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

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

A Proof of Theorem 1, Proposition 1 and Corollary 1

Let $\mathbb{1}_A(a)$ be the indicator function that returns 1 if a certain input a is included in the set A, and 0 otherwise. Let binary variables $z(\mathbf{x})$ and $w(\mathbf{x})$ as $z(\mathbf{x}) := \mathbb{1}_{\mathbb{R}^+}(\hat{f}(\mathbf{x}) - \theta)$, and $w(\mathbf{x}) := \mathbb{1}_{\mathcal{E}}(\hat{f}(\mathbf{x}) - \theta)$, respectively.

To prove the theorem 1, proposition 1 and corollary 1, we introduce the following two lemmas.

Lemma 1. *Let the candidate point set be $\mathcal{X}$, then, the following holds for $\gamma(\mathbf{x}) = \max\{p^{\min}(\mathbf{x}), \Pr(\mathbf{x} \in U_\theta)\}$ and $\eta(\mathbf{x}) = \mathbb{1}_{\mathbb{R}^+}(\Pr(\mathbf{x} \in U_\theta) - p^{\max}(\mathbf{x}))$:*

$$\Pr(\forall \mathbf{x} \in \mathcal{X}, |z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \gamma(\mathbf{x}), w(\mathbf{x}) \geq \eta(\mathbf{x})) \geq 1 - \sum_{\mathbf{x} \in X} r^{\min}(\mathbf{x}).$$

Proof. For notational simplicity, we introduce shorthand notations. $p^0(\mathbf{x}) := \Pr(\mathbf{x} \in L_\theta) = \Pr(z(\mathbf{x}) = 0)$, $p^1(\mathbf{x}) := \Pr(z(\mathbf{x}) = 1) = 1 - p^0(\mathbf{x})$ and $q^1(\mathbf{x}) := \Pr(\mathbf{x} \in U_\theta) = \Pr(w(\mathbf{x}) = 1), q^0(\mathbf{x}) := \Pr(w(\mathbf{x}) = 0) = 1 - q^1(\mathbf{x})$. Let's consider the joint distribution of $z(\mathbf{x})$ and $w(\mathbf{x})$. Specifically, let $p^{00}(\mathbf{x}) := \Phi\left(\frac{\theta - \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right), p^{01}(\mathbf{x}) := \Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) - \Phi\left(\frac{\theta - \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right), p^{11}(\mathbf{x}) := \Phi\left(\frac{\theta + \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) - \Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right), p^{10}(\mathbf{x}) := 1 - \Phi\left(\frac{\theta + \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right)$. Then, the joint distribution of $z(\mathbf{x})$ and $w(\mathbf{x})$ is given as

$$\Pr(z(\mathbf{x}), w(\mathbf{x})) = \prod_{i=0, j=0}^{1,1} (p^{ij}(\mathbf{x}))^{\delta(i - z(\mathbf{x}))\delta(j - w(\mathbf{x}))}, \tag{1}$$

where $\delta(a) = 1$ if and only if $a = 0$. This distribution considers the probability of $p(f(\mathbf{x}) \mid \mathbf{y})$ when partitioning the range of $f(\mathbf{x})$ into four regions: $f(\mathbf{x}) \leq \theta - \epsilon/2$,

$\theta - \epsilon/2 < f(\mathbf{x}) \leq \theta$, $\theta < f(\mathbf{x}) \leq \theta + \epsilon/2$, and $f(\mathbf{x}) > \theta + \epsilon/2$. Integrating the value of $p(f(\mathbf{x}) \mid \mathbf{y})$ over each region results in the probability in that region. Furthermore, considering the marginal probabilities for each region, we can show that $p^0(\mathbf{x}) = p^{00}(\mathbf{x}) + p^{01}(\mathbf{x})$ and $p^1(\mathbf{x}) = p^{10}(\mathbf{x}) + p^{11}(\mathbf{x})$. Similarly, $q^0(\mathbf{x}) = p^{00}(\mathbf{x}) + p^{10}(\mathbf{x})$ and $q^1(\mathbf{x}) = p^{01}(\mathbf{x}) + p^{11}(\mathbf{x})$ hold.

The possible values for $z(\mathbf{x})$ are only 0 or 1, and given that $\Pr(z(\mathbf{x}) = 0, w(\mathbf{x}) = 0) = p^{00}(\mathbf{x})$ and $\Pr(z(\mathbf{x}) = 1, w(\mathbf{x}) = 0) = p^{10}(\mathbf{x})$, the cumulative distribution function for $z(\mathbf{x})$ can be expressed as follows:

$$\Pr(z(\mathbf{x}) \leq b, w(\mathbf{x}) = 0) = \begin{cases} 0 & b < 0, \\ p^{00}(\mathbf{x}) & 0 \leq b < 1, \\ q^0(\mathbf{x}) & b \geq 1. \end{cases} \tag{2}$$

Similarly, we have

$$\Pr(z(\mathbf{x}) \geq a, w(\mathbf{x}) = 0) = \begin{cases} q^0(\mathbf{x}) & a \leq 0, \\ p^{10}(\mathbf{x}) & 0 < a \leq 1, \\ 0 & a > 1. \end{cases} \tag{3}$$

For $\gamma(\mathbf{x}) > 0$, using the notation $b = \mathbb{E}[z(\mathbf{x})] + \gamma(\mathbf{x})$, from $\mathbb{E}[z(\mathbf{x})] = p^1(\mathbf{x})$ we have

$$\Pr(z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})] \leq \gamma(\mathbf{x}), w(\mathbf{x}) = 0) = \begin{cases} 0 & p^1(\mathbf{x}) + \gamma(\mathbf{x}) < 0, \\ p^{00}(\mathbf{x}) & 0 \leq p^1(\mathbf{x}) + \gamma(\mathbf{x}) < 1, \\ q^0(\mathbf{x}) & p^1(\mathbf{x}) + \epsilon \geq 1, \end{cases} \tag{4}$$

$$= \begin{cases} 0 & \gamma(\mathbf{x}) < -p^1(\mathbf{x}), \\ p^{00}(\mathbf{x}) & -p^1(\mathbf{x}) \leq \gamma(\mathbf{x}) < 1 - p^1(\mathbf{x}), \\ q^0(\mathbf{x}) & \gamma(\mathbf{x}) \geq 1 - p^1(\mathbf{x}). \end{cases} \tag{5}$$

From $\gamma(\mathbf{x}) > 0$ and $p^0(\mathbf{x}) = 1 - p^1(\mathbf{x})$, we also have

$$\Pr(z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})] \leq \gamma(\mathbf{x}), w(\mathbf{x}) = 0) = \begin{cases} p^{00}(\mathbf{x}) & 0 < \gamma(\mathbf{x}) < p^0(\mathbf{x}), \\ q^0(\mathbf{x}) & \gamma(\mathbf{x}) \geq p^0(\mathbf{x}). \end{cases} \tag{6}$$

Similarly, with $a = \mathbb{E}[z(\mathbf{x})] - \gamma(\mathbf{x})$, we have

$$\Pr(z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})] \geq -\gamma(\mathbf{x}), w(\mathbf{x}) = 0) = \begin{cases} q^0(\mathbf{x}) & p^1(\mathbf{x}) - \gamma(\mathbf{x}) \leq 0, \\ p^{10}(\mathbf{x}) & 0 < p^1(\mathbf{x}) - \gamma(\mathbf{x}) \leq 1, \\ 0 & p^1(\mathbf{x}) - \gamma(\mathbf{x}) > 1 \end{cases} \tag{7}$$

$$= \begin{cases} q^0(\mathbf{x}) & \gamma(\mathbf{x}) \geq p^1(\mathbf{x}), \\ p^{10}(\mathbf{x}) & -p^0(\mathbf{x}) \leq \gamma(\mathbf{x}) < p^1(\mathbf{x}), \\ 0 & \gamma(\mathbf{x}) < -p^0(\mathbf{x}). \end{cases} \tag{8}$$

Then, from $\gamma(\mathbf{x}) > 0$,

$$\Pr(z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})] \geq -\gamma(\mathbf{x}), w(\mathbf{x}) = 0) = \begin{cases} q^0(\mathbf{x}) & \gamma(\mathbf{x}) \geq p^1(\mathbf{x}), \\ p^{10}(\mathbf{x}) & 0 < \gamma(\mathbf{x}) < p^1(\mathbf{x}), \end{cases} \tag{9}$$

We also have the following relationship:

$$\begin{aligned}
&\Pr(a \le z(\mathbf{x}) \le b, w(\mathbf{x}) = 0) + \Pr(z(\mathbf{x}) < a, w(\mathbf{x}) = 0) + \Pr(z(\mathbf{x}) > b, w(\mathbf{x}) = 0) \\
&= \Pr(w(\mathbf{x}) = 0), \qquad (10)\\
&\Pr(z(\mathbf{x}) < a, w(\mathbf{x}) = 0) + \Pr(z(\mathbf{x}) \ge a, w(\mathbf{x}) = 0) = \Pr(w(\mathbf{x}) = 0), \qquad (11)\\
&\Pr(z(\mathbf{x}) > b, w(\mathbf{x}) = 0) + \Pr(z(\mathbf{x}) \le b, w(\mathbf{x}) = 0) = \Pr(w(\mathbf{x}) = 0). \qquad (12)
\end{aligned}$$

