Multiscale Gaussian Process Level Set Estimation

Shubhanshu Shekhar
University of California, San Diego
shshekha@eng.ucsd.edu

Tara Javidi
University of California, San Diego
tjavidi@eng.ucsd.edu

Abstract

In this paper, the problem of estimating the level set of a black-box function from noisy and expensive evaluation queries is considered. A new algorithm for this problem in the Bayesian framework with a Gaussian Process (GP) prior is proposed. The proposed algorithm employs a hierarchical sequence of partitions to explore different regions of the search space at varying levels of detail depending upon their proximity to the level set boundary. It is shown that this approach results in the algorithm having a low complexity implementation whose computational cost is significantly smaller than the existing algorithms for higher dimensional search space $\mathcal{X}$. Furthermore, high probability bounds on a measure of discrepancy between the estimated level set and the true level set for the the proposed algorithm are obtained, which are shown to be strictly better than the existing guarantees for a large class of GPs. In the process, a tighter characterization of the information gain of the proposed algorithm is obtained which takes into account the structured nature of the evaluation points. This approach improves upon the existing technique of bounding the information gain with maximum information gain.

1 Introduction

Suppose $f : \mathcal{X} \to \mathbb{R}$ is an unknown black-box function which can only be accessed through its noisy observations

$$y = f(x) + \eta. \tag{1}$$

For some $\tau > 0$, we define the τ (super-)level set of f as $S_\tau = \{x \in \mathcal{X} : f(x) \geq \tau\}$. Given a budget of n function evaluations, our goal is to design an adaptive query point selection strategy in order to efficiently construct an estimate $\hat{S}_\tau$ of the τ level set of f. The accuracy of an estimate $\hat{S}_\tau$ is measured by the term

$$\mathcal{L}(\hat{S}_\tau, S_\tau) = \sup_{x \in \hat{S}_\tau \triangle S_\tau} |f(x) - \tau|,$$

where $\hat{S}_\tau \Delta S_\tau = \left(\hat{S}_\tau \setminus S_\tau\right) \bigcup \left(S_\tau \setminus \hat{S}_\tau\right)$ denotes the symmetric difference of the true and estiamted level sets. This problem of estimating level sets of unknown functions from noisy evaluations arises naturally in a wide range of applications. These applications include monitoring environmental parameters such as humidity and solar radiation (Gotovos et al., 2013), analyzing geospatial data and medical imaging (Willett and Nowak, 2007).

In this paper, we propose a new algorithm for level set estimation which utilizes ideas from existing algorithms in the areas of global optimization (Bubeck et al., 2011; Munos, 2011) and Bayesian Optimization (Wang et al., 2014; Shekhar and Javidi, 2017). Compared to the state of the art, we show that our proposed algorithm has better computational complexity as well as tighter convergence guarantees.

1.1 Related Work

Bryan et al. (2006) first considered the level set estimation with a GP prior and studied several heuristics for selecting the evaluation points based on variance, classification probability, information gain and straddle heuristic. They empirically compared the performance of these methods and concluded that the straddle heuristic outperformed other methods of selecting evaluation points.

Gotovos et al. (2013) built upon the work of Bryan et al. (2006) and proposed the LSE algorithm which uses a search strategy inspired by the GP-UCB algorithm of Srinivas et al. (2012) and derived theoretical bounds on the convergence rate of the estimation error. Bogunovic et al. (2016) further highlighted the

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

connection between Bayesian Optimization and level set estimation by studying these problems in a unified framework. Their proposed algorithm, TruVAR, can also deal with non-uniform observation costs and heterostedastic noise. For the case of fixed noise and cost model, their bounds match those of Gotovos et al. (2013).

1.2 Contributions

For the case of $\mathcal{X} = [0,1]^D$, all the algorithms mentioned above have two drawbacks: first, the computational cost of implementing them exactly increases exponentially with D, and second, their theoretical convergence guarantees depend on the maximum mutual information gain γ_n. Some recent results in Bayesian Optimization literature (Scarlett et al., 2017; Scarlett, 2018) suggest that bounds based on γ_n can be quite loose, especially for the Mátern family of kerenls. The main contributions of this paper address these issues:

- We propose a new algorithm for level set estimation which explores the search space by employing a hierarchical sequence of partitions of $\mathcal{X}$, and show that the computational complexity of the algorithm with a given evaluation budget n only has linear dependence on the dimension D.
- We also derive theoretical guarantees on the estimation error of the proposed algorithm which improve upon the theoretical guarantees for existing algorithms.
- Finally, by exploiting the structured nature of the points evaluated by our algorithm, we obtain a more refined characterization of the information gain of our algorithm. In particular, we obtain a tighter bound on the information gain for all members of the widely used Mátern family of kernels.

2 Preliminaries

A Gaussian Process (GP) is a collection of random variables whose finite subcollections are jointly Gaussian, that is, all linear combinations of any finite subcollection are univariate Gaussian random variables. Gaussian Processes with index set $\mathcal{X}$ are completely specified by their mean function $\mu : \mathcal{X} \to \mathbb{R}$ and covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We also note that a zero mean Gaussian Process with a non-degenerate covariance function k induces the *canonical metric* d_k on the index set defined as $d_k(x_1, x_2) = \big(k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)\big)^{1/2}$.

