Objective Bound Conditional Gaussian Process for Bayesian Optimization

Taewon Jeong [1] **Heeyoung Kim** [1]

Abstract

A Gaussian process is a standard surrogate model for an unknown objective function in Bayesian optimization. In this paper, we propose a new surrogate model, called the objective bound conditional Gaussian process (OBCGP), to condition a Gaussian process on a bound on the optimal function value. The bound is obtained and updated as the best observed value during the sequential optimization procedure. Unlike the standard Gaussian process, the OBCGP explicitly incorporates the existence of a point that improves the best known bound. We treat the location of such a point as a model parameter and estimate it jointly with other parameters by maximizing the likelihood using variational inference. Within the standard Bayesian optimization framework, the OBCGP can be combined with various acquisition functions to select the next query point. In particular, we derive cumulative regret bounds for the OBCGP combined with the upper confidence bound acquisition algorithm. Furthermore, the OBCGP can inherently incorporate a new type of prior knowledge, i.e., the bounds on the optimum, if it is available. The incorporation of this type of prior knowledge into a surrogate model has not been studied previously. We demonstrate the effectiveness of the OBCGP through its application to Bayesian optimization tasks, such as the sequential design of experiments and hyperparameter optimization in neural networks.

1. Introduction

Bayesian optimization (BO) (Snoek et al., 2012) is a widely used technique for maximizing or minimizing black-box objective functions. It is typically used to find the global optimum of a nonconvex function, whose derivative information is unavailable. BO is particularly useful when the objective function is expensive to evaluate. In this case, grid or random search (e.g., simulated annealing and tabu search) (Bertsimas et al., 1993; Glover and Laguna, 2013) can be highly inefficient. BO has been applied to various problems, including experimental design and hyperparameter tuning. BO incorporates a prior distribution over the objective function and updates the prior with sequentially acquired data to form a posterior distribution that better approximates the objective function. The posterior distribution is subsequently used to construct an acquisition function. By maximizing the acquisition function, we can determine the next query point and evaluate the objective function. Then, the posterior distribution is updated again according to the augmented data with the new point. BO iterates these steps. It commonly uses a Gaussian process (GP) for the prior over the objective function. GPs provide a flexible and analytically tractable family of prior distributions, working as a surrogate model that captures the behavior of the unknown objective function. The acquisition function used to determine the next query point quantifies the utility of candidate points and balances the trade-off between exploitation (local search) and exploration (global search). Various acquisition functions have been proposed to maximize the probability of gaining information during the sequential procedure, including the expected improvement (EI) (Jones et al., 1998), probability of improvement (Brochu et al., 2010), upper confidence bound (UCB) (Srinivas et al., 2009), entropy search (Hennig and Schuler, 2012), predictive entropy search (Hernández-Lobato et al., 2014), output-space predictive entropy search (Hoffman and Ghahramani, 2015), and max-value entropy search (Wang and Jegelka, 2017).

In this paper, we propose a new surrogate model, called the objective bound conditional Gaussian process (OBCGP), to explicitly condition a GP on a bound on the optimum. The bound is initially obtained as the best observed function value from the initial experiment using a space-filling design, and then it is updated as the current best function value during the sequential optimization procedure. The OBCGP is developed based on the idea that explicit incorporation of the knowledge of the existence of a better point than the current best point helps to infer the objective function more accurately. The location of such a point, which improves the best known bound, is included as a model parameter in

[1] Department of Industrial and Systems Engineering, KAIST, Daejeon, Republic of Korea. Correspondence to: Heeyoung Kim <heeyoungkim@kaist.ac.kr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

the OBCGP. More specifically, we introduce a candidate optimum location, x_M, which improves the best known bound, and the corresponding function value, $f(x_M)$. For example, if we assume a maximization problem, x_M satisfies $f(x_M) > l_p$, where l_p is the best known lower bound. Here, x_M and $f(x_M)$ are not yet observed but treated as a model parameter and a latent variable, respectively, by employing a conditional GP, given x_M and $f(x_M)$. We estimate the unknown parameters of the OBCGP, including x_M, by maximizing the likelihood using variational inference (Blei et al., 2017).

The OBCGP can be combined with various acquisition functions to select the next query point. In particular, we derive the cumulative regret bounds for the OBCGP combined with the UCB acquisition algorithm. Moreover, the OBCGP can inherently incorporate a new type of prior knowledge, i.e., the bounds on the optimal function values, if it is available. The incorporation of this type of prior knowledge into a surrogate model has not been studied previously, although other types of prior knowledge, such as smoothness, periodicity, convexity, and monotonicity of objective functions, have been incorporated in certain studies (Osborne et al., 2009; Jauch and Peña, 2016). In practice, prior knowledge of the bounds on the optimal function values is often available from past experiments, engineering knowledge, and expert knowledge (Kim et al., 2017). Such prior knowledge can provide useful information on the objective function. Without the incorporation of this prior knowledge, the surrogate model obtained using only experimental data may not conform to the prior knowledge.

2. Methodology

2.1. Review of a Gaussian Process

A stochastic process, $f : \Omega \subseteq R^d \rightarrow R$, is called a GP if, for any finite subset of Ω, $X_n = \{x_1, \ldots, x_n\} \subset \Omega$, $f(x_1), \ldots, f(x_n)$ have a multivariate Gaussian distribution. A GP is specified by a mean function, $\mu(x)$, and a covariance or kernel function, $k(x, x')$, where $x, x' \in \Omega$: $f(x) \sim GP(\mu(x), k(x, x'))$, where $\mu(x)$ defines the mean of $f(x)$, and $k(x, x')$ specifies the covariance between $f(x)$ and $f(x')$. In this paper, we assume a stationary kernel for the GP, i.e., $k(x, x') = k(\|x - x'\|)$, and without loss of generality, $\mu(x) = 0$. In BO, the GP typically provides a surrogate model for $f(x)$ by computing the posterior distribution over the objective function.

