum over steps $h = 1$ to H, we obtain the following bound on the regret incurred in episode k:

$$V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \leq c \sum_{h=1}^{H} \xi_{k-1}(s_{h,k}, a_{h,k}).$$

Fourth step: Bounding the total regret.

Summing the episode regret over $k = 1$ to K episodes, we have:

$$\text{Regret}(T) \leq c \sum_{k=1}^{K} \sum_{h=1}^{H} \xi_{k-1}(s_{h,k}, a_{h,k}).$$

Replacing ξ_k, we get:

$$\begin{aligned}\text{Regret}(T) \leq \beta_K(\delta/Td) \sum_{k=1}^{K} \sum_{h=1}^{H} (\sigma_{R,k-1}(s_{h,k}, a_{h,k}) + u_G \left\|\sigma_{S,k-1}(s_{h,k}, a_{h,k})\right\|) \\ + \frac{u_H \beta_K^2(\delta/Td)}{2} \sum_{k=1}^{K} \sum_{h=1}^{H} \left\|\sigma_{S,k-1}(s_{h,k}, a_{h,k})\right\|^2 \end{aligned} \tag{12}$$

The sum of sequentially conditioned standard deviations of a sequence of observations from a GP often appears in the analysis of regret bounds in Bayesian optimization. A classical result shows that in the case of a single output GP, for a sequence of inputs $z_1, z_2, \ldots, z_T$, we have:

$$\sum_{i=1}^{n} \sigma_{i-1}^2(z_i) \leq C \log\det\left(\mathbf{I}_n + \frac{1}{\lambda^2}\mathbf{K}_n\right),$$

where λ is the GP noise parameter, $\mathbf{K}_n$ is the kernel matrix over the observed inputs and $C = 2/\log(1 + 1/\lambda^2)$ is a constant [8]. This result is sometimes referred to as the elliptical potential lemma, especially in the special case of linear kernels [20].

A direct application of this result to our setting faces two key challenges: *i)* We model transitions using a multi-output GP. Naively applying the bound to each output dimension separately results in a regret bound that scales suboptimally with d_S, the dimension of the state space. *ii)* Our regret decomposition involves a double sum over episodes k and steps h, but within each episode, the observations across h are not sequential updates. This structure leads to an additional scaling with the episode length H.

We address both challenges. First, we derive a new elliptical potential lemma tailored for multi-output GPs, which improves the dependence on d_S in the regret bound. That is given in Lemma 1. Second, to improve the H dependence, we leverage tools and techniques from the analysis of GPs under batch observations and delayed feedback [47] to tighten the bound with respect to H (that roughly speaking can be understood as delay in updating the GP model).

Lemma 2 (Elliptical potential for multi-output GPs with delayed updates)**.** *Let $f : \mathcal{Z} \to \mathbb{R}^d$ be a d-dimensional function modeled as a multi-output GP with a matrix-valued kernel k. Let $z_1, \ldots, z_T \in \mathcal{Z}$ be a sequence of input points and suppose the GP posterior is only updated every H steps, so that the standard deviation at time t is $\sigma_{H\lfloor(t-1)/H\rfloor}(z_t) \in \mathbb{R}^d$. Then,*

$$\sum_{t=1}^{T} \left\|\sigma_{H\lfloor(t-1)/H\rfloor}(z_t)\right\| \leq \sqrt{\frac{4\Gamma(T)}{\log(1+\lambda^{-2})}\left(T + \frac{4H^2\Gamma(T/H)}{\log(1+\lambda^{-2})}\right)}.$$

Applying Lemma 2 to the regret bound in terms of uncertainties (12), we obtain

$$\text{Regret}(T) = \mathcal{O}\left(\log(Td/\delta)\sqrt{T\Gamma(T)}\right). \tag{13}$$

C Auxiliary proofs

C.1 Proof of Corollary 1

We prove the lower bound by induction; the upper bound can be shown similarly.

As the base case, observe that $V_{H+1} = V_{H+1}^{\star} = V_{H+1}^{l} = 0$.