As mentioned earlier, in this paper we work under the Bayesian framework in which we assume that the black-box function f is a sample from a zero mean Gaussian Process, $GP(0, k)$, with known covariance function k. Furthermore, we also assume that the observation noise η is distributed as $N(0, \sigma^2)$ and the variance σ^2 is known to the algorithm. Given observations $\mathcal{D}_t = \{(x_i, y_i) \mid 1 \le i \le t\}$, the posterior distribution at any $x \in \mathcal{X}$ is again a univariate Gaussian with parameters

$$\mu_t(x) = k(x, x_{\mathcal{D}_t}) J_t^{-1} y_{\mathcal{D}_t}$$
$$\sigma_t^2(x) = k(x, x) + k(x, x_{\mathcal{D}_t}) J_t^{-1} k(x_{\mathcal{D}_t}, x).$$

In the above display, $x_{\mathcal{D}_t}$ and $y_{\mathcal{D}_t}$ denote the vectors of evaluation points and their corresponding observations. The terms $k(x, x_{\mathcal{D}_t})$ and $k(x_{\mathcal{D}_t}, x_{\mathcal{D}_t})$ denote the vector and matrix of pairwise covariance values respectively. Finally, the term J_t is equal to $\big(k(x_{\mathcal{D}_t}, x_{\mathcal{D}_t}) + \sigma^2 E_t\big)$ and E_t is the $t \times t$ identity matrix.

Next, we introduce some definitions regarding the properties of the index space $\mathcal{X}$.

Definition 1. Given a set $\mathcal{X}$ with an associated metric d, we define the *metric dimension* of $\mathcal{X}$ with respect to d, denoted by D_m, as follows:

$$D_m \coloneqq \inf\{a > 0 \mid \ \exists C < \infty : N(\mathcal{X}, r, d) \le C r^{-a} \ \forall r \ge 0\}$$

where $N(\mathcal{X}, r, d)$ is the $r-$covering number of $\mathcal{X}$ with respect to the metric d defined as:

$$N(\mathcal{X}, r, d) \coloneqq \min\{|\mathcal{Z}| \mid \mathcal{Z} \subset \mathcal{X},\ \mathcal{X} \subset \cup_{z \in \mathcal{Z}} B(z, r, d)\}.$$

A related notion is the $r-$packing number of a set $\mathcal{X}$ with respect to a metric d, denoted by $M(\mathcal{X}, r, d)$, which is defined as:

$$M(\mathcal{X}, r, d) \coloneqq \max\{|\mathcal{Z}| \mid \mathcal{Z} \subset \mathcal{X}, \quad d(z_1, z_2) \ge r \ \forall z_1, z_2 \in \mathcal{Z}\}.$$

Finally, we introduce a local notion of dimensionality of the metric space.

Definition 2. Suppose $\mathcal{P}(\mathcal{X})$ denotes the power set of $\mathcal{X}$, and $\zeta : (0, \infty) \mapsto \mathcal{P}(\mathcal{X})$ represents a mapping from the positive real numbers to subsets of $\mathcal{X}$. Then we define the dimension of $(\mathcal{X}, d)$ associated with the mapping $\zeta(\cdot)$ as

$$D_\zeta \coloneqq \inf\{a > 0 \ \mid \exists C < \infty : \ M(\zeta(r), r, d) \le C r^{-a} \ \forall r > 0\}.$$

The above definition of dimension is a simple generalization of some existing definitions such as the *near-optimality dimension* of (Bubeck et al., 2011; Munos, 2011; Shekhar and Javidi, 2017) and *zooming dimension* of (Kleinberg et al., 2013). For instance, the

c-near-optimality dimension of (Bubeck et al., 2011) is obtained by selecting $\zeta(r) = \{x \in \mathcal{X} \mid f(x) \geq f(x^*) - cr\}$ for some $c > 0$, where $f(x^*)$ denotes the maximum value of f.

3 Main Results

We begin by stating the assumptions on the metric space $(\mathcal{X}, d)$ and the covariance function in Section 3.1, followed by a high level description of our proposed algorithm in Section 3.2 and then present the details and the theoretical analysis in Section 3.3.

3.1 Assumptions

We assume that the set $\mathcal{X}$ is a compact metric space with associated metric d, and that $\mathcal{X}$ has a finite metric dimension, D_m, with respect to d. We also assume that the metric space $(\mathcal{X}, d)$ admits a *tree of partitions* (Bubeck et al., 2011) which is a sequence of finite subsets $(\mathcal{X}_h)_{h\geq 0}$ of $\mathcal{X}$ such that

X1 $|\mathcal{X}_h| = 2^h$ and the elements of $\mathcal{X}_h$ are denoted by $x_{h,i}$ for $1 \leq i \leq 2^h$.

X2 To each $x_{h,i}$ is associated a cell $\mathcal{X}_{h,i}$ such that $\cup_i \mathcal{X}_{h,i} = \mathcal{X}$ for all h and $\mathcal{X}_{h+1,2i-1} \cup \mathcal{X}_{h+1,2i} = \mathcal{X}_{h,i}$ for all (h, i) pairs.

