Appendix

A DETAILS ON ALGORITHMS

A.1 Pseudo-codes of MT-KB and MT-BKB

Algorithm 1 Multi-task kernelized bandits (MT-KB)

Require: Kernel Γ, distribution P_λ, scalarization s_λ, time budget T, parameters η, $\{\beta_t\}_{t=0}^{T-1}$
Initialize $\mu_0(x) = 0$ and $\Gamma_0(x, x') = \Gamma(x, x')$
for round $t = 1, 2, 3, \ldots, T$ **do**
  Compute acquisition function $u_t(x) = \mathbb{E}\left[s_\lambda\left(\mu_{t-1}(x)\right)\right] + L \cdot \beta_{t-1} \left\|\Gamma_{t-1}(x, x)\right\|^{1/2}$
  Select point $x_t \in \operatorname{argmax}_{x\in\mathcal{X}} u_t(x)$
  Get vector-valued output $y_t = f(x_t) + \varepsilon_t$
  Compute

$$G_t(x) = \left[\Gamma(x_1, x)^\top, \ldots, \Gamma(x_t, x)^\top\right]^\top,\ G_t = \left[\Gamma(x_i, x_j)\right]_{i,j=1}^t,\ Y_t = \left[y_1^\top, \ldots, y_t^\top\right]^\top$$

  Update the model

$$\mu_t(x) = G_t(x)^\top (G_t + \eta I_{nt})^{-1} Y_t$$
$$\Gamma_t(x, x) = \Gamma(x, x) - G_t(x)^\top (G_t + \eta I_{nt})^{-1} G_t(x)$$

end for

Algorithm 2 Multi-task budgeted kernelized bandits (MT-BKB)

Require: Kernel Γ, distribution P_λ, scalarization s_λ, time budget T, parameters η, q, $\{\tilde\beta_t\}_{t=0}^{T-1}$
Initialize $\tilde\mu_0(x) = 0$ and $\tilde\Gamma_0(x, x') = \Gamma(x, x')$
for round $t = 1, 2, 3, \ldots, T$ **do**
  Compute acquisition function $\tilde u_t(x) = \mathbb{E}\left[s_\lambda\left(\tilde\mu_{t-1}(x)\right)\right] + L \cdot \tilde\beta_{t-1} \|\tilde\Gamma_{t-1}(x, x)\|^{1/2}$
  Select point $x_t \in \operatorname{argmax}_{x\in\mathcal{X}} \tilde u_t(x)$
  Get vector-valued output $y_t = f(x_t) + \varepsilon_t$
  Initialize dictionary $\mathcal{D}_t = \emptyset$
  for $i = 1, 2, 3, \ldots, t$ **do**
    Set inclusion probability $p_{t,i} = \min\left\{q\|\tilde\Gamma_{t-1}(x_i, x_i)\|, 1\right\}$
    Draw $z_{t,i} \sim$ Bernoulli$(p_{t,i})$
    if $z_{t,i} = 1$ **then**
      Update $\mathcal{D}_t = \mathcal{D}_t \cup \{x_i\}$
    end if
  end for
  Set $m_t = |\mathcal{D}_t|$, enumerate $\mathcal{D}_t = \{x_{i_1}, \ldots, x_{i_{m_t}}\}$ and compute

$$\tilde G_t(x) = \left[\frac{1}{\sqrt{p_{t,i_1}}}\Gamma(x_{i_1}, x)^\top, \ldots, \frac{1}{\sqrt{p_{t,i_{m_t}}}}\Gamma(x_{i_{m_t}}, x)^\top\right]^\top,\ \tilde G_t = \left[\frac{1}{\sqrt{p_{t,i_u} p_{t,i_v}}}\Gamma(x_{i_u}, x_{i_v})\right]_{u,v=1}^{m_t}$$

  Find Nyström embeddings $\tilde\Phi_t(x) = \left(\tilde G_t^{1/2}\right)^+ \tilde G_t(x)$
  Compute $\tilde V_t = \sum_{s=1}^t \tilde\Phi_t(x_s)\tilde\Phi_t(x_s)^\top$ and update

$$\tilde\mu_t(x) = \tilde\Phi_t(x)^\top (\tilde V_t + \eta I_{nm_t})^{-1} \sum_{s=1}^t \tilde\Phi_t(x_s) y_s$$
$$\tilde\Gamma_t(x, x) = \Gamma(x, x) - \tilde\Phi_t(x)^\top \tilde\Phi_t(x) + \eta \tilde\Phi_t(x)^\top (\tilde V_t + \eta \cdot I_{nm_t})^{-1} \tilde\Phi_t(x)$$

end for

A.2 Computational Complexity under ICM (Separable) Kernels

In this section, we describe the time complexities of MT-KB and MT-BKB for the intrinsic coregionalization model (ICM) $\Gamma(x,x') = k(x,x')B$. As discussed earlier, we assume that an efficient oracle to optimize the acquisition function is provided to us, and the per step cost comes only from computing it. To this end, we first describe simplified model updates under ICM kernel using the eigen-system of B and then detail out the time required for computing the updates. We note here that the eigen decomposition, which is $O(n^3)$, needs to be computed only once at the beginning and can be used at every step of the algorithms.

Per-step Complexity of MT-KB Let $B = \sum_{i=1}^n \xi_i u_i u_i^\top$ denotes the eigen decomposition of the positive semi-definite matrix B. Then, $\Gamma(x,x) = \sum_{i=1}^n \xi_i k(x,x) u_i u_i^\top$. From the definition of the Kronecker product, we now have $G_t = \sum_{i=1}^n \xi_i K_t \otimes u_i u_i^\top$ and $G_t(x) = \sum_{i=1}^n \xi_i k_t(x) \otimes u_i u_i^\top$, where $K_t = [k(x_i,x_j)]_{i,j=1}^t$ and $k_t(x) = [k(x_1,x),\ldots,k(x_t,x)]^\top$. Since $\{u_i\}_{i=1}^n$ yields an orthonormal basis of $\mathbb{R}^n$, the output $y_t \in \mathbb{R}^n$ can be written as $y_t = \sum_{i=1}^n y_t^\top u_i \cdot u_i$. We then have $Y_t = \sum_{i=1}^n Y_t^i \otimes u_i$, where $Y_t^i = \left[y_1^\top u_i,\ldots,y_t^\top u_i\right]^\top$. We also note that $I_{nt} = \sum_{i=1}^n I_t \otimes u_i u_i^\top$, and, therefore $G_t + \eta I_{nt} = \sum_{i=1}^n (\xi_i K_t + \eta I_t) \otimes u_i u_i^\top$. Now, let $K_t = \sum_{j=1}^t \alpha_j w_j w_j^\top$ denotes the eigen decomposition of the (positive semi-definite) kernel matrix K_t. We then have

$$G_t + \eta I_{nt} = \sum_{i=1}^n \sum_{j=1}^t (\xi_i \alpha_j + \eta) w_j w_j^\top \otimes u_i u_i^\top = \sum_{i=1}^n \sum_{j=1}^t (\xi_i \alpha_j + \eta)(w_j \otimes u_i)(w_j \otimes u_i)^\top. \tag{4}$$

By the properties of tensor product $(w_j \otimes u_i)^\top (w_{j'} \otimes u_{i'}) = (w_j^\top w_{j'}) \cdot (u_i^\top u_{i'})$, which is equal to 1 if $i = i'$, $j = j'$, and is equal to 0 otherwise. Therefore, (4) denotes the eigen decomposition of $G_t + \eta I_{nt}$. Hence

$$(G_t + \eta I_{nt})^{-1} = \sum_{i=1}^n \sum_{j=1}^t \frac{1}{\xi_i \alpha_j + \eta} w_j w_j^\top \otimes u_i u_i^\top = \sum_{i=1}^n (\xi_i K_t + \eta I_t)^{-1} \otimes u_i u_i^\top. \tag{5}$$

By the orthonormality of $\{u_i\}_{i=1}^n$ and the mixed product property of Kronecker product, we now obtain $(G_t + \eta I_{nt})^{-1} Y_t = \sum_{i=1}^n (\xi_i K_t + \eta I_t)^{-1} Y_t^i \otimes u_i$, and thus, in turn,

$$\mu_t(x) = G_t(x)^\top (G_t + \eta I_{nt})^{-1} Y_t = \sum_{i=1}^n \xi_i k_t(x)^\top (\xi_i K_t + \eta I_t)^{-1} Y_t^i \cdot u_i. \tag{6}$$

Similarly, we get $G_t(x)^\top (G_t + \eta I_{nt})^{-1} G_t(x) = \sum_{i=1}^n \xi_i^2 k_t(x)^\top (\xi_i K_t + \eta I_t)^{-1} k_t(x) \cdot u_i u_i^\top$ and therefore,

$$\|\Gamma_t(x,x)\| = \max_{1 \leqslant i \leqslant n} \xi_i \left(k(x,x) - \xi_i k_t(x)^\top (\xi_i K_t + \eta I_t)^{-1} k_t(x) \right). \tag{7}$$

Let us now discuss the time required to compute $\mu_t(x)$ and $\|\Gamma_t(x,x)\|$. Given the eigen decomposition, updating $\{Y_t^i\}_{i=1}^n$ re-using those already computed at the previous step requires projecting the current output y_t onto all coordinates, and thus, takes $O(n^2)$ time. Now, since the kernel matrix K_t is rescaled by the eigenvalues ξ_i, we can find the eigen decomposition of K_t once and reuse those to compute $\{(\xi_i K_t + \eta I_t)^{-1}\}_{i=1}^n$ in $O(t^3)$ time. Next, computing n matrix-vector multiplications and vector inner products of the form $k_t(x)^\top (\xi_i K_t + \eta I_t)^{-1} k_t(x)$ and $k_t(x)^\top (\xi_i K_t + \eta I_t)^{-1} Y_t^i$ take $O(nt^2)$ time. Finally, the sum in (6) and the max in (7) can be computed in $O(n^2)$ and $O(n)$ time, respectively. Therefore, the overall cost to compute $\mu_t(x)$ and $\|\Gamma_t(x,x)\|$ are $O\left(n^2 + nt^2 + t^3\right) = O\left(n^2 + t^2(n+t)\right)$.

Per-step Complexity of MT-BKB Let $\tilde{\varphi}_t(x) = \left(\tilde{K}_t^{1/2}\right)^+ \tilde{k}_t(x) \in \mathbb{R}^{m_t}$ denotes the Nyström embedding of the scalar kernel k, where $\tilde{k}_t(x) = \left[\frac{1}{\sqrt{p_{t,i_1}}} k(x_{i_1},x),\ldots,\frac{1}{\sqrt{p_{t,i_{m_t}}}} k(x_{i_{m_t}},x)\right]^\top$ and $\tilde{K}_t = \left[\frac{1}{\sqrt{p_{t,i_u} p_{t,i_v}}} k(x_{i_u},x_{i_v})\right]_{u,v=1}^{m_t}$. Then the eigen decomposition $B = \sum_{i=1}^n \xi_i u_i u_i^\top$ yields $\tilde{G}_t = \sum_{i=1}^n \xi_i \tilde{K}_t \otimes u_i u_i^\top$ and $\tilde{G}_t(x) = \sum_{i=1}^n \xi_i \tilde{k}_t(x) \otimes u_i u_i^\top$. A similar argument as in (4) and (5) now implies $\left(\tilde{G}_t^{1/2}\right)^+ = \sum_{i=1}^n \frac{1}{\sqrt{\xi_i}} \left(\tilde{K}_t^{1/2}\right)^+ \otimes u_i u_i^\top$. Therefore, the Nyström embeddings for the multi-task kernel Γ can be computed using the embeddings for the scalar kernel k as

$$\tilde{\Phi}_t(x) = \left(\tilde{G}_t^{1/2}\right)^+ \tilde{G}_t(x) = \sum_{i=1}^n \sqrt{\xi_i} \left(\tilde{K}_t^{1/2}\right)^+ \tilde{k}_t(x) \otimes u_i u_i^\top = \sum_{i=1}^n \sqrt{\xi_i} \tilde{\varphi}_t(x) \otimes u_i u_i^\top.$$

We now have

$$\tilde{V}_t = \sum_{s=1}^t \tilde{\Phi}_t(x_s) \tilde{\Phi}_t(x_s)^\top = \sum_{s=1}^t \sum_{i=1}^n \xi_i \tilde{\varphi}_t(x_s) \tilde{\varphi}_t(x_s)^\top \otimes u_i u_i^\top = \sum_{i=1}^n \xi_i \tilde{v}_t \otimes u_i u_i^\top,$$

where $\tilde{v}_t = \sum_{s=1}^{t} \tilde{\varphi}_t(x_s)\tilde{\varphi}_t(x_s)^\top$. A similar argument as in (4) and (5) then implies

$$(\tilde{V}_t + \eta I_{nm_t})^{-1} = \sum_{i=1}^{n} (\xi_i \tilde{v}_t + \eta I_{m_t})^{-1} \otimes u_i u_i^\top .$$

