Efficient Large-Scale Gaussian Process Bandits by Believing only Informative Actions

Amrit Singh Bedi AMRIT0714@GMAIL.COM
CISD, U.S. Army Research Laboratory 2800 Powder Mill Rd. Adelphi, MD 20783

Dheeraj Peddireddy ME12B1028@IITH.AC.IN
School of IE, Purdue University, 315 N. Grant Street, West Lafayette, IN 47907

Vaneet Aggarwal VANEET@PURDUE.EDU
School of IE and ECE, Purdue University, 315 N. Grant Street, West Lafayette, IN 47907

Alec Koppel ALEC.E.KOPPEL.CIV@MAIL.MIL
CISD, U.S. Army Research Laboratory 2800 Powder Mill Rd. Adelphi, MD 20783

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

Abstract

In this work, we cast Bayesian optimization as a multi-armed bandit problem, where the payoff function is sampled from a Gaussian process (GP). Further, we focus on action selections via the GP upper confidence bound (UCB). While numerous prior works use GPs in bandit settings, they do not apply to settings where the total number of iterations T may be large-scale, as the complexity of computing the posterior parameters scales cubically with the number of past observations. To circumvent this computational burden, we propose a simple statistical test: only incorporate an action into the GP posterior when its conditional entropy exceeds an ϵ threshold. Doing so permits us to derive sublinear regret bounds of GP bandit algorithms up to factors depending on the compression parameter ϵ for both discrete and continuous action sets. Moreover, the complexity of the GP posterior remains provably finite. Experimentally, we observe state of the art accuracy and complexity tradeoffs for GP bandit algorithms on various hyper-parameter tuning tasks, suggesting the merits of managing the complexity of GPs in bandit settings.

Keywords: multi-armed bandits, Bayesian optimization, Gaussian Processes, adaptive control.

1. Introduction

Bayesian optimization is a framework for global optimization of a black box function via noisy evaluations (Frazier, 2018), and provides an alternative to simulated annealing (Kirkpatrick et al., 1983; Bertsimas and Tsitsiklis, 1993) or exhaustive search (Davis, 1991). These methods have proven adept at hyper-parameter tuning of machine learning models (Snoek et al., 2012; Li et al., 2017), nonlinear system identification (Srivastava et al., 2013), experimental design (Chaloner and Verdinelli, 1995; Press, 2009), and semantic mapping (Shotton et al., 2008). More specifically, denote function $f : \mathcal{X} \to \mathbb{R}$ we seek to optimize via noisy samples, i.e., for a given $\mathbf{x}_t \in \mathcal{X}$, we observe $y_t = f(\mathbf{x}_t) + \epsilon_t$ sequentially. Our goal is to select a sequence of actions $\{\mathbf{x}_t\}$ that are competitive in performance with respect to the optimal selection $\mathbf{x}^* = \operatorname{argmax}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. For sequential decision making, a canonical performance metric is *regret*:

$$\mathbf{Reg}_T := \sum_{t=1}^{T} f(\mathbf{x}^*) - f(\mathbf{x}_t). \tag{1}$$

which quantifies the performance of a sequence of decisions $\{\mathbf{x}_t\}$ as compared with the optimal $\mathbf{x}^*$. We make no assumptions for now on the properties of f, other than each function evaluation must be selected judiciously. Regret in (1) is natural because at each time we quantify how far decision $\mathbf{x}_t$ was from optimal through the difference $r_t := f(\mathbf{x}^*) - f(\mathbf{x}_t)$. An algorithm eventually learns the optimal strategy if it is no-regret: $\mathbf{Reg}_T/T \to 0$ as $T \to \infty$.

In this work, we focus on Bayesian optimization, which hypothesizes a likelihood on the relationship between the unknown function $f(\mathbf{x})$ and action selection $\mathbf{x} \in \mathcal{X}$. Then upon selecting an action $\mathbf{x}$, one tracks a posterior distribution, or *belief model* (Powell and Ryzhov, 2012), over possible outcomes $y = f(\mathbf{x}) + \epsilon$ which informs how the next action is selected. In classical Bayesian inference, posterior distributions do not influence which samples $(\mathbf{x}, y)$ are observed next (Ghosal et al., 2000). In contrast, in multi-armed bandits, action selection $\mathbf{x}$ determines which observations form the posterior, which is why it is also referred to as *active learning* (Jamieson et al., 2015). The key distinguishing questions in this setting are the specification of a (i) likelihood and (ii) action selection strategy. These choices come with their own merits and drawbacks in terms of optimality and computational efficiency. Regarding (i) the likelihood model, when the action space $\mathcal{X}$ is discrete and of moderate size $X = |\mathcal{X}|$, one may track a probability for each element of $\mathcal{X}$, as in Thompson (posterior) sampling (Russo et al., 2018), Gittins indices (Gittins et al., 2011), and the Upper Confidence Bound (UCB) (Auer et al., 2002). These methods differ in their manner of action selection, but not distributional representation.

