An (ϵ, δ)-accurate level set estimation with a stopping criterion

Hideaki Ishibashi[1] (✉), Kota Matsui[2,4], Kentaro Kutsukake[2], and Hideitsu Hino[3]

[1] Kyushu Institute of Technology, Kitakyushu Fukuoka 808-0196, Japan
ishibashi@brain.kyutech.ac.jp
[2] Nagoya University, Nagoya Aichi 464-8601, Japan
matsui.kota.x3@f.mail.nagoya-u.ac.jp,
kutsukake.kentaro.c3@f.mail.nagoya-u.ac.jp
[3] The Institute of Statistical Mathematics, Tachikawa Tokyo, 190-0014, Japan
hino@ism.ac.jp
[4] RIKEN AIP, Chuo Tokyo, 103-0027, Japan

Abstract. The level set estimation problem seeks to identify regions within a set of candidate points where an unknown and costly to evaluate function's value exceeds a specified threshold, providing an efficient alternative to exhaustive evaluations of function values. Traditional methods often use sequential optimization strategies to find ϵ-accurate solutions, which permit a margin around the threshold contour but frequently lack effective stopping criteria, leading to excessive exploration and inefficiencies. This paper introduces an acquisition strategy for level set estimation that incorporates a stopping criterion, ensuring the algorithm halts when further exploration is unlikely to yield improvements, thereby reducing unnecessary function evaluations. We theoretically prove that our method satisfies ϵ-accuracy with a confidence level of $1 - \delta$, addressing a key gap in existing approaches. Furthermore, we show that this also leads to guarantees on the lower bounds of performance metrics such as F-score. Numerical experiments demonstrate that the proposed acquisition function achieves comparable precision to existing methods while confirming that the stopping criterion effectively terminates the algorithm once adequate exploration is completed.

Keywords: Level set estimation · Stopping criterion · Gaussian process · Adaptive experimental design.

1 Introduction

Adaptive experimental design is a data-driven approach to planning experiments that determines the next experimental conditions based on data obtained so far. It is applied in various fields of experimental sciences such as drug discovery [20] and the development of new materials [47]. For example, in manufacturing industries, identifying defective areas where the physical properties of materials do not meet the desired quality is a crucial issue. Such defective areas are often

determined by measuring the physical properties, using techniques like X-ray diffraction, in various parts of refined materials and determining whether they exceed acceptable lower limits. This problem can be formulated as a *Level Set Estimation* (LSE) by considering a black-box function that takes the coordinates of measurement locations as input and outputs the physical properties at each measurement location. LSE is a problem aimed to identify regions on the input space where the output values of a black-box function are greater (or smaller) than a certain threshold, and active learning (AL; [44]) approach has been proposed specifically to perform LSE with as few experimental iterations as possible [9,18,12,51].

In practical scenarios of adaptive experimental design, determining "when to stop the experiment" is crucially important. If stopping is not done appropriately, it can lead to wasteful experiments and the squandering of various costs. In experimental sciences, there are situations where an upper limit on the number of experiments that can be conducted. A naïve approach often involves conducting experiments up to such a "budget limit" and then stopping. For LSE, few theoretical guarantees exist on the consistency [7] or sample complexity [5], and there also exist some theoretical results on the finite-time guarantee of LSE [18,36], but to the best of the authors' knowledge, research on the stopping criterion for LSE is limited, with one example being the F-score sampling criterion [42]. This method stops when the 5th percentile of the sampled F-scores exceeds the desired F-score, and it allows for intuitive parameter setting and can be applied to a wide range of acquisition functions. However, in many applications, it is often unclear what is the maximum possible F-score for the problem, and the actual F-score at the stopping point may not exceed the desired F-score, making it difficult to stop the LSE procedure by specifying the F-score.

Contributions In this paper, we propose an acquisition function for LSE based on the distribution of a random variable that represents the difficulty of classification. The proposed acquisition function entails a natural stopping criterion, probabilistically ensuring that the algorithm can be appropriately stopped when the LSE is accurately performed when used with the proposed acquisition function. Furthermore, our method probabilistically guarantees the lower bounds of performance metrics such as the F-score, accuracy, recall, precision, and specificity. Experiments using test functions and real-world data on the quality of silicon ingots demonstrate that the proposed method performs at least as well as existing methods in terms of the F-score, and can stop the algorithm when sufficient estimation accuracy is achieved.

Related Works For active learning, stopping criteria based on various perspectives have been proposed. For example, it is investigated in [48] that the use of classifier confidence to determine that there are no informative instances remaining in the candidate point set and to stop AL. In [40], an intrinsic stopping criterion based on the exhaustiveness of the candidate point set is proposed, that does not depend on a predefined threshold parameter. A stopping criteria based on *Stabilizing Predictions* is proposed in [11], that checks the stability of the

current model's predictions on the validation set and decides whether to stop the AL. In [29,10,2], stopping criterion based on the change in the F-score is considered. Criteria called *TotalConf* and *LeastConf* are proposed in [37], which stop the AL based on the amount of change in the classification confidence (i.e., prediction uncertainty) for unlabeled data. A method to stop AL based on upper bound of the generalization error is proposed in [24]. For Bayesian optimization (BO), stopping criteria based on regret have been proposed [33,25,50]. Note that each of these studies concerns stopping criteria for active learning and adaptive experimental design for classification, regression and optimization tasks, and are not directly applicable to the LSE problem.

Similar to the LSE problem, the estimation of the excursion set (which is also known as the probability of failure of a system in the industrial world) has also been considered, and different approaches such as sequential experimental design and kriging have been employed to tackle it with criteria targeted to reduce the uncertainty about the level set [8,3,15]. Contour finding, which identifies the contour where a black-box function equals a given threshold, has been developed independently of LSE but is closely related to it [17,35,31]. Several extensions of LSE to various situations are also considered, such as LSE under input uncertainty [16,23,26], heavy-tailed output noise [32] or heteroscedasticity of outputs [52], settings that aim at distributionally robust LSE [22], dealing with Bernoulli observations [30], considering control over type-I and type-II errors [4], and the setting where the input is composed of both deterministic and uncertain parts [1].