X3 There exist constants $0 < v_2 \leq 1 \leq v_1$ and $0 < \rho < 1$ such that for all (h, i) pairs

$$B(x_{h,i}, v_2\rho^h, d) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, v_1\rho^h, d)$$

Remark 1. As a concrete example, consider $\mathcal{X} = [0, 1]^D$ for some $D > 0$ and let d be the Euclidean metric. In this case, the metric dimension D_m is equal to D, the dimension of the space. Now, let $\mathcal{X}_0 = (0.5, 0.5, \ldots, 0.5)$ and the associated cell $\mathcal{X}_{0,1} = \mathcal{X}$. For any $h \geq 1$ the cells are constructed by dividing cells from the level $h-1$ equally along the longest side (breaking ties arbitrarily), and the set $\mathcal{X}_h$ is defined as the center points of the cells so obtained. This tree of partitions satisfies the assumptions $X1 - X3$ with parameters $\rho = 2^{-1/D}$, $v_1 = 2\sqrt{D}$ and $v_2 = 1/2$.

Next, we state our assumptions on the covariance function k and the metric d_k it induces on $\mathcal{X}$:

C1 There exists a non-decreasing continuous function $g : \mathbb{R}^+ \to \mathbb{R}^+$, with $g(0) = 0$, such that $d_k(x_1, x_2) \leq g(d(x_1, x_2))$ for all $x_1, x_2 \in \mathcal{X}$.

C2 There exists a $\delta_k > 0$ such that for all $r \leq \delta_k$, we have for constants $C_k > 0$ and $0 < \alpha \leq 1$ satisfying $g(r) \leq C_k r^\alpha$.

Remark 2. These two assumptions are satisfied by all commonly used covariance functions such as Squared Exponential (SE), Mátern family and Rational Quadratic kernels. For example, for the case of SE kernel with scale and length parameters a_s and a_l respectively, we have $d_k(x_1, x_2) = \sqrt{2a_s\big(1 - \exp(-d(x_1, x_2)^2/a_l)\big)} \leq \sqrt{2a_s/a_l}d(x_1, x_2)$ which implies the assumptions $C1 - C2$ are satisfied for $\delta_k = diam(\mathcal{X})$ and $\alpha = 1$.

Remark 3. It is easy to check that the class of covariance functions satisfying $C1$ and $C2$, denoted by $\mathcal{K}$, is closed under finite linear combinations. Hence it includes various GP models useful in practical applications which are constructed by combining commonly used covariance functions (Duvenaud, 2014).

3.2 General Outline

We now present a high level outline of the proposed algorithm for level set estimation.

- At any time t, we maintain an *active* set of points $\mathcal{X}_t$, and their associated cells.
- For every point $x \in \mathcal{X}_t$ we compute bounds on the maximum and minimum function value in the associated cell.
- In each iteration, we choose a candidate point x_t from $\mathcal{X}_t$ which has the highest deviation from the threshold τ.
- We take one of two actions:
  - If the selected point has been explored enough, we *refine* the cell.
  - Otherwise, we evaluate the function at the point x_t.

In our proposed algorithm, the upper and lower bounds on the function value in each cell consist of two terms: an uncertainty term due to the observation noise and another term which estimates the variation of the function in the cell. When the uncertainty due to observation noise is smaller than the variation, it implies that the cell has been sufficiently explored at the current scale, and we proceed to refine it into smaller cells. On the other hand, if the uncertainty due to noise is larger than variation, it means that the cell requires more function evaluations at the current scale.

3.3 Algorithm for GP level set estimation

The steps of our proposed algorithm for level set estimation with GP prior assumptions are shown in Algorithm 1. Besides the budget n, the threshold τ and

the tree of partitions $(\mathcal{X}_h)_{h\geq 0}$, the algorithm also requires as input several other parameters β_n, $(V_h)_{h\geq 0}$ and h_{max}. The term β_n is the scaling factor used in computing the posterior confidence intervals, and V_h is a high probability upper bound on the variation of the unknwon function in any cell $\mathcal{X}_{h,i}$. The term h_{max} denotes the largest depth that the algorithm should explore in the tree of partitions $(\mathcal{X}_h)_{h\geq 0}$.

At any time t, the algorithm maintains two sets, $\hat{S}_t$ and $\hat{R}_t$, which contain points that do not require further consideration. More specifically, set $\hat{S}_t$ contains points whose lower bounds are greater than or equal to τ and thus with high probability we have $\hat{S}_t \subset S_\tau$. Similarly, we also have $\hat{R}_t \subset S_\tau^c$ for all values of t.

Algorithm 1: Level set estimation with GP prior

Input: n, τ, $(\mathcal{X}_h)_{h\geq 0}$, $(V_h)_{h\geq 0}$, β_n, h_{max}
Initialize $t = 1$, $n_e = 0$, $\hat{S}_t = \phi$, $\hat{R}_t = \phi$
while $n_e \leq n$ **do**
  for $x_{h,i} \in \mathcal{X}_t$ **do**
    if $\bar{l}_t(x_{h,i}) \geq \tau$ **then**
      $\hat{S}_t \leftarrow \hat{S}_t \cup \mathcal{X}_{h,i}$;
      $\mathcal{X}_t \leftarrow \mathcal{X}_t \setminus \{x_{h,i}\}$;
    else if $\bar{u}_t(x_{h,i}) < \tau$ **then**
      $\hat{R}_t \leftarrow \hat{R}_t \cup \mathcal{X}_{h,i}$;
      $\mathcal{X}_t \leftarrow \mathcal{X}_t \setminus \{x_{h,i}\}$;
  end
  $x_{h_t,i_t} \in$
   $\arg\max_{x_{h,i}\in\mathcal{X}_t} \max\left(\bar{u}_t(x_{h,i}) - \tau,\ \tau - \bar{l}_t(x_{h,i})\right)$;
  if $\beta_n\sigma_{t-1}(x_{h_t,i_t}) < V_{h_t}$ *AND* $h_t \leq h_{max}$ **then**
    $\mathcal{X}_t \leftarrow \mathcal{X}_t \setminus \{x_{h_t,i_t}\}$;
    $\mathcal{X}_t \leftarrow \mathcal{X}_t \cup \{x_{h_t+1,j} \mid p(x_{h_t+1,j}) = x_{h_t,i_t}\}$;
  else
    evaluate $y_t = f(x_{h_t,i_t}) + \eta_t$;
    update $\mu_t(\cdot)$, $\sigma_t(\cdot)$;
    $n_e \leftarrow n_e + 1$;
  end
  $t \leftarrow t + 1$;
end
Output: $\hat{S}_t$