However, when the range of possibilities X is large, computational challenges arise. This is because the number of parameters one needs to define a posterior distribution over $\mathcal{X}$ is proportional to X, an instance of the curse of dimensionality in nonparametric statistics. One way to circumvent this issue for continuous spaces is to discretize the action space according to a pre-defined time-horizon that determines the total number of selected actions (Bubeck et al., 2011; Magureanu et al., 2014), and carefully tune the discretization to the time-horizon T. The drawback of these approaches is that as $T \to \infty$, the number of parameters in the posterior grows intractably large.

An alternative is to define a history-dependent distribution directly over the large (possibly continuous) space using, e.g., Gaussian Processes (GPs) (Rasmussen, 2004) or Monte Carlo (MC) methods (Smith, 2013). Bandit action selection strategies based on such distributional representations have been shown to be no-regret in recent years – see (Srinivas et al., 2012; Gopalan et al., 2014). While MC methods permit the most general priors on the unknown function f, computational and technical challenges arise when the prior/posterior no longer posses conjugacy properties (Gopalan et al., 2014). By contrast, GPs, stochastic processes any finite collection of realizations of which are jointly Gaussian (Krige, 1951), have a conjugate prior and posterior, and thus their parametric updates admit a closed form – see (Rasmussen, 2004)[Ch. 2].

This attribute of GPs has driven the development of bandit strategies of various kinds, such as the upper-confidence bound (UCB) (Srinivas et al., 2012; De Freitas et al., 2012), expected improvement (EI) (Wang and de Freitas, 2014; Nguyen et al., 2017), and step-wise uncertainty reduction (SUR) (Villemonteix et al., 2009), including knowledge gradient (Frazier et al., 2008), whose convergence in terms of regret or statistical consistency (Bect et al., 2019) may be established.

However, these convergence results hinge upon requiring use of the dense GP whose posterior distribution, through the mean and covariance (4), has complexity cubic in T due to the inversion of a Gram (kernel) matrix formed from the entire training set. This is an instance of the curse of dimensionality in nonparametric statistics. Numerous efforts to reduce the complexity of GPs exist in the literature – see (Csató and Opper, 2002; Bauer et al., 2016; Bui et al., 2017). These methods

Algorithm 1 Compressed GP-Bandits (COB)

```
for t = 1,2... do
  Select action $\mathbf{x}_t$ via UCB (3) or EI (3): $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x})$
  Sample: $y_t = f(\mathbf{x}_t) + \epsilon_t$
  If conditional entropy exceeds $\epsilon$ threshold $\mathbf{H}(y_t|\mathbf{y}_{t-1}) = \frac{1}{2}\log\left(2\pi e(\sigma^2 + \sigma^2_{\mathbf{D}_{t-1}}(\mathbf{x}_t))\right) > \epsilon$
    Augment dictionary $\mathbf{D}_t = [\mathbf{D}_{t-1}; \mathbf{x}_t]$
    Append $y_t$ to target vector $\mathbf{y}_{\mathbf{D}_t} = [\mathbf{y}_{\mathbf{D}_{t-1}}; y_t]$
    Update posterior mean $\mu_{\mathbf{D}_t}(\mathbf{x})$ & variance $\sigma_{\mathbf{D}_t}(\mathbf{x})$

        $\mu_{\mathbf{D}_t}(\mathbf{x}) = \boldsymbol{k}_{\mathbf{D}_t}(\mathbf{x})^T(\mathbf{K}_{\mathbf{D}_t} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_{\mathbf{D}_t}$
        $\sigma^2_{\mathbf{D}_t}(\mathbf{x}) = \kappa(\mathbf{x},\mathbf{x}') - \boldsymbol{k}_{\mathbf{D}_t}(\mathbf{x})^T(\mathbf{K}_{\mathbf{D}_t,\mathbf{D}_t} + \sigma^2\mathbf{I})^{-1}\boldsymbol{k}_{\mathbf{D}_t}(\mathbf{x}')$

  else  Fix dict. $\mathbf{D}_t = \mathbf{D}_{t-1}$, target $\mathbf{y}_{\mathbf{D}_t} = \mathbf{y}_{\mathbf{D}_{t-1}}$, & GP.

        $(\mu_{\mathbf{D}_t}(\mathbf{x}), \sigma_{\mathbf{D}_t}(\mathbf{x}), \mathbf{D}_t) = (\mu_{\mathbf{D}_{t-1}}(\mathbf{x}), \sigma_{\mathbf{D}_{t-1}}(\mathbf{x}), \mathbf{D}_{t-1})$

end for
```

all fix the complexity of the posterior and "project" all additional points onto a fixed likelihood "subspace." Doing so, however, may cause uncontrollable statistical bias and divergence.