From these equations, we have

$$\begin{aligned}
&\Pr(a \le z(\mathbf{x}) \le b, w(\mathbf{x}) = 0) \\
&= \Pr(z(\mathbf{x}) \ge a, w(\mathbf{x}) = 0) + \Pr(z(\mathbf{x}) \le b, w(\mathbf{x}) = 0) - \Pr(w(\mathbf{x}) = 0), \qquad (13)
\end{aligned}$$

and $k = \arg\max_{i \in \{0,1\}} \{p^i(\mathbf{x})\}$ leads us to

$$\begin{aligned}
&\Pr(|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \le \gamma(\mathbf{x}), w(\mathbf{x}) = 0) \\
&= \begin{cases} p^{10}(\mathbf{x}) + p^{00}(\mathbf{x}) - q^0(\mathbf{x}) & 0 < \gamma(\mathbf{x}) < p^{\min}(\mathbf{x}), \\ p^{k0}(\mathbf{x}) + q^0(\mathbf{x}) - q^0(\mathbf{x}) & p^{\min}(\mathbf{x}) \le \gamma(\mathbf{x}) < p^{\max}(\mathbf{x}), \\ q^0(\mathbf{x}) + q^0(\mathbf{x}) - q^0(\mathbf{x}) & \gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x}) \end{cases} \\
&= \begin{cases} 0 & 0 < \gamma(\mathbf{x}) < p^{\min}(\mathbf{x}), \\ p^{k0}(\mathbf{x}) & p^{\min}(\mathbf{x}) \le \gamma(\mathbf{x}) < p^{\max}(\mathbf{x}), \\ q^0(\mathbf{x}) & \gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x}). \end{cases} \qquad (14)
\end{aligned}$$

With the similar argument for $\Pr(|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \le \gamma(\mathbf{x}), w(\mathbf{x}) = 1)$, we have

$$\Pr(|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \le \gamma(\mathbf{x}), w(\mathbf{x}) = 1) = \begin{cases} 0 & 0 < \gamma < p^{\min}(\mathbf{x}), \\ p^{k1}(\mathbf{x}) & p^{\min}(\mathbf{x}) \le \gamma < p^{\max}(\mathbf{x}), \\ q^1(\mathbf{x}) & \gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x}). \end{cases} \qquad (15)$$

Then, from Eqs. (14)and (15), defining

$$\delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})) = \begin{cases} 1 & (0 < \gamma(\mathbf{x}) < p^{\min}(\mathbf{x})) \wedge (\eta(\mathbf{x}) \le 0), \\ 1 & (0 < \gamma(\mathbf{x}) < p^{\min}(\mathbf{x})) \wedge (0 < \eta(\mathbf{x}) \le 1), \\ 1 & (0 < \gamma(\mathbf{x}) < p^{\min}(\mathbf{x})) \wedge (\eta_m > 1), \\ p^{\min}(\mathbf{x}) & (p^{\min}(\mathbf{x}) \le \gamma(\mathbf{x}) < p^{\max}(\mathbf{x})) \wedge (\eta(\mathbf{x}) \le 0), \\ 1 - p^{k1}(\mathbf{x}) & (p^{\min}(\mathbf{x}) \le \gamma(\mathbf{x}) < p^{\max}(\mathbf{x})) \wedge (0 < \eta(\mathbf{x}) \le 1), \\ 1 & (p^{\min}(\mathbf{x}) \le \epsilon < p^{\max}(\mathbf{x})) \wedge (\eta(\mathbf{x}) > 1), \\ 0 & (\gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x})) \wedge (\eta(\mathbf{x}) \le 0), \\ q^0(\mathbf{x}) & (\gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x})) \wedge (0 < \eta(\mathbf{x}) \le 1), \\ 1 & (\gamma(\mathbf{x}) \ge p^{\max}(\mathbf{x})) \wedge (\eta(\mathbf{x}) > 1), \end{cases} \qquad (16)$$

we have

$$\Pr(|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \le \gamma(\mathbf{x}), w(\mathbf{x}) \ge \eta(\mathbf{x})) = 1 - \delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})).$$

Let $E(\mathbf{x})$ be the event $\{(|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \le \gamma(\mathbf{x})) \cap (w(\mathbf{x}) \ge \eta(\mathbf{x}))\}$, and its complement is denoted by $\overline{E(\mathbf{x})}$. Since $\Pr(E(\mathbf{x})) = 1 - \Pr(\overline{E(\mathbf{x})})$, we have

$$\Pr(\overline{E(\mathbf{x})}) = \delta\left(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})\right). \qquad (17)$$

By Boole's inequality, the following inequalities hold:

$$\Pr\left(\cup_{\mathbf{x}\in\mathcal{X}}\overline{E(\mathbf{x})}\right) \leq \sum_{\mathbf{x}\in\mathcal{X}} \delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})). \tag{18}$$

By De Morgan's laws, we have

$$\Pr\left(\cap_{\mathbf{x}\in\mathcal{X}} E(\mathbf{x})\right) \geq 1 - \sum_{\mathbf{x}\in\mathcal{X}} \delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})). \tag{19}$$

Therefore, the following inequality is established:

$$\Pr\left(\forall \mathbf{x} \in \mathcal{X}, |z(\mathbf{x}) - \mathbb{E}\left[z(\mathbf{x})\right]| \leq \gamma(\mathbf{x}), w(\mathbf{x}) \geq \eta(\mathbf{x})\right) \geq 1 - \sum_{m=1}^{M} \delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})).$$

Here, let $\gamma(\mathbf{x}) = \max\{p^{\min}(\mathbf{x}), q^1(\mathbf{x})\}$ and $\eta(\mathbf{x}) = \mathbb{1}_{\mathbb{R}^+}(q^1(\mathbf{x}) - p^{\max}(\mathbf{x}))$, then $\delta(\mathbf{x}, \gamma(\mathbf{x}), \eta(\mathbf{x})) = r^{\min}(\mathbf{x})$. Thus, the lemma is proved.

Lemma 2. *Let* $r^{\max}(\mathbf{x}) = \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}$. *For any event* $E(\mathbf{x}) := \{(|z(\mathbf{x}) - \mathbb{E}\left[z(\mathbf{x})\right]| \leq \gamma(\mathbf{x})) \cap (w(\mathbf{x}) \geq \eta(\mathbf{x}))\}$, *the following relationships hold:*

1. *If* $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)$, *then* $E(\mathbf{x}) = (f(\mathbf{x}) - \theta > 0)$.
2. *If* $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in L_\theta)$, *then* $E(\mathbf{x}) = (f(\mathbf{x}) - \theta \leq 0)$.
3. *If* $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in U_\theta)$, *then* $E(\mathbf{x}) = \left(-\frac{\epsilon}{2} < f(\mathbf{x}) - \theta \leq \frac{\epsilon}{2}\right)$.

Proof. We first prove that $E(\mathbf{x}) = (f(\mathbf{x}) - \theta > 0)$ holds when $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)$. Since $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)$, we can derive following inequalities:

$$\begin{aligned}
p^{\max}(\mathbf{x}) =& \Pr(\mathbf{x} \in H_\theta),\\
p^{\min}(\mathbf{x}) =& \Pr(\mathbf{x} \in L_\theta),\\
\gamma(\mathbf{x}) =& \max\{p^{\min}(\mathbf{x}), \Pr(\mathbf{x} \in U_\theta)\}\\
=& \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\} (< p^{\max}(\mathbf{x})),\\
\eta(\mathbf{x}) =& \mathbb{1}_{\mathbb{R}^+}(\Pr(\mathbf{x} \in U_\theta) - \Pr(\mathbf{x} \in H_\theta))\\
=& 0.
\end{aligned}$$

From these, the event $E(\mathbf{x})$ can be transformed as follows:

$$E(\mathbf{x}) = \{(|z(\mathbf{x}) - \mathbb{E}\left[z(\mathbf{x})\right]| \leq \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}) \cap (w(\mathbf{x}) \geq 0)\}.$$

Since $w(\mathbf{x}) \in \{0, 1\}$, $w(\mathbf{x}) \geq 0$ is always true. Hence,

$$\begin{aligned}
&E(\mathbf{x})\\
=&\{|z(\mathbf{x}) - \mathbb{E}\left[z(\mathbf{x})\right]| \leq \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}\}\\
=&\{\mathbb{E}\left[z(\mathbf{x})\right] - \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\} \leq z(\mathbf{x}) \leq \mathbb{E}\left[z(\mathbf{x})\right] + \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}\}.
\end{aligned}$$

Applying $\mathbb{E}[z(\mathbf{x})] = \Pr(\mathbf{x} \in H_\theta)$, we have

$$\begin{aligned} &E(\mathbf{x}) \\ =&\{\Pr(\mathbf{x} \in H_\theta) - \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\} \leq z(\mathbf{x}) \leq \max\{1, \Pr(\mathbf{x} \in H_\theta) + \Pr(\mathbf{x} \in U_\theta)\}\} \end{aligned}$$

Since $\max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\} < \Pr(\mathbf{x} \in H_\theta)$, the following inequalities hold.

$$\begin{aligned} \Pr(\mathbf{x} \in H_\theta) - \max\{\Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\} > 0, \\ \max\{1, \Pr(\mathbf{x} \in H_\theta) + \Pr(\mathbf{x} \in U_\theta)\} \geq 1. \end{aligned}$$

From the above, the event $E(\mathbf{x})$ is always true only when $z(\mathbf{x}) = 1$. Since $z(\mathbf{x}) := \mathbb{1}_{\mathbb{R}^+}(\hat{f}(\mathbf{x}) - \theta)$, $z(\mathbf{x}) = 1$ holds when $f(\mathbf{x}) - \theta > 0$. Therefore, when $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)$, the following relationship holds.

$$E(\mathbf{x}) = \{f(\mathbf{x}) - \theta > 0\}$$

Similarly, when $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in L_\theta)$, we have the following equations.