Now consider the inductive step. We compare the value of $Q^l_{h,k}(s,a)$ to $Q^\star_h(s,a)$:

$$\begin{aligned}
Q^l_{h,k}(s,a) - Q^\star_h(s,a) &= \mu_{R,k}(s,a) + V^l_{h+1,k}(\mu_{S,k}(s,a)) - \xi_k(s,a) - \big(f_{\mathrm{R}}(s,a) + V^\star_{h+1}(f_{\mathrm{S}}(s,a))\big) \\
&= \underbrace{\mu_{R,k}(s,a) - f_{\mathrm{R}}(s,a) + V^\star_{h+1}(\mu_{S,k}(s,a)) - V^\star_{h+1}(f_{\mathrm{S}}(s,a)) - \xi_k(s,a)}_{\text{Term 1}} \\
&\quad + \underbrace{V^l_{h+1,k}(\mu_{S,k}(s,a)) - V^\star_{h+1}(\mu_{S,k}(s,a))}_{\text{Term 2}} \\
&\leq 0.
\end{aligned}$$

The first line follows from the definitions of $Q^l_{h,k}$ and $Q^\star_h$. The second line is obtained by adding and subtracting $V^\star_{h+1}(\mu_{S,k}(s,a))$, followed by regrouping terms. The inequality holds since Term 1 is non-positive due to the confidence bounds in Theorem 1, and Term 2 is non-positive by the induction hypothesis.

Next, we extend the argument to the value function:

$$\begin{aligned}
V^l_h(s) - V^\star_h(s) &= \max_{a\in\mathcal{A}} Q^l_h(s,a) - \max_{a\in\mathcal{A}} Q^\star_h(s,a) \\
&\leq \max_{a\in\mathcal{A}} \left[Q^l_h(s,a) - Q^\star_h(s,a)\right] \\
&\leq 0,
\end{aligned}$$

which completes the inductive proof that $Q^l_{h,k}(s,a) \leq Q^\star_h(s,a)$ and $V^l_h(s) \leq V^\star_h(s)$ for all s,a,h.

We know that $Q^l_{h,k}(s,a) \leq Q_{h,k}(s,a)$ for all (s,a,h,k). For the inductive step, consider,

$$\begin{aligned}
&Q^l_{h,k}(s,a) - Q_{h,k}(s,a) \\
&= \mu_{R,k}(s,a) + V^l_{h+1,k}(\mu_{S,k}(s,a)) - \xi_k(s,a) - (f_{R,k}(s,a) + V_{h+1,k}(f_{S,k}(s,a))) \\
&= \underbrace{\mu_{R,k}(s,a) - f_{R,k}(s,a) + V_{h+1,k}(\mu_{S,k}(s,a)) - V_{h+1,k}(f_{S,k}(s,a)) - \xi_k(s,a)}_{\text{Term 1}} \\
&\quad + \underbrace{V^l_{h+1,k}(\mu_{S,k}(s,a)) - V_{h+1,k}(\mu_{S,k}(s,a))}_{\text{Term 2}} \\
&\leq 0.
\end{aligned}$$

The decomposition follows by adding and subtracting $V_{h+1,k}(\mu_{S,k}(s,a))$ and regrouping terms. Term 1 is non-positive due to the high-probability confidence bound (Theorem 1). Term 2 is non-positive by the induction hypothesis.

Hence, we conclude $Q^l_{h,k}(s,a) \leq Q_{h,k}(s,a)$ for all (s,a).

Extending to the value function:

$$\begin{aligned}
V^l_{h,k}(s) - V_{h,k}(s) &= \max_{a\in\mathcal{A}} Q^l_{h,k}(s,a) - \max_{a\in\mathcal{A}} Q_{h,k}(s,a) \\
&\leq \max_{a\in\mathcal{A}} \big(Q^l_{h,k}(s,a) - Q_{h,k}(s,a)\big) \\
&\leq 0.
\end{aligned}$$

This completes the inductive proof that $Q^l_{h,k}(s,a) \leq Q_{h,k}(s,a)$ and $V^l_{h,k}(s) \leq V_{h,k}(s)$ for all s,a,h. The upper bounds, i.e., $Q^\star_h(s,a), Q_{h,k}(s,a) \leq Q^u_{h,k}(s,a)$ for all s,a,h, are proven analogously using similar argument.

