Active Learning of Conditional Mean Embeddings via Bayesian Optimisation (Appendix)

This document presents supplementary theoretical results and a discussion on applications of the proposed CME-UCB method to policy search in reinforcement learning (see Appendix D).

A AUXILIARY RESULTS FOR THE MAIN PROOFS

The following are auxiliary results referred by the proofs in the main paper.

Lemma A.1 (Chowdhury and Gopalan (2019))**.** *If $c(\mathbf{u},\mathbf{u}) \leq 1$ for all $\mathbf{u} \in \mathcal{U}$, then the following hold:*

$$\sigma_{t-1}^2(\mathbf{u}) \leq (1+1/\eta)\sigma_t^2(\mathbf{u}) \,,$$
$$\sum_{i=1}^{t} \sigma_i^2(\mathbf{u}_i) \leq \log\det\left(\eta^{-1}\mathbf{C}_t + \mathbf{I}\right) \,.$$

Similarly, if $k(\mathbf{x},\mathbf{x}) \leq 1$ for all $\mathbf{x} \in \mathcal{X}$, then

$$\sum_{i=1}^{t} s_i^2(\mathbf{x}_i) \leq \log\det\left(\lambda^{-1}\mathbf{K}_t + \mathbf{I}\right) \,.$$

Lemma A.2. *The conditional mean embedding operator $\Theta : \mathcal{H}_c \to \mathcal{H}_k$ is bounded with $\|\Theta\| < \infty$ if and only if, for any $f \in \mathcal{H}_k$, exists $g \in \mathcal{H}_c$ such that:*

$$\forall \mathbf{u} \in \mathcal{U}, \quad g(\mathbf{u}) = \mathbb{E}_{P_\mathbf{u}}[f] \,. \tag{1}$$

Proof. First, recall that any bounded linear operator between Hilbert spaces $\mathrm{M} : \mathcal{H}_c \to \mathcal{H}_k$ has an adjoint $\mathrm{M}^\mathsf{T} : \mathcal{H}_k \to \mathcal{H}_c$, which is such that:

$$\langle f, \mathrm{M}\, g\rangle_k = \langle \mathrm{M}^\mathsf{T} f, g\rangle_c \,, \tag{2}$$

for any $f \in \mathcal{H}_k$ and any $g \in \mathcal{H}_c$ (Kreyszig, 1978, Thm. 3.9-2). Therefore, if the conditional mean embedding operator Θ, as previously defined, is a bounded linear operator, we also have that:

$$\begin{aligned}\forall \mathbf{u} \in \mathcal{U}, \quad \langle f, \Theta\, \phi_c(\mathbf{u})\rangle_k &= \langle \Theta^\mathsf{T} f, \phi_c(\mathbf{u})\rangle_c \\ &= \langle g, \phi_c(\mathbf{u})\rangle_c = g(\mathbf{u}) \,,\end{aligned}$$

where we set $g := \Theta^\mathsf{T} f \in \mathcal{H}_c$. Conversely, if there is a $g \in \mathcal{H}_c$, such that:

$$\forall \mathbf{u} \in \mathcal{U}, \quad \langle f, \Theta\phi_c(\mathbf{u})\rangle_k = \langle g, \phi_c(\mathbf{u})\rangle_c \,, \tag{3}$$

the mapping $\mathrm{M} : f \mapsto g$ is linear due to the linearity of the expectation. The latter also implies that:

$$\forall h \in \mathcal{H}_c, \quad \langle g, h\rangle_c = \langle f, \Theta\, h\rangle_k \,, \tag{4}$$

which in turn implies that $\|\Theta\, h\|_k < \infty$, for all $h \in \mathcal{H}_c$, as $\Theta\, h \in \mathcal{H}_k$. Therefore, Θ must be bounded. □

B OTHER AUXILIARY RESULTS

This section presents a few auxiliary results which assist in a practical implementation of the algorithm. For the next derivations, we use $\mathrm{eig}(\mathbf{A})$ to denote the set of eigenvalues of a matrix $\mathbf{A}$. For positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, we use $\mathbf{A} \preccurlyeq \mathbf{B}$ to indicate that $\mathbf{B} - \mathbf{A}$ is positive semi-definite. We also use $\|\mathbf{v}\|_\mathbf{A} = \sqrt{\mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v}}$ to denote the Mahalanobis norm of a vector $\mathbf{v} \in \mathbb{R}^m$, which is a valid norm whenever $\mathbf{A}$ is positive-definite.