$$\begin{aligned} p^{\max}(\mathbf{x}) =& \Pr(\mathbf{x} \in L_\theta), \\ p^{\min}(\mathbf{x}) =& \Pr(\mathbf{x} \in H_\theta), \\ \gamma(\mathbf{x}) =& \max\{p^{\min}(\mathbf{x}), = \Pr(\mathbf{x} \in U_\theta)\}, \\ =& \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in U_\theta)\} (< p^{\max}(\mathbf{x})), \\ \eta(\mathbf{x}) =& \mathbb{1}_{\mathbb{R}^+}(\Pr(\mathbf{x} \in U_\theta) - \Pr(\mathbf{x} \in L_\theta)) \\ =& 0. \end{aligned}$$

By applying a derivation similar to the previous one, we have

$$E(\mathbf{x}) = \{f(\mathbf{x}) - \theta \leq 0\}.$$

Next, we prove that $E(\mathbf{x}) = \left(-\frac{\epsilon}{2} < f(\mathbf{x}) - \theta \leq \frac{\epsilon}{2}\right)$ when $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in U_\theta)$. From the assumption, the following equations hold.

$$\begin{aligned} \gamma(\mathbf{x}) =& \max\{p^{\min}(\mathbf{x}), \Pr(\mathbf{x} \in U_\theta)\}, \\ =& \Pr(\mathbf{x} \in U_\theta) (> p^{\max}(\mathbf{x})) \\ \eta(\mathbf{x}) =& \mathbb{1}_{\mathbb{R}^+}(\Pr(\mathbf{x} \in U_\theta) - p^{\max}(\mathbf{x})) \\ =& 1. \end{aligned}$$

By using $\mathbb{E}[z(\mathbf{x})] = \Pr(\mathbf{x} \in H_\theta)$, $|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \Pr(\mathbf{x} \in U_\theta)$ is transformed as follows:

$$\Pr(\mathbf{x} \in H_\theta) - \Pr(\mathbf{x} \in U_\theta) \leq z(\mathbf{x}) \leq \Pr(\mathbf{x} \in H_\theta) + \Pr(\mathbf{x} \in U_\theta).$$

Since $\Pr(\mathbf{x} \in U_\theta) > p^{\max}(\mathbf{x}) = \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta)\}$, the following inequalities hold.

$$\begin{aligned} \Pr(\mathbf{x} \in H_\theta) - \Pr(\mathbf{x} \in U_\theta) <& 0, \\ \Pr(\mathbf{x} \in H_\theta) + \Pr(\mathbf{x} \in U_\theta) >& 1. \end{aligned}$$

From these relationships, $|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \Pr(\mathbf{x} \in U_\theta)$ is always true. Therefore, the event $E(\mathbf{x})$ holds when $w(\mathbf{x}) = 1$. Since $w(\mathbf{x}) := \mathbb{1}_{\mathcal{E}}(\hat{f}(\mathbf{x}) - \theta)$, we can derive as follows:

$$E(\mathbf{x}) = \left\{ -\frac{\epsilon}{2} < f(\mathbf{x}) - \theta \leq \frac{\epsilon}{2} \right\}.$$

Thus, the lemma has been proven.

A.1 Proof of theorem 1

Events $f(x) - \theta > 0$, $f(x) - \theta < 0$ and $-\epsilon/2 < f(x) - \theta \leq \epsilon/2$ are denoted by $E_{H_\theta}(\mathbf{x})$, $E_{L_\theta}(\mathbf{x})$, and $E_{U_\theta}(\mathbf{x})$, respectively. From the definition of the ϵ-accuracy, the following relationship holds.

$$\begin{aligned} &\Pr((\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta) \text{ is } \epsilon\text{-accurate}) \\ &= \Pr\left(\left(\bigcap_{\mathbf{x} \in \tilde{H}_\theta} E_{H_\theta}(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{L}_\theta} E_{L_\theta}(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{U}_\theta} E_{U_\theta}(\mathbf{x}) \right) \right). \end{aligned} \tag{20}$$

On the other hand, from the lemma 1, the following equation holds.

$$p\left(\forall \mathbf{x} \in \mathcal{X}, |z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \gamma(\mathbf{x}), w(\mathbf{x}) \geq \eta(\mathbf{x})\right) \geq 1 - \sum_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x}). \tag{21}$$

Therefore, we just prove the equivalence between the left side hand of Eq.(21) and the right side hand of Eq. (20). The left side hand of Eq.21 is transformed as follows:

$$\Pr(\forall \mathbf{x} \in \mathcal{X}, |z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \gamma(\mathbf{x}), w(\mathbf{x}) \geq \eta(\mathbf{x})) = \Pr\left(\bigcap_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \right).$$

If we determine $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ using Eqs. (10), (11), (12), we can divide the $\mathcal{X}$ to $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ as follows:

$$\Pr\left(\bigcap_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \right) = \Pr\left(\left(\bigcap_{\mathbf{x} \in \tilde{H}_\theta} E(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{L}_\theta} E(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{U}_\theta} E(\mathbf{x}) \right) \right).$$

Since $(\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta)$ is determined by Eqs. (10), (11), and (12), we can apply lemma 2 to the above equation. Therefore, the following equation holds.

$$\begin{aligned} &\Pr\left(\left(\bigcap_{\mathbf{x} \in \tilde{H}_\theta} E(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{L}_\theta} E(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{U}_\theta} E(\mathbf{x}) \right) \right) \\ &= \Pr\left(\left(\bigcap_{\mathbf{x} \in \tilde{H}_\theta} E_{H_\theta}(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{L}_\theta} E_{L_\theta}(\mathbf{x}) \right) \cap \left(\bigcap_{\mathbf{x} \in \tilde{U}_\theta} E_{U_\theta}(\mathbf{x}) \right) \right) \\ &= \Pr\left((\tilde{H}_\theta, \tilde{L}_\theta, \tilde{U}_\theta) \text{ is } \epsilon\text{-accurate} \right). \end{aligned}$$

From the above, we proved that the theorem 1.

A.2 Proof of proposition 1

In this subsection, we show the lower bound of performance measures such as accuracy, recall, precision, specificity, and F-score. For the performance measures, the following relationships hold.

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

$$\begin{aligned}
\textit{Accuracy} &\geq \frac{|\tilde{H}_\theta|+|\tilde{L}_\theta|}{|\tilde{H}_\theta|+|\tilde{L}_\theta|+|\tilde{U}_\theta|},\\
\textit{Precision} &\geq \frac{|\tilde{H}_\theta|}{|\tilde{H}_\theta|+|\tilde{U}_\theta|},\\
\textit{Recall} &\geq \frac{|\tilde{H}_\theta|}{|\tilde{H}_\theta|+|\tilde{U}_\theta|},\\
\textit{Specificity} &\geq \frac{|\tilde{L}_\theta|}{|\tilde{L}_\theta|+|\tilde{U}_\theta|},\\
\textit{F-score} &\geq \frac{2|\tilde{H}_\theta|}{2|\tilde{H}_\theta|+|\tilde{U}_\theta|}.
\end{aligned}$$

Proof. From theorem 1, the following equation holds with probability $1-\sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$,

$$\Pr\left((\forall x\in\tilde{H}_\theta, f(x)-\theta>0)\cap(\forall x\in\tilde{L}_\theta, f(x)-\theta<0)\cap(\forall x\in\tilde{U}_\theta, -\epsilon/2<f(x)-\theta\leq\epsilon/2)\right).$$

We assume that candidate points are considered positive if it is included in the upper-level set and negative if it is included in the lower-level set. Then, the event $(\forall x\in\tilde{H}_\theta, f(x)-\theta>0)$ means that all candidate points predicted as positive are truly positive. Therefore, $|\tilde{H}_\theta|$ is counted as true positive (TP). Similarly, the event $(\forall x\in\tilde{L}_\theta, f(x)-\theta<0)$ means that all candidate points predicted as negative are truly negative. Therefore, $|\tilde{L}_\theta|$ is counted as true negative (TN). It is unclear whether the elements of $\tilde{U}_\theta$ are counted as TP, TN, False Positive (FP), or False Negative (FN). In the worst-case scenario, all elements of $|\tilde{U}_\theta|$ are included in either FP or FN. Then, $|\tilde{U}_\theta|$ is counted as $FP+FN$. Therefore, we can derive the following lower bound.

$$\text{F-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}=\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}\geq\frac{2|\tilde{H}_\theta|}{2|\tilde{H}_\theta|+|\tilde{U}_\theta|}.$$

Since $(\tilde{H}_\theta,\tilde{L}_\theta,\tilde{U}_\theta)$ is ϵ-accurate with probability $1-\sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$, the above lower bound also holds with probability $1-\sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$.

Similarly, for Accuracy, the following inequalities hold with $1-\sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$.

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\geq\frac{|\tilde{H}_\theta|+|\tilde{L}_\theta|}{|\tilde{H}_\theta|+|\tilde{L}_\theta|+|\tilde{U}_\theta|},$$

For the recall, we assume that all elements of $\tilde{U}_\theta$ are included in FN as the worst-case scenario. Then, the following lower bound holds with probability $1 - \sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \geq \frac{|\tilde{H}_\theta|}{|\tilde{H}_\theta| + |\tilde{U}_\theta|}.$$

Similarly, assuming that all elements of $\tilde{U}_\theta$ are included in FP as the worst-case scenario for the precision and specificity, the following inequality holds with probability $1 - \sum_{\mathbf{x}\in\mathcal{X}} r^{\min}(\mathbf{x})$,

$$\begin{aligned}\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \geq \frac{|\tilde{H}_\theta|}{|\tilde{H}_\theta| + |\tilde{U}_\theta|},\\ \text{Specificity} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \geq \frac{|\tilde{L}_\theta|}{|\tilde{L}_\theta| + |\tilde{U}_\theta|}.\end{aligned}$$

Therefore, the proposition is proved.