Suppose that we are given a collection of data points, $D_n = \{(x_i, f(x_i))\}_{i=1}^n$. Let $\mathbf{f}_n = (f(x_1), \ldots, f(x_n))^T$. Assuming a prior for f as a GP, we can write the posterior distribution of $f(x^*)$ as $f(x^*)|D_n \sim N(\mu_{GP}(x^*; D_n), \sigma^2_{GP}(x^*; D_n))$, where $\mu_{GP}(x^*; D_n) = \mathbf{k}^T K^{-1} \mathbf{f}_n$ and $\sigma^2_{GP}(x^*; D_n) = k(x^*, x^*) - \mathbf{k}^T K^{-1} \mathbf{k}$, where $\mathbf{k} = (k(x_1, x^*), \ldots, k(x_n, x^*))^T$ is an n-dimensional vector of covariances, and K is an $n \times n$ matrix having element $k(x_i, x_j)$ in the ith row and jth column.

2.2. Objective Bound Conditional Gaussian Process

We propose a new surrogate model, the OBCGP, to incorporate a bound on the optimal function value. Suppose that we consider a maximization problem. In this study, we assume noiseless observations; the extension to the noisy case is discussed in Section 7. As mentioned in Section 1, x_M is a candidate optimum location whose function value, $f(x_M)$, is greater than a lower bound, l_p, which could be the maximum among the observed function values or a previously known bound. As the iterations proceed, l_p becomes tighter, and the surrogate model is expected to incorporate a more useful bound. Our introduced x_M and $f(x_M)$ are similar to an inducing input and an inducing variable, respectively, employed for efficient computation in sparse GPs (Titsias, 2009). Inducing variables, treated as latent variables, are unobserved function values evaluated at auxiliary pseudo inputs (inducing points), treated as model parameters. However, unlike in previous studies, we set $f(x_M)$ to have a specific support according to a bound of the optimal function value. Consequently, the function values at the inducing points do not follow a Gaussian distribution. This is a different point of the OBCGP from that of the standard GP. The OBCGP is constructed by setting a conditional GP, given x_M and $f(x_M)$. This setting emanates from a simple idea: if we assign a function value ($f(x_M)$) at a specific location (x_M) to be in a tight range (according to a bound, rather than to $[-\infty, \infty]$ in the standard GP) and restrict the objective function to interpolate the point $(x_M, f(x_M))$, the shape of the objective function must be constrained accordingly.

More specifically, we incorporate x_M and $f(x_M)$ into the OBCGP as a model parameter and a latent variable, respectively, by using a conditional GP model of $\mathbf{f}_n$ given x_M and $f(x_M)$ as follows:

$$p(\mathbf{f}_n | x_M, f(x_M)) = N(\mathbf{f}_n | \boldsymbol{\mu}_n, \Sigma_{n \times n}), \tag{1}$$

where

$$\boldsymbol{\mu}_n = f(x_M) k(x_M, x_M)^{-1} \mathbf{k}_M \tag{2}$$

and

$$\Sigma_{n \times n} = K - \mathbf{k}_M k^{-1}(x_M, x_M) \mathbf{k}_M^T, \tag{3}$$

where $\mathbf{k}_M = (k(x_1, x_M), \ldots, k(x_n, x_M))^T$ is an n-dimensional vector.

For the OBCGP in Eq.(1), the inducing point, x_M, is a model parameter, not a latent variable, and it plays a role in connecting $\mathbf{f}_n$ and $f(x_M)$. By estimating the model parameters, including x_M, we fit our surrogate model.

In Eq.(1), we assign a probability distribution to $f(x_M)$ with an appropriate support according to the bound on the optimal function value. Depending on the bound type (upper or lower), we assign different probability distributions. Let x_{opt} and $f(x_{opt})$ denote the true optimum location and the corresponding optimal function value, respectively. Additionally, let l_p and u_p denote a lower bound and an upper bound on the optimal function value, respectively.

Case 1. Incorporating a lower bound: $f(x_{opt}) \geq l_p$

As described earlier, we define x_M as a candidate optimum location where the objective function has a value greater than or equal to the lower bound, i.e., $f(x_M) \geq l_p$, where l_p is obtained as the largest observed value during the sequential optimization procedure or from prior knowledge. In Eq.(2), we assign a probability distribution to $f(x_M)$ with support $[l_p, \infty]$ as follows:

$$f(x_M) = l_p + Z_M, \quad Z_M \sim \text{exponential}(\lambda), \tag{4}$$

where the prior distribution for Z_M is an exponential distribution with hyperparameter λ. Here, we assign an exponential distribution for computational tractability in inference procedures. However, other distributions can also be used if the support, $[l_p, \infty]$, is satisfied. Note that the lower bound, l_p, can be updated during the sequential optimization process. If $f(x_{best}) \geq l_p$, where x_{best} is the best location among the already evaluated locations (i.e., $x_{best} = \arg\max_{x \in X_n} f(x)$, where $X_n = \{x_i\}_{i=1}^n$ is the set of evaluated locations), we replace the lower bound, l_p, with $l_p = f(x_{best})$.

Case 2. Incorporating an upper bound: $f(x_{opt}) \leq u_p$

Because we assume a maximization problem, an upper bound, u_p, cannot be obtained from observations but is assumed to be available from prior knowledge. Case 2 involves a situation where both the lower and the upper bounds for the optimal objective value are available because the optimal objective value should be in range $[f(x_{best}), u_p]$ as we collect data. Thus, Case 2 deals with the knowledge, $l_p = f(x_{best}) \leq f(x_{opt}) \leq u_p$. In Eq.(2), we assign a probability distribution to $f(x_M)$ with support $[l_p, u_p]$ as follows:

$$f(x_M) = l_p + (u_p - l_p) Z_M, \quad Z_M \sim \text{beta}(1, \lambda), \tag{5}$$

where the prior distribution for Z_M is a beta distribution with hyperparameter λ. Here, we choose a beta distribution with support $[l_p, u_p]$ for computational tractability in inference procedures. However, other distributions can be used if the support, $[l_p, u_p]$, is satisfied.