By contrast, in this work, we propose a statistical test for the GP that explicitly trades off memory and regret (1), motivated by compression routines permit flexible representational complexity of nonparametric models (Koppel, 2019; Koppel et al., 2019; Elvira et al., 2016). Specifically, we propose a simple statistical test that operates inside GP-UCB which incorporates actions into the posterior only when conditional entropy exceeds an ϵ threshold (Sec. 2). We call this method Compressed GP-UCB, or *CUB* (Algorithm 1). Next, derive sublinear regret bounds of GP bandit algorithms up to factors depending on the compression parameter ϵ for both discrete and continuous action sets (Sec. 3). We further establish that the complexity of the GP posterior remains provably finite (Sec. 3). In the end, we experimentally employ these approaches for optimizing some simple non-convex functions and tuning the regularizer and step-size of a logistic regressor, which obtains a state of the art trade off in regret versus computational efficiency (Sec. 4).

2. Gaussian Process Bandits

Information Gain and Upper-Confidence Bound To find $\mathbf{x}^* = \text{argmax}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ when f is unknown, one may first globally approximate f well, and then evaluate it at the maximizer. In order to formalize this approach, we propose to quantify how informative a collection of points $\{\mathbf{x}_u\} \subset \mathcal{X}$ is through information gain (Cover and Thomas, 2012), a standard quantity that tracks the mutual information between f and observations $y_u = f(\mathbf{x}_u) + \epsilon_u$ all indices u in some sampling set, defined as $I(\{y_u\}; f) = H(\{y_u\}) - H(\{y_u\} \,\big|\, f)$ where $H(\{y_u\})$ denotes the entropy of observations $\{y_u\}$ and $H(\{y_u\} \,\big|\, f)$ denotes the entropy conditional on f. For a Gaussian $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance Σ, the entropy is given as $H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\log|2\pi e\Sigma|$ and the information gain is given in closed form as $I(\{y_u\}; f) = \frac{1}{2}\log|2 + \sigma^{-2}\mathbf{K}_t|$.

Suppose we are tasked with finding a subset of K points $\{\mathbf{x}_u\}_{u\leq T}$ that maximize the information gain. This amounts to a challenging subset selection problem whose exact solution cannot be found in polynomial time (Ko et al., 1995). However, near-optimal solutions may be obtained via greedy maximization, as information gain is submodular (Krause et al., 2008). Maximizing information gain, i.e., selecting $\mathbf{x}_t = \text{argmax}_{\mathbf{x}\in\mathcal{X}} I(\{y_u\}; f)$, is equivalent to (Srinivas et al., 2012)

$$\mathbf{x}_t = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}}\, \sigma_{\mathbf{X}_{t-1}}(\mathbf{x}) \tag{2}$$

where $\sigma_{\mathbf{X}_{t-1}}(\mathbf{x})$ is the empirical standard deviation associated with a matrix $\mathbf{X}_{t-1}$ of data points $\mathbf{X}_{t-1} := [\mathbf{x}_1 \ \cdots \mathbf{x}_{t-1}] \in \mathbb{R}^{d\times(t-1)}$. We note that (2) may be shown to obtain the near-optimal selection of points in the sense that after T rounds, executing (2) guarantees $I(\{y_u\}_{u=1}^T; f) \geq (1-1/e)I(\{y_u\}_{u=1}^K; f)$ for some $K \leq T$ points using the theory of submodular functions discussed in (Nemhauser et al., 1978). Indeed, selecting points based upon (2) permits one to efficiently *explore* f globally. However, it dictates that action selection does not move towards the actual maximizer $\mathbf{x}^*$ of f. For this, $\mathbf{x}_t$ should be chosen according to prior knowledge about the function f, *exploiting* information about where f is large. To balance between these two extremes, a number of different acquisition functions $\alpha(\mathbf{x})$ are possible based on the Gaussian Process posterior – see (Powell and Ryzhov, 2012; Nguyen et al., 2017). Here we propose an upper-confidence bound (UCB) based action selection with exploration parameter β_t as

$$\mathbf{x}_t = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}}\, \mu_{\mathbf{X}_{t-1}}(\mathbf{x}) + \sqrt{\beta_t}\sigma_{\mathbf{X}_{t-1}}(\mathbf{x}) \tag{3}$$

where the mean and standard deviation are given by a GP to be defined next.

