Appendix A. Supplements for Lemmas

Lemma 2 (Generalized version of Lemma 7 of Nguyen et al. (2017)) *The sum of the predictive variances is bounded by the maximum information gain γ_T. That is for $\forall x \in \mathcal{X}$, it holds that*

$$\sum_{t=1}^{T} \sigma_{t-1}^2(x) \leq \frac{2}{\log(1+\upsilon_{\max}^{-2})}\gamma_T \tag{34}$$

where $\upsilon_{\max} = \max(\upsilon_1, \cdots, \upsilon_T)$ is the maximum standard deviation of the additive Gaussian observation noise.

Proof Let us define $G(x) = \frac{x}{\log(1+x)}$, we notice that $G(x)$ is monotonically increasing when $x \geq 0$ with its minimum value $G(0) = 0$. Utilizing this property, we have that $\frac{s}{\log(1+s)} \leq \frac{\upsilon_t^{-2}}{\log(1+\upsilon_t^{-2})}$ for $s \in [0, \upsilon_t^{-2}]$ and $s = \upsilon_t^{-2}\sigma_{t-1}^2(x) \leq \upsilon_t^{-2}$ since $\sigma_{t-1}^2(x) \leq k(x,x) \leq 1$. Define $\upsilon_{\max} = \max(\upsilon_1, \cdots, \upsilon_T)$ as the maximum standard deviation of the additive Gaussian observation noise, then we can derive that

$$\begin{aligned}
\forall x \in \mathcal{X}, \sum_{t=1}^{T} \sigma_{t-1}^2(x) &= \sum_{t=1}^{T} \upsilon_t^2 \underbrace{\upsilon_t^{-2}\sigma_{t-1}^2(x)}_{s} \leq \sum_{t=1}^{T} \upsilon_t^2 \left(\frac{\upsilon_t^{-2}\log(1+s)}{\log(1+\upsilon_t^{-2})}\right) \\
&= \sum_{t=1}^{T} \frac{\log(1+\upsilon_t^{-2}\sigma_{t-1}^2(x))}{\log(1+\upsilon_t^{-2})} \leq \frac{2}{\log(1+\upsilon_{\max}^{-2})}\frac{1}{2}\sum_{t=1}^{T}\log(1+\upsilon_t^{-2}\sigma_{t-1}^2(x)) \\
&\leq \frac{2}{\log(1+\upsilon_{\max}^{-2})}\gamma_T
\end{aligned} \tag{35}$$

where the last inequality is led by the definition of γ_T. ■

Lemma 5 *Let $\delta \in (0,1)$. For $x \in \mathcal{X}, t \in \mathcal{N}$, set $I_t^C(x) = \max\{0, f(x) - f(x_t^+)\}$, then with probability at least $1 - 2\delta$ we have*

$$\alpha_t^C(x) \geq \max\{I_t^C(x) - \sqrt{\beta_t}\left(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^+)\right), 0\}. \tag{36}$$

Proof If $\tilde{\sigma}_{t-1}(x) = 0$, then we have $\alpha_t^C(x) = I_t^C(x) = 0$. We now assume $\tilde{\sigma}_{t-1}(x) > 0$. Set $q = \frac{f(x)-f(x_t^+)}{\tilde{\sigma}_{t-1}(x)}$ and $\tilde{z} = \frac{\mu_{t-1}(x)-\mu_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}$, then we have $\tilde{z} - q = \frac{f(x_t^+)-\mu_{t-1}(x_t^+)-(f(x)-\mu_{t-1}(x))}{\tilde{\sigma}_{t-1}(x)}$. By Lemma 1, we have that $|\tilde{z} - q| \leq \frac{\sigma_{t-1}(x)+\sigma_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}\sqrt{\beta_t}$ holds with probability $1 - 2\delta$. Denote $m(x) = \frac{\sigma_{t-1}(x)+\sigma_{t-1}(x_t^+)}{\tilde{\sigma}_{t-1}(x)}$, thus $q - m(x)\sqrt{\beta_t} \leq \tilde{z}$. If $I_t^C(x) = 0$, then the lower bound is trivial as $\alpha_t^C(x)$ is non-negative. Thus suppose $I_t^C(x) > 0$. Set $\tau(z) = z\Phi(z) + \phi(z)$, since $\tau(z)$ is non-decreasing for all z, we have that

$$\begin{aligned}
\alpha_t^C(x) \geq \tilde{\sigma}_{t-1}(x)\tau(q - m(x)\sqrt{\beta_t}) \geq \tilde{\sigma}_{t-1}(x)(q - m(x)\sqrt{\beta_t}) \quad &\text{by } \tau(z) \geq z \\
= I_t^C(x) - \sqrt{\beta_t}\left(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^+)\right). &
\end{aligned} \tag{37}$$

■

Lemma 6 *Let $\delta \in (0,1)$. Then with a probability of at least $1-2\delta$, we have*

$$f(x^*) - f(x_t^+) \leq \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right) + \tilde{\sigma}_{t-1}(x_t)\tau(z_{t-1}(x_t)). \tag{38}$$