We now complete the description of the algorithm by specifying the choice of the terms β_n, $(V_h)_{h\geq 0}$, h_{max}, $\bar{l}_t$ and $\bar{u}_t$. The detailed reasoning for these choices are provided in Appendix A.1.

- For any $\delta > 0$, we select $\beta_n = \sqrt{2\log n\left(2n^{1+2/(2\alpha\log(1/\bar{\rho}))}\right) + 2\log(1/\delta)}$, where $\bar{\rho} = \min\{\rho, 1/2\}$, and α is the parameter introduced in Assumption **C2**. This choice ensures that $\forall t \geq 1$ and $\forall x \in \mathcal{X}_t$, we have $|f(x) - \mu_{t-1}(x)| \leq \beta_n\sigma_{t-1}(x)$ with probability $\geq 1 - \delta$.
- For any $\delta > 0$, we select

$$V_h = g(v_1\rho^h)\Bigg(\big(C_2 + 2\log(1/\delta_h) + (4D'_m)\log(1/v_1\rho^h)\big)^{1/2} + C_3\Bigg)$$

  where C_2 and C_3 are constants whose exact expressions are given in Appendix A.1, $\delta_h = \delta/(2^h h_{\max})$ where $h_{\max}$ is introduced below, and $D'_m = D_m/\alpha$ where D_m is the metric dimension of $(\mathcal{X}, d)$ and α is the parameter introduced in Assumption **C2**. With this choice of V_h, we have with probability at least $1 - \delta$,

$$\sup_{x\in\mathcal{X}_{h,i}} |f(x) - f(x_{h,i}| \leq V_h \quad \forall h, i$$

  The expression for V_h is obtained by using classical chaining arguments(van Handel, 2014, § 5.3) along with the assumptions on the covariance function.
- We choose the value of h_{max} to be $\log(n)/(2\alpha\log(1/\bar{\rho}))$ where $\bar{\rho} = \min\{\rho, 1/2\}$. This choice of h_{max} along with the finite metric dimension assumption ensures that the size of $\mathcal{X}_t$ for all t is at most polynomial in n which allows us to construct tight confidence bounds on the function values at all points in $\mathcal{X}_t$.

It now remains to define the terms $\bar{u}_t$ and $\bar{l}_t$. To compute the lower bound $\bar{l}_t(x_{h,i})$ on the function value in a cell $\mathcal{X}_{h,i}$, we first obtain a lower bound on the function value at $x_{h,i}$ and then subtract V_h from it. The lower bound on $f(x_{h,i})$ is obtained by computing two lower bounds and taking the maximum. The term $\bar{u}_t$ is computed in a similar manner as well. The details of the computations are as follows:

$$\bar{l}_t(x_{h,i}) = \max\{\bar{l}_{t-1}(x_{h,i}), l_t(x_{h,i})\},$$

where

$$l_t(x_{h,i}) = \max\{\mu_t(x_{h,i}) - \beta_n\sigma_t(x_{h,i}), \mu_t(p(x_{h,i})) - \beta_n\sigma_t(p(x_{h,i})) - V_{h-1}\} - V_h,$$

and

$$\bar{u}_t(x_{h,i}) = \min\{\bar{u}_{t-1}(x_{h,i}), u_t(x_{h,i})\},$$

where

$$u_t(x_{h,i}) = \min\{\mu_t(x_{h,i}) + \beta_n\sigma_t(x_{h,i}), \mu_t(p(x_{h,i})) + \beta_n\sigma_t(p(x_{h,i})) + V_{h-1}\} + V_h.$$

In the above display, for any $h \geq 1$ and $1 \leq i \leq 2^h$, we use $p(x_{h,i})$ to denote the parent node of $x_{h,i}$ in the tree of partitions, i.e., $p(x_{h,i}) = \arg\min_{x\in\mathcal{X}_{h-1}} d(x, x_{h,i})$.

We now proceed to the theoretical analysis of Algorithm 1 and begin by presenting a lemma which characterizes the properties of the points which are evaluated the algorithm.

