Practical Multi-fidelity Bayesian Optimization for Hyperparameter Tuning Supplementary Material

Jian Wu
Operations Research
& Information Eng.
Cornell University
Ithaca, NY 14850

Saul Toscano-Palmerin
Operations Research
& Information Eng.
Cornell University
Ithaca, NY 14850

Peter I. Frazier
Operations Research
& Information Eng.
Cornell University
Ithaca, NY 14850

Andrew Gordon Wilson
Courant Institute
of Mathematical Sciences
New York University
New York, NY 10003

1 Background: Gaussian processes

We put a Gaussian process (GP) prior [5] on the function g. The GP prior is defined by its mean function $\mu_0 : A \times [0,1]^m \mapsto \mathbb{R}$ and kernel function $K_0 : \{A \times [0,1]^m\} \times \{A \times [0,1]^m\} \mapsto \mathbb{R}$. These mean and kernel functions have hyperparameters, whose inference we discuss below.

We assume that evaluations of $g(\mathbf{x}, \mathbf{s})$ are subject to additive independent normally distributed noise with common variance σ^2. We treat the parameter σ^2 as a hyperparameter of our model, and also discuss its inference below. Our assumption of normally distributed noise with constant variance is common in the BO literature [2].

Here we use $\mathbf{z} = (\mathbf{x}, \mathbf{s})$ to refer more briefly to a point, fidelity pair. The posterior distribution on g after observing n function values at points $\mathbf{z}_{(1:n)} := \{(\mathbf{x}_{(1)}, \mathbf{s}_{(1)}), (\mathbf{x}_{(2)}, \mathbf{s}_{(2)}), \cdots, (\mathbf{x}_{(n)}, \mathbf{s}_{(n)})\}$ with observed values $y_{(1:n)} := \{y_{(1)}, y_{(2)}, \cdots, y_{(n)}\}$ remains a Gaussian process [5], and $g \mid \mathbf{z}_{(1:n)}, y_{(1:n)} \sim \mathrm{GP}(\mu_n, K_n)$ with μ_n and K_n as follows, where I is an identity matrix:

$$\mu_n(\mathbf{z}) = \mu_0(\mathbf{z}) \\ + K_0(\mathbf{z}, \mathbf{z}_{1:n}) \left(K_0(\mathbf{z}_{1:n}, \mathbf{z}_{1:n}) + \sigma^2 I\right)^{-1} (y_{1:n} - \mu_0(\mathbf{z}_{1:n}))$$

$$K_n(\mathbf{z}, \mathbf{z}') = K_0(\mathbf{z}, \mathbf{z}') \\ - K_0(\mathbf{z}, \mathbf{z}_{1:n}) \left(K_0(\mathbf{z}_{1:n}, \mathbf{z}_{1:n}) + \sigma^2 I\right)^{-1} K_0(\mathbf{z}_{1:n}, \mathbf{z}').$$

We should note that taKG may choose to retain more than one observations per evaluation because a single evaluation of g provides additional trace observations, and so n may be larger than the number of evaluations.

This statistical approach contains several hyperparameters: the variance σ^2, and any parameters in the mean and kernel functions. We treat these hyperparameters in a Bayesian way as proposed in Snoek et al. [7]. We analogously train a separate GP on the logarithm of the cost of evaluating $g(x, s)$.

Now, using the notation of the paper, let C_n be the Cholesky factor of the covariance matrix $K_n((\mathbf{x}, S), (\mathbf{x}, S)) + \sigma^2 I$. Thus, by the previous equations,

$$\mathbb{E}_n[g(\mathbf{x}', \mathbf{1}) | \mathbf{y}(\mathbf{x}, S)] = \mu_n(\mathbf{x}') \\ + K_n((\mathbf{x}', \mathbf{1}), (\mathbf{x}, S))(C_n^T)^{-1}(C_n)^{-1} \\ (\mathbf{y}(\mathbf{x}, S) - \mathbb{E}_n(\mathbf{y}(\mathbf{x}, S))), \tag{1.1}$$

and $(C_n)^{-1}(\mathbf{y}(\mathbf{x}, S) - \mathbb{E}_n(\mathbf{y}(\mathbf{x}, S)))$ follows an independent standard normal random distribution, which shows that

$$\mathbb{E}_n[g(\mathbf{x}', \mathbf{1}) | \mathbf{y}(\mathbf{x}, S)] = \mu_n(\mathbf{x}') \\ + \tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S)\mathbf{w}, \tag{1.2}$$

where $\mathbf{w}$ is an independent standard normal random vector.

