CSE599i: Online and Adaptive Machine Learning **Winter 2018**

Lecture 13: Gaussian Process Optimization

Lecturer: Kevin Jamieson *Scribes: Rahul Nadkarni, Brian Hou, Aditya Mandalika*

In these notes, we will introduce the Gaussian Process Upper Confidence Bound (GP-UCB) algorithm and bound the regret of the algorithm. First, we introduce the property of *submodularity* in Section 1.1, one of the tools that is necessary to prove these regret bounds. Next, we review *Gaussian processes* in Section 1.2.

1 Preliminaries

1.1 Submodularity

Consider a finite set of objects Ω and an objective function f that assigns each subset $S \subseteq \Omega$ a non-negative value $f(S)$. Finding the subset that maximizes f is intractable in general because all $2^{|\Omega|}$ subsets would need to be checked, but if f is both *submodular* and *monotone*, we can prove that a greedy algorithm will achieve at least $(1 - \frac{1}{e})$ of the optimum.[1]

Definition 1 (Discrete Derivative)**.** *For a set function f, $S \subseteq \Omega$, and $x \in \Omega$, let $\Delta_f(x|S) = f(S \cup \{x\}) - f(S)$ be the discrete derivative of f at S with respect to x. This is the marginal gain of adding x to the set S.*

Definition 2 (Submodularity)**.** *A set function f is submodular if for all $S \subseteq T \subseteq \Omega$ and $x \in \Omega \setminus T$,*

$$\Delta_f(x|S) \geq \Delta_f(x|T)$$

Intuitively, submodularity is equivalent to diminishing returns: the marginal benefit of adding a new element x to a set T is no greater than the marginal benefit of adding that element to a subset S.

Definition 3 (Monotonicity)**.** *A set function f is monotone if for all $S \subseteq T \subseteq \Omega$,*

$$f(S) \leq f(T)$$

A function is monotone if all of its discrete derivatives are non-negative, i.e. $\Delta_f(x|S) \geq 0 \ \forall x \in \Omega$. Intuitively, this means that adding an element to the set will never reduce the objective, which is important if we are greedily optimizing the objective as described in the following section.

Definition 4 (Monotone Submodularity)**.** *A set function f is monotone submodular if for all $S \subseteq T \subseteq \Omega$ and $x \in \Omega$,*

$$\Delta_f(x|S) \geq \Delta_f(x|T)$$

This is slightly different than the statement in Definition 2 because x is permitted to be in T.

[1] This section is adapted from Krause and Golovin [2012] and Kun [2014].

1.1.1 Submodular Optimization

Suppose we want to find the subset of at most $k < |\Omega|$ elements, $S_k^* \subset \Omega$, that maximizes a non-negative monotone submodular set function f. Algorithm 1 describes a greedy algorithm for computing S_k.

Algorithm 1 A greedy algorithm for submodular optimization.

Input: Monotone submodular function f; Ω; k
$S_0 \leftarrow \{\}$
for $t = 1, 2, \ldots, k$ **do**
  $x_t \leftarrow \arg\max_{x \in \Omega \setminus S_{t-1}} \Delta_f(x|S_{t-1})$
  $S_t \leftarrow S_{t-1} \cup \{x_t\}$
end for
return S_k

To compare the objective values for the set S_k returned by the greedy algorithm and S_k^*, we will instead prove a more general case. We permit the greedy algorithm to run for ℓ iterations (yielding $S_1, S_2, \ldots, S_\ell$) where ℓ does not necessarily equal k.

Theorem 1 (Nemhauser et al. 1978)**.** *Let f be a monotone submodular set function and S_k^* be the subset of at most k elements that maximizes f. After running the greedy algorithm for ℓ iterations,*

$$f(S_\ell) \geq (1 - e^{-\ell/k}) f(S_k^*)$$

Proof. Let $S_k^* = \{x_1^*, \ldots, x_k^*\}$ be the optimal set of size k.

$$\begin{aligned}
f(S_k^*) &\leq f(S_t \cup \{x_1^* \ldots x_k^*\}) && (1)\\
&= f(S_t \cup \{x_1^* \ldots x_k^*\}) - f(S_t \cup \{x_1^* \ldots x_{k-1}^*\}) + f(S_t \cup \{x_1^* \ldots x_{k-1}^*\}) && (2)\\
&= \Delta_f(x_k^* | S_t \cup \{x_1^* \ldots x_{k-1}^*\}) + f(S_t \cup \{x_1^* \ldots x_{k-1}^*\}) && (3)\\
&= f(S_t) + \sum_{j=1}^{k} \Delta(x_j^* | S_t \cup \{x_1^*, \ldots, x_{j-1}^*\}) && (4)\\
&\leq f(S_t) + \sum_{x \in S_k^*} \Delta_f(x|S_t) && (5)\\
&\leq f(S_t) + \sum_{x \in S_k^*} (f(S_{t+1}) - f(S_t)) && (6)\\
&\leq f(S_t) + k(f(S_{t+1}) - f(S_t)) && (7)
\end{aligned}$$

Equation (1) follows from monotonicity of f. Equations (2) to (4) are a telescoping sum. Equation (5) follows from submodularity of f. Equation (6) is because S_{t+1} is greedily constructed by maximizing $\Delta_f(x|S_t)$. Equation (7) follows because S_k^* contains at most k elements.