Proof By Lemma 5 and $I_t^M(x) = \max\{0, f(x) - f(x_t^+)\}$, we have that

$$f(x^*) - f(x_t^+) \leq I_t^M(x^*) \leq \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right) + \alpha_t^C(x^*) \tag{39}$$

where the second inequality is provided by Lemma 5. By the definition of $x_t = \arg\max_{x \in \mathcal{X}} \alpha_t^M(x)$, we obtain

$$\alpha_t^C(x^*) \leq \alpha_t^C(x_t) = \tilde{\sigma}_{t-1}(x_t)\tau(z_{t-1}(x_t)) \tag{40}$$

Thus, we derive the following result by combining the above two inequalities

$$f(x^*) - f(x_t^+) \leq \sqrt{\beta_t}\left(\sigma_{t-1}(x^*) + \sigma_{t-1}(x_t^+)\right) + \tilde{\sigma}_{t-1}(x_t)\tau(z_{t-1}(x_t)). \tag{41}$$

This final inequality holds with probability $1-2\delta$. ■

Appendix B. Additional simulation results: Functions sampled from Gaussian kernel

In this part, we constructed our functions based on the samples drawn from a squared exponential kernel with length scale $\ell = 3$ and amplitude $\sigma = 1$. As indicated by Figure 5 (b), the covariances between neighbor samples are not all zero given a relatively large length scale parameter. We created 30 sample sets $S_1, \cdots, S_{30}$ of 4000 data points from this kernel function as shown in Figure 5(a). For each sample set S_i, the function f_i is defined as $f_i(x) = f(x_j) + \varepsilon$ where $x_j = \arg\min_{x_j \in S_i} \|x - x_j\|$ and the observation noise ε is Gaussian distributed with mean 0 and standard deviation $\upsilon = 0.16$. We deployed BO with EI or Corrected EI to optimize these functions and the kernel was set to be the same as the sampled kernel. The performance of the acquisition function is evaluated through $f(x_t)$ corresponding to the same κ that equals 1 percent of the maximum difference over five samples. A two-sided Wilcoxon sign rank test is performed to test the null hypothesis that the corrected EI is not different from EI under our noisy settings. The test gave a p-value equal to 0.013, indicating we should reject our null hypothesis. The scatter plot as shown in Figure 6 also indicates this fact.

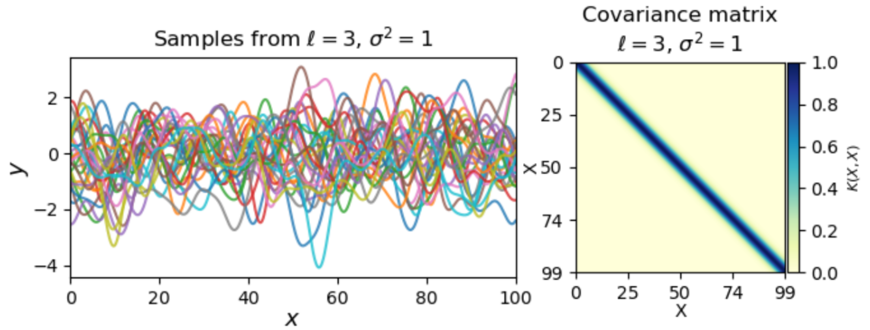


Figure 5: (Left) Samples from an exponential quadratic kernel with a length scale $\ell = 3$ and an amplitude $\sigma = 1$. Each line is formed by 4000 samples. (Right) Visual representation of the kernel matrix for the samples. The diagonal indicates the variances of the noise terms and the blue oblique region implies a strong correlation between the neighbor points.

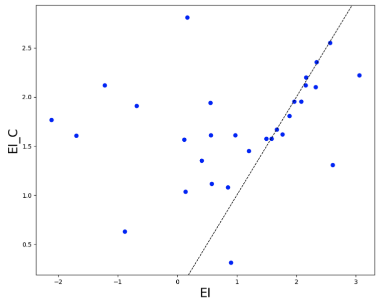
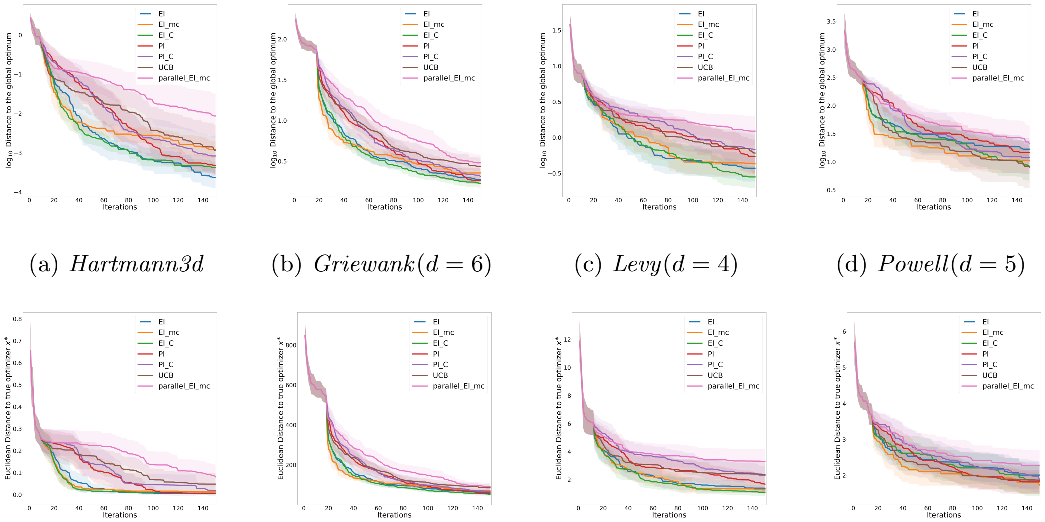