A.3 Proof of corollary 1

In the standard classification rule, $\mathbf{x}$ is classified into $\tilde{H}_\theta$ when $\mu_N(\mathbf{x}) - \beta\sigma_N(\mathbf{x}) > \theta$, $\mathbf{x}$ is classified into $\tilde{L}_\theta$ when $\mu_N(\mathbf{x}) + \beta\sigma_N(\mathbf{x}) < \theta$, and $\mathbf{x}$ is classified into $\tilde{U}_\theta$ when the both of conditions are not satisfied. Regarding $\Phi(\beta)$ as $\Pr(\mathbf{x} \in U_\theta)$, the standard classification rule is equivalent to the classification rule of Eqs. (10), (11), and (12) under the assumption that $\Phi(\beta) > 0.5$. Let $\tilde{r}^{\min}(\mathbf{x}) = \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), 1 - \Phi(\beta)\}$. Then, β can be converted to $\epsilon(\mathbf{x}) = g^{-1}(\Phi(\beta \mid \mathbf{x})$. In this case, the theorem 1 is reformulated as follows:

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

Therefore, the corollary 1 is proved.

B Detailed Algorithm of the Proposed Method

B.1 How to determine the margin

We have considered ϵ as a given, but in practice, ϵ is not necessarily provided and must be determined based on some criteria. However, the appropriate value for ϵ can change depending on the variance of the prior distribution and the magnitude of the noise, making it challenging to set ϵ appropriately without prior knowledge of these factors. In the following, we describe a method to adaptively determine ϵ according to the variance of the prior distribution and the size of the noise, even when no prior knowledge about ϵ is available. When stopping LSE based on Eq. (13), the average probability r^{avg} that must be ensured per candidate point

is $r^{\text{avg}} = \frac{1-\delta}{|\mathcal{X}|}$. Moreover, from $r^{\min}(\mathbf{x}) := \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \notin U_\theta)\}$, since the margin affects only $\Pr(\mathbf{x} \notin U_\theta)$, focusing solely on $\Pr(\mathbf{x} \notin U_\theta)$, for it to be $\frac{1-\delta}{|\mathcal{X}|}$, the probability that the true function is included in the margin, $\Pr(\mathbf{x} \in U_\theta)$, must be $1 - \frac{1-\delta}{|\mathcal{X}|}$. Since the most difficult points for classification are those where the surrogate model's mean function values are near the threshold, considering the worst-case scenario where $\theta = \mu(\mathbf{x})$, we have

$$\epsilon = 2\sigma_N(\mathbf{x})\Phi^{-1}\left(1 - \frac{1-\delta}{2|\mathcal{X}|}\right). \tag{22}$$

By setting ϵ such that this equation holds, ϵ can be set to make $\Pr(\mathbf{x} \in U_\theta)$ larger than $1 - \frac{1-\delta}{|\mathcal{X}|}$ when $\sigma_N(\mathbf{x})$ reaches the desired level.

Since $\sigma_N(\mathbf{x})$ is influenced by the variance of the prior distribution and the precision parameter of observation noise λ, it is difficult to predetermine the desirable $\sigma_N(\mathbf{x})$. However, the posterior variance depends solely on the input and is fixed once the input is determined. Thus, the posterior variance of $\mathbf{x}$ can be considered as an indicator of *how much data have been collected* at $\mathbf{x}$. This is an idea similar to the effective sample size (ESS) used in the context of MCMC [13] and survey sampling [27]. If the user specifies the number of measurements allowed per point, the corresponding posterior variance can be derived. This calculation is complex, but the relationship between the data count-like quantity and the posterior variance can be considered by looking at the decrease in the posterior variance $\sigma_L(\mathbf{x})$ when observing the same candidate point L times. In this situation, $\sigma_L(\mathbf{x})$ is equivalent to the posterior variance of the 1D Gaussian distribution when N samples are observed with the variance of the prior distribution as $k(\mathbf{x}, \mathbf{x})$ and the accuracy parameter of the observation noise as λ, that is, $\sigma_L^2(\mathbf{x}) = \frac{\lambda^{-1}k(\mathbf{x},\mathbf{x})}{\lambda^{-1}+Lk(\mathbf{x},\mathbf{x})}$. Finally, ϵ is estimated using the following equation.

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

If k is a non-stationary kernel, ϵ varies depending on the candidate point, whereas for a stationary kernel, it remains the same across all candidate points. L is a parameter that the user decides according to the problem; if the user wishes to reduce data acquisition costs at the expense of prediction accuracy, L should be set smaller, and if the user wants to ensure prediction accuracy even at higher data acquisition costs, L should be set larger. Empirically, for the 2D case, which is common in LSE, we confirmed that L values between 1 and 5 work effectively.

In the Appendix C.5, we demonstrate that our method for determining ϵ is less dependent on objective functions and noise variances, compared to directly setting ϵ. Furthermore, we present the differences in LSE efficiency and compare the stopping times as L, the parameter used to determine ϵ, is varied.

B.2 The pseudo code

The pseudo code of the proposed method based on the method for determining ϵ is shown in Algorightm 1.

Algorithm 1 Proposed LSE Algorithm with Stopping Criterion

Require: observed data S_N, GP prior, a set of candidate points $\mathcal{X}$, and hyperparameters $\delta > 0$ and $L \in \mathbb{N}$ (or $\varepsilon > 0$).
Initialize: $\tilde{H}_\theta, \tilde{L}_\theta \leftarrow \emptyset$, $\tilde{U}_\theta \leftarrow \mathcal{X}$, $\delta \leftarrow 0$
while $\delta < \tilde{\delta}$ **do**
Step 1 Construct the predictive mean $\mu_N(\mathbf{x})$ and standard deviation $\sigma_N(\mathbf{x})$ defined in (2) and (3). Define $r^{\max}(\mathbf{x}) := \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta)\}$.
Step 2 Classify for all $\mathbf{x} \in \mathcal{X}$ as follows:
if $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in H_\theta)$ **then**
update the upper-level set: $\tilde{H}_\theta \leftarrow \tilde{H}_\theta \cup \{\mathbf{x}\}$
else if $r^{\max}(\mathbf{x}) = \Pr(\mathbf{x} \in L_\theta)$ **then**
update the lower-level set: $\tilde{L}_\theta \leftarrow \tilde{L}_\theta \cup \{\mathbf{x}\}$
else
update the unclassified set $\tilde{U}_\theta \leftarrow \tilde{U}_\theta \cup \{\mathbf{x}\}$.
end if
Step 3 Calculate $r^{\min}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ where

- $r^{\min}(\mathbf{x}) = \min\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta), \Pr(\mathbf{x} \in U_\theta))\}$
- $\Pr(\mathbf{x} \in H_\theta) = 1 - \Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right)$
- $\Pr(\mathbf{x} \in L_\theta) = \Phi\left(\frac{\theta - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right)$
- $\Pr(\mathbf{x} \notin U_\theta) = 1 - \Phi\left(\frac{\theta + \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right) + \Phi\left(\frac{\theta - \epsilon/2 - \mu_N(\mathbf{x})}{\sigma_N(\mathbf{x})}\right)$
- $\epsilon = 2\sqrt{\frac{\lambda^{-1} k(\mathbf{x},\mathbf{x})}{\lambda^{-1} + L k(\mathbf{x},\mathbf{x})}} \Phi^{-1}\left(1 - \frac{1-\delta}{2|\mathcal{X}|}\right)$

Step 4 Select the next evaluation point by maximizing the acquisition function as $\mathbf{x}_N = \arg\max_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x})$ and observe $y_N = f(\mathbf{x}_N) + \eta$, $\eta \sim \mathcal{N}(0, \lambda^{-1})$.
Step 5 Update the dataset: $S_N \leftarrow S_N \cup \{(\mathbf{x}_N, y_N)\}$
Step 6 Calculate $\tilde{\delta} \leftarrow 1 - \sum_{\mathbf{x} \in \mathcal{X}} r^{\min}(\mathbf{x})$
end while

C Additional experimental results

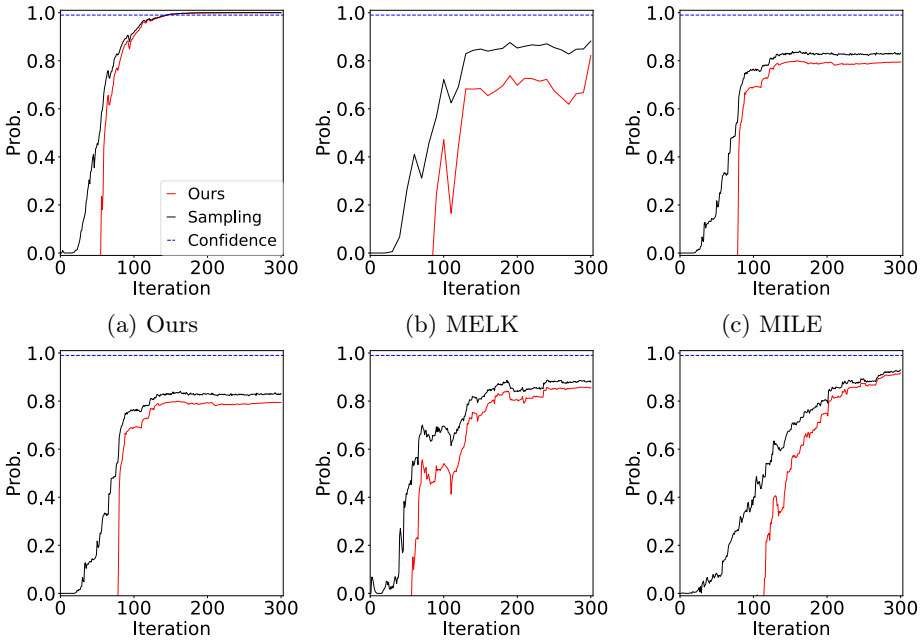
C.1 The relationship between true probability of Theorem 1 and its lower bound

From Lemma 1, the left-hand side of Eq. (14) can be calculated by evaluating the following equation,

$$\Pr(\forall \mathbf{x} \in \mathcal{X}, |z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \gamma(\mathbf{x}), w(\mathbf{x}) \geq \eta(\mathbf{x})).$$

This can be approximated by generating sample paths according to the Gaussian process posterior distribution. Specifically, by considering the generated sample paths as the true function and counting the number of times the equation is satisfied, we can compute this probability. In this subsection, we compare the probability obtained from sampling with its lower bound proposed in this study and discuss the tightness of the proposed lower bound.