We can rearrange Equation (7) to describe how the gap between the optimal set S_k^* and the greedily-constructed set S_t shrinks as the greedy algorithm runs for more iterations. This inequality holds for all $t \geq 0$.

$$\begin{aligned}
f(S_k^*) - f(S_t) &\leq k(f(S_k^*) - f(S_t)) - k(f(S_k^*) - f(S_{t+1})) && (8)\\
f(S_k^*) - f(S_{t+1}) &\leq \left(1 - \frac{1}{k}\right)(f(S_k^*) - f(S_t)) && (9)
\end{aligned}$$

Starting from $t = 0$ and applying Equation (9) ℓ times yields

$$f(S_k^*) - f(S_\ell) \leq \left(1 - \frac{1}{k}\right)^\ell (f(S_k^*) - f(S_0)) \qquad (10)$$

Because f is non-negative, we can describe the gap directly in terms of $f(S_k^*)$.

$$f(S_k^*) - f(S_\ell) \leq \left(1 - \frac{1}{k}\right)^\ell f(S_k^*) \tag{11}$$

Finally, we apply the common inequality $1 - x \leq e^{-x}$ to simplify, rearranging to yield our desired result.

$$f(S_k^*) - f(S_\ell) \leq e^{-\ell/k} f(S_k^*) \tag{12}$$

$$f(S_\ell) \geq (1 - e^{-\ell/k}) f(S_k^*) \tag{13}$$

Setting $\ell = k$ shows that the greedy algorithm (Algorithm 1) is approximately optimal.

$$f(S_k) \geq \left(1 - \frac{1}{e}\right) f(S_k^*) \tag{14}$$

□

1.2 Gaussian Processes

Gaussian processes generalize the concept of a Gaussian distribution over discrete random variables to the idea of a Gaussian distribution over continuous functions.[2]

1.2.1 Intuition and Motivation

Suppose that we are given some data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, as well as an estimate of the Gaussian noise model at each $\mathbf{x}_i$. If we fit a continuous function to those data points, we may also hope to estimate the noise at interpolated points. Intuitively, nearby points along the fitted function will have a similar noise distribution, but adjusted by the distance between them.

Just as a Gaussian distribution is fully described by its mean and covariance matrix, a Gaussian process is fully described by its mean and covariance function. See Figure 1 for an example of how the mean and covariance change as additional points are provided to the Gaussian process.

A Gaussian process is parametrized by a *kernel function*, which determines which $\mathbf{x}_i$ are close. One common choice for the kernel function is the squared exponential, or radial basis function (RBF).

$$k_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma_s^2} \|\mathbf{x} - \mathbf{x}'\|^2\right)$$

Figure 2 shows Gaussian processes with $\sigma_s = 0.1$, 1, and 2, respectively. For the RBF kernel, a higher choice of σ_s results in smoother samples from the Gaussian process.

Gaussian processes allow us to model nonlinear functions in a particularly appealing way: the covariance function provides a metric of confidence in the estimated mean function's accuracy. This nonlinearity can be used to extend the linear contextual bandit scenario to nonlinear contextual bandits. In addition, the covariance function yields an Upper Confidence Bound (UCB) algorithm that is analogous to the linear setting (LinUCB) described in Lecture 10.

[2] This section is adapted from Rasmussen and Williams [2006, Ch. 2].

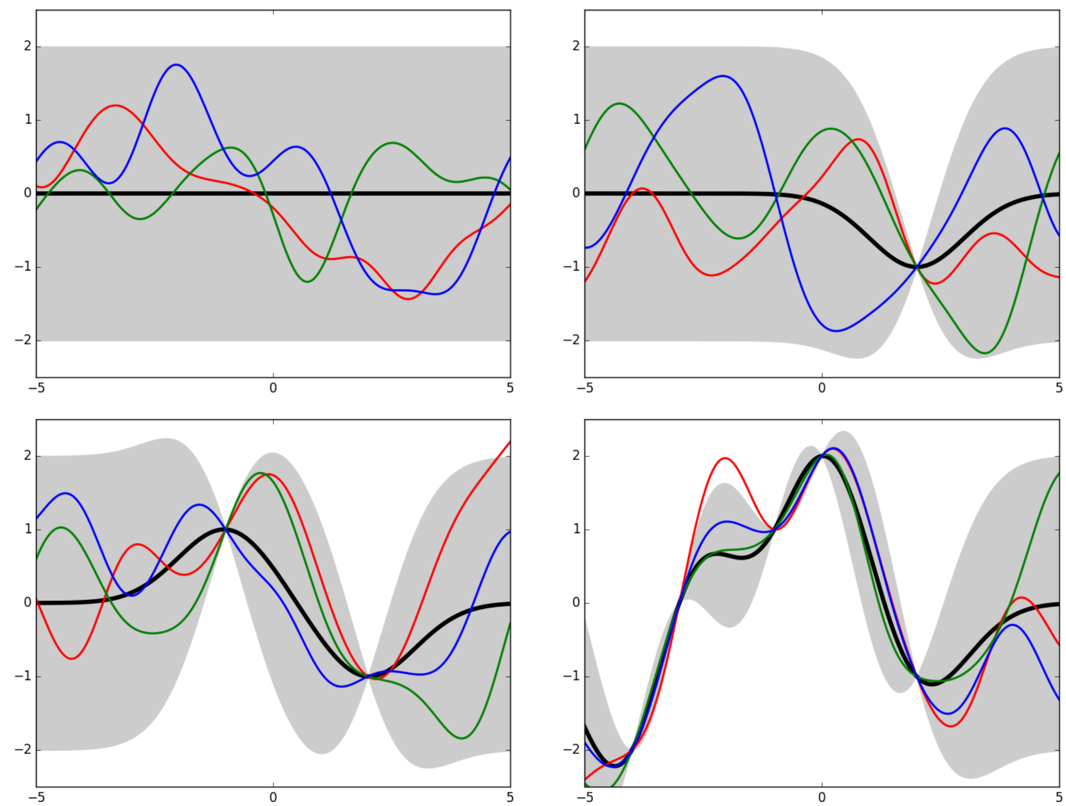

Figure 1: An example of the behavior of Gaussian process regression as additional data points are added. All figures show the mean function of the Gaussian process in black, as well as the region within two standard deviations in gray. The red, green, and blue functions are samples from the Gaussian process. (Note that the samples are not necessarily always within the gray region; there is a 5% probability that a sample from a Gaussian distribution is outside of two standard deviations.)

