Appendix

A PRELIMINARIES

A.1 Relevant Results on Gaussian Process Multi-armed Bandits

We first review some relevant definitions and results from the Gaussian process multi-armed bandits literature, which will be useful in the analysis of our algorithms. We first begin with the definition of *Maximum Information Gain*, first appeared in Srinivas et al. [2009], which basically measures the reduction in uncertainty about the unknown function after some noisy observations (rewards).

For a function $f : \mathcal{X} \to \mathbb{R}$ and any subset $A \subset \mathcal{X}$ of its domain, we use $f_A := [f(x)]_{x \in A}$ to denote its restriction to A, i.e., a vector containing f's evaluations at each point in A (under an implicitly understood bijection from coordinates of the vector to points in A). In case f is a random function, f_A will be understood to be a random vector. For jointly distributed random variables X, Y, $I(X;Y)$ denotes the Shannon mutual information between them.

Definition 1 (Maximum Information Gain (MIG)) *Let $f : \mathcal{X} \to \mathbb{R}$ be a (possibly random) real-valued function defined on a domain $\mathcal{X}$, and t a positive integer. For each subset $A \subset \mathcal{X}$, let Y_A denote a noisy version of f_A obtained by passing f_A through a channel $\mathbb{P}[Y_A|f_A]$. The Maximum Information Gain (MIG) about f after t noisy observations is defined as*

$$\gamma_t(f, \mathcal{X}) := \max_{A \subset \mathcal{X}: |A| = t} I(f_A; Y_A).$$

(We omit mentioning explicitly the dependence on the channels for ease of notation.)

MIG will serve as a key instrument to obtain our regret bounds by virtue of Lemma 1.

For a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and points $x, x_1, \ldots, x_s \in \mathcal{X}$, we define the vector $k_s(x) := [k(x_1, x), \ldots, k(x_s, x)]^T$ of kernel evaluations between x and $x_1, \ldots, x_s$, and $K_{\{x_1,\ldots,x_s\}} \equiv K_s := [k(x_i, x_j)]_{1 \le i,j \le s}$ be the kernel matrix induced by the x_is. Also for each $x \in \mathcal{X}$ and $\lambda > 0$, let $\sigma_s^2(x) := k(x, x) - k_s(x)^T (K_s + \lambda I)^{-1} k_s(x)$.

Lemma 1 (Information Gain and Predictive Variances under GP prior and additive Gaussian noise) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive semi-definite kernel and $f \sim GP_{\mathcal{X}}(0, k)$ a sample from the associated Gaussian process over $\mathcal{X}$. For each subset $A \subset \mathcal{X}$, let Y_A denote a noisy version of f_A obtained by passing f_A through a channel that adds iid $\mathcal{N}(0, \lambda)$ noise to each element of f_A. Then,*

$$\gamma_t(f, \mathcal{X}) = \max_{A \subset \mathcal{X}: |A| = t} \frac{1}{2} \ln \left| I + \lambda^{-1} K_A \right|, \tag{13}$$

and

$$\gamma_t(f, \mathcal{X}) = \max_{\{x_1,\ldots,x_t\} \subset \mathcal{X}} \frac{1}{2} \sum_{s=1}^{t} \ln \left(1 + \lambda^{-1} \sigma_{s-1}^2(x_s)\right). \tag{14}$$

Proof The proofs follow from Srinivas et al. [2009]. ■

Remark. Note that the right hand sides of (13) and (14) depend only on the kernel function k, domain $\mathcal{X}$, constant λ and number of observations t. Further, as shown in Theorem 8 of Srinivas et al. [2009], the dependency on λ is only of $\tilde{O}(1/\lambda)$. Hence to indicate these dependencies on k and $\mathcal{X}$ more explicitly, we denote the Maximum Information Gain $\gamma_t(f, \mathcal{X})$ in the setting of Lemma 1 as $\gamma_t(k, \mathcal{X})$.

Lemma 2 (Sum of Predictive variances is bounded by MIG) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive semi-definite kernel such that it has bounded variance, i.e. $k(x, x) \le 1$ for all $x \in \mathcal{X}$ and $f \sim GP_{\mathcal{X}}(0, k)$ be a sample from the associated Gaussian process over $\mathcal{X}$, then for all $s \ge 1$ and $x \in \mathcal{X}$,*

$$\sigma_{s-1}^2(x) \le (1 + \lambda^{-1}) \sigma_s^2(x), \tag{15}$$

and

$$\sum_{s=1}^{t} \sigma_{s-1}^2(x_s) \le (2\lambda + 1) \gamma_t(k, \mathcal{X}). \tag{16}$$

Proof From our assumption $k(x,x) \leq 1$, we have $0 \leq \sigma^2_{s-1}(x) \leq 1$ for all $x \in \mathcal{X}$, and hence $\sigma^2_{s-1}(x_s) \leq \ln\big(1+\lambda^{-1}\sigma^2_{s-1}(x_s)\big)/\ln(1+\lambda^{-1})$ since $\alpha/\ln(1+\alpha)$ is non-decreasing for any $\alpha \in [0,\infty)$. Therefore

$$\sum_{s=1}^{t} \sigma^2_{s-1}(x_s) \leq 2/\ln(1+\lambda^{-1}) \sum_{s=1}^{t} \frac{1}{2}\ln\big(1+\lambda^{-1}\sigma^2_{s-1}(x_s)\big) \leq 2\gamma_t(k,\mathcal{X})/\ln(1+\lambda^{-1}),$$

where the last inequality follows from (14). Now see that $2/\ln(1+\lambda^{-1}) \leq (2+\lambda^{-1})/\lambda^{-1} = 2\lambda+1$, since $\ln(1+\alpha) \geq 2\alpha/(2+\alpha)$ for any $\alpha \in [0,\infty)$. Hence $\sum_{s=1}^{t}\sigma^2_{s-1}(x_s) \leq (2\lambda+1)\gamma_t(k,\mathcal{X})$.

Further from Appendix F in Chowdhury and Gopalan [2017], see that $\sigma^2_s(x) = \sigma^2_{s-1}(x) - k^2_{s-1}(x_s,x)/\big(\lambda+\sigma^2_{s-1}(x_s)\big)$ for all $x \in \mathcal{X}$, where $k_s(x,x') := k(x,x') - k_s(x)^T(K_s+\lambda I)^{-1}k_s(x')$. Since $k_{s-1}(x,\cdot), x \in \mathcal{X}$ lie in the reproducing kernel Hilbert space (RKHS) of k_{s-1}, the reproducing property implies that $k_{s-1}(x_s,x) = \langle k_{s-1}(x_s,\cdot), k_{s-1}(x,\cdot)\rangle_{k_{s-1}}$. Hence by Cauchy-Schwartz inequality $k^2_{s-1}(x_s,x) \leq \|k_{s-1}(x_s,\cdot)\|^2_{k_{s-1}} \|k_{s-1}(x,\cdot)\|^2_{k_{s-1}} = k_{s-1}(x_s,x_s)k_{s-1}(x,x) = \sigma^2_{s-1}(x_s)\sigma^2_{s-1}(x)$, where the second last step follows from the reproducing property and the last step is due to $\sigma^2_s(x) = k_s(x.x)$. Therefore $\sigma^2_s(x) \geq \sigma^2_{s-1}(x)\Big(1 - \frac{\sigma^2_{s-1}(x_s)}{\lambda+\sigma^2_{s-1}(x_s)}\Big) = \lambda\sigma^2_{s-1}(x)/\big(\lambda+\sigma^2_{s-1}(x_s)\big)$. Further by the bounded variance assumption, $\sigma^2_{s-1}(x_s) \leq 1$ and hence $\lambda/\big(\lambda+\sigma^2_{s-1}(x_s)\big) \geq \lambda/(1+\lambda)$. This implies $\sigma^2_s(x)/\sigma^2_{s-1}(x) \geq \lambda/(1+\lambda)$ and hence $\sigma^2_{s-1}(x) \leq (1+\lambda^{-1})\sigma^2_s(x)$. ■

Lemma 3 (Ratio of predictive variances is bounded by Information Gain Kandasamy et al. [2018]) *Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a symmetric, positive-semidefinite kernel and $f \sim GP_{\mathcal{X}}(0,k)$. Further, let A and B be finite subsets of $\mathcal{X}$, and for a positive constant λ, let σ_A and $\sigma_{A\cup B}$ be the posterior standard deviations conditioned on queries A and $A\cup B$ respectively (similarly defined as in Lemma 1). Also, let $\gamma_(k,\mathcal{X})$ denote the maximum information gain after t noisy observations. Then the following holds for all $x \in \mathcal{X}$:*

$$\max_{A,B\subset\mathcal{X}:|B|=t} \frac{\sigma_A(x)}{\sigma_{A\cup B}(x)} \leq \exp\big(\gamma_t(k,\mathcal{X})\big). \tag{17}$$

Proof The proof can be figured out from Desautels et al. [2014], but we include it here for completeness. Let Y_A and Y_B are vectors of noisy observations when we query f at A and B respectively, and $I\left(f(x);Y_B \;\middle|\; Y_A\right)$ denotes the *mutual information* between $f(x)$ and Y_B, conditioned on Y_A. Note that

$$I\big(f(x);Y_B \;\big|\; Y_A\big) = H\big(f(x) \;\big|\; Y_A\big) - H\big(f(x) \;\big|\; Y_{A\cup B}\big) = \frac{1}{2}\ln\big(2\pi e\sigma^2_A(x)\big) - \frac{1}{2}\ln\big(2\pi e\sigma^2_{A\cup B}(x)\big) = \ln\left(\frac{\sigma_A(x)}{\sigma_{A\cup B}(x)}\right).$$

Hence for all $x \in \mathcal{X}$ and for all finite subsets A, B of $\mathcal{X}$, we have

$$\frac{\sigma_A(x)}{\sigma_{A\cup B}(x)} = \exp\Big(I\big(f(x);Y_B \;\big|\; Y_A\big)\Big). \tag{18}$$

Now, by monotonicity of *mutual information*, we have $I\left(f(x);Y_B \;\middle|\; Y_A\right) \leq I\left(f;Y_B \;\middle|\; Y_A\right)$ for all $x \in \mathcal{X}$. Further, if $|B| = t$, we have $I\left(f;Y_B \;\middle|\; Y_A\right) \leq \max_{B\subset\mathcal{X}:|B|=t} I\left(f;Y_B \;\middle|\; Y_A\right)$. Thus for all $x \in \mathcal{X}$ and for all finite subset B of $\mathcal{X}$ for which $|B| = t$, we have

$$I\left(f(x);Y_B \;\middle|\; Y_A\right) \leq \max_{B\subset\mathcal{X}:|B|=t} I\left(f;Y_B \;\middle|\; Y_A\right).$$

Now by submodularity of *conditional mutual information*, for all finite subset A of $\mathcal{X}$, we have

$$\max_{B\subset\mathcal{X}:|B|=t} I\left(f;Y_B \;\middle|\; Y_A\right) \leq \max_{B\subset\mathcal{X}:|B|=t} I\left(f;Y_B\right).$$

Further see that $I\left(f;Y_B\right) = I\left(f_B;Y_B\right)$, since $H(Y_B \;|\; f) = H(Y_B \;|\; f_B)$. This implies, for all $x \in \mathcal{X}$ and for all finite subsets A, B of $\mathcal{X}$ for which $|B| = t$, that

$$I\left(f(x);Y_B \;\middle|\; Y_A\right) \leq \max_{B\subset\mathcal{X}:|B|=t} I\left(f_B;Y_B\right) = \gamma_t(k,\mathcal{X}). \tag{19}$$

Now the result follows by combining (18) and (19). ■

Bound on Maximum Information Gain Srinivas et al. [2009] proved upper bounds over $\gamma_t(k, \mathcal{X})$ for three commonly used kernels, namely *Linear*, *Squared Exponential* and *Matérn*, defined respectively as

$$\begin{aligned} k_{Linear}(x, x') &= x^T x', \\ k_{SE}(x, x') &= \exp\left(-s^2/2l^2\right), \\ k_{Matérn}(x, x') &= \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{s\sqrt{2\nu}}{l}\right)^{\nu} B_\nu \left(\frac{s\sqrt{2\nu}}{l}\right), \end{aligned}$$

where $l > 0$ and $\nu > 0$ are hyper-parameters of the kernels, $s = \|x - x'\|_2$ encodes the similarity between two points $x, x' \in \mathcal{X}$ and B_ν denotes the *modified Bessel function*. The bounds are given in Lemma 4.

Lemma 4 (MIG for common kernels) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive semi-definite kernel and $f \sim GP_{\mathcal{X}}(0, k)$. Let $\mathcal{X}$ be a compact and convex subset of $\mathbb{R}^d$ and the kernel k satisfies $k(x, x') \leq 1$ for all $x, x' \in \mathcal{X}$. Then for*

- *Linear kernel:* $\gamma_t(k_{Linear}, \mathcal{X}) = \tilde{O}(d \ln t)$.
- *Squared Exponential kernel:* $\gamma_t(k_{SE}, \mathcal{X}) = \tilde{O}\left((\ln t)^d\right)$.
- *Matérn kernel:* $\gamma_t(k_{Matérn}, \mathcal{X}) = \tilde{O}\left(t^{d(d+1)/(2\nu + d(d+1))} \ln t\right)$.

Note that, the Maximum Information Gain $\gamma_t(k, \mathcal{X})$ depends only *poly-logarithmically* on the number of observations t for all these kernels.

Reproducing kernel Hilbert spaces (RKHS) A Reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k(\mathcal{X})$ is a complete subspace of the space of square integrable functions $L_2(\mathcal{X})$ defined over the domain $\mathcal{X}$. It includes functions of the form $f(x) = \sum_i \alpha_i k(x, x_i)$ with $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$, where k is a symmetric, positive- definite kernel function. The RKHS has an inner product $\langle \cdot, \cdot \rangle_k$, which obeys the reproducing property: $f(x) = \langle f, k(x, \cdot) \rangle_k$ for all $f \in \mathcal{H}_k(\mathcal{X})$, and the induced RKHS norm $\|f\|_k^2 = \langle f, f \rangle_k$ measures smoothness of f with respect to the kernel k. Lemma 5 gives a concentration bound for a member f of $\mathcal{H}_k(\mathcal{X})$. A (slightly) modified version of Lemma 5 has appeared independently in Chowdhury and Gopalan [2017] and Durand et al. [2017].

Lemma 5 (Concentration of an RKHS member) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, positive-semidefinite kernel and $f : \mathcal{X} \to \mathbb{R}$ be a member of the RKHS $\mathcal{H}_k(\mathcal{X})$ of real-valued functions on $\mathcal{X}$ with kernel k. Let $\{x_t\}_{t \geq 1}$ and $\{\varepsilon_t\}_{t \geq 1}$ be stochastic processes such that $\{x_t\}_{t \geq 1}$ form a predictable process, i.e., $x_t \in \sigma(\{x_s, \varepsilon_s\}_{s=1}^{t-1})$ for each t, and $\{\varepsilon_t\}_{t \geq 1}$ is conditionally R-sub-Gaussian for a positive constant R, i.e.,*

$$\forall t \geq 0, \;\; \forall \lambda \in \mathbb{R}, \;\; \mathbb{E}\left[e^{\lambda \varepsilon_t} \;\middle|\; \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right),$$

where $\mathcal{F}_{t-1}$ is the σ-algebra generated by $\{x_s, \varepsilon_s\}_{s=1}^{t-1}$ and x_t. Let $\{y_t\}_{t \geq 1}$ be a sequence of noisy observations at the query points $\{x_t\}_{t \geq 1}$, where $y_t = f(x_t) + \varepsilon_t$. For $\lambda > 0$ and $x \in \mathcal{X}$, let

$$\begin{aligned} \mu_{t-1}(x) &:= k_{t-1}(x)^T (K_{t-1} + \lambda I)^{-1} Y_{t-1}, \\ \sigma_{t-1}^2(x) &:= k(x, x) - k_{t-1}(x)^T (K_{t-1} + \lambda I)^{-1} k_{t-1}(x), \end{aligned}$$

where $Y_{t-1} := [y_1, \ldots, y_{t-1}]^T$ denotes the vector of observations at $\{x_1, \ldots, x_{t-1}\}$. Then, for any $0 < \delta \leq 1$, with probability at least $1 - \delta$, uniformly over $t \geq 1, x \in \mathcal{X}$,

$$|f(x) - \mu_{t-1}(x)| \leq \left(\|f\|_k + \frac{R}{\sqrt{\lambda}} \sqrt{2\Big(\ln(1/\delta) + \frac{1}{2} \sum_{s=1}^{t-1} \ln\left(1 + \lambda^{-1} \sigma_{s-1}^2(x_s)\right)\Big)} \right) \sigma_{t-1}(x).$$

Proof The proof follows from the proof of Theorem 2.1 in Durand et al. [2017]. ■

A.2 Relevent Results on Episodic Continuous Markov Decision Processes

Definition 2 (Bellman operator) *For any MDP $M = \{\mathcal{S}, \mathcal{A}, R_M, P_M, H\}$, any policy $\pi : \mathcal{S} \times \{1, \dots, H\} \to \mathcal{A}$, any period $1 \le h \le H$, any value function $V : \mathcal{S} \to \mathbb{R}$ and any state $s \in \mathcal{S}$, the Bellman operator $T^M_{\pi,h}$ is defined as*

$$\big(T^M_{\pi,h} V\big)(s) = \overline{R}_M\big(s, \pi(s,h)\big) + \mathbb{E}_{s'}\big[V(s')\big],$$

where the subscript s' implies that $s' \sim P_M\big(s, \pi(s,h)\big)$ and $\overline{R}_M$ denotes the mean reward function.