C.2 Proof of Lemma 1

We consider a d-dimensional GP $f : \mathcal{Z} \to \mathbb{R}^d$ and a sequence of inputs $z_1, \ldots, z_T$. Define the full observation vector as:

$$y_{1:T} = \begin{bmatrix} f(z_1) + \varepsilon_1 \\ \vdots \\ f(z_T) + \varepsilon_T \end{bmatrix} \in \mathbb{R}^{Td}, \quad \varepsilon_t \sim \mathcal{N}(0, \lambda^2 I_d).$$

The mutual information between $y_{1:T}$ and the latent function values is:

$$\mathcal{I}_T = I(y_{1:T}; f(z_1), \ldots, f(z_T)) = \frac{1}{2} \log \det \left(\mathbf{I}_{Td} + \lambda^{-2} \mathbf{K}_T \right),$$

where $\mathbf{K}_T \in \mathbb{R}^{Td \times Td}$ is the prior kernel matrix over all outputs.

Let $\sigma_{t-1}(z_t) \in \mathbb{R}^d$ be the vector of posterior standard deviations at z_t given observations up to time $t-1$. Define the total uncertainty:

$$S_T := \sum_{t=1}^{T} \|\sigma_{t-1}(z_t)\|^2 = \sum_{t=1}^{T} \sum_{j=1}^{d} \sigma_{t-1,j}^2(z_t).$$

We apply the scalar inequality (used in [8]):

$$\sigma^2 \leq \frac{1}{\log(1+\lambda^{-2})} \log \left(1 + \frac{\sigma^2}{\lambda^2} \right), \quad \text{for all } \sigma^2 \in [0, 1].$$

Applying this to each term $\sigma_{t-1,j}^2(z_t)$ and summing, we obtain:

$$S_T \leq \frac{1}{\log(1+\lambda^{-2})} \sum_{t=1}^{T} \sum_{j=1}^{d} \log \left(1 + \frac{\sigma_{t-1,j}^2(z_t)}{\lambda^2} \right).$$

Since $y_{1:T}$ is jointly Gaussian, the total sum of these log-terms is bounded by $2\mathcal{I}_T$, giving:

$$S_T \leq \frac{2}{\log(1+\lambda^{-2})} \mathcal{I}_T.$$

Thus, we conclude:

$$\sum_{t=1}^{T} \|\sigma_{t-1}(z_t)\|^2 \leq C\mathcal{I}_T, \quad \text{with } C = \frac{2}{\log(1+\lambda^{-2})}.$$

C.3 Proof of Lemma 2

We begin by naively applying the elliptical potential lemma to the same step index across episodes

$$\begin{aligned} \sum_{t=1}^{T} \|\sigma_{H\lfloor (t-1)/H \rfloor}(z_t)\| &\leq \sum_{h=1}^{H} \sum_{j=1}^{K} \|\sigma_{H(j-1)+h}(z_{Hj+h})\|^2 \\ &\leq CH\Gamma(K) \end{aligned} \tag{14}$$

To improve on this, we use the following inequality proven in Lemma 3, for any z and $t' < t$,

$$\|\sigma_{t'}(z)\|^2 \leq \|\sigma_t(z)\|^2 \left(1 + \sum_{j=t'+1}^{t} \|\sigma_{t'}(z_j)\|^2 \right).$$

Applying this to the case where the model is updated every H steps, for each $t \in [T]$, we let $t' = H\lfloor (t-1)/H \rfloor$. Then,

$$\|\sigma_{t'}(z_t)\| \leq \|\sigma_t(z_t)\| \sqrt{1 + \sum_{j=t'+1}^{t} \|\sigma_{t'}(z_j)\|^2}.$$

