Multi-task Bayesian Optimization via Gaussian Process Upper Confidence Bound

Sihui Dai SIHUID@PRINCETON.EDU
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

Jialin Song JSSONG@CALTECH.EDU
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91106, USA

Yisong Yue YYUE@CALTECH.EDU
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91106, USA

Abstract

Bayesian optimization is a popular method for optimizing black-box functions. In this work, we consider multi-task generalization of Bayesian optimization for optimizing multiple related black-box functions simultaneously. We use multi-task Gaussian processes to model the joint distribution among all tasks and propose a simple acquisition function based on GP-UCB (Srinivas et al., 2012). Extending the notion of regret to multi-task Bayesian optimization, we provide a theoretical analysis to show that our proposed algorithm is no-regret under mild assumptions. We provide experimental results on a wide variety of applications, from synthetic test function optimizations to hyperparameter tuning of machine learning models and nanophotonic structure designs. Our experiments show that multi-task GP-UCB is able to achieve smaller simple regret compared to optimizing each task individually using single-task GP-UCB.

Keywords: Bayesian Optimization, Multi-task Optimization, Optimal Experimental Design

1. Introduction

Optimizing expensive blackbox functions has many applications from experimental design (Romero et al., 2013; Fleischman et al., 2017; Yang et al., 2019; Zhang et al., 2020) to hyperparameter tuning of machine learning models (Snoek et al., 2012; Swersky et al., 2013). In many settings, a collection of related functions need to be optimized simultaneously. For example, a scientist might want to design a collection of light filters, each of which is targeted at a different wavelength. While the ultimate design of each filter is different, they might share common structures and materials. One could optimize each task separately, ignoring the connections among them. A better approach is to optimize all tasks simultaneously and make use of the shared information to speed up the optimization process.

The core research problem we consider in this paper is how to model connections among tasks and how to perform the joint optimization. To answer the first question, we use the multi-task Gaussian process (Bonilla et al., 2008), an extension of Gaussian processes, to model inter-task correlations. We propose MT-GP-UCB, a novel multi-task Bayesian optimization algorithm to balance exploration-exploitation across tasks. We show that under mild assumptions, MT-GP-UCB is a no-regret learning algorithm. Experimental results on a wide variety of applications are provided, from synthetic test function optimization to real world nanophotonic structure designs to showcase strong performances of the proposed algorithm.

2. Related Works

2.1 Gaussian Processes

Gaussian processes (GP) (Williams and Rasmussen, 2006) are compact models for an infinite collection $\mathcal{X}$ of random variables such that any finite subset of them is distributed jointly as a multivariate Gaussian. A GP is specified by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a covariance, or kernel, function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For each $x \in \mathcal{X}$, the prior mean is $\mu(x) = \mathbb{E}[f(x)]$. For $x, x' \in \mathcal{X}$, their prior covariance is $k(x, x') = \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))]$. GP models enable close-form posterior inference given a set of noisy observations. Assume that $f \sim GP(\mu, k)$ and $S = \{(x_i, y_i\}_{i=1}^n$ are noisy observations, where $y_i = f(x_i) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then we can compute the posterior mean and covariance in terms of the observed values $y_S = [y_i]_{1 \leq i \leq n}$, the sample covariance matrix $K_S = [k(x_i, x_j) + \sigma^2]_{1 \leq i,j \leq n}$ and $k_S(x) = [k(x_i, x)]_{1 \leq i \leq n}$ as follows:

$$\mu_S(x) = \mu(x) + k_S(x)^T K_S^{-1} y_S$$
$$k_S(x, x') = k(x, x') - k_S(x)^T K_S^{-1} k_S(x')$$

2.2 Bayesian Optimization with Gaussian Processes

With Gaussian process models, Bayesian optimization is a popular approach for optimizing expensive blackbox functions. A classical method with theoretical guarantee is the GP-UCB algorithm (Srinivas et al., 2012) which takes a multi-armed bandit view of the optimization problem and proposes an upper-confidence bound based acquisition function for query selections. Other related approaches include Max-value entropy search (Wang and Jegelka, 2017), entropy search (Hennig and Schuler, 2012), predictive entropy search (Hernández-Lobato et al., 2014) and expected improvement algorithm (Jones et al., 1998).