Gaussian Processes A Gaussian Process (GP) is a stochastic process for which every finite collection of realizations is jointly Gaussian. We hypothesize a Gaussian Process prior for $f(\mathbf{x})$, which is specified by a mean function $\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and covariance kernel defined as $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - \mu(\mathbf{x}))^T(f(\mathbf{x}') - \mu(\mathbf{x}'))\right]$. Subsequently, we assume the prior is zero-mean $\mu(\mathbf{x}) = 0$.

GPs play multiple roles in this work: as a way of specifying smoothness and a prior for unknown function f, as well as characterizing regret when f is a sample from a known GP $GP(\mathbf{0}; \kappa(\mathbf{x}; \mathbf{x}'))$. GPs exhibit the ability to admit a closed form for their maximum a posteriori conditional mean and covariance given training set $\mathcal{S}_t = \{\mathbf{x}_u, \mathbf{y}_u\}_{u\leq t}$ as (Rasmussen, 2004)[Ch. 2].

$$\mu_{\mathbf{X}_t}(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{y}_t, \;\; \sigma^2_{\mathbf{X}_t}(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x}')^T \tag{4}$$

where $\mathbf{k}_t(\mathbf{x}) = [\kappa(\mathbf{x}_1, \mathbf{x}), \cdots, \kappa(\mathbf{x}_t, \mathbf{x})]$ denotes the empirical kernel map and $\mathbf{K}_t$ denotes the gram matrix of kernel evaluations whose entries are $\kappa(\mathbf{x}, \mathbf{x}')$ for $\mathbf{x}, \mathbf{x}' \in \{\mathbf{x}_u\}_{u\leq t}$. The $\mathbf{X}_t$ subscript underscores its role in parameterizing the mean and covariance. Further, note that (4) depends upon a linear observation model $y_t = f(\mathbf{x}_t) + \epsilon_t$ with Gaussian noise prior $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. The parametric updates (4) depend on past actions $\{\mathbf{x}_u\}_{\{u\leq t\}}$, which causes the *kernel dictionary* $\mathbf{X}_t$ to grow by one at each iteration, i.e., $\mathbf{X}_{t+1} = [\mathbf{X}_t \ ; \ \mathbf{x}_{t+1}] \in \mathbb{R}^{d\times t}$, and that the posterior at time $t+1$ uses *all past observations* $\{\mathbf{x}_u\}_{u\leq t}$. Subsequently, we refer to the number of columns in the dictionary matrix as the *model order*. The GP posterior at time $t+1$ has model order t.

The resulting action selection strategy (3) using the GP (4) is called GP-UCB, and its regret (1) is established in (Srinivas et al., 2012)[Theorem 1 and 2] as sublinear with high probability up to factors depending on the maximum information gain γ_T over T points, which is defined as

$$\gamma_T := \max_{\{\mathbf{x}_u\}} I(\{y_u\}_{u=1}; f) \;\; \text{such that} \;\; |\{\mathbf{x}_u\}| = T. \tag{5}$$

Compression Statistic The fundamental role of information gain in the regret of GP-UCB provides a conceptual basis for finding a parsimonious GP posterior that nearly preserves no-regret properties of (3) - (4). To define our compression rule, first we define some key quantities related to approximate GPs. Suppose we select some other kernel dictionary $\mathbf{D} \in \mathbb{R}^{d\times M}$ rather than $\mathbf{X}_t$ at time t, where M is the *model order* of the Gaussian Process. Then, the only difference is that the kernel matrix $\mathbf{K}_t$ in (4) and the empirical kernel map $\mathbf{k}_t(\cdot)$ are substituted by $\mathbf{K}_{\mathbf{DD}}$ and $\mathbf{k}_{\mathbf{D}}(\cdot)$, respectively, where the entries of $[\mathbf{K}_{\mathbf{DD}}]_{mn} = \kappa(\mathbf{d}_m, \mathbf{d}_n)$ and $\{\mathbf{d}_m\}_{m=1}^M \subset \{\mathbf{x}_u\}_{u\leq t}$. Further, $\mathbf{y}_{\mathbf{D}}$ denotes the sub-vector of $\mathbf{y}_t$ associated with only the indices of training points in matrix $\mathbf{D}$. We denote the training subset associated with these indices as $\mathcal{S}_{\mathbf{D}} := \{\mathbf{x}_u, y_u\}_{u=1}^M$. If we rewrite (4) with $\mathbf{D}$ as the dictionary rather than $\mathbf{X}_{t+1}$, we obtain