This operator returns the expected value of the state s, where we follow the policy $\pi(s,h)$ for one step under P_M.

Lemma 6 (Bellman equation) *For any MDP $M = \{\mathcal{S}, \mathcal{A}, R_M, P_M, H\}$, any policy $\pi : \mathcal{S} \times \{1, \dots, H\} \to \mathcal{A}$ and any period $1 \le h \le H$, the value functions $V^M_{\pi,h}$ satisfy*

$$V^M_{\pi,h}(s) = \big(T^M_{\pi,h} V^M_{\pi,h+1}\big)(s)$$

for all $s \in \mathcal{S}$, with $V^M_{\pi,H+1} := 0$.

Proof For any MDP $M = \{\mathcal{S}, \mathcal{A}, R_M, P_M, H, \}$ and policy $\pi : \mathcal{S} \times \{1, \dots, H\} \to \mathcal{A}$, period $h \in \{1, \dots, H\}$ and state $s \in \mathcal{S}$, recall the finite horizon, undiscounted, value function

$$V^M_{\pi,h}(s) := \mathbb{E}_{M,\pi}\left[\sum_{j=h}^{H} \overline{R}_M(s_j, a_j) \;\middle|\; s_h = s\right],$$

where the subscript π indicates the application of the learning policy π, i.e., $a_j = \pi(s_j, j)$, and the subscript M explicitly references the MDP environment M, i.e., $s_{j+1} \sim P_M(s_j, a_j)$, for all $j = h, \dots, H$. See that, by definition, $V^M_{\pi,H+1}(s) = 0$ for all $s \in \mathcal{S}$. Further $V^M_{\pi,h}(s)$ can be rewritten as

$$\begin{aligned}
V^M_{\pi,h}(s) &= \overline{R}_M\big(s, \pi(s,h)\big) + \mathbb{E}_{M,\pi}\left[\sum_{j=h+1}^{H} \overline{R}_M(s_j, a_j) \;\middle|\; s_h = s\right] \\
&= \overline{R}_M\big(s, \pi(s,h)\big) + \mathbb{E}_{s'}\left[\mathbb{E}_{M,\pi}\left[\sum_{j=h+1}^{H} \overline{R}_M(s_j, a_j) \;\middle|\; s_{h+1} = s'\right]\right] \\
&= \overline{R}_M(s, \pi(s,h)) + \mathbb{E}_{s'}\Big[V^M_{\pi,h+1}(s')\Big],
\end{aligned}$$

where the subscript s' implies that $s' \sim P_M\big(s, \pi(s,h)\big)$. Now the result follows from Definition 2. ■

Lemma 7 (Bounds on deviations of rewards and transitions imply bounds on deviation of the value function) *Let $M_l, l \ge 1$ be a sequence of MDPs and for each $l \ge 1$ and π_l be the optimal policy for the MDP M_l. Let $M_\star$ be an MDP with the transition function $P_\star$ and $s_{l,h+1} \sim P_\star(s_{l,h}, a_{l,h})$, where $a_{l,h} = \pi_l(s_{l,h}, h)$. Now for all $l \ge 1$ and $1 \le h \le H$, define*

$$\Delta_{l,h} := \mathbb{E}_{s' \sim P_\star(z_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\Big] - \Big(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\Big),$$

where $z_{l,h} := (s_{l,h}, a_{l,h})$. Then for any $\tau \ge 1$,

$$\sum_{l=1}^{\tau}\Big(V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\Big) \le \sum_{l=1}^{\tau}\sum_{h=1}^{H}\Big(\big|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\big| + L_{M_l}\big\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\big\|_2 + \Delta_{l,h}\Big),$$

where L_{M_l} is defined to be the global Lipschitz constant (1) of one step future value function for MDP M_l.

Proof The arguments in this proof borrow ideas from Osband et al. [2013]. Applying Lemma 6 for $h = 1$, $s = s_{l,1}$ and two MDP-policy pairs (M_l, π_l) and $(M_\star, \pi_l)$, we have

$$\begin{aligned}
V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}) &= \big(T^{M_l}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) - \big(T^{M_\star}_{\pi_l,1} V^{M_\star}_{\pi_l,2}\big)(s_{l,1}) \\
&= \big(T^{M_l}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) - \big(T^{M_\star}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) + \big(T^{M_\star}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) - \big(T^{M_\star}_{\pi_l,1} V^{M_\star}_{\pi_l,2}\big)(s_{l,1}).
\end{aligned}$$

Further using Definition 2 for $M = M_\star$, $\pi = \pi_l$, $h = 1$, $V = V^{M_l}_{\pi_l,2}$ and $s = s_{l,1}$, we have

$$\big(T^{M_\star}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) = \overline{R}_\star\big(s_{l,1}, \pi_l(s_{l,1},1)\big) + \mathbb{E}_{s'\sim P_\star(s_{l,1},\pi_l(s_{l,1},1))}\Big[V^{M_l}_{\pi_l,2}(s')\Big], \tag{20}$$

where $R_\star$ and $P_\star$ denote the reward and transition functions of the MDP $M_\star$ respectively. Again using Definition 2 for $M = M_\star$, $\pi = \pi_l$, $h = 1$, $V = V^{M_\star}_{\pi_l,2}$ and $s = s_{l,1}$, we have

$$\big(T^{M_\star}_{\pi_l,1} V^{M_\star}_{\pi_l,2}\big)(s_{l,1}) = \overline{R}_\star\big(s_{l,1}, \pi_l(s_{l,1},1)\big) + \mathbb{E}_{s'\sim P_\star(s_{l,1},\pi_l(s_{l,1},1))}\Big[V^{M_\star}_{\pi_l,2}(s')\Big]. \tag{21}$$

Subtracting (21) from (20), we have

$$\begin{aligned}\big(T^{M_\star}_{\pi_l,1} V^{M_l}_{\pi_l,2}\big)(s_{l,1}) - \big(T^{M_\star}_{\pi_l,1} V^{M_\star}_{\pi_l,2}\big)(s_{l,1}) &= \mathbb{E}_{s'\sim P_\star(s_{l,1},\pi_l(s_{l,1},1))}\Big[V^{M_l}_{\pi_l,2}(s') - V^{M_\star}_{\pi_l,2}(s')\Big] \\ &= V^{M_l}_{\pi_l,2}(s_{l,2}) - V^{M_\star}_{\pi_l,2}(s_{l,2}) + \Delta_{l,1},\end{aligned}$$

where $\Delta_{l,1} := \mathbb{E}_{s'\sim P_\star(s_{l,1},\pi_l(s_{l,1},1))}\Big[V^{M_l}_{\pi_l,2}(s') - V^{M_\star}_{\pi_l,2}(s')\Big] - \Big(V^{M_l}_{\pi_l,2}(s_{l,2}) - V^{M_\star}_{\pi_l,2}(s_{l,2})\Big)$. Then (20) implies

$$V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}) = V^{M_l}_{\pi_l,2}(s_{l,2}) - V^{M_\star}_{\pi_l,2}(s_{l,2}) + (T^{M_l}_{\pi_l,1} V^{M_l}_{\pi_l,2})(s_{l,1}) - (T^{M_\star}_{\pi_l,1} V^{M_l}_{\pi_l,2})(s_{l,1}) + \Delta_{l,1}.$$

Now since $V^M_{\pi,H+1}(s) = 0$ for any MDP M, policy π and state s, an inductive argument gives

$$V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}) = \sum_{h=1}^{H}\Big(\big(T^{M_l}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) - \big(T^{M_\star}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) + \Delta_{l,h}\Big), \tag{22}$$

where $\Delta_{l,h} := \mathbb{E}_{s'\sim P_\star(s_{l,h},\pi_l(s_{l,h},h))}\Big[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\Big] - \Big(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\Big)$.

Now using Definition 2 respectively for $M = M_l$ and $M = M_\star$ with $\pi = \pi_l$, $V = V^{M_l}_{\pi_l,h+1}$ and $s = s_{l,h}$, we have

$$\begin{aligned}\big(T^{M_l}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) - \big(T^{M_\star}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) = &\Big(\overline{R}_{M_l}\big(s_{l,h}, \pi_l(s_{l,h},h)\big) + \mathbb{E}_{s'\sim P_{M_l}(s_{l,h},\pi_l(s_{l,h},h))}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big]\Big) \\ &-\Big(\overline{R}_\star\big(s_{l,h}, \pi_l(s_{l,h},h)\big) + \mathbb{E}_{s'\sim P_\star(s_{l,h},\pi_l(s_{l,h},h))}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big]\Big).\end{aligned}$$

Further using the fact that $a_{l,h} = \pi_l(s_{l,h}, h))$ and defining $z_{l,h} = (s_{l,h}, a_{l,h})$, we have

$$\begin{aligned}&\big(T^{M_l}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) - \big(T^{M_\star}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) \\ &= \overline{R}_{M_l}(s_{l,h}, a_{l,h}) - \overline{R}_\star(s_{l,h}, a_{l,h}) + \mathbb{E}_{s'\sim P_{M_l}(s_{l,h},a_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big] - \mathbb{E}_{s'\sim P_\star(s_{l,h},a_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big] \\ &= \overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h}) + \mathbb{E}_{s'\sim P_{M_l}(z_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big] - \mathbb{E}_{s'\sim P_\star(z_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s')\Big].\end{aligned}$$

and $\Delta_{l,h} = \mathbb{E}_{s'\sim P_\star(z_{l,h})}\Big[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\Big] - \Big(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\Big)$.

Now for an MDP M, a distribution φ over $\mathcal{S}$ and for every period $1 \le h \le H$, recall that the *one step future value function* is defined as

$$U^M_h(\varphi) := \mathbb{E}_{s'\sim\varphi}\Big[V^M_{\pi_M,h+1}(s')\Big],$$

where π_M denotes the optimal policy for the MDP M. Observe that π_l is the optimal policy for the MDP M_l. This implies

$$\big(T^{M_l}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) - \big(T^{M_\star}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) = \overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h}) + U^{M_l}_h\big(P_{M_l}(z_{l,h})\big) - U^{M_l}_h\big(P_\star(z_{l,h})\big).$$

Further (1) implies

$$U^{M_l}_h\big(P_{M_l}(z_{l,h})\big) - U^{M_l}_h\big(P_\star(z_{l,h})\big) \le L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2,$$

where L_{M_l} is defined to be the global Lipschitz constant (1) of one step future value function for MDP M_l. Hence we have

$$\big(T^{M_l}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) - \big(T^{M_\star}_{\pi_l,h} V^{M_l}_{\pi_l,h+1}\big)(s_{l,h}) \le \left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2. \tag{23}$$

Now the result follows by plugging (23) back in (22) and summing over $l = 1, \ldots, \tau$. ■

B ANALYSIS OF GP-UCRL AND PSRL IN THE KERNELIZED MDPs

B.1 Preliminary Definitions and Results

Now we define the *span* of an MDP, which is crucial to measure the difficulties in learning the optimal policy of the MDP [Jaksch et al., 2010, Bartlett and Tewari, 2009].

Definition 3 (Span of an MDP) *The span of an MDP M is the maximum difference in value of any two states under the optimal policy, i.e.*

$$\Psi_M := \max_{s,s' \in \mathcal{S}} V^M_{\pi_M,1}(s) - V^M_{\pi_M,1}(s').$$

Now define $\Psi_{M,h} := \max_{s,s' \in \mathcal{S}} V^M_{\pi_M,h}(s) - V^M_{\pi_M,h}(s')$ as the span of M at period h and let $\tilde{\Psi}_M := \max_{h \in \{1,\dots,H\}} \Psi_{M,h}$ as the maximum possible span in an episode. Clearly $\Psi_M \leq \tilde{\Psi}_M$.

Definition 4 *A sequence of random variables $\{Z_t\}_{t \geq 1}$ is called a martingale difference sequence corresponding to a filtration $\{\mathcal{F}_t\}_{t \geq 0}$, if for all $t \geq 1$, Z_t is $\mathcal{F}_t$-measurable, and for all $t \geq 1$,*

$$\mathbb{E}\left[Z_t \,\middle|\, \mathcal{F}_{t-1}\right] = 0.$$

Lemma 8 (Azuma-Hoeffding Inequality) *If a martingale difference sequence $\{Z_t\}_{t \geq 1}$, corresponding to filtration $\{\mathcal{F}_t\}_{t \geq 0}$, satisfies $|Z_t| \leq \alpha_t$ for some constant α_t, for all $t = 1, \dots, T$, then for any $0 < \delta \leq 1$,*

$$\mathbb{P}\left[\sum_{t=1}^{T} Z_t \leq \sqrt{2\ln(1/\delta)\sum_{t=1}^{T}\alpha_t^2}\right] \geq 1 - \delta.$$

Lemma 9 (Bound on Martingale difference sequence) *Let $\mathcal{M}_l$ be the set of plausible MDPs constructed by GP-UCRL (Algorithm 1) at episode $l, l \geq 1$ and $M_l, l \geq 1$ be a sequence of MDPs such that $M_l \in \mathcal{M}_l$ for each $l \geq 1$ and π_l be the optimal policy for the MDP M_l for each $l \geq 1$. Let $M_\star$ be an MDP with reward function $R_\star$ and transition function $P_\star$. Let $s_{l,h+1} \sim P_\star(s_{l,h}, a_{l,h})$, where $a_{l,h} = \pi_l(s_{l,h}, h)$. Now define $\Delta_{l,h} := \mathbb{E}_{s' \sim P_\star(z_{l,h})}\left[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\right] - \left(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\right)$, with $z_{l,h} := (s_{l,h}, a_{l,h})$. Then for any $0 \leq \delta \leq 1$ and $\tau \geq 1$, with probability at least $1 - \delta$,*

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \Delta_{l,h} \leq (LD + 2CH)\sqrt{2\tau H \ln(1/\delta)},$$

where $D := \max_{s,s' \in \mathcal{S}} \|s - s'\|_2$ is the diameter of the state space $\mathcal{S}$, C is a uniform upper bound over the absolute value of the mean reward function $\overline{R}_\star$, i.e. $\left|\overline{R}_\star(z)\right| \leq C$ for all $z \in \mathcal{Z}$ and L is an upper bound over the global Lipschitz constant (1) of one step future value function for MDP $M_\star$, i.e. $L_{M_\star} \leq L$.