2.3 Multi-task Bayesian Optimization

A generalization of Bayesian optimization to the multi-task setting has many practical applications. In (Swersky et al., 2013), the authors focus on hyperparameter tunings of machine learning models. They study the problem of utilizing already tuned models to speed up new model tunings for related tasks and propose an acquisition function based on entropy search. In (Metzen, 2016), the author designs an algorithm that aims on minimizing regret directly and shows it generalizes to the multi-task setting. The most closely related to our study is the work of contextual Gaussian process bandit optimization (Krause and

Ong, 2011) where contextual information can be associated with tasks. Their proposed CGP-UCB approach is also an extension of the GP-UCB acquisition function (Srinivas et al., 2012) but their focus, based on the contextual regret definition, does not capture the goal of optimizing each task, which is the focus of this paper. Another point of distinction is that our method operates in query batches instead of sequentially like previous approaches, which enables a more flexible information gathering procedure across tasks. Other multi-task Bayesian optimization algorithms based on information theory (Groves et al., 2018; Pearce and Branke, 2018; Ramachandran et al., 2018) and knowledge gradient (Poloczek et al., 2016) are also proposed.

3. The Multi-task GP-UCB Algorithm

Given a set of n tasks $f^{[n]} := \{f^1, f^2, ..., f^n\}$, we would like to find a set $\{x^1, x^2, ..., x^n\}$ such that f^i achieves its maximum value at x^i. The true forms of f^i are unknown, and there is some nonzero cost in querying tasks, so we would like to minimize the number of times we query. We assume that the cost of querying each task is the same across tasks. At each iteration, we query $\{f^1, f^2, ..., f^n\}$ at $\{x^1, x^2, ..., x^n\}$ (not necessarily the same locations), and we receive noisy observations $y^i = f^i(x^i) + \epsilon^i$ where $\epsilon^i \sim \mathcal{N}(0, \sigma_i^2)$. We assume the Gaussian noises are mutually independent across tasks. Notationally, we use x_t^i to denote the query point selected for task f^i at step t. $x_t^{[n]} := \{x_t^1, x_t^2, ..., x_t^n\}$ are all query points selected at step t and $x_{[t]}^{[n]} := \{x_1^{[n]}, x_2^{[n]}, ..., x_t^{[n]}\}$ are all query points selected up until step t. Finally, $y_{[t]}^{[n]}$ are all noisy observations at query points up until step t.

To model the joint distribution across tasks, we employ a multi-task Gaussian process model proposed in (Bonilla et al., 2008). Specifically, we consider a GP model with a composition kernel structure $k = k_X \otimes k_T$ such that $k(x^i, x'^j) = k_X(x, x')k_T(i, j)$. Here k_X models covariance between query locations and k_T captures inter-task covariances.

Let μ_T, Σ_T denote the posterior mean and covariance of the joint GP after $x_{[T]}^{[n]}$ has been queried. Let μ_T^i be the mean of task i at time T and σ_T^i be the variance for task i obtained after marginalization. Let $x_{T+1}^{[n]}$ denote the the set $\{x_{T+1}^1, x_{T+1}^2, ..., x_{T+1}^n\}$ to be queried at time $T+1$. We define multi-task GP-UCB (MT-GP-UCB) acquisition function as:

$$x_{T+1}^i := \arg\max_{x^i} \mu_T^i(x^i) + \beta_{T+1}^{-\frac{1}{2}} \sigma_T^i(x^i) \quad \forall i \in [1, n]$$

The main difference from GP-UCB is that the posterior mean and covariance are computed using all observations instead of from a single task only. We aim to derive a regret bound for MT-GP-UCB and show that it is no-regret. Define instantaneous multi-task regret as $r_t := \frac{1}{n}\sum_{i=1}^n (f^i(x_t^{*i}) - f^i(x_t^i))$ and simple regret as $r_t^* := \frac{1}{n}\sum_{i=1}^n (f^i(x_t^{*i}) - \max_{j \leq i} f^i(x_t^j))$. The cumulative regret is $R(t) := \sum_t r_t$.

4. Theoretical Analysis

We provide a regret analysis for the simplified case where the domain $\mathcal{X}$ is discrete with a cardinality of N. Proofs for continous domains are more involved but use similar core ideas. All proofs are deferred to the Appendix due to limited space. Similar to the GP-UCB regret analysis (Srinivas et al., 2012), our bound depends on a term capturing the mutual information between query points and blackbox functions. Specifically, we define $\gamma_T := \max_{x_{[T]}^{[n]} \subset \mathcal{X}} I(y_{[T]}^{[n]}; f^{[n]})$ is the maximum mutual information between T rounds of batched queries and $f^{[n]} = \{f^1, f^2, ..., f^n\}$. In our analysis, we will also use the single task version of the maximum mutual information. For each function f^i, we define $\gamma_T^i := \max_{x_{[T]}^i \subset \mathcal{X}} I(y_{[T]}^i; f^i)$ to be the maximum mutual information obtained with a single task.