$$\boldsymbol{\mu}_{\mathbf{D}}(\mathbf{x}) = \mathbf{k}_{\mathbf{D}}(\mathbf{x}_{t+1})[\mathbf{K}_{\mathbf{D},\mathbf{D}} + \sigma^2\mathbf{I}]^{-1}\mathbf{y}_{\mathbf{D}}, \;\; \sigma^2_{\mathbf{D}}(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}') - \boldsymbol{k}_{\mathbf{D}}(\mathbf{x})^T(\mathbf{K}_{\mathbf{D},\mathbf{D}} + \sigma^2\mathbf{I})^{-1}\boldsymbol{\kappa}_{\mathbf{D}}(\mathbf{x}'). \quad (6)$$

The question, then, is how to select a sequence of dictionaries $\mathbf{D}_t \in \mathbb{R}^{p\times M_t}$ whose M_t columns comprise a subset of those of $\mathbf{X}_t$ in such a way to approximately preserve the regret bounds of (Srinivas et al., 2012)[Theorem 1 and 2] while ensuring the model order is moderate $M_t \ll t$.

We propose using conditional entropy as a statistic to compress against, i.e., a new data point should be appended to the Gaussian process posterior only when its conditional entropy is at least ϵ, which results in the following update rule for the dictionary $\mathbf{D}_t \in \mathbb{R}^{p\times M_t}$:

$$\begin{aligned} &\mathbf{D}_t = [\mathbf{D}_{t-1} \; ; \; \mathbf{x}_t] && \text{whenever } \mathbf{H}(y_t|\hat{\mathbf{y}}_{t-1}) = \frac{1}{2}\log\left(2\pi e(\sigma^2 + \sigma^2_{\mathbf{D}_{t-1}}(\mathbf{x}_t))\right) > \epsilon \\ &\mathbf{D}_t = \mathbf{D}_{t-1} && \text{otherwise} \end{aligned} \quad (7)$$

where we define ϵ as the compression budget. This amounts to a statistical test of whether the action $\mathbf{x}_t$ yielded an informative sample y_t in the sense that its conditional entropy exceeds an ϵ threshold. Therefore, uninformative past decisions are dropped from belief formation about the present. The modification of GP-UCB, called *Compressed GP-UCB*, or *CUB* for short, uses (3) with the lazy GP belief model (6) defined by dictionary updates (7), and is summarized as Algorithm 1. Next, we rigorously establish how Algorithm 1 trades off regret and memory through the ϵ threshold on conditional entropy for whether a point $(\mathbf{x}_t, y_t)$ should be included in the GP.

3. Balancing Regret and Complexity

In this section, we establish that Algorithm 1 attains comparable regret (1) to the standard GP-UCB approach to bandit optimization under the canonical settings of the action space $\mathcal{X}$ being a discrete finite set and a continuous compact Euclidean subset. We build upon techniques pioneered in (Srinivas et al., 2012). The points of departure in our analysis are: (i) the characterization of statistical bias induced by the compression rule (7) in the regret bounds, and (ii) the relating of properties of the posterior (7) and action selections (3) to topological properties of the action space $\mathcal{X}$ to ensure the model order of the GP defined by (6) is at-worst finite for all t. Next we present our main convergence result regarding Algorithm 1.

Theorem 1 **(Regret of Compressed GP-UCB)** *Fix $\delta \in (0, 1)$ and suppose the Gaussian Process prior for f has zero mean with covariance kernel $\kappa(\mathbf{x}, \mathbf{x}')$. Define constant $C := 8/\log(1 + \sigma^{-2})$ Then under the following parameter selections and conditions on the data domain $\mathcal{X}$, we have:*

i. ***(Finite decision set)*** *For finite cardinality $|\mathcal{X}| = X$, with exploration parameter β_t selected as $\beta_t = 2\log(Xt^2\pi^2/6\delta)$, the accumulated regret is sublinear regret with probability $1-\delta$. This implies that $\mathbb{P}\left\{\textbf{\textit{Reg}}_T \leq \sqrt{C_1 T \beta_T \hat{\gamma}_T} + \sqrt{\epsilon} T\right\} \geq 1 - \delta$ where ϵ is the compression budget.*

ii. ***(General decision set)*** *For continuous set $\mathcal{X} \subset [0, r]^d$, assume the derivative of the GP sample paths are bounded with high probability, i.e., for constants a, b, $\mathbb{P}\left\{\sup_{\mathbf{x}\in\mathcal{X}} |\partial f / \partial \mathbf{x}_j| > L\right\} \leq ae^{-(L/b)^2}$ for $j = 1, .., d$. Then, under exploration parameter $\beta_t = 2\log(Xt^2\pi^2/3\delta) + 2d\log(t^2 dbr\sqrt{\log(4da/\delta)})$, we have $\mathbb{P}\left\{\textbf{\textit{Reg}}_T \leq \sqrt{C_1 T \beta_T \hat{\gamma}_T} + \sqrt{\epsilon} T + \frac{\pi^2}{6}\right\} \geq 1 - \delta$.*