2 Level Set Estimation

Consider an unknown function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is a finite set of input $\mathbf{x}$. This is a so-called pool-based problem. The objective of LSE is to classify, given a threshold $\theta \in \mathbb{R}$, whether the outputs $\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ corresponding to a given candidate point set $\mathcal{X}$ exceed θ, using as few datasets as possible. The upper/lower level sets are defined as $H_\theta = \{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) > \theta\}$ and $L_\theta = \{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) \leq \theta\}$, respectively. In LSE, the following procedure is iteratively performed to achieve this objective: i) Estimate the surrogate function $\hat{f}$ from the obtained dataset. ii) Utilize the surrogate function to classify each candidate point into any one of the upper-level set, the lower-level set, or the undetermined set. iii) Select the next search point based on the surrogate function. iv) Query the oracle for the corresponding output of the selected point. v) Add the obtained point to the dataset.

The surrogate function is often modeled by the Gaussian process regression (GPR; [49]). Consider a set of input-output pairs $S_N = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$. In GPR, we assume that the function $\hat{f}$ is generated by a Gaussian process (GP) with a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$. Additionally, the observed output y is assumed to have Gaussian noise with precision parameter λ added to the generated function $\hat{f}$. Therefore, in GPR, we consider the following generative model: $\hat{f}(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, and $y \mid \mathbf{x} \sim \mathcal{N}(\hat{f}(\mathbf{x}), \lambda^{-1})$.

Denoting $\mathbf{y} := (y_1, y_2, \ldots, y_N)$, the joint distribution of $\mathbf{y}$ and the output $\hat{f}(\mathbf{x}^*)$ for a new input $\mathbf{x}^*$ can be expressed by the following equation.

$$\begin{bmatrix} \mathbf{y} \\ \hat{f}(\mathbf{x}^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ m(\mathbf{x}^*) \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{K}} & \mathbf{k}(\mathbf{x}^*) \\ \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right), \tag{1}$$

where $\tilde{\mathbf{K}} = \mathbf{K} + \lambda^{-1}\mathbf{I}$, $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}(\mathbf{x}^*) = (k(\mathbf{x}_n, \mathbf{x}^*))_{n=1}^N \in \mathbb{R}^N$, and $\mathbf{m} = (m(\mathbf{x}_n))_{n=1}^N \in \mathbb{R}^N$. Therefore, the posterior distribution when observing the dataset S_N is given by $p(\hat{f}(\mathbf{x}^*) \mid \mathbf{y}) = \mathcal{N}(\hat{f}(\mathbf{x}^*) \mid \mu_N(\mathbf{x}^*), \sigma_N^2(\mathbf{x}^*))$. Here,

$$\mu_N(\mathbf{x}^*) = m(\mathbf{x}^*) + \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\tilde{\mathbf{K}}^{-1}(\mathbf{y} - \mathbf{m}), \tag{2}$$

$$\sigma_N^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\tilde{\mathbf{K}}^{-1}\mathbf{k}(\mathbf{x}^*). \tag{3}$$

In LSE using GPR, the next exploration point is determined based on the posterior distribution. Specifically, if we define the acquisition function $\alpha : \mathcal{X} \to \mathbb{R}$ parameterized by $p(\hat{f} \mid \mathbf{y})$, the next exploration point is determined by $\mathbf{x}^{\mathrm{new}} = \arg\max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x}; p(\hat{f} \mid \mathbf{y}), \theta)$. Although there are various types of acquisition functions, such as those based on confidence bounds [18] and expected improvement for level set estimation [51], we focus on a typical approach based on misclassification probability [14]. Assuming that the true function f is generated from the posterior distribution $p(\hat{f} \mid \mathbf{y})$, the probability $\Pr(\mathbf{x} \in L_\theta)$ can be expressed as follows, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian:

$$\Pr(\mathbf{x} \in L_\theta) = \int_{-\infty}^{\theta} p(\hat{f}(\mathbf{x}) \mid \mathbf{y}) d\hat{f}(\mathbf{x}) = \Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})} \right). \tag{4}$$

Similarly, $\Pr(\mathbf{x} \in H_\theta) = 1 - \Pr(\mathbf{x} \in L_\theta)$. Then, $p^{\min}(\mathbf{x}) = \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta)\}$ represents the difficulty of classifying the candidate point $\mathbf{x}$; hence we call this "misclassification probability" [14]. Similarly, we call $p^{\max}(\mathbf{x}) = \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta)\}$ "classification probability". Therefore, the following acquisition function selects the candidate points that are difficult to classify as the next points of evaluation:

$$\mathbf{x}^{\mathrm{new}} = \arg\max_{\mathbf{x} \in \mathcal{X}} \ p^{\min}(\mathbf{x}). \tag{5}$$

When classifying candidate points, the standard method is the classification rule based on confidence intervals proposed by [18]. Let $\tilde{H}_\theta$ and $\tilde{L}_\theta$ be estimated upper-level set and lower-level set, respectively. We further introduce an undetermined set $\tilde{U}_\theta$. Then, a candidate point $\mathbf{x}$ is classified according to the following classification rule:

$$\tilde{H}_\theta = \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}, \mu_N(\mathbf{x}) - \beta\sigma_N(\mathbf{x}) > \theta\}, \tag{6}$$

$$\tilde{L}_\theta = \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}, \mu_N(\mathbf{x}) + \beta\sigma_N(\mathbf{x}) < \theta\}, \tag{7}$$

$$\tilde{U}_\theta = \mathcal{X} \backslash \{\tilde{H}_\theta \cup \tilde{L}_\theta\}, \tag{8}$$

where β is the parameter that controls the exploration-exploitation trade-off in the acquisition function.

3 Proposed Acquisition Function and Stopping Criterion

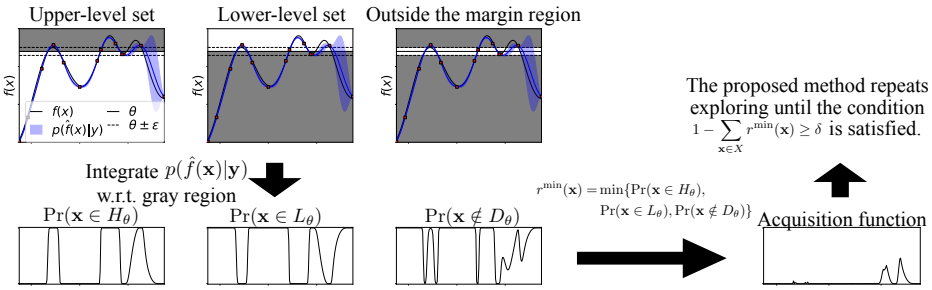


Fig. 1: The proposed method selects a candidate point that is difficult to classify and has a low probability of containing the true function value in the margin region and stops LSE when the condition is satisfied.

This section describes the acquisition function for LSE proposed in this paper and its stopping criterion. The pseudocode of the proposed LSE procedure is shown in Appendix B.2.