The following provides an exact expression for the RKHS norm of the conditional mean embedding operator Θ when the control space $\mathcal{U}$ is finite and c is a strictly positive-definite kernel (Sriperumbudur et al., 2011).

Lemma B.1 (Bound on $\|\Theta\|_\mathrm{op}$ for finite $\mathcal{U}$)**.** *Let $\mathcal{U} := \{\mathbf{u}_i\}_{i=1}^m$, $m \in \mathbb{N}$. Assume $c : \mathcal{U}\times\mathcal{U} \to \mathbb{R}$ is a strictly positive-definite kernel. Then we have that:*

$$\|\Theta\|_\mathrm{op} = \sqrt{\max \mathrm{eig}\left(\mathbf{C}^{-1/2}\hat{\mathbf{K}}\mathbf{C}^{-1/2}\right)} \,,$$

where $\mathbf{C} := [c(\mathbf{u}_i,\mathbf{u}_j)]_{i,j=1}^m$ and $\hat{\mathbf{K}} := [\langle \vartheta(\mathbf{u}_i), \vartheta(\mathbf{u}_j)\rangle_k]_{i,j=1}^m$.

Proof. As $\mathcal{U}$ is finite, we can express any function in the corresponding RKHS $\mathcal{H}_c$ as:

$$g \in \mathcal{H}_c, \quad g = \sum_{i=1}^{m} \alpha_i c(\cdot, \mathbf{u}_i) \,, \tag{5}$$

where $\alpha_i \in \mathbb{R}$, for $i \in \{1, \dots, m\}$. The RKHS norm can be expressed as $\|g\|_c = \sqrt{\boldsymbol{\alpha}^\mathsf{T} \mathbf{C} \boldsymbol{\alpha}}$, where $[\boldsymbol{\alpha}]_i = \alpha_i$. In addition, from the definition of ϑ, applying Θ to any $g \in \mathcal{H}_c$ we have:

$$\Theta\, g = \sum_{i=1}^{m} \alpha_i \vartheta(\mathbf{u}_i)\,, \tag{6}$$

which has RKHS norm $\|\Theta\, g\|_k = \sqrt{\boldsymbol{\alpha}^\mathsf{T} \hat{\mathbf{K}} \boldsymbol{\alpha}}$.

Plugging equations (5) and (6) into the definition of the operator norm, we have:

$$\begin{aligned} \|\Theta\|_{\mathrm{op}} &= \sup_{g \in \mathcal{H}_c : \|g\|_c = 1} \|\Theta\, g\|_k \\ &= \sup_{\boldsymbol{\alpha} \in \mathbb{R}^m : \sqrt{\boldsymbol{\alpha}^\mathsf{T} \mathbf{C} \boldsymbol{\alpha}} = 1} \sqrt{\boldsymbol{\alpha}^\mathsf{T} \hat{\mathbf{K}} \boldsymbol{\alpha}}\,. \end{aligned} \tag{7}$$

As c is strictly positive-definite on $\mathcal{U}$, $\mathbf{C}$ is a positive-definite matrix, ensuring its inverse exists. Making a change of variable $\boldsymbol{\alpha}' = \mathbf{C}^{1/2} \boldsymbol{\alpha}$, we obtain:

$$\begin{aligned} \|\Theta\|_{\mathrm{op}} &= \sup_{\boldsymbol{\alpha}' \in \mathbb{R}^m} \frac{\sqrt{\boldsymbol{\alpha}'^\mathsf{T} \mathbf{C}^{-1/2} \hat{\mathbf{K}} \mathbf{C}^{-1/2} \boldsymbol{\alpha}'}}{\sqrt{\boldsymbol{\alpha}'^\mathsf{T} \boldsymbol{\alpha}'}} \\ &= \sqrt{\max \mathrm{eig}\left(\mathbf{C}^{-1/2} \hat{\mathbf{K}} \mathbf{C}^{-1/2} \right)}\,, \end{aligned} \tag{8}$$

where the latter follows from the Rayleigh-Ritz theorem (Horn and Johnson, 1985, Theorem 4.2.2), i.e. $\max_{\boldsymbol{\alpha} \neq \mathbf{0}} \frac{\boldsymbol{\alpha}^\mathsf{T} \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{\alpha}} = \max \mathrm{eig}(\mathbf{M})$, combined with the monotonicity of the square root. □

For our experiments, we synthesised state noise as Gaussian random variables. In this case, the following result allows us to compute the operator norm of the conditional embedding operator when combined with Lemma B.1.