We further have

$$\sum_{s=1}^{t} \tilde{\Phi}_t(x_s) y_s = \sum_{s=1}^{t}\sum_{i=1}^{n} \sqrt{\xi_i} \cdot y_s^\top u_i \cdot \tilde{\varphi}_t(x_s) \otimes u_i = \sum_{i=1}^{n} \sqrt{\xi_i} \left(\sum_{s=1}^{t} y_s^\top u_i \cdot \tilde{\varphi}_t(x_s) \right) \otimes u_i .$$

Similar to (6), we therefore obtain

$$\tilde{\mu}_t(x) = \sum_{i=1}^{n} \xi_i \tilde{\varphi}_t(x)^\top \left(\xi_i \tilde{v}_t + \eta I_{m_t}\right)^{-1} \left(\sum_{s=1}^{t} y_s^\top u_i \cdot \tilde{\varphi}_t(x_s) \right) \cdot u_i . \tag{8}$$

We now note that $\tilde{\Phi}_t(x)^\top \tilde{\Phi}_t(x) = \sum_{i=1}^{n} \xi_i \tilde{\varphi}_t(x)^\top \tilde{\varphi}_t(x) \cdot u_i u_i^\top$. Similar to (7), we then obtain

$$\left\| \tilde{\Gamma}_t(x,x) \right\| = \max_{1 \leqslant i \leqslant n} \xi_i \left(k(x,x) - \tilde{\varphi}_t(x)^\top \tilde{\varphi}_t(x) + \eta \tilde{\varphi}_t(x)^\top \left(\xi_i \tilde{v}_t + \eta I_{m_t}\right)^{-1} \tilde{\varphi}_t(x) \right) . \tag{9}$$

We now discuss the time required to compute the scalar kernel embedding $\tilde{\varphi}_t(x)$. Sampling the dictionary $\mathcal{D}_t$, as we reuse the variances from the previous round, takes $O(t)$ time. We now compute the embedding $\tilde{\varphi}_t(x)$ in $O(m_t^3 + m_t^2)$ time, which corresponds to an inversion of $\tilde{K}_t^{1/2}$ and a matrix-vector product of dimension m_t, the size of the dictionary. Given the embedding function, let us now find the time required to compute $\tilde{\mu}_t(x)$ and $\|\tilde{\Gamma}_t(x,x)\|$. We first construct the matrix $\tilde{v}_t$ from scratch using all the points selected so far, which takes $O(m_t^2 t)$ time. Then the inverses $\{(\xi_i \tilde{v}_t + \eta I_{m_t})^{-1}\}_{i=1}^n$ can be computed in $O(m_t^3)$ time and the matrix-vector multiplications $\{(\xi_i \tilde{v}_t + \eta I_{m_t})^{-1} \tilde{\varphi}_t(x)\}_{i=1}^n$ in $O(nm_t^2)$ time. Similar to MT-KB, projecting the current output onto every direction takes $O(n^2)$ time. The projections can then be used to compute n vectors of the form $\sum_{s=1}^{t} y_s^\top u_i \cdot \tilde{\varphi}_t(x_s)$ in $O(nm_t t)$ time. Finally, n vector inner products of dimension m_t can be computed in $O(nm_t)$ time. Therefore, the overall cost to compute (8) and (9) is $O(n^2 + nm_t t + nm_t^2 + m_t^3 + m_t^2 t) = O\left(n^2 + m_t t(n+t)\right)$, since the dictionary size $m_t \leqslant t$.

B MULTI-TASK CONCENTRATION

We first introduce some notations. For any two Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$ with respective inner products $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we denote by $\mathcal{L}(\mathcal{G}, \mathcal{H})$ the space of all bounded linear operators from $\mathcal{G}$ to $\mathcal{H}$, with the operator norm $\|A\| := \sup_{\|g\|_{\mathcal{G}} \leqslant 1} \|Ag\|_{\mathcal{H}}$. We also denote, for any $A \in \mathcal{L}(\mathcal{G}, \mathcal{H})$, by $A^\top$ its adjoint, which is the unique operator such that $\langle A^\top h, g \rangle_{\mathcal{G}} = \langle h, Ag \rangle_{\mathcal{H}}$ for all $g \in \mathcal{G}$, $h \in \mathcal{H}$. In the case $\mathcal{G} = \mathcal{H}$, we denote $\mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{H}, \mathcal{H})$. We now state the following lemma about operators, which we will use several times.

Lemma 2 (Operator identities) *Let $A \in \mathcal{L}(\mathcal{G}, \mathcal{H})$. Then, for any $\eta > 0$, the following hold*

$$(A^\top A + \eta I)^{-1} A^\top = A^\top (AA^\top + \eta I)^{-1},$$
$$I - A^\top (AA^\top + \eta I)^{-1} A = \eta (A^\top A + \eta I)^{-1}.$$

We now present the main result of this appendix, which is stated and proved using the feature map of the multi-task kernel.

B.1 Feature Map of Multi-task Kernel

We assume the multi-task kernel Γ to be continuous relative to the operator norm on $\mathcal{L}(\mathbb{R}^n)$, the space of bounded linear operators from $\mathbb{R}^n$ to itself. Then the RKHS $\mathcal{H}_\Gamma(\mathcal{X})$ associated with the kernel Γ is a subspace of the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}^n$, and hence, Γ is a Mercer kernel Carmeli et al. (2010). Let μ be a probability measure on the (compact) set $\mathcal{X}$. Since Γ is a Mercer kernel on $\mathcal{X}$ and $\sup_{x \in \mathcal{X}} \|\Gamma(x,x)\| < \infty$, the RKHS $\mathcal{H}_\Gamma(\mathcal{X})$ is a subspace of $L^2(\mathcal{X}, \mu; \mathbb{R}^n)$, the Banach space of measurable functions $g : \mathcal{X} \to \mathbb{R}^n$ such that $\int_{\mathcal{X}} \|g(x)\|^2 d\mu(x) < \infty$, with norm $\|g\|_{L^2} = \left(\int_{\mathcal{X}} \|g(x)\|^2 d\mu(x) \right)^{1/2}$. Since $\Gamma(x,x) \in \mathcal{L}(\mathbb{R}^n)$ is a compact operator[6], by the Mercer theorem for multi-task

[6]An operator $A \in \mathcal{L}(\mathcal{H})$ is said to be compact if the image of each bounded set under A is relatively compact.

kernels Carmeli et al. (2010), there exists an at most countable sequence $\{(\psi_i, \nu_i)\}_{i\in\mathbb{N}}$ such that

$$\Gamma(x,x') = \sum_{i=1}^{\infty} \nu_i \psi_i(x)\psi_i(x')^\top \quad \text{and}$$
$$\|g\|_\Gamma^2 = \sum_{i=1}^{\infty} \frac{\langle g, \psi_i\rangle_{L^2}^2}{\nu_i}, \quad g \in L^2(\mathcal{X}, \mu; \mathbb{R}^n)\ ,$$

where $\nu_i \geqslant 0$ for all i, $\lim_{i\to\infty} \nu_i = 0$ and $\{\psi_i : \mathcal{X} \to \mathbb{R}^n\}_{i\in\mathbb{N}}$ is an orthonormal basis of $L^2(\mathcal{X}, \mu; \mathbb{R}^n)$. In particular $g \in \mathcal{H}_\Gamma(\mathcal{X})$ if and only if $\|g\|_\Gamma < \infty$. Note that $\{\sqrt{\nu_i}\psi_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_\Gamma(\mathcal{X})$. Then, we can represent the objective function $f \in \mathcal{H}_\Gamma(\mathcal{X})$ as

$$f = \sum_{i=1}^{\infty} \theta_i^\star \sqrt{\nu_i}\psi_i$$

for some $\theta^\star := (\theta_1^\star, \theta_2^\star, \ldots) \in \ell^2$, the Hilbert space of square-summable sequences of real numbers, such that $\|f\|_\Gamma = \|\theta^\star\|_2 := \left(\sum_{i=1}^{\infty} |\theta_i^\star|^2\right)^{1/2} < \infty$. We now define a feature map $\Phi : \mathcal{X} \to \mathcal{L}(\mathbb{R}^n, \ell^2)$ of the multi-task kernel Γ by

$$\Phi(x)y := \left(\sqrt{\nu_1}\psi_1(x)^\top y, \sqrt{\nu_2}\psi_2(x)^\top y, \ldots\right), \quad \forall\, x \in \mathcal{X},\ y \in \mathbb{R}^n.$$

We then have $f(x) = \Phi(x)^\top\theta^\star$ and $\Gamma(x,x') = \Phi(x)^\top\Phi(x')$ for all $x, x' \in \mathcal{X}$.

B.2 Martingale Control in ℓ^2 Space

Let us define $S_t = \sum_{s=1}^{t} \Phi(x_s)\varepsilon_s$, where $\varepsilon_1, \ldots, \varepsilon_t$ are the random noise vectors in $\mathbb{R}^n$. Now consider $\mathcal{F}_{t-1}$, the σ-algebra generated by the random variables $\{x_s, \varepsilon_s\}_{s=1}^{t-1}$ and x_t. Observe that S_t is $\mathcal{F}_t$-measurable and $\mathbb{E}\left[S_t \mid \mathcal{F}_{t-1}\right] = S_{t-1}$. The process $\{S_t\}_{t\geqslant 1}$ is thus a martingale with values[7] in the ℓ^2 space. We now define a map $\Phi_{\mathcal{X}_t} : \ell^2 \to \mathbb{R}^{nt}$ by

$$\Phi_{\mathcal{X}_t}\theta := \left[\left(\Phi(x_1)^\top\theta\right)^\top, \ldots, \left(\Phi(x_t)^\top\theta\right)^\top\right]^\top, \quad \forall\, \theta \in \ell^2.$$

We also let $V_t := \Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t}$ be a map from ℓ^2 to itself and I be the identity operator in ℓ^2. In Lemma 3, we measure the deviation of S_t by the norm weighted by $(V_t + \eta I)^{-1}$, which is itself derived from S_t. Lemma 3 represents the multi-task generalization of the result of Durand et al. (2018), and we recover their result under the single-task setting ($n = 1$).

Lemma 3 (Self-normalized martingale control) *Let the noise vectors $\{\varepsilon_t\}_{t\geqslant 1}$ be σ-sub-Gaussian. Then, for any $\eta > 0$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$, the following holds uniformly over all $t \geqslant 1$:*

$$\|S_t\|_{(V_t+\eta I)^{-1}} \leqslant \sigma\sqrt{2\log(1/\delta) + \log\det\left(I + \eta^{-1}V_t\right)}\ .$$

Proof For any sequence of real numbers $\theta = (\theta_1, \theta_2, \ldots)$ such that $\left\|\sum_{i=1}^{\infty} \theta_i\sqrt{\nu_i}\psi_i(x)\right\|_2 < \infty$, let us define $\Phi(x)^\top\theta := \sum_{i=1}^{\infty} \theta_i\sqrt{\nu_i}\psi_i(x)$ and

$$M_t^\theta = \prod_{s=1}^{t} D_s^\theta, \quad D_s^\theta = \exp\left(\frac{\varepsilon_s^\top\Phi(x_s)^\top\theta}{\sigma} - \frac{1}{2}\left\|\Phi(x_s)^\top\theta\right\|_2^2\right)\ .$$

Since the noise vectors $\{\varepsilon_t\}_{t\geqslant 1}$ are conditionally σ-sub-Gaussian, i.e.,

$$\forall \alpha \in \mathbb{R}^n, \forall t \geqslant 1, \quad \mathbb{E}\left[\exp(\varepsilon_t^\top\alpha) \mid \mathcal{F}_{t-1}\right] \leqslant \exp\left(\sigma^2\|\alpha\|_2^2/2\right),$$

we have $\mathbb{E}\left[D_t^\theta|\mathcal{F}_{t-1}\right] \leqslant 1$ and hence $\mathbb{E}\left[M_t^\theta|\mathcal{F}_{t-1}\right] \leqslant M_{t-1}^\theta$. Therefore, it is immediate that $\{M_t^\theta\}_{t=0}^{\infty}$ is a non-negative super-martingale and actually satisfies $\mathbb{E}\left[M_t^\theta\right] \leqslant 1$.

Now, let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. By the convergence theorem for non-negative super-martingales, $M_\infty^\theta = \lim_{t\to\infty} M_t^\theta$ is almost surely well-defined, and thus M_τ^θ is well-defined as well irrespective of whether $\tau < \infty$ or not. Let $Q_t^\theta = M_{\min\{\tau,t\}}^\theta$ be a stopped version of $\{M_t^\theta\}_t$. Then, by Fatou's lemma,

$$\mathbb{E}\left[M_\tau^\theta\right] = \mathbb{E}\left[\liminf_{t\to\infty} Q_t^\theta\right] \leqslant \liminf_{t\to\infty} \mathbb{E}\left[Q_t^\theta\right] = \liminf_{t\to\infty} \mathbb{E}\left[M_{\min\{\tau,t\}}^\theta\right] \leqslant 1\ , \tag{10}$$

since the stopped super-martingale $\left\{M_{\min\{\tau,t\}}^\theta\right\}_{t\geqslant 1}$ is also a super-martingale.