3.1 Proposed Acquisition Function

The acquisition function based on misclassification probability (5) is an intuitive and natural choice, where points with Bernoulli distribution parameters close to 0.5, and therefore difficult to classify, are selected as candidates for the next observation. In this formulation, noting that the cumulative distribution function of the standard Gaussian is $\Phi(0) = 0.5$, Eq. (4) suggests two possible scenarios for the selected candidate points. The first case occurs when the true function value at the candidate point is far from the threshold, but due to insufficient data observed near the candidate point, the posterior distribution's variance is large, making classification difficult ($\Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) \to 0.5$ as $\sigma_N(\mathbf{x}) \to \infty$). In this case, exploring the candidate point reduces the variance of the posterior distribution and increases the classification probability, making it less likely to be explored in subsequent searches. This is the case the exploration offers reasonable information.

On the other hand, the second case is problematic. The acquisition function (5) would select points at which the true function values are close to the threshold ($\Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) \to 0.5$ as $\mu_N(\mathbf{x}) \to \theta$). In this scenario, the same candidate point is repeatedly explored while other candidate points are ignored. For this issue, the previous study has heuristically used the product of the misclassification probability and the posterior variance as the acquisition function [14].

In this study, we assume that a margin $\epsilon > 0$ is given[5], which indicates a tolerance of the accuracy of estimation. If the gap between the true function value $f(\mathbf{x})$ and threshold θ for a candidate point $\mathbf{x}$ lies within the range $\mathcal{E} := (-\epsilon/2, \epsilon/2]$, it is considered as a difficult to classify, and that exploring this candidate point will not increase certainty about the classification, and thus the point is removed from the candidate point set. To put it another way, for candidate points that are difficult to classify, we decide to make a concession and perform an ϵ-accurate classification. The notion of the margin is essentially equivalent to ϵ-accuracy introduced in [12], but we explicitly utilize it as information for determining the next experimental condition. The difficult-to-classify set is defined as $U_\theta = \{\mathbf{x} \in \mathcal{X} \mid (f(\mathbf{x}) - \theta) \in \mathcal{E}\}$, and the solution triplet $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ is ϵ-accurate if $\forall \mathbf{x} \in \tilde{H}_\theta$ is in H_θ, $\forall \mathbf{x} \in \tilde{L}_\theta$ is in L_θ, and $\forall \mathbf{x} \in \tilde{U}_\theta$ is in U_θ.

The probability of $\mathbf{x} \in U_\theta$ is given by

$$\begin{aligned}\Pr(\mathbf{x} \in U_\theta) &= \int_{\theta-\epsilon/2}^{\theta+\epsilon/2} p(\hat{f}(\mathbf{x}) \mid \mathbf{y}) d\hat{f}(\mathbf{x}) \\ &= \Phi\left(\frac{\theta + \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) - \Phi\left(\frac{\theta - \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right).\end{aligned}$$

Similarly, $\Pr(\mathbf{x} \notin U_\theta) = 1 - \Pr(\mathbf{x} \in U_\theta)$. Then, with $r^{\min}(\mathbf{x}) := \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \notin U_\theta)\}$, we redefine the acquisition function as

$$\mathbf{x}^{\text{new}} = \underset{\mathbf{x} \in \mathcal{X}}{\arg\max}\ r^{\min}(\mathbf{x}). \tag{9}$$

As shown in Fig. 1, this acquisition function evaluates not only the probability that a candidate point belongs to the upper/lower level sets but also the probability $\Pr(\mathbf{x} \notin U_\theta)$ that the gap does not fall within the range $\mathcal{E}$. For a candidate point $\mathbf{x}$ where the gap $f(\mathbf{x}) - \theta$ is within $\mathcal{E}$, if the area around the candidate point has not been well explored, $\Pr(\mathbf{x} \notin U_\theta)$ increases, and if $p^{\min}(\mathbf{x})$ is also large, then $r^{\min}(\mathbf{x})$ increases, leading to the selection of $\mathbf{x}$. Conversely, if the area around the candidate point has been thoroughly explored, the posterior variance decreases, thus increasing the probability that the gap $f(\mathbf{x}) - \theta$ falls within $\mathcal{E}$ and decreasing $\Pr(\mathbf{x} \notin U_\theta)$. Therefore, even if $p^{\min}(\mathbf{x})$ is large and classification is difficult, $r^{\min}(\mathbf{x})$ becomes small, making it less likely to be chosen as the next point of evaluation.

As similar approaches to the misclassification-based approach, there are entropy-based and variance-based approaches [14,17]. These acquisition functions share the fundamental idea with the one in (5) and therefore inherit similar issues to those mentioned at the beginning of this section regarding (5). Several studies have discussed approaches to address the issues of these acquisition functions [35,41], but none provide theoretical guarantees on stopping performance, leaving the evaluation of this aspect to empirical analysis. In contrast, our acquisition function addresses the aforementioned issues while also providing theoretical guarantees on stopping performance, as discussed in the next section.

[5] Here, the margin is assumed to be given, but a method for setting the margin based on the observed data are discussed in Appendix. B.1.

3.2 Classification rule and stopping criterion

We describe the classification rule and stopping method for LSE using the acquisition function Eq. (9). Letting $r^{\max}(\mathbf{x}) := \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}$, the proposed classification rule is as follows:

$$\tilde{H}_\theta := \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}, r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)\}, \tag{10}$$

$$\tilde{L}_\theta := \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}, r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in L_\theta)\}, \tag{11}$$

$$\tilde{U}_\theta := \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}, r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in U_\theta)\}. \tag{12}$$

As will be discussed later, this classification rule can be considered equivalent to the classification rule of Eqs. (6), (7), and (8) under certain conditions.

The proposed stopping criterion uses a confidence parameter δ $(0 < \delta < 1)$ as a threshold, and LSE is stopped when the following inequality is satisfied:

$$1 - \sum_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x}) \geq \delta. \tag{13}$$

That is, LSE is stopped when the sum of the acquisition function values for all candidate points becomes small enough. At the point of stopping LSE, the following probability inequality holds:

Theorem 1. *If we assume that $\tilde{H}_\theta$,$\tilde{L}_\theta$ and $\tilde{U}_\theta$ are determined by using the classification rule of Eqs. (10),(11) and (12), then the following inequality holds:*

$$\Pr((\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta) \text{ is } \epsilon\text{-accurate}) \geq 1 - \sum_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x}). \tag{14}$$

The proof is shown in Appendix A. From this theorem, when Eq. (13) is satisfied, stopping LSE guarantees that $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ is ϵ-accurate with a probability of at least δ. Therefore, we refer to the LSE that uses the combination of the proposed acquisition function and stopping criterion as the (ϵ, δ)-*accurate* LSE. Since the left-hand side of Eq. (14) can be evaluated by sampling functions according to the GP posterior distribution, we provide the tightness of the proposed lower bound in the Appendix C.1.