Lemma 1. *For the choice of parameters described above, we have for any $\delta > 0$, with probability at least $1 - 2\delta$:*

- *If at time t a point x_{h_t,i_t} is evaluated by the algorithm, then the maximum deviation from τ of the function value in the cell $\mathcal{X}_{h_t,i_t}$ can be upper bounded as follows:*

$$\sup_{x \in \mathcal{X}_{h_t,i_t}} |f(x) - \tau| \leq 10 V_{h_t} \tag{2}$$

- *If the evaluated point x_{h_t,i_t} also satisfies the condition that $h_t < h_{\max}$, then we can bound the maximum devitation from τ in another way using the posterior standard deviation at x_{h_t,i_t}:*

$$\sup_{x \in \mathcal{X}_{h_t,i_t}} |f(x) - \tau| \leq 4\beta_n \sigma_t(x_{h_t,i_t}) \tag{3}$$

- *A point $x_{h,i}$, with $h < h_{\max}$, may be evaluated no more than q_h times before it is expanded, where*

$$q_h = \frac{\sigma^2 \beta_n^2}{V_h^2}.$$

and for h large enough so that $v_1 \rho^h \leq \delta_k$, we have

$$q_h = \mathcal{O}\left(\frac{\sigma^2 \beta_n^2}{(v_1 \rho^h)^{2\alpha}} \right)$$

Proof. We prove the three statements separately.

- We observe that if a point is evaluated by the algorithm, then we must have $\bar{l}_t(x_{h_t,i_t}) \leq \tau \leq \bar{u}_t(x_{h_t,i_t})$. This implies that $\max\{\bar{u}_t(x_{h_t,i_t}) - \tau, \tau - \bar{l}_t(x_{h_t,i_t})\} \leq \bar{u}_t(x_{h_t,i_t}) - \bar{l}_t(x_{h_t,i_t})$. Now, using the fact that $\bar{u}_t(x_{h_t,i_t}) \leq \mu_t(p(x_{h_t,i_t})) + \beta_n \sigma_t(p(x_{h_t,i_t})) + V_{h_t - 1} + V_{h_t}$, and $\bar{l}_t(x_{h_t,i_t}) \geq \mu_t(p(x_{h_t,i_t})) - \beta_n \sigma_t(p(x_{h_t,i_t})) - V_{h_t - 1} - V_{h_t}$ we get for any $x \in \mathcal{X}_{h_t,i_t}$.

$$\begin{aligned} |f(x) - \tau| &\leq \bar{u}_t(x_{h_t,i_t}) - \bar{l}_t(x_{h_t,i_t}) \\ &\leq 2\beta_n \sigma_t(p(x_{h_t,i_t})) + 2V_{h_t - 1} + 2V_{h_t} \\ &\overset{(a)}{\leq} 4V_{h_t - 1} + 2V_{h_t} \overset{(b)}{\leq} 10 V_{h_t} \end{aligned}$$

  where (a) follows from the fact that $\beta_n \sigma_t(p(x_{h_t,i_t}))$ must be smaller than $V_{h_t - 1}$ for the cell associated with $p(x_{h_t,i_t})$ to be refined and (b) follows from the fact that $V_{h_t - 1} \leq 2V_{h_t}$ (see Remark 5 in Appendix A.1).

- Assume that a point x_{h_t,i_t} is evaluated by the algorithm at time t. Then for any $x \in \mathcal{X}_{h_t,i_t}$ we have

$$\begin{aligned} |f(x) - \tau| &\leq \max\{\bar{u}_t(x_{h_t,i_t}) - \tau, \tau - \bar{l}_t(x_{h_t,i_t})\} \\ &\leq \bar{u}_t(x_{h_t,i_t}) - \bar{l}_t(x_{h_t,i_t}) \\ &\overset{(a)}{\leq} 2\beta_n \sigma_t(x_{h_t,i_t}) + 2V_{h_t} \\ &\overset{(b)}{\leq} 4\beta_n \sigma_t(x_{h_t,i_t}) \end{aligned}$$

  where (a) follows from the fact that by definition $\bar{u}_t(x_{h_t,i_t}) \leq \mu_t(x_{h_t,i_t}) + \beta_n \sigma_t(x_{h_t,i_t}) + V_{h_t}$ and $\bar{l}_t(x_{h_t,i_t}) \geq \mu_t(x_{h_t,i_t}) - \beta_n \sigma_t(x_{h_t,i_t}) - V_{h_t}$ and (b) follows from the condition for function evaluation at the point x_{h_t,i_t}.

- We observe that from the first part of Proposition 3 of (Shekhar and Javidi, 2017), if a point $x_{h,i}$ has been evaluated $n_{h,i}$ times by the algorithm, then we must have $\sigma_t(x_{h,i}) \leq \sigma/(\sqrt{n_{h,i}})$. Using this fact, we can obtain an upper bound on the number of times the algorithm evaluates a point $x_{h,i}$ before refining, denoted by q_h, as follows:

$$q_h = \min\{m : \beta_n \frac{\sigma}{\sqrt{m}} \leq V_h\}.$$

  On simplifying, we get the required result $q_h \leq \left(\sigma^2 \beta_n^2\right)/V_h^2$.

□

The first statement in the above lemma tells us that the points evaluated by the algorithm which lie deeper in the tree of partitions have smaller deviation from the threshold τ, or alternatively, the algorithm discretizes the search space coarsely in the regions far from the threshold an constructs finer partitions in the regions close to the threshold. The second statement provides a bound on the deviation of the evaluated points in terms of the posterior standard deviation, and thus combined with the first statement described how the algorithm balances exploration (evaluating points with high $\sigma_t(\cdot)$) and exploitation. Finally, the last satement of Lemma 1 tells us that the algorithm evaluates more points in the deeper parts of the tree of partitions.