2 Proofs Details

In this section we prove the theorems of the paper. We first show some smoothness properties of $\tilde{\sigma}_n$, μ_n and c_n in the following lemma.

Lemma 1. *We assume that the domain A is compact, μ_0 is a constant, the kernel K_0 is continuously differentiable, and the prior parameters on $\log c$ continuously differentiable. We then have that*

1. *Fix any $\mathbf{x}$ and S. Then $\mu_n(\mathbf{x}')$ and $\tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S)$ are both continuously differentiable in $\mathbf{x}'$.*

2. *Fix any* $\mathbf{x}'$ *and number of fidelities* $|S|$. *Then* $\tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S)$ *is continuously differentiable in* $\mathbf{x}$ *and each element of* S.

3. c_n *is continuously differentiable.*

4. $\max_{1 \leq i \leq q} c_n(x_i, s_i)$ *is differentiable in* $\mathbf{x}$ *and* $\mathbf{s}$ *if* $\left|\text{argmax}_{1 \leq i \leq q} c_n(x_i, s_i)\right| = 1$.

Proof. The posterior parameters of the Gaussian process on $\log c$ are continuously differentiable if its prior parameters are continuously differentiable (this proves (3)).

By (1.1), we know that that $\tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S) = K_n((\mathbf{x}', 1), (\mathbf{x}, S))(C_n^T)^{-1}$ where $(\mathbf{x}, S) := \{(\mathbf{x}, \mathbf{s}) : \mathbf{s} \in S\}$ and C_n is the Cholesky factor of the covariance matrix $K_n((\mathbf{x}, S), (\mathbf{x}, S)) + \sigma^2 I$. Thus, (1) follows from continuous differentiability of K_n.

To prove (2) we only need to show that $(C_n^T)^{-1}$ is continuously differentiable with respect to $\mathbf{x}$ and the components of S. This follows from the fact that multiplication, matrix inversion (when the inverse exists) , and Cholesky factorization [6] preserve continuous differentiability.

(4) follows easily from (3). □

We now prove Theorem 1.

Proof of Theorem 1. Recall the intuitive explanation of Theorem 1 given in the body of the paper:

$$\begin{aligned}
&\nabla_{\mathbf{x},S}\, \mathbb{E}_n\left[\min_{\mathbf{x}'}\left(\mu_n(\mathbf{x}', \mathbf{1}) + \tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S) \cdot \mathbf{w}\right)\right] \\
&= \mathbb{E}_n\left[\nabla_{\mathbf{x},S} \min_{\mathbf{x}'}\left(\mu_n(\mathbf{x}', \mathbf{1}) + \tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S) \cdot \mathbf{w}\right)\right] \\
&= \mathbb{E}_n\left[\nabla_{\mathbf{x},S}\left(\mu_n(\mathbf{x}^*, \mathbf{1}) + \tilde{\sigma}_n(\mathbf{x}^*, \mathbf{x}, S) \cdot \mathbf{w}\right)\right] \\
&= \mathbb{E}_n\left[\nabla_{\mathbf{x},S}\, \tilde{\sigma}_n(\mathbf{x}^*, \mathbf{x}, S) \cdot \mathbf{w}\right],
\end{aligned}$$

where $\mathbf{x}^*$ is a global minimum (over $\mathbf{x}' \in A$) of $h(\mathbf{x}', \mathbf{x}, S) := \mu_n(\mathbf{x}', \mathbf{1}) + \tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S) \cdot \mathbf{w}$, $\mathbf{w}$ is a standard normal random vector, and $\nabla_{\mathbf{x},S}$ indicates the gradient with respect to $\mathbf{x}$ and S holding $\mathbf{x}^*$ fixed.

To complete the proof, we need to justify the interchange of expectation and the gradient (the second line) and ignoring the dependence of $\mathbf{x}^*$ on $\mathbf{x}$ and S when taking the gradient (the third line).