Lemma B.2. *Let $k(\mathbf{x}, \mathbf{x}') := \exp(-\|\mathbf{x} - \mathbf{x}'\|^2_{\mathbf{L}^{-1}})$ be the squared exponential kernel, with $\mathbf{x}, \mathbf{x}' \in \mathcal{X} = \mathbb{R}^D$. Let state distributions be Gaussian $P_{\mathbf{u}} = N(\hat{\mathbf{x}}(\mathbf{u}), \boldsymbol{\Sigma}(\mathbf{u}))$, for $\mathbf{u} \in \mathcal{U}$. Then the inner product between conditional mean embeddings defines the following kernel:*

$$\begin{aligned} \hat{k}(P_{\mathbf{u}}, P_{\mathbf{u}'}) &:= \langle \vartheta(\mathbf{u}), \vartheta(\mathbf{u}') \rangle_k \\ &= \frac{\exp\left(-\frac{1}{2} \|\hat{\mathbf{x}}(\mathbf{u}) - \hat{\mathbf{x}}(\mathbf{u}')\|^2_{\mathbf{L}(\mathbf{u}, \mathbf{u}')^{-1}} \right)}{\det(\mathbf{I} + \mathbf{L}^{-1}(\boldsymbol{\Sigma}(\mathbf{u}) + \boldsymbol{\Sigma}(\mathbf{u})))^{1/2}}\,, \end{aligned} \tag{9}$$

where $\mathbf{L}(\mathbf{u}, \mathbf{u}') := \mathbf{L} + \boldsymbol{\Sigma}(\mathbf{u}) + \boldsymbol{\Sigma}(\mathbf{u}')$.

Proof. This result simply follows from the definition of the conditional mean embeddings and a closed-form solution for the expected value of the squared-exponential kernel under independent Gaussian inputs (Girard, 2004, Eq. 3.53). □

C PROOFS FOR THE REFINED CME-UCB

In this section we provide proofs for the main results regarding the improved CME-UCB algorithm. We first introduce auxiliary results which we will use for the proofs.

We start by adapting a general result by Abbasi-Yadkori (2012, Corollary 3.15) to our settings. For these results, we use:

$$\mathrm{W}_t := \lambda \mathbf{I} + \boldsymbol{\Phi}_k(\mathbf{X}_t) \boldsymbol{\Phi}_k(\mathbf{X}_t)^\mathsf{T}\,, \tag{10}$$

which is a positive-definite operator on $\mathcal{H}_k$. Note that, if we reverse the second term on the right-hand side above, we recover $\mathbf{K}_t = \boldsymbol{\Phi}_k(\mathbf{X}_t)^\mathsf{T} \boldsymbol{\Phi}_k(\mathbf{X}_t)$, which is the observed states kernel matrix. With the definitions above, we obtain the following concentration inequality on the RKHS distance between the objective function f and the least-squares estimator $\hat{\mu}_t$.

Lemma C.1. *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, uniformly over all $t \geq 0$,*

$$\|f - \hat{\mu}_t\|_{\mathrm{W}_t} \leq \beta_{k,t}(\delta)\,,$$

where $\beta_{k,t}(\delta)$ is given by Lemma 2.