Now, summing over all $t = 1, \ldots, T$, we apply the Cauchy–Schwarz inequality:

$$\begin{aligned} \sum_{t=1}^{T} \|\sigma_{H\lfloor (t-1)/H \rfloor}(z_t)\| &\leq \sum_{t=1}^{T} \|\sigma_t(z_t)\| \sqrt{1 + \sum_{j=H\lfloor (t-1)/H \rfloor + 1}^{t} \|\sigma_{H\lfloor (t-1)/H \rfloor}(z_j)\|^2} \\ &\leq \left(\sum_{t=1}^{T} \|\sigma_t(z_t)\|^2 \right)^{1/2} \left(T + H \sum_{t=1}^{T} \|\sigma_{H\lfloor (t-1)/H \rfloor}(z_t)\|^2 \right)^{1/2}. \end{aligned}$$

We now substitute the same term with the looser bound given earlier in (14),

$$\sum_{t=1}^{T} \|\sigma_{H\lfloor(t-1)/H\rfloor}(z_t)\| \leq \sqrt{\frac{2\Gamma(T)}{\log(1+\lambda^{-2})}\left(T + \frac{2H^2\Gamma(K)}{\log(1+\lambda^{-2})}\right)}.$$

This completes the proof.

Lemma 3 (Variance ratio inequality for multi-output GPs)**.** *Let $f : \mathcal{Z} \to \mathbb{R}^d$ be a d-dimensional function modeled as a multi-output GP with a matrix-valued kernel k. Let $\sigma_t(z) \in \mathbb{R}^d$ denote the vector of posterior marginal standard deviations at point z given t observations. Then, for any $z \in \mathcal{Z}$ and $t' < t$,*

$$1 \leq \frac{\|\sigma_{t'}(z)\|^2}{\|\sigma_t(z)\|^2} \leq 1 + \sum_{j=t'+1}^{t} \|\sigma_{t'}(z_j)\|^2.$$

Proof. For each output dimension $j \in [d]$, the scalar variance update satisfies (see, e.g., Lemma 4 of [47] and Proposition A.1 of [48])

$$\frac{\sigma^2_{t',j}(z)}{\sigma^2_{t,j}(z)} \leq 1 + \sum_{i=t'+1}^{t} \sigma^2_{t',j}(z_i).$$

Now summing over $j = 1, \ldots, d$ we get:

$$\frac{\|\sigma_{t'}(z)\|^2}{\|\sigma_t(z)\|^2} = \frac{\sum_{j=1}^{d} \sigma^2_{t',j}(z)}{\sum_{j=1}^{d} \sigma^2_{t,j}(z)} \leq 1 + \sum_{i=1}^{d} \sum_{j=t'+1}^{t} \sigma^2_{t',i}(z_j) = 1 + \sum_{j=t'+1}^{t} \|\sigma_{t'}(z_j)\|^2.$$

Therefore,

$$\frac{\|\sigma_{t'}(z)\|^2}{\|\sigma_t(z)\|^2} \leq 1 + \sum_{j=t'+1}^{t} \|\sigma_{t'}(z_j)\|^2.$$

The lower bound $1 \leq \|\sigma_{t'}(z)\|^2/\|\sigma_t(z)\|^2$ holds since variance decreases monotonically as more data is observed. This completes the proof. □

D Discussion on the linear model of coregionalization

A widely used and computationally convenient special case of multi-output GPs is the *linear model of coregionalization* (LMC) [49, 50]. In this model, the vector-valued function $f : \mathcal{Z} \to \mathbb{R}^d$ is expressed as a linear combination of L independent latent Gaussian processes:

$$f_j(z) = \sum_{\ell=1}^{L} \alpha_{j\ell} g_\ell(z), \quad g_\ell \sim \mathrm{GP}(0, k^{(g)}), \tag{15}$$

where $k^{(g)} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a shared scalar kernel, and $\alpha \in \mathbb{R}^{d \times L}$ is a matrix of output mixing weights. This induces a matrix-valued kernel:

$$k(z, z') = \mathbf{A}\, k^{(g)}(z, z'), \quad \text{where } \mathbf{A} = \alpha\alpha^\top \in \mathbb{R}^{d \times d}. \tag{16}$$

Under this kernel structure, the block kernel matrix over the training data admits a Kronecker product decomposition:

$$\mathbf{K}_n = \mathbf{A} \otimes \mathbf{K}_n^{(g)}, \tag{17}$$

where $\mathbf{K}_n^{(g)} \in \mathbb{R}^{n \times n}$ is the input kernel matrix with $[\mathbf{K}_n^{(g)}]_{ij} = k^{(g)}(z_i, z_j)$. The cross-covariance matrix between test point z and training data becomes:

$$\mathbf{k}_n(z) = \mathbf{A} \otimes \mathbf{k}_n^{(g)}(z) \in \mathbb{R}^{nd \times d}, \tag{18}$$

with $[\mathbf{k}_n^{(g)}(z)]_i = k^{(g)}(z_i, z)$.