We first justify the third line. By Lemma 1, h is continuously differentiable. Thus, by the envelope theorem (see Corollary 4 of Milgrom and Segal 4), even though $\mathbf{x}^*$ depends on $\mathbf{x}$ and S, this dependence can be ignored when computing $\nabla_{\mathbf{x},S}\, h(\mathbf{x}^*, \mathbf{x}, S)$ (observe that we assume that $\mathbf{x}^*$ is unique in the statement of the theorem).

We now justify the fourth line. Recall that A is compact, components of $\mathbf{s}$ have domain $[0, 1]$, and gradients with respect to S are taken assuming that $|S|$ is held fixed. Thus, the domain of $\mathbf{x}, S$ is compact. Also $\tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S)$ is continuously differentiable with respect to $\mathbf{x}, S$ by Lemma 1. Thus $\|\tilde{\sigma}_n(\mathbf{x}', \mathbf{x}, S)\|$ is bounded. Consequently, Corollary 5.9 of Bartle [1] implies that we can interchange the gradient and the expectation.

□

The following corollary follows from the previous proof.

Corollary 1. *Under the assumptions of the previous theorem,* $L_n(\mathbf{x}, S)$ *is continuous.*

We now prove Theorem 2.

Proof. We prove this theorem using Theorem 2.3 of Section 5 of Kushner and Yin [3], which depends on the structure of the stochastic gradient G of the objective function. In addition, we simplify the notation and denote $(\mathbf{x}_t, S_t)$ by Z_t.

The theorem from Kushner and Yin [3], requires the following hypotheses:

1. $\epsilon_t \to 0$, $\sum_{t=1}^{\infty} \epsilon_t = \infty$, and $\sum_t \epsilon_t^2 < \infty$.

2. $\sup_t \mathbb{E}\left[|G(Z_t)|^2\right] < \infty$

3. There exist uniformly continuous functions $\{\lambda_t\}_{t \geq 0}$ of Z, and random vectors $\{\beta_t\}_{t \geq 0}$, such that $\beta_t \to 0$ almost surely and
$$E_n[G(Z_t)] = \lambda_t(Z_t) + \beta_t.$$
Furthermore, there exists a continuous function $\bar{\lambda}$, such that for each $Z \in A^q$,
$$\lim_n \left|\sum_{i=1}^{m(r_m+s)} \epsilon_i\left[\lambda_i(Z) - \bar{\lambda}(Z)\right]\right| = 0$$
for each $s \geq 0$, where $m(r)$ is the unique value of k such that $t_k \leq t < t_{k+1}$, where $t_0 = 0$,$t_k = \sum_{i=0}^{k-1} \epsilon_i$.

4. There exists a continuously differentiable real-valued function ϕ, such that $\bar{\lambda} = -\nabla\phi$ and it is constant on each connected subset of stationary points.

5. The constraint functions defining A are continuously differentiable.

We now prove that our problem satisfies these hypotheses. (1) is true by the hypothesis of the lemma.

We now prove (2). Letting $\mathbf{x}^*$ be defined in terms of $\mathbf{w}$ as in Theorem 2 and choosing a generic fixed Z,

$$\mathbb{E}\left[\left|\nabla_{\mathbf{x}}\,\tilde{\sigma}_n\left(\mathbf{x}^*, Z\right)\cdot\mathbf{w}\right|^2\right]$$
$$\leq \mathbb{E}\left[\left\|\nabla\tilde{\sigma}_n\left(\mathbf{x}^*, Z\right)\right\|^2\left\|\mathbf{w}\right\|^2\right] \leq M|S|$$

where $M := \sup_{\mathbf{x},\mathbf{z}}\left\|\nabla\tilde{\sigma}_n\left(\mathbf{x},\mathbf{z}\right)\right\|^2$ and $|S|$ is the dimensionality of $\mathbf{w}$. M is finite because the domain of the problem is compact and $\nabla\tilde{\sigma}_n\left(\mathbf{x},\mathbf{z}\right)$ is continuous by Lemma 1. Since c_n is continuously differentiable and bounded below, we conclude that the supremum over Z of $\mathbb{E}\left[\left|G\left(Z\right)\right|^2\right]$ is bounded.