[7] We ignore issues of measurability here.

Let $\mathcal{F}_\infty$ be the σ-algebra generated by $\{\mathcal{F}_t\}_{t=0}^\infty$, and $\Theta = (\Theta_1, \Theta_2, \ldots)$, $\Theta_i \sim \mathcal{N}(0, 1/\eta)$ be an infinite i.i.d. Gaussian random sequence which is independent of $\mathcal{F}_\infty$. Since $\Gamma(x,x) \in \mathcal{L}(\mathbb{R}^n)$ has finite trace, we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^\infty \Theta_i \sqrt{\nu_i}\psi_i(x)\right\|_2^2\right] = \frac{1}{\eta}\sum_{i=1}^\infty \nu_i \left\|\psi_i(x)\right\|_2^2 = \frac{1}{\eta}\operatorname{tr}\left(\Gamma(x,x)\right) < \infty\ .$$

Therefore, $\left\|\sum_{i=1}^\infty \Theta_i \sqrt{\nu_i}\psi_i(x)\right\|_2 < \infty$ almost surely and thus M_t^Θ is well-defined. Now, thanks to the sub-Gaussian property, $\mathbb{E}\left[M_t^\Theta | \Theta\right] \leqslant 1$ almost surely, and thus $\mathbb{E}\left[M_t^\Theta\right] \leqslant 1$ for all t.

Let $M_t := \mathbb{E}\left[M_t^\Theta | \mathcal{F}_\infty\right]$ be a mixture of non-negative super-martingales M_t^Θ. Then $\{M_t\}_{t=0}^\infty$ is also a non-negative super-martingale adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Hence, by a similar argument as in (10), M_τ is almost surely well-defined and $\mathbb{E}\left[M_\tau\right] = \mathbb{E}\left[M_\tau^\Theta\right] \leqslant 1$. Let us now compute the mixture martingale M_t. We first note for any $\theta \in \ell^2$ that $M_t^\theta = \exp\left(\langle \theta, S_t/\sigma\rangle_2 - \frac{1}{2}\left\|\theta\right\|_{V_t}^2\right)$. The difficulty however lies in the handling of possibly infinite dimension. To this end, we follow Durand et al. (2018) to consider the first d dimensions for each $d \in \mathbb{N}$. Let Θ_d denote the restriction of Θ to the first d components. Thus $\Theta_d \sim \mathcal{N}(0, \frac{1}{\eta}I_d)$. Similarly, let $S_{t,d}$, $V_{t,d}$ and $M_{t,d}$ denote the corresponding restrictions of S_t, V_t and M_t, respectively. Following the steps from Chowdhury and Gopalan (2017), we then obtain that

$$\begin{aligned} M_{t,d} &= \frac{\det(\eta I_d)^{1/2}}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \exp\left(\langle \alpha, S_{t,d}/\sigma\rangle_2 - \frac{1}{2}\left\|\alpha\right\|_{V_{t,d}}^2\right)\exp\left(-\frac{\eta}{2}\left\|\alpha\right\|_2^2\right) d\alpha \\ &= \frac{1}{\det(I_d + \eta^{-1}V_{t,d})^{1/2}}\exp\left(\frac{1}{2\sigma^2}\left\|S_{t,d}\right\|_{(V_{t,d}+\eta I_d)^{-1}}^2\right). \end{aligned}$$

Note that $M_{\tau,d}$ is also almost surely well defined and $\mathbb{E}\left[M_{\tau,d}\right] \leqslant 1$ for all $d \in \mathbb{N}$. We now fix a $\delta \in (0,1]$. An application of Markov's inequality and Fatou's Lemma then yields

$$\begin{aligned} \mathbb{P}\left[\left\|S_\tau\right\|_{(V_\tau+\eta I)^{-1}}^2 > 2\sigma^2 \log\left(\frac{\det(I+\eta^{-1}V_\tau)^{1/2}}{\delta}\right)\right] &= \mathbb{P}\left[\frac{\exp\left(\frac{1}{2\sigma^2}\left\|S_\tau\right\|_{(V_\tau+\eta I)^{-1}}^2\right)}{\frac{1}{\delta}\det(I+\eta^{-1}V_\tau)^{1/2}} > 1\right] \\ &= \mathbb{P}\left[\lim_{d\to\infty}\frac{\exp\left(\frac{1}{2\sigma^2}\left\|S_{\tau,d}\right\|_{(V_{\tau,d}+\eta I_d)^{-1}}^2\right)}{\frac{1}{\delta}\det(I_d+\eta^{-1}V_{\tau,d})^{1/2}} > 1\right] \\ &\leqslant \mathbb{E}\left[\lim_{d\to\infty}\frac{\exp\left(\frac{1}{2\sigma^2}\left\|S_{\tau,d}\right\|_{(V_{\tau,d}+\eta I_d)^{-1}}^2\right)}{\frac{1}{\delta}\det(I_d+\eta^{-1}V_{\tau,d})^{1/2}}\right] \\ &\leqslant \delta \lim_{d\to\infty}\mathbb{E}\left[M_{\tau,d}\right] \leqslant \delta\ . \end{aligned}$$

We now define a random stopping time τ following Chowdhury and Gopalan (2017), by

$$\tau = \min\left\{t \geqslant 0 : \left\|S_t\right\|_{(V_t+\eta I)^{-1}}^2 > 2\sigma^2 \log\left(\frac{\det(I+\eta^{-1}V_t)^{1/2}}{\delta}\right)\right\}\ .$$

We then have

$$\mathbb{P}\left[\exists\, t \geqslant 1 :\ \left\|S_t\right\|_{(V_t+\eta I)^{-1}}^2 > 2\sigma^2 \log\left(\frac{\det(I+\eta^{-1}V_t)^{1/2}}{\delta}\right)\right] = \mathbb{P}\left[\tau < \infty\right] \leqslant \delta\ ,$$

which concludes the proof. ■

B.3 Concentration Bound for the Vector-valued Estimate (Proof of Theorem 1)

We first reformulate $\mu_t(x)$ in terms of the feature map $\Phi(x)$ as

$$\begin{aligned}
\mu_t(x) &= G_t(x)^\top \left(G_t + \eta I_{nt}\right)^{-1} Y_t \\
&= \Phi(x)^\top \Phi_{\mathcal{X}_t}^\top \left(\Phi_{\mathcal{X}_t} \Phi_{\mathcal{X}_t}^\top + \eta I_{nt}\right)^{-1} Y_t \\
&= \Phi(x)^\top \left(\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \eta I\right)^{-1} \Phi_{\mathcal{X}_t}^\top Y_t \\
&= \Phi(x)^\top \left(V_t + \eta I\right)^{-1} \sum_{s=1}^{t} \Phi(x_s) y_s \\
&= \Phi(x)^\top \left(V_t + \eta I\right)^{-1} \sum_{s=1}^{t} \Phi(x_s)(f(x_s) + \varepsilon_s) \\
&= \Phi(x)^\top \left(V_t + \eta I\right)^{-1} \sum_{s=1}^{t} \Phi(x_s) \left(\Phi(x_s)^\top \theta^\star + \varepsilon_s\right) \\
&= \Phi(x)^\top \theta^\star - \eta \Phi(x)^\top (V_t + \eta I)^{-1} \theta^\star + \Phi(x)^\top (V_t + \eta I)^{-1} S_t \\
&= f(x) + \Phi(x)^\top (V_t + \eta I)^{-1} \left(S_t - \eta \theta^\star\right),
\end{aligned}$$

where the third step follows from Lemma 2. We now obtain, from the definition of operator norm, the following

$$\begin{aligned}
\left\|f(x) - \mu_t(x)\right\|_2 &\leqslant \left\|\Phi(x)^\top (V_t + \eta I)^{-1/2}\right\| \left\|(V_t + \eta I)^{-1/2} \left(S_t - \eta \theta^\star\right)\right\|_2 \\
&\leqslant \left\|(V_t + \eta I)^{-1/2} \Phi(x)\right\| \left(\left\|S_t\right\|_{(V_t+\eta I)^{-1}} + \eta \left\|\theta^\star\right\|_{(V_t+\eta I)^{-1}}\right) \\
&\leqslant \left\|\Phi(x)^\top (V_t + \eta I)^{-1} \Phi(x)\right\|^{1/2} \left(\left\|S_t\right\|_{(V_t+\eta I)^{-1}} + \eta^{1/2} \left\|f\right\|_\Gamma\right),
\end{aligned}$$

where the last step is controlled as $\left\|\theta^\star\right\|_{(V_t+\eta I)^{-1}} \leqslant \eta^{-1/2} \left\|\theta^\star\right\|_2 = \eta^{-1/2} \left\|f\right\|_\Gamma$. A simple application of Lemma 2 now yields

$$\begin{aligned}
\eta \Phi(x)^\top (V_t + \eta I)^{-1} \Phi(x) &= \eta \Phi(x)^\top (\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \eta I)^{-1} \Phi(x) \\
&= \Phi(x)^\top \Phi(x) - \Phi(x)^\top \Phi_{\mathcal{X}_t}^\top (\Phi_{\mathcal{X}_t} \Phi_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} \Phi_{\mathcal{X}_t} \Phi(x) \\
&= \Gamma(x,x) - G_t(x)^\top (G_t + \eta I_{nt})^{-1} G_t(x) = \Gamma_t(x,x). \qquad (11)
\end{aligned}$$

We then have $\left\|\Phi(x)^\top (V_t + \eta I)^{-1} \Phi(x)\right\|^{1/2} = \eta^{-1/2} \left\|\Gamma_t(x,x)\right\|^{1/2}$. We conclude the proof from Lemma 3 and using Sylvester's identity to get

$$\det\left(I + \eta^{-1} V_t\right) = \det\left(I + \eta^{-1} \Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t}\right) = \det\left(I_{nt} + \eta^{-1} \Phi_{\mathcal{X}_t} \Phi_{\mathcal{X}_t}^\top\right) = \det\left(I_{nt} + \eta^{-1} G_t\right). \qquad (12)$$

C REGRET ANALYSIS OF MT-KB

C.1 Properties of Multi-task Predictive GP Variance

Lemma 4 (Sum of predictive variances) *For any $\eta > 0$ and $t \geqslant 1$,*

$$\frac{1}{\eta}\sum_{s=1}^{t} \operatorname{tr}\left(\Gamma_s(x_s, x_s)\right) = \log\det\left(I_{nt} + \eta^{-1} G_t\right) = \sum_{s=1}^{t} \log\det\left(I_n + \eta^{-1}\Gamma_{s-1}(x_s, x_s)\right).$$

Proof For the first part, we observe from (11) that

$$\begin{aligned}
\frac{1}{\eta}\sum_{s=1}^{t} \operatorname{tr}\left(\Gamma_s(x_s, x_s)\right) &= \sum_{s=1}^{t} \operatorname{tr}\left(\Phi(x_s)^\top (V_s + \eta I)^{-1}\Phi(x_s)\right) \\
&= \sum_{s=1}^{t} \operatorname{tr}\left((V_s + \eta I)^{-1}\Phi(x_s)\Phi(x_s)^\top\right) \\
&= \sum_{s=1}^{t} \operatorname{tr}\left((V_s + \eta I)^{-1}\left((V_s + \eta I) - (V_{s-1} + \eta I)\right)\right) \\
&\leqslant \sum_{s=1}^{t} \log\left(\frac{\det(V_s + \eta I)}{\det(V_{s-1} + \eta I)}\right) \\
&= \log\det\left(I + \eta^{-1} V_t\right) = \log\det\left(I_{nt} + \eta^{-1} G_t\right).
\end{aligned}$$

Here, the last equality follows from (12). The inequality follows from the fact that for two p.d. matrices A and B such that $A - B$ is p.s.d., $\operatorname{tr}\left(A^{-1}(A - B)\right) \leqslant \log\left(\frac{\det(A)}{\det(B)}\right)$ Calandriello et al. (2019).