In this experiment, we use two test functions, `Branin` and `Rosenbrock`, and compare the left-hand side and right-hand side of Eq. (14) for each acquisition



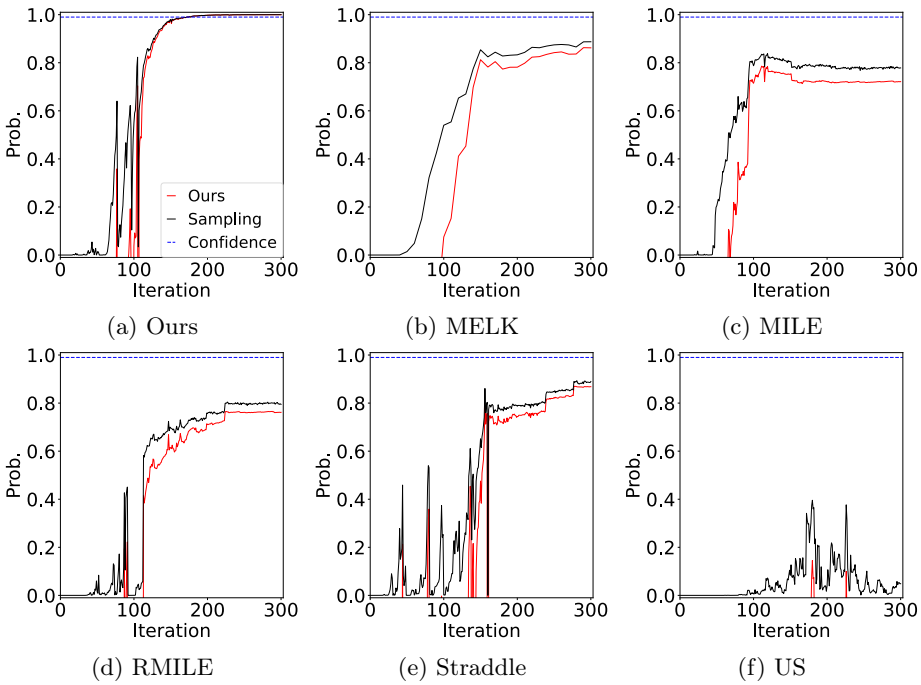
(a) Ours (b) MELK (c) MILE

(d) RMILE (e) Straddle (f) US

Fig. 1: True probability and its lower bound in `Branin` function.

function. The experimental settings are the same as the experiment in the main text, and we set the number of sample paths generated for evaluating the left-hand side to $10,000$. There are numerical considerations when calculating the left-hand side of Eq. (14). Since $z(\mathbf{x})$ is a binary variable, $\mathbb{E}[z(\mathbf{x})] = \Pr(\mathbf{x} \in H_\theta)$, meaning that $|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]|$ can only take $\Pr(\mathbf{x} \in H_\theta)$ or $\Pr(\mathbf{x} \in L_\theta)$. Therefore, for $\gamma(\mathbf{x}) = p^{\min}(\mathbf{x})$, $|z(\mathbf{x}) - \mathbb{E}[z(\mathbf{x})]| \leq \gamma(\mathbf{x})$ only holds when the equality holds. Although it is correct to use Eq. (14) theoretically, calculating whether equality holds numerically can lead to numerical errors. To avoid these errors, we modify the actual calculation by setting $\gamma(\mathbf{x}) = \max\{(p^{\min}(\mathbf{x}) + p^{\max}(\mathbf{x}))/2, \Pr(\mathbf{x} \in U_\theta)\}$.

The experimental results are shown in Figs. 1 and 2. The results indicate that when the value of the left-hand side is small, the lower bound becomes loose, but as the left-hand side approaches 1, the lower bound becomes tighter. Notably, when the lower bound reaches the confidence parameter $\delta = 0.99$ used in this study, the bound is considerably tight, and there is no difference in the stopping timing whether it is computed approximately using sampling or using the lower bound. This is because the lower bound becomes tighter as the probabilities of each candidate point become sufficiently high, given that no bounds other than the uniform bound are used to calculate the lower bound.



(a) Ours (b) MELK (c) MILE

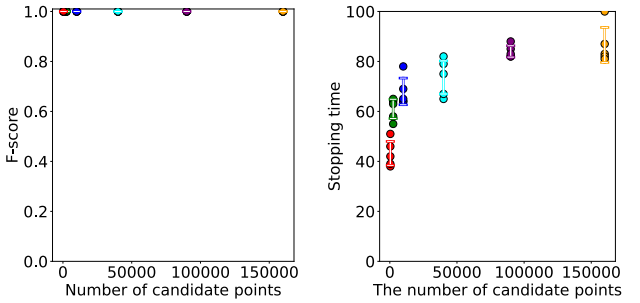
(d) RMILE (e) Straddle (f) US

Fig. 2: True probability and its lower bound in `Rosenbrock` function.

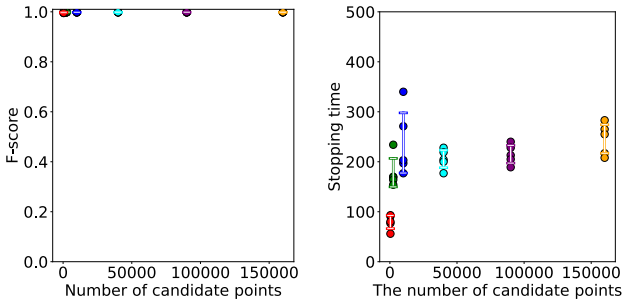
Next, we compare the left-hand side of Eq. (14) when using the proposed acquisition function versus other acquisition functions. In this case, the left-hand side of Eq. (14) converges to 1 for the proposed acquisition function, while it does not for other acquisition functions. Especially, in the case of US for `Rosenbrock`, even if LSE progresses, the true probability remains small. Therefore, its lower bound takes a negative value, resulting in a trivial lower bound. This is because other acquisition functions do not search further to improve the classification accuracy of candidate points once it exceeds a certain threshold, whereas the proposed acquisition function continues to search to increase the probability of ϵ-accuracy, even for candidate points that have already achieved a certain level of classification accuracy. Actually, we show that the proposed stopping criterion cannot stop LSE when we use the other acquisition functions in Appendix C.4. Therefore, it can be said that the combination of the proposed acquisition function and stopping criterion realizes efficient LSE.

C.2 The effect of the number of candidate points on the stopping timing

Since the proposed stopping criterion uses Boolean inequalities, it is possible that the stopping timing may be delayed as the number of candidate points increases. In this experiment, we evaluate the impact of increasing the number of candidate



(a) F1 score($\sigma_{\text{noise}} = 0$) (b) Stopping time($\sigma_{\text{noise}} = 0$)



(c) F1 score($\sigma_{\text{noise}} = 10$) (b) Stopping time($\sigma_{\text{noise}} = 10$)

Fig. 3: True probability and its lower bound in `Rosenbrock` function.

points on the stopping timing of LSE and demonstrate that the proposed method can still stop even as the number of candidates increases. For this experiment, the `rosenbrock` function is used as the test function, and the LSE threshold is set to $\theta = 100$. Additionally, two cases are evaluated: one without observation noise and the other with Gaussian noise having a variance of $\sigma^2_{\text{noise}} = 10^2$.

The experimental results are shown in Fig. 3. From these results, it can be observed that the F-score of the stopping timing shows almost no difference even as the number of candidate points increases, and the increase in stopping timing is slight with respect to the increase in candidate points. This is because the proposed stopping criterion does not rely on any lower bounds other than Boole's inequalities. Specifically, the probability that each individual candidate point is ϵ-accurate can be calculated precisely. When the probability that each candidate point is ϵ-accurate is sufficiently high, the lower bound using Boolean inequalities also becomes tight. Therefore, it can be concluded that the proposed stopping criterion is slight affected by the increase in the number of candidate points.

C.3 Other test functions

In this subsection, we demonstrate that the proposed method can stop LSE when applied to various test functions. We also show results without adding

observation noise, illustrating that the standard stopping criterion (fully classified: FC) can stop LSE in noise-free scenarios, but fails when noise is present. In this experiment, we use the `Sphere` function, `Rosenbrock` function, and `Cross in tray` function, as well as the `Booth` function, `Branin` function, and `Holder table` function. The thresholds for these test functions are set as follows: $\theta = 20$ for the `Sphere` function, $\theta = 100$ for the `Rosenbrock` function, $\theta = -1.5$ for the `Cross in tray` function, $\theta = 500$ for the `Booth` function, $\theta = 100$ for the `Branin` function, and $\theta = -3$ for the `Holder table` function. When adding observation noise, the noise levels are set to $\sigma_{\text{noise}} = 2$ for the `Sphere` function, $\sigma_{\text{noise}} = 30$ for the `Rosenbrock` function, $\sigma_{\text{noise}} = 0.01$ for the `Cross in tray` function, $\sigma_{\text{noise}} = 30$ for the `Booth` function, $\sigma_{\text{noise}} = 20$ for the `Branin` function, and $\sigma_{\text{noise}} = 0.3$ for the `Holder table` function. Other experimental settings remain the same as in the main text.