Proof. The proof simply follows by verifying that our assumptions on the observation process are equivalent to those of Abbasi-Yadkori (2012, Corollary 3.15) and that:

$$b\sqrt{\lambda} + \sigma_\zeta \sqrt{2 \log\left(\frac{\det(\mathbf{I} + \lambda^{-1} \mathbf{K}_t)^{1/2}}{\delta} \right)} \leq \beta_{k,t}(\delta)\,, \tag{11}$$

since $\frac{1}{2} \log\det(\mathbf{I} + \lambda^{-1} \mathbf{K}_t) \leq \gamma_{k,t}$. □

Now consider the posterior state kernel:

$$k_t(\mathbf{x}, \mathbf{x}') := k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\mathsf{T} (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{k}_t(\mathbf{x})\,, \tag{12}$$

and its expected value under arbitrary state distributions $P, P' \in \mathcal{P}$:

$$k_t(P, P') := \int_{\mathcal{X}} \int_{\mathcal{X}} k_t(\mathbf{x}, \mathbf{x}')\, \mathrm{d}P(\mathbf{x})\, \mathrm{d}P'(\mathbf{x}')\,, \tag{13}$$

where we allow for a slight abuse of notation. Similarly, let us define:

$$s_t(P) := \sqrt{\lambda^{-1} k_t(P, P)}\,. \tag{14}$$

Let $\hat{P}^t_{\mathbf{u}}$ be a t sample empirical approximation to the conditional distribution $P_{\mathbf{u}}$ in the sense that $\mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f] = \langle f, \hat{\vartheta}_t(\mathbf{u}) \rangle_k$ for any $f \in \mathcal{H}_k$. Then, with Lemma C.1, we obtain the following UCB on the expected value of the objective function under the learnt conditional mean embedding, which is a restatement of Lemma 5.

Lemma C.2 (Restatement of Lemma 5)**.** *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, uniformly over all $t \geq 0$ and $\mathbf{u} \in \mathcal{U}$*

$$|\mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[\hat{\mu}_t]| \leq \beta_{k,t}(\delta) s_t(\hat{P}^t_{\mathbf{u}}) \, .$$

Proof. First, observe that:

$$\begin{aligned} \left|\mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[\hat{\mu}_t]\right| &= \left|\langle f - \hat{\mu}_t, \hat{\vartheta}_t(\mathbf{u})\rangle_k\right| \\ &= \left|\langle \mathrm{W}_t^{1/2}(f - \hat{\mu}_t), \mathrm{W}_t^{-1/2}\, \hat{\vartheta}_t(\mathbf{u})\rangle_k\right| \\ &\leq \|f - \hat{\mu}_t\|_{\mathrm{W}_t} \|\hat{\vartheta}_t(\mathbf{u})\|_{\mathrm{W}_t^{-1}} \, , \end{aligned} \tag{15}$$

which follows by the Cauchy-Schwarz inequality and the fact that W_t is self-adjoint. Concerning the second term on the final right-hand side, using the Woodbury inverse matrix identity, we know that:

$$\lambda \, \mathrm{W}_t^{-1} = \mathbf{I} - \boldsymbol{\Phi}_k(\mathbf{X}_t)(\mathbf{K}_t + \lambda \mathbf{I})\boldsymbol{\Phi}_k(\mathbf{X}_t)^{\mathsf{T}} \, . \tag{16}$$

From Equation 12, we then have that:

$$k_t(\mathbf{x}, \mathbf{x}') = \lambda \langle \phi_k(\mathbf{x}), \mathrm{W}_t^{-1}\, \phi_k(\mathbf{x}')\rangle_k \, . \tag{17}$$

Finally, applying the definition of the conditional mean embedding to Equation 13 yields:

$$\begin{aligned} \|\hat{\vartheta}_t(\mathbf{u})\|^2_{\mathrm{W}_t^{-1}} &= \langle \hat{\vartheta}_t(\mathbf{u}), \mathrm{W}_t^{-1}\, \hat{\vartheta}_t(\mathbf{u})\rangle_k \\ &= \lambda^{-1} k_t(\hat{P}^t_{\mathbf{u}}, \hat{P}^t_{\mathbf{u}}) \\ &= s_t^2(\hat{P}^t_{\mathbf{u}}) \end{aligned} \tag{18}$$