By using theorem 1, we can also guarantee the lower bound of performance measures. Here, we present only the lower bound of the F-score as follows.

Proposition 1. *If we assume that $\tilde{H}_\theta$,$\tilde{L}_\theta$ and $\tilde{U}_\theta$ are determined by using the classification rule of Eqs. (10),(11) and (12), then the inequality*

$$\textit{F-score} \geq \frac{2|\tilde{H}_\theta|}{2|\tilde{H}_\theta| + |\tilde{U}_\theta|}$$

holds with probability $1 - \sum_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x})$.

The lower bound of other performance measures such as accuracy, recall, precision, and specificity, and their proofs are shown in Appendix A. In [43], the

lower bound of the F-score could not be analytically computed, so it is estimated using sampling. In contrast, our method provides lower bounds for the various measures such as F-score.

The standard classification rule and the proposed classification rule can be considered the same under certain conditions. The standard classification rule can be interpreted as follows: $\mathbf{x}$ is classified into upper-level set when $\Pr(\mathbf{x} \in H_\theta) > \Phi(\beta)$ is satisfied, $\mathbf{x}$ is classified into lower-level set when $\Pr(\mathbf{x} \in L_\theta) > \Phi(\beta)$ is satisfied, and $\mathbf{x}$ is classified into undetermined set when the both of conditions are not satisfied. Regarding $\Pr(\mathbf{x} \in U_\theta)$ as $\Phi(\beta)$, the standard classification is equivalent to the proposed classification rule under the assumption that $\Phi(\beta) > 0.5$[6].

By using the above relationship, we can show that the triplet $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ is ϵ-accurate in the case of the standard classification rule. Here, β in the standard classification rule and ϵ in the proposed classification rule can be mutually converted. Note that β also changes for each $\mathbf{x}$ in general even if ϵ is common to all $\mathbf{x}$ since $\Pr(\mathbf{x} \in U_\theta)$ varies depending on $\mathbf{x}$. We denote $\Pr(\mathbf{x} \in U_\theta)$ as $g(\epsilon \mid \mathbf{x})$, then there is an inverse mapping of $g(\epsilon \mid \mathbf{x})$ because $g(\cdot \mid \mathbf{x}) : \mathbb{R}^+ \to (0, 1)$ is a strictly increasing function with respect to $\epsilon \in \mathbb{R}^+$. Therefore, the following mutual conversions between ϵ and β hold:

$$\beta = \Phi^{-1}\left(g(\epsilon \mid \mathbf{x})\right), \quad \epsilon = g^{-1}(\Phi(\beta) \mid \mathbf{x}).$$

With these conversions, we can show that the triplet $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ is ϵ-accurate when we use the standard classification rule as follows:

Corollary 1. *We assume that $\tilde{H}_\theta$,$\tilde{L}_\theta$ and $\tilde{U}_\theta$ are determined by using the classification rule of Eqs. (6), (7), and (8) with β and $\Phi(\beta) > 0.5$. Let $\tilde{r}^{\min}(\mathbf{x}) = \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), 1 - \Phi(\beta)\}$. Then, the following inequality holds:*

$$\Pr((\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta) \text{ is } g^{-1}(\Phi(\beta) \mid \mathbf{x})\text{-accurate}) \geq 1 - \sum_{\mathbf{x} \in \mathcal{X}} \tilde{r}^{\min}(\mathbf{x}). \tag{15}$$

The proof is shown in Appendix A.

3.3 Choice of ϵ and δ

We explain how to determine the parameters ϵ and δ, and its sensitivity. Regarding δ, the proposed lower bound tends to increase monotonically, and the bound becomes tighter as the true probability of ϵ-accuracy approaches 1 as shown in Appendix. C.1. Therefore, we just have to set δ close to 1, such as $\delta = 0.99$. Since the stopping time does not change significantly when δ is close to 1, we can say that the stopping timing tends to be insensitive to the choice of δ.

[6] The condition $\Phi(\beta) > 0.5$ is added because in the standard classification rule, when $\Phi(\beta) \leq 0.5$, there is a possibility that $\mathbf{x}$ belongs to both the upper-level and lower-level sets. The standard classification rule often uses values such as $\beta = 1.96$, which corresponds to $\Phi(\beta) = 0.975$, implicitly assuming $\Phi(\beta) > 0.5$.

On the other hand, it is difficult to set ϵ appropriately, since it depends on the range of the objective function and the noise variance. In this study, instead of determining ϵ directly, we determine ϵ as follows:[7]

$$\epsilon(\mathbf{x}) = 2\sqrt{\frac{\lambda^{-1}k(\mathbf{x},\mathbf{x})}{\lambda^{-1}+Lk(\mathbf{x},\mathbf{x})}}\Phi^{-1}\left(1-\frac{1-\delta}{2|\mathcal{X}|}\right), \tag{16}$$

where L is a parameter set by the user instead of ϵ, and it can be interpreted as the minimum number of observations required per candidate point, even in cases where classification is not possible. As shown in the Appendix. C.6, L is less sensitive to the range of the function and the noise variance than directly specifying ϵ, making it a more robust parameter.

3.4 Computational cost

The proposed stopping criterion only requires the cumulative distribution function (CDF) of the standard normal distribution, and it does not require any sampling. Since the CDF of the standard normal distribution can be efficiently computed using libraries, the computational cost increases only linearly with the number of candidate points. In contrast, F-scores sampling (FS) [43] requires sampling functions from the posterior distribution, resulting in a quadratic increase in computational cost with respect to the number of candidate points. Therefore, compared to FS, the proposed stopping criterion remains computationally feasible even as the number of candidate points increases.

4 Experimental results

In this section, we demonstrate the effectiveness of the proposed acquisition function using both synthetic data and a practical application for estimating the red zone in silicon ingots.[8] In all experiments, the threshold θ that defines the level set is a pre-fixed value, but the results of setting θ to several different values are also shown in the Appendix C.7. Note that consistent results are obtained even when different thresholds are used.

4.1 LSE for test functions

Aiming at demonstrating the applicability of the proposed method across functions with various shapes, we evaluate the proposed method using test functions commonly used as benchmarks in the study of optimization algorithms. The test functions employed in this experiment are the `Rosenbrock`, `Branin`, and

[7] When using a stationary kernel, ϵ is independent of $\mathbf{x}$. Please refer to the Appendix. B.1 for the detailed derivation of this equation.