The first image shows the zero-mean prior with uniform spread at each point. As more points are added, the mean function shifts to accommodate the data point. Furthermore, the gray region is pinched around the point, indicating that nearby points also have reduced uncertainty. In this example, we have assumed that there is no measurement noise σ_n, so there is no uncertainty at the given data point. If this is not the case, then the gray region will not be pinched to zero.

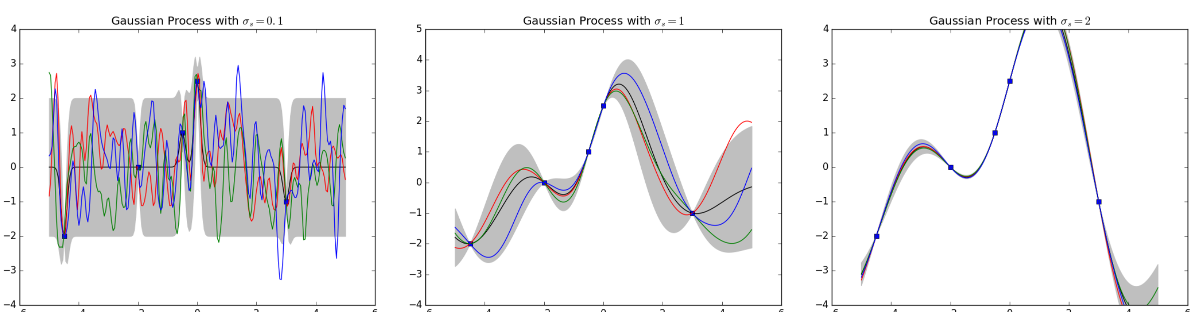

Figure 2: Gaussian processes using RBF kernel with $\sigma_s = 0.1$, 1, and 2, respectively. For the RBF kernel, increasing σ_s results in smoother samples from the Gaussian process.

1.2.2 Gaussian Process Regression

The goal of Gaussian process regression is to recover a (nonlinear) function f given data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i = f(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. We can combine $\mathbf{x}_i$ into the design matrix $\mathbf{X}$ and combine y_i into the vector $\mathbf{Y}$.

A Gaussian process is completely specified by its mean function $\mu(\mathbf{x})$ (which, for simplicity, is assumed to be the zero function) and a positive definite covariance function $k(\mathbf{x}, \mathbf{x}')$.

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{15}$$
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))] \tag{16}$$

We can write the Gaussian process as

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{17}$$

One formal definition of a Gaussian process is that it is an infinite collection of random variables, any finite subset of which is jointly Gaussian. This means that given training input $\mathbf{X}$ with additive independent identically distributed Gaussian measurement noise $\epsilon_i \in \mathcal{N}(0, \sigma_n^2)$, training output $\mathbf{Y}$, and query input $\mathbf{X}_*$, the query output $\mathbf{Y}_*$ (which is equal to $f(\mathbf{X}_*)$ in the absence of noise) is given by the following Gaussian distribution:

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \tag{18}$$

where $K(\cdot, \cdot)$ is a matrix of pairwise kernel values. (If there are n training pairs and n_* query pairs, then $K(\mathbf{X}, \mathbf{X}_*)$ is an $n \times n_*$ matrix.)

The multivariate Gaussian distribution has the property that any conditional distribution is also Gaussian. Therefore, the distribution $\mathbf{Y}_* | \mathbf{X}, \mathbf{Y}, \mathbf{X}_*$ can be fully described with a mean and covariance matrix. We can describe that mean and covariance using the standard multivariate Gaussian conditional formula:

$$\mathbf{Y}_* | \mathbf{X}, \mathbf{Y}, \mathbf{X}_* \sim \mathcal{N}\left(\mu_{\mathbf{Y}_*}, \Sigma_{\mathbf{Y}_*}\right) \tag{19}$$
$$\mu_{\mathbf{Y}_*} = K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} \mathbf{Y} \tag{20}$$
$$\Sigma_{\mathbf{Y}_*} = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} K(\mathbf{X}, \mathbf{X}_*) \tag{21}$$

2 Gaussian Process Optimization

2.1 Problem setup

We consider the problem of maximizing a real-valued function $f(\cdot)$ over a domain D, such that we want to choose a sequence of T points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ to maximize the sum $\sum_{t=1}^{T} f(\boldsymbol{x}_t)$. We can compare this to choosing the best point in hindsight, i.e. $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in D} f(\boldsymbol{x})$, and therefore define the instantaneous regret at time t as $r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$ and the cumulative regret as $R_T = \sum_{t=1}^{T} r_t$ with the goal of minimizing R_T. An additional issue is that we can only observe noisy estimates of the function, meaning that when we choose a point $\boldsymbol{x}_t$ we observe $y_t = f(\boldsymbol{x}_t) + \epsilon_t$, where $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 \boldsymbol{I})$ for some known variance quantity σ^2. We can approach this problem by modeling the underlying function $f(\cdot)$ as a Gaussian process, with a mean function $\mu(\cdot)$ capturing our estimate of the function across the domain and a variance function $\sigma^2(\cdot)$ capturing our uncertainty in this estimate.

2.2 The Gaussian Process Upper Confidence Bound (GP-UCB) Algorithm

We can use Gaussian processes to describe a natural UCB algorithm for optimizing an unknown nonlinear function. β_t will be defined later, but it is an increasing function in t and captures the number of standard deviations in the upper confidence bounds.