First we show that the cumulative regret is bounded, with high probability, by a quantity related to the maximum mutual information γ_T:

Theorem 1 *Let $(f^1, f^2, ..., f^n)$ be samples from a multi-task Gaussian process with a mean function $\mu(x)$ and a bounded covariance function $k(x, x') \leq 1$. Let $\delta \in (0, 1)$ and $\beta_t = 2\log(nNt^2\pi^2/6\delta)$. By running MT-GP-UCB for T rounds, the cumulative regret satisfies the following bound with high probability. Precisely,*

$$\mathbb{P}(R_T \leq \sqrt{CT\beta_T\gamma_T}, \forall T \geq 1) \geq 1 - \delta$$

where $C := \frac{8}{n\log(1+\sigma^{-2})}$ and σ^2 is the common noise variance across tasks.

To show MT-GP-UCB is no-regret, it suffices to show that $R_T/T \to 0$ as $T \to \infty$. We make use of the following lemma to establish the relationship between γ_T and the single task maximum mutual information.

Lemma 2 *The maximum mutual information across functions γ_T is upper bounded by the sum of maximum mutual information of each individual function if each function optimization is conducted separately. That is $\gamma_T \leq \sum_{i=1}^{n} \gamma_T^i$.*

In (Srinivas et al., 2012), the authors show that for common kernels such as RBF and Matérn kernels, γ_T^i grows sublinearly in T. It follows that γ_T also grows sublinearly in T. Combining with Theorem 1, MT-GP-UCB is a no-regret algorithm.

5. Experimental Results

We present experiments on both discrete and continuous datasets. We compare the simple regret from MT-GP-UCB to single task Bayesian optimization using GP-UCB (ST-GP-UCB) and random sampling. For MT-GP-UCB, we use a GP with a multi-task kernel to model the joint distribution over tasks. We use an RBF kernel to model covariance between points and an index kernel to model covariance between tasks. For ST-GP-UCB, we use a separate GP with RBF kernel for each task. We perform experiments on synthetic test functions including the Hartman-3D, Currin exponential, and Borehole functions and their transformations as tasks. Additionally, we experiment with optimizing over discrete data from nanophotonics and tuning hyperparameters for various models on the iris dataset (Dua and Graff, 2017).

5.1 Synthetic Examples

We experiment with the Hartman-3D, Currin exponential, and Borehole functions. These functions are defined in Appendix B. For all functions, we scale the each dimension of the input to lie in the range $[0, 1]$. We run with 2 tasks. The first task is the function itself and the second is a transformation of the function. We test 3 transformations, one where the second task is identical to the first, one where the second task is a shift from the first (all inputs shifted by 0.01 in each dimension), and one where the second task is a scaled version of the first (all inputs scaled by 0.9). We compare MT-GP-UCB with ST-GP-UCB and choose $\beta = 10$ for both methods. We assume that each task has a cost of 0.5. Plots of simple regrets are shown in Figure 1. Plots of average regret R_T/T are also provided in Appendix B. We can see that for most of functions and their transformations, MT-GP-UCB is able to achieve the smallest simple regret as more points are queried.