For the second part, we obtain from Schur's determinant identity that

$$\begin{aligned}
&\det\left(I_{nt} + \eta^{-1} G_t\right) \\
&= \det\left(I_{n(t-1)} + \eta^{-1} G_{t-1}\right) \times \\
&\qquad \det\left(I_n + \eta^{-1}\Gamma(x_t, x_t) - \eta^{-1} G_{t-1}(x_t)^\top \left(I_{n(t-1)} + \eta^{-1} G_{t-1}\right)^{-1} \eta^{-1} G_{t-1}(x_t)\right) \\
&= \det\left(I_{n(t-1)} + \eta^{-1} G_{t-1}\right) \det\left(I_n + \eta^{-1}\Gamma_{t-1}(x_t, x_t)\right) \\
&= \ldots \\
&= \prod_{s=1}^{t} \det\left(I_n + \eta^{-1}\Gamma_{s-1}(x_s, x_s)\right).
\end{aligned}$$

We conclude the proof by applying logarithm on both sides. ■

Lemma 5 (Predictive variance geometry) *Let $\|\Gamma(x, x)\| \leqslant \kappa$. Then, for any $\eta > 0$ and $t \geqslant 1$,*

$$\Gamma_t(x, x) \preceq \Gamma_{t-1}(x, x) \preceq (1 + \kappa/\eta)\, \Gamma_t(x, x).$$

Proof Let us define $\overline{V}_t = V_t + \eta I$ for all $t \geqslant 0$. We then have from (11) that

$$\begin{aligned}
\Gamma_t(x, x) &= \eta\Phi(x)^\top \overline{V}_t^{-1}\Phi(x) \\
&= \eta\Phi(x)^\top \left(\overline{V}_{t-1} + \Phi(x_t)\Phi(x_t)^\top\right)^{-1}\Phi(x) \\
&= \eta\Phi(x)^\top \overline{V}_{t-1}^{-1}\Phi(x) - \\
&\qquad \eta\Phi(x)^\top \overline{V}_{t-1}^{-1}\Phi(x_t)\left(I_n + \Phi(x_t)^\top \overline{V}_{t-1}^{-1}\Phi(x_t)\right)^{-1}\Phi(x_t)^\top \overline{V}_{t-1}^{-1}\Phi(x) \\
&= \Gamma_{t-1}(x, x) - \eta^{-1}\Gamma_{t-1}(x_t, x)^\top \left(I_n + \eta^{-1}\Gamma_{t-1}(x_t, x_t)\right)^{-1}\Gamma_{t-1}(x_t, x) \\
&\preceq \Gamma_{t-1}(x, x).
\end{aligned}$$

Here in the third step, we have used the Sherman-Morrison formula and in the last step, we have used the positive semi-definite property of multi-task kernels. To prove the second part, we first note that

$$\begin{aligned}\frac{1}{\eta}\Gamma_t(x,x) &= \Phi(x)^\top \left(\overline{V}_{t-1} + \Phi(x_t)\Phi(x_t)^\top\right)^{-1} \Phi(x)\\ &= \Phi(x)^\top \overline{V}_{t-1}^{-1/2} \left(I + \overline{V}_{t-1}^{-1/2}\Phi(x_t)\Phi(x_t)^\top \overline{V}_{t-1}^{-1/2}\right)^{-1} \overline{V}_{t-1}^{-1/2}\Phi(x)\ . \end{aligned} \tag{13}$$

Further, since $\|\Gamma(x,x)\| \leqslant \kappa$, we have $\lambda_{\max}\left(\Gamma(x,x)\right) \leqslant \kappa$, and hence,

$$\Gamma_t(x,x) \preceq \Gamma_{t-1}(x,x) \preceq \Gamma_{t-2}(x,x) \preceq \ldots \Gamma_0(x,x) = \Gamma(x,x) \preceq \kappa I_n. \tag{14}$$

Since $\overline{V}_{t-1}^{-1/2}\Phi(x_t)\Phi(x_t)^\top \overline{V}_{t-1}^{-1/2}$ and $\Phi(x_t)^\top \overline{V}_{t-1}^{-1}\Phi(x_t)$ have same set of non-zero eigenvalues, we now obtain from (14) that $\overline{V}_{t-1}^{-1/2}\Phi(x_t)\Phi(x_t)^\top \overline{V}_{t-1}^{-1/2} \preceq \frac{\kappa}{\eta} I$. Then (13) implies that

$$\Gamma_t(x,x) \succeq \eta\ \Phi(x)^\top \overline{V}_{t-1}^{-1}\Phi(x)/\left(1+\kappa/\eta\right) = \Gamma_{t-1}(x,x)/\left(1+\kappa/\eta\right),$$

which completes the proof. ■

C.2 Regret Bound for MT-KB (Proof of Theorem 2)

Since the scalarization functions s_λ is L-Lipschitz in the ℓ_2 norm, we have

$$\left|s_\lambda\left(f(x)\right) - s_\lambda\left(\mu_{t-1}(x)\right)\right| \leqslant L \left\|f(x) - \mu_{t-1}(x)\right\|_2 .$$

Since $\mu_0(x) = 0$, $\Gamma_0(x,x) = \Gamma(x,x)$ and $\|f\|_\Gamma \leqslant b$, we have

$$\forall \lambda \in \Lambda, \quad \left\|f(x) - \mu_0(x)\right\|_2 = \left\|\Gamma_x^\top f\right\|_2 \leqslant \|f\|_\Gamma \left\|\Gamma_x\right\| = \|f\|_\Gamma \left\|\Gamma_x^\top \Gamma_x\right\|^{1/2} \leqslant b \left\|\Gamma_0(x,x)\right\|^{1/2}.$$

Then, from Theorem 1 and Lemma 4, the following holds with probability at least $1-\delta$:

$$\forall t \geqslant 1, \forall x \in \mathcal{X}, \forall \lambda \in \Lambda, \quad \left|s_\lambda\left(f(x)\right) - s_\lambda\left(\mu_{t-1}(x)\right)\right| \leqslant L\beta_{t-1}\left\|\Gamma_{t-1}(x,x)\right\|^{1/2}, \tag{15}$$

where $\beta_t = b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\log(1/\delta) + \sum_{s=1}^t \log\det\left(I_n + \eta^{-1}\Gamma_{s-1}(x_s,x_s)\right)}$, $t \geqslant 0$. We can now upper bound the *instantaneous regret* at time $t \geqslant 1$ as

$$\begin{aligned} r_t &:= \mathbb{E}\left[s_\lambda\left(f(x^\star)\right)\right] - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\ &\leqslant \mathbb{E}\left[s_\lambda\left(\mu_{t-1}(x^\star)\right)\right] + L\beta_{t-1}\left\|\Gamma_{t-1}(x^\star,x^\star)\right\|^{1/2} - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\ &\leqslant \mathbb{E}\left[s_\lambda\left(\mu_{t-1}(x_t)\right)\right] + L\beta_{t-1}\left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2} - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\ &\leqslant 2L\beta_{t-1}\left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2}. \end{aligned}$$

Here in the first and third step, we have used (15). The second step follows from the choice of x_t. Since β_t is a monotonically increasing function in t, we have the cumulative regret

$$R_C(T) := \sum_{t=1}^T r_t \leqslant 2L\beta_T \sum_{t=1}^T \left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2} \leqslant 2L\beta_T\sqrt{(1+\kappa/\eta)T\sum_{t=1}^T \left\|\Gamma_t(x_t,x_t)\right\|},$$

where the last step is due to the Cauchy-Schwartz inequality and Lemma 5. We now obtain from Lemma 4 that $\beta_T \leqslant b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\left(\log(1/\delta) + \gamma_{nT}(\Gamma,\eta)\right)}$, which concludes the proof.

C.3 Inter-task Structure in Regret for Separable Kernels (Proof of Lemma 1)

For separable multi-task kernels $\Gamma(x,x') = k(x,x')B$, the kernel matrix is given by $G_T = K_T \otimes B$, where K_T is kernel matrix corresponding to the scalar kernel k and $\otimes$ denotes the Kronecker product. Let $\{\alpha_t\}_{t=1}^T$ denote the eigenvalues of K_T. Then the eigenvalues of G_T are given by $\alpha_t\xi_i$, $1 \leqslant t \leqslant T$, $1 \leqslant i \leqslant n$, where ξ_i's are the eigenvalues of B. We now have

$$\log\det(I_{nT}+\eta^{-1}G_T)=\sum_{t=1}^{T}\sum_{i=1}^{n}\log(1+\alpha_t\xi_i/\eta)$$
$$=\sum_{i\in[n]:\xi_i>0}\sum_{t=1}^{T}\log(1+\alpha_t\xi_i/\eta)$$
$$=\sum_{i\in[n]:\xi_i>0}\log\det\left(I_T+(\eta/\xi_i)^{-1}K_T\right).$$

Taking supremum over all possible subsets $\mathcal{X}_T$ of $\mathcal{X}$, we then obtain that $\gamma_{nT}(\Gamma,\eta)\leqslant\sum_{i\in[n]:\xi_i>0}\gamma_T(k,\eta/\xi_i)$.

To prove the second part, we use the feature representation of the scalar kernel k. To this end, we let $\varphi:\mathcal{X}\to\ell^2$ be a feature map of the scalar kernel k, so that $k(x,x')=\varphi(x)^\top\varphi(x')$ for all $x,x'\in\mathcal{X}$. We now define a map $\varphi_{\mathcal{X}_t}:\ell^2\to\mathbb{R}^t$ by

$$\varphi_{\mathcal{X}_t}\theta:=\left[\varphi(x_1)^\top\theta,\ldots,\varphi(x_t)^\top\theta\right]^\top,\quad\forall\,\theta\in\ell^2.$$

We also let $v_t:=\varphi_{\mathcal{X}_t}^\top\varphi_{\mathcal{X}_t}$ be a map from ℓ^2 to itself. For any $\alpha>0$, we then obtain from Lemma 2 that

$$\alpha\,\varphi(x)^\top(v_t+\alpha I)^{-1}\varphi(x)=\alpha\,\varphi(x)^\top(\varphi_{\mathcal{X}_t}^\top\varphi_{\mathcal{X}_t}+\alpha I)^{-1}\varphi(x)$$
$$=\varphi(x)^\top\varphi(x)-\varphi(x)^\top\varphi_{\mathcal{X}_t}^\top(\varphi_{\mathcal{X}_t}\varphi_{\mathcal{X}_t}^\top+\alpha I_t)^{-1}\varphi_{\mathcal{X}_t}\varphi(x)$$
$$=k(x,x)-k_t(x)^\top(K_t+\alpha I_t)^{-1}k_t(x),$$

where $k_t(x)=[k(x_1,x),\ldots,k(x_t,x)]^\top$ and $K_t=[k(x_i,x_j)]_{1,j=1}^{t}$. We then have from (7) that

$$\|\Gamma_t(x,x)\|=\max_{1\leqslant i\leqslant n}\xi_i\left(k(x,x)-k_t(x)^\top\left(K_t+\frac{\eta}{\xi_i}I_t\right)^{-1}k_t(x)\right)$$
$$=\max_{1\leqslant i\leqslant n}\xi_i\cdot\frac{\eta}{\xi_i}\varphi(x)^\top\left(v_t+\frac{\eta}{\xi_i}I\right)^{-1}\varphi(x)$$
$$\leqslant\eta\,\varphi(x)^\top\left(v_t+\frac{\eta}{\kappa}I\right)^{-1}\varphi(x).$$

Here, in the last step we have used that $\xi_i\leqslant\kappa$ for all $i\in[n]$. This holds from our hypothesis $\|\Gamma(x,x)\|\leqslant\kappa$ and $k(x,x)=1$. We now observe that $\left(v_t+\frac{\eta}{\kappa}I\right)^{-1}\preceq(v_t+\eta I)^{-1}$ for $\kappa\leqslant1$ and $\left(v_t+\frac{\eta}{\kappa}I\right)^{-1}\preceq\kappa\,(v_t+\eta I)^{-1}$ for $\kappa\geqslant1$. Therefore

$$\|\Gamma_t(x,x)\|\leqslant\eta\max\{\kappa,1\}\varphi(x)^\top\left(v_t+\eta I\right)^{-1}\varphi(x).$$

A simple application of Lemma 4 for $n=1$ and $\Gamma(\cdot,\cdot)=k(\cdot,\cdot)$ now yields

$$\sum_{t=1}^{T}\|\Gamma_t(x,x)\|\leqslant\eta\max\{\kappa,1\}\sum_{t=1}^{T}\varphi(x_t)^\top\left(v_t+\eta I\right)^{-1}\varphi(x_t)$$
$$=\eta\max\{\kappa,1\}\log\det\left(I_T+\eta^{-1}K_T\right)\leqslant2\eta\max\{\kappa,1\}\gamma_T(k,\eta),$$

which completes the proof.

C.4 Inter-task Structure in Regret for Sum of Separable Kernels

We now present a generalization of Lemma 1 for multi-task kernels of the form $\Gamma(x,x')=\sum_{j=1}^{M}k_j(x,x')B_j$. This class of kernels is called the sum of separable (SoS) kernel and includes the diagonal kernel $\Gamma(x,x')=\mathrm{diag}\left(k_1(x,x'),\ldots,k_n(x,x')\right)$ as a special case.