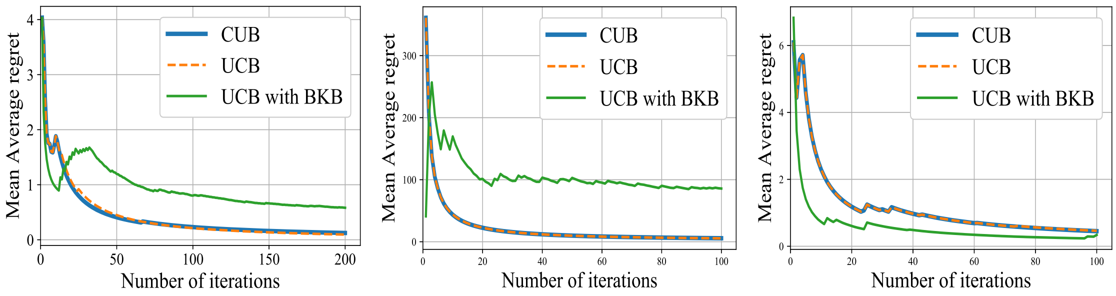
Theorem 1, whose proof is detailed in (Bedi et al., 2020), establishes that Algorithm 1 attains sublinear regret with high probability when the action space $\mathcal{X}$ is discrete and finite, as well as when it is a continuous compact subset of Euclidean space, up to factors depending on the maximum information gain (5) and the compression budget ϵ in (7). The sublinear dependence of the information gain on T in terms of the parameter dimension d is derived in (Srinivas et al., 2012)[Sec. V-B] for common kernels such as the linear, Gaussian, and Matérn.

The proof follows a path charted in (Srinivas et al., 2012)[Appendix I], except that we must contend with the compression-induced error. Specifically, we begin by computing the confidence interval for each action $\mathbf{x}_t$ taken by the proposed algorithm at time t. Then, we bound the instantaneous regret $r_t := f(\mathbf{x}^*) - f(\hat{\mathbf{x}}_t)$ in terms of the problem parameters such as β_t, δ, C, compression budget ϵ, and information gain γ_T using the fact that the upper-confidence bound overshoots the maximizer. By summing over time with Cauchy-Schwartz, we build an upper-estimate of cumulative regret based on instantaneous regret r_t. Unsurprisingly, an additional term appears due to our compression budget ϵ in the final regret bounds, which for $\epsilon = 0$ reduces to (Srinivas et al., 2012)[Theorem 1 and 2]. However, rather than permitting the complexity of the GP to grow unbounded with T, instead it grows only when informative actions are taken, and preserves the sublinear growth of regret for any ϵ such that $\sqrt{\epsilon}T = o(T)$ such as $\epsilon = T^{2(p-1)}$ for any $p < 1$.

Next, we establish the main merit of doing this statistical test inside a bandit algorithm is that it controls the complexity of the belief model that decides action selections. In particular, Theorem 2 formalizes that the dictionary $\mathbf{D}_t$ defined by (7) in Algorithm 1 will always have finite number of elements $M_T(\epsilon)$ even if $T \to \infty$, which is stated next.

Theorem 2 *Suppose that the conditional entropy $H(\{y_t\} \mid f)$ is bounded for all t. Then, the number of elements in the dictionary $\mathbf{D}_T$ denoted by $M_T(\epsilon)$ in the GP posterior of Algorithm 1 is finite as $T \to \infty$ for fixed compression threshold ϵ.*

The implications of Theorem 2 (see (Bedi et al., 2020) for proof) are that the algorithm only the stores important actions in the belief model and drops extraneous points. Interestingly, this result states that despite infinitely many actions being taken in the limit, only finite many of them are ϵ-informative. In principle, one could make ϵ adaptive with t to improve performance, but analyzing such a choice becomes complicated as relating the worst-cast model complexity to the covering number of the space $\mathcal{X}$ would then depend on variable sets whose conditional entropy is at least ϵ. In the next section, we evaluate the merit of these conceptual results on experimental settings involving black box non-convex optimization and hyper-parameter tuning of linear logistic regressors.