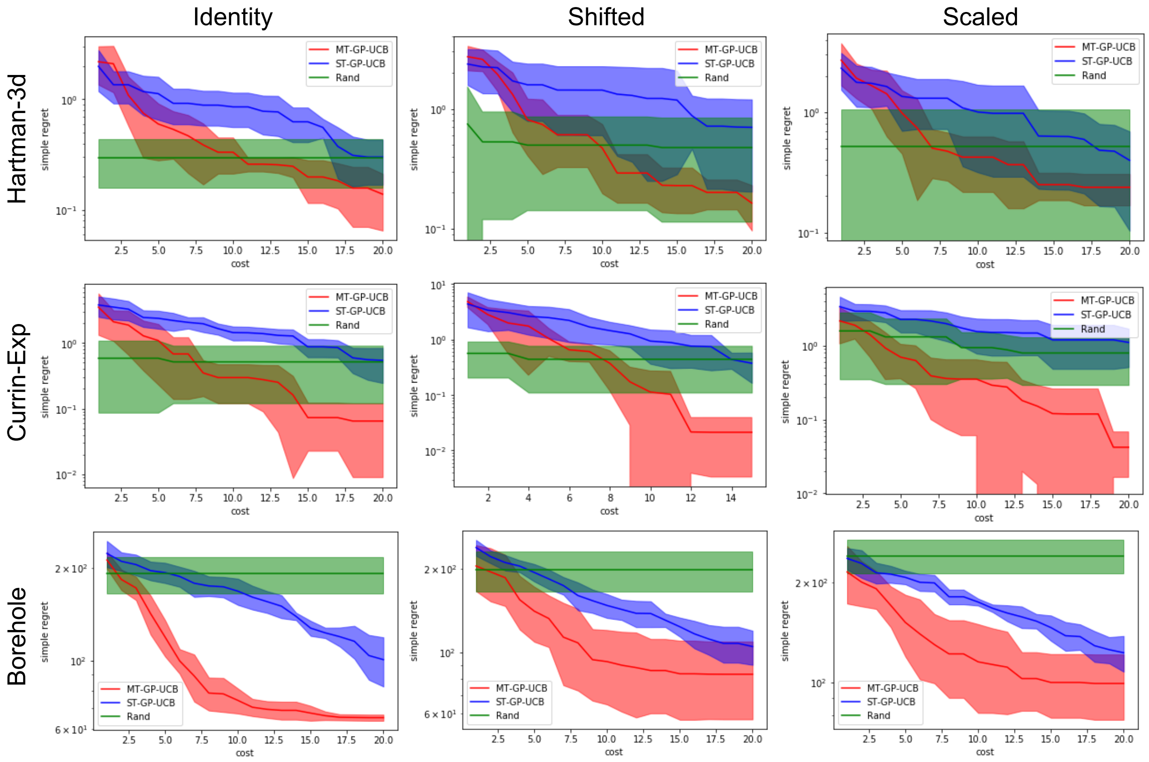

Figure 1: Simple regret of MT-GP-UCB (red), ST-GP-UCB (blue), and random sampling (green) on test functions with mean and standard deviation computed over 5 runs. For each experiment, we run with 2 tasks: one which is the untransformed function and the other is the function under either an identity, shift, or scaling transformation.

5.2 Nanophotonics

We experiment with a discrete nanophotonics design dataset presented in Song et al. (2018). The optimization objective is to find a nanophotonic structure which best satisfies a desired color filtering property. We are given a set of 499702 structures specified by 5 geometric properties and aim to minimize a measure called figure-of-merit (FOM), which measures how well a structure satisfies the color filtering property. We experiment with three design tasks for filtering light with wavelengths of 550 nm, 650 nm and 750 nm and assume that each task has cost $\frac{1}{3}$. We compare MT-GP-UCB against ST-GP-UCB and random sampling, and use $\beta = 1$ for both MT-GP-UCB and ST-GP-UCB. The simple regrets are shown in Figure 2A. We can see that the average simple regret over the 5 runs for MT-GP-UCB is lower than that of random sampling and ST-GP-UCB.

5.3 Hyperparameter tuning

Finally, we consider hyperparameter tunings of different machine learning models on the UCI iris dataset (Dua and Graff, 2017). Our tasks are tuning hyperparameters on 4 models: a KNN classifier, a support vector classifier (SVC) with RBF kernel, a SVC with linear kernel, and logistic regression. For the KNN classifier, we tune the number of neighbors, optimizing over the range $[2, 20]$. For SVC and logistic regression, we tune the regularization parameter, optimizing over the range $[0.01, 500]$. Our objective is to maximize the 10-fold cross validation (CV) accuracy of the models on each dataset. To compute regret, we assume the optimal value of the CV accuracy is 1. All tasks have $\frac{1}{4}$ cost. A comparison between MT-GP-UCB and ST-GP-UCB ($\beta = 1$) is shown in Figure 2B. We exclude random in our graph due to it's significantly worse performance (simple regret of about 0.08). We can see that MT-GP-UCB is able to achieve smaller simple regret compared to ST-GP-UCB.