3. Inference

For both Case 1 and Case 2 (with the OBCGP), the marginal distribution, $p(\mathbf{f}_n) = \int p(\mathbf{f}_n | Z_M) p(Z_M)\ dZ_M$, is not tractable for computation, because the distribution of Z_M is not Gaussian. Thus, instead of optimizing the likelihood directly, we apply variation inference by introducing a variational distribution, $q(Z_M)$, which approximates the posterior distribution with variational parameters (Blei et al., 2017).

3.1. Inference for Case 1

Let $\theta = \{\psi, x_M\}$ denote a set of model parameters to be estimated, where ψ denotes a set of kernel parameters. To apply variational inference, we take a variational distribution as a gamma, $q_\phi(Z_M) = \text{gamma}(\alpha, \beta)$, where $\phi = \{\alpha, \beta\}$ is the set of variational parameters to be estimated. A gamma distribution provides various shapes of a density function depending on the parameters. Therefore, it can approximate the posterior distribution over Z_M flexibly. Using this variational distribution, we can express the variational lower bound on $\log p_\theta(\mathbf{f}_n)$ as follows:

$$\begin{aligned} &\log p_\theta(\mathbf{f}_n; X_n) \geq L(\theta, \phi; D_n) \\ &= E_{q_\phi(Z_M)}[\log p_\theta(\mathbf{f}_n | Z_M)] - KL(q_\phi(Z_M) || p(Z_M)), \end{aligned}$$

where $KL(q_\phi(Z_M) || p(Z_M))$ is the Kullback–Leibler (KL) divergence, which can be computed analytically as

$$\begin{aligned} &KL(q_\phi(Z_M) || p(Z_M)) \\ &= (\alpha - 1)\gamma(\alpha) - \log \Gamma(\alpha) + \log \lambda - \log \beta + \frac{\beta - \lambda}{\lambda}, \end{aligned}$$

where $\Gamma(\cdot)$ and $\gamma(\cdot)$ are the gamma and digamma functions, respectively. The first term of $L(\theta, \phi; D_n)$ can be computed using a closed-form expression by utilizing Eqs.(1), (2), and (3), as follows:

$$\begin{aligned} &E_{q_\phi(Z_M)}[\log p_\theta(\mathbf{f}_n | Z_M)] \\ &\propto -\frac{1}{2} \log |\Sigma_{n \times n}| - \frac{1}{2} \mathbf{f}_n^T \Sigma_{n \times n}^{-1} \mathbf{f}_n \\ &- \frac{(l_p^2 + 2 l_p E_1^q + E_2^q)}{2} \frac{k(x_M, \boldsymbol{X}_n)}{k(x_M, x_M)} \Sigma_{n \times n}^{-1} \frac{k(\boldsymbol{X}_n, x_M)}{k(x_M, x_M)} \\ &+ (l_p + E_1^q) \frac{k(x_M, \boldsymbol{X}_n)}{k(x_M, x_M)} \Sigma_{n \times n}^{-1} \mathbf{f}_n, \end{aligned}$$

where E_1^q and E_2^q are the first and second moments of $q_\phi(Z_M)$, respectively. By maximizing $L(\theta, \phi; D_n)$ with respect to both θ and ϕ, we can estimate the parameters. For both Case 1 and Case 2, we maximize $L(\theta, \phi; D_n)$ using the adaptive moment estimation (Adam) (Kingma and Ba, 2015), which is a popular updating method based on stochastic gradient descent.

The computation of the posterior distribution, $p(f(x^*)|D_n)$, for a new location, x^*, is analytically intractable. However, if we focus only on the posterior mean and variance of $\mathbf{f}_n$, we can approximate them analytically. For $p(f(x^*)|D_n, Z_M)$, we have the following:

$$f(x^*)|D_n, Z_M \sim N(\hat{\mu}(x^*; D_n, Z_M), \hat{\sigma}^2(x^*; D_n, Z_M)),$$

where $\hat{\mu}(x^*; D_n, Z_M) = \mathcal{A} + \tau Z_M$ and $\hat{\sigma}^2(x^*; D_n, Z_M) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, where $\mathcal{A} = l_p \frac{k(x^*, x_M)}{k(x_M, x_M)} - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{f}_n - l_p \frac{k(X_n, x_M)}{k(x_M, x_M)})$, $\tau = \frac{k(x^*, x_M) - \Sigma_{12}\Sigma_{22}^{-1} k(X_n, x_M)}{k(x_M, x_M)}$ and $\Sigma([x^*, X_n], [x^*, X_n]) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then, we can easily compute the posterior mean and variance of $f(x^*)$, given D_n, as follows:

$$\begin{aligned} \tilde{\mu}(x^*; D_n) &= E[f(x^*)|D_n] = E[E[f(x^*)|D_n, Z_M]|D_n] \\ &\approx \mathcal{A} + \tau E(Z_M|D_n) = \mathcal{A} + \tau E_1^q, \\ \tilde{\sigma}^2(x^*; D_n) &= V[f(x^*)|D_n] \\ &= E[f(x^*)^2|D_n] - (E[f(x^*)|D_n])^2 \\ &\approx \hat{\sigma}^2(x^*; D_n, Z_M)) + \tau^2(E_2^q - (E_1^q)^2). \end{aligned}$$

3.2. Inference for Case 2

In Case 2, we consider beta(α, β) as a variational distribution over Z_M. A beta distribution provides various shapes of a density function depending on the parameters. Therefore, it can approximate $p(Z_M|D_n)$ flexibly. Similar to Case 1, we can approximate the posterior mean and variance of $f(x^*)$, given D_n, as follows:

$$\begin{aligned} \tilde{\mu}(x^*; D_n) &= E[f(x^*)|D_n] \approx \mathcal{A} + \tau E_1^q, \\ \tilde{\sigma}^2(x^*; D_n) &= Var[f(x^*)|D_n] \\ &\approx \hat{\sigma}^2(x^*; D_n, Z_M)) + \tau^2(E_2^q - (E_1^q)^2), \end{aligned}$$

where $\mathcal{A} = l_p \frac{k(x^*, x_M)}{k(x_M, x_M)} - \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{f}_n - l_p \frac{k(X_n, x_M)}{k(x_M, x_M)})$ and $\tau = (u_p - l_p)\frac{k(x^*, x_M) - \Sigma_{12}\Sigma_{22}^{-1} k(X_n, x_M)}{k(x_M, x_M)}$, and $\hat{\sigma}^2(x^*; D_n, Z_M)$ is defined as in Case 1.