Figure 4 shows the F-scores for the iteration (number of evaluation) when different acquisition function is used for exploration. From these results, it can be seen that there are no significant differences between the acquisition functions, except for the baseline method US. Additionally, comparing the cases with and without noise, while the convergence point of the F-score changes due to noise, the efficiency of the acquisition functions is not significantly affected by the presence of noise. Regarding the timing of each stop criterion, all stopping criteria can stop LSE in the noise-free case. On the other hand, in the presence of noise, similar to the experimental results in the main text, the FC criterion fails to stop LSE, whereas both the FS criterion and the proposed criterion can stop it. While the FS criterion tends to stop LSE aggressively, there are cases, such as in Fig. 4(k), where it fails to do so. In contrast, the proposed criterion stops LSE more conservatively but is able to stop LSE even in cases like Fig. 4(k).

C.4 Applicability of the proposed stopping criterion to other acquisition functions

The proposed stopping criterion terminates LSE when the classification probability $p^{\max}(\mathbf{x}) = \max\{\Pr(\mathbf{x} \in H_\theta), \Pr(\mathbf{x} \in L_\theta)\}$ of all candidate points becomes large enough, or the probability that the true function is included in $\mathcal{E}$ is large enough. Since this stopping criterion does not assume anything about the acquisition function, it can be, in principle, applied to other acquisition functions as well. In this subsection, we verify whether LSE can be stopped using the proposed stopping criterion with other acquisition functions. The test functions, noise levels, and other experimental settings are the same as in the Appendix C.3.

Figure 5 shows the stopping times for each acquisition function. The results indicate that, in the absence of noise, the proposed stopping criterion can stop LSE at the timing when the F-score converges, similar to the standard stopping criterion (FC). However, when observation noise is added, most acquisition functions, except the proposed one, fail to stop LSE. This is because the proposed stopping criterion is conservative. Specifically, in this experiment, the threshold for the stopping criterion is set to $\delta = 0.99$, meaning that the average classifica-

tion probability for each candidate point needs to be $1-(1-\delta)/|\mathcal{X}| = 0.999975$ for LSE to stop.

In contrast, typical LSE classifies a candidate point to the upper level set when $\mu_N(\mathbf{x}) - \beta\sigma_N(\mathbf{x}) > \theta$, which is equivalent to classifying the point when $\Pr(\mathbf{x} \in H_\theta) = \Phi((\mu_N(\mathbf{x}) - \theta)/\sigma_N(\mathbf{x})) > \Phi(\beta)$. Similarly, for the lower level set, a point is classified when $\Pr(\mathbf{x} \in L_\theta) = \Phi((\theta - \mu_N(\mathbf{x}))/\sigma_N(\mathbf{x})) > \Phi(\beta)$. In this experiment, $\beta = 1.96$, so a candidate point is classified into the upper or lower level set when its classification probability exceeds $\Phi(\beta) \approx 0.975$.

Typically, once a candidate point is classified into the upper or lower level set, it no longer needs to be explored, making it difficult to meet the stopping criterion requirements, even with repeated exploration. Therefore, it is hard for acquisition functions other than the proposed method, which prioritizes exploring the candidate points with the lowest classification probability, to stop LSE using the proposed stopping criterion.

C.5 The robustness of our margin-setting method

In this section, we demonstrate that the method for determining ϵ proposed in the Appendix B.1 is less dependent on noise variance and the range of the objective function than directly setting ϵ. To illustrate this, we compare the stopping timings using five different ϵ values $\{2, 5, 10, 30, 50\}$ with the proposed method. The test functions used for this evaluation are the `rosenbrock` and `sphere` functions, and the experiments assess the impact of varying Gaussian noise variance in five patterns for each function. For the `rosenbrock` function, the noise variances are set to $\{0^2, 10^2, 20^2, 30^2, 40^2\}$, and for the `sphere` function, the variances are set to $\{0^2, 0.5^2, 1.0^2, 1.5^2, 2.0^2\}$. In the experiments, the proposed method for determining ϵ sets $L = 3$, and all other parameters follow the settings used for the experiment of the test functions.

The experimental results for the `rosenbrock` function are shown in Figs. 6(a) and (b). From Fig. 6(a), we can observe that, regardless of the method for determining ϵ, no reduction in F-scores occurs, and the LSE does not stop until the F-scores are close enough to their limit, even when noise is varied. However, from Fig. 6(b), we see that when $\epsilon = 2$, $\epsilon = 5$, and $\epsilon = 10$, the LSE can stop before using the entire budget in the noise-free case, but as the noise increases, it fails to stop even when the entire budget is used. Furthermore, for $\epsilon = 30$ and $\epsilon = 50$, the LSE stops earlier than the proposed method when the noise variance is up to 20^2 and 30^2, respectively, but as the noise increases further, the LSE fails to stop even after using the entire budget. Next, in the results for the `sphere` function shown in Figs. 6(c) and (d), we find that, unlike in the `rosenbrock` function, where $\epsilon = 30$ and $\epsilon = 50$ successfully stopped the LSE, they stop prematurely before the F-scores converge in the `sphere` function. In contrast, $\epsilon = 5$, which failed to stop the LSE in the `rosenbrock` function, successfully stops the LSE in the `sphere` function. This indicates that the appropriate ϵ must be chosen depending on the target function and the level of noise when directly setting ϵ.

On the other hand, the proposed method can stop the LSE with relatively consistent accuracy across different functions and noise levels, despite using the same parameter settings. Thus, it has been demonstrated that the method for determining ϵ proposed in Appendix B.1 is less dependent on data noise or the range of the objective function than directly setting ϵ.

In conclusion, if an acceptable classification error ϵ is given, we use that value; otherwise, we can use L as an easy-to-use knob.

C.6 Effects of parameter L

Next, we evaluate the impact of changing the parameter L of the proposed acquisition function and stopping criterion using test functions. In this experiment, experimental settings are the same as the Appendix C.3. Figure 7 shows the F-scores and the stopping timing when L is varied as $L = 1, 2, 3, 4, 5$. Although there are rare cases where the increase of F-score is lower with some parameters, there are no significant differences in the F-scores when L is varied, regardless of the test function used or whether observation noise is added. Thus, it can be said that the parameter L does not significantly affect the acquisition function.

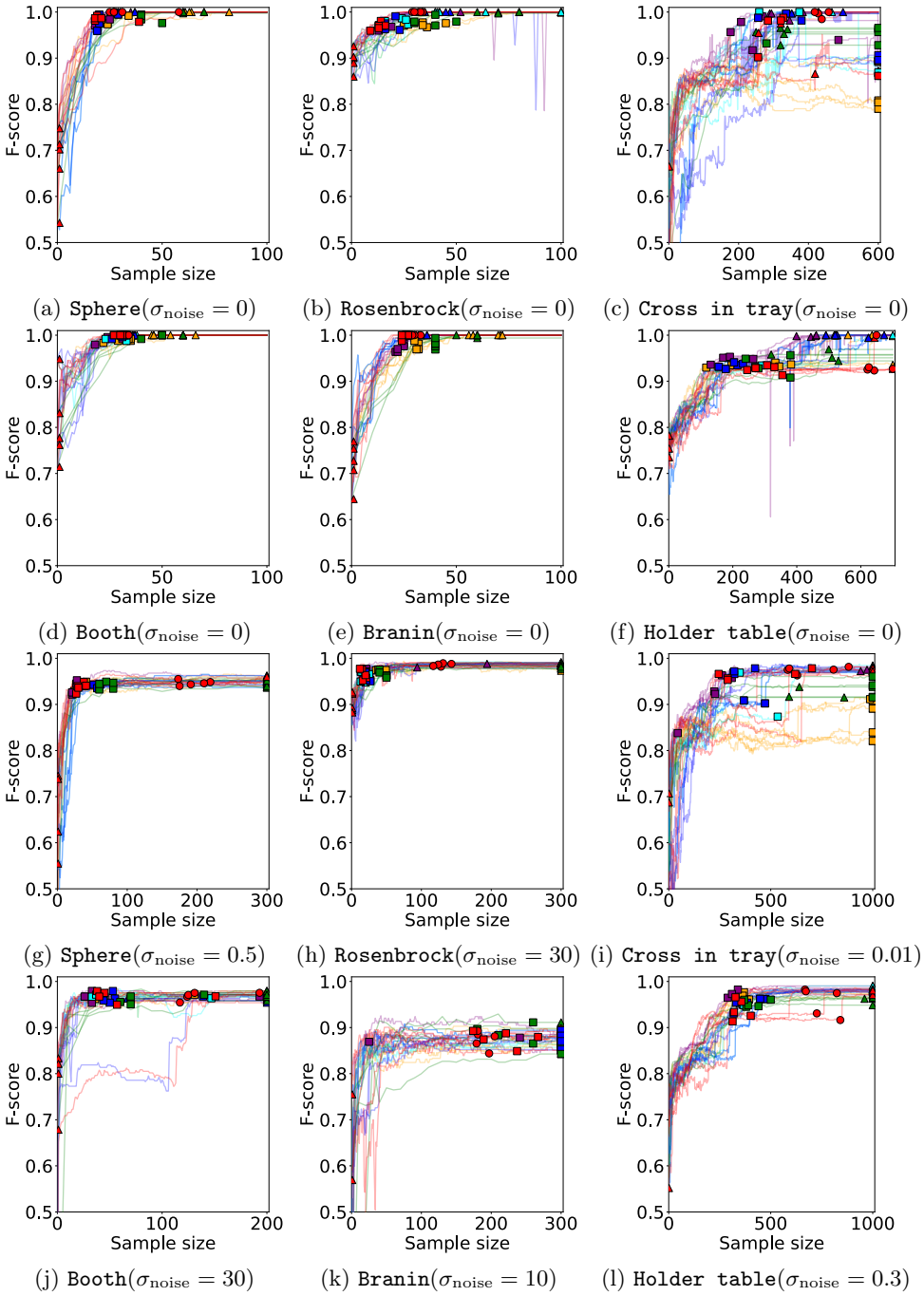
On the other hand, L has a large impact on the stopping timing. Specifically, for test functions other than the `Cross in tray` function and `Holder table` function, the impact of L is small without noise, but with noise, smaller L values lead to earlier stopping. Therefore, in most cases, setting $L = 1$ is sufficient if early stopping is desired even at the expense of accuracy, while typically setting L around 3 is adequate. However, for the `Cross in tray` and `Holder table` functions, LSE stops before the F-score converges adequately when L is small. This occurs because these functions have complex shapes, requiring appropriate hyperparameter estimation by the GP, which in turn requires a sufficient amount of data. With limited data, there is a tendency to estimate a larger kernel width. As a result, in situations with insufficient data, the GP may converge to a function different from the true function, meeting the stopping criterion and stopping LSE prematurely. In such cases, it is necessary to increase L to prevent LSE from stopping until appropriate hyperparameters are estimated.