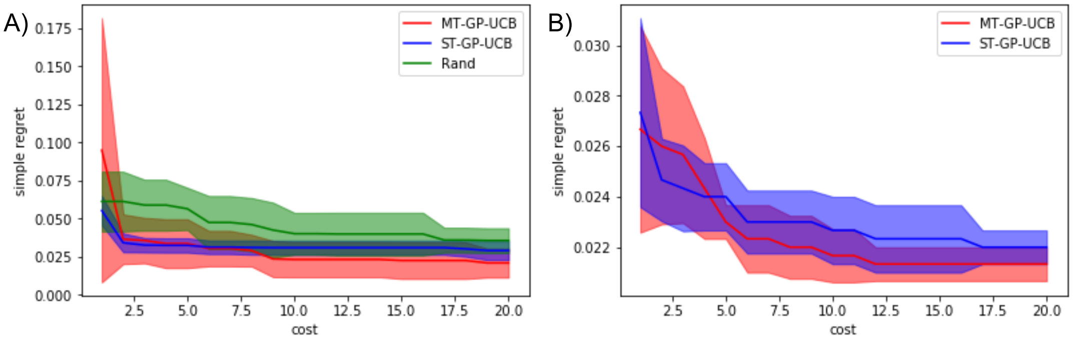

Figure 2: A) Simple regret of MT-GP-UCB, ST-GP-UCB, and random sampling on nanophotonics dataset with mean and standard deviation calculated over 5 runs. B) Simple regret of MT-GP-UCB and ST-GP-UCB on tuning hyperparameters for 4 different models trained on the iris dataset with mean and standard deviation calculated over 5 runs.

6. Conclusion and Future Work

We present a multi-task generalization of Bayesian optimization using GP-UCB, which utilizes multi-task Gaussian processes to model the joint distribution across tasks. We show that under mild assumptions, MT-GP-UCB is no-regret and provide experimental results on synthetic functions, nanophotonic design, and hyperparameter tuning. In the future, we would like to compare MT-GP-UCB to other multi-task Bayesian optimization methods such as contextual Gaussian process bandit optimization (Krause and Ong, 2011).

Appendix A. Proofs

We build our proofs on a few lemmas from (Srinivas et al., 2012), specifically Lemma 5.1 and 5.2 which establish a bound on instantaneous regret in terms of posterior variance. We state them here for completeness.

Lemma 3 *(Lemma 5.1 (Srinivas et al., 2012)) Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(nN\pi_t/\delta)$ where $\sum_{t\geq 1} \pi_t^{-1} = 1, \pi_t > 0$. Then*

$$|f^i(x) - \mu^i_{t-1}(x)| \leq \beta_t^{1/2}\sigma^i_{t-1}(x), \forall x \in \mathcal{X}, t \geq 1, 1 \leq i \leq n$$

holds with probability $\geq 1 - \delta$.

Lemma 4 *(Lemma 5.2 (Srinivas et al., 2012)) Fix $t \geq 1$. If $|f^i(x) - \mu^i_{t-1}(x)| \leq \beta_t^{1/2}\sigma^i_{t-1}(x)$, $\forall x \in \mathcal{X}, 1 \leq i \leq n$, then the regret $r_t \leq \frac{2}{n}\beta_t^{1/2}\sum_{i=1}^n \sigma^i_{t-1}(x^i_t)$.*

The original lemmas are proven for ST-GP-UCB. Since MT-GP-UCB acquisition rule is applied to each function separately while conditioned on all observations, the instantaneous regret above also holds. Next we connect the mutual information with the posterior variances.

Lemma 5 *The information gain from all queried points*

$$I(y^{[n]}_{[T]}; f^{[n]}) = \sum_{t=1}^T \sum_{i=1}^n \frac{1}{2}\log(1 + \sigma_i^{-2}(\sigma^i_{t-1})^2(x^i_t))$$

Proof The mutual information is obtained by revealing noisy evaluations of $f^{[n]}$ at chosen points $x^{[n]}_{[T]}$.

$$I(y^{[n]}_{[T]}; f^{[n]}) = I(y^{[n]}_{[T]}; f^{[n]}_{[T]}) = H(y^{[n]}_{[T]}) - H(y^{[n]}_{[T]}|f^{[n]}_{[T]})$$

As we assume that noise across tasks are mutually independent,

$$H(y^{[n]}_{[T]}|f^{[n]}_{[T]}) = \sum_{t=1}^T \sum_{i=1}^n \frac{1}{2}\log(2\pi e\sigma_i^2) \tag{1}$$