4. Acquisition Functions for the OBCGP

As discussed in Section 3, we can obtain the approximated posterior mean and variance of the OBCGP at unobserved points. Based on these, we can approximate the posterior distribution of the OBCGP as a Gaussian distribution using moment matching (Hoffman and Ghahramani, 2015). More specifically, $p(f(x^*)|D_n) \approx N(\tilde{\mu}(x^*; D_n), \tilde{\sigma}^2(x^*; D_n))$, where $\tilde{\mu}(x^*; D_n)$ and $\tilde{\sigma}^2(x^*; D_n)$ are the posterior mean and variance of $f(x^*)$, respectively, which were derived in Sections 3.1 and 3.2. Applying this approximation, we can readily use popular acquisition functions with the OBCGP. For example, the EI and UCB are formulated in terms of the GP posterior mean and variance, which reflect exploitation and exploration, respectively. Therefore, if we use the Gaussian approximation for the OBCGP posterior based on moment matching, the acquisition functions can be easily evaluated by simply replacing the GP posterior moments with the OBCGP posterior moments.

Alternatively, we can evaluate the acquisition functions using Monte Carlo samples from the OBCGP posterior without the normal approximation. Let us consider an example of the EI, which is defined as $EI(x) = E[(f(x) - f_{best})^+|D_n]$, where $a^+ = \max(a, 0)$. Note that, given D_n and $f(x_M)$, $f(x)$ follows a normal distribution according to the definition of the OBCGP in Eq.(1).Thus, if we rewrite $EI(x) = E[E[(f(x) - f_{best})^+|D_n, f(x_M)]|D_n]$, the inner expectation term can be computed in the same way as for the GP. The outer expectation term denotes the expectation over $f(x_M)$, given D_n, which can be approximated using Monte Carlo samples from the OBCGP posterior. In summary, the EI for the OBCGP can be approximated using the posterior samples as follows:

$$EI(x) = \frac{1}{N}\sum_{i=1}^{N} E[(f(x) - f_{best})^+|D_n, f^i(x_M)],$$

where $f^i(x_M)$ are the samples from $p(f(x_M)|D_n)$. For sampling, we use the estimated posterior distribution with variational inference. For example, in Case 1, $f^i(x_M) = l_p + Z_M^i$, where Z_M^i is a sample from $q_\phi(Z_M)$.

Evaluating the acquisition functions according to the moment-matching normal approximation is computationally more efficient than the Monte Carlo sampling method without the normal approximation. Moreover, we found that the difference in the BO performance of the two approaches was negligible in our experiments. In Section 6, we report the results of BO with the OBCGP based on the moment-matching approximation. The results based on the sampling method are also provided in the Supplementary Materials, along with more details regarding the inference for the OBCGP and the derivation of the closed-form expression of the EI with the OBCGP.

5. Regret Bounds

After estimating the objective function using the OBCGP, we select the next query point using an acquisition function. Various acquisition functions can be used with the OBCGP. In particular, we analyze the case of the OBCGP combined with the UCB (OBCGP-UCB). Using the OBCGP-UCB, we choose the next query point (the ith query point) as

$$x_i = \text{argmax}_{x \in \Omega} \tilde{\mu}_{i-1}(x) + \beta_i^{1/2}\hat{\sigma}_{i-1}(x), \qquad (6)$$

where $\tilde{\mu}_{i-1}(x) = \tilde{\mu}(x; D_{i-1})$, $\hat{\sigma}_{i-1}(x) = \sqrt{\hat{\sigma}^2(x; D_{i-1}, Z_M)}$, and β_i are the appropriate constants. We derive the cumulative regret bounds of the OBCGP-UCB in Theorem 1 in the same manner as in Srinivas et al. (2009). The cumulative regret (R_n) is defined as the sum of instantaneous regrets: $R_n = \sum_{i=1}^{n} |f(x_{opt}) - f(x_i)|$.

Theorem 1. *Suppose* $|\Omega| < \infty$. *Let* $\delta \in (0,1)$ *and* $\tilde{\beta}_i = 2\log(\frac{|\Omega| i^2 \pi^2}{6\delta})$. *With the assumption that* $\max(P(Z_M < E[Z_M|D_{i-1}]|D_{i-1}), P(Z_M > E[Z_M|D_{i-1}]|D_{i-1})) = \gamma_i \leq \gamma \in [0,1)$ *for* $\forall i \geq 1$ *and* β_i *in Eq.(6) set as* $\beta_i = \tilde{\beta}_i - 2\log(1-\gamma_i)$*, the following holds with probability* $\geq 1-\delta$*:*

$$R_n \leq \sqrt{(\tilde{\beta}_n - 2\log(1-\gamma))C_1 n \eta_n},$$

where $C_1 = \frac{8k_0}{\log(1+k_0^{-2})}$ *and* $\eta_n = \max_{A \subset \Omega; |A|=n} \frac{1}{2}\log|\mathbf{I} + \frac{K(A,A)}{k_0^3}|$*, where* $k_0 = k(x,x) > 0$ *is a constant for* $\forall x \in \Omega$ *for a stationary kernel* k*.*