Lemma 6 (Inter-task structure in regret for SoS kernel) *Let $\Gamma(x,x')=\sum_{j=1}^{M}k_j(x,x')B_j$ and $B_j\in\mathbb{R}^{n\times n}$ be positive semi-definite. Then*

$$\gamma_{nT}(\Gamma,\eta)\leqslant\sum_{j=1}^{M}\rho_{B_j}\max\{\xi_{B_j},1\}\gamma_T(k_j,\eta),\quad\sum_{t=1}^{T}\|\Gamma_t(x_t,x_t)\|\leqslant2\eta\sum_{j=1}^{M}\max\{\xi_{B_j},1\}\gamma_T(k_j,\eta),$$

where ρ_{B_j} and ξ_{B_j} denote the rank and the maximum eigenvalue of B_j, respectively and $\gamma_T(k_j)$ is the maximum information gain corresponding to scalar kernel k_j. Moreover, if $\Gamma(x,x')=\mathrm{diag}\left(k_1(x,x'),\ldots,k_n(x,x')\right)$ and each k_j is a stationary *kernel, then*

$$\gamma_{nT}(\Gamma,\eta) \leqslant \sum_{j=1}^{n} \gamma_T(k_j,\eta), \qquad \sum_{t=1}^{T} \|\Gamma_t(x_t,x_t)\| \leqslant 2\eta \max_{1\leqslant j\leqslant n} \gamma_T(k_j,\eta).$$

Proof We let, for each scalar kernel k_j, a feature map $\varphi_j : \mathcal{X} \to \ell^2$, so that $k_j(x,x') = \varphi_j(x)^\top \varphi_j(x')$. We now define the feature map $\Phi : \mathcal{X} \to \mathcal{L}(\mathbb{R}^n, \ell^2)$ of the multi-task kernel $\Gamma(x,x') = \sum_{j=1}^{M} k_j(x,x')B_j$ by

$$\Phi(x)y := \left(\varphi_1(x)\otimes B_1^{1/2}y,\ldots,\varphi_M(x)\otimes B_M^{1/2}y\right), \quad \forall\, x\in\mathcal{X},\ y\in\mathbb{R}^n\ ,$$

with the inner product

$$\Phi(x)^\top\Phi(x') := \sum_{j=1}^{M}\left(\varphi_j(x)\otimes B_j^{1/2}\right)^\top\left(\varphi_j(x')\otimes B_j^{1/2}\right) = \sum_{j=1}^{M}\varphi_j(x)^\top\varphi_j(x')\cdot B_j.$$

We then have

$$V_t := \sum_{s=1}^{t}\Phi(x_s)\Phi(x_s)^\top = \sum_{s=1}^{t}\sum_{j=1}^{M}\varphi_j(x_s)\varphi_j(x_s)^\top\otimes B_j = \sum_{j=1}^{M} v_{t,j}\otimes B_j,$$

where $v_{t,j} := \sum_{s=1}^{t}\varphi_j(x_s)\varphi_j(x_s)^\top$. We further obtain from (11) that

$$\Gamma_t(x,x) = \sum_{j=1}^{M}\eta\left(\varphi_j(x)\otimes B_j^{1/2}\right)^\top\left(\sum_{j=1}^{M} v_{t,j}\otimes B_j + \eta I\right)^{-1}\left(\varphi_j(x)\otimes B_j^{1/2}\right).$$

Now each B_j is a positive semi-definite matrix and so is $v_{t,j}\otimes B_j$. Hence, for for all $j\in[M]$, $\left(\sum_{j=1}^{M} v_{t,j}\otimes B_j + \eta I\right)^{-1} \preceq (v_{t,j}\otimes B_j+\eta I)^{-1}$. Therefore

$$\Gamma_t(x,x) \preceq \sum_{j=1}^{M}\eta\left(\varphi_j(x)\otimes B_j^{1/2}\right)^\top(v_{t,j}\otimes B_j+\eta I)^{-1}\left(\varphi_j(x)\otimes B_j^{1/2}\right) = \sum_{j=1}^{M}\Gamma_{t,j}(x,x), \tag{16}$$

where $\Gamma_{t,j}(x,x) := \eta\left(\varphi_j(x)\otimes B_j^{1/2}\right)^\top(v_{t,j}\otimes B_j+\eta I)^{-1}\left(\varphi_j(x)\otimes B_j^{1/2}\right)$. Now, let $(\xi_{j,i},u_{j,i})$ denotes the i-th eigen-pair of B_j. A similar argument as in (5) then yields

$$(v_{t,j}\otimes B_j+\eta I)^{-1} = \sum_{i=1}^{n}(\xi_{j,i}v_{t,j}+\eta I)^{-1}\otimes u_{j,i}u_{j,i}^\top\ .$$

We then have from the mixed product property of Kronecker product and the orthonormality of $\{u_{j,i}\}_{i=1}^{n}$ that

$$\begin{aligned}\Gamma_{t,j}(x,x) &= \sum_{i=1}^{n}\eta\,\xi_{j,i}\varphi_j(x)^\top(\xi_{j,i}v_{t,j}+\eta I)^{-1}\varphi_j(x)\cdot u_{j,i}u_{j,i}^\top\\ &= \sum_{i=1}^{n}\eta\,\varphi_j(x)^\top\left(v_{t,j}+\frac{\eta}{\xi_{j,i}}I\right)^{-1}\varphi_j(x)\cdot u_{j,i}u_{j,i}^\top\ .\end{aligned}$$

Since $\left(v_t+\frac{\eta}{\xi_{j,i}}I\right)^{-1}\preceq(v_t+\eta I)^{-1}$ for $\xi_{j,i}\leqslant 1$ and $\left(v_t+\frac{\eta}{\xi_{j,i}}I\right)^{-1}\preceq\xi_{j,i}\,(v_t+\eta I)^{-1}$ for $\xi_{j,i}\geqslant 1$, we now have

$$\begin{aligned}\mathrm{tr}(\Gamma_{t,j}(x,x)) &\leqslant \eta\sum_{i\in[n]:\xi_{j,i}>0}\max\{\xi_{j,i},1\}\varphi_j(x)^\top(v_{t,j}+\eta I)^{-1}\varphi_j(x)\\ &\leqslant \eta\,\rho_{B_j}\max\{\xi_{B_j},1\}\varphi_j(x)^\top(v_{t,j}+\eta I)^{-1}\varphi_j(x).\end{aligned}$$

Similarly

$$\begin{aligned}\|\Gamma_{t,j}(x,x)\| &\leqslant \eta\max_{1\leqslant i\leqslant n}\max\{\xi_{j,i},1\}\varphi_j(x)^\top(v_{t,j}+\eta I)^{-1}\varphi_j(x)\\ &\leqslant \eta\max\{\xi_{B_j},1\}\varphi_j(x)^\top(v_{t,j}+\eta I)^{-1}\varphi_j(x).\end{aligned}$$

Let $K_{T,j} = [k_j(x_p,x_q)]_{p,q=1}^{T}$ denotes the kernel matrix corresponding to the scalar kernel k_j. An application of Lemma 4 for $n = 1$ and $\Gamma(\cdot,\cdot) = k_j(\cdot,\cdot)$ now yields

$$\sum_{t=1}^{T} \operatorname{tr}\left(\Gamma_{t,j}(x_t, x_t)\right) \leqslant \eta\, \rho_{B_j} \max\{\xi_{B_j}, 1\} \log\det\left(I_T + \eta^{-1} K_{T,j}\right) \quad \text{and}$$

$$\sum_{t=1}^{T} \left\| \Gamma_{t,j}(x_t, x_t) \right\| \leqslant \eta\, \max\{\xi_{B_j}, 1\} \log\det\left(I_T + \eta^{-1} K_{T,j}\right).$$

We then have from (16) and Lemma 4 that

$$\begin{aligned}
\log\det\left(I_{nT} + \eta^{-1} G_T\right) &= \frac{1}{\eta} \sum_{t=1}^{T} \operatorname{tr}\left(\Gamma_t(x_t, x_t)\right) \\
&\leqslant \frac{1}{\eta} \sum_{j=1}^{M} \sum_{t=1}^{T} \operatorname{tr}\left(\Gamma_{t,j}(x_t, x_t)\right) \\
&\leqslant \sum_{j=1}^{M} \rho_{B_j} \max\{\xi_{B_j}, 1\} \log\det\left(I_T + \eta^{-1} K_{T,j}\right).
\end{aligned}$$

Taking supremum over all possible subsets $\mathcal{X}_T$ of $\mathcal{X}$, we now obtain that $\gamma_{nT}(\Gamma, \eta) \leqslant \sum_{j=1}^{M} \rho_{B_j} \max\{\xi_{B_j}, 1\} \gamma_T(k_j, \eta)$. We further have from (16) that

$$\sum_{t=1}^{T} \left\| \Gamma_t(x_t, x_t) \right\| \leqslant \sum_{j=1}^{M} \sum_{t=1}^{T} \left\| \Gamma_{t,j}(x_t, x_t) \right\| \leqslant 2\eta \sum_{j=1}^{M} \max\{\xi_{B_j}, 1\} \gamma_T(k_j, \eta),$$

which completes the proof for the first part.

For the diagonal kernel, $M = n$ and each B_j is a diagonal matrix with 1 in the j-th diagonal entry and 0 in all others. In this case, we have

$$\Gamma_t(x, x) = \eta \sum_{j=1}^{n} \varphi_j(x)^\top \left(v_{t,j} + \eta I\right)^{-1} \varphi_j(x) \cdot B_j\ .$$

We then have from Lemma 4 that

$$\begin{aligned}
\log\det\left(I_{nT} + \eta^{-1} G_T\right) &= \frac{1}{\eta} \sum_{t=1}^{T} \operatorname{tr}\left(\Gamma_t(x_t, x_t)\right) \\
&= \sum_{t=1}^{T} \sum_{j=1}^{n} \varphi_j(x_t)^\top \left(v_{t,j} + \eta I\right)^{-1} \varphi_j(x_t) \cdot \operatorname{tr}\left(B_j\right) \\
&= \sum_{j=1}^{n} \sum_{t=1}^{T} \varphi_j(x_t)^\top \left(v_{t,j} + \eta I\right)^{-1} \varphi_j(x_t) \\
&= \sum_{j=1}^{n} \log\det\left(I_T + \eta^{-1} K_{T,j}\right).
\end{aligned}$$

Taking supremum over all possible subsets $\mathcal{X}_T$ of $\mathcal{X}$, we now obtain that $\gamma_{nT}(\Gamma, \eta) \leqslant \sum_{j=1}^{n} \gamma_T(k_j, \eta)$. We further have

$$\left\| \Gamma_t(x, x) \right\| = \max_{1 \leqslant j \leqslant n} \eta\, \varphi_j(x)^\top \left(v_{t,j} + \eta I\right)^{-1} \varphi_j(x)\ .$$

Let $j^\star(x) = \operatorname{argmax}_{1 \leqslant j \leqslant n} k_j(x, x)$. Since each k_j is stationary, i.e., $k_j(x, x') = k_j(x - x')$, we have $j^\star(x)$ is independent of x. We now let $j^\star = j^\star(x)$ for all x. Then it can be easily checked that

$$\left\| \Gamma_t(x, x) \right\| = \eta\, \varphi_{j^\star}(x)^\top \left(v_{t,j^\star} + \eta I\right)^{-1} \varphi_{j^\star}(x)\ .$$

We now obtain from Lemma 4 that

$$\begin{aligned}
\sum_{t=1}^{T} \left\| \Gamma_t(x_t, x_t) \right\| &= \eta \sum_{t=1}^{T} \varphi_{j^\star}(x_t)^\top \left(v_{t,j^\star} + \eta I\right)^{-1} \varphi_{j^\star}(x_t) \\
&= \eta\, \log\det\left(I_T + \eta^{-1} K_{T,j^\star}\right) \leqslant 2\eta \max_{1 \leqslant j \leqslant n} \gamma_T(k_j, \eta)\ ,
\end{aligned}$$

which completes the proof for the second part. ■

D ANALYSIS OF MT-BKB

D.1 Trading-off Approximation Accuracy and Size

Given a dictionary $\mathcal{D}_t = \{x_{i_1}, \ldots, x_{i_{m_t}}\}$, we define a map $\Phi_{\mathcal{D}_t} : \ell^2 \to \mathbb{R}^{nm_t}$ by

$$\Phi_{\mathcal{D}_t}\theta := \left[\frac{1}{\sqrt{p_{t,i_1}}}\left(\Phi(x_{i_1})^\top\theta\right)^\top, \ldots, \frac{1}{\sqrt{p_{t,i_{m_t}}}}\left(\Phi(x_{i_{m_t}})^\top\theta\right)^\top\right]^\top, \quad \forall\, \theta \in \ell^2, \tag{17}$$

where $p_{t,i_j} = \min\left\{q\left\|\tilde{\Gamma}_{t-1}(x_{i_j}, x_{i_j})\right\|, 1\right\}$ for all $j \in [m_t]$.