(a) Example func. (b) Rosenbrock func. (c) Hyper. para. tuning

Figure 1: Figure shows mean average regret vs iteration performance comparison for the proposed algorithm under different experimental settings.

4. Experiments

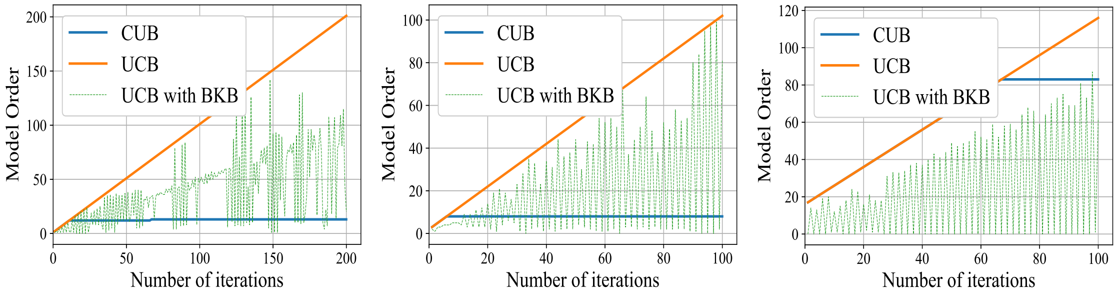
In this section, we evaluate the performance of the proposed statistical compression method under a few different action selections (acquisition functions). Specifically, Algorithm 1 employs the Upper Confidence Bound (UCB) acquisition function, but the key insight here is a modification of the GP posterior, not the action selection. Thus, we validate its use for Most Probable Improvement (MPI) (Wang and de Freitas, 2014) as well, defined as $\alpha^{\text{MPI}}(\mathbf{x}) = \sigma_{t-1}\phi(z) + [\mu_{t-1}(\mathbf{x}) - \xi]\Phi(z)$ and $\xi = \operatorname{argmax}_{\mathbf{x}} \mu_{t-1}(\mathbf{x})$, where $\phi(z)$ and $\Phi(z)$ denote the standard Gaussian density and distribution functions, and $z = (\mu_{t-1}(\mathbf{x}) - \xi)/\sigma_{t-1}(\mathbf{x})$ is the centered z-score. We further compare the compression scheme against Budgeted Kernel Bandits (BKB) proposed by (Calandriello et al., 2019) which proposes to randomly add or drop points according to a distribution that is inversely proportional to the posterior variance, also on the aforementioned acquisition functions.

Unless otherwise specified, the squared exponential kernel is used to represent the correlation between the input, the lengthscale is set to θ = 1.0, the noise prior is set to σ^2 = 0.001, the compression budget $\epsilon = 10^{-4}$, and the confidence bounds hold with probability of at least δ = 0.9. As a common practice across all three problems, we initialize the Gaussian priors with 2^d training data randomly collected from the input domain, where d is the input dimension. We quantify the performance using Mean Average Regret over the iterations. In addition, the model order, or number of points defining the GP posterior, is visualized over time to characterize the compression of the training dictionary. To ensure fair evaluations, all the listed simulations were performed on a PC with 1.8 GHz Intel Core i7 CPU and 16 GB memory. The details about the different experimental settings is provided below.

1) **Example Function:** First, we evaluate our proposed method on an example function given by $f(x) = \sin(x) + \cos(x) + 0.1x$. A zero mean unit variance random Gaussian noise is induced at every observation of f, to emulate the black box scenarios.

2) **Rosenbrock Function:** For the second simulation, we compare the compressed variants with their baseline algorithm on a 2 dimensional non-convex function popularly known as the Rosenbrock function, given by $f(x, y) = (a - x)^2 + b(y - x^2)^2$, with $a = 1$ and $b = 10$.

3) **Hyperparameter Tuning in Logistic Regression:** We propose to use bandit algorithms to automate the hyper-parameter tuning of machine learning algorithms. More specifically, we propose using Algorithm 1 and variants with different acquisition functions to tune the following hyper-parameters of a supervised learning scheme, whose concatenation forms the action space:



(*a*) Example func. (*b*) Rosenbrock func. (*c*) Hyper. para. tuning

Figure 2: Figure shows model order vs iteration performance comparison for the proposed algorithm under different experimental settings. We note the significant performance benefit in terms of the model order.

the learning rate, batch size, dropout of the inputs, and the ℓ_2 regularization constant. The specific supervised learning problem we focus on is the training of a multi-class logistic regressors over the MNIST training set (LeCun and Cortes, 2010) for classifying hand written digits. The instantaneous reward here is the statistical accuracy on a hold-out validation set.