This formulation is particularly useful when jointly modeling structured outputs such as reward and transition functions in reinforcement learning, as it captures both intra- and inter-output correlations while enabling scalable inference. We provide a brief discussion on the regret bounds with such a structured kernel.

D.1 Information gain and regret bounds for LMC

We analyze how the structure of the LMC affects the information gain term $\Gamma(T)$ appearing in the regret bound. Recall that $\Gamma(T)$ upper bounds the quantity $\frac{1}{2}\log\det(\mathbf{I}_{Td} + \frac{1}{\lambda^2}\mathbf{K}_T)$, where $\mathbf{K}_T$ is the kernel matrix of the multi-output GP. Under the LMC structure, $\mathbf{K}_T = \mathbf{A} \otimes \mathbf{K}_T^{(g)}$, where $\mathbf{A}$ captures output correlations and $\mathbf{K}_T^{(g)}$ is the kernel matrix corresponding to a shared latent GP. Using properties of Kronecker products and letting λ_i denote the eigenvalues of $\mathbf{A}$, we obtain:

$$\begin{aligned}\log\det\left(\mathbf{I}_{Td} + \frac{1}{\lambda^2}\mathbf{K}_T\right) &= \log\det\left(\mathbf{I}_{Td} + \frac{1}{\lambda^2}\mathbf{A} \otimes \mathbf{K}_T^{(g)}\right) \\ &= \sum_{i=1}^{d} \log\det\left(\mathbf{I}_T + \frac{\lambda_i}{\lambda^2}\mathbf{K}_T^{(g)}\right).\end{aligned}$$

For the Matérn family of kernels, the information gain of the latent scalar GP is known to satisfy [46, 51]:

$$\Gamma^{(g)}(T) = \tilde{\mathcal{O}}\left(\left(\frac{T}{\lambda^2}\right)^{\alpha}\right),$$

where $\alpha = \frac{d}{2\nu+d} < 1$ depends on the input dimension d and the kernel smoothness parameter ν.

Substituting into the sum, we obtain:

$$\begin{aligned}\Gamma(T) &= \sum_{i=1}^{d} \tilde{\mathcal{O}}\left(\left(\frac{T\lambda_i}{\lambda^2}\right)^{\alpha}\right) \\ &= \tilde{\mathcal{O}}\left(\left(\sum_{i=1}^{d}\lambda_i^{\alpha}\right)\left(\frac{T}{\lambda^2}\right)^{\alpha}\right). \end{aligned} \tag{19}$$

In the general case without structure, a standard upper bound is given by:

$$\Gamma(T) = \tilde{\mathcal{O}}\left(\left(\frac{Td}{\lambda^2}\right)^{\alpha}\right).$$

Comparing the two, we observe that when $\mathbf{A}$ has low-rank behavior, specifically, when $\sum_{i=1}^{d}\lambda_i^{\alpha} \leq d^{\alpha}$, the LMC-based bound in (19) can be tighter. In particular, the regret bound becomes:

$$\text{Regret} = \tilde{\mathcal{O}}\left(\left(\sum_{i=1}^{d}\lambda_i^{\frac{d}{2\nu+d}}\right)T^{\frac{d}{2\nu+d}}\right). \tag{20}$$

This shows the effect of shared latent structure and output correlations on the regret bounds.

E Additional experiments

In this section, we provide further information about our experiments and additional results.

Model training. We model the reward and transition functions using a multi-output sparse variational Gaussian process, trained by maximizing the evidence lower bound (ELBO) with the Adam optimizer [52]. A shared base kernel (either RBF or Matérn) is used across outputs, and the outputs are linearly mixed according to a matrix following a linear model of coregionalization (LMC). The model uses 100 inducing points per output dimension, a zero mean function, and is optimized for 20 steps per iteration using GPyTorch [53]. Full code is included in the supplementary material for reproducibility.