Before proceeding to the convergence analysis of Algorithm 1, we need to introduce some definitions. For any $r > 0$, define $h_r := \max\{h \geq 0 : v_1 \rho^h \geq r\}$. Then we define the dimensionality of the region of $\mathcal{X}$ at which the function f takes values close to τ, as $\tilde{D} := D_\zeta$, where $\zeta(r) = \{x \in \mathcal{X} \mid |f(x) - \tau| \leq 10 V_{h_r}\}$ and D_ζ was introduced in Definition 2.

We note that by definition, the random variable $\tilde{D}$ is almost surely bounded by the metric dimension D_m of the metric space $(\mathcal{X}, d)$. More specifically for the case of $\mathcal{X} \subset [0,1]^D$, we have $\tilde{D} \leq D$ almost surely.

Finally, we introduce the term J_n which is the sum of the posterior variance of the points evaluated by the algorithm, defined as

$$J_n \coloneqq \sum_{t \in Q_n} \sigma_t^2(x_{h_t,i_t}),$$

where Q_n is the set of times at which the algorithm performed function evaluations.

We can now state the main result of this section, which bounds the approximation error of our proposed algorithm in two ways with high probability. The first bound is in terms of $\tilde{D}$, while the second bound is in terms of J_n.

Theorem 1. *Assuming that f is a sample from $GP(0,k)$ with $k \in \mathcal{K}$, the following two statements are true have with probability at least $1-2\delta$,*

$$\mathcal{L}(\hat{S}_\tau, S_\tau) = \tilde{\mathcal{O}}\left(n^{-\frac{\alpha}{\tilde{D}+2\alpha}}\right) \tag{4}$$

$$\mathcal{L}(\hat{S}_\tau, S_\tau) = \tilde{\mathcal{O}}\left(\beta_n \sqrt{J_n/n}\right) \tag{5}$$

where $\tilde{\mathcal{O}}$ suppresses the polylogarithmic factors. The term J_n in (5) can be further upper bounded by a constant times $I(y_{\mathcal{D}_n}; f_{\mathcal{D}_n})$, the mutual information between the function and the observations at the points of evaluation $\mathcal{D}_n$.

Proof Outline. The proof of this theorem combines ideas from the proofs of (Gotovos et al., 2013, Theorem 1) and from results in global optimization literature such as (Munos, 2011). More specifically, by definition of the terms $\bar{u}_t(\cdot)$ and $\bar{l}_t(\cdot)$, the upper bound on the deviation of the points chosen by the algorithm, $\bar{u}_t(x_{h_t,i_t}) - \bar{l}_t(x_{h_t,i_t})$, is monotonically non-increasing in t. Thus the maximum deviation from τ at any time t can be upper bounded by the average of the deviations of all the points evaluated by the algorithm up to that time. Lemma 1 gives us two ways of bounding the maximum deviation from τ of the evaluated points, one in terms of the posterior standard deviation of the evaluated points, and another in terms of the variation V_{h_t}, the variation in the function value in the cell. Using the standard deviation bounds, and proceeding as in (Srinivas et al., 2012; Gotovos et al., 2013), we can obtain the bound given in (5). Finally, combining the V_{h_t} based bound with our assumption on the metric space $(\mathcal{X}, d)$, we can obtain the dimension type bound given in (4) by using counting arguments similar to those used in (Munos, 2011; Wang et al., 2014). The details of the proof are given in Appendix A.2

Remark 4. The standard approach of obtaining explicit bounds in terms of n for $I(y_{\mathcal{D}_n}; f_{\mathcal{D}_n})$, as laid out in (Srinivas et al., 2012), consists of two steps: first bound $I(y_{\mathcal{D}_n}; f_{\mathcal{D}_n})$ by γ_n, the maximum information gain with n observations, defined as $\gamma_n \coloneqq \sup_{G \subset \mathcal{X}: |G| = n} I(\boldsymbol{y_G}; f)$, and then employ the bounds on γ_n derived in (Srinivas et al., 2012, Theorem 5) for some commonly used covariance functions to get the required bounds on the estimation error of the algorithm. This is also the approach followed to obtain the existing convergence guarantees for GP level set estimation (Gotovos et al., 2013; Bogunovic et al., 2016). In Section 3.3.1 we provide a more refined approach to bounding the term J_n for Algorithm 1.

Low Complexity Implementation: The computational complexity of Algorithm 1 in the worst case can be $\mathcal{O}(n^{\alpha\tilde{D}+3})$ which can be infeasible for large $\tilde{D}$. However, we can construct a low complexity version of Algorithm 1 with slightly weaker theoretical guarantees by the following modifications:

- Replace Line 11 in Algorithm 1 with the following selection rule: $x_{h_t,i_t} \in \arg\max_{x_{h,i} \in \mathcal{X}_t} |\tau - \mu_t(x_{h,i})| + \beta_n \sigma_t(x_{h,i}) + V_h$
- Remove the refinement rules in Lines 12-15 and refine a cell $\mathcal{X}_{h,i}$ if $x_{h,i}$ has been evaluated q_h times, where q_h is given in Lemma 1.