The end result then follows by Lemma C.1. □

The predictive variance $s_t^2(\hat{P}^t_{\mathbf{u}})$ can be computed in practice as:

$$s_t^2(\hat{P}^t_{\mathbf{u}}) = \lambda^{-1} \mathbf{v}_t(\mathbf{u})^{\mathsf{T}} k_t(\mathbf{X}_t, \mathbf{X}_t) \mathbf{v}_t(\mathbf{u}) \, , \tag{19}$$

where $\mathbf{v}_t(\mathbf{u}) := (\mathbf{C}_t + \eta \mathbf{I})^{-1} \mathbf{c}_t(\mathbf{u})$ and

$$k_t(\mathbf{X}_t, \mathbf{X}_t) := \mathbf{K}_t - \mathbf{K}_t(\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{K}_t \, , \tag{20}$$

which is the GP posterior covariance matrix on the observed states. It is worth noting that the predictive variance $s_t^2(\hat{P}^t_{\mathbf{u}})$ is always smaller than $\sigma_t^2(\mathbf{u})$, the Mahalanobis norm square of the control features. To see this, we first note that:

$$\begin{aligned} s_t^2(\hat{P}^t_{\mathbf{u}}) &= \mathbf{v}_t(\mathbf{u})^{\mathsf{T}} \mathbf{K}_t (\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{v}_t(\mathbf{u}) \\ &\leq \|(\mathbf{C}_t + \eta \mathbf{I})^{-1} \mathbf{c}_t(\mathbf{u})\|_2^2 \, . \end{aligned} \tag{21}$$

We now define:

$$\mathrm{V}_t := \eta \mathbf{I} + \boldsymbol{\Phi}_c(\mathbf{U}_t) \boldsymbol{\Phi}_c(\mathbf{U}_t)^{\mathsf{T}} \, ,$$

which is a positive-definite operator on $\mathcal{H}_c$. Note that, if we reverse the second term on the right-hand side above, we recover $\mathbf{C}_t = \boldsymbol{\Phi}_c(\mathbf{U}_t)^{\mathsf{T}} \boldsymbol{\Phi}_c(\mathbf{U}_t)$ and $\mathbf{c}_t(\mathbf{u}) = \boldsymbol{\Phi}_c(\mathbf{U}_t)^{\mathsf{T}} \phi_c(\mathbf{u})$. With the definitions above, we now obtain:

$$(\mathbf{C}_t + \eta \mathbf{I})^{-1} \mathbf{c}_t(\mathbf{u}) = \boldsymbol{\Phi}_c(\mathbf{U}_t)^{\mathsf{T}} \, \mathrm{V}_t^{-1} \, \phi_c(\mathbf{u}) \, .$$

Substituting this to Equation 21, we then upper bound the predictive variance as:

$$\begin{aligned} s_t^2(\hat{P}^t_{\mathbf{u}}) &\leq \phi_c(\mathbf{u})^{\mathsf{T}} \, \mathrm{V}_t^{-1} \, \boldsymbol{\Phi}_c(\mathbf{U}_t) \boldsymbol{\Phi}_c(\mathbf{U}_t)^{\mathsf{T}} \, \mathrm{V}_t^{-1} \, \phi_c(\mathbf{u}) \\ &= \phi_c(\mathbf{u})^{\mathsf{T}} \, \mathrm{V}_t^{-1} \, \phi_c(\mathbf{u}) - \eta \, \phi_c(\mathbf{u})^{\mathsf{T}} \, \mathrm{V}_t^{-2} \, \phi_c(\mathbf{u}) \\ &= \sigma_t^2(\mathbf{u}) - \eta \, \|\mathrm{V}_t^{-1} \, \phi_c(\mathbf{u})\|_c^2 \\ &\leq \sigma_t^2(\mathbf{u}) \, . \end{aligned} \tag{22}$$

Now, combining Lemma C.2 with Theorem 1 in the main paper, we obtain the following tighter confidence interval as compared to Lemma 3.

Proposition C.3 (Restatement of Proposition 6)**.** *For any $\delta \in (0, 1]$, let $\beta_{c,t}(\delta)$ and $\beta_{k,t}(\delta)$ be as given in Theorem 1 and Lemma 2, respectively. Then, with probability at least $1 - \delta$, the following holds uniformly over all $t \geq 0$ and $\mathbf{u} \in \mathcal{U}$:*

$$|\mathbb{E}_{P_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[\hat{\mu}_t]| \leq b\beta_{c,t}(\delta/2)\sigma_t(\mathbf{u}) + \beta_{k,t}(\delta/2) s_t(\hat{P}^t_{\mathbf{u}}) \, .$$