Kernel complexity. We study the effect of kernel complexity on regret using synthetic MDPs generated from 200 different GP-sampled environments. For all GPs, we fix a random linear mixing matrix (see Appendix D),

$$\mathbf{A} = \begin{pmatrix} 0.9926 & 0.2082 & 0.4968 \\ -0.3196 & 0.8869 & 0.1603 \\ 0.1557 & -1.4231 & -1.3905 \end{pmatrix}.$$

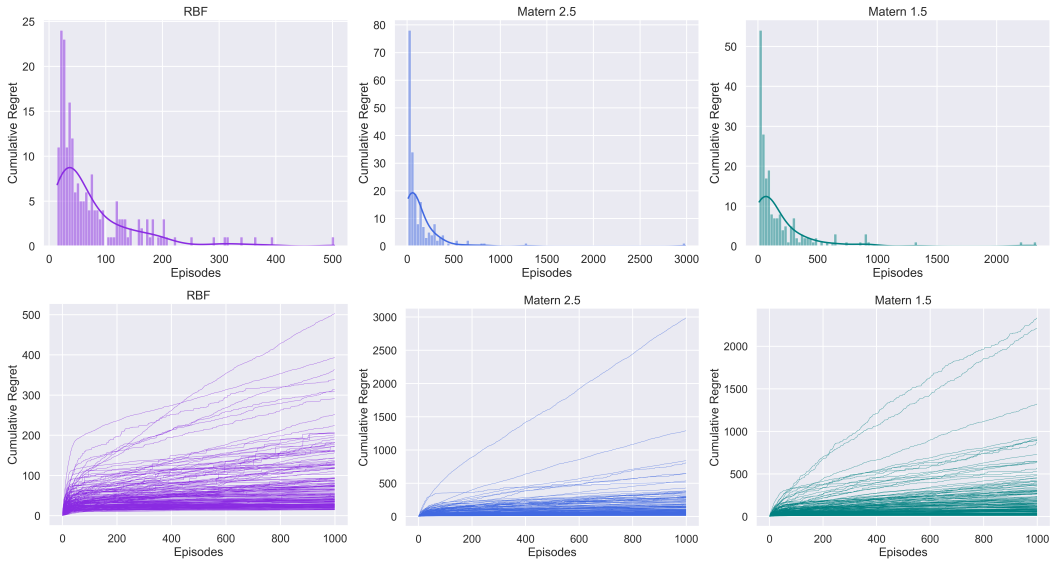

Figure 2: The first row shows the histogram of cumulative regrets over all trials on the last episode of RL-GPS training with each kernel. The second row shows the regret growth curves from all trials for each kernel.

In each trial, we sample new ground-truth reward and transition functions from the multi-output GP to define an MDP. The optimal value function is then computed for this MDP via finite-horizon value iteration with horizon $H = 20$. We train RL-GPS (Algorithm 1) for $K = 1000$ episodes in each environment and record the cumulative regret. Since the mixing matrix is fixed, randomness within a trial arises only from the starting state distribution during rollouts. Across trials, randomness stems from variation in the sampled environments, which can differ significantly in difficulty. As a result, we observe a substantial right-skew in cumulative regret, with a few environments producing particularly challenging instances. To demonstrate this variability, we show the cumulative regret at the final episode across all trials for each kernel as well as the individual regret curves for each trial in Figure 2. All trials for each experiment run within 24 hours on one NVIDIA GeForce RTX 2080 Ti GPU.