Lemma 7 (Approximation properties) *For any $T \geqslant 1$, $\varepsilon \in (0,1)$ and $\delta \in (0,1]$, set $\rho = \frac{1+\varepsilon}{1-\varepsilon}$ and $q = \frac{6\rho\ln(2T/\delta)}{\varepsilon^2}$. Then, for any $\eta > 0$, with probability at least $1-\delta$, the following hold uniformly over all $t \in [T]$:*

$$(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t} - \varepsilon\eta I \preceq \Phi_{\mathcal{D}_t}^\top\Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t} + \varepsilon\eta I\ ,$$

$$m_t \leqslant 6\rho q\,(1+\kappa/\eta)\sum_{s=1}^{t}\left\|\Gamma_s(x_s, x_s)\right\|.$$

Proof Let S_t be an nt-by-nt block diagonal matrix with i-th diagonal block $[S_t]_i = \frac{1}{\sqrt{p_{t,i}}}I_n$ if $x_i \in \mathcal{D}_t$, and $[S_t]_i = 0$ if $x_i \notin \mathcal{D}_t$, $1 \leqslant i \leqslant t$. We then have $\Phi_{\mathcal{D}_t}^\top\Phi_{\mathcal{D}_t} = \Phi_{\mathcal{X}_t}^\top S_t^\top S_t\Phi_{\mathcal{X}_t}$. The proof now can be completed by following Calandriello et al. (2019, Theorem 1). ∎

Remark 5 *Note that although tuning the approximation trade-off parameter q requires the knowledge of the time horizon T in advance, Lemma 7 is quite robust to the uncertainty on T. If the horizon is not known, then after the T-th step, one can increase q according to the new desired horizon, and update the dictionary with this new value of q. Combining this with a standard doubling trick preserve the approximation properties Calandriello et al. (2019).*

Constructing Approximating Confidence Sets We now focus on the dictionary $\mathcal{D}_t$ chosen by MT-BKB at each step and discuss a principled approach to compute the approximations $\tilde{\mu}_t(x)$ and $\tilde{\Gamma}_t(x,x)$. To this end, we let

$$P_t = \Phi_{\mathcal{D}_t}^\top\left(\Phi_{\mathcal{D}_t}\Phi_{\mathcal{D}_t}^\top\right)^+\Phi_{\mathcal{D}_t} \tag{18}$$

denote the symmetric orthogonal projection operator on the subspace of $\mathcal{L}\left(\mathbb{R}^n, \ell^2\right)$ that is spanned by $\Phi(x_{i_1}), \ldots, \Phi(x_{i_{m_t}})$. We also let $\widehat{\Phi}_t(x) = P_t\Phi(x)$ denote the projection of $\Phi(x)$. We now define a map $\widehat{\Phi}_{\mathcal{X}_t} : \ell^2 \to \mathbb{R}^{nt}$ by

$$\widehat{\Phi}_{\mathcal{X}_t}\theta := \left[\left(\widehat{\Phi}_t(x_1)^\top\theta\right)^\top, \ldots, \left(\widehat{\Phi}_t(x_t)^\top\theta\right)^\top\right]^\top, \quad \forall\, \theta \in \ell^2.$$

We then have $\widehat{\Phi}_{\mathcal{X}_t} = \Phi_{\mathcal{X}_t}P_t$ and $\widehat{\Phi}_{\mathcal{X}_t}\widehat{\Phi}_{\mathcal{X}_t}^\top = \Phi_{\mathcal{X}_t}P_t\Phi_{\mathcal{X}_t}^\top$.

Lemma 8 (Approximation as given by projection) *For any $\eta > 0$ and $t \geqslant 1$, we have*

$$\tilde{\mu}_t(x) = \Phi(x)^\top\left(\widehat{V}_t + \eta I\right)^{-1}\sum_{s=1}^{t}\widehat{\Phi}_t(x_s)y_s \quad \textit{and} \quad \tilde{\Gamma}_t(x,x) = \eta\Phi(x)^\top\left(\widehat{V}_t + \eta I\right)^{-1}\Phi(x),$$

where $\widehat{V}_t := \widehat{\Phi}_{\mathcal{X}_t}^\top\widehat{\Phi}_{\mathcal{X}_t}$.

Proof We first note that

$$\tilde{\Phi}_t(x)^\top\tilde{\Phi}_t(x') = \tilde{G}_t(x)^\top\tilde{G}_t^+\tilde{G}_t(x') = \Phi(x)^\top P_t\Phi(x').$$

We now define an $nt \times nm_t$ matrix $\tilde{\Phi}_{\mathcal{X}_t} = \left[\tilde{\Phi}_t(x_1), \ldots, \tilde{\Phi}_t(x_t)\right]^\top$. We then have

$$\tilde{\Phi}_{\mathcal{X}_t}\tilde{\Phi}_t(x) = \Phi_{\mathcal{X}_t}P_t\Phi(x) = \widehat{\Phi}_{\mathcal{X}_t}\Phi(x), \quad \tilde{\Phi}_{\mathcal{X}_t}\tilde{\Phi}_{\mathcal{X}_t}^\top = \Phi_{\mathcal{X}_t}P_t\Phi_{\mathcal{X}_t}^\top = \widehat{\Phi}_{\mathcal{X}_t}\widehat{\Phi}_{\mathcal{X}_t}^\top, \tag{19}$$

where P_t is the projection operator as defined in (18). We also have $\tilde{V}_t := \sum_{s=1}^t \tilde{\Phi}_t(x_s)\tilde{\Phi}_t(x_s)^\top = \tilde{\Phi}_{\mathcal{X}_t}^\top \tilde{\Phi}_{\mathcal{X}_t}$. Therefore

$$\begin{aligned}
\tilde{\mu}_t(x) &= \tilde{\Phi}_t(x)^\top (\tilde{\Phi}_{\mathcal{X}_t}^\top \tilde{\Phi}_{\mathcal{X}_t} + \eta I_{nm_t})^{-1} \sum_{s=1}^t \tilde{\Phi}_t(x_s) y_s \\
&= \tilde{\Phi}_t(x)^\top (\tilde{\Phi}_{\mathcal{X}_t}^\top \tilde{\Phi}_{\mathcal{X}_t} + \eta I_{nm_t})^{-1} \tilde{\Phi}_{\mathcal{X}_t}^\top Y_t \\
&= \tilde{\Phi}_t(x)^\top \tilde{\Phi}_{\mathcal{X}_t}^\top (\tilde{\Phi}_{\mathcal{X}_t} \tilde{\Phi}_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} Y_t \\
&= \Phi(x)^\top \widehat{\Phi}_{\mathcal{X}_t}^\top (\widehat{\Phi}_{\mathcal{X}_t} \widehat{\Phi}_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} Y_t \\
&= \Phi(x)^\top (\widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} + \eta I)^{-1} \widehat{\Phi}_{\mathcal{X}_t}^\top Y_t = \Phi(x)^\top (\widehat{V}_t + \eta I)^{-1} \sum_{s=1}^t \widehat{\Phi}_t(x_s) y_s \ ,
\end{aligned}$$

where in third and fifth step, we have used Lemma 2, and in fourth step, we have used (19). Further

$$\begin{aligned}
\tilde{\Gamma}_t(x,x) &= \Gamma(x,x) - \tilde{\Phi}_t(x)^\top \tilde{\Phi}_t(x) + \eta \tilde{\Phi}_t(x)^\top (\tilde{\Phi}_{\mathcal{X}_t}^\top \tilde{\Phi}_{\mathcal{X}_t} + \eta I_{nm_t})^{-1} \tilde{\Phi}_t(x) \\
&= \Gamma(x,x) - \tilde{\Phi}_t(x)^\top \left(I_{nm_t} - \eta (\tilde{\Phi}_{\mathcal{X}_t}^\top \tilde{\Phi}_{\mathcal{X}_t} + \eta I_{nm_t})^{-1} \right) \tilde{\Phi}_t(x) \\
&= \Gamma(x,x) - \tilde{\Phi}_t(x)^\top \tilde{\Phi}_{\mathcal{X}_t}^\top (\tilde{\Phi}_{\mathcal{X}_t} \tilde{\Phi}_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} \tilde{\Phi}_{\mathcal{X}_t} \tilde{\Phi}_t(x) \\
&= \Phi(x)^\top \Phi(x) - \Phi(x)^\top \widehat{\Phi}_{\mathcal{X}_t}^\top (\widehat{\Phi}_{\mathcal{X}_t} \widehat{\Phi}_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} \widehat{\Phi}_{\mathcal{X}_t} \Phi(x) \\
&= \Phi(x)^\top \left(I - \widehat{\Phi}_{\mathcal{X}_t}^\top (\widehat{\Phi}_{\mathcal{X}_t} \widehat{\Phi}_{\mathcal{X}_t}^\top + \eta I_{nt})^{-1} \widehat{\Phi}_{\mathcal{X}_t} \right) \Phi(x) \\
&= \eta \Phi(x)^\top (\widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} + \eta I)^{-1} \Phi(x) = \eta \Phi(x)^\top (\widehat{V}_t + \eta I)^{-1} \Phi(x),
\end{aligned}$$

where in third and sixth step, we have used Lemma 2, and in fourth step, we have used (19). ■

Lemma 9 (Multi-task concentration under Nyström approximation) *Let $f \in \mathcal{H}_\Gamma(\mathcal{X})$ and the noise vectors $\{\varepsilon_t\}_{t \geqslant 1}$ be σ-sub-Gaussian. Further, for any $\eta > 0$, $\varepsilon \in (0,1)$ and $t \geqslant 1$, let $(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} - \varepsilon \eta I \preceq \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \varepsilon \eta I$. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$, the following holds uniformly over all $x \in \mathcal{X}$ and $t \geqslant 1$:*

$$\left\| f(x) - \tilde{\mu}_t(x) \right\|_2 \leqslant \left(c_\varepsilon \left\| f \right\|_\Gamma + \frac{\sigma}{\sqrt{\eta}} \sqrt{2\log(1/\delta) + \log\det(I_{nt} + \eta^{-1} G_t)} \right) \left\| \tilde{\Gamma}_t(x,x) \right\|^{1/2},$$

where $c_\varepsilon = 1 + \frac{1}{\sqrt{1-\varepsilon}}$.

Proof Let us first define $\tilde{\alpha}_t(x) := \Phi(x)^\top \left(\widehat{V}_t + \eta I\right)^{-1} \sum_{s=1}^t \widehat{\Phi}_t(x_s) f(x_s)$, where $\widehat{V}_t = \widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t}$. We now note that $f(x) = \Phi(x)^\top \theta^\star$ and $\tilde{\alpha}_t(x) = \Phi(x)^\top \left(\widehat{V}_t + \eta I\right)^{-1} \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \theta^\star$ for some $\theta^\star \in \ell^2$, so that $\left\|f\right\|_\Gamma = \left\|\theta^\star\right\|_2$. We then have

$$\begin{aligned}
\left\| f(x) - \tilde{\alpha}_t(x) \right\|_2 &= \left\| \Phi(x)^\top \left(\theta^\star - \left(\widehat{V}_t + \eta I\right)^{-1} \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \theta^\star \right) \right\|_2 \\
&\leqslant \left\| \Phi(x)^\top \left(\widehat{V}_t + \eta I\right)^{-1/2} \right\| \left\| \theta^\star - \left(\widehat{V}_t + \eta I\right)^{-1} \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \theta^\star \right\|_{(\widehat{V}_t + \eta I)} \\
&= \left\| \Phi(x)^\top \left(\widehat{V}_t + \eta I\right)^{-1} \Phi(x) \right\|^{1/2} \left\| \left(\widehat{V}_t + \eta I\right) \theta^\star - \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \theta^\star \right\|_{\left(\widehat{V}_t + \eta I\right)^{-1}} \\
&= \eta^{-1/2} \left\| \tilde{\Gamma}_t(x,x) \right\|^{1/2} \left\| \eta \theta^\star - \widehat{\Phi}_{\mathcal{X}_t}^\top \left(\Phi_{\mathcal{X}_t} - \widehat{\Phi}_{\mathcal{X}_t} \right) \theta^\star \right\|_{\left(\widehat{V}_t + \eta I\right)^{-1}} \\
&\leqslant \eta^{-1/2} \left\| \tilde{\Gamma}_t(x,x) \right\|^{1/2} \left(\eta \left\| \theta^\star \right\|_{\left(\widehat{V}_t + \eta I\right)^{-1}} + \left\| \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \left(I - P_t\right) \theta^\star \right\|_{\left(\widehat{V}_t + \eta I\right)^{-1}} \right) \\
&\leqslant \left(\left\| \theta^\star \right\|_2 + \eta^{-1/2} \left\| \left(\widehat{V}_t + \eta I\right)^{-1/2} \widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} \left(I - P_t\right) \theta^\star \right\|_2 \right) \left\| \tilde{\Gamma}_t(x,x) \right\|^{1/2}.
\end{aligned}$$