For the first term, we apply the chain rule for jointly entropy

$$\begin{aligned} H(y^{[n]}_{[T]}) &= H(y^{[n]}_T|y^{[n]}_{[T-1]}) + H(y^{[n]}_{[T-1]}) \\ &= \sum_{i=1}^n H(y^i_T|y^{[n]}_{[T-1]}) + H(y^{[n]}_{[T-1]}) \\ &= \sum_{i=1}^n \frac{1}{2}\log(2\pi e(\sigma_i^2 + (\sigma^i_T)^2(x^i_T))) + H(y^{[n]}_{[T-1]}) \\ &= \sum_{t=1}^T \sum_{i=1}^n \frac{1}{2}\log(2\pi e(\sigma_i^2 + (\sigma^i_t)^2(x^i_t))) \end{aligned} \tag{2}$$

with $C_1 :=$ Combining the results of Equations 1 and 2, we prove this lemma. ■

Now we present the proof for Theorem 1.

Proof For simplicity, we assume the noise variances are the same across tasks as σ^2. The modification to different variances are straightforward. By Lemma 3 and 4, we have that $r_t^2 \leq \frac{4}{n^2}\beta_t(\sum_{i=1}^n \sigma_{t-1}^i(x_t^i))^2, \forall t \geq 1$ with probability at least $1-\delta$.

From the AM-GM inequality, we know that

$$(\sum_{i=1}^n \sigma_{t-1}^i(x_t^i))^2 \leq n\sum_{i=1}^n (\sigma_{t-1}^i)^2(x_t^i)$$

it follows that

$$\begin{aligned} r_t^2 &\leq \frac{4}{n}\beta_t \sum_{i=1}^n (\sigma_{t-1}^i)^2(x_t^i) \\ &\leq \frac{4}{n}\beta_T \sum_{i=1}^n \sigma^2(\sigma^{-2}(\sigma_{t-1}^i)^2(x_t^i) \\ &\leq \frac{4}{n}\beta_T \sum_{i=1}^n \sigma^2 C_1 \log(1+\sigma^{-2}(\sigma_{t-1}^i)^2(x_t^i)) \\ &= \frac{4\sigma^2}{n}C_1\beta_T \sum_{i=1}^n \log(1+\sigma^{-2}(\sigma_{t-1}^i)^2(x_t^i)) \end{aligned}$$

with $C_1 := \frac{\sigma^{-2}}{\log(1+\sigma^{-2})} \geq 1$

As $R_T = \sum_{t=1}^T r_t$, we can apply the Cauchy-Schwarz inequality

$$\begin{aligned} R_T^2 &\leq T\sum_{t=1}^T r_t \\ &\leq T\sum_{t=1}^T \frac{4\sigma^2}{n}C_1\beta_T \sum_{i=1}^n \log(1+\sigma^{-2}(\sigma_{t-1}^i)^2(x_t^i)) \\ &= \frac{4\sigma^2}{n}C_1 T\beta_T \sum_{t=1}^T\sum_{i=1}^n \log(1+\sigma^{-2}(\sigma_{t-1}^i)^2(x_t^i)) \\ &= \frac{8\sigma^2}{n}C_1 T\beta_T I(y_{[T]}^{[n]}; f^{[n]}) \\ &\leq \frac{8\sigma^2}{n}C_1 T\beta_T\gamma_T \end{aligned}$$

So

$$R_T \leq \sqrt{CT\beta_T\gamma_T}$$

with $C := \frac{8\sigma^2}{n}C_1 = \frac{8\sigma^2}{n}\cdot\frac{\sigma^{-2}}{\log(1+\sigma^{-2})} = \frac{8}{n\log(1+\sigma^{-2})}$ ■

Finally, we provide the proof for Lemma 2.

Proof For any fixed $x_{[T]}^{[n]}$,

$$\begin{aligned} I(y_{[T]}^{[n]}; f^{[n]}) &= H(y_{[T]}^{[n]}) - H(y_{[T]}^{[n]} | f_{[T]}^{[n]}) \\ &= \sum_{i=1}^{n} H(y_{[T]}^{i} | y_{[T]}^{[i-1]}) - H(y_{[T]}^{[n]} | f_{[T]}^{[n]}) \\ &\leq \sum_{i=1}^{n} H(y_{[T]}^{i}) - H(y_{[T]}^{[n]} | f_{[T]}^{[n]}) \\ &= \sum_{i=1}^{n} H(y_{[T]}^{i}) - \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{1}{2} \log(2\pi e \sigma_i^2) \\ &= \sum_{i=1}^{n} H(y_{[T]}^{i}) - \sum_{i=1}^{n} H(y_{[T]}^{i} | f^i) \\ &= \sum_{i=1}^{n} I(y_{[T]}^{i}; f^i) \end{aligned}$$

It follows that the maximum mutual information γ_T is upper bounded by the sum of individual maximum mutual information $\sum_{i=1}^{n} \gamma_T^i$. ■

Appendix B. Additional Experiment Details and Results

B.1 Function Definitions

Below are the definitions of the synthetic example functions used in our experiments.