[8] In these experiments, we use a Macbook Pro with Apple M1 Max (10-core CPU, 32-core GPU and 32GB memory), and implemented with Python and library GPy [19]. The code is available at `https://github.com/hideaki-ishibashi/stopping_LSE`

`Cross in tray` functions[9], each representing a different landscape: one with a single local minimum with a spherical vicinity, one with a valley-like structure, and one with multiple local minima. Additional results are discussed in the Appendix C.3. For each function, thresholds are set, and candidate points exceeding these thresholds are considered part of the true upper set, while those below are viewed as the true lower set. The thresholds are set as follows: $\theta = 100$ for the `Rosenbrock` function and the `Branin` function, $\theta = -1.5$ for the `Cross in tray` function. Although these test functions have continuous domains, they are discretized into a grid of $20 \times 20 = 400$ points, which serve as observation candidates. In the Appendix C.2, we show that the stopping timing of the proposed method tends not to change even if the number of candidate points increases. Any of these points may be selected by acquisition functions, and repeated selections of the same points is allowed. Gaussian noise is added to the observations, with standard deviations set according to the range of each test function: $\sigma_{\text{noise}} = 30$ for the `Rosenbrock` function, $\sigma_{\text{noise}} = 20$ for the `Branin` function, and $\sigma_{\text{noise}} = 0.01$ for the `Cross in tray` function.

The proposed method is evaluated based on both the performance of the acquisition function and the efficiency achieved by early stopping. Generally, the performance of the LSE acquisition function is assessed using the F-score, which compares the predicted upper/lower level sets to the true upper/lower level sets over the candidate points. Not all candidate points may be classified in every search due to the classification rules. For the evaluation purpose, unclassified candidate points are assigned to the upper or lower level sets only for the F-score calculation if the posterior mean of GP exceeds or falls below the threshold, respectively. The performance of the acquisition function is evaluated based on the mean and variance of the convergence speed of the F-score when the LSE algorithm is executed using five randomly selected initial points. On the other hand, the effectiveness of the stopping criterion is evaluated based on the stopping time and the F-score at that moment. In other words, a good stopping criterion allows the algorithm to stop with fewer observations while achieving a high F-score.

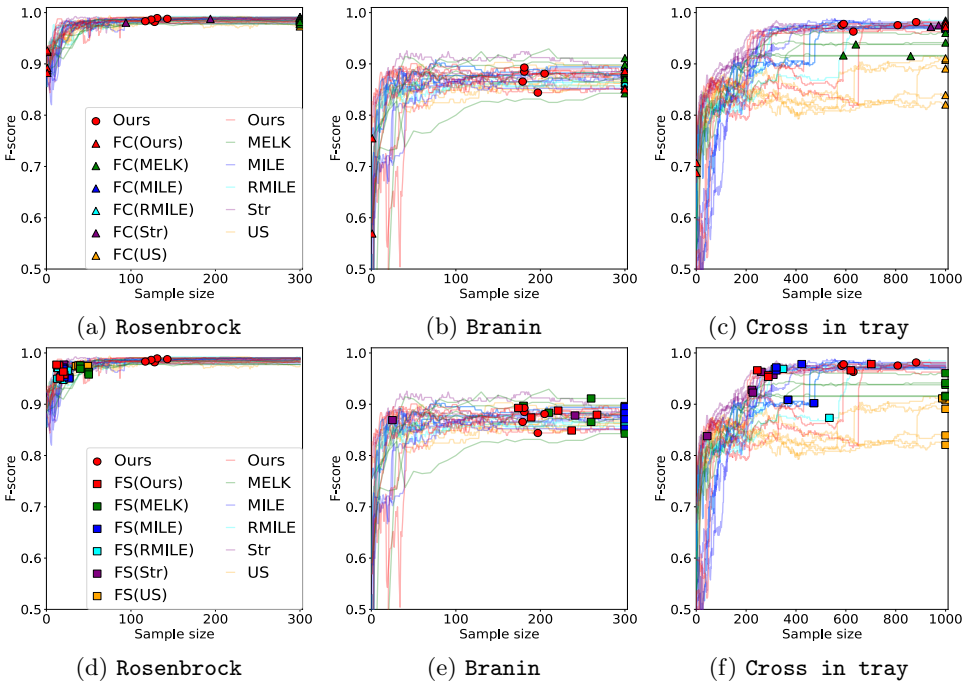
Comparison methods The level set estimation problem is also related to Bayesian optimization [38] and bandit problems [46], and its applications range from brain science [34] to astronomy [6,39] for example. Various algorithms (acquisition functions) have been proposed, but in this study, we compare those that are considered particularly major types and important in terms of practical performance: in addition to uncertainty sampling (US), which selects points that maximize the predictive variance of the Gaussian process as a baseline, we consider Straddle [14], MILE [18], RMILE [18], and MELK [36], which is a recently proposed sampling method based on experimental design. Although many other methods exist, they do not consistently outperform those mentioned here. It should also be noted that the main focus of this paper is the proposal of an acquisition function equipped with a stopping criterion. In these acquisition func-

[9] `https://www.sfu.ca/~ssurjano/optimization.html`

tions, candidate points are classified according to Eqs. (6), (7), and (8), where $\beta = 1.96$. On the other hand, in the proposed method, candidate points are classified according to Eqs. (10), (11), and (12). The same value of $\beta = 1.96$ is used for Straddle, MILE, and RMILE. MELK assumes that candidate points are not reclassified and that multiple points are sampled simultaneously. Following the settings of the previous study [36], MELK samples 10 points at a time without reclassifying candidate points. In contrast, other methods reclassify candidate points and sample one point at a time. RMILE's robust adjustment parameter ν is set to $\nu = 0.1$ according to the previous studies [51,22]. The proposed acquisition function also requires setting a parameter L, which is conservatively set to $L = 5$ to address the complex shape function based on the experimental results in Appendix C.6. The threshold for the proposed stopping criterion is set at $\delta = 0.99$. To evaluate the stopping criterion of the proposed method, we consider two stopping criteria. One is a standard stopping criterion which stops LSE when all candidate points are classified' (we call this the *fully classified (FC)* criterion), and the other is a stopping criteria based on sampling F-scores [43] (the criterion referred to as F-score Sampling (FS)). The stopping times of these stopping criteria are compared with the stopping times when using the proposed acquisition function and stopping criterion. In the FS criterion, as hyperparameters, we need to set the desired F-score and the probability of exceeding that F-score. In this experiment, we set the desired F-score and the probability to 0.95 and 95% (that is, 5th percentile).