Proof First assume that $M_\star$ is fixed in advance. For each $l \geq 1$ and $h \in \{1, \dots H\}$, we define $\mathcal{H}_{l-1} := \{s_{j,k}, a_{j,k}, r_{j,k}, s_{j,k+1}\}_{1 \leq j \leq l-1, 1 \leq k \leq H}$ as the history of all observations till episode $l-1$ and $\mathcal{G}_{l,h} := \mathcal{H}_{l-1} \cup \{s_{l,k}, a_{l,k}, r_{l,k}, s_{l,k+1}\}_{1 \leq k \leq h}$ as the history of all observations till episode l and period h. See that $\mathcal{H}_0 = \emptyset$ and $\mathcal{H}_l = G_{l,H}$ for all $l \geq 1$. Further defining $\mathcal{G}_{l,0} := \mathcal{H}_{l-1} \cup \{s_{l,1}\}$, we see that $\mathcal{G}_{l,h} = \mathcal{G}_{l,h-1} \cup \{a_{l,h}, r_{l,h}, s_{l,h+1}\}$ for all $h \in \{1, \dots H\}$. Clearly the sets $\mathcal{G}_{l,h}$ satisfy $G_{l,0} \subset G_{l,1} \subset G_{l,2} \subset \dots \subset G_{l,H} \subset G_{l+1,0}$ for all $l \geq 1$. Hence, the sequence of sets $\{\mathcal{G}_{l,h}\}_{l \geq 1, 0 \leq h \leq H}$ defines a filtration.

Now by construction in Algorithm 1, $\mathcal{M}_l$ is deterministic given $\mathcal{H}_{l-1}$. Hence, M_l and π_l are also deterministic given $\mathcal{H}_{l-1}$. This implies $\Delta_{l,h} = \mathbb{E}_{s' \sim P_\star(z_{l,h})}\left[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\right] - \left(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\right)$ is $G_{l,h}$ - measurable. Further note that $a_{l,h} = \pi_l(s_{l,h}, h)$ is deterministic given $\mathcal{G}_{l,h-1}$, as both π_l and $s_{l,h}$ are deterministic given $\mathcal{G}_{l,h-1}$. This implies

$$\begin{aligned}\mathbb{E}\left[\Delta_{l,h} \,\middle|\, \mathcal{G}_{l,h-1}\right] &= \mathbb{E}_{s' \sim P_\star(z_{l,h})}\left[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\right] - \mathbb{E}_{s_{l,h+1} \sim P_\star(z_{l,h})}\left[V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\right] \\ &= 0. \end{aligned} \tag{24}$$

Further, observe that $|\Delta_{l,h}| \leq \big(\max_s V^{M_l}_{\pi_l,h+1}(s) - \min_s V^{M_l}_{\pi_l,h+1}(s)\big) + \big(\max_s V^{M_\star}_{\pi_l,h+1}(s) - \min_s V^{M_\star}_{\pi_l,h+1}(s)\big)$. The first term $\max_s V^{M_l}_{\pi_l,h+1}(s) - \min_s V^{M_l}_{\pi_l,h+1}(s)$ is upper bounded by $\tilde{\Psi}_{M_l}$, which is an upper bound over the *span* of the MDP M_l (Definition 3). Now from (1), we get $\Psi_{M_l} \leq L_{M_l} D$, where $D := \max_{s,s' \in \mathcal{S}} \|s - s'\|_2$ is the diameter of the state space $\mathcal{S}$ and L_{M_l} is a global Lipschitz constant for the one step future value function. Further by construction of the set of plausible MDPs $\mathcal{M}_l$ and as $M_l \in \mathcal{M}_l$, we have $L_{M_l} \leq L$. Hence, we have $\max_s V^{M_l}_{\pi_l,h+1}(s) - \min_s V^{M_l}_{\pi_l,h+1}(s) \leq LD$. Now, since by our hypothesis $\big|\overline{R}_\star(z)\big| \leq C$ for all $z \in \mathcal{Z}$, see that $V^{M_\star}_{\pi,h}(s) \leq CH$ for all π, $1 \leq h \leq H$ and $s \in \mathcal{S}$. Hence, we have $\max_s V^{M_\star}_{\pi_l,h+1}(s) - \min_s V^{M_\star}_{\pi_l,h+1}(s) \leq 2CH$.

Therefore the sequence of random variables $\{\Delta_{l,h}\}_{l \geq 1, 1 \leq h \leq H}$ is a martingale difference sequence (Definition 4) with respect to the filtration $\{\mathcal{G}_{l,h}\}_{l \geq 1, 0 \leq h \leq H}$, with $|\Delta_{l,h}| \leq LD + 2CH$ for all $l \geq 1$ and $1 \leq h \leq H$. Thus, by Lemma 8, for any $\tau \geq 1$ and $0 < \delta \leq 1$, we have with probability at least $1 - \delta$,

$$\sum_{l=1}^{\tau} \sum_{h=1}^{H} \Delta_{l,h} \leq \sqrt{2 \ln(1/\delta) \sum_{l=1}^{\tau} \sum_{h=1}^{H} (LD + 2CH)^2} = (LD + 2CH)\sqrt{2\tau H \ln(1/\delta)}. \tag{25}$$

Now consider the case when $M_\star$ is random. Then we define $\tilde{\mathcal{H}}_{l-1} := \mathcal{H}_{l-1} \cup M_\star$ and $\tilde{\mathcal{G}}_{l,h} := \mathcal{G}_{l,h} \cup M_\star$. Then $\{\Delta_{l,h}\}_{l \geq 1, 1 \leq h \leq H}$ is a martingale difference sequence with respect to the filtration $\{\tilde{\mathcal{G}}_{l,h}\}_{l \geq 1, 0 \leq h \leq H}$, and hence (25) holds in this case also. ■

B.2 Analysis of GP-UCRL in Kernelized MDPs

Recall that at each episode $l \geq 1$, GP-UCRL constructs confidence sets $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ as

$$\begin{aligned} \mathcal{C}_{R,l} &= \big\{f : \mathcal{Z} \to \mathbb{R} \;\big|\; |f(z) - \mu_{R,l-1}(z)| \leq \beta_{R,l}\sigma_{R,l-1}(z) \forall z \in \mathcal{Z}\big\}, \\ \mathcal{C}_{P,l} &= \big\{f : \mathcal{Z} \to \mathbb{R}^m \;\big|\; \|f(z) - \mu_{P,l-1}(z)\|_2 \leq \beta_{P,l} \|\sigma_{P,l-1}(z)\|_2 \, \forall z \in \mathcal{Z}\big\}, \end{aligned} \tag{26}$$

where $\mu_{R,0}(z) = 0$, $\sigma^2_{R,0}(z) = k_R(z,z)$ and for each $l \geq 1$,

$$\begin{aligned} \mu_{R,l}(z) &= k_{R,l}(z)^T (K_{R,l} + HI)^{-1} R_l, \\ \sigma^2_{R,l}(z) &= k_R(z,z) - k_{R,l}(z)^T (K_{R,l} + HI)^{-1} k_{R,l}(z). \end{aligned} \tag{27}$$

Here H is the number of periods, I is the $(lH) \times (lH)$ identity matrix, $R_l = [r_{1,1}, \ldots, r_{l,H}]^T$ is the vector of rewards observed at $\mathcal{Z}_l = \{z_{j,k}\}_{1 \leq j \leq l, 1 \leq k \leq H} = \{z_{1,1}, \ldots, z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{R,l}(z) = [k_R(z_{1,1}, z), \ldots, k_R(z_{l,H}, z)]^T$ is the vector kernel evaluations between z and elements of the set $\mathcal{Z}_l$ and $K_{R,l} = [k_R(u,v)]_{u,v \in \mathcal{Z}_l}$ is the kernel matrix computed at $\mathcal{Z}_l$. Further $\mu_{P,l}(z) = [\mu_{P,l-1}(z,1), \ldots, \mu_{P,l-1}(z,m)]^T$ and $\sigma_{P,l}(z) = [\sigma_{P,l-1}(z,1), \ldots, \sigma_{P,l-1}(z,m)]^T$, where $\mu_{P,0}(z,i) = 0$, $\sigma_{P,0}(z,i) = k_P\big((z,i),(z,i)\big)$ and for each $l \geq 1$,

$$\begin{aligned} \mu_{P,l}(z,i) &= k_{P,l}(z,i)^T (K_{P,l} + mHI)^{-1} S_l, \\ \sigma^2_{P,l}\big((z,i)\big) &= k_P((z,i),(z,i)) - k_{P,l}(z,i)^T (K_{P,l} + mHI)^{-1} k_{P,l}(z,i). \end{aligned} \tag{28}$$

Here m is the dimension of the state space, H is the number of periods, I is the $(mlH) \times (mlH)$ identity matrix, $S_l = [s^T_{1,2}, \ldots, s^T_{l,H+1}]^T$ denotes the vector of state transitions at $\mathcal{Z}_l = \{z_{1,1}, \ldots, z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{P,l}(z,i) = \Big[k_P\big((z_{1,1},1),(z,i)\big), \ldots, k_P\big((z_{l,H},m),(z,i)\big)\Big]^T$ is the vector of kernel evaluations between (z,i) and elements of the set $\tilde{\mathcal{Z}}_l = \big\{(z_{j,k},i)\big\}_{1 \leq j \leq l, 1 \leq k \leq H, 1 \leq i \leq m} = \big\{(z_{1,1},1), \ldots, (z_{l,H},m)\big\}$ and $K_{P,l} = \big[k_P(u,v)\big]_{u,v \in \tilde{\mathcal{Z}}_l}$ is the kernel matrix computed at $\tilde{\mathcal{Z}}_l$. Here for any $0 < \delta \leq 1$, $B_R, B_P, \sigma_R, \sigma_P > 0$, $\beta_{R,l} := B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3/\delta) + \gamma_{(l-1)H}(k_R, \lambda_R, \mathcal{Z})\big)}$ and $\beta_{P,l} := B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3/\delta) + \gamma_{m(l-1)H}(k_P, \lambda_P, \tilde{\mathcal{Z}})\big)}$ are properly chosen confidence widths of $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ respectively.

Lemma 10 (Concentration of mean reward and mean transition functions) *Let $M_\star = \{\mathcal{S}, \mathcal{A}, R_\star, P_\star, H\}$ be an MDP with period H, state space $\mathcal{S} \subset \mathbb{R}^m$ and action space $\mathcal{A} \subset \mathbb{R}^n$. Let the mean reward function $\overline{R}_\star$ be a member of the*

RKHS $\mathcal{H}_{k_R}(\mathcal{Z})$ corresponding to the kernel $k_R : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ and let the noise variables $\varepsilon_{R,l,h}$ be conditionally σ_R-sub-Gaussian (6). Let the mean transition function $\overline{P}_\star$ be a member of the RKHS $\mathcal{H}_{k_P}(\tilde{\mathcal{Z}})$ corresponding to the kernel $k_P : \mathcal{Z} \times \tilde{\mathcal{Z}} \to \mathbb{R}$, where $\tilde{\mathcal{Z}} := \mathcal{Z} \times \{1, \dots, m\}$ and let the noise variables $\varepsilon_{P,l,h}$ be conditionally component-wise independent and σ_P-sub-Gaussian (7). Further let $\|\overline{R}_\star\|_{k_R} \leq B_R$ and $\|\overline{P}_\star\|_{k_P} \leq B_P$. Then, for any $0 < \delta \leq 1$, the following holds:

$$\mathbb{P}\left[\forall z \in \mathcal{Z}, \forall l \geq 1, \left|\overline{R}_\star(z) - \mu_{R,l-1}(z)\right| \leq \beta_{R,l}\, \sigma_{R,l-1}(z)\right] \geq 1 - \delta/3, \tag{29}$$

$$\mathbb{P}\left[\forall z \in \mathcal{Z}, \forall l \geq 1, \left\|\overline{P}_\star(z) - \mu_{P,l-1}(z)\right\|_2 \leq \beta_{P,l}\, \|\sigma_{P,l-1}(z)\|_2\right] \geq 1 - \delta/3, \tag{30}$$

Proof First fix any $0 < \delta \leq 1$. Now for all $l \geq 1$ and $1 \leq h \leq H+1$, let us define $\mathcal{Z}_{l,h-1} = \{z_{j,k}\}_{1 \leq j \leq l-1, 1 \leq k \leq H} \cup \{z_{l,k}\}_{1 \leq k \leq h-1} = \{z_{1,1}, \dots, z_{1,H}, \dots, z_{l,1}, \dots, z_{l,h-1}\}$ as the set of all state-action pairs available till period $h-1$ of episode l. Further, let $R_{l,h-1} = [r_{1,1}, \dots, r_{1,H}, \dots, r_{l,1}, \dots, r_{l,h-1}]^T$ denotes the vector of rewards observed at $\mathcal{Z}_{l,h-1}$. See that $\mathcal{Z}_{l,0} = \mathcal{Z}_{l-1,H}$ and $R_{l,0} = R_{l-1,H}$ for all $l \geq 2$. Also $R_{1,0} = 0$ and $\mathcal{Z}_{1,0} = \emptyset$. Now we define, for all $z \in \mathcal{Z}$, $l \geq 1$ and $1 \leq h \leq H+1$, the following:

$$\begin{aligned} \mu_{R,l,h-1}(z) &= k_{R,l,h-1}(z)^T (K_{R,l,h-1} + HI)^{-1} R_{l,h-1}, \\ \sigma^2_{R,l,h-1}(z) &= k_R(z,z) - k_{R,l,h-1}(z)^T (K_{R,l,h-1} + HI)^{-1} k_{R,l,h-1}(z), \end{aligned} \tag{31}$$

where $k_{R,l,h-1}(z) = [k_R(z_{1,1}, z), \dots, k_R(z_{1,H}, z), \dots, k_R(z_{l,1}, z), \dots, k_R(z_{l,h-1}, z)]^T$ is the vector kernel evaluations between z and elements of the set $\mathcal{Z}_{l,h-1}$, $K_{R,l,h-1} = [k_R(z,z')]_{z,z' \in \mathcal{Z}_{l,h-1}}$ is the kernel matrix computed at $\mathcal{Z}_{l,h-1}$. See that $\mu_{R,l,0}(z) = \mu_{R,l-1,H}(z)$, and $\sigma_{R,l,0}(z) = \sigma_{R,l-1,H}(z)$ for all $l \geq 2$ and $z \in \mathcal{Z}$. Also $\mu_{R,1,0}(z) = 0$ and $\sigma_{R,1,0} = k_R(z,z)$ for all $z \in \mathcal{Z}$.

At the state-action pair $z_{l,h}$, the reward observed is $r_{l,h} = \overline{R}_\star(z_{l,h}) + \varepsilon_{R,l,h}$. Here, by our hypothesis, the mean reward function $\overline{R}_\star \in \mathcal{H}_{k_R}(\mathcal{Z})$ and the noise sequence $\{\varepsilon_{R,l,h}\}_{l \geq 1, 1 \leq h \leq H}$ is conditionally σ_R-sub-Gaussian. Now Lemma 5 implies that, with probability at least $1 - \delta/3$, uniformly over all $z \in \mathcal{Z}$, $l \geq 1$ and $1 \leq h \leq H$,

$$\left|\overline{R}_\star(z) - \mu_{R,l,h-1}(z)\right| \leq \left(\|\overline{R}_\star\|_{k_R} + \frac{\sigma_R}{\sqrt{H}} \sqrt{2\Big(\ln(3/\delta) + \frac{1}{2} \sum_{(j,k)=(1,1)}^{(l,h-1)} \ln\left(1 + H^{-1} \sigma^2_{R,j,k-1}(z_{j,k})\right)\Big)} \right) \sigma_{R,l,h-1}(z).$$

Again, from Lemma 1, we have

$$\frac{1}{2} \sum_{(j,k)=(1,1)}^{(l,h-1)} \ln\left(1 + H^{-1} \sigma^2_{R,j,k-1}(z_{j,k})\right) \leq \gamma_{(l-1)H+h-1}(k_R, \mathcal{Z}),$$

where $\gamma_t(k_R, \mathcal{Z})$ denotes the maximum information gain about an $f \sim GP_{\mathcal{Z}}(0, k_R)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0, H)$. Therefore, with probability at least $1 - \delta/3$, uniformly over all $z \in \mathcal{Z}$, $l \geq 1$ and $1 \leq h \leq H$,

$$\left|\overline{R}_\star(z) - \mu_{R,l,h-1}(z)\right| \leq \Big(B_R + \frac{\sigma_R}{\sqrt{H}} \sqrt{2\big(\ln(3/\delta) + \gamma_{(l-1)H+h-1}(k_R, \mathcal{Z})\big)} \Big) \sigma_{R,l,h-1}(z), \tag{32}$$

since by our hypothesis $\|\overline{R}_\star\|_{k_R} \leq B_R$. Now see that $\mu_{R,l,0} = \mu_{R,l-1}$ and $\sigma_{R,l,0} = \sigma_{R,l-1}$ and $\beta_{R,l} = B_R + \frac{\sigma_R}{\sqrt{H}} \sqrt{2\big(\ln(3/\delta) + \gamma_{(l-1)H}(k_R, \lambda_R, \mathcal{Z})\big)}$ for every $l \geq 1$. Hence (29) follows by using (32) with $h = 1$.