Algorithm 2 The GP-UCB algorithm.

Input: Function to maximize, f; Function input space D; GP Prior $\mu_0 = 0, \sigma_0, k(\boldsymbol{x}, \boldsymbol{x}') \leq 1\, \forall \boldsymbol{x}, \boldsymbol{x}'$
for $t = 1, 2, \ldots$ **do**
  Choose $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in D} \mu_{t-1}(\boldsymbol{x}) + \sqrt{\beta_t}\sigma_{t-1}(\boldsymbol{x})$
  Examine $y_t = f(\boldsymbol{x}_t) + \epsilon_t,\ \epsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 \boldsymbol{I})$
  Perform Gaussian process Bayesian update with $(\boldsymbol{x}_t, y_t)$ to obtain μ_t and σ_t (Eq. 20 and 21)
end for

Next, we describe how to choose β_t to achieve sublinear regret bounds.

2.3 Regret Bounds for GP-UCB

We will analyze the GP-UCB algorithm for solving the problem described in Section 2.1, and in particular derive bounds on the cumulative regret under this algorithm.

2.3.1 Case 1: Finite D

For the first case, assume the domain D that we are working in is of finite size $|D|$. In a practical setting, we can imagine taking a bounded set D that is in a continuous space and discretizing it to allow for the following regret bound to hold. The regret bound is defined as follows

Theorem 2. *Let $\delta \in (0,1)$ and $\beta_t = 2\log(|D|t^2\pi^2/6\delta)$. Running GP-UCB with β_t for a function f sampled from a GP with mean function zero and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, we obtain a regret bound of $\mathcal{O}^*(\sqrt{T\gamma_T \log|D|})$ with high probability (probability $1-\delta$). Precisely,*

$$Pr\left\{R_T \leq \sqrt{C_1 T \beta_T \gamma_T} \quad \forall T \geq 1\right\} \geq 1 - \delta$$

where $C_1 = 8/\log(1+\sigma^{-2})$ and T is the number of rounds of sampling an individual point.

Here, γ_T is the *maximum information gain* after T rounds, defined as

$$\gamma_T := \max_{A \subset D: |A| = T} I(\boldsymbol{y}_A; \boldsymbol{f}_A) = \max_{A \subset D: |A| = T} H(\boldsymbol{y}_A) - H(\boldsymbol{y}_A \mid \boldsymbol{f}_A)$$

For marginal entropy $H(\boldsymbol{y}_A)$ of the observations $\boldsymbol{y}_A$ and conditional entropy $H(\boldsymbol{y}_A \mid \boldsymbol{f}_A)$ of the observations $\boldsymbol{y}_A$ given the corresponding function values $\boldsymbol{f}_A$. The quantity γ_T describes the maximum reduction in uncertainty about the unknown function f from revealing the observations $\boldsymbol{y}_A$, and is a problem-dependent quantity relying on the properties of both the choice of kernel and the input space.

We prove Theorem 2 via the following sequence of lemmas. We begin by showing that we can bound the deviation between the true function and our estimated mean function by a scaled version of the estimated variance function. We then use this to bound the instantaneous regret $r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$ at each timepoint t, where $\boldsymbol{x}^*$ refers to the point(s) in the domain D which maximizes the function f. We describe an alternative way of writing the maximum information gain γ_T, and rewrite the bound on the instantaneous regret to come up with a bound on the sum of squared instantaneous regrets in terms of the maximum information gain. We then use this final result to bound the cumulative regret $R_T = \sum_{t=1}^{T} r_t$ in terms of the information gain, variables chosen for GP-UCB, and constant terms. We begin with the following lemma.

Lemma 1. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(|D|\pi_t/\delta)$, for some π_t such that $\sum_{t\geq 1} \pi_t^{-1} = 1$, $\pi_t > 0$. Then,*

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in D, \forall t \geq 1$$

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

Proof. The outline of this proof is that we start with the expression for $\Pr\{r > c\}$ for a random variable $r \sim N(0,1)$, based on the definition of the Gaussian cumulative distribution function. We then manipulate that expression to an alternate form, perform the substitutions $r = (f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x}))/\sigma_{t-1}(\boldsymbol{x})$ and $c = \beta_t^{1/2}$, and apply the union bound over all points in our domain and all timepoints to prove the lemma.

To start, fix $t \geq 1$ and $\boldsymbol{x} \in D$. Conditioned in $\boldsymbol{y}_{t-1} = (y_1, \ldots, y_{t-1})$, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}\}$ are deterministic, and $f(\boldsymbol{x}) \sim N(\mu_{t-1}(\boldsymbol{x}), \sigma_{t-1}^2(\boldsymbol{x}))$. For a variable $r \sim N(0,1)$, we can use the Gaussian cumulative distribution function to define

$$\begin{aligned}
\Pr\{r > c\} &= \frac{1}{\sqrt{2\pi}} \int_c^{\infty} e^{-r^2/2}\, dr = \frac{1}{\sqrt{2\pi}} \int_c^{\infty} e^{-r^2/2 + c^2/2}\, e^{-c^2/2}\, dr \\
&= e^{-c^2/2}\, \frac{1}{\sqrt{2\pi}} \int_c^{-\infty} e^{(-r^2/2 + rc - c^2/2) + (-rc - c^2)}\, dr \\
&= e^{-c^2/2}\, \frac{1}{\sqrt{2\pi}} \int_c^{-\infty} e^{-(r-c)^2/2}\, e^{-c(r-c)}\, dr
\end{aligned}$$