Multi-output kernel structure. In these experiments, we further study how the multi-output kernel structure affects regret in sparse navigation tasks. The state space $\mathcal{S} = [0, 1]^2$ is discretized into 25 equally spaced bins per dimension and the action space $\mathcal{A}$ consists of 9 discrete actions that move the agent in the cardinal or diagonal directions or allow it to remain stationary. The agent receives a reward of $+1$ when within 0.1 of the destination and a penalty of -0.01 otherwise. We study two settings: free movement within the grid and constrained navigation through a maze. As before, the optimal value function $V^\star$ is computed using finite-horizon value iteration. Algorithm 1 runs for $K = 1000$ episodes with a sparse LMC multi-output GP posterior. The results over 200 trials are shown in Figure 3 and demonstrate sublinear regret growth, consistent with our theory. Notably, in these sparse, less smooth environments, the RBF kernel accumulates regret more rapidly than the Matérn kernels. This is due to the RBF kernel's strong smoothness assumptions, which lead to model misspecification and slower adaptation when the ground truth is rougher [45]. The results also highlight the value of modeling output correlations using the LMC. Specifically, the Matérn 1.5 kernel without LMC incurs substantially higher regret in the maze environment compared to its LMC-enabled counterpart. This indicates that explicitly capturing output dependencies can improve sample efficiency and reduce regret, especially in structured or high-dimensional tasks.

For this set of experiments, we train the multi-output GP using a Matérn kernel with $\nu = 1.5$, comparing two modeling approaches: independent GPs and the linear model of coregionalization (LMC). When using LMC, the mixing weights are learned during training, so the randomness across trials arises only from the random initialization of the mixing weights and the starting states for rollouts. In contrast to the earlier experiments, the environment itself is fixed across all trials. For the unconstrained navigation experiments, the horizon used is $H = 20$ and the GP is updated for 20

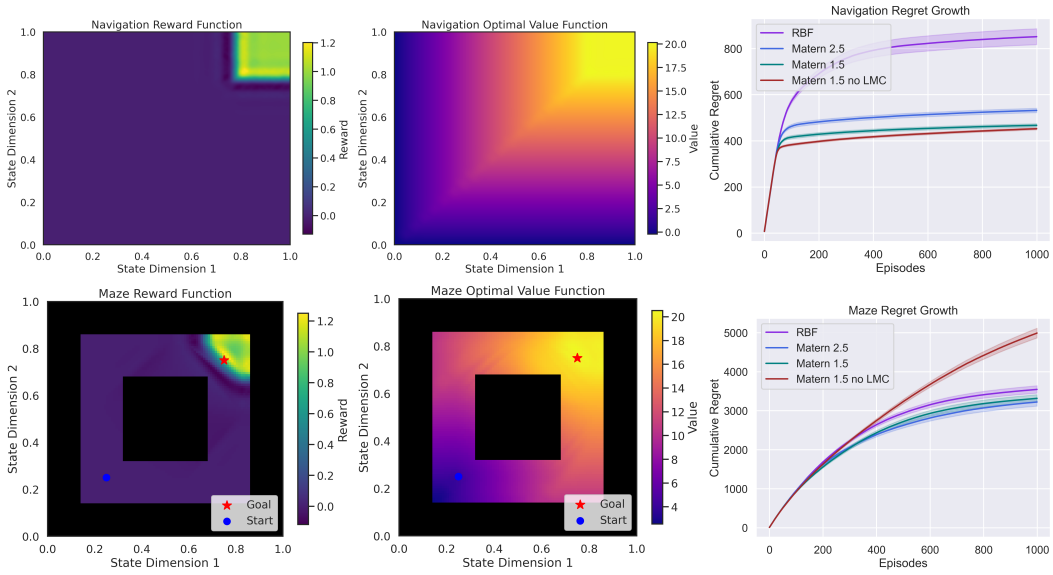

Figure 3: The reward function (left column), optimal value function (middle column), and cumulative regret of TS with a multi-output GP with confidence regions representing the standard error over 200 random trials (right column) are given. The first row corresponds to the sparse navigation task and the second row corresponds to the sparse maze problem.

optimization steps at each iteration. For the maze experiments, the horizon used is also $H = 20$ and, due to the increased modeling difficulty, the GP is updated for 50 optimization steps at each iteration. All trials for each unconstrained navigation experiment run within 24 hours on one NVIDIA GeForce RTX 2080 Ti GPU. Due to the greater number of GP updates in the maze setting, trials take longer to run and all 200 trials complete within 80 hours on the same hardware. Note that all experiments are easily parallelizable.