Further for all $l \geq 1$ and $1 \leq h \leq H+1$, let $S_{l,h} = [s^T_{1,2}, \dots, s^T_{1,H+1}, \dots, s^T_{l,2}, \dots, s^T_{l,h}]^T$ denotes the vector of state transitions at $\mathcal{Z}_{l,h-1} = \{z_{1,1}, \dots, z_{1,H}, \dots, z_{l,1}, \dots, z_{l,h-1}\}$, where every state $s_{j,h} = [s_{j,h}(1), \dots, s_{j,h}(m)]^T, 1 \leq j \leq l, 1 \leq h \leq H+1$, is an m - dimensional vector. Further for all $l \geq 1$, $1 \leq h \leq H$ and $1 \leq b \leq m+1$, define the following set:

$$\begin{aligned} \tilde{\mathcal{Z}}_{l,h,b-1} &= \big\{(z_{j,k}, i)\big\}_{1 \leq j \leq l-1, 1 \leq k \leq H, 1 \leq i \leq m} \cup \big\{(z_{l,k}, i)\big\}_{1 \leq k \leq h-1, 1 \leq i \leq m} \cup \big\{(z_{l,h}, i)\big\}_{1 \leq i \leq b-1} \\ &= \big\{\mathcal{Z}_{l,h-1} \times \{1, \dots, m\}\big\} \cup \big\{\{z_{l,h}\} \times \{1, \dots, b-1\}\big\} \\ &= \big\{(z_{1,1}, 1), \dots, (z_{1,1}, m), \dots, (z_{l,h-1}, 1), \dots, (z_{l,h-1}, m), (z_{l,h}, 1), \dots, (z_{l,h}, b-1)\big\}, \end{aligned}$$

and the following vector

$$\begin{aligned} S_{l,h,b-1} &= [S_{l,h}^T, s_{l,h+1}(1), \ldots, s_{l,h+1}(b-1)]^T \\ &= [s_{1,2}^T, \ldots, s_{1,H+1}^T, \ldots, s_{l,2}^T, \ldots, s_{l,h}^T, s_{l,h+1}(1), \ldots, s_{l,h+1}(b-1)]^T \\ &= [s_{1,2}(1), \ldots, s_{1,2}(m), \ldots, s_{l,h}(1), \ldots, s_{l,h}(m), s_{l,h+1}(1), \ldots, s_{l,h+1}(b-1)]^T. \end{aligned}$$

See that $\mathcal{Z}_{l,h,0} = \mathcal{Z}_{l,h-1,m}$ and $S_{l,h,0} = S_{l,h-1,m}$ for all $l \geq 1$ and $2 \leq h \leq H$. Further $\mathcal{Z}_{l,1,0} = \mathcal{Z}_{l-1,H,m}$ and $S_{l,1,0} = S_{l-1,H,m}$ for all $l \geq 2$. Also $\mathcal{Z}_{1,1,0} = 0$ and $S_{1,1,0} = 0$. Now we define, for all $z \in \mathcal{Z}$, $1 \leq i \leq m$, $l \geq 1$ and $1 \leq h \leq H+1$, the following:

$$\begin{aligned} \mu_{P,l,h,b-1}(z,i) &= k_{P,l,h,b-1}(z,i)^T (K_{P,l,h,b-1} + mHI)^{-1} S_{l,h,b-1}, \\ \sigma^2_{P,l,h,b-1}(z,i) &= k_P\big((z,i),(z,i)\big) - k_{P,l,h,b-1}(z,i)^T (K_{P,l,h,b-1} + mHI)^{-1} k_{P,l,h,b-1}(z,i), \end{aligned} \tag{33}$$

where $k_{P,l,h,b-1}(z,i) = \Big[k_P\big((z_{1,1},1),(z,i)\big), \ldots, k_P\big((z_{l,h},b-1),(z,i)\big)\Big]^T$ is the vector of kernel evaluations between (z,i) and elements of the set $\tilde{\mathcal{Z}}_{l,h,b-1}$, $K_{P,l,h,b-1} = \big[k_P(z,z')\big]_{z,z' \in \tilde{\mathcal{Z}}_{l,h,b-1}}$ is the kernel matrix computed at $\tilde{\mathcal{Z}}_{l,h,b-1}$. See that $\mu_{P,l,h,0} = \mu_{P,l,h-1,m}$ and $\sigma_{P,l,h,0} = \sigma_{P,l,h-1,m}$ for all $l \geq 1$ and $2 \leq h \leq H$. Further, $\mu_{P,l,1,0} = \mu_{P,l-1,H,m}$ and $\sigma_{P,l,1,0} = \sigma_{P,l-1,H,m}$ for all $l \geq 2$. Also $\mu_{P,1,1,0}(z,i) = 0$ and $\sigma_{P,1,1,0} = k_P\big((z,i),(z,i)\big)$ for all $z \in \mathcal{Z}$ and $1 \leq i \leq m$.

At the state-action pair $z_{l,h}$, the MDP transitions to the state $s_{l,h+1}$, where $s_{l,h+1}(i) = \overline{P}_\star(z_{l,h},i) + \varepsilon_{P,l,h}(i)$, $1 \leq i \leq m$. Thus, we can view $s_{l,h+1}(i)$ as a noisy observation of $\overline{P}_\star$ at the query $(z_{l,h},i) \in \tilde{\mathcal{Z}}$. Here, by our hypothesis, the mean transition function $\overline{P}_\star \in \mathcal{H}_{\tilde{k}_P}(\tilde{\mathcal{Z}})$ and the noise sequence $\{\varepsilon_{P,l,h}(i)\}_{l \geq 1, 1 \leq h \leq H, 1 \leq i \leq m}$ is conditionally σ_P-sub-Gaussian. Now Lemma 5 implies that, with probability at least $1-\delta/3$, uniformly over all $z \in \mathcal{Z}$, $1 \leq i \leq m$, $l \geq 1$, $1 \leq h \leq H$ and $1 \leq b \leq m$:

$$\big|\overline{P}_\star(z,i) - \mu_{P,l,h,b-1}(z,i)\big| \leq \left(\big\|\overline{P}_\star\big\|_{k_P} + \frac{\sigma_P}{\sqrt{mH}} \sqrt{2\Big(\ln(3/\delta) + \frac{1}{2} \sum_{(j,k,q)=(1,1,1)}^{(l,h,b-1)} \ln\Big(1 + \frac{\sigma^2_{P,j,k,q-1}(z_{j,k},q)}{mH}\Big)\Big)} \right) \sigma_{P,l,h,b-1}(z,i).$$

Again, from Lemma 1, we have

$$\frac{1}{2} \sum_{(j,k,q)=(1,1,1)}^{(l,h,b-1)} \ln\Big(1 + \frac{\sigma^2_{P,j,k,q-1}(z_{j,k},q)}{mH}\Big) \leq \gamma_{m(l-1)H+m(h-1)+b-1}(k_P, \tilde{\mathcal{Z}}),$$

where $\gamma_t(k_P, \tilde{\mathcal{Z}})$ denotes the maximum information gain about an $f \sim GP_{\tilde{\mathcal{Z}}}(0,k_P)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0,mH)$. Therefore, with probability at least $1-\delta/3$, uniformly over all $z \in \mathcal{Z}$, $1 \leq i \leq m$, $l \geq 1$, $1 \leq h \leq H$ and $1 \leq b \leq m$,

$$\big|\overline{P}_\star(z,i) - \mu_{P,l,h,b-1}(z,i)\big| \leq \Big(B_P + \frac{\sigma_P}{\sqrt{mH}} \sqrt{2\big(\ln(3/\delta) + \gamma_{m(l-1)H+m(h-1)+b-1}(k_P, \tilde{\mathcal{Z}})\big)} \Big) \sigma_{P,l,h,b-1}(z,i), \tag{34}$$

since by our hypothesis $\big\|\overline{P}_\star\big\|_{k_P} \leq B_P$. Now see that $\mu_{P,l,1,0} = \mu_{P,l-1}$ and $\sigma_{P,l,1,0} = \sigma_{P,l-1}$ for every $l \geq 1$. Hence using (34) for $h=1$ and $b=1$, see that, with probability at least $1-\delta/3$, uniformly over all $z \in \mathcal{Z}$, $1 \leq i \leq m$, and $l \geq 1$,

$$\big|\overline{P}_\star(z,i) - \mu_{P,l-1}(z,i)\big| \leq \Big(B_P + \frac{\sigma_P}{\sqrt{mH}} \sqrt{2\big(\ln(3/\delta) + \gamma_{m(l-1)H}(k_P, \tilde{\mathcal{Z}})\big)} \Big) \sigma_{P,l-1}(z,i).$$

Now recall that $\beta_{P,l} = B_P + \frac{\sigma_P}{\sqrt{mH}} \sqrt{2\big(\ln(3/\delta) + \gamma_{m(l-1)H}(k_P, \lambda_P, \tilde{\mathcal{Z}})\big)}$, $\overline{P}_\star(z) = [\overline{P}_\star(z,1), \ldots, \overline{P}_\star(z,m)]^T$, $\mu_{P,l-1}(z) = [\mu_{P,l-1}(z,1), \ldots, \mu_{P,l-1}(z,m)]^T$ and $\sigma_{P,l-1}(z) = [\sigma_{P,l-1}(z,1), \ldots, \sigma_{P,l-1}(z,m)]^T$. Then, with probability at least $1-\delta/3$, uniformly over all $z \in \mathcal{Z}$ and $l \geq 1$,

$$\big\|\overline{P}_\star(z) - \mu_{P,l-1}(z)\big\|_2 \leq \sqrt{\sum_{i=1}^m \beta^2_{P,l} \sigma^2_{P,l-1}(z,i)} = \beta_{P,l} \sqrt{\sum_{i=1}^m \sigma^2_{P,l-1}(z,i)} = \beta_{P,l} \left\|\sigma_{P,l-1}(z)\right\|_2,$$

and hence (30) follows. ■

Lemma 11 (Sum of predictive variances upper bounded by Maximum Information Gain) *Let $\sigma_{R,l}$ and $\sigma_{P,l}$ be defined as in 27 and 28 respectively and let the kernels k_R and k_P satisfy $k_R(z,z) \leq 1$ and $k_P((z,i),(z,i)) \leq 1$ for all $z \in \mathcal{Z}$ and $1 \leq i \leq m$. Then for any $\tau \geq 1$,*

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}(z_{l,h}) \leq \sqrt{2e\tau H^2 \gamma_{\tau H}(k_R, \mathcal{Z})} \tag{35}$$

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}(z_{l,h})\|_2 \leq \sqrt{2em\tau H^2 \gamma_{m\tau H}(k_P, \tilde{\mathcal{Z}})}, \tag{36}$$

where $\gamma_t(k_R, \mathcal{Z})$ denotes the maximum information gain about an $f \sim GP_{\mathcal{Z}}(0, k_R)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0, H)$ and $\gamma_t(k_P, \tilde{\mathcal{Z}})$ denotes the maximum information gain about an $f \sim GP_{\tilde{\mathcal{Z}}}(0, k_P)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0, mH)$.

Proof Note that $\sigma_{R,l-1}(z) = \sigma_{R,l,0}(z)$, where $\sigma_{R,l,0}(z)$ is defined in (31). Now from (15), see that $\sigma^2_{R,l,0}(z) \leq (1 + 1/H)\sigma^2_{R,l,1}(z) \leq (1+1/H)^2\sigma^2_{R,l,2}(z) \leq \cdots \leq (1+1/H)^{H-1}\sigma^2_{R,l,H-1}(z)$, i.e. $\sigma^2_{R,l,0}(z) \leq (1+1/H)^{h-1}\sigma^2_{R,l,h-1}(z)$ for all $z \in \mathcal{Z}$ and $1 \leq h \leq H$. This implies

$$\begin{aligned}
\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l-1}(z_{l,h}) = \sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l,0}(z_{l,h}) &\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}(1+1/H)^{h-1}\sigma^2_{R,l,h-1}(z_{l,h}) \\
&\leq (1+1/H)^{H-1}\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sigma^2_{R,l,h-1}(z_{l,h}) \\
&\leq (1+1/H)^{H-1}(2H+1)\gamma_{\tau H}(k_R, \mathcal{Z}) \\
&\leq 2eH\gamma_{\tau H}(k_R, \mathcal{Z}),
\end{aligned} \tag{37}$$

where the second last inequality follows from (16) and last inequality is due to the fact that $(1+1/\alpha)^{\alpha} \leq e$ and $(1+1/\alpha)^{-1}(2\alpha+1) \leq 2\alpha$ for all $\alpha > 0$. Further by Cauchy-Schwartz inequality

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}(z_{l,h}) \leq \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sigma^2_{R,l-1}(z_{l,h})}. \tag{38}$$

Now (35) follows by combining (37) and (38).

Similarly Note that $\sigma_{P,l-1}(z,i) = \sigma_{P,l,1,0}(z,i)$, where $\sigma_{P,l,1,0}(z,i)$ is defined in (33). Now from (15), see that $\sigma^2_{P,l,1,0}(z,i) \leq (1+1/mH)^{m(h-1)+b-1}\sigma^2_{P,l,h,b-1}(z,i)$ for all $z \in \mathcal{Z}$, $1 \leq i \leq m$, $1 \leq h \leq H$ and $1 \leq b \leq m$. This implies

$$\begin{aligned}
\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l-1}(z_{l,h},b) = \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l,1,0}(z_{l,h},b) &\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m}(1+1/mH)^{m(h-1)+b-1}\sigma^2_{P,l,h,b-1}(z_{l,h},b) \\
&\leq (1+1/mH)^{m(H-1)+m-1}\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m}\sigma^2_{P,l,h,b-1}(z_{l,h},b) \\
&\leq (1+1/mH)^{mH-1}(2mH+1)\gamma_{m\tau H}(k_P, \tilde{\mathcal{Z}}) \\
&\leq 2emH\gamma_{m\tau H}(k_P, \tilde{\mathcal{Z}}),
\end{aligned} \tag{39}$$

where the second last inequality follows from (16) and last inequality is due to the fact that $(1+1/\alpha)^{\alpha} \leq e$ and $(1+1/\alpha)^{-1}(2\alpha+1) \leq 2\alpha$ for all $\alpha > 0$. Further by Cauchy-Schwartz inequality

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}(z_{l,h})\|_2 \leq \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H}\|\sigma_{P,l-1}(z_{l,h})\|_2^2} = \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m}\sigma^2_{P,l-1}(z_{l,h},b)}. \tag{40}$$

Now (36) follows by combining (39) and (40). ■

B.2.1 Frequentist Regret Bound for GP-UCRL in Kernelized MDPs: Proof of Theorem 1

Note that at every episode l, GP-UCRL (Algorithm 1) selects the policy π_l such that

$$V^{M_l}_{\pi_l,1}(s_{l,1}) = \max_{\pi} \max_{M \in \mathcal{M}_l} V^{M}_{\pi,1}(s_{l,1}), \tag{41}$$

where $s_{l,1}$ is the initial state, $\mathcal{M}_l$ is the family of MDPs constructed by GP-UCRL and M_l is the most optimistic realization from $\mathcal{M}_l$. Further see that the mean reward function $R_\star$ of the unknown MDP $M_\star$ lies in the RKHS $\mathcal{H}_{k_R}(\mathcal{Z})$. Thus for all $z \in \mathcal{Z}$,

$$\left|\overline{R}_\star(z)\right| = \left|\langle \overline{R}_\star, k_R(z,\cdot)\rangle_{k_R}\right| \leq \left\|\overline{R}_\star\right\|_{k_R} k_R(z,z) \leq B_R, \tag{42}$$

where the first equality is due to the reproducing property of RKHS, the first inequality is the Cauchy-Schwartz inequality and the final inequality is due to hypothesis that $\left\|\overline{R}_\star\right\|_{k_R} \leq B_R$ and $k_R(z,z) \leq 1$ for all $z \in \mathcal{Z}$. Now, (42), Lemma 7 and Lemma 9 together imply that for any $0 < \delta \leq 1$ and $\tau \geq 1$, with probability at least $1 - \delta/3$,

$$\sum_{l=1}^{\tau} \left(V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \leq \sum_{l=1}^{\tau}\sum_{h=1}^{H} \Big(\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L_{M_l} \left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 \Big) + (LD + 2B_R H)\sqrt{2\tau H \ln(3/\delta)}. \tag{43}$$