The proof for Theorem 1 is provided in the Supplementary Materials. According to Srinivas et al. (2009), η_n is the maximum mutual information that can be gained about the objective function f from revealing n noisy observations with a noise variance of k_0^3. Therefore, Theorem 1 shows that the cumulative regret is bounded in terms of the maximum information gain. For the squared exponential kernel, $\eta_n = \mathcal{O}((\log n)^{d+1})$. For the Matérn kernel, $\eta_n = \mathcal{O}(n^{d(d+1)/(2\nu+d(d+1))}\log n)$, where $\nu > 1$ is a smoothness parameter of the kernel (Srinivas et al., 2009). Integrating these results of η_n with Theorem 1, we can show that the average regret, R_n/n, for the squared exponential kernel and the Matérn kernel is $R_n/n = \mathcal{O}(\frac{\sqrt{(\log n)^{d+2}}}{\sqrt{n}})$ and $R_n/n = \mathcal{O}(\frac{\log n}{n^{\nu/(2\nu+d(d+1))}})$, respectively. From these results, we conclude that $\lim_{n\to\infty} R_n/n = 0$. Because the simple regret, $|f(x_{opt}) - f(x_{best})|$, is smaller than the average regret, the simple regret of the OBCGP-UCB also converges to 0 with high probability. In general, the assumption of $\gamma_i \leq \gamma \in [0,1)$ for $\forall i \geq 1$ in Theorem 1 can be easily satisfied by setting $\gamma = \max_{i \leq n} \gamma_i$ because we have $\gamma_i = 0$ for $i > n$ with sufficiently large n owing to the consistency of the posterior distributions (Doob, 1961). Although the OBCGP-UCB has the same rate of convergence as the GP-UCB, the finite sample performance of the OBCGP-UCB was found to be superior in our experiments in Section 6.

Similar to Srinivas et al. (2009), we can generalize the results in Theorem 1 to any compact and convex $\Omega \subset R^d$ under mild assumptions on the kernel function k. These results are provided in the Supplementary Materials, along with the proof.

6. Empirical Studies

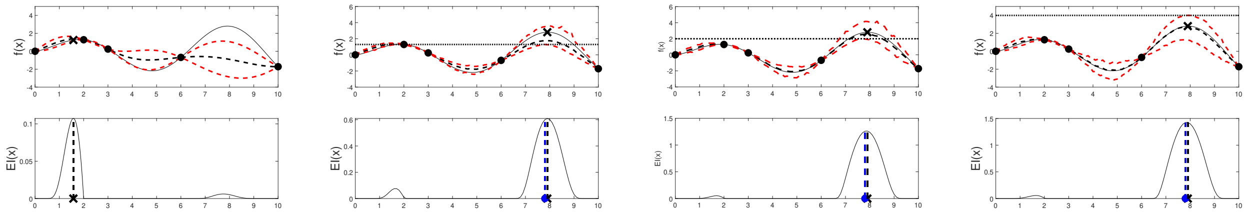
6.1. Behavior of the OBCGP

The behavior of the OBCGP may be understood by looking at how x_M is estimated. Let $f^q(x_M)$ denote the random variable following the posterior distribution of $f(x_M)$, i.e., $p(f(x_M)|D_n)$. For example, for Case 1, $f^q(x_M) = l_p + Z_M$, $Z_M \sim q(Z_M)$. Given all other parameters being fixed, the estimate of x_M, $\hat{x}_M$, is determined as follows:

$$\hat{x}_M = \arg\max_x E[\log N(f^q(x_M)|\mu_{GP}(x; D_n), \sigma^2_{GP}(x; D_n))], \tag{7}$$

where $\mu_{GP}(x; D_n)$ and $\sigma^2_{GP}(x; D_n)$ are the GP posterior mean and variance at x, respectively (see Section 2.1), and $N(\cdot|\cdot,\cdot)$ represents the normal probability density function. The details of the derivation of Eq.(7) are provided in the Supplementary Materials. In Eq.(7), $\hat{x}_M$ is closely related to the GP posterior mean and variance. This indicates that, although x_M is a parameter of the OBCGP, we may understand its behavior using the GP posterior results. Eq.(7) implies that $\hat{x}_M$ is chosen such that the following two preferences are balanced: (1) $\mu_{GP}(\hat{x}_M; D_n)$ is close to $E[f^q(x_M)]$ and (2) $\sigma^2_{GP}(\hat{x}_M; D_n)$ well covers the gap between $\mu_{GP}(\hat{x}_M; D_n)$ and $E[f^q(x_M)]$. This argument helps to understand the possible candidates for $\hat{x}_M$. For example, suppose we apply the OBCGP without any prior bound knowledge. Then, we set $l_p = f(x_{best})$, and $f^q(x_M)$ is in the range of $[f(x_{best}), \infty]$ (thus, $E[f^q(x_M)] > f(x_{best})$). Then, we can consider the locations near x_{best} (denoted as x^{near}_{best}) as the candidates of $\hat{x}_M$ because $\mu_{GP}(x^{near}_{best}; D_n)$ would be closer to $f(x_{best})$, and thus, $E[f^q(x_M)]$, than the other locations are. However, a gap should exist between $\mu_{GP}(x^{near}_{best}; D_n)$ and $E[f^q(x_M)]$. If $\sigma^2_{GP}(x^{near}_{best}; D_n)$ is too small to cover the gap, then x^{near}_{best} may not be chosen for $\hat{x}_M$. Instead, other locations with a larger posterior variance can be selected for $\hat{x}_M$.