Here in the fourth step, we have used Lemma 8 and in the second last step, we have used $\widehat{\Phi}_{\mathcal{X}_t} = \Phi_{\mathcal{X}_t} P_t$, where P_t is the projection operator as defined in (18). The last step is controlled as $\left\|\theta^\star\right\|_{(\widehat{V}_t+\eta I)^{-1}} \leqslant \eta^{-1/2}\left\|\theta^\star\right\|_2$. We now have

$$\begin{aligned}\left\|\left(\widehat{V}_t+\eta I\right)^{-1/2}\widehat{\Phi}_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t}\left(I-P_t\right)\theta^\star\right\|_2 &\leqslant \left\|\left(\widehat{V}_t+\eta I\right)^{-1/2}\widehat{\Phi}_{\mathcal{X}_t}^\top\right\|\left\|\Phi_{\mathcal{X}_t}\left(I-P_t\right)\right\|\left\|\theta^\star\right\|_2\\ &\leqslant \left\|\Phi_{\mathcal{X}_t}(I-P_t)\Phi_{\mathcal{X}_t}^\top\right\|^{1/2}\left\|\theta^\star\right\|_2,\end{aligned}$$

where we have used that $\left\|(\widehat{V}_t+\eta I)^{-1/2}\widehat{\Phi}_{\mathcal{X}_t}^\top\right\| = \left\|\widehat{\Phi}_{\mathcal{X}_t}(\widehat{\Phi}_{\mathcal{X}_t}^\top\widehat{\Phi}_{\mathcal{X}_t}+\eta I)^{-1}\widehat{\Phi}_{\mathcal{X}_t}^\top\right\|^{1/2} \leqslant 1$ and $(I-P_t)^2 = I-P_t$. We now observe from Lemma 2 and our hypothesis $(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t} - \varepsilon\eta I \preceq \Phi_{\mathcal{D}_t}^\top\Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t}+\varepsilon\eta I$ that

$$I-P_t \preceq I-\Phi_{\mathcal{D}_t}^\top(\Phi_{\mathcal{D}_t}\Phi_{\mathcal{D}_t}^\top+\eta I_{nm_t})^{-1}\Phi_{\mathcal{D}_t} = \eta(\Phi_{\mathcal{D}_t}^\top\Phi_{\mathcal{D}_t}+\eta I)^{-1} \preceq \frac{\eta}{1-\varepsilon}(\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t}+\eta I)^{-1},$$

and therefore, $\left\|\Phi_{\mathcal{X}_t}(I-P_t)\Phi_{\mathcal{X}_t}^\top\right\|^{1/2} \leqslant \sqrt{\frac{\eta}{1-\varepsilon}}\left\|\Phi_{\mathcal{X}_t}(\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t}+\eta I)^{-1}\Phi_{\mathcal{X}_t}^\top\right\|^{1/2} \leqslant \sqrt{\frac{\eta}{1-\varepsilon}}$. Putting it all together, we now have

$$\left\|f(x)-\tilde{\alpha}_t(x)\right\|_2 \leqslant \left\|\theta^\star\right\|_2\left(1+\frac{1}{\sqrt{1-\varepsilon}}\right)\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2} = c_\varepsilon\left\|f\right\|_\Gamma\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2}, \tag{20}$$

where we have used that $\left\|\theta^\star\right\|_2 = \left\|f\right\|_\Gamma$ and $c_\varepsilon = 1+\frac{1}{\sqrt{1-\varepsilon}}$. We further obtain from Lemma 8 that

$$\begin{aligned}\left\|\tilde{\mu}_t(x)-\tilde{\alpha}_t(x)\right\|_2 &= \left\|\Phi(x)^\top\left(\widehat{V}_t+\eta I\right)^{-1}\sum_{s=1}^t\widehat{\Phi}_t(x_s)(y_s-f(x_s))\right\|_2\\ &\leqslant \left\|\Phi(x)^\top\left(\widehat{V}_t+\eta I\right)^{-1/2}\right\|\left\|\sum_{s=1}^t\widehat{\Phi}_t(x_s)\varepsilon_s\right\|_{\left(\widehat{V}_t+\eta I\right)^{-1}}\\ &= \left\|\Phi(x)^\top\left(\widehat{V}_t+\eta I\right)^{-1}\Phi(x)\right\|^{1/2}\left\|\widehat{\Phi}_{\mathcal{X}_t}^\top E_t\right\|_{\left(\widehat{V}_t+\eta I\right)^{-1}}\\ &= \eta^{-1/2}\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2}\left\|\widehat{\Phi}_{\mathcal{X}_t}^\top E_t\right\|_{\left(\widehat{V}_t+\eta I\right)^{-1}},\end{aligned}$$

where $E_t = \left[\varepsilon_1^\top,\ldots,\varepsilon_t^\top\right]^\top$ denotes an $nt\times 1$ vector formed by concatenating the noise vectors $\varepsilon_i, 1\leqslant i\leqslant t$. We now have

$$\begin{aligned}\left\|\widehat{\Phi}_{\mathcal{X}_t}^\top E_t\right\|^2_{\left(\widehat{V}_t+\eta I\right)^{-1}} &= E_t^\top\widehat{\Phi}_{\mathcal{X}_t}\left(\widehat{\Phi}_{\mathcal{X}_t}^\top\widehat{\Phi}_{\mathcal{X}_t}+\eta I\right)^{-1}\widehat{\Phi}_{\mathcal{X}_t}^\top E_t\\ &= E_t^\top\left(I_{nt}-\eta\left(\widehat{\Phi}_{\mathcal{X}_t}\widehat{\Phi}_{\mathcal{X}_t}^\top+\eta I_{nt}\right)^{-1}\right)E_t\\ &\leqslant E_t^\top\left(I_{nt}-\eta\left(\Phi_{\mathcal{X}_t}\Phi_{\mathcal{X}_t}^\top+\eta I_{nt}\right)^{-1}\right)E_t\\ &= E_t^\top\Phi_{\mathcal{X}_t}\left(\Phi_{\mathcal{X}_t}^\top\Phi_{\mathcal{X}_t}+\eta I\right)^{-1}\Phi_{\mathcal{X}_t}^\top E_t = \left\|\Phi_{\mathcal{X}_t}^\top E_t\right\|^2_{(V_t+\eta I)^{-1}},\end{aligned}$$

where in second and fourth step, we have used Lemma 2, and in third step, we have used $\widehat{\Phi}_{\mathcal{X}_t}\widehat{\Phi}_{\mathcal{X}_t}^\top = \Phi_{\mathcal{X}_t}P_t\Phi_{\mathcal{X}_t}^\top \preceq \Phi_{\mathcal{X}_t}\Phi_{\mathcal{X}_t}^\top$. We then have

$$\begin{aligned}\left\|\tilde{\mu}_t(x)-\tilde{\alpha}_t(x)\right\|_2 &\leqslant \eta^{-1/2}\left\|\sum_{s=1}^t\Phi(x_s)\varepsilon_s\right\|_{(V_t+\eta I)^{-1}}\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2}\\ &= \eta^{-1/2}\left\|S_t\right\|_{(V_t+\eta I)^{-1}}\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2},\end{aligned} \tag{21}$$

where $S_t := \sum_{s=1}^t\Phi(x_s)\varepsilon_s$. Combining (20) and (21) together, we now obtain

$$\begin{aligned}\left\|f(x)-\tilde{\mu}_t(x)\right\|_2 &\leqslant \left\|f(x)-\tilde{\alpha}_t(x)\right\|_2+\left\|\tilde{\alpha}_t(x)-\tilde{\mu}_t(x)\right\|_2\\ &\leqslant \left(c_\varepsilon\left\|f\right\|_\Gamma+\eta^{-1/2}\left\|S_t\right\|_{(V_t+\eta I)^{-1}}\right)\left\|\tilde{\Gamma}_t(x,x)\right\|^{1/2}.\end{aligned}$$

We now conclude the proof using Lemma 3. ■

D.2 Controlling Approximate Predictive Variance

We now show that an accurate dictionary helps us to control the approximate predictive variances

Lemma 10 (Approximate predictive variance control) *For any $\eta > 0$ and $\varepsilon \in (0,1)$, let $\rho = (1+\varepsilon)/(1-\varepsilon)$ and $(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} - \varepsilon\eta I \preceq \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \varepsilon\eta I$. Then*

$$\frac{1}{\rho}\Gamma_t(x,x) \preceq \tilde{\Gamma}_t(x,x) \preceq \rho\Gamma_t(x,x).$$

Proof We first note that $\widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} = P_t \Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} P_t$, where P_t is the projection operator as defined in (18). Then our hypothesis $(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} - \varepsilon\eta I \preceq \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \varepsilon\eta I$ can be re-formulated as

$$\frac{1}{1+\varepsilon} P_t \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} P_t - \frac{\varepsilon\eta}{1+\varepsilon} P_t \preceq \widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} \preceq \frac{1}{1-\varepsilon} P_t \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} P_t + \frac{\varepsilon\eta}{1-\varepsilon} P_t.$$

Since, by definition, $P_t \Phi_{\mathcal{D}_t}^\top = \Phi_{\mathcal{D}_t}^\top$ and $P_t \preceq I$, we have

$$\frac{1}{1+\varepsilon} \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} - \frac{\varepsilon\eta}{1+\varepsilon} \preceq \widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} \preceq \frac{1}{1-\varepsilon} \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} + \frac{\varepsilon\eta}{1-\varepsilon},$$

and, thus, in turn

$$\frac{1}{1+\varepsilon}\left(\Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} + \eta I\right) \preceq \widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} + \eta I \preceq \frac{1}{1-\varepsilon}\left(\Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} + \eta I\right).$$

We now obtain from our hypothesis that

$$\frac{1-\varepsilon}{1+\varepsilon}\left(\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \eta I\right) \preceq \widehat{\Phi}_{\mathcal{X}_t}^\top \widehat{\Phi}_{\mathcal{X}_t} + \eta I \preceq \frac{1+\varepsilon}{1-\varepsilon}\left(\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \eta I\right).$$

This further implies that

$$\frac{1-\varepsilon}{1+\varepsilon}\Phi(x)^\top (V_t+\eta I)^{-1}\Phi(x) \preceq \Phi(x)^\top \left(\widehat{V}_t + \eta I\right)^{-1}\Phi(x) \preceq \frac{1+\varepsilon}{1-\varepsilon}\Phi(x)^\top (V_t+\eta I)^{-1}\Phi(x),$$

which completes the proof. ■

D.3 Regret and Complexity Bounds for MT-BKB (Proof of Theorem 3)

Since the scalarization functions s_λ is L-Lipschitz in the ℓ_2 norm, we have

$$\forall \lambda \in \Lambda, \quad \left|s_\lambda\left(f(x)\right) - s_\lambda\left(\tilde{\mu}_{t-1}(x)\right)\right| \leqslant L \left\|f(x) - \tilde{\mu}_{t-1}(x)\right\|_2.$$

Since $\tilde{\mu}_0(x) = 0$, $\tilde{\Gamma}_0(x,x) = \Gamma(x,x)$ and $\|f\|_\Gamma \leqslant b$, we have

$$\|f(x) - \tilde{\mu}_0(x)\|_2 = \left\|\Gamma_x^\top f\right\|_2 \leqslant \|f\|_\Gamma \|\Gamma_x\| = \|f\|_\Gamma \left\|\Gamma_x^\top \Gamma_x\right\|^{1/2} \leqslant b\left\|\tilde{\Gamma}_0(x,x)\right\|^{1/2}.$$

Further, since $\log(1+ax) \leqslant a\log(1+x)$ holds for any $a \geqslant 1$ and $x \geqslant 0$, we obtain from Lemma 4 and Lemma 10 that

$$\begin{aligned}\log\det\left(I_{nt} + \eta^{-1}G_t\right) &= \sum_{s=1}^t \log\det\left(I_n + \eta^{-1}\Gamma_{s-1}(x_s,x_s)\right) \\ &\leqslant \rho\sum_{s=1}^t \log\det\left(I_n + \eta^{-1}\tilde{\Gamma}_{s-1}(x_s,x_s)\right), \end{aligned} \tag{22}$$

where $\rho = \frac{1+\varepsilon}{1-\varepsilon}$. Let us now assume, for any $t \geqslant 1$, that

$$(1-\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} - \varepsilon\eta I \preceq \Phi_{\mathcal{D}_t}^\top \Phi_{\mathcal{D}_t} \preceq (1+\varepsilon)\Phi_{\mathcal{X}_t}^\top \Phi_{\mathcal{X}_t} + \varepsilon\eta I. \tag{23}$$