Notice that the integral term scaled by $1/\sqrt{2\pi}$ resembles the Gaussian density integrated from c to ∞ for a random variable r with mean c and unit standard deviation, which would ordinarily integrate to 1/2 (i.e., integrating over the Gaussian density for all values greater than the mean). However, the integrand is scaled by the term $e^{-c(r-c)}$. For $c > 0$ and $r \geq c$, we have that $e^{-c(r-c)} \leq 1$, so the integrand is actually scaled down. Therefore, we have that

$$e^{-c^2/2}\, \frac{1}{\sqrt{2\pi}} \int_c^{-\infty} e^{-(r-c)^2/2}\, e^{-c(r-c)}\, dr \leq \frac{1}{2} e^{-c^2/2}$$

Using this expression with $r = (f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x}))/\sigma_{t-1}(\boldsymbol{x})$ and $c = \beta_t^{1/2}$, we get that for any given $\boldsymbol{x} \in D$ and timepoint $t \geq 1$

$$\Pr\{|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\} \leq e^{-\beta_t/2}$$

Applying the union bound over all $\boldsymbol{x} \in D$, we get

$$\Pr\left\{\bigcup_{\boldsymbol{x}\in D} |f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\right\} \leq \sum_{\boldsymbol{x}\in D} \Pr\{|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\} \leq |D| e^{-\beta_t/2}$$

To further generalize this to all timepoints $t \geq 1$, we can redefine $|D|e^{-\beta_t/2} = \delta/\pi_t$. To satisfy the conditions on π_t from the lemma statement, we can use $\pi_t = \pi^2 t^2/6$ since $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$. We then apply the union bound again over all timepoints $t \in \mathbb{N}$ to get

$$\begin{aligned}\Pr\left\{\bigcup_{t=1}^{\infty} |f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}), \forall \boldsymbol{x} \in D\right\} &\leq \sum_{t=1}^{\infty} \Pr\{|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}), \forall \boldsymbol{x} \in D\} \\ &\leq \sum_{t=1}^{\infty} \frac{\delta}{\pi_t} = \delta\end{aligned}$$

Changing the direction of the inequality to upper bound the absolute difference $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})|$ for all $\boldsymbol{x} \in D$ and all $t \geq 1$ gives us that

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in D, \forall t \geq 1$$

Holds with probability $\geq 1 - \delta$, which gives us the original statement of the lemma. This completes the proof of Lemma 1. □

Lemma 2. *Fix $t \geq 1$. If $|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$, then the instantaneous regret $r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$ (where $\boldsymbol{x}^*$ is the point that maximizes the true function f) is bounded as*

$$r_t \leq 2\beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t).$$

Proof. We have already shown that the deviation of the true function value at $\boldsymbol{x}$ from our estimated mean function is bounded by a scaled version of our estimated variance function at that point. If we choose our next point according to the argmax expression given by the GP-UCB algorithm, we can be sure that the upper confidence bound at that point is greater than both the function evaluated at that point and at the optimal point $\boldsymbol{x}^*$ which maximizes the function. This allows us to bound the deviation between the function value at our chosen point and the maximum function value (i.e. the instantaneous regret) by a scaled version of our estimated variance.

Since GP-UCB specifies choosing $\boldsymbol{x}_t$ as the argmax of $\mu_{t-1}(\boldsymbol{x}_t) + \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t)$ at each timestep, we have that: $\mu_{t-1}(\boldsymbol{x}_t) + \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) \geq \mu_{t-1}(\boldsymbol{x}^*) + \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}^*) \geq f(\boldsymbol{x}^*)$. In other words, the expression within our argmax must be larger at our chosen point than at the optimal point $\boldsymbol{x}^*$, otherwise we wouldn't have chosen it. This in turn must also be larger than the actual function evaluated at $\boldsymbol{x}^*$ due to the previously-proven upper confidence bound. Using this fact, we can bound the instantaneous regret r_t as

$$r_t = f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) + \mu_{t-1}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t) \leq 2\beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t).$$

□

Lemma 3. *The information gain for the points selected can be expressed in terms of the predictive variances. For observations $\boldsymbol{y}_T = (y_1, \ldots, y_T)$ and function values $\boldsymbol{f}_T = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_T))$:*

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = \frac{1}{2}\sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)).$$

Proof. Recall that we defined the maximum information gain in Theorem 2 in terms of the information gain. The purpose of this lemma is to derive an alternate expression for the information gain $I(\boldsymbol{y}_A; \boldsymbol{f}_A)$ which we can incorporate into the regret bound from the previous lemma, so that we can derive a regret bound expression in terms of the maximum information gain across all sets of points A of size given by the number of rounds T. For a Gaussian with covariance matrix $\sigma^2\boldsymbol{I}$, the expression for the information gain can be written as

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = H(\boldsymbol{y}_T) - H(\boldsymbol{y}_T \mid \boldsymbol{f}_T) = H(\boldsymbol{y}_T) - \frac{1}{2}\log|2\pi e\sigma^2\boldsymbol{I}| \tag{22}$$

Using the definition of the entropy for a $N(0, \sigma^2 \boldsymbol{I})$ random variable. The term σ^2 is defined to be the known variance of the additive Gaussian observation noise, i.e. $y_t = f(\boldsymbol{x}_t) + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma^2)$. Since the determinant of a diagonal matrix is the product of the diagonal elements, we also have that

$$\frac{1}{2} \log|2\pi e \sigma^2 \boldsymbol{I}| = \frac{1}{2} \sum_{t=1}^{T} \log(2\pi e \sigma^2) \tag{23}$$

We can expand out $H(\boldsymbol{y}_T)$ as

$$H(\boldsymbol{y}_T) = H(\boldsymbol{y}_{T-1}) + H(y_T \mid \boldsymbol{y}_{T-1}) = H(\boldsymbol{y}_{T-1}) + \frac{1}{2} \log(2\pi e(\sigma^2 + \sigma_{T-1}^2(\boldsymbol{x}_T)))$$