Hyper-parameter setting In this experiment, we consider a GP with the mean function set to θ and the covariance function defined by a Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \rho \exp(-\frac{1}{2l^2}\|\mathbf{x} - \mathbf{x}'\|^2)$. The mean function is set to θ to ensure that, in the absence of any observed data, the probability of unobserved candidate points being classified into either the upper or lower level set is 50%. This setting can be adjusted based on any prior knowledge available in real applications. The variance ρ of the Gaussian kernel, the kernel width l, and the noise precision λ are hyperparameters, which are estimated by maximizing the marginal likelihood of the observed data each time a search is conducted using LSE. To prevent large fluctuations in the hyperparameters with each search, gamma priors are placed on ρ and l. Additionally, the noise precision λ is constrained within the range $[10^{-6}, 10^{6}]$ to prevent it from becoming infinite.

Results The F-scores for each acquisition function and the respective stopping timings are shown in Figs. 2. Although there are slight differences between individual test functions, no acquisition function, including the proposed method, significantly outperforms the baseline US or is markedly inefficient. Thus, it is crucial to stop LSE at the right moment when the F-scores have converged to enhance the efficiency of LSE. Regarding the stopping timings, the fully classified criterion often fails to stop the LSE even when the budget is fully utilized except for MELK in `Cross in tray` function. The inability of the FC criterion to stop is due to the occurrence of difficult-to-classify candidate points when function values at candidate points equal the threshold, and classification be-



(a) `Rosenbrock` (b) `Branin` (c) `Cross in tray`

(d) `Rosenbrock` (e) `Branin` (f) `Cross in tray`

Fig. 2: F-scores using each acquisition function and stoped timings with the proposed (Ours), F-score sampling (FS) and fully classified (FC) criteria for test functions. (a)–(c) Stopping time of FC and Ours. (d)–(f) Stopping times of FS and Ours.

comes more challenging as noise is added to the data, distancing the function values from the threshold. In MELK, it is sometimes possible to stop LSE even when the fully classified stopping criteria are used, as shown in Fig. 2(c), since it does not reclassify candidate points. However, the F-score of MELK may be lower than that of other methods, as it cannot correct candidate points that have been misclassified. In the FS criterion, as shown in Fig. 2(d), when the F-score converges, LSE can be stopped regardless of the acquisition function used. However, as shown in Figs. 2(f), despite the F-score not having converged, FS stops LSE. This is because the desired F-score is set to 0.95 in this experiment. This value was suitable when the F-score converged to 1, as in case `Rosenbrock`. However, it was not suitable when the noise was high, and the F-score did not converge to 1, as in cases of `Branin` function. Moreover, when noise was low, FS stopped LSE before the F-score converged to 1. Therefore, it is necessary to set the appropriate desired F-score according to the situation in the FS. Furthermore, under the FS criterion, despite setting the desired threshold to 0.95, the actual F-scores at the stopping time tend to be lower than 0.95 as shown in Fig. 2(e) and (f). Therefore, in practical applications, the desired value should

be set higher than expected. In contrast, the proposed method, despite using the same parameters, can stop at the time when convergence occurs, regardless of where the F-score converges. This demonstrates that, compared to FS, the proposed method does not require the parameters of the stopping criterion to be finely tuned to the specific problem.

4.2 Red-zone estimation of silicon ingots

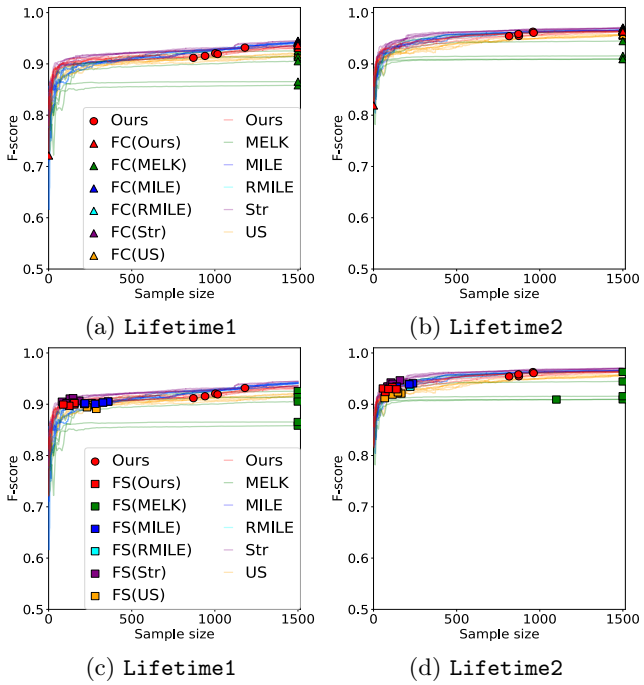
We demonstrate the effectiveness of the proposed method when using LSE to estimate the red zone in silicon ingots used in solar cells. The objective in this problem is to estimate regions contaminated with impurities (called the red zone) that are unsuitable for solar cell production. Typically, red zone estimation is performed through spatial mapping with measurement points placed in a regular grid, which is very time-consuming. Recently, the efficiency of red zone estimation using LSE has been proposed [21]. The data used in the experiments consist of lifetime measurements taken at grid points on two different types of silicon ingots, with each ingot measured at a grid of 161×121 points [28]. Hereinafter, the lifetime data from the first silicon ingot will be referred to as `Lifetime1`, and from the second ingot as `Lifetime2`. In both cases, the threshold is set to $\theta = 230$.

The performance of LSE methods are evaluated, in the similar manner to the test functions, by observing the F-scores, and the F-score at the stopping timing, with initial values changed randomly five times. The same methodology as for the test functions was used for comparison, but once a candidate point is selected, it is not selected again to estimate the noise because we have only one observation for each point. The parameters of the acquisition functions, the GP prior, and the hyperparameter estimation method were employed in the same manner as for the test functions.

As shown in Figs. 3, the transition of the F-score shows no significant differences regardless of the acquisition function used, except for MELK. In MELK, the F-score tends to converge to a low value. This is because MELK does not reclassify candidate points. In the perspective of the stopping timings, the FC stopping criterion fails to stop even after the entire budget is used. On the other hand, the proposed criterion allows for early stopping once the F-scores converge. This is likely due to measurement noise, leading to difficult-to-classify candidate points. Thus, the proposed method effectively stops the LSE in red zone estimation. FS criterion stops LSE earlier than the proposed criterion. However, the F-score continues to gradually increase even after FS stops, and the appropriate stopping point varies depending on the situation.