We now prove (3). Our definition of λ_t will be the same for all t. Define

$$\lambda_t\left(Z\right) = \mathbb{E}\left[\frac{c_n\left(Z\right)\nabla\tilde{\sigma}_n\left(\mathbf{x}^*, Z\right)\mathbf{w}}{c_n\left(Z\right)^2}\right] - \mathbb{E}\left[\frac{\nabla c_n\left(Z\right)}{c_n\left(Z\right)^2}\left(\mu_n\left(\mathbf{x}^*, \mathbf{1}\right) + \tilde{\sigma}_n\left(\mathbf{x}^*, Z\right)\mathbf{w}\right)\right].$$

We will prove that λ_t is continuous. In the proof of Theorem 1, we show that $\nabla\tilde{\sigma}_n\left(\mathbf{x}^*, Z\right)\mathbf{w}$ is continuous in Z. Furthermore,

$$\begin{aligned}\left\|\nabla\tilde{\sigma}_n\left(y_1, Z\right)\mathbf{w}^1\right\| &\leq \left\|\nabla\tilde{\sigma}_n\left(y_1, Z\right)\right\|\left\|\mathbf{w}\right\| \\ &\leq L\left\|\mathbf{w}\right\|.\end{aligned}$$

Consequently $\mathbb{E}\left[\nabla\tilde{\sigma}_n\left(Y, Z\right)\mathbf{w}\right]$ is continuous by Corollary 5.7 of Bartle [1]. In Theorem 1, we also show that $\mathbb{E}\left[\left(\mu_n\left(Y, \mathbf{1}\right) + \tilde{\sigma}_n\left(Y, Z\right)\mathbf{w}\right)\right]$ is continuous in Z. Since c_n is continuously differentiable, we conclude that λ_t is continuous. By defining $\beta_t = 0$ for all t, and $\bar{\lambda} = \lambda_1$, we conclude the proof of (3).

Finally, define $\phi\left(Z\right) = -\mathbb{E}\left[\frac{\mu_n(Y,\mathbf{1})+\tilde{\sigma}_n(Y,Z)\mathbf{w}}{c_n(Z)}\right]$. Observe that in Lemma 2, we show that we can interchange the expectation and the gradient in $\mathbb{E}\left[\nabla\left(\mu_n\left(Y\right) + \tilde{\sigma}_n\left(Y, Z\right)\mathbf{w}\right)\right]$, and so $\lambda_m\left(Z\right) = -\nabla\phi\left(Z\right)$. In a connected subset of stationary points, we have that $\lambda_m\left(Z\right) = 0$, and so $\phi\left(Z\right)$ is constant. This ends the proof of the theorem. ☐

Proof of Proposition 1. Since

$$VOI_n(x,s) := \mathbb{E}_n[\mu^*(x,1) - \min_{x'}(\mu_n(x') + C_n(x',(x,s))W]$$

where W is a standard normal random variable. By Jensen's inequality, we have

$$\begin{aligned}VOI_n(x,s) &:= \mathbb{E}_n[\mu^*(x,1) - \min_{x'}(u_n\left(x',s,W\right))] \\ &\geq \mu^*(x,1) - \min_{x'}\mathbb{E}_n(u_n\left(x',s,W\right)) = 0.\end{aligned}$$

where $u_n\left(x,s,W\right) := \mu^n(x',1) + C_n((x',1),(x,s))W)$. The inequality becomes equal only if $\min_{x'}(\mu_n(x') + C_n(x',(x,s))W$ is a linear function of W for any fixed (x,s), i.e the argmin for the inner optimization function doesn't change as we vary W, which is not true if $K_n((x',1),(x,s)) > 0$ i.e. evaluating at (x,s) provides value to determine the argmin of the surface $(x,1)$. ☐

Proof of Proposition 2. The proof follows a very similar argument than the previous proof. By Jensen's inequality, we have that

$$\mathbb{E}_n\left[\min_{x'}\mathbb{E}_n\left[g(x',1) \mid \mathbf{y}(x,S)\right]\right] \geq \mathbb{E}_n\left[\min_{x'}\mathbb{E}_n\left[g(x',1) \mid \mathbf{y}(x,S\bigcup C(S))\right]\right]$$

The inequality becomes equal only if the argmin for the inner optimization function doesn't change as we vary the normal random vector, which is not true under our assumptions