Proof. Using the triangle inequality, we have:

$$\begin{aligned} |\mathbb{E}_{P_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[\hat{\mu}_t]| \leq & |\mathbb{E}_{P_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f]| \\ & + |\mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[\hat{\mu}_t]| \, . \end{aligned} \tag{23}$$

Applying Theorem 1 and the Cauchy-Schwarz inequality to the first term on the right-hand side of Equation 23, it holds with probability at least $1 - \delta$:

$$\begin{aligned} |\mathbb{E}_{P_{\mathbf{u}}}[f] - \mathbb{E}_{\hat{P}^t_{\mathbf{u}}}[f]| &= |\langle f, \vartheta(\mathbf{u}) - \hat{\vartheta}_t(\mathbf{u})\rangle_k| \\ &\leq \|f\|_k \|\vartheta(\mathbf{u}) - \hat{\vartheta}_t(\mathbf{u})\|_k \\ &\leq \|f\|_k \beta_{c,t}(\delta) \sigma_t(\mathbf{u}) \, . \end{aligned} \tag{24}$$

The result in Proposition C.3 then follows by applying Lemma C.2 and substituting Equation 24 into Equation 23. □

Refined algorithm and its regret bound Given past observations $\mathcal{D}_{t-1} = \{(\mathbf{u}_i, \mathbf{x}_i, y_i)\}_{i=1}^{t-1}$, we define a refined UCB acquisition function:

$$\tilde{h}(\mathbf{u}|\mathcal{D}_{t-1}) = \langle \hat{\mu}_{t-1}, \hat{\vartheta}_{t-1}(\mathbf{u})\rangle_k + \beta_{t-1}(\mathbf{u}) \, ,$$

where $\beta_t(\mathbf{u}) := b\beta_{c,t}(\delta/2)\sigma_t(\mathbf{u}) + \beta_{k,t}(\delta/2) s_t(\hat{P}^t_{\mathbf{u}})$, for any $t \geq 0$. As before, we choose $\mathbf{u}_t$ that maximises this refined acquisition function. We now derive an upper bound on the cumulative regret of this refined version of the CME-UCB algorithm.

Theorem C.4 (Restatement of Theorem 7)**.** *Fix any $\delta \in (0, 1]$. Then, under the same hypothesis of Proposition C.3, the refined CME-UCB, enjoys, with probability at least $1 - \delta$, the regret bound:*

$$R_n \leq 2\left(b\beta_{c,n}(\delta/2) + \beta_{k,n}(\delta/2)\right)\sqrt{2(1+1/\eta)\gamma_{c,n}\, n}\,.$$

Proof. Let us assume that:

$$\forall t \geq 0,\ \forall \mathbf{u} \in \mathcal{U},\ \left|\langle f, \vartheta(\mathbf{u})\rangle_k - \langle \hat{\mu}_t, \hat{\vartheta}_t(\mathbf{u})\rangle_k\right| := \left|\mathbb{E}_{P_\mathbf{u}}[f] - \mathbb{E}_{\hat{P}^t_\mathbf{u}}[\hat{\mu}_t]\right| \leq \beta_t(\mathbf{u})\,. \quad (25)$$

Then the instantaneous regret at time $t \geq 1$ is:

$$\begin{aligned} r_t &:= \mathbb{E}[f(\mathbf{x})|\mathbf{u}^\star] - \mathbb{E}[f(\mathbf{x})|\mathbf{u}_t] \\ &= \langle f, \vartheta(\mathbf{u}^\star)\rangle_k - \langle f, \vartheta(\mathbf{u}_t)\rangle_k \\ &\leq \langle \hat{\mu}_{t-1}, \hat{\vartheta}_{t-1}(\mathbf{u}^\star)\rangle_k + \beta_{t-1}(\mathbf{u}^\star) - \langle f, \vartheta(\mathbf{u}_t)\rangle_k \\ &\leq \langle \hat{\mu}_{t-1}, \hat{\vartheta}_{t-1}(\mathbf{u}_t)\rangle_k + \beta_{t-1}(\mathbf{u}_t) - \langle f, \vartheta(\mathbf{u}_t)\rangle_k \\ &\leq 2\beta_{t-1}(\mathbf{u}_t) \\ &\leq 2\left(b\beta_{c,t-1}(\delta/2) + \beta_{k,t-1}(\delta/2)\right)\sigma_{t-1}(\mathbf{u}_t)\,, \end{aligned}$$