5 Discussion and Conclusion

In this paper, we proposed an acquisition function for level set estimation by directly modeling the difficulty of classifying into upper/lower sets. The proposed acquisition function is based on the notion of the ϵ-accuracy [12], and an adaptive



(a) `Lifetime1` (b) `Lifetime2`

(c) `Lifetime1` (d) `Lifetime2`

Fig. 3: F-scores using each acquisition function and stoped timings with the proposed (Ours), F-score sampling (FS) and fully classified (FC) criteria for red zone estimation. (a), (b) Stopping time of FC and Ours. (c), (d) Stopping times of FS and Ours.

determination method of the ϵ parameter is proposed. A stopping criterion for the algorithm was also proposed. When applied to both synthetic and real data, the proposed method performed comparably to existing methods in terms of acquisition function performance and was able to stop the algorithm early, even in the presence of observation noise. Empirically, the proposed method tends to be conservative. This behavior can be beneficial in some cases but harmful in others. Balancing theoretical guarantees with more aggressive stopping remains an open problem for future work. Another direction for future work is extending the method to query-based problems where evaluation points are selected from a continuous domain [45]. Additionally, extending the method to high-dimensional problems is also an important issue.

Acknowledgments

This work was supported by Grants-in-Aid from the Japan Society for the Promotion of Science (JSPS) for Scientific Research (KAKENHI grant nos.

JP23K28146 and JP24K20836 to K.M., JP24K15088 to H.I., and JP23K24909, 25H01494 and JPMJMI21G2 to H.H.).

References

1. Ait Abdelmalek-Lomenech, R., Bect, J., Chabridon, V., Vazquez, E.: Bayesian sequential design of computer experiments for quantile set inversion. Technometrics pp. 1–10 (2024). https://doi.org/10.1080/00401706.2024.2394475
2. Altschuler, M., Bloodgood, M.: Stopping active learning based on predicted change of f measure for text classification. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC). pp. 47–54 (Jan 2019). https://doi.org/10.1109/ICOSC.2019.8665646
3. Azzimonti, D., Bect, J., Chevalier, C., Ginsbourger, D.: Quantifying uncertainties on excursion sets under a Gaussian random field prior. SIAM/ASA journal on uncertainty quantification **4**(1), 850–874 (Jan 2016). https://doi.org/10.1137/141000749
4. Azzimonti, D., Ginsbourger, D., Chevalier, C., Bect, J., Richet, Y.: Adaptive design of experiments for conservative estimation of excursion sets. Technometrics: a journal of statistics for the physical, chemical, and engineering sciences **63**(1), 13–26 (Jan 2021). https://doi.org/10.1080/00401706.2019.1693427
5. Bachoc, F., Cesari, T., Gerchinovitz, S.: The sample complexity of level set approximation. In: The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 130, pp. 424–432. PMLR (2021)
6. Beaky, M.M., Scherrer, R.J., Villumsen, J.V.: Topology of large-scale structure in seeded hot dark matter models. The astrophysical journal **387**, 443 (Mar 1992). https://doi.org/10.1086/171097
7. Bect, J., Bachoc, F., Ginsbourger, D.: A supermartingale approach to Gaussian process based sequential design of experiments. Bernoulli **25**(4A), 2883 – 2919 (2019). https://doi.org/10.3150/18-BEJ1074
8. Bect, J., Ginsbourger, D., Li, L., Picheny, V., Vazquez, E.: Sequential design of computer experiments for the estimation of a probability of failure. Statistics and Computing **22**, 773–793 (2012)
9. Bichon, B.J., Eldred, M.S., Swiler, L.P., Mahadevan, S., McFarland, J.M.: Efficient global reliability analysis for nonlinear implicit performance functions. AIAA journal **46**(10), 2459–2468 (Oct 2008). https://doi.org/10.2514/1.34321
10. Bloodgood, M., Grothendieck, J.: Analysis of stopping active learning based on stabilizing predictions. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. pp. 10–19. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013)
11. Bloodgood, M., Vijay-Shanker, K.: A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). pp. 39–47. Association for Computational Linguistics, Boulder, Colorado (Jun 2009)
12. Bogunovic, I., Scarlett, J., Krause, A., Cevher, V.: Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In: Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)

13. Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics **7**(4), 434–455 (1998). https://doi.org/10.1080/10618600.1998.10474787
14. Bryan, B., Nichol, R.C., Genovese, C.R., Schneider, J., Miller, C.J., Wasserman, L.: Active learning for identifying function threshold boundaries. Advances in neural information processing systems **18** (2005)
15. Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., Richet, Y.: Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics: a journal of statistics for the physical, chemical, and engineering sciences **56**(4), 455–465 (Oct 2014). https://doi.org/10.1080/00401706.2013.860918
16. Chevalier, C., Ginsbourger, D., Bect, J., Molchanov, I.: Estimating and quantifying uncertainties on level sets using the vorob'ev expectation and deviation with Gaussian process models. In: Contributions to Statistics, pp. 35–43. Contributions to statistics, Springer International Publishing, Heidelberg (2013). https://doi.org/10.1007/978-3-319-00218-7_5
17. D. Austin Cole, Robert B. Gramacy, J.E.W.G.F.B.P.E.L., Leser, W.P.: Entropy-based adaptive design for contour finding and estimating reliability. Journal of Quality Technology **55**(1), 43–60 (2023). https://doi.org/10.1080/00224065.2022.2053795
18. Gotovos, A., Casati, N., Hitz, G., Krause, A.: Active learning for level set estimation. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. AAAI Press (2013)
19. GPy: GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy (since 2012)
20. Griffiths, R.R., Hernández-Lobato, J.M.: Constrained bayesian optimization for automatic chemical design using variational autoencoders. Chemical science **11**(2), 577–586 (2020)
21. Hozumi, S., Kutsukake, K., Matsui, K., Kusakawa, S., Ujihara, T., Takeuchi, I.: Adaptive defective area identification in material surface using active transfer learning-based level set estimation. ArXiv **abs/2304.01404** (2023)
22. Inatsu, Y., Iwazaki, S., Takeuchi, I.: Active learning for distributionally robust level-set estimation. In: International Conference on Machine Learning. pp. 4574–4584. PMLR (2021)
23. Inatsu, Y., Karasuyama, M., Inoue, K., Takeuchi, I.: Active learning for level set estimation under input uncertainty and its extensions. Neural Computation **32**(12), 2486–2531 (2020)
24. Ishibashi, H., Hino, H.: Stopping criterion for active learning based on deterministic generalization bounds. In: The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]. Proceedings of Machine Learning Research, vol. 108, pp. 386–397. PMLR (2020)
25. Ishibashi, H., Karasuyama, M., Takeuchi, I., Hino, H.: A stopping criterion for bayesian optimization by the gap of expected minimum simple regrets. In: International Conference on Artificial Intelligence and Statistics. pp. 6463–6497. PMLR (2023)
26. Iwazaki, S., Inatsu, Y., Takeuchi, I.: Bayesian experimental design for finding reliable level set under input uncertainty. IEEE Access **8**, 203982–203993 (2020)
27. Kish, L.: Survey Sampling. Wiley (1965)