Now for each $l \geq 1$, we define the following events:

$$\begin{aligned} E_{R,l} &:= \left\{\forall z \in \mathcal{Z}, \left|\overline{R}_\star(z) - \mu_{R,l-1}(z)\right| \leq \beta_{R,l}\sigma_{R,l-1}(z)\right\}, \\ E_{P,l} &:= \left\{\forall z \in \mathcal{Z}, \left\|\overline{P}_\star(z) - \mu_{P,l-1}(z)\right\|_2 \leq \beta_{P,l} \left\|\sigma_{P,l-1}(z)\right\|_2 \right\}. \end{aligned}$$

By construction of the set of MDPs $\mathcal{M}_l$ in Algorithm 1, it follows that when both the events $E_{R,l}$ and $E_{P,l}$ hold for all $l \geq 1$, the unknown MDP $M_\star$ lies in $\mathcal{M}_l$ for all $l \geq 1$. Thus (41) implies $V^{M_l}_{\pi_l,1}(s_{l,1}) \geq V^{M_\star}_{\pi_\star,1}(s_{l,1})$ for all $l \geq 1$. This in turn implies, for every episode $l \geq 1$,

$$V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}) \leq V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}). \tag{44}$$

Further when $E_{R,l}$ holds for all $l \geq 1$, then

$$\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| \leq \left|\overline{R}_{M_l}(z_{l,h}) - \mu_{R,l-1}(z_{l,h})\right| + \left|\overline{R}_\star(z_{l,h}) - \mu_{R,l-1}(z_{l,h})\right| \leq 2\beta_{R,l}\, \sigma_{R,l-1}(z_{l,h}), \tag{45}$$

since the mean reward function $\overline{R}_{M_l}$ lies in the confidence set $\mathcal{C}_{R,l}$ (26). Similarly when $E_{P,l}$ holds for all $l \geq 1$,

$$\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 \leq \left\|\overline{P}_{M_l}(z_{l,h}) - \mu_{P,l-1}(z_{l,h})\right\|_2 + \left\|\overline{P}_\star(z_{l,h}) - \mu_{P,l-1}(z_{l,h})\right\|_2 \leq 2\beta_{P,l} \left\|\sigma_{P,l-1}(z_{l,h})\right\|_2, \tag{46}$$

since the mean transition function $\overline{P}_{M_l}$ lies in the confidence set $\mathcal{C}_{P,l}$ (26).

Now combining (43), (44), (45) and (46), when both the events $E_{R,l}$ and $E_{P,l}$ hold for all $l \geq 1$, then with probability at least $1 - \delta/3$,

$$\sum_{l=1}^{\tau} \left(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \leq 2\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left(\beta_{R,l}\sigma_{R,l-1}(z_{l,h}) + L_{M_l}\beta_{P,l} \left\|\sigma_{P,l-1}(z_{l,h})\right\|_2\right) + (LD + 2B_R H)\sqrt{2\tau H \ln(3/\delta)}.$$

Now Lemma 10 implies that $\mathbb{P}\left[\forall l \geq 1, E_{R,l}\right] \geq 1 - \delta/3$ and $\mathbb{P}\left[\forall l \geq 1, E_{P,l}\right] \geq 1 - \delta/3$. Hence, by a union bound, for any $\tau \geq 1$, with probability at least $1 - \delta$,

$$\sum_{l=1}^{\tau} \left(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \leq 2\beta_{R,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}(z_{l,h}) + 2L\beta_{P,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left\|\sigma_{P,l-1}(z_{l,h})\right\|_2 + (LD + 2B_R H)\sqrt{2\tau H \ln(3/\delta)}. \tag{47}$$

Here we have used the fact that both $\beta_{R,l}$ and $\beta_{P,l}$ are non-decreasing with the number of episodes l and that $L_{M_l} \leq L$ by construction of $\mathcal{M}_l$ (and since $M_l \in \mathcal{M}_l$). Now from Lemma 11, we have $\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}(z_{l,h}) \leq$

$\sqrt{2e\tau H^2\gamma_{\tau H}(k_R,\mathcal{Z})}$ and $\sum_{l=1}^{\tau}\sum_{h=1}^{H}\|\sigma_{P,l-1}(z_{l,h})\|_2 \leq \sqrt{2em\tau H^2\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})}$. Therefore with probability at least $1-\delta$, the cumulative regret of GP-UCRL after τ episodes, i.e. after $T=\tau H$ timesteps is

$$\begin{aligned} Regret(T) &= \sum_{l=1}^{\tau}\left(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \\ &\leq 2\beta_{R,\tau}\sqrt{2eH\gamma_T(k_R,\mathcal{Z})T} + 2L\beta_{P,\tau}\sqrt{2emH\gamma_{mT}(k_P,\tilde{\mathcal{Z}})T} + (LD+2B_RH)\sqrt{2T\ln(3/\delta)}, \end{aligned}$$

where $\beta_{R,\tau} = B_R + \frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3/\delta)+\gamma_{(\tau-1)H}(k_R,\mathcal{Z})\big)}$ and $\beta_{P,\tau} = B_P + \frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3/\delta)+\gamma_{m(\tau-1)H}(k_P,\tilde{\mathcal{Z}})\big)}$. Now the result follows by defining $\gamma_T(R) := \gamma_T(k_R,\mathcal{Z})$ and $\gamma_{mT}(P) := \gamma_{mT}(k_P,\tilde{\mathcal{Z}})$.

B.3 Bayes Regret of PSRL under RKHS Priors: Proof of Theorem 2

$\Phi \equiv (\Phi_R, \Phi_P)$ is the distribution of the unknown MDP $M_\star = \{\mathcal{S},\mathcal{A},R_\star,P_\star,H\}$, where Φ_R and Φ_P are specified by distributions over real valued functions on $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ respectively with a sub-Gaussian noise model in the sense that

- The reward distribution is $R_\star : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, with mean $\overline{R}_\star \in H_{k_R}(\mathcal{Z})$, $\left\|\overline{R}_\star\right\|_{k_R} \leq B_R$ and additive σ_R-sub-Gaussian noise.
- The transition distribution is $P_\star : \mathcal{S}\times\mathcal{A}\to\mathcal{S}$, with mean $\overline{P}_\star \in H_{\tilde{k}_P}(\tilde{\mathcal{Z}})$, $\left\|\overline{P}_\star\right\|_{k_P} \leq B_P$ and component-wise additive and independent σ_P-sub-Gaussian noise.

At the start of episode l, PSRL samples an MDP M_l from Φ_l, where $\Phi_l \equiv (\Phi_{R,l},\Phi_{P,l})$ is the corresponding posterior distribution conditioned on the history of observations $\mathcal{H}_{l-1} := \{s_{j,k},a_{j,k},r_{j,k}\}_{1\leq j\leq l-1,1\leq k\leq H}$. Therefore, conditioned on $\mathcal{H}_{l-1}$, both $M_\star$ and M_l are identically distributed. Hence for any $\sigma(\mathcal{H}_{l-1})$ measurable function g, $\mathbb{E}\left[g(M_\star)\;\middle|\;\mathcal{H}_{l-1}\right] = \mathbb{E}\left[g(M_l)\;\middle|\;\mathcal{H}_{l-1}\right]$ and hence by the *tower property*,

$$\mathbb{E}\left[g(M_\star)\right] = \mathbb{E}\left[g(M_l)\right]. \tag{48}$$

See that, conditioned on $\mathcal{H}_{l-1}$, the respective optimal policies $\pi_\star$ and π_l of $M_\star$ and M_l are identically distributed. Since $s_{l,1}$ is deterministic, (48) implies that $\mathbb{E}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1})\right] = \mathbb{E}\left[V^{M_l}_{\pi_l,1}(s_{l,1})\right]$. Hence for every episode $l\geq 1$,

$$\begin{aligned} \mathbb{E}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right] &= \mathbb{E}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_l}_{\pi_l,1}(s_{l,1})\right] + \mathbb{E}\left[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right] \\ &= \mathbb{E}\left[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right]. \end{aligned} \tag{49}$$

Now, from Lemma 7, for any $\tau\geq 1$,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\left[V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right]\right] \leq \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left[\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 + \Delta_{l,h}\right]\right], \tag{50}$$

where $z_{l,h} := (s_{l,h},a_{l,h})$ and $\Delta_{l,h} := \mathbb{E}_{s'\sim P_\star(z_{l,h})}\left[V^{M_l}_{\pi_l,h+1}(s') - V^{M_\star}_{\pi_l,h+1}(s')\right] - \left(V^{M_l}_{\pi_l,h+1}(s_{l,h+1}) - V^{M_\star}_{\pi_l,h+1}(s_{l,h+1})\right)$. From (24), see that $\mathbb{E}\left[\Delta_{l,h}\;\middle|\;\mathcal{G}_{l,h-1},M_\star,M_l\right] = 0$, where $\mathcal{G}_{l,h-1} := \mathcal{H}_{l-1}\cup\{s_{l,k},a_{l,k},r_{l,k},s_{l,k+1}\}_{1\leq k\leq h-1}$ denotes the history of all observations till episode l and period $h-1$. Now by tower property $\mathbb{E}\left[\Delta_{l,h}\right]=0, l\geq 1, 1\leq h\leq H$. Hence, combining (49) and (50), for any $\tau\geq 1$,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right]\right] \leq \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left[\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2\right]\right]. \tag{51}$$

Now fix any $0<\delta\leq 1$ and for each $l\geq 1$, define two events $E_\star := \left\{\overline{R}_\star\in\mathcal{C}_{R,l},\overline{P}_\star\in\mathcal{C}_{P,l}\;\forall l\geq 1\right\}$ and $E_M := \left\{\overline{R}_{M_l}\in\mathcal{C}_{R,l},\overline{P}_{M_l}\in\mathcal{C}_{P,l}\;\forall l\geq 1\right\}$, where $\mathcal{C}_{R,l},\mathcal{C}_{P,l}, l\geq 1$ are the confidence sets constructed by GP-UCRL as defined in (26). Now from Lemma 10, $\mathbb{P}\left[E_\star\right]\geq 1-2\delta/3$ and hence by (48) $\mathbb{P}\left[E_M\right]\geq 1-2\delta/3$. Further define $E := E_\star\cap E_M$ and by union bound, see that

$$\mathbb{P}\left[E^c\right] \leq \mathbb{P}\left[E_\star^c\right] + \mathbb{P}\left[E_M^c\right] \leq 4\delta/3. \tag{52}$$

(51) and (52) together imply,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1})-V^{M_\star}_{\pi_l,1}(s_{l,1})\right]\right] \;\leq\; \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left[\left|\overline{R}_{M_l}(z_{l,h})-\overline{R}_\star(z_{l,h})\right| \;\middle|\; E\right]\right. \\ \left.+\mathbb{E}\left[L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h})-\overline{P}_\star(z_{l,h})\right\|_2\right] \;\middle|\; E\right]+8\delta B_R\tau H/3, \tag{53}$$

where we have used that $V^{M_\star}_{\pi_\star,1}(s_{l,1})-V^{M_\star}_{\pi_l,1}(s_{l,1})\leq 2B_RH$, since $\left|\overline{R}_\star(z)\right|\leq B_R$ for all $z\in\mathcal{Z}$. Now from Lemma 10 and construction of $\mathcal{C}_{R,l}, l\geq 1$,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left|\overline{R}_{M_l}(z_{l,h})-\overline{R}_\star(z_{l,h})\right| \;\middle|\; E\right] \;\leq\; \sum_{l=1}^{\tau}\sum_{h=1}^{H}2\beta_{R,l}\sigma_{R,l-1}(z_{l,h}) \\ \leq\; 2\beta_{R,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sigma_{R,l-1}(z_{l,h}) \tag{54} \\ \leq\; 2\beta_{R,\tau}\sqrt{2e\tau H^2\gamma_{\tau H}(k_R,\mathcal{Z})}, \tag{55}$$

where the last step follows from Lemma 11. Now form (48), $\mathbb{E}\left[L_{M_l}\right]=\mathbb{E}\left[L_\star\right]$ and therefore $\mathbb{E}\left[L_{M_l}\;\middle|\;E\right]\leq \mathbb{E}\left[L_{M_l}\right]/\mathbb{P}\left[E\right]\leq \mathbb{E}\left[L_\star\right]/(1-4\delta/3)$. Similarly from Lemma 10 and construction of $\mathcal{C}_{P,l}, l\geq 1$,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}L_{M_l}\left\|\overline{P}_{M_l}(z_{l,h})-\overline{P}_\star(z_{l,h})\right\|_2 \;\middle|\; E\right] \;\leq\; \sum_{l=1}^{\tau}\sum_{h=1}^{H}\mathbb{E}\left[L_{M_l}\;\middle|\;E\right]2\beta_{P,l}\left\|\sigma_{P,l-1}(z_{l,h})\right\|_2 \\ \leq\; \frac{\mathbb{E}\left[L_\star\right]}{1-4\delta/3}2\beta_{P,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H}\left\|\sigma_{P,l-1}(z_{l,h})\right\|_2 \\ \leq\; \frac{\mathbb{E}\left[L_\star\right]}{1-4\delta/3}2\beta_{P,\tau}\sqrt{2em\tau H^2\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})},, \tag{56}$$

where the last step follows from Lemma 11. Combining (53), (55) and (56), for any $0<\delta\leq 1$ and $\tau\geq 1$,

$$\mathbb{E}\left[\sum_{l=1}^{\tau}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1})-V^{M_\star}_{\pi_l,1}(s_{l,1})\right]\right]\leq 2\beta_{R,\tau}\sqrt{2e\tau H^2\gamma_{\tau H}(k_R,\mathcal{Z})}+\frac{\mathbb{E}\left[L_\star\right]}{1-4\delta/3}2\beta_{P,\tau}\sqrt{2em\tau H^2\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})} \\ +8\delta B_R\tau H/3,$$

where $\beta_{R,\tau}=B_R+\frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3/\delta)+\gamma_{(\tau-1)H}(k_R,\mathcal{Z})\big)}$ and $\beta_{P,\tau}=B_P+\frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3/\delta)+\gamma_{m(\tau-1)H}(k_P,\tilde{\mathcal{Z}})\big)}$. See that the left hand side of the above is independent of δ. Now using $\delta=1/\tau H$, the Bayes regret of PSRL after τ episodes, i.e. after $T=\tau H$ timesteps is

$$\mathbb{E}\left[Regret(T)\right] \;=\; \sum_{l=1}^{\tau}\mathbb{E}\left[V^{M_\star}_{\pi_\star,1}(s_{l,1})-V^{M_\star}_{\pi_l,1}(s_{l,1})\right] \\ \leq\; 2\alpha_{R,\tau}\sqrt{2eH\gamma_T(k_R,\mathcal{Z})T}+3\mathbb{E}\left[L_\star\right]\alpha_{P,\tau}\sqrt{2emH\gamma_{mT}(k_P,\tilde{\mathcal{Z}})T}+3B_R,$$

since $1/(1-4/3\tau H)\leq 3/2$ as $\tau\geq 2$, $H\geq 2$. Here $\alpha_{R,\tau}:=B_R+\frac{\sigma_R}{\sqrt{H}}\sqrt{2\big(\ln(3T)+\gamma_{(\tau-1)H}(k_R,\mathcal{Z})\big)}$, $\alpha_{P,\tau}=B_P+\frac{\sigma_P}{\sqrt{mH}}\sqrt{2\big(\ln(3T)+\gamma_{m(\tau-1)H}(k_P,\tilde{\mathcal{Z}})\big)}$. Now the result follows by defining $\gamma_T(R):=\gamma_T(k_R,\mathcal{Z})$ and $\gamma_{mT}(P):=\gamma_{mT}(k_P,\tilde{\mathcal{Z}})$.