where the second inequality is due the choice of $\mathbf{u}_t$ in the refined algorithm, the first and the third inequalities are due to Equation 25 and the last inequality is due to Equation 22. Now, by Proposition C.3, Equation 25 holds with probability at least $1 - \delta$. Then, with probability at least $1 - \delta$, we have the cumulative regret:

$$\begin{aligned} R_n &\leq \sum_{t=1}^{n} 2\left(b\beta_{c,t-1}(\delta/2) + \beta_{k,t-1}(\delta/2)\right)\sigma_{t-1}(\mathbf{u}_t) \\ &\leq 2\left(b\beta_{c,n}(\delta/2) + \beta_{k,n}(\delta/2)\right)\sum_{t=1}^{n}\sigma_{t-1}(\mathbf{u}_t) \\ &\leq 2\left(b\beta_{c,n}(\delta/2) + \beta_{k,n}(\delta/2)\right)\sqrt{n\sum_{t=1}^{n}\sigma^2_{t-1}(\mathbf{u}_t)}\,, \end{aligned}$$

where the last step is due to the the Cauchy-Schwartz inequality and the second last step is due to the monotonicity of $\beta_{c,t}$ and $\beta_{k,t}$. Now the result follows from the definition of $\gamma_{c,t}$ along with the identities $\sigma^2_{t-1}(\mathbf{u}) \leq (1 + 1/\eta)\sigma^2_t(\mathbf{u})$ and $\sum_{t=1}^{n}\sigma^2_t(\mathbf{u}_t) = \log\det\left(\lambda^{-1}\mathbf{C}_n + \mathbf{I}\right)$ (Lemma A.1). □

D APPLICATION TO REINFORCEMENT LEARNING

Following a similar strategy to the likelihood-free inference application, we can let $\mathbf{u} \in \mathcal{U}$ represent the parameters of a policy $\pi_\mathbf{u}$ in reinforcement learning. When executed, policies generate trajectories $\boldsymbol{\xi}$, which are associated with a task-specific total reward $J[\boldsymbol{\xi}]$. A common objective is to maximise the expected return, defined as $J[\pi] = \mathbb{E}_{\boldsymbol{\xi}}[J[\boldsymbol{\xi}]]$, with a slight abuse of notation.

To apply our framework to the policy search problem, we use $\vartheta(\mathbf{u})$ to model $p(\boldsymbol{\xi}|\mathbf{u})$. Trajectories are usually high-dimensional, but filled with information that can be compressed into summary statistics $\mathbf{x}_{\boldsymbol{\xi}}$ (Ramos et al., 2019). We then model $f : \mathcal{X} \to \mathbb{R}$, such that $f(\mathbf{x}_{\boldsymbol{\xi}}) = J[\boldsymbol{\xi}]$, as an element of $\mathcal{H}_k$. Our results then allow for an upper bound on the regret of a policy-search algorithm following the proposed UCB strategy.

One can also apply CME-UCB to learn a likelihood model $p(\boldsymbol{\xi}_o|\mathbf{u})$ for a imitation learning objective, trying to match an observed trajectory $\boldsymbol{\xi}_o$ (Hussein et al., 2017). In this case, the problem reverts to maximum likelihood estimation in a likelihood-free inference framework.

References

Yasin Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. Phd, University of Alberta, 2012.

Sayak Ray Chowdhury and Aditya Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 13790–13801, 2019.

Agathe Girard. *Approximate methods for propagation of uncertainty with Gaussian process models*. Ph. d, University of Glasgow, 2004.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2), April 2017. ISSN 0360-0300. doi: 10.1145/3054912. URL `https://doi.org/10.1145/3054912`.

Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1978.

Fabio Ramos, Rafael Carvalhaes Possas, and Dieter Fox. BayesSim : adaptive domain randomization via probabilistic inference for robotics simulators. In *Robotics: Science and Systems (RSS)*, Freiburg im Breisgau, Germany, 2019.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research (JMLR)*, 12:2389–2410, 2011.