Where we can write the variance term in the entropy expression as a sum of variances due to the fact that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ are deterministic conditioned on $\boldsymbol{y}_{T-1}$, and the conditional variance $\sigma_{T-1}^2(\boldsymbol{x}_T)$ does not depend on $\boldsymbol{y}_{T-1}$. Continuing to expand the entropy terms gives us

$$H(\boldsymbol{y}_T) = \frac{1}{2} \sum_{t=1}^{T} \log(2\pi e(\sigma^2 + \sigma_{t-1}^2(\boldsymbol{x}_t))) \tag{24}$$

Finally, we can substitute the expressions from (23) and (24) into (22) to get

$$\begin{aligned}
I(\boldsymbol{y}_T; \boldsymbol{f}_T) &= H(\boldsymbol{y}_T) - H(\boldsymbol{y}_T \mid \boldsymbol{f}_T) \\
&= \frac{1}{2} \sum_{t=1}^{T} \log(2\pi e(\sigma^2 + \sigma_{t-1}^2(\boldsymbol{x}_t))) - \frac{1}{2} \sum_{t=1}^{T} \log(2\pi e \sigma^2) \\
&= \frac{1}{2} \sum_{t=1}^{T} \log\left(\frac{2\pi e(\sigma^2 + \sigma_{t-1}^2(\boldsymbol{x}_t)))}{2\pi e \sigma^2}\right) \\
&= \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2} \sigma_{t-1}^2(\boldsymbol{x}_t))
\end{aligned}$$

This concludes the proof. $\square$

Lemma 4. *Pick $\delta \in (0, 1)$ and let β_t be defined as in Lemma 1 as $\beta_t = 2\log(|D| t^2 \pi^2 / 6\delta)$. Then, the following holds with probability $\geq 1 - \delta$:*

$$\sum_{t=1}^{T} r_t^2 \leq C_1 \beta_T I(\boldsymbol{y}_T; \boldsymbol{f}_T) \leq C_1 \beta_T \gamma_T \quad \forall T \geq 1,$$

where $C_1 := 8/\log(1 + \sigma^{-2}) \geq 8\sigma^2$, β_T is β_t defined at $t = T$, and γ_T is the maximum information gain over all sets of chosen points of size T (as defined in the statement of Theorem 2).

Proof. By Lemma 1 and Lemma 2 we have that $r_t^2 \leq 4\beta_t \sigma_{t-1}^2(\boldsymbol{x}_t), \forall t \geq 1$ with probability $\geq 1 - \delta$. Since β_t is defined in such a way that it is nondecreasing for increasing t and $t \leq T$, we can write

$$4\beta_t \sigma_{t-1}^2(\boldsymbol{x}_t) \leq 4\beta_T \sigma_{t-1}^2(\boldsymbol{x}_t)$$

To further manipulate this inequality, we rely on the fact that $\sigma_{t-1}^2(\boldsymbol{x}_t) = k(\boldsymbol{x}_t, \boldsymbol{x}_t)$ by definition, and our restriction in the statement of GP-UCB that $k(\boldsymbol{x}, \boldsymbol{x}') \leq 1$ for all $\boldsymbol{x}, \boldsymbol{x}'$ means that $\sigma_{t-1}^2(\boldsymbol{x}_t) \leq 1$ for all t. We can combine this with the fact that the function $s/\log(1 + s)$ is positive and monotonically increasing for positive s to show that

$$\frac{\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)}{\log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t))} \leq \frac{\sigma^{-2}}{\log(1+\sigma^{-2})}$$

$$\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t) \leq \frac{\sigma^{-2}}{\log(1+\sigma^{-2})}\log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t))$$

$$4\beta_T\sigma_{t-1}^2(\boldsymbol{x}_t) = 4\beta_T\sigma^2(\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)) \leq 4\beta_T\sigma^2 C_2 \log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t))$$

with $C_2 = \sigma^{-2}/\log(1+\sigma^{-2})$. We can combine this inequality with the expression we got in Lemma 2, squaring both sides and summing over all T timepoints to get

$$\begin{aligned}\sum_{t=1}^T r_t^2 \leq 4\beta_T \sum_{t=1}^T \sigma_{t-1}^2(\boldsymbol{x}_t) &\leq 4\beta_T\sigma^2 C_2 \sum_{t=1}^T \log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)) \\ &= 8\sigma^2 C_2\beta_T\left(\frac{1}{2}\sum_{t=1}^T \log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t))\right) \\ &= C_1\beta_T I(\boldsymbol{y}_T;\boldsymbol{f}_T) \leq C_1\beta_T\gamma_T\end{aligned}$$

Where $C_1 = 8\sigma^2 C_2 = 8/\log(1+\sigma^{-2})$. Note that the constants C_1 and C_2 are not meaningful other than to collect constant terms to simplify the notation. The final expression above is the same as stated in the lemma, and concludes the proof. □

We prove Theorem 2 by combining the results of the previous lemmas. As an overview, we began by defining Lemma 1 to show that we can bound the deviation of the actual function value from our estimated mean function by a scaled version of our estimated variance function, at all points $\boldsymbol{x}$ in our domain D and all timepoints t. We then showed in Lemma 2 that this fact can be used to bound the instantaneous regret r_t. Lemma 3 showed that we can write an expression for the information gain $I(\boldsymbol{y}_T;\boldsymbol{f}_T)$ in terms of our estimated variance function and the observation noise variance σ^2. Lemma 4 combined the facts introduced in Lemma 2 and Lemma 3 to bound the sum of squared instantaneous regrets by a quantity that involves the maximum information gain. Finally, we conclude the proof of Theorem 2 by referring back to the definition of the total regret R_T as the sum of instantaneous regrets r_t over all timepoints $t = 1,\dots,T$, using our bound on the sum of squared instantaneous regrets from Lemma 4, and making use of the Cauchy-Schwarz inequality:

$$R_T^2 = \left(\sum_{t=1}^T r_t\right)^2 = \left(\sum_{t=1}^T r_t\cdot 1\right)^2 \leq \left(\sum_{t=1}^T r_t^2\right)\left(\sum_{t=1}^T 1^2\right) = T\sum_{t=1}^T r_t^2 \implies R_T \leq \sqrt{T\sum_{t=1}^T r_t^2} \leq \sqrt{C_1 T\beta_T\gamma_T}$$

Where the inequality above again holds with probability $\geq 1-\delta$. This concludes the proof of Theorem 2.