C BAYES REGRET UNDER GAUSSIAN PROCESS PRIORS

In this section, we develop the Bayesian RL analogue of Gaussian process bandits, i.e., learning under the assumption that MDP dynamics and reward behavior are sampled according to Gaussian process priors.

Regularity and Noise assumptions Each of our results in this section will assume that the mean reward function $\overline{R}_\star$ and the mean transition function $\overline{P}_\star$ are randomly sampled from from Gaussian processes $GP_{\mathcal{Z}}(0, k_R)$ and $GP_{\tilde{\mathcal{Z}}}(0, k_P)$, respectively, where $\mathcal{Z} := \mathcal{S}\times\mathcal{A}$ and $\tilde{\mathcal{Z}} := \mathcal{Z}\times\{1, \ldots, m\}$. Further, we will assume the noise sequences $\{\varepsilon_{R,l,h}\}_{l\geq 1, 1\leq h\leq H}$ are iid Gaussian $\mathcal{N}(0, \lambda_R)$ and $\{\varepsilon_{P,l,h}\}_{l\geq 1, 1\leq h\leq H}$ are iid Gaussian $\mathcal{N}(0, \lambda_P I)$. Note that the same GP priors and noise models were used to design our algorithms (see Section 3.1). Thus, in this case the algorithm is assumed to have exact knowledge of the data generating process (the 'fully Bayesian' setup).

Further, in order to achieve non-trivial regret for continuous state/action MDPs, we need the following smoothness assumptions similar to those made by Srinivas et al. [2009] on the kernels. We assume that $\mathcal{S} \subseteq [0, c_1]^m$ and $\mathcal{A} \subseteq [0, c_2]^n$ are compact and convex, and that the kernels k_R and k_P satisfy[10] the following high probability bounds on the derivatives of GP sample paths $\overline{R}_\star$ and $\overline{P}_\star$, respectively:

$$\mathbb{P}\left[\sup_{z\in\mathcal{Z}} \left|\partial \overline{R}_\star(z)/\partial z_j\right| > L_R\right] \leq a_R e^{-(L_R/b_R)^2} \tag{57}$$

holds for all $1 \leq j \leq m+n$ and for any $L_R > 0$ corresponding to some $a_R, b_R > 0$, and

$$\mathbb{P}\left[\sup_{z\in\mathcal{Z}} \left|\partial \overline{P}_\star(z,i)/\partial z_j\right| > L_P\right] \leq a_P e^{-(L_P/b_P)^2} \tag{58}$$

holds for al $1 \leq j \leq m+n$, $1 \leq i \leq m$ and for any $L_P > 0$ corresponding to some $a_P, b_P > 0$. Also we assume that

$$\mathbb{P}\left[\sup_{z\in\mathcal{Z}} \left|\overline{R}_\star(z)\right| > L\right] < a e^{-(L/b)^2}, \tag{59}$$

holds for any $L \geq 0$ for some corresponding $a, b > 0$[11].

Choice of confidence sets for GP-UCRL For any fixed $0 < \delta \leq 1$, at the beginning of each episode l, GP-UCRL construct the confidence set $\mathcal{C}_{R,l}$ as

$$\mathcal{C}_{R,l} = \left\{f : |f(z) - \mu_{R,l-1}([z]_l)| \leq \beta_{R,l}\sigma_{R,l-1}([z]_l) + 1/l^2, \forall z\right\}, \tag{60}$$

where $\mu_{R,l-1}(z)$, $\sigma_{R,l-1}(z)$ are defined as in (2) and $\beta_{R,l} := \sqrt{2\ln\left(|\mathcal{S}_l|\,|\mathcal{A}_l|\,\pi^2 l^2/\delta\right)}$. Here $(\mathcal{S}_l)_{l\geq 1}$ and $(\mathcal{A}_l)_{l\geq 1}$ are suitable discretizations of state space $\mathcal{S}$ and action space $\mathcal{A}$ respectively, $[z]_l := ([s]_l, [a]_l)$, where $[s]_l$ is the closest point in $\mathcal{S}_l$ to s and $[a]_l$ is the closest point in $\mathcal{A}_l$ to a. Also $|\mathcal{S}_l| = \max\left\{\left(2c_1 m l^2 b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}\right)^m, \left(2c_1 m l^2 b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}\right)^m\right\}$ and $|\mathcal{A}_l| = \max\left\{\left(2c_2 n l^2 b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}\right)^n, \left(2c_2 n l^2 b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}\right)^n\right\}$.

Similarly GP-UCRL construct the confidence set $\mathcal{C}_{P,l}$ as

$$\mathcal{C}_{P,l} = \left\{f : \|f(z) - \mu_{P,l-1}([z]_l)\|_2 \leq \beta_{P,l}\,\|\sigma_{P,l-1}([z]_l)\|_2 + \frac{\sqrt{m}}{l^2}, \forall z\right\}, \tag{61}$$

where $\beta_{P,l} := \sqrt{2\ln\left(|\mathcal{S}_l|\,|\mathcal{A}_l|\,m\pi^2 l^2/\delta\right)}$, $\mu_{P,l}(z) := [\mu_{P,l-1}(z,1), \ldots, \mu_{P,l-1}(z,m)]^T$ and $\sigma_{P,l}(z) := [\sigma_{P,l-1}(z,1), \ldots, \sigma_{P,l-1}(z,m)]^T$ with $\mu_{P,l-1}(z,i)$ and $\sigma_{P,l-1}(z,i)$ be defined as in (3)

Theorem 3 (Bayesian regret bound for GP-UCRL under GP prior) *Let $M_\star = \{\mathcal{S}, \mathcal{A}, R_\star, P_\star, H\}$ be an MDP with period H, state space $\mathcal{S} \subseteq [0, c_1]^m$ and action space $\mathcal{A} \subseteq [0, c_2]^n$, $m, n \in \mathbb{N}, c_1, c_2 > 0$. Let $\mathcal{S}$ and $\mathcal{A}$ be compact and convex. Let the mean reward function $\overline{R}_\star$ be a sample from $GP_{\mathcal{Z}}(0, k_R)$, where $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ and let the noise*

[10] This assumption holds for stationary kernels $k(z, z') \equiv k(z - z')$ that are four times differentiable, such as SE and Matérn kernels with $\nu \geq 2$.

[11] This is a mild assumption on the kernel k_R, since $\overline{R}_\star(z)$ is Gaussian and thus has exponential tails.

variables $\varepsilon_{R,l,h}$ be iid Gaussian $\mathcal{N}(0,\lambda_R)$. Let the mean transition function $\overline{P}_\star$ be a sample from $GP_{\tilde{\mathcal{Z}}}(0,k_P)$, where $\tilde{\mathcal{Z}} := \mathcal{Z}\times\{1,\dots,m\}$ and let the noise variables $\varepsilon_{P,l,h}$ be iid Gaussian $\mathcal{N}(0,\lambda_P I)$. Further let the kernel k_R satisfy (57), (59) and the kernel k_P satisfy (58). Also let $k_R(z,z)\le 1$, $k_P((z,i),(z,i))\le 1$ for all $z\in\mathcal{Z}$ and $1\le i\le m$. Then for any $0\le\delta\le 1$, GP-UCRL, with confidence sets (60) and (61), enjoys, with probability at least $1-\delta$, the regret bound

$$Regret(T) \le 2\beta_{R,\tau}\exp\left(\gamma_{H-1}(R)\right)\sqrt{(2\lambda_R+1)\gamma_T(R)T} + 2L\beta_{P,\tau}\exp\left(\gamma_{mH-1}(P)\right)\sqrt{(2\lambda_P+1)\gamma_{mT}(P)T} + (L\sqrt{m}+1)H\pi^2/3 + (LD+2CH)\sqrt{2T\ln(6/\delta)},$$

where $C := b\sqrt{\ln(6a/\delta)}$, $\beta_{R,l} := \sqrt{2\ln\left(|\mathcal{S}_l|\,|\mathcal{A}_l|\,\pi^2 l^2/\delta\right)}$ and $\beta_{P,l} := \sqrt{2\ln\left(|\mathcal{S}_l|\,|\mathcal{A}_l|\,m\pi^2 l^2/\delta\right)}$.

Theorem 4 (Bayes regret of PSRL under GP prior) *Let $M_\star$ be an MDP as in Theorem 3 and Φ be a (known) prior distibution over MDPs $M_\star$. Then the Bayes regret of PSRL (Algorithm 2) satisfies*

$$\mathbb{E}\left[Regret(T)\right] \le 2\alpha_{R,\tau}\exp\left(\gamma_{H-1}(R)\right)\sqrt{(2\lambda_R+1)\gamma_T(R)T} + 3\,\mathbb{E}\left[L_\star\right]\alpha_{P,\tau}\exp\left(\gamma_{mH-1}(P)\right)\sqrt{(2\lambda_P+1)\gamma_{mT}(P)T} + 3C + (1+\sqrt{m}\mathbb{E}\left[L_\star\right])\pi^2 H,$$

where $C = \mathbb{E}\left[\sup_{z\in\mathcal{Z}}\left|\overline{R}_\star(z)\right|\right]$, $\alpha_{R,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,\pi^2\tau^2 T\right)}$ and $\alpha_{P,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,m\pi^2\tau^2 T\right)}$.

C.1 Detail Analysis

Here the state space $\mathcal{S}\subseteq[0,c_1]^m$ and the action space $\mathcal{A}\subseteq[0,c_2]^n$ for $c_1,c_2>0$. Both $\mathcal{S}$ and $\mathcal{A}$ are assumed to be compact and convex. Then at every round l, we can construct (by Lemma 15 of Desautels et al. [2014]) two discretization sets $\mathcal{S}_l$ and $\mathcal{A}_l$ of $\mathcal{S}$ and $\mathcal{A}$ respectively, with respective sizes $\mathcal{S}_l$ and $\mathcal{A}_l$, such that for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$, the following holds:

$$\begin{aligned}\|s-[s]_l\|_1 &\le c_1 m/|\mathcal{S}_l|^{1/m},\\ \|a-[a]_l\|_1 &\le c_2 n/|\mathcal{A}_l|^{1/n},\end{aligned}$$

where $[s]_l := \operatorname{argmin}_{s'\in\mathcal{S}_l}\|s-s'\|_1$, is the closest point in S_l to s and $[a]_l := \operatorname{argmin}_{a'\in\mathcal{A}_l}\|a-a'\|_1$ is the closest point in A_l to a (in the sense of 1-norm). Now for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$, we define $z := [s^T,a^T]^T$ and correspondingly $[z]_l := \left[[s]_l^T,[a]_l^T\right]^T$. Further define $\mathcal{Z} := \mathcal{S}\times\mathcal{A} := \{z=[s^T,a^T]^T : s\in\mathcal{S},a\in\mathcal{A}\}$ and $\mathcal{Z}_l := \mathcal{S}_l\times\mathcal{A}_l := \{z=[s^T,a^T]^T : s\in\mathcal{S}_l,a\in\mathcal{A}_l\}$. See that $z,[z]_l\in\mathbb{R}^{m+n}$ and $\mathcal{Z},\mathcal{Z}_l\subset\mathbb{R}^{m+n}$.

Lemma 12 (Samples from GPs are Lipschitz) *Let $\mathcal{S}\subseteq[0,c_1]^m$ and $\mathcal{A}\subseteq[0,c_2]^n$ be compact and convex, $m,n\in\mathbb{N},c_1,c_2>0$. Let $\overline{R}_\star$ be a sample from $GP_{\mathcal{Z}}(0,k_R)$, where $\mathcal{Z}:=\mathcal{S}\times\mathcal{A}$, $\overline{P}_\star$ be a sample from $GP_{\tilde{\mathcal{Z}}}(0,k_P)$, where $\tilde{\mathcal{Z}}:=\mathcal{Z}\times\{1,\dots,m\}$. Further let the kernels k_R and k_P satisfy (57) and (58) respectively. Then, for any $0<\delta\le 1$, the following holds:*

$$\mathbb{P}\left[\forall z\in\mathcal{Z},\forall l\ge 1\left|\overline{R}_\star(z)-\overline{R}_\star([z]_l)\right|\le 1/l^2\right]\ge 1-\delta/6, \tag{62}$$

$$\mathbb{P}\left[\forall z\in\mathcal{Z},\forall 1\le i\le m,\forall l\ge 1\left|\overline{P}_\star(z,i)-\overline{P}_\star([z]_l,i)\right|\le 1/l^2\right]\ge 1-\delta/6. \tag{63}$$

Proof From (57), recall the assumption on kernel k_R:

$$\mathbb{P}\left[\sup_{z\in\mathcal{Z}}\left|\partial\overline{R}_\star(z)/\partial z_j\right|>L_R\right]\le a_R e^{-(L_R/b_R)^2}, 1\le j\le m+n,$$

holds for any $L_R>0$ for some corresponding $a_R,b_R>0$. Now using union bound,

$$\mathbb{P}\left[\forall 1\le j\le m+n\ \sup_{z\in\mathcal{Z}}\left|\partial\overline{R}_\star(z)/\partial z_j\right|\le L_R\right]\ge 1-(m+n)a_R e^{-(L_R/b_R)^2}.$$

From Mean-Value Theorem, this implies that with probability at least $1-(m+n)a_R e^{-(L_R/b_R)^2}$,

$$\forall z,z'\in\mathcal{Z},\left|\overline{R}_\star(z)-\overline{R}_\star(z')\right|\le L_R\|z-z'\|_1.$$

Therefore, with probability at least $1-(m+n)a_R e^{-(L_R/b_R)^2}$,

$$\begin{aligned}\forall l\geq 1, \forall z\in\mathcal{Z}, \left|\overline{R}_\star(z)-\overline{R}_\star([z]_l)\right| &\leq L_R\left\|z-[z]_l\right\|_1\\ &= L_R\big(\left\|s-[s]_l\right\|_1+\left\|a-[a]_l\right\|_1\big)\\ &\leq L_R\big(c_1 m/\left|\mathcal{S}_l\right|^{1/m}+c_2 n/\left|\mathcal{A}_l\right|^{1/n}\big).\end{aligned}$$

Now for any $0<\delta\leq 1$ choose $L_R=b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}$. Then with probability at least $1-\delta/6$,

$$\forall l\geq 1, \forall z\in\mathcal{Z}, \left|\overline{R}_\star(z)-\overline{R}_\star([z]_l)\right|\leq b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}\Big(c_1 m/\left|\mathcal{S}_l\right|^{1/m}+c_2 n/\left|\mathcal{A}_l\right|^{1/n}\Big). \tag{64}$$

Similarly from (58), recall the assumption on kernel k_P:

$$\mathbb{P}\left[\sup_{z\in\mathcal{Z}}\left|\partial\overline{P}_\star(z,i)/\partial z_j\right|>L_P\right]\leq a_P e^{-(L_P/b_P)^2}, 1\leq j\leq m+n, 1\leq i\leq m,$$

holds for any $L_P>0$ for some corresponding $a_P, b_P>0$. Now using union bound,

$$\mathbb{P}\left[\forall 1\leq j\leq m+n, \forall 1\leq i\leq m \ \sup_{z\in\mathcal{Z}}\left|\partial\overline{P}_\star(z,i)/\partial z_j\right|\leq L_P\right]\geq 1-m(m+n)a_P e^{-(L_P/b_P)^2}.$$

Now from Mean-Value Theorem, this implies that with probability at least $1-m(m+n)a_P e^{-(L_P/b_P)^2}$,

$$\forall z,z'\in\mathcal{Z}, \forall 1\leq i\leq m, \left|\overline{P}_\star(z,i)-\overline{P}_\star(z',i)\right|\leq L_P\left\|z-z'\right\|_1.$$

Therefore, with probability at least $1-m(m+n)a_P e^{-(L_P/b_P)^2}$, we have

$$\begin{aligned}\forall l\geq 1, \forall z\in\mathcal{Z}, \forall 1\leq i\leq m, \left|\overline{P}_\star(z,i)-\overline{P}_\star([z]_l,i)\right| &\leq L_P\left\|z-[z]_l\right\|_1\\ &= L_P\big(\left\|s-[s]_l\right\|_1+\left\|a-[a]_l\right\|_1\big)\\ &\leq L_P\big(c_1 m/\left|\mathcal{S}_l\right|^{1/m}+c_2 n/\left|\mathcal{A}_l\right|^{1/n}\big).\end{aligned}$$