Hartman-3D The domain of this function is the unit cube $[0, 1]^3$. Hartman-3D is defined as

$$f(x) = -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{3} A_{ij} (x_j - P_{ij})^2 \right)$$

where $\alpha = (1.0, 1.2, 3.0, 3.2)^\intercal$, $A = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$, $P = 10^{-4} \begin{pmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{pmatrix}$

Currin Exponential Function The domain of the Currin exponential function is the unit square $[0, 1]^2$. The Currin exponential function is defined as

$$f(x) = \left(1 - \exp\left(\frac{-1}{2x_2} \right) \right) \left(\frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20} \right)$$

Borehole Function The input of the Borehole function has 8 dimensions with domain $[0.05, 0.15; 100, 50000; 6307, 115600; 990, 1110; 63.1, 116; 700, 820; 1120, 1680; 9855, 12045]$, which we scale to lie in $[0,1]^8$. The Borehole function is defined as

$$f(x) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}$$

B.2 Average Regret

Synthetic Functions Plots of the average regrets for each set of tasks over 5 runs are shown in Figure 3. We can see that MT-GP-UCB is able to achieve smaller average regret compared to ST-GP-UCB and random sampling.

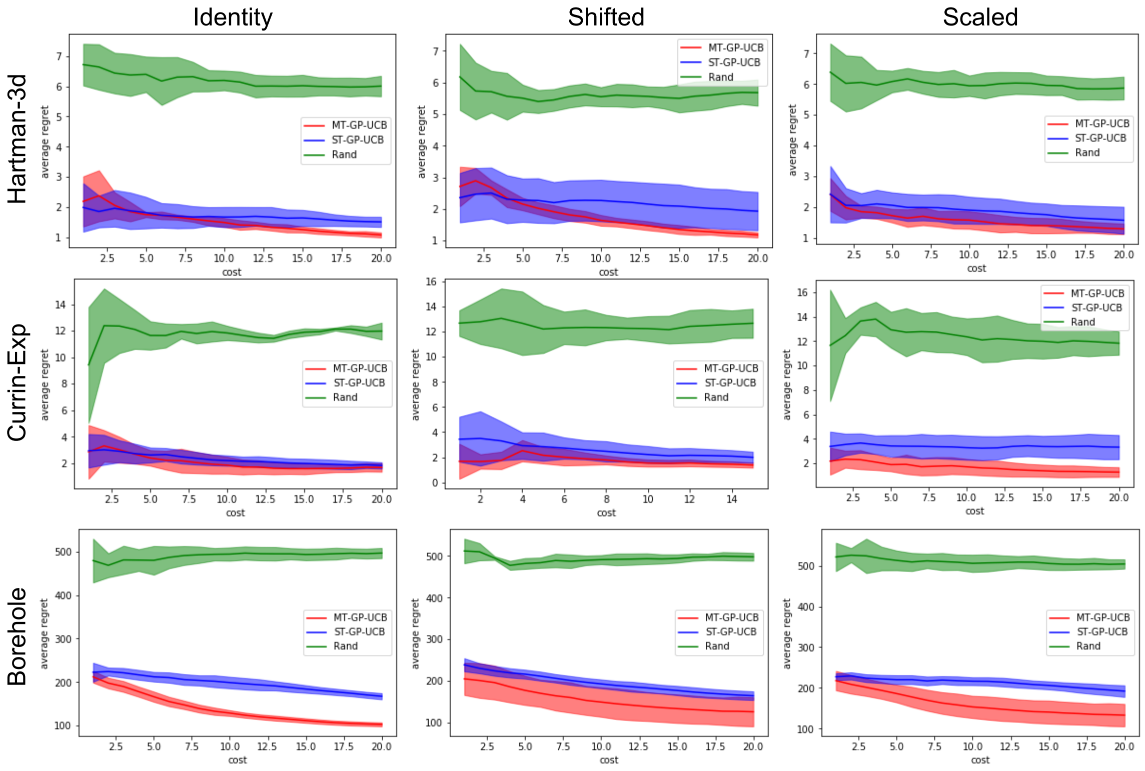

Figure 3: Average regret of MT-GP-UCB, ST-GP-UCB, and random sampling on transformed synthetic examples