2.3.2 Case 2: Generalization

We now generalize the previous theorem to any compact and convex $D \subset \mathbb{R}^d$. However as we shall see, this generalization comes with a few assumptions on the kernel function. Let us now first state the theorem:

Theorem 3. *Let $D \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}, r > 0$. Suppose that the kernel $k(x,x')$ satisfies the following high probability bound on the derivatives of GP sample paths f: for some $a, b > 0$*

$$Pr\{\sup_{\boldsymbol{x}\in D}|\partial f/\partial x_j| > L\} \leq ae^{-(L/b)^2},\ j = 1,\dots,d.$$

Pick $\delta \in (0,1)$ and define

$$\beta_t = 2\log\left(t^2 2\pi^2/(3\delta)\right) + 2d\log\left(t^2 dbr\sqrt[2]{\log(4da/\delta)}\right)$$

Running the GP-UCB algorithm with β_t for a sample f of a GP with mean function zero and covariance function $k(\boldsymbol{x}, \boldsymbol{x'})$, we obtain a regret bound of $\mathcal{O}^*(\sqrt[2]{dT\gamma_T})$ with high probability. Precisely with $C_1 = 8/\log\left(1+\sigma^{-2}\right)$, we have

$$Pr\{R_T \leq \sqrt[2]{C_1 T \beta_T \gamma_T} + 2 \quad \forall T \geq 1\} \geq 1-\delta$$

where, T is the number of rounds of sampling an individual point and γ_T is the maximum information gain at the end of T rounds

$$\gamma_T := \underset{A \subset D : |A| = T}{max} I(\boldsymbol{y}_A; \boldsymbol{f}_A)$$

Note that the assumption (weaker than Lipschitz condition) on the kernel function imposes smoothness conditions, specifically, that the slope of the function at any given point exceeds a constant L with very low probability. This naturally disqualifies GPs with sample paths f that are not differentiable or vary erratically. It is to be noted that this assumption is not very strict and many kernels, such as the Radial Basis Function (RBF) kernel, used in practice satisfy this condition.

Before we continue to prove the theorem, note that the sequence of lemmas constructed for Theorem 2 breaks at Lemma 1 which sets the value of β_t as a function of size of D. Since we are considering an infinite D, we break away from the previous lemmas albeit maintaining a similar proof strategy.

To give an overview, this is the sequence of steps we shall follow to obtain the regret bound for continuous decision set D: we will show in Lemma 5 that we can extend Lemma 1 to the continuous case i.e. we can bound the deviation between the true function and our estimated mean function by a scaled version of the estimated variance function. We shall note that the scaling is appropriately chosen to not involve the size of the decision set as was done in the discrete case. Once this is formalized, we need to obtain a confidence interval on x^*. For this we will take advantage of discretization where we use a particular discretization D_t at given time step t. In each of these subsets, using the same procedure as in the previous discrete case, we obtain confidence bounds for a given subset, a variant of Lemma 1. With this set up, for the next steps where we obtain a bound on the instantaneous regret, we will pick a particular discretization for Lemmas 7 and 8. We will notice that we obtain similar expressions as in the discrete case by this point and using the procedure as in Lemma 4, we prove the theorem and conclude.

As stated, our first step is to show that all decisions made are of high confidence.

Lemma 5. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(2\pi t/\delta)$, where $\sum_{t\geq 1} \pi_t^{-1} = 1, \pi_t > 0$. Then,*

$$|f(\boldsymbol{x}_t)\mu_{t-1}\boldsymbol{x}_t| \leq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t), \quad \forall t \geq 1$$

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

Proof. The proof is exactly as in Lemma 1 with only the difference in the definition of β_t since we no longer can define it as a function of size of D. We fix $t \geq 1$. Given the sequence of function values $\{y_1, \ldots, y_{t-1}\}$, we have that from the algorithm, $\{x_1, \ldots, x_{t-1}\}$ is deterministic and $f(\boldsymbol{x}) \sim N(\mu_{t-1}(\boldsymbol{x}), \sigma_{t-1}^2(\boldsymbol{x}))$. As in Lemma 1,

$$\Pr\{|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| > \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x})\} \leq e^{-\beta_t/2}$$

But from the definition of β_t, we have $e^{-\beta_t/2} = \delta/2\pi_t$. Using the union bound over t and the fact that $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$, we set $\pi_t = \pi^2 t^2/12$, to prove the lemma. □

Note that in the current setting, we could also have discretized the space D into finite sets D_t, where D_t is used at time t. This gives us the same confidence bound as Lemma 1. For the sake of completeness in exposition, we will formally state (proof can be borrowed from Lemma 1):

Lemma 6. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(|D_t|\pi_t/\delta)$, where $\sum_{t\geq 1} \pi_t^{-1} = 1$, $\pi_t > 0$. Then,*

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \geq \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in D_t, \forall t \geq 1$$

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

Consider a discretization D_t of size $(\tau_t)^d$ such that:

$$||\boldsymbol{x} - [\boldsymbol{x}]_t|| \leq rd/\tau_t, \quad \forall \boldsymbol{x} \in D_t \tag{25}$$

where $[\boldsymbol{x}]_t$ denotes the closest point in D_t to $\boldsymbol{x}$. This essentially implies that we are discretizing the space finely enough such that for any point $\boldsymbol{x}$ considered in D, there exists at least one point $[\boldsymbol{x}]_t$ in the discretized space D_t contained within the ball of radius rd/τ_t centered at $\boldsymbol{x}$. With this framework in place, we are now ready to bound the confidence on $\boldsymbol{x}^*$.