Now choose $L_P=b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}$. Then with probability at least $1-\delta/6$,

$$\forall l\geq 1, \forall z\in\mathcal{Z}, \forall 1\leq i\leq m, \left|\overline{P}_\star(z,i)-\overline{P}_\star([z]_l,i)\right|\leq b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}\big(c_1 m/\left|\mathcal{S}_l\right|^{1/m}+c_2 n/\left|\mathcal{A}_l\right|^{1/n}\big). \tag{65}$$

Now by using $|\mathcal{S}_l|=\max\Big\{\Big(2c_1 ml^2 b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}\Big)^m, \Big(2c_1 ml^2 b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}\Big)^m\Big\}$ and $|\mathcal{A}_l|=\max\Big\{\Big(2c_2 nl^2 b_R\sqrt{ln\big(6(m+n)a_R/\delta\big)}\Big)^n, \Big(2c_2 nl^2 b_P\sqrt{ln\big(6m(m+n)a_P/\delta\big)}\Big)^n\Big\}$ in (64) and (65), we get 62 and 63 respectively. ■

Now recall that at each episode $l\geq 1$, GP-UCRL constructs confidence sets $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ as

$$\begin{aligned}\mathcal{C}_{R,l}&=\big\{f:\mathcal{Z}\to\mathbb{R}\ \big|\ \forall z\in\mathcal{Z}, |f(z)-\mu_{R,l-1}([z]_l)|\leq\beta_{R,l}\sigma_{R,l-1}([z]_l)+1/l^2\big\},\\ \mathcal{C}_{P,l}&=\big\{f:\mathcal{Z}\to\mathbb{R}^m\ \big|\ \forall z\in\mathcal{Z}, \left\|f(z)-\mu_{P,l-1}([z]_l)\right\|_2\leq\beta_{P,l}\left\|\sigma_{P,l-1}([z]_l)\right\|_2+\sqrt{m}/l^2\big\},\end{aligned} \tag{66}$$

where $\mu_{R,0}(z)=0$, $\sigma^2_{R,0}(z)=k_R(z,z)$ and for each $l\geq 1$,

$$\begin{aligned}\mu_{R,l}(z)&=k_{R,l}(z)^T(K_{R,l}+\lambda_R I)^{-1}R_l,\\ \sigma^2_{R,l}(z)&=k_R(z,z)-k_{R,l}(z)^T(K_{R,l}+\lambda_R I)^{-1}k_{R,l}(z).\end{aligned} \tag{67}$$

Here I is the $(lH)\times(lH)$ identity matrix, $R_l=[r_{1,1},\ldots,r_{l,H}]^T$ is the vector of rewards observed at $\mathcal{Z}_l=\{z_{j,k}\}_{1\leq j\leq l, 1\leq k\leq H}=\{z_{1,1},\ldots,z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{R,l}(z)=[k_R(z_{1,1},z),\ldots,k_R(z_{l,H},z)]^T$ is the vector kernel evaluations between z and elements of the set $\mathcal{Z}_l$ and $K_{R,l}=$

$[k_R(u,v)]_{u,v\in\mathcal{Z}_l}$ is the kernel matrix computed at $\mathcal{Z}_l$. Further $\mu_{P,l}(z) = [\mu_{P,l-1}(z,1),\ldots,\mu_{P,l-1}(z,m)]^T$ and $\sigma_{P,l}(z) = [\sigma_{P,l-1}(z,1),\ldots,\sigma_{P,l-1}(z,m)]^T$, where $\mu_{P,0}(z,i)=0$, $\sigma_{P,0}(z,i) = k_P\big((z,i),(z,i)\big)$ and for each $l \geq 1$,

$$\begin{aligned}\mu_{P,l}(z,i) &= k_{P,l}(z,i)^T(K_{P,l}+\lambda_P I)^{-1}S_l,\\ \sigma^2_{P,l}\big((z,i)\big)\big) &= k_P((z,i),(z,i)) - k_{P,l}(z,i)^T(K_{P,l}+\lambda_P I)^{-1}k_{P,l}(z,i).\end{aligned} \tag{68}$$

Here I is the $(mlH)\times(mlH)$ identity matrix, $S_l = [s_{1,2}^T,\ldots,s_{l,H+1}^T]^T$ denotes the vector of state transitions at $\mathcal{Z}_l = \{z_{1,1},\ldots,z_{l,H}\}$, the set of all state-action pairs available at the end of episode l. $k_{P,l}(z,i) = \Big[k_P\big((z_{1,1},1),(z,i)\big),\ldots,k_P\big((z_{l,H},m),(z,i)\big)\Big]^T$ is the vector of kernel evaluations between (z,i) and elements of the set $\tilde{\mathcal{Z}}_l = \big\{(z_{j,k},i)\big\}_{1\leq j\leq l,1\leq k\leq H,1\leq i\leq m} = \big\{(z_{1,1},1),\ldots,(z_{l,H},m)\big\}$ and $K_{P,l} = \big[k_P(u,v)\big]_{u,v\in\tilde{\mathcal{Z}}_l}$ is the kernel matrix computed at $\tilde{\mathcal{Z}}_l$. Here for any $0<\delta\leq 1$, $\beta_{R,l} := \sqrt{2\ln\big(|\mathcal{S}_l|\,|\mathcal{A}_l|\,\pi^2 l^2/\delta\big)}$ and $\beta_{P,l} := \sqrt{2\ln\big(|\mathcal{S}_l|\,|\mathcal{A}_l|\,m\pi^2 l^2/\delta\big)}$ are properly chosen confidence parameters of $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ respectively, where both $|\mathcal{S}_l|$ and $|\mathcal{A}_l|$ are approximately $O\Big(\big(l^2\ln(1/\delta)\big)^d\Big)$ with $d=\max\{m,n\}$.

Lemma 13 (Posterior Concentration of Gaussian Processes) *Let $M_\star = \{\mathcal{S},\mathcal{A},R_\star,P_\star,H\}$ be an MDP with period H, state space $\mathcal{S}\subseteq[0,c_1]^m$ and action space $\mathcal{A}\subseteq[0,c_2]^n$, $m,n\in\mathbb{N}, c_1,c_2>0$. Let $\mathcal{S}$ and $\mathcal{A}$ be compact and convex. Let the mean reward function $\overline{R}_\star$ be a sample from $GP_{\mathcal{Z}}(0,k_R)$, where $\mathcal{Z}:=\mathcal{S}\times\mathcal{A}$ and let the noise variables $\varepsilon_{R,l,h}$ be iid Gaussian $\mathcal{N}(0,\lambda_R)$. Let the mean transition function $\overline{P}_\star$ be a sample from $GP_{\tilde{\mathcal{Z}}}(0,k_P)$, where $\tilde{\mathcal{Z}}:=\mathcal{Z}\times\{1,\ldots,m\}$ and let the noise variables $\varepsilon_{P,l,h}$ be iid Gaussian $\mathcal{N}(0,\lambda_P I)$. Further let the kernels k_R and k_P satisfy (57) and (58) respectively. Then, for any $0<\delta\leq 1$, the following holds:*

$$\mathbb{P}\left[\forall z\in\mathcal{Z},\forall l\geq 1,\left|\overline{R}_\star(z)-\mu_{R,l-1}([z]_l)\right|\leq\beta_{R,l}\,\sigma_{R,l-1}([z]_l)+1/l^2\right] \;\geq\; 1-\delta/3, \tag{69}$$

$$\mathbb{P}\left[\forall z\in\mathcal{Z},\forall l\geq 1,\left\|\overline{P}_\star(z)-\mu_{P,l-1}([z]_l)\right\|_2\leq\beta_{P,l}\left\|\sigma_{P,l-1}([z]_l)\right\|_2+\sqrt{m}/l^2\right] \;\geq\; 1-\delta/3, \tag{70}$$

Proof Note that conditioned on $\mathcal{H}_{l-1} := \{s_{j,k},a_{j,k},r_{j,k},s_{j,k+1}\}_{1\leq j\leq l-1,1\leq k\leq H}$, $\overline{R}_\star(z)\sim\mathcal{N}\big(\mu_{R,l-1}(z),\sigma^2_{R,l-1}(z)\big)$. If $a\sim\mathcal{N}(0,1)$, $c\geq 0$, then $\mathbb{P}\left[|a|\geq c\right]\leq\exp(-c^2/2)$. Using this Gaussian concentration inequality and a union bound over all $l\geq 1$ and all $z\in\mathcal{Z}_l$, with probability at least $1-\delta/6$, we have

$$\forall l\geq 1,\forall z\in\mathcal{Z}_l,\left|\overline{R}_\star(z)-\mu_{R,l-1}(z)\right|\leq\beta_{R,l}\sigma_{R,l-1}(z). \tag{71}$$

Now as $[z]_l\in\mathcal{Z}_l$, using union bound in (62) and (71), we have with probability at least $1-\delta/3$,

$$\forall l\geq 1,\forall z\in\mathcal{Z},\left|\overline{R}_\star(z)-\mu_{R,l-1}([z]_l)\right|\leq\beta_{R,l}\sigma_{R,l-1}([z]_l)+1/l^2.$$

Similarly, conditioned on $\mathcal{H}_{l-1}$, $\overline{P}_\star(z,i)\sim\mathcal{N}\big(\mu_{R,l-1}(z,i),\sigma^2_{P,l-1}(z,i)\big)$ for all $z\in\mathcal{Z}$ and $1\leq i\leq m$. Then using the Gaussian concentration inequality and a union bound over all $l\geq 1$, all $z\in\mathcal{Z}_l$ and all $i=1,\ldots,m$, with probability at least $1-\delta/6$, we have

$$\forall l\geq 1,\forall z\in\mathcal{Z}_l,\forall 1\leq i\leq m,\left|\overline{P}_\star(z,i)-\mu_{P,l-1}(z,i)\right|\leq\beta_{P,l}\sigma_{P,l-1}(z,i). \tag{72}$$

Now as $[z]_l\in\mathcal{Z}_l$, using union bound with (63) and (72), we have with probability at least $1-\delta/3$,

$$\forall l\geq 1,\forall z\in\mathcal{Z},\forall 1\leq i\leq m,\left|\overline{P}_\star(z,i)-\mu_{P,l-1}([z]_l,i)\right|\leq\beta_{P,l}\sigma_{P,l-1}([z]_l,i)+1/l^2.$$

Now Recall that $\overline{P}_\star(z) = [\overline{P}_\star(z,1),\ldots,\overline{P}_\star(z,m)]^T$, $\mu_{P,l-1}(z) = [\mu_{P,l}(z,1),\ldots,\mu_{P,l-1}(z,m)]^T$ and $\sigma_{P,l-1}(z) = [\sigma_{P,l}(z,1),\ldots,\tilde{\sigma}_{P,l-1}(z,m)]^T$. Then with probability at least $1-\delta/3$, for all $l\geq 1$ and for all $z\in\mathcal{Z}$,

$$\begin{aligned}\left\|\overline{P}_\star(z)-\mu_{P,l-1}([z]_l)\right\|_2 \leq \sqrt{\sum_{i=1}^m\Big(\beta_{P,l}\sigma_{P,l-1}([z]_l,i)+\frac{1}{l^2}\Big)^2} &\leq \sqrt{\sum_{i=1}^m\beta^2_{P,l}\sigma^2_{P,l-1}([z]_l,i)}+\sqrt{\sum_{i=1}^m\frac{1}{l^4}}\\ &= \beta_{P,l}\left\|\sigma_{P,l-1}([z]_l)\right\|_2+\frac{\sqrt{m}}{l^2}.\end{aligned}$$

■

Lemma 14 (Sum of predictive variances upper bounded by Maximum Information Gain) *Let $\sigma_{R,l}$ and $\sigma_{P,l}$ be defined as in 67 and 68 respectively and let the kernels k_R and k_P satisfy $k_R(z,z) \leq 1$ and $k_P((z,i),(z,i)) \leq 1$ for all $z \in \mathcal{Z}$ and $1 \leq i \leq m$. Then, for any $\tau \geq 1$,*

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}([z_{l,h}]_l) \leq \exp\big(\gamma_{H-1}(k_R,\mathcal{Z})\big)\sqrt{(2\lambda_R+1)\tau H \gamma_{\tau H}(k_R,\mathcal{Z})} \tag{73}$$

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}([z_{l,h}]_l)\|_2 \leq \exp\big(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\big)\sqrt{(2\lambda_P+1)\tau H \gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})}, \tag{74}$$

where $\gamma_t(k_R,\mathcal{Z})$ denotes the maximum information gain about an $f \sim GP_{\mathcal{Z}}(0,k_R)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0,\lambda_R)$ and $\gamma_t(k_P,\tilde{\mathcal{Z}})$ denotes the maximum information gain about an $f \sim GP_{\tilde{\mathcal{Z}}}(0,k_P)$ after t noisy observations with iid Gaussian noise $\mathcal{N}(0,\lambda_P)$.

Proof Note that $\sigma_{R,l-1}(z) = \sigma_{R,l,0}(z)$, where $\sigma_{R,l,0}(z)$ is defined in (31). From Lemma 3, $\frac{\sigma_{R,l,0}(z)}{\sigma_{R,l,h-1}(z)} \leq \exp\big(\gamma_{h-1}(k_R,\mathcal{Z})\big)$ for all $z \in \mathcal{Z}$ and $1 \leq h \leq H$. This implies

$$\begin{aligned}
\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l-1}([z_{l,h}]_l) = \sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l,0}([z_{l,h}]_l) &\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H} \exp\big(2\gamma_{h-1}(k_R,\mathcal{X})\big)\sigma^2_{R,l,h-1}([z_{l,h}]_l) \\
&\leq \exp\big(2\gamma_{H-1}(k_R,\mathcal{Z})\big)\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l,h-1}([z_{l,h}]_l) \\
&\leq \exp\big(2\gamma_{H-1}(k_R,\mathcal{Z})\big)(2\lambda_R+1)\gamma_{\tau H}(k_R,\mathcal{Z}),
\end{aligned} \tag{75}$$

where the second last inequality follows from (16). Further by Cauchy-Schwartz inequality

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}([z_{l,h}]_l) \leq \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma^2_{R,l-1}([z_{l,h}]_l)}. \tag{76}$$

Now (73) follows by combining (75) and (76).