The results of this implementation are given in Figure 1 and Figure 2. Observe that the compression technique (7) yields algorithms whose regret is comparable to the dense GP, with a significant reduction in model complexity that eventually settles to a constant. This constant is a fundamental measure of the complexity of the action space required for finding a no-regret policy. Overall, then, one can run Algorithm 1 on the back-end of any training scheme for supervised learning in order to automate the selection of hyper-parameters in perpetuity without worrying about eventual slowdown.

5. Conclusions

We considered bandit problems whose action spaces are discrete but have large cardinality, or are continuous. Following a number of previous works for bandits with large action spaces, we parameterized the action distribution as a Gaussian Process in order to have a closed form expression for the a posteriori variance. Unfortunately, Gaussian Processes exhibit complexity challenges when operating ad infinitum: the complexity of computing posterior parameters grows cubically with the time index. While numerous previous memory-reduction methods exist for GPs, designing compression for bandit optimization is relatively unexplored. Within this gap, we proposed a compression rule for the GP posterior explicitly derived by information-theoretic regret bounds, where the conditional entropy encapsulates the per-step progress of the bandit algorithm. This compression only includes past actions whose conditional entropy exceeds an ϵ-threshold to enter into the posterior.

As a result, we derived explicit tradeoffs between model complexity and information-theoretic regret. Moreover, the complexity of the resulting GP posterior is at worst finite and depends on the covering number (metric entropy) of the action space, a fundamental constant that determines the bandit problem's difficulty. In experiments, we observed a favorable tradeoff between regret, model complexity, and iteration index/clock time for a couple toy non-convex optimization problems as well as the actual problem of how to tune hyper-parameters of a supervised machine learning model.

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse gaussian process approximations. In *Advances in neural information processing systems*, pages 1533–1541, 2016.

Julien Bect, François Bachoc, David Ginsbourger, et al. A supermartingale approach to gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

Amrit Singh Bedi, Dheeraj Peddireddy, Vaneet Aggarwal, and Alec Koppel. Efficient gaussian process bandits by believing only informative actions. *arXiv preprint arXiv:2003.10550*, 2020.

Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011.

Thang D Bui, Cuong Nguyen, and Richard E Turner. Streaming sparse gaussian process approximations. In *Advances in Neural Information Processing Systems*, pages 3301–3309, 2017.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557, 2019.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Lehel Csató and Manfred Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

Lawrence Davis. Handbook of genetic algorithms. 1991.

Nando De Freitas, Alex Smola, and Masrour Zoghi. Exponential regret bounds for gaussian process bandits with deterministic observations. *arXiv preprint arXiv:1206.6457*, 2012.

Víctor Elvira, Joaquín Míguez, and Petar M Djurić. Adapting the number of particles in sequential monte carlo methods through an online scheme for convergence assessment. *IEEE Transactions on Signal Processing*, 65(7):1781–1794, 2016.

Y. Engel, S. Mannor, and R. Meir. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*, 52(8):2275–2285, Aug 2004. ISSN 1941-0476. doi: 10.1109/TSP.2004.830985.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

Subhashis Ghosal, Jayanta K Ghosh, Aad W Van Der Vaart, et al. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.

Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in neural information processing systems*, pages 2656–2664, 2015.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220 (4598):671–680, 1983.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Learning curve prediction with bayesian neural networks. In *ICLR*, 2017.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

Alec Koppel. Consistent online gaussian process regression without the sample complexity bottleneck. In *2019 American Control Conference (ACC)*, pages 3512–3518. IEEE, 2019.

Alec Koppel, Amrit Singh Bedi, Ketan Rajawat, and Brian M Sadler. Optimally compressed nonparametric online learning. *arXiv preprint arXiv:1909.11555*, 2019.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: a novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18 (1):6765–6816, 2017.

Stefan Magureanu, Richard Combes, and Alexandre Proutière. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *COLT 2014*, 2014.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions?i. *Mathematical programming*, 14(1):265–294, 1978.

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Asian Conference on Machine Learning*, pages 279–294, 2017.

Warren B Powell and Ilya O Ryzhov. *Optimal learning*, volume 841. John Wiley & Sons, 2012.

William H Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392, 2009.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

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

Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

Adrian Smith. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 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.

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, 2012.

Vaibhav Srivastava, Paul Reverdy, and Naomi E Leonard. On optimal foraging and multi-armed bandits. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 494–499. IEEE, 2013.

Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509, 2009.

Ziyu Wang and Nando de Freitas. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014.