Then, from (22) and Lemma 9, the following holds with probability at least $1-\delta/2$:

$$\forall t \geqslant 1, \forall x \in \mathcal{X}, \forall \lambda \in \Lambda, \qquad \left|s_\lambda\left(f(x)\right) - s_\lambda\left(\tilde{\mu}_{t-1}(x)\right)\right| \leqslant L\tilde{\beta}_{t-1}\left\|\tilde{\Gamma}_{t-1}(x,x)\right\|^{1/2}, \tag{24}$$

where $\tilde{\beta}_t = c_\varepsilon b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\log(2/\delta) + \rho\sum_{s=1}^{t}\log\det\left(I_n + \eta^{-1}\tilde{\Gamma}_{s-1}(x_s,x_s)\right)}$, $t \geqslant 0$ and $c_\varepsilon = 1 + \frac{1}{\sqrt{1-\varepsilon}}$. We can now upper bound the *instantaneous regret* at time $t \geqslant 1$ as

$$\begin{aligned}
r_t &:= \mathbb{E}\left[s_\lambda\left(f(x^\star)\right)\right] - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\
&\leqslant \mathbb{E}\left[s_\lambda\left(\tilde{\mu}_{t-1}(x^\star)\right)\right] + L\tilde{\beta}_{t-1}\left\|\tilde{\Gamma}_{t-1}(x^\star,x^\star)\right\|^{1/2} - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\
&\leqslant \mathbb{E}\left[s_\lambda\left(\tilde{\mu}_{t-1}(x_t)\right)\right] + L\tilde{\beta}_{t-1}\left\|\tilde{\Gamma}_{t-1}(x_t,x_t)\right\|^{1/2} - \mathbb{E}\left[s_\lambda\left(f(x_t)\right)\right]\\
&\leqslant 2L\tilde{\beta}_{t-1}\left\|\tilde{\Gamma}_{t-1}(x_t,x_t)\right\|^{1/2}.
\end{aligned}$$

Here in the first and third step, we have used (24). The second step follows from the choice of x_t. Since $\tilde{\beta}_t$ is a monotonically increasing function in t, we now have

$$\begin{aligned}
R_C(T) := \sum_{t=1}^{T} r_t \leqslant 2L\tilde{\beta}_T\sum_{t=1}^{T}\left\|\tilde{\Gamma}_{t-1}(x_t,x_t)\right\|^{1/2} &\leqslant 2L\tilde{\beta}_T\sqrt{\rho T\sum_{t=1}^{T}\left\|\Gamma_{t-1}(x_t,x_t)\right\|}\\
&\leqslant 2L\tilde{\beta}_T\sqrt{\rho(1+\kappa/\eta)T\sum_{t=1}^{T}\left\|\Gamma_t(x_t,x_t)\right\|},
\end{aligned}$$

where the second last step is due to the Cauchy-Schwartz inequality and Lemma 10, and the last step is due to Lemma 5. A similar argument as in (22) now yields

$$\begin{aligned}
\sum_{t=1}^{T}\log\det\left(I_n + \eta^{-1}\tilde{\Gamma}_{t-1}(x_t,x_t)\right) &\leqslant \rho\sum_{t=1}^{T}\log\det\left(I_n + \eta^{-1}\Gamma_{t-1}(x_t,x_t)\right)\\
&= \rho\log\det\left(I_{nT} + \eta^{-1}G_T\right) \leqslant 2\rho\gamma_{nT}(\Gamma,\eta).
\end{aligned}$$

We then have $\tilde{\beta}_T \leqslant c_\varepsilon b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\left(\log(2/\delta) + \rho^2\gamma_{nT}(\Gamma,\eta)\right)}$. Setting $q = \frac{6\rho\ln(4T/\delta)}{\varepsilon^2}$, we now have from Lemma 7, that with probability at least $1-\delta/2$, uniformly across all $t \in [T]$, the dictionary size $m_t \leqslant 6\rho q\left(1+\kappa/\eta\right)\sum_{s=1}^{t}\left\|\Gamma_s(x_s,x_s)\right\|$ and (23) is true. Using a union bound argument, we then obtain, with probability at least $1-\delta$, the cumulative regret

$$R_C^{\text{MT-BKB}}(T) \leqslant 2L\left(c_\varepsilon b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\left(\log(1/\delta) + \rho^2\gamma_{nT}(\Gamma,\eta)\right)}\right)\sqrt{\rho(1+\kappa/\eta)T\sum_{t=1}^{T}\left\|\Gamma_t(x_t,x_t)\right\|}\ .$$

We conclude the proof by noting that $\rho = \frac{1+\varepsilon}{1-\varepsilon} > 1$ and $c_\varepsilon = 1 + \frac{1}{\sqrt{1-\varepsilon}} \leqslant 2\rho$.

E ON PARETO OPTIMALITY AND RANDOM SCALARIZATIONS

In this section, we show that our algorithms can be adapted to achieve a low Bayes regret. We recall that for a set of points $\mathcal{X}_T = \{x_1,\ldots,x_T\}$, the *Bayes regret* is defined as

$$R_B(T) := \mathbb{E}\left[r_\lambda(T)\right]\ ,\quad \text{where}\ \ r_\lambda(T) := \max_{x\in\mathcal{X}} s_\lambda\left(f(x)\right) - \max_{x\in\mathcal{X}_T} s_\lambda\left(f(x)\right)\ .$$

If s_λ is Lipschitz continuous and montonically increasing, then a low value of Bayes' regret implies that $f(\mathcal{X}_T)$ spans the high probability regions (w.r.t. the prior P_λ) of the Pareto front $f(\mathcal{X}_f)$. To see this, we first note that monotonicity ensures $x^\star_\lambda := \operatorname{argmax}_{x\in\mathcal{X}} s_\lambda\left(f(x)\right)$, the maximizer of the scalarized objective is a Pareto optimal point, i.e., $x^\star_\lambda \in \mathcal{X}_f$ (Roijers et al., 2013). Thus, the prior P_λ defines a probability distribution over the Pareto optimal set $\mathcal{X}_f$, and thus, in turn, over the Pareto front $f(\mathcal{X}_f)$. Next, we observe that it requires the point-wise regret $r_\lambda(T)$ to be low for all $\lambda \in \Lambda$ that has high mass, to achieve a low Bayes regret. Now, the point-wise regret $r_\lambda(T) = 0$ if $x^\star_\lambda \in \mathcal{X}_T$. Then, by the Lipschitz continuity, a low value of $R_B(T)$ will essentially imply $f(\mathcal{X}_T)$ to "span" the high probability regions of $f(\mathcal{X}_f)$.

Controlling the Bayes Regret Following Paria et al. (2020), we bound the Bayes regret by a surrogate regret measure, defined as

$$R'(T) := \sum_{t=1}^{T}\mathbb{E}\big[s_{\lambda_t}\left(f(x^\star_{\lambda_t})\right) - s_{\lambda_t}\left(f\left(x_t\right)\right)\big]\ ,\quad \text{where}\ \ \lambda_t \overset{\text{i.i.d.}}{\sim} P_\lambda,\ \forall t \leqslant T\ .$$

Paria et al. (2020) show that under some mild conditions (Λ is a bounded set and s_λ is Lipschitz continuous in λ), then $R_B(T) \leqslant \frac{1}{T} R'(T) + o(1)$. A sub-linear growth of $R'(T)$ with T then implies that $R_B(T) \to 0$ as $T \to \infty$. We now adapt MT-KB with random scalarizations to ensure a sub-linear growth of $R'(T)$. (A similar analysis follows for MT-BKB.) At each round t, we modify the acquisition function for MT-KB as

$$u'_t(x) = s_{\lambda_t}\left(\mu_{t-1}(x)\right) + L\beta_{t-1}\left\|\Gamma_{t-1}(x,x)\right\|^{1/2} \;, \quad \text{where} \;\; \lambda_t \sim P_\lambda \,.$$

We then select the point x_t that maximizes this modified acquisition function u'_t.

Now, since the scalarization function s_λ is L-Lipschitz in the ℓ_2 norm, we have with probability one, the following:

$$\left|s_{\lambda_t}\left(f(x)\right) - s_{\lambda_t}\left(\mu_{t-1}(x)\right)\right| \leqslant L\left\|f(x) - \mu_{t-1}(x)\right\|_2 \;.$$

Then, from Theorem 1 and Lemma 4, the following holds with probability at least $1-\delta$:

$$\forall t \geqslant 1, \forall x \in \mathcal{X}, \quad \left|s_{\lambda_t}\left(f(x)\right) - s_{\lambda_t}\left(\mu_{t-1}(x)\right)\right| \leqslant L\beta_{t-1}\left\|\Gamma_{t-1}(x,x)\right\|^{1/2}, \tag{25}$$

where $\beta_t = b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\log(1/\delta) + \sum_{s=1}^{t}\log\det\left(I_n + \eta^{-1}\Gamma_{s-1}(x_s,x_s)\right)}$, $t \geqslant 0$. We can now upper bound the *instantaneous surrogate regret* at time $t \geqslant 1$ as

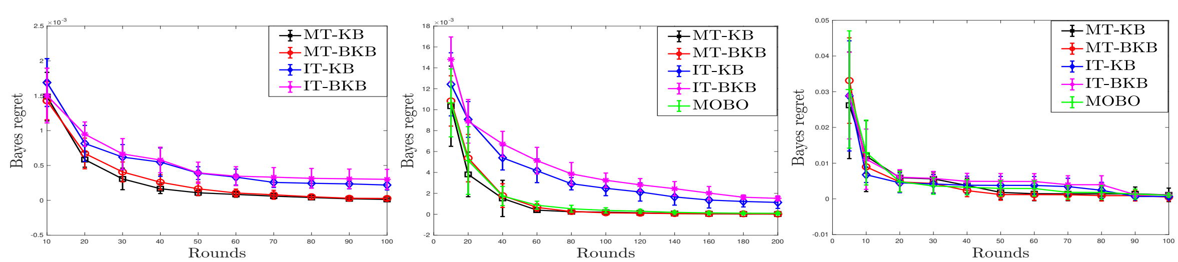
$$\begin{aligned} r'(t) &:= s_{\lambda_t}\left(f(x^\star_{\lambda_t})\right) - s_{\lambda_t}\left(f(x_t)\right) \\ &\leqslant s_{\lambda_t}\left(\mu_{t-1}(x^\star_{\lambda_t})\right) + L\beta_{t-1}\left\|\Gamma_{t-1}(x^\star_{\lambda_t}, x^\star_{\lambda_t})\right\|^{1/2} - s_{\lambda_t}\left(f(x_t)\right) \\ &\leqslant s_{\lambda_t}\left(\mu_{t-1}(x_t)\right) + L\beta_{t-1}\left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2} - s_{\lambda_t}\left(f(x_t)\right) \\ &\leqslant 2L\beta_{t-1}\left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2}. \end{aligned}$$

Here in the first and third step, we have used (25). The second step follows from the choice of x_t. Since β_t is a monotonically increasing function in t, we have

$$\begin{aligned} R'(T) := \mathbb{E}\left[\sum_{t=1}^{T} r'(t)\right] &\leqslant 2L\beta_T \sum_{t=1}^{T}\left\|\Gamma_{t-1}(x_t,x_t)\right\|^{1/2} \\ &\leqslant 2L\beta_T\sqrt{(1+\kappa/\eta)T\sum_{t=1}^{T}\left\|\Gamma_t(x_t,x_t)\right\|} \\ &\leqslant 2L\beta_T\sqrt{2(\kappa+\eta)T\gamma_{nT}(\Gamma,\eta)}\,, \end{aligned}$$

where the second last step is due to the Cauchy-Schwartz inequality and Lemma 5, and the last step follows from Lemma 4. We further obtain from Lemma 4 that $\beta_T \leqslant b + \frac{\sigma}{\sqrt{\eta}}\sqrt{2\left(\log(1/\delta) + \gamma_{nT}(\Gamma,\eta)\right)}$, yielding the desired sub-linear growth of $R'(T)$ with T.

Comparison of Bayes Regret We compare the Bayes regret $R_B(T)$ of MT-KB and MT-BKB (using random scalarizations) with independent task benchmarks IT-KB, IT-BKB and MOBO in Figure 4. We observe that learning the tasks together yields better if not similar performance compared to learning the tasks independently.



(a) RKHS function (b) Perturbed sine function (c) Sensor measurements

Figure 4: Comparison of Bayes regret of MT-KB and MT-BKB with IT-KB, IT-BKB and MOBO using Chebyshev scalarization.

F ADDITIONAL DETAILS ON EXPERIMENTS

Comments on Parameters Used We set the confidence radii (i.e., β_t and $\tilde{\beta}_t$) of MT-KB and MT-BKB exactly as given in Theorem 2 and Theorem 3, respectively. Similarly, for IT-KB and IT-BKB, we use respective choices of radii given in

Chowdhury and Gopalan (2017) and Calandriello et al. (2019) in the context of single task BO and suitably blow those up by a $\sqrt{n}$ factor to account for n tasks. For MOBO, we use the UCB acquistion function and set the radius as specified in Paria et al. (2020). To make the comparison uniform across all experiments, we do not tune any hyper-parameter for any algorithm and for a particular hyperparameter, we always use the same value in all algorithms. The hyper-paramter choices are specified in Section 5. We though believe that careful tuning of hyper-parameters might lead to better performance in practice.

A Note on the Sensor Data The data was collected at 30 second intervals for 5 consecutive days starting Feb. 28th 2004 from 54 sensors deployed in the Intel Berkeley Research lab. We have downloaded the data previously from the webpage `http://db.csail.mit.edu/labdata/labdata`. But the link appears to be broken now. We can share a copy of our downloaded version if asked to do so.