Similarly Note that $\sigma_{P,l-1}(z,i) = \sigma_{P,l,1,0}(z,i)$, where $\sigma_{P,l,1,0}(z,i)$ is defined in (33). Now from Lemma 3, we have $\frac{\sigma_{P,l,1,0}(z,i)}{\sigma_{P,l,h,b-1}(z,i)} \leq \exp\big(\gamma_{m(h-1)+b-1}(k_P,\lambda_P,\tilde{\mathcal{Z}})\big)$ for all $z \in \mathcal{Z}$, $1 \leq i \leq m$, $1 \leq h \leq H$ and $1 \leq b \leq m$. This implies

$$\begin{aligned}
\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l-1}([z_{l,h}]_l,b) &= \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l,1,0}([z_{l,h}]_l,b) \\
&\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \exp\big(2\gamma_{m(h-1)+b-1}(k_P,\tilde{\mathcal{Z}})\big)\sigma^2_{P,l,h,b-1}([z_{l,h}]_l,b) \\
&\leq \exp\big(2\gamma_{m(H-1)+m-1}(k_P,\tilde{\mathcal{Z}})\big)\sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l,h,b-1}([z_{l,h}]_l,b) \\
&\leq \exp\big(2\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\big)(2\lambda_P+1)\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}}),
\end{aligned} \tag{77}$$

where the second last inequality follows from (16). Further by the Cauchy-Schwartz inequality

$$\sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}([z_{l,h}]_l)\|_2 \leq \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}([z_{l,h}]_l)\|_2^2} = \sqrt{\tau H \sum_{l=1}^{\tau}\sum_{h=1}^{H}\sum_{b=1}^{m} \sigma^2_{P,l-1}([z_{l,h}]_l,b)}. \tag{78}$$

Now (74) follows by combining (77) and (78). ■

C.2 Bayesian Regret Bound for GP-UCRL under GP prior: Proof of Theorem 3

Note that at every episode l, GP-UCRL (Algorithm 1) selects the policy π_l such that

$$V^{M_l}_{\pi_l,1}(s_{l,1}) = \max_{\pi} \max_{M \in \mathcal{M}_l} V^{M}_{\pi,1}(s_{l,1}) \tag{79}$$

where $s_{l,1}$ is the initial state, $\mathcal{M}_l$ is the family of MDPs constructed by GP-UCRL and M_l is the most optimistic realization from $\mathcal{M}_l$. Further from 59

$$\mathbb{P}\left[\sup_{z \in \mathcal{Z}} \left|\overline{R}_\star(z)\right| > L\right] < ae^{-(L/b)^2},$$

holds for any $L \geq 0$ for some corresponding $a, b > 0$. Thus for any $0 < \delta \leq 1$, setting $L = b\sqrt{\ln(6a/\delta)}$, with probability at least $1 - \delta/6$, for all $z \in \mathcal{Z}$

$$\left|\overline{R}_\star(z)\right| \leq b\sqrt{\ln(6a/\delta)}. \tag{80}$$

Now (80), Lemma 7 and Lemma 9 together with an union bound imply that for any $\tau \geq 1$, with probability at least $1-\delta/3$,

$$\sum_{l=1}^{\tau} \left(V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \leq \sum_{l=1}^{\tau}\sum_{h=1}^{H} \Big(\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| + L_{M_l} \left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 \Big) + (LD + 2CH)\sqrt{2\tau H \ln(6/\delta)}, \tag{81}$$

where $C := b\sqrt{\ln(6a/\delta)}$. Now for each $l \geq 1$, we define the following events:

$$\begin{aligned} E_{R,l} &:= \{\forall z \in \mathcal{Z}, \left|\overline{R}_\star(z) - \mu_{R,l-1}([z]_l)\right| \leq \beta_{R,l}\, \sigma_{R,l-1}([z]_l) + 1/l^2\}, \\ E_{P,l} &:= \{\forall z \in \mathcal{Z}, \left|\overline{P}_\star(z) - \mu_{P,l-1}([z]_l)\right| \leq \beta_{P,l} \left\|\sigma_{P,l-1}([z]_l)\right\|_2 + \sqrt{m}/l^2\}. \end{aligned}$$

By construction of the set of MDPs $\mathcal{M}_l$ in Algorithm 1, it follows that when both the events $E_{R,l}$ and $E_{P,l}$ hold for all $l \geq 1$, the unknown MDP $M_\star$ lies in $\mathcal{M}_l$ for all $l \geq 1$. Thus (79) implies $V^{M_l}_{\pi_l,1}(s_{l,1}) \geq V^{M_\star}_{\pi_\star,1}(s_{l,1})$ for all $l \geq 1$. This in turn implies, for every episode $l \geq 1$,

$$V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}) \leq V^{M_l}_{\pi_l,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1}). \tag{82}$$

Further when $E_{R,l}$ holds for all $l \geq 1$, then

$$\left|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\right| \leq \left|\overline{R}_{M_l}(z_{l,h}) - \mu_{R,l-1}([z_{l,h}]_l)\right| + \left|\overline{R}_\star(z_{l,h}) - \mu_{R,l-1}([z_{l,h}]_l)\right| \leq 2\beta_{R,l}\sigma_{R,l-1}([z_{l,h}]_l) + 2/l^2, \tag{83}$$

since the mean reward function $\overline{R}_{M_l}$ lies in the confidence set $\mathcal{C}_{R,l}$ as defined in (66). Similarly when $E_{P,l}$ holds for all $l \geq 1$,

$$\begin{aligned} \left\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\right\|_2 &\leq \left\|\overline{P}_{M_l}(z_{l,h}) - \mu_{P,l-1}([z_{l,h}]_l)\right\|_2 + \left\|\overline{P}_\star(z_{l,h}) - \mu_{P,l-1}([z_{l,h}]_l)\right\|_2 &(84)\\ &\leq 2\beta_{P,l} \left\|\sigma_{P,l-1}([z_{l,h}]_l)\right\|_2 + 2\sqrt{m}/l^2, &(85) \end{aligned}$$

since the mean transition function $\overline{P}_{M_l}$ lies in the confidence set $\mathcal{C}_{P,l}$ as defined in (66). Now combining (81), (82), (83) and (84), when both the events $E_{R,l}$ and $E_{P,l}$ hold for all $l \geq 1$, then with probability at least $1 - \delta/3$,

$$\sum_{l=1}^{\tau} \left(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) \leq 2\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left(\beta_{R,l}\sigma_{R,l-1}([z_{l,h}]_l) + 1/l^2 + L_{M_l}\beta_{P,l} \left\|\sigma_{P,l-1}([z_{l,h}]_l)\right\|_2 + L_{M_l}\sqrt{m}/l^2\right) + (LD + 2CH)\sqrt{2\tau H \ln(6/\delta)}.$$

Now Lemma 13 implies that $\mathbb{P}\left[\forall l \geq 1, E_{R,l}\right] \geq 1 - \delta/3$ and $\mathbb{P}\left[\forall l \geq 1, E_{P,l}\right] \geq 1 - \delta/3$. Hence, by a union bound, for any $\tau \geq 1$, with probability at least $1 - \delta$,

$$\begin{aligned} \sum_{l=1}^{\tau} \left(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\right) &\leq 2\beta_{R,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}([z_{l,h}]_l) + 2L\beta_{P,\tau}\sum_{l=1}^{\tau}\sum_{h=1}^{H} \left\|\sigma_{P,l-1}([z_{l,h}]_l)\right\|_2 \\ &\quad + (L\sqrt{m} + 1)H\pi^2/3 + (LD + 2CH)\sqrt{2\tau H \ln(6/\delta)}. \end{aligned} \tag{86}$$

Here we have used the fact that both $\beta_{R,l}$ and $\beta_{P,l}$ are non-decreasing with the number of episodes l, $\sum_{l=1}^{\tau} 1/l^2 \leq \pi^2/6$ and that $L_{M_l} \leq L$ by construction of $\mathcal{M}_l$ (and since $M_l \in \mathcal{M}_l$). Now from Lemma 14, we have $\sum_{l=1}^{\tau}\sum_{h=1}^{H} \sigma_{R,l-1}([z_{l,h}]_l) \leq \exp\big(\gamma_{H-1}(k_R,\mathcal{Z})\big)\sqrt{(2\lambda_R+1)\tau H \gamma_{\tau H}(k_R,\mathcal{Z})}$ and $\sum_{l=1}^{\tau}\sum_{h=1}^{H} \|\sigma_{P,l-1}([z_{l,h}]_l)\|_2 \leq \exp\big(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\big)\sqrt{(2\lambda_P+1)\tau H \gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})}$. Therefore with probability at least $1-\delta$, the cumulative regret of GP-UCRL after τ episodes, i.e. after $T=\tau H$ timesteps is

$$\begin{aligned} Regret(T) &= \sum_{l=1}^{\tau}\big(V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\big) \\ &\leq 2\beta_{R,\tau}\exp\big(\gamma_{H-1}(k_R,\mathcal{Z})\big)\sqrt{(2\lambda_R+1)\gamma_T(k_R,\mathcal{Z})T} \\ &\quad +2L\beta_{P,\tau}\exp\big(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\big)\sqrt{(2\lambda_P+1)\gamma_{mT}(k_P,\tilde{\mathcal{Z}})T} \\ &\quad +(L\sqrt{m}+1)H\pi^2/3 + (LD+2CH)\sqrt{2T\ln(6/\delta)}, \end{aligned}$$

where $C := b\sqrt{\ln(6a/\delta)}$, $\beta_{R,\tau} := \sqrt{2\ln\big(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,\pi^2\tau^2/\delta\big)}$ and $\beta_{P,\tau} := \sqrt{2\ln\big(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,m\pi^2\tau^2/\delta\big)}$. Now the result follows by defining $\gamma_T(R) := \gamma_T(k_R,\mathcal{Z})$ and $\gamma_{mT}(P) := \gamma_{mT}(k_P,\tilde{\mathcal{Z}})$.

C.3 Bayes Regret of PSRL under GP prior: Proof of Theorem 4

$\Phi \equiv (\Phi_R, \Phi_P)$ is the distribution of the unknown MDP $M_\star = \{\mathcal{S},\mathcal{A},R_\star,P_\star,H\}$, where Φ_R and Φ_P are specified by GP priors $GP_{\mathcal{Z}}(0,k_R)$ are $GP_{\tilde{\mathcal{Z}}}(0,k_P)$ respectively with a Gaussian noise model in the sense that

- The reward distribution is $R_\star : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, with mean $\overline{R}_\star \sim GP_{\mathcal{Z}}(0,k_R)$, and additive $\mathcal{N}(0,\lambda_R)$ Gaussian noise.
- The transition distribution is $P_\star : \mathcal{S}\times\mathcal{A}\to\mathcal{S}$, with mean $\overline{P}_\star \sim GP_{\tilde{\mathcal{Z}}}(0,\tilde{k}_P)$, and component-wise additive and independent $\mathcal{N}(0,\lambda_R)$ Gaussian noise.

Conditioned on the history of observations $\mathcal{H}_{l-1} := \{s_{j,k},a_{j,k},r_{j,k}\}_{1\leq j\leq l-1,1\leq k\leq H}$ both $M_\star$ and M_l are identically distributed with $\Phi_l \equiv (\Phi_{R,l},\Phi_{P,l})$, where $\Phi_{R,l}$ and $\Phi_{P,l}$ are specified by GP posteriors $GP_{\mathcal{Z}}(\mu_{R,l-1},k_{R,l-1})$ and $GP_{\tilde{\mathcal{Z}}}(\mu_{P,l-1},k_{P,l-1})$ respectively.

In this case we use the confidence sets $\mathcal{C}_{R,l}$ and $\mathcal{C}_{P,l}$ as given in 66 and define an event $E := E_\star \cap E_M$, where $E_\star := \big\{\overline{R}_\star \in \mathcal{C}_{R,l}, \overline{P}_\star \in \mathcal{C}_{P,l}\ \forall l\geq 1\big\}$ and $E_M := \big\{\overline{R}_{M_l} \in \mathcal{C}_{R,l}, \overline{P}_{M_l} \in \mathcal{C}_{P,l}\ \forall l\geq 1\big\}$. Now from Lemma 13, $\mathbb{P}[E_\star] \geq 1-2\delta/3$ and hence $\mathbb{P}[E] \geq 1-4\delta/3$ similarly as in the proof of Theorem 2. Further (51) implies

$$\begin{aligned} \mathbb{E}\left[\sum_{l=1}^{\tau}\Big[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\Big]\right] &\leq \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\Big[\,\big|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\big| \;\Big|\; E\right] \\ &\quad +\mathbb{E}\Big[L_{M_l}\big\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\big\|_2\Big] \;\Big|\; E\Big] + 8\delta C\tau H/3, \end{aligned} \tag{87}$$

where we have used that $\mathbb{E}\Big[V^{M_\star}_{\pi_\star,1}(s_{l,1}) - V^{M_\star}_{\pi_l,1}(s_{l,1})\Big] \leq 2CH$, where $C = \mathbb{E}\big[\sup_{z\in\mathcal{Z}}\big|\overline{R}_\star(z)\big|\big]$. From Lemma 13 and construction of $\mathcal{C}_{R,l}, l\geq 1$,

$$\begin{aligned} \mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H}\big|\overline{R}_{M_l}(z_{l,h}) - \overline{R}_\star(z_{l,h})\big| \;\Big|\; E\right] &\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}\Big(2\beta_{R,l}\sigma_{R,l-1}([z_{l,h}]_l) + 2/l^2\Big) \\ &\leq 2\beta_{R,\tau}\exp\big(\gamma_{H-1}(k_R,\mathcal{Z})\big)\sqrt{(2\lambda_R+1)\tau H\gamma_{\tau H}(k_R,\mathcal{Z})} + \frac{\pi^2 H}{3}, \end{aligned} \tag{88}$$

where the last step follows from Lemma 14. Similarly from Lemma 13 and construction of $\mathcal{C}_{P,l}, l\geq 1$,

$$\begin{aligned} &\mathbb{E}\left[\sum_{l=1}^{\tau}\sum_{h=1}^{H} L_{M_l}\big\|\overline{P}_{M_l}(z_{l,h}) - \overline{P}_\star(z_{l,h})\big\|_2 \;\Big|\; E\right] \\ &\leq \sum_{l=1}^{\tau}\sum_{h=1}^{H}\mathbb{E}\left[L_{M_l} \;\middle|\; E\right]\Big(2\beta_{P,l}\,\|\sigma_{P,l-1}([z_{l,h}]_l)\|_2 + 2\sqrt{m}/l^2\Big). \\ &\leq \frac{\mathbb{E}[L_\star]}{1-4\delta/3}\Big(2\beta_{P,\tau}\exp\big(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\big)\sqrt{(2\lambda_P+1)\tau H\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})} + \sqrt{m}\pi^2 H/3\Big), \end{aligned} \tag{89}$$

where the last step follows from Lemma 14 and from the proof of Theorem 2. Combining (87), (88) and (89), for any $0 < \delta \leq 1$ and $\tau \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\sum_{l=1}^{\tau}\left[V_{\pi_\star,1}^{M_\star}(s_{l,1}) - V_{\pi_l,1}^{M_\star}(s_{l,1})\right]\right] \quad &\leq \quad 2\beta_{R,\tau}\exp\left(\gamma_{H-1}(k_R,\mathcal{Z})\right)\sqrt{(2\lambda_R+1)\tau H\gamma_{\tau H}(k_R,\lambda_R,\mathcal{Z})} \\ &\quad +2\beta_{P,\tau}\frac{\mathbb{E}\left[L_\star\right]}{1-4\delta/3}\exp\left(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\right)\sqrt{(2\lambda_P+1)\tau H\gamma_{m\tau H}(k_P,\tilde{\mathcal{Z}})} \\ &\quad +8\delta C\tau H/3 + \frac{\left(1+\frac{\mathbb{E}[L^\star]}{1-4\delta/3}\sqrt{m}\right)\pi^2 H}{3},\end{aligned}$$

where $\beta_{R,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,\pi^2\tau^2/\delta\right)}$ and $\beta_{P,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,m\pi^2\tau^2/\delta\right)}$. See that the left hand side is independent of δ. Now using $\delta = 1/\tau H$, the Bayes regret of PSRL after τ episodes, i.e. after $T = \tau H$ timesteps is

$$\begin{aligned}\mathbb{E}\left[Regret(\tau)\right] \quad &= \quad \sum_{l=1}^{\tau}\mathbb{E}\left[V_{\pi_\star,1}^{M_\star}(s_{l,1}) - V_{\pi_l,1}^{M_\star}(s_{l,1})\right] \\ &\leq \quad 2\alpha_{R,\tau}\exp\left(\gamma_{H-1}(k_R,\mathcal{Z})\right)\sqrt{(2\lambda_R+1)\gamma_T(k_R,\mathcal{Z})T} \\ &\quad +3\,\mathbb{E}\left[L_\star\right]\alpha_{P,\tau}\exp\left(\gamma_{mH-1}(k_P,\tilde{\mathcal{Z}})\right)\sqrt{(2\lambda_P+1)\gamma_{mT}(k_P,\tilde{\mathcal{Z}})T} + 3C + (1+\sqrt{m}\mathbb{E}\left[L_\star\right])\pi^2 H,\end{aligned}$$

since $1/(1-4/3\tau H) \leq 3/2$ as $\tau \geq 2$, $H \geq 2$. Here $C = \mathbb{E}\left[\sup_{z\in\mathcal{Z}}\left|\overline{R}_\star(z)\right|\right]$, $\alpha_{R,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,\pi^2\tau^2T\right)}$, $\alpha_{P,\tau} := \sqrt{2\ln\left(|\mathcal{S}_\tau|\,|\mathcal{A}_\tau|\,m\pi^2\tau^2T\right)}$. Now the result follows by defining $\gamma_T(R) := \gamma_T(k_R,\mathcal{Z})$ and $\gamma_{mT}(P) := \gamma_{mT}(k_P,\tilde{\mathcal{Z}})$.