For this modified algorithm, it is easy to show that we can obtain dimension-type bounds on the estimation error given by (4). However, since we do not take the posterior standard deviation into account in the selection rule, we cannot obtain the information-type bound for this algorithm. On the other hand, the size of the active set at time t for this algorithm satisfies $|\mathcal{X}_t| \leq t$. Hence the computational cost of implementing this algorithm is dominated by the posterior calculation step which is a $\mathcal{O}(t^3)$ operation for any time t. Furthermore, since the cost of refining a cell is $\mathcal{O}(D)$ and there can be no more than n cell refinements, the total cost of implementing this algorithm is $\mathcal{O}(n^4 + Dn)$

Comparison with existing algorithms: Compared to the existing algorithms for level set estimation in the Bayesian framework, our algorithm has lower computational complexity as well as tighter guarantees on the estimation error.

The existing level set estimation algorithms with theoretical guarantees on their performance such (Gotovos et al., 2013; Bogunovic et al., 2016) assume that the search space is finite. They can, however, be easily extended to continuous search spaces by selecting query points from a sequence of increasing finite subsets of the search space $\mathcal{X}$ as suggested by Srinivas et al.

(2012). More specifically, if $\mathcal{X} \subset [0,1]^D$, then the existing algorithms at any time t, select a query point by solving an optimization problem over a uniform grid of size $\mathcal{O}(t^{2D})$. Thus with a budget of n function evaluations, the computational cost of implementing these algorithms is at least $\mathcal{O}\big(n^{2D+3}\big)$. The exponential dependence on D makes the application of these algorithms to higher dimensions infeasible. Practical implementations of these algorithms in higher dimensions must employ certain heuristics and approximations, which do not come with theoretical guarantees. In contrast, our algorithm admits a low complexity version with theoretical guarantees on the estimation error for which the cost of implementation has only a linear dependence on the dimension of the search space $\mathcal{X}$. Thus for larger values of D, the cost of implementing the low complexity version of our algorithm can be significantly smaller than the state of the art.

In addition to the computational benefits, the convergence guarantees presented in Theorem 1 for Algorithm 1 also improve upon results of (Gotovos et al., 2013) for the Mátern family of kernels in two ways:

- The bounds provided by (Gotovos et al., 2013) are only valid for $\nu > 1$ since no explicit bounds on γ_n are known for the Mátern kernel with $\nu = 1/2$. The dimension type bound of Theorem 1, in contrast, is valid for all $\nu \geq 1/2$. Thus by putting $\alpha = 1/2$ for the Mátern 1/2 kernel, we obtain an explicit upper bound on the estimation error of the form $\mathcal{O}(n^{-1/(2D+2)})$ when $\mathcal{X} \subset [0,1]^D$.

- For the case of $\nu > 1$, a sufficient condition under which the dimension type bounds given in (4) are tighter than those of (Gotovos et al., 2013) is when $D \geq \nu - 1$. This implies that for the two most commonly used kernels in machine learning applications, Mátern kernels with $\nu = 3/2$ and $\nu = 5/2$, the bounds of Theorem 1 are tighter than prior work for almost all dimensions. In Section 3.3.1, we will further relax this condition, by obtaining tighter bounds for all values of ν and D.

3.3.1 Tighter bounds on Information Gain

As mentioned earlier, the standard approach of bounding the information gain of the n evaluation points, as proposed by Srinivas et al. (2012), is to first bound it with γ_n, and then use the explicit bounds on γ_n derived in Theorem 5 of Srinivas et al. (2012). This approach does not utlize any knowledge about the distribution of the evaluation points in the space $\mathcal{X}$. In the case of Algorithm 1, however, since we know the evaluation points are only selected from the set $\cup_{h\geq 0}\mathcal{X}_h$, we can use this to provide a more fine grained characterization of the information gain.

Theorem 2. *Suppose Q_n denotes the times at which the algorithm performns function evaluations. Then for $J_n = \sum_{t\in Q_n} \sigma_t^2\left(x_{h_t,i_t}\right)$, we have the following with probability at least $(1-\delta)$:*

$$J_n \leq \sum_{h\geq 0} \big(\mathcal{I}_h(n_h, T_h) + \mathcal{O}(1)\big) \tag{6}$$

where n_h is the number of function evaluations performed by Algorithm 1 on points in $\mathcal{X}_h$, $T_h \in \{1,2,\ldots,n_h\}$ and the term $\mathcal{I}_h(n_h,T_h)$ is defined as follows:

$$\mathcal{I}_h(n_h,T_h) = \max_{1\leq s\leq n_h} \Bigg(T_h \log(sm_h/\sigma^2) + \sigma^{-2}(n_h - s)\sum_{i=T_h+1}^{m_h} \hat{\lambda}_i\Bigg). \tag{7}$$

In the above display, $m_h = 2^h\left(\log\left(\frac{2^h h_{\max}}{\delta}\right)\right)$ and $\hat{\lambda}_i$ denotes the i^{th} largest eigenvalue of the empirical covariance matrix computed at m_h points uniformly sampled from the set $\mathcal{Z}_h := \bigcup_{x_{h,i}\in\mathcal{X}_h} B(x_{h,i},\epsilon_h,d)$ for $\epsilon_h = \min\{v_2\rho^h, 1/n_h\}$.