Lemma 7. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(2\pi_t/\delta) + 4d\log\left(dtbr\sqrt[2]{\log(2da/\delta)}\right)$, where $\sum_{t\geq 1}\pi_t^{-1} = 1$, $\pi_t > 0$. Let $\tau_t = dt^2br\sqrt[2]{\log(2da/\delta)}$. Let $[\boldsymbol{x}^*]_t$ denote the closest point in D_t to $\boldsymbol{x}^*$. Then,*

$$|f(\boldsymbol{x}^*) - \mu_{t-1}([\boldsymbol{x}^*]_t)| \leq \beta_t^{1/2}\sigma_{t-1}([\boldsymbol{x}^*]_t) + \frac{1}{t^2} \quad \forall t \geq 1$$

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

Proof. We now invoke the assumption we made for the kernel function to satisfy with high probability on the GP sample paths. With the assumption and union bound, we have:

$$Pr\{\forall j, \forall \boldsymbol{x} \in D, |\partial f/\partial x_j| < L\} \geq 1 - dae^{-L^2/b^2}.$$

In other words, we have with a probability greater than $1 - dae^{-L^2/b^2}$ that for all,

$$\forall \boldsymbol{x} \in D \ |f(x) - f(x')| \leq L\,||x - x'||_1 \tag{26}$$

Now, let $\delta/2 = dae^{-L^2/b^2}$. This implies $L = b\sqrt[2]{\log 2ad/\delta}$. Substituting in Eq. 26, we have

$$\forall \boldsymbol{x} \in D \ |f(x) - f(x')| \leq b\sqrt[2]{\log 2ad/\delta}\,||x - x'||_1$$

Using Eq. 25, we obtain for each time-step and D_t,

$$\forall \boldsymbol{x} \in D_t \ |f(x) - f(x')| \leq rdb\sqrt[2]{\log 2ad/\delta}/\tau_t$$

Now, choose the discretization such that $\tau_t = dt^2br\sqrt[2]{\log 2da/\delta}$ and we therefore get,

$$\forall \boldsymbol{x} \in D_t \ |f(x) - f(x')| \leq 1/t^2$$

This implies that $|D_t| = (dt^2br\sqrt[2]{\log(2da/\delta)})^d$ from the definition of τ_t considered. Using $\delta/2$ in Lemma 6, we can apply the confidence bound to $[\boldsymbol{x}^*]$ to obtain the result. □

Lemma 8. *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(4\pi_t/\delta) + 4d\log\left(dtbr\sqrt[2]{\log(4da/\delta)}\right)$, where $\sum_{t\geq 1}\pi_t^{-1} = 1$, $\pi_t > 0$. Then the regret is bounded as,*

$$r_t \leq 2\beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) + \frac{1}{t^2} \quad \forall t \geq 1$$

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

Proof. Since by the definition, $\boldsymbol{x}_t = argmax_{\boldsymbol{x}\in D}\mu_{t-1} + \beta_t^{1/2}\sigma_{t-1}$, we have that

$$\mu_{t-1}(\boldsymbol{x}_t) + \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) \geq \mu_{t-1}([\boldsymbol{x}^*]_t) + \beta_t^{1/2}\sigma_{t-1}([\boldsymbol{x}^*]_t)$$

Also notice that from Lemma 7, we have:

$$\mu_{t-1}([\boldsymbol{x}^*]_t) + \beta_t^{1/2}\sigma_{t-1}([\boldsymbol{x}^*]_t) + 1/t^2 \geq f(\boldsymbol{x}^*)$$

From the above two equations, we can compute the regret at time t as:

$$\begin{aligned} r_t &= f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \\ &\leq \mu_{t-1}([\boldsymbol{x}^*]_t) + \beta_t^{1/2}\sigma_{t-1}([\boldsymbol{x}^*]_t) + 1/t^2 - f(\boldsymbol{x}) \\ &\leq \mu_{t-1}(\boldsymbol{x}_t) + \beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) + 1/t^2 - f(\boldsymbol{x}) \end{aligned}$$

Finally from Lemma 5, we get,

$$r_t \leq 2\beta_t^{1/2}\sigma_{t-1}(\boldsymbol{x}_t) + 1/t^2. \tag{27}$$

which completes the proof. □

Note that we have the same framework as before when D was finite except that we have an addition $1/t^2$ term which sums to a constant ($\pi^2/6$) when a summation over the time-step is considered. Following the procedure described in Lemma 4, we get with a probability greater than $1 - \delta$,

$$R_T = \sum_{t=1}^{T} r_t \leq \sqrt[2]{C_1 T \beta_T \gamma_T} + \pi^2/6 \quad \forall T \geq 1$$

which implies Theorem 3.

References

Andreas Krause and Daniel Golovin. Submodular function maximization., 2012.

Jeremy Kun. When greedy algorithms are good enough: Submodularity and the (1 1/e)-approximation. `https://jeremykun.com/2014/07/07/when-greedy-algorithms-are-good-enough-submodularity\-and-the-1-1e-approximation/`, 2014.

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

Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. *The MIT Press, Cambridge, MA, USA*, 38:715–719, 2006.