Nanophotonics and Hyperparameter Tuning Plots of the average regrets for running MT-GP-UCB, ST-GP-UCB, and random sampling on the nanophotonics and hyperparameter tuning optimization problems are shown in Figure 4. We can see that ST-GP-UCB actually achieves smaller average regret despite higher simple regret, suggesting that MT-GP-UCB might explore more on that dataset. For the hyperparameter tuning optimization problem, we can see that the average regret of MT-GP-UCB is on average lower than that of ST-GP-UCB.

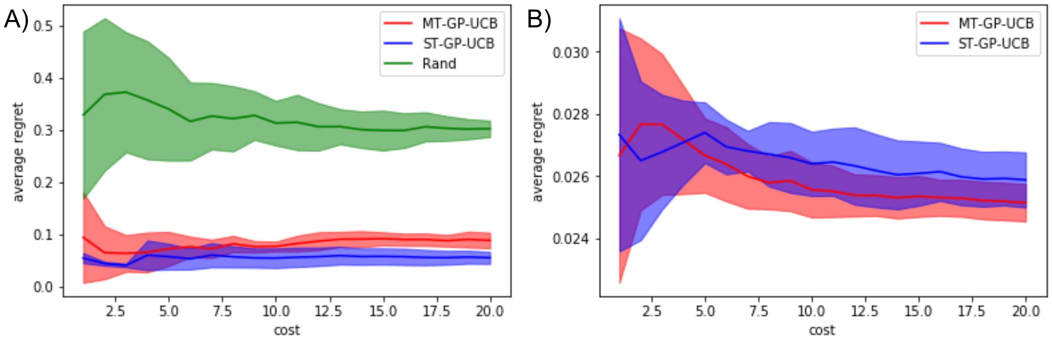

Figure 4: A) Average regret of MT-GP-UCB, ST-GP-UCB, and random sampling on nanophotonics dataset with mean and standard deviation calculated over 5 runs. B) Average regret of MT-GP-UCB and ST-GP-UCB on tuning hyperparameters for 4 different models trained on the iris dataset with mean and standard deviation calculated over 5 runs.

References

Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2008.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dagny Fleischman, Luke A Sweatlock, Hirotaka Murakami, and Harry Atwater. Hyper-selective plasmonic color filters. *Optics express*, 25(22):27386–27395, 2017.

Matthew Groves, Michael Pearce, and Juergen Branke. On parallelizing multi-task bayesian optimization. In *2018 Winter Simulation Conference (WSC)*, pages 1993–2002. IEEE, 2018.

Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837, 2012.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pages 2447–2455, 2011.

Jan Hendrik Metzen. Minimum regret search for single-and multi-task optimization. In *International Conference on Machine Learning*, pages 192–200, 2016.

Michael Pearce and Juergen Branke. Continuous multi-task bayesian optimisation with correlation. *European Journal of Operational Research*, 270(3):1074–1085, 2018.

Matthias Poloczek, Jialei Wang, and Peter I Frazier. Warm starting bayesian optimization. In *2016 Winter Simulation Conference (WSC)*, pages 770–781. IEEE, 2016.

Anil Ramachandran, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Information-theoretic transfer learning framework for bayesian optimisation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 827–842. Springer, 2018.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110 (3):E193–E201, 2013.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Jialin Song, Yury S. Tokpanov, Yuxin Chen, Dagny Fleischman, Kate T. Fountaine, Harry A. Atwater, and Yisong Yue. Optimizing photonic nanostructures via multi-fidelity gaussian processes, 2018.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2182033. URL `http://dx.doi.org/10.1109/TIT.2011.2182033`.

Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in neural information processing systems*, pages 2004–2012, 2013.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3627–3635. JMLR. org, 2017.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Kevin K Yang, Yuxin Chen, Alycia Lee, and Yisong Yue. Batched stochastic bayesian optimization via combinatorial constraints design. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3410–3419, 2019.

Yichi Zhang, Daniel W Apley, and Wei Chen. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific Reports*, 10(1):1–13, 2020.