Proof Outline. The proof of this theorem proceeds similarly to the proof of Theorem 8 of Srinivas et al. (2012) by relating the information gain to the spectrum of the covariance matrix computed at some finite subset of $\mathcal{X}$ and then further approximating it the spectrum of the corresponding Hilbert-Schmidt operator associated with the covariance function. However, one key difference is that instead of computing the covariance matrix over a uniform grid over $\mathcal{X}$ (as in Lemma 7.7 of (Srinivas et al., 2012)), we construct a sequence of uniform discretiztions by sampling points uniformly from sets of the form $\cup_{x\in\mathcal{X}_h} B(x,\epsilon_h,d)$ for all $h \geq 0$ and appropriate choice of ϵ_h. Due to this, we can replace the approximation error term (the last term in the statement of (Srinivas et al., 2012, Theorem 8)) with a $\mathcal{O}(1)$ term in the statement of our Theorem. The details are given in Appendix A.3

We now instantiate the bound described in Theorem 2 for the special case of Matern kernels with $\nu > 1$.

Theorem 3. *Suppose $\mathcal{X} \subset [0,1]^D$ and $I(y_{\mathcal{D}_n}; f_{\mathcal{D}_n})$ denotes the information gain for the set of points evaluated by Algorithm 1. Then if f is sampled from $GP(0,k)$ where k is a Mátern kernel with smoothness parameter $\nu > 1$, we have*

$$I(y_{\mathcal{D}_n}; f_{\mathcal{D}_n}) = \tilde{\mathcal{O}}\big(n^a\big) \tag{8}$$

where

$$a = \frac{D^2 + 3D}{4\nu + D^2 + 5D}.$$

Proof Outline. For proving the above theorem, we partition the evaluated points into two sets depending on whether their depth is more than some value $H \leq h_{max}$ or not. For the set of points with $h \leq H$, we bound the corresponding $\mathcal{I}_h$ values by making appropriate choice of the parameter T_h which balances the two terms of $\mathcal{I}_h$. The term n_h can be upper bounded by a $\mathcal{O}\left(\rho^{-h(2\alpha+D)}\right)$ term, and m_h can be upper bounded by a $\mathcal{O}\left(2^h \log(n)\right)$ term. For the set of points with $h > H$ we use a bound on posterior standard deviation using Lemma 1 and the cell refining rule of Algorithm 1. Finally, the depth H is chosen to balance the contributions of the terms with $h \leq H$ and $h > H$. The details are given in Appendix A.4.

The bound given by the above theorem is tighter than the existing bound on γ_n provided in Theorem 5 of (Srinivas et al., 2012) for all values of $\nu > 1$ and $D \geq 1$. Thus, in addition to the dimension dependent bound for $\nu = 1/2$, by employing the above result we have obtained tighter characterization of the estimation error of our algorithm for all Mátern kernels with half integer values of ν and for all values of D.

With some small modifications to the result of Theorem 3, similar bounds on the information gain can be derived for the Gaussian Process bandit algorithms in (Shekhar and Javidi, 2017), thus proving tighter characterization of the cumulative regret for all Mátern kernels.

4 Conclusion and Future work

In this paper we considered the problem of level set estimation of a black-box function from noisy observations. We proposed an algorithm for this problem in the Bayesian framework with GP prior and analyzed its performance. We showed that our proposed algorithm has lower computational complexity as well as tighter theoretical guarantees than existing algorithms. In the process, we also obtained tighter characterization of the information gain from n function evaluations for our proposed algorithm. Finally, we also considered the problem of level set estimation in the non-Bayesian framework with certain smoothness assumptions, and proposed an algorithm which does not require the knowledge of the smoothness parameters.

There are several directions along which the work presented in this paper can be extended. We conjecture that the bounds on the information gain of our algorithm obtained in Theorem 3 can be further improved by employing more careful counting arguments. Another important direction is to study the problem of level set estimation in the non-Bayesian setting, for example under Hölder continuity assumptions, and design computationally efficient algorithms which can automatically adapt to the unknown smoothness parameters.

Acknowledgements

The authors thank the three anonymous reviewers for their helpful feedback.

References

I. Bogunovic, J. Scarlett, A. Krause, and V. Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems*, pages 1507–1515, 2016.

B. Bryan, R. C. Nichol, C. R. Genovese, J. Schneider, C. J. Miller, and L. Wasserman. Active learning for identifying function threshold boundaries. In *Advances in neural information processing systems*, pages 163–170, 2006.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

D. Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

A. Gotovos, N. Casati, G. Hitz, and A. Krause. Active learning for level set estimation. In *IJCAI*, pages 1344–1350, 2013.

R. Kleinberg, A. Slivkins, and E. Upfal. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277*, 2013.

R. Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in neural information processing systems*, pages 783–791, 2011.

J. Scarlett. Tight regret bounds for bayesian optimization in one dimension. *arXiv preprint arXiv:1805.11792*, 2018.

J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. *arXiv preprint arXiv:1706.00090*, 2017.

M. W. Seeger, S. M. Kakade, and D. P. Foster. Information consistency of nonparametric gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

S. Shekhar and T. Javidi. Gaussian process bandits with adaptive discretization. *arXiv preprint arXiv:1712.01447*, 2017.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5): 3250–3265, 2012.

R. van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Z. Wang, B. Shakibi, L. Jin, and N. Freitas. Bayesian multi-scale optimistic optimization. In *Artificial Intelligence and Statistics*, pages 1005–1014, 2014.

R. M. Willett and R. D. Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.