28. Kutsukake, K., Deura, M., Ohno, Y., Yonenaga, I.: Characterization of silicon ingots: Mono-like versus high-performance multicrystalline. Japanese Journal of Applied Physics **54**(8S1), 08KD10 (jul 2015). https://doi.org/10.7567/JJAP.54.08KD10
29. Laws, F., Schütze, H.: Stopping criteria for active learning of named entity recognition. In: Proceedings of the 22nd International Conference on Computational Linguistics. vol. 1, pp. 465–472. Association for Computational Linguistics, Morristown, NJ, USA (2008). https://doi.org/10.3115/1599081.1599140
30. Letham, B., Guan, P., Tymms, C., Bakshy, E., Shvartsman, M.: Look-ahead acquisition functions for Bernoulli level set estimation. In: AISTATS. Proceedings of Machine Learning Research, vol. abs/2203.09751, pp. 8493–8513. PMLR (Mar 2022). https://doi.org/10.48550/arXiv.2203.09751
31. Li, J., Li, J., Xiu, D.: An efficient surrogate-based method for computing rare failure probability. Journal of Computational Physics **230**(24), 8683–8697 (2011). https://doi.org/https://doi.org/10.1016/j.jcp.2011.08.008
32. Lyu, X., Binois, M., Ludkovski, M.: Evaluating Gaussian process metamodels and sequential designs for noisy level set estimation. Statistics and computing **31**(4), 1–21 (Jul 2021). https://doi.org/10.1007/s11222-021-10014-w
33. Makarova, A., Shen, H., Perrone, V., Klein, A., Faddoul, J.B., Krause, A., Seeger, M., Archambeau, C.: Automatic termination for hyperparameter optimization. In: International Conference on Automated Machine Learning. pp. 7–1. PMLR (2022)
34. Marchini, J., Presanis, A.: Comparing methods of analyzing fmri statistical parametric maps. NeuroImage **22**(3), 1203–1213 (2004). https://doi.org/https://doi.org/10.1016/j.neuroimage.2004.03.030
35. Marques, A., Lam, R., Willcox, K.: Contour location via entropy reduction leveraging multiple information sources. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
36. Mason, B., Jain, L., Mukherjee, S., Camilleri, R., Jamieson, K.G., Nowak, R.D.: Nearly optimal algorithms for level set estimation. In: International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event. Proceedings of Machine Learning Research, vol. 151, pp. 7625–7658. PMLR (2022)
37. McDonald, G., Macdonald, C., Ounis, I.: Active learning stopping strategies for technology-assisted sensitivity review. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2053–2056. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401267
38. Nguyen, Q.P., Low, B.K.H., Jaillet, P.: An information-theoretic framework for unifying active learning problems. arXiv [cs.LG] (Dec 2020)
39. Nikakhtar, F., Ayromlou, M., Baghram, S., Rahvar, S., Rahimi Tabar, M.R., Sheth, R.K.: The excursion set approach: Stratonovich approximation and cholesky decomposition. Monthly notices of the Royal Astronomical Society **478**(4), 5296–5300 (Aug 2018). https://doi.org/10.1093/mnras/sty1415
40. Olsson, F., Tomanek, K.: An intrinsic stopping criterion for committee-based active learning. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 138–146. CoNLL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
41. Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R.T., Kim, N.H.: Adaptive designs of experiments for accurate approximation of a target region. Journal of mechanical design (New York, N.Y.: 1990) **132**(7), 071008 (Jul 2010). https://doi.org/10.1115/1.4001873

42. Qing, J., Knudde, N., Garbuglia, F., Spina, D., Couckuyt, I., Dhaene, T.: Adaptive sampling with automatic stopping for feasible region identification in engineering design. Eng. with Comput. **38**(Suppl 3), 1955–1972 (Aug 2022). https://doi.org/10.1007/s00366-021-01341-7
43. Qing, J., Knudde, N., Garbuglia, F., Spina, D., Couckuyt, I., Dhaene, T.: Adaptive sampling with automatic stopping for feasible region identification in engineering design. Engineering with Computers **38**(3), 1955–1972 (2022). https://doi.org/10.1007/s00366-021-01341-7
44. Settles, B.: Active Learning Literature Survey. Machine Learning **15**(2), 201–221 (2010). https://doi.org/10.1.1.167.4245
45. Shekhar, S., Javidi, T.: Multiscale gaussian process level set estimation. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 3283–3291. PMLR (16–18 Apr 2019)
46. Srinivas, N., Krause, A., Kakade, S., Seeger, M.: Gaussian process optimization in the bandit setting: no regret and experimental design. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. p. 1015–1022. ICML'10, Omnipress, Madison, WI, USA (2010)
47. Ueno, T., Rhone, T.D., Hou, Z., Mizoguchi, T., Tsuda, K.: Combo: An efficient bayesian optimization library for materials science. Materials discovery **4**, 18–21 (2016)
48. Vlachos, A.: A stopping criterion for active learning. Comput. Speech Lang. **22**(3), 295–312 (Jul 2008). https://doi.org/10.1016/j.csl.2007.12.001
49. Williams, C., Rasmussen, C.: Gaussian processes for regression. In: Advances in Neural Information Processing Systems. vol. 8, pp. 514–520. MIT Press (1995)
50. Wilson, J.: Stopping bayesian optimization with probabilistic regret bounds. In: Advances in Neural Information Processing Systems. vol. 37, pp. 98264–98296. Curran Associates, Inc. (2024)
51. Zanette, A., Zhang, J., Kochenderfer, M.J.: Robust super-level set estimation using gaussian processes. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18. pp. 276–291. Springer (2019)
52. Zhang, Y., Chen, X.: Sequential metamodel-based approaches to level-set estimation under heteroscedasticity. Statistical analysis and data mining **17**(3) (Jun 2024). https://doi.org/10.1002/sam.11697