C.7 Experiments with different threshold values

In this subsection, we evaluate the effect of changing the threshold on the stopping timing for four test functions: `Sphere`, `Branin`, `Rosenbrock`, and `Booth`. We examine three different threshold settings for each test function. For the `Sphere` function, the thresholds are set to $\theta = 10, 20, 30$; for the `Branin` function, $\theta = 50, 75, 100$; for the `Rosenbrock` function, $\theta = 100, 300, 500$; and for the `Booth` function, $\theta = 100, 300, 500$. All other experimental settings are the same as those described in Appendix C.3, except for the threshold.

The results are shown in Fig.8. Consistent with the main text, it is observed that the FC criterion fails to stop, the FS criterion tends to stop LSE aggressively, and the proposed criterion stops LSE conservatively. However, as seen in Figs.8(c), (e), (d), and (f), there are cases where the FS criterion fails to stop

LSE, and as shown in Figs. 8(h) and (i), there are cases where it stops before the F-score exceeds the threshold. In contrast, the proposed criterion consistently stops LSE even in such cases.



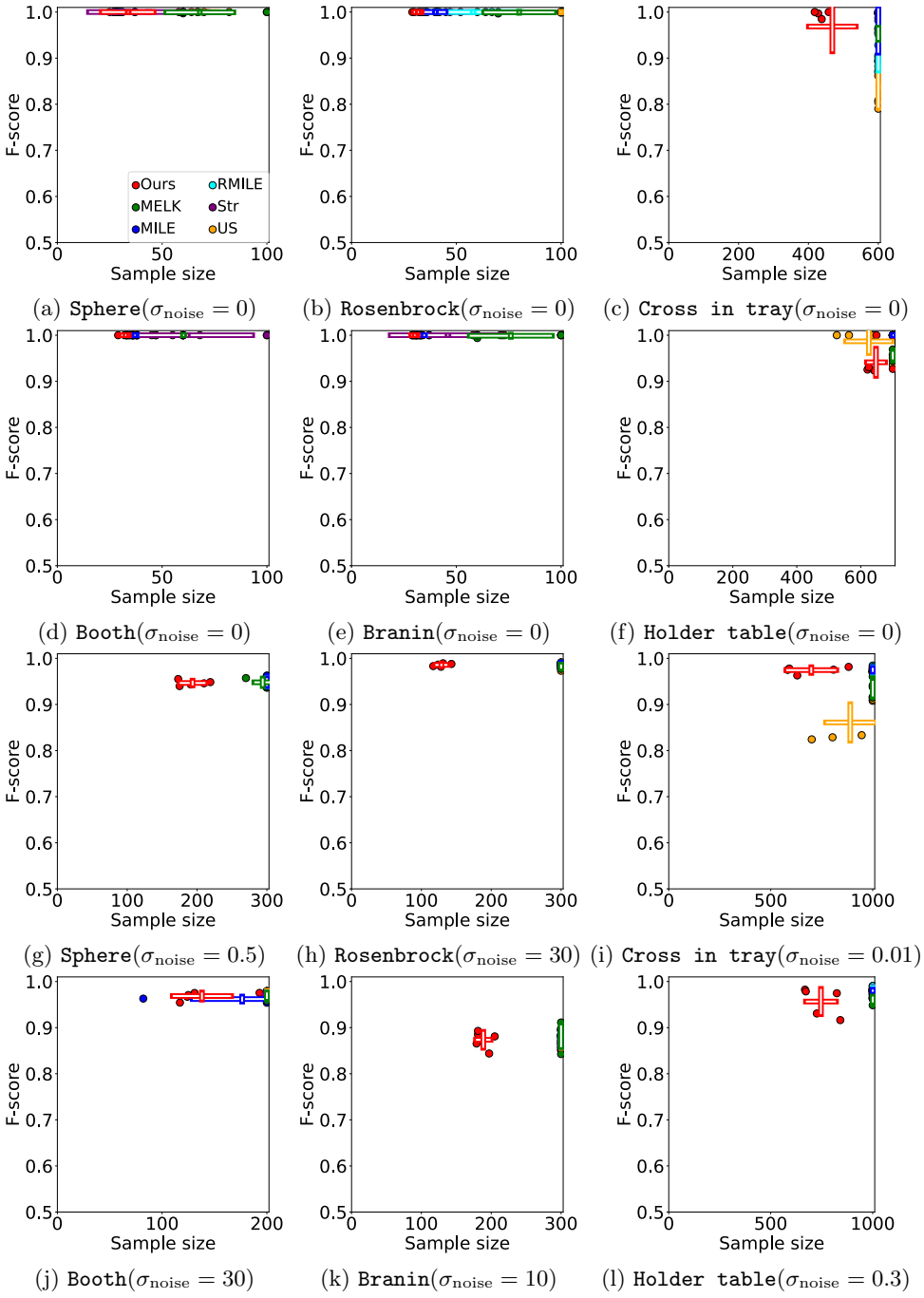
(a) `Sphere`($\sigma_{\mathrm{noise}} = 0$) (b) `Rosenbrock`($\sigma_{\mathrm{noise}} = 0$) (c) `Cross in tray`($\sigma_{\mathrm{noise}} = 0$)

(d) `Booth`($\sigma_{\mathrm{noise}} = 0$) (e) `Branin`($\sigma_{\mathrm{noise}} = 0$) (f) `Holder table`($\sigma_{\mathrm{noise}} = 0$)

(g) `Sphere`($\sigma_{\mathrm{noise}} = 0.5$) (h) `Rosenbrock`($\sigma_{\mathrm{noise}} = 30$) (i) `Cross in tray`($\sigma_{\mathrm{noise}} = 0.01$)

(j) `Booth`($\sigma_{\mathrm{noise}} = 30$) (k) `Branin`($\sigma_{\mathrm{noise}} = 10$) (l) `Holder table`($\sigma_{\mathrm{noise}} = 0.3$)

Fig. 4: F-score and stopped time using each acquisition function with proposed (Our) and fully classified (FC) criteria. (a) – (f) noise-free case. (g) – (l) noise addition case. Error bars mean standard deviation. Each label has the same meaning as in Fig. 2



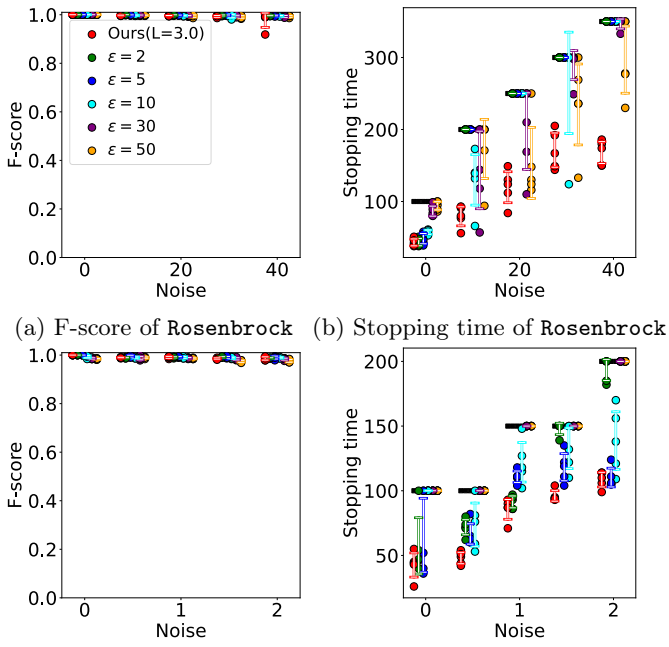
(a) `Sphere`($\sigma_{\text{noise}} = 0$) (b) `Rosenbrock`($\sigma_{\text{noise}} = 0$) (c) `Cross in tray`($\sigma_{\text{noise}} = 0$)

(d) `Booth`($\sigma_{\text{noise}} = 0$) (e) `Branin`($\sigma_{\text{noise}} = 0$) (f) `Holder table`($\sigma_{\text{noise}} = 0$)

(g) `Sphere`($\sigma_{\text{noise}} = 0.5$) (h) `Rosenbrock`($\sigma_{\text{noise}} = 30$) (i) `Cross in tray`($\sigma_{\text{noise}} = 0.01$)

(j) `Booth`($\sigma_{\text{noise}} = 30$) (k) `Branin`($\sigma_{\text{noise}} = 10$) (l) `Holder table`($\sigma_{\text{noise}} = 0.3$)