Figure 6: Comparison of EI and corrected EI over $f(x_t)$ for same termination criterion κ. The dotted line represents the function $y = x$. More points are shown to be above this line, indicating that corrected EI is more likely to select the point that returns a higher value on the objective function than standard EI.

Appendix C. Supplements for benchmark results

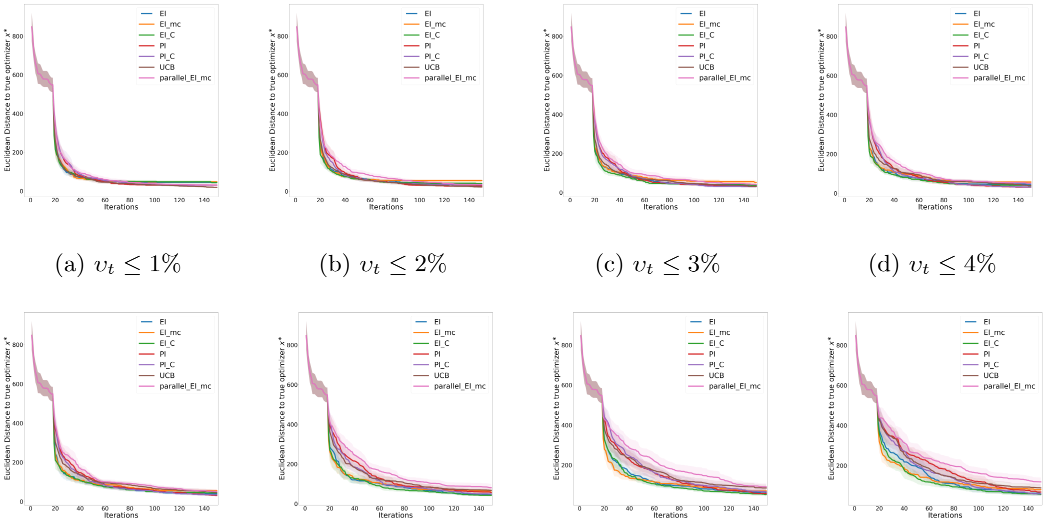
We present additional results in Figure 7 and Figure 8. Figure 7 shows the optimization performance when the noise standard deviation υ_t is less than or equal to 15% of the range of the objective function. Notably, our proposed method shows competitive performance compared to other acquisition functions. Figure 8 shows the sequential optimization performance of our proposed method relative to other acquisition functions under increasing noise levels. We observe that all methods experience a decline in performance as the noise level increases, however, the corrected EI exhibits excellent performance relative to EI even in the high-noise regime.



(a) *Hartmann3d* (b) *Griewank*$(d = 6)$ (c) *Levy*$(d = 4)$ (d) *Powell*$(d = 5)$

(e) *Hartmann3d* (f) *Griewank*$(d = 6)$ (g) *Levy*$(d = 4)$ (h) *Powell*$(d = 5)$

Figure 7: Comparison of methods for Benchmark objective functions under the case that observation noise standard deviation υ_t is less than or equal to 15% of the range of the objective function. Figures (a)∼(d) show how the mean and 95% confidence bound (shaded region) of the distance between the best feasible objective and the global optimum changes with each iteration of optimization. Figures (e)∼(f) visualize the variation of the L_2 distance between the best point and the global optimizer x^*.



(a) $\upsilon_t \leq 1\%$ (b) $\upsilon_t \leq 2\%$ (c) $\upsilon_t \leq 3\%$ (d) $\upsilon_t \leq 4\%$

(e) $\upsilon_t \leq 5\%$ (f) $\upsilon_t \leq 10\%$ (g) $\upsilon_t \leq 15\%$ (h) $\upsilon_t \leq 20\%$

Figure 8: Optimization performance under increasing noise levels υ_t on a *Griewank*$(d = 6)$ function. We define the noise level as a percentage of the range of the objective function and evaluate performance by measuring the L_2 distance between the best point and the global optimizer x^*.