We confirm this argument by comparing the behaviors of the OBCGP and the GP on some BO examples. Figure 1 shows a comparison between the estimated posterior mean function and the next query point using the standard GP in (a), the OBCGP without any prior bound knowledge in (b), the OBCGP with a given lower bound in (c), and the OBCGP with a given upper bound in (d). The EI was used for the acquisition function. In each figure in the upper panel of Figure 1, the five black circles represent the observations, solid line represents the true function, black dashed line denotes the estimated posterior mean function, red dashed lines indicate the 95 % confidence intervals, cross mark represents the next query point, and horizontal dotted line denotes the bound as the current best value (in (b)) or from prior knowledge (in (c) and (d)). The lower panel in Figure 1 shows a plot of the EI for each potential next point to sample, where the blue and black vertical lines represent the location of $\hat{x}_M$ and the next query point determined using



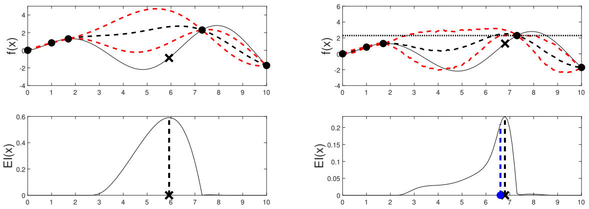
(a) GP-EI (b) OBCGP-EI without a known bound (c) OBCGP-EI with LB (d) OBCGP-EI with UB

Figure 1. Upper panel: comparison between the GP and the OBCGP; solid line: true function; circles: observations; black dashed line: posterior mean function; red dashed lines: 95 % confidence intervals; horizontal dotted line: the best observed function value in (b), a known lower bound (LB) in (c), and an upper bound (UB) in (d); cross marks: the next query point. Lower panel: the EI for each potential next point; blue vertical line: $\hat{x}_M$; black vertical line: the next query point determined using the EI.

the EI, respectively.

We can see that the estimated posterior mean functions that incorporate the bound information (either the current best value or a known bound) using the OBCGP significantly differ from those using the pure GP, although they are based on the same observations. As shown in Figures 1(a) and 1(b), even without any prior bound knowledge, the predicted function using the OBCGP is considerably different from that using the GP. In Figure 1(a), the posterior variances near x_{best} are very small; therefore, in Figure 1(b), the OBCGP determines $\hat{x}_M$ (blue vertical line) far from x_{best}. In Figure 1(b), we can see that $\hat{x}_M$ is located to be near the true maximum, and at this point, the posterior mean of $f(\hat{x}_M)$ (i.e., $E[f^q(x_M)]$) is slightly larger than $f(x_{best})$ (the horizontal dotted line). In Figures 1(c) and 1(d), $\hat{x}_M$ is also determined to be near the true maximum; however, $E[f^q(x_M)]$ is slightly larger than the lower bound in Figure 1(c) or smaller than the upper bound but larger than $f(x_{best})$ in Figure 1(d), for each bound type. These results show that the OBCGP is an effective tool that can incorporate the bound information into a surrogate model. We found the predicted function using the OBCGP to be significantly different from that using the GP. This difference in the objective functions using the GP and the OBCGP produced different results in selecting the next query point, although the same acquisition function was used. In the case of the EI with the GP (Figure 1(a)), a point near x_{best} was chosen as the next query point. In contrast, the next query point selected by the EI with OBCGP (Figures 1(b)-1(d)) was the one located far away from x_{best}. The next query point determined using the OBCGP was located close to $\hat{x}_M$.

To illustrate another scenario, Figure 2 shows the results of similar experiments as those in Figure 1 but with a different set of observations. In this case, as shown in Figure 2(a), the GP posterior mean near x_{best} is considerably large compared the previous scenario. Therefore, the gap between $\mu_{GP}(x_{best}^{near}; D_n)$ and $E[f^q(x_M)]$ may be covered by $\sigma^2_{GP}(x_{best}^{near}; D_n)$. This may lead to the OBCGP with $l_p = f(x_{best})$ determining $\hat{x}_M$ to be near x_{best} as shown in Figure 2(b).



(a) GP-EI (b) OBCGP-EI without a known bound

Figure 2. Similar experiments as those in Figure 1 but with a different set of observations.

6.2. Applications to Simulated Functions

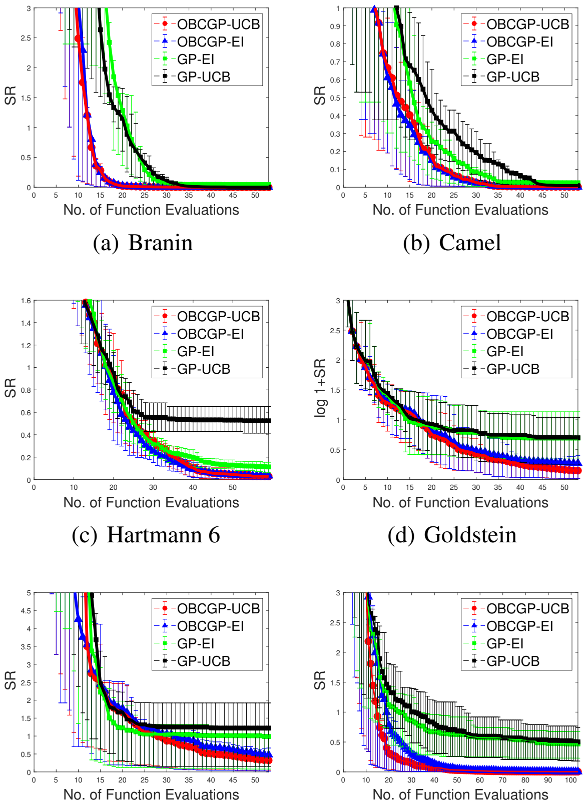
We evaluated the performance of BO with the OBCGP and compared it with that of BO with the GP. We used five test functions, namely, the Branin, six-hump camel, Hartman 6, Goldstein price, and Rosenbrock functions (Jamil and Yang, 2013). The Hartman 6 function is six-dimensional, whereas the other functions are two-dimensional. Furthermore, to consider a high-dimensional case, we embedded the Branin function in a 20-dimensional space (denoted as "Branin-20D") by adding 18 surplus dimensions that do not affect the function value at all. All functions deal with minimization problems (i.e., our task is to find $x_{opt} = \arg\min_{x\in\mathcal{X}} f(x)$). Suppose that we have knowledge of $f(x_{opt}) < u_p$ or $f(x_{opt}) > l_p$. By changing the problem to $x_{opt} = \arg\max_{x\in\mathcal{X}} -f(x)$ and replacing u_p and l_p with $-u_p$ and $-l_p$, respectively, we can formulate the cases of $f(x_{opt}) < u_p$ and $f(x_{opt}) > l_p$ as Case 1 and Case 2, respectively. For each test function, we first evaluated the initial design points determined using Latin hypercube sampling, which is one of the most popular sampling methods for an initial design (Kim et al., 2015). We used nine initial points for the Hartman 6 function and five initial points for all the other functions. The same initial points were used for both the OBCGP and the GP for each test function.