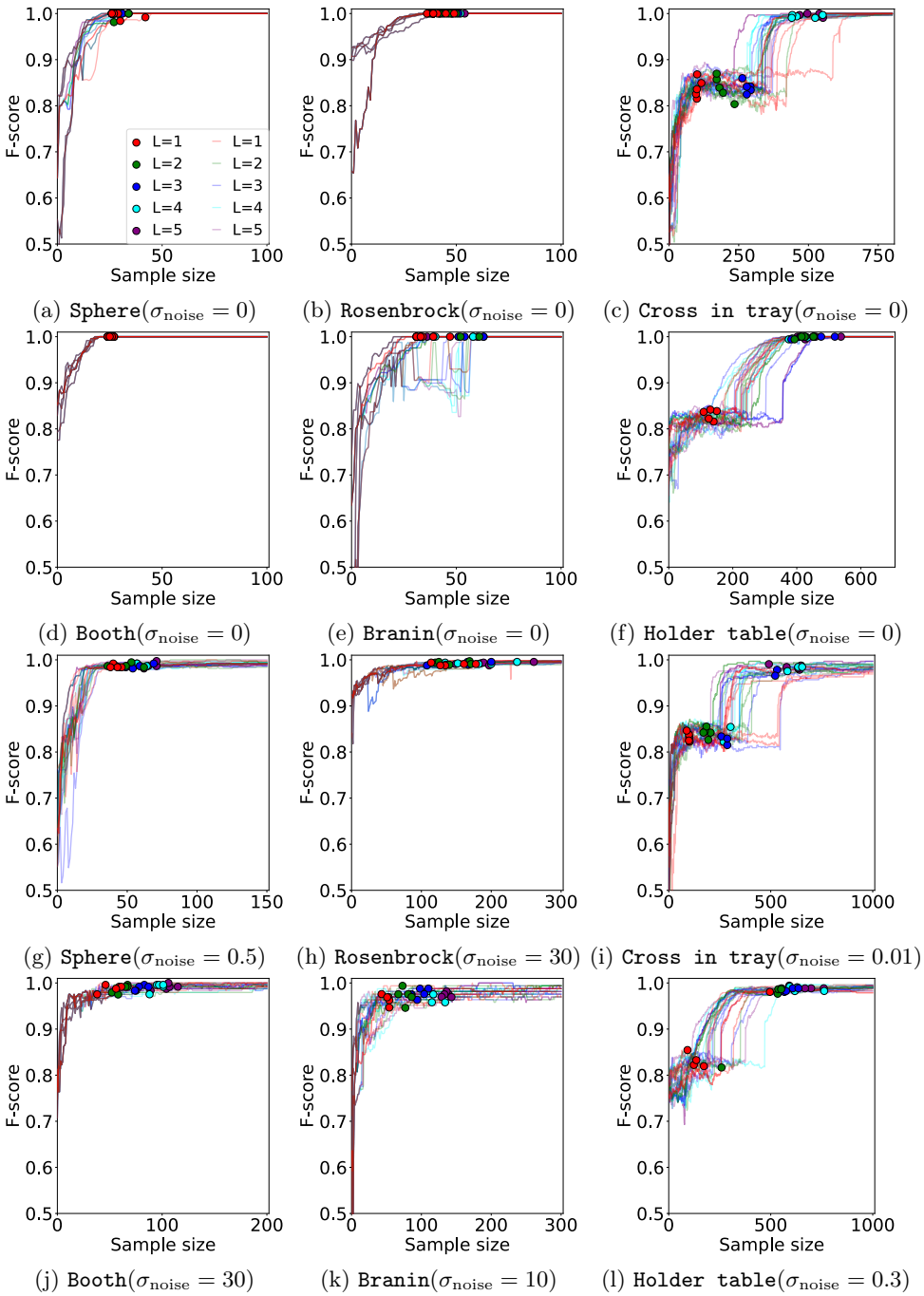
Fig. 5: Stopped time using each acquisition function with the proposed stopping criterion. (a) – (f) noise-free case. (g) – (l) noise addition case. Error bars mean standard deviation. Error bars mean standard deviation.



(a) F-score of `Rosenbrock` (b) Stopping time of `Rosenbrock`

(c) F-score of `Sphere` (d) Stopping time of `Sphere`

Fig. 6: The impact of noise variance and the range of function in the case of the proposed method and the method for setting ϵ directly. The black line of (b) and (d) means budget. To make it easier to understand, a jitter is added to the x-coordinate to prevent overlapping of the drawings.



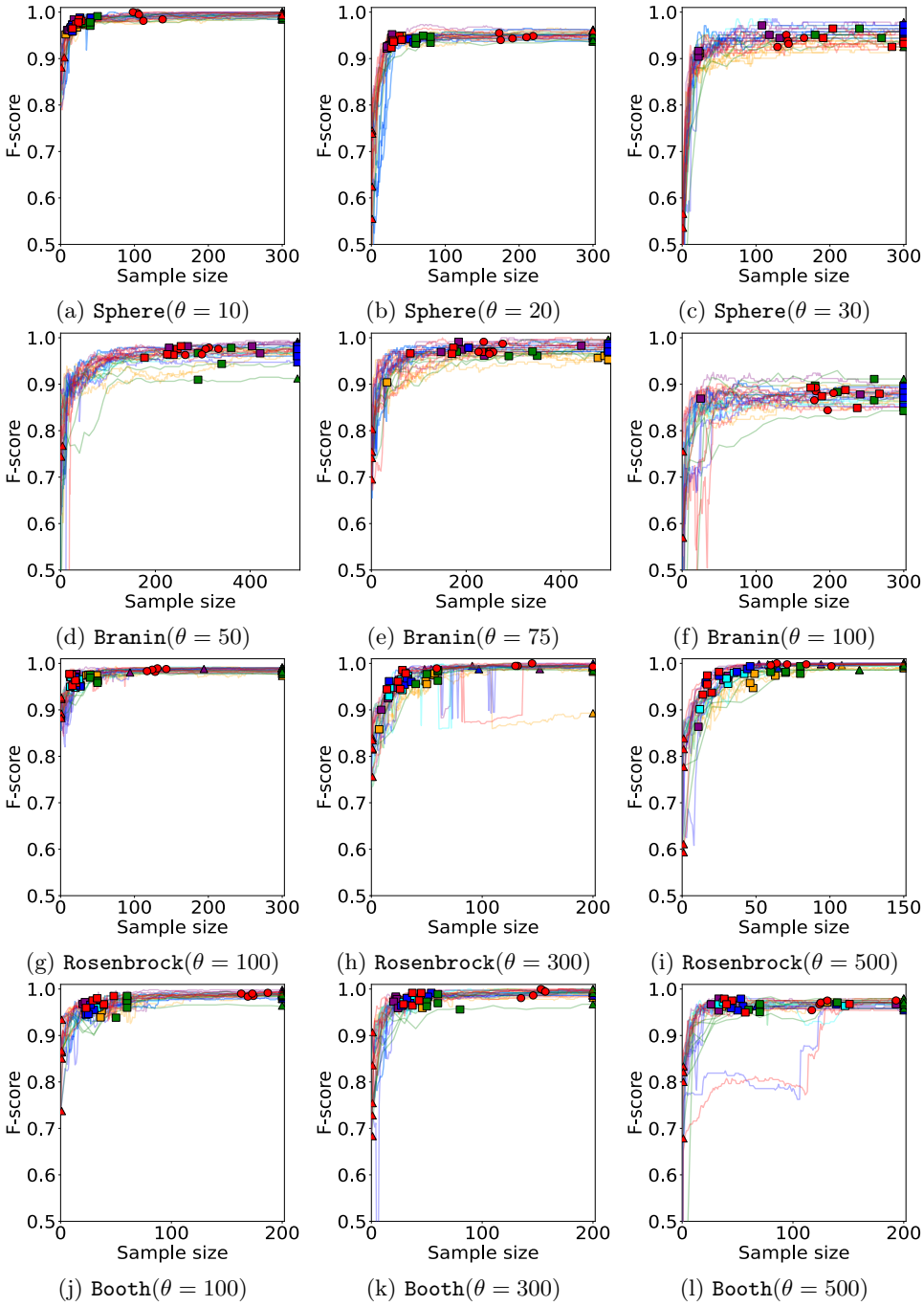
(a) `Sphere`($\sigma_{\text{noise}} = 0$) (b) `Rosenbrock`($\sigma_{\text{noise}} = 0$) (c) `Cross in tray`($\sigma_{\text{noise}} = 0$)

(d) `Booth`($\sigma_{\text{noise}} = 0$) (e) `Branin`($\sigma_{\text{noise}} = 0$) (f) `Holder table`($\sigma_{\text{noise}} = 0$)

(g) `Sphere`($\sigma_{\text{noise}} = 0.5$) (h) `Rosenbrock`($\sigma_{\text{noise}} = 30$) (i) `Cross in tray`($\sigma_{\text{noise}} = 0.01$)

(j) `Booth`($\sigma_{\text{noise}} = 30$) (k) `Branin`($\sigma_{\text{noise}} = 10$) (l) `Holder table`($\sigma_{\text{noise}} = 0.3$)

Fig. 7: F-scores and Stopped time using each acquisition function. (a) – (f) noise-free case. (g) – (l) noise addition case.



(a) `Sphere`($\theta = 10$) (b) `Sphere`($\theta = 20$) (c) `Sphere`($\theta = 30$)

(d) `Branin`($\theta = 50$) (e) `Branin`($\theta = 75$) (f) `Branin`($\theta = 100$)

(g) `Rosenbrock`($\theta = 100$) (h) `Rosenbrock`($\theta = 300$) (i) `Rosenbrock`($\theta = 500$)

(j) `Booth`($\theta = 100$) (k) `Booth`($\theta = 300$) (l) `Booth`($\theta = 500$)

Fig. 8: F-scores using each acquisition function and stopped timings with the proposed (Ours) and fully classified (FC) criteria for datasets with different thresholds.