For a fair comparison between the OBCGP and the GP, we assumed that there was no prior knowledge of the bound on the optimal function value. Therefore, for the OBCGP, we set an initial upper bound on the optimal function value as the minimum value of the initially evaluated points, i.e., the best observed function value. We set hyperparameter λ to 0.1 (i.e., $E[Z_M] = 0.1$ in Eq.(4)). Then, we estimated the function using the OBCGP or GP and selected the next query point using the EI or UCB. Subsequently, we evaluated the function at the new point. We repeated the estimation of the function and the selection of the next query until we evaluated 50 more points. Then, we estimated the optimum as $f(x_{best})$, based on all the evaluated points. Using different sets of initial points, we repeated the optimization procedure 200 times. For each test function, we measured the average and the first and third quartiles of the simple regret over 200 experiments with the number of function evaluations and compared these quantities between the OBCGP and the GP. Figure 3 presents the results. We can see that although the OBCGP does not use any known bound on the optimal function value, BO with the OBCGP outperforms the traditional BO, particularly for highly volatile functions such as the Goldstein and Rosenbrock functions. Moreover, we can see that BO with the OBCGP can find the optimum of high-dimensional functions (i.e., Hartmann 6 and Branin-20D) after a sufficient number of function evaluations, whereas BO with the GP cannot.

The OBCGP can inherently incorporate a new type of prior knowledge, i.e., the bounds on the optimal function values, if it is available. To demonstrate this advantage, we conducted further experiments by assuming that certain upper or lower bounds for each test function were available from expert knowledge or past experiments. Detailed analysis and discussions are provided in the Supplementary Materials. The code for BO with the OBCGP is available at https://github.com/twj-KAIST/OBCGP-BO.

6.3. Sensitivity Analysis

We performed a sensitivity analysis of λ in Eq.(4) to investigate the impact of λ on the BO results. Figure 4 shows a comparison of the BO performance of the OBCGP with different λ values of 0.1 and 0.01 on the Branin and Hartmann 6 functions, together with the BO performance of the GP. We considered the UCB and EI as the acquisition functions. We can see that the OBCGP with different λ values showed a slight difference in the speed to reach the optimum, but equally found the true optimum very well after sufficient iterations. In particular, compared with the GP under the same acquisition function, the OBCGP was substantially better at finding the optimum regardless of λ values. Additional results of the sensitivity analysis on the other test functions are provided in the Supplementary Materials.



(a) Branin (b) Camel

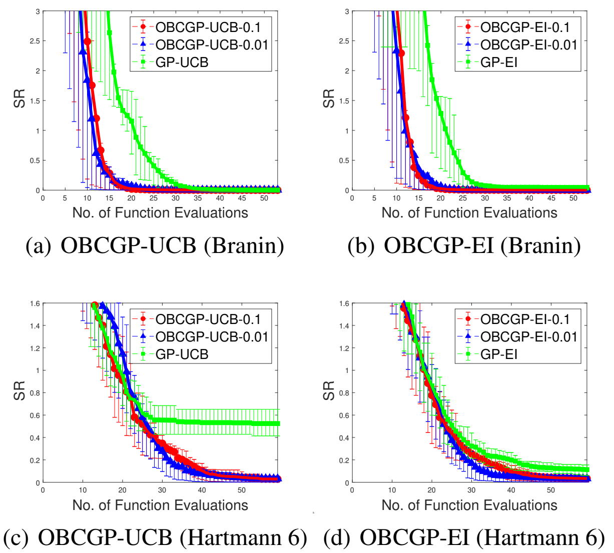
(c) Hartmann 6 (d) Goldstein

(e) Rosenbrock (f) Branin-20D

Figure 3. Comparison between the performances of BO with the OBCGP and BO with the GP: simple regret (SR) vs. the number of function evaluations.

6.4. Hyperparameter Optimization in Neural Networks

BO is one of the most widely used frameworks for tuning hyperparameters in neural networks. We applied the OBCGP to optimize the hyperparamters in a multilayer perceptron (MLP) (LeCun et al., 2015) to classify a popular MNIST dataset (LeCun and Cortes, 2010). Two acquisition functions of the UCB and EI were considered. We trained 60,000 images and tested 10,000 images using the MLP with two hidden layers of 100 and 50 hidden units, each stacked with a sigmoid activation function. To avoid overfitting, we considered ℓ_1 regularization of the weights in the MLP (Phaisangittisagul, 2016). Moreover, we considered adding noise to the input data to train the MLP more robustly (Zhang et al., 2017; Vincent et al., 2010). The regularization coefficient and the variance of the injected noise, together with the learning rate, were hyperparameters that



(a) OBCGP-UCB (Branin) (b) OBCGP-EI (Branin)

(c) OBCGP-UCB (Hartmann 6) (d) OBCGP-EI (Hartmann 6)

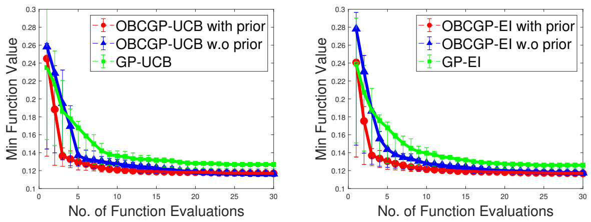
Figure 4. Sensitivity analysis of λ for OBCGP (Case 1).

were optimized using the OBCGP. The three hyperparameters were searched over the same region from 0.0001 to 0.1. We calculated the validation loss with the cross-entropy loss function, which is a black-box function over the hyperparameters and expensive to evaluate. By minimizing the validation loss using BO with the OBCGP, we optimized the hyperparameters.

We conducted two sets of experiments using BO with the OBCGP: with and without the incorporation of the prior knowledge of a bound on the optimal function value. The experiments with the incorporated prior knowledge showed the unique advantage of our proposed OBCGP in serving as a framework to incorporate a new type of information, i.e., the bounds on the optimal function value, into a surrogate model. To assume a plausible prior bound, not a cherry-picking bound, we assumed that another MLP classifier with only one hyperparameter, namely, the learning rate, was already implemented, with the learning rate being determined via the GP-EI. We used the validation loss of this MLP with the optimized learning rate as our prior knowledge of an upper bound for optimal validation loss. This setting of an upper bound was based on the idea that the MLP with two more hyperparameters would perform more effectively because more hyperparameters, if they are tuned optimally, would refine the network for better performance. We also performed experiments without the incorporation of the the prior knowledge of the bound on the optimal function value for a fair comparison of the BO performance between the OBCGP and the GP.

We first evaluated five initial points determined using Latin hypercube sampling. We set an initial upper bound on the optimal function value as the minimum value of the five evaluated points. We performed BO with the OBCGP until we evaluated 20 additional points, updating the bound during the experiments. We recorded the minimum validation loss based on the evaluated points for each iteration. Using different sets of initial points, we repeated this procedure 100 times and evaluated the average of the minimum validation loss at each iteration based on the results of the 100 experiments. Similarly, to consider the prior knowledge, we evaluated the average of the minimum validation loss at each iteration on the basis of the results of the 100 experiments for the MLP with one hyperparameter optimized using the GP-EI and used the minimum validation loss at the final iteration as an upper bound on the minimum validation loss for the MLP with three hyperparameters.

Figures 5(a) and 5(b) show the results of the MLP with the three hyperparameters optimized using the OBCGP-UCB and OBCGP-EI, respectively, where the red lines represent the OBCGP with the incorporation of the prior knowledge and the blue lines represent the OBCGP without the incorporation of the prior knowledge, along with the results of the GP in the green line. With or without the incorporation of the prior knowledge, the OBCGP-UCB and OBCGP-EI reached the minimum validation loss more rapidly than the GP-UCB and GP-EI. Moreover, the value of the minimum validation loss at the final iteration was significantly smaller using the OBCGP than that using the GP. Furthermore, with the incorporation of the prior knowledge, the OBCGP-UCB and OBCGP-EI reached the minimum validation loss even more rapidly.



(a) MLP example: UCB (b) MLP example: EI

Figure 5. Minimum validation loss of the MLP.

7. Conclusion

In this paper, we proposed the OBCGP, a new surrogate model for BO. The OBCGP incorporates important information: the existence of a point that is better than the current best point. This information is particularly useful when the standard GP suffers from a small number of evaluated points. By incorporating this information into the GP, we can estimate the objective function more accurately and search for optimal points more efficiently. The OBCGP can be used with various acquisition functions to select the next query point. In particular, we derived the cumulative regret bound of the OBCGP-UCB and showed that the average regret with the commonly used kernel functions converges to 0. Through the experimental results, we demonstrated that the OBCGP with popular acquisition functions is more effective for optimizing black-box functions than the standard GP. Furthermore, the OBCGP can inherently incorporate a new type of prior knowledge, i.e., the bounds on the optimal function values, if it is available. The incorporation of this type of prior knowledge into a surrogate model has not been studied previously. Through the experimental results, we also confirmed that the OBCGP can effectively incorporate this type of prior knowledge. In this study, we assumed noiseless observations. A simple method to consider noisy observations, $y(x) = f(x) + \epsilon$, $\epsilon \sim N(0, \sigma_\epsilon^2)$, is to replace $f(x_{best})$ with $y(x_{best}) - z_{\alpha/2}\sigma_\epsilon$, where α is the significance level, for the maximization problem. More sophisticated approaches can be studied in future work.

References

Bertsimas, D., Tsitsiklis, J., et al. (1993). Simulated annealing. *Statistical science*, 8(1):10–15.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Doob, J. (1961). Notes on martingale theory. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob*, volume 2, pages 95–102.

Glover, F. and Laguna, M. (2013). Tabu search. In *Handbook of Combinatorial Optimization*, pages 3261–3362. Springer.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926.

Hoffman, M. W. and Ghahramani, Z. (2015). Output-space predictive entropy search for flexible global optimization. In *NIPS workshop on Bayesian Optimization*.

Jamil, M. and Yang, X.-S. (2013). A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150–194.

Jauch, M. and Peña, V. (2016). Bayesian optimization with shape constraints. *NIPS 2016 Bayesian Optimization Workshop*.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.

Kim, H., Vastola, J. T., Kim, S., Lu, J.-C., and Grover, M. A. (2017). Incorporation of engineering knowledge into the modeling process: a local approach. *International Journal of Production Research*, 55(20):5865–5880.

Kim, S., Kim, H., Lu, R. W., Lu, J.-C., Casciato, M. J., and Grover, M. A. (2015). Adaptive combined space-filling and d-optimal designs. *International Journal of Production Research*, 53(17):5354–5368.

Kingma, D. P. and Ba, J. L. (2015). Adam: Amethod for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Osborne, M. A., Garnett, R., and Roberts, S. J. (2009). Gaussian processes for global optimization. In *3rd international conference on learning and intelligent optimization (LION3)*, pages 1–15.

Phaisangittisagul, E. (2016). An analysis of the regularization between l2 and dropout in single hidden layer neural network. In *Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference on*, pages 174–179. IEEE.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3627–3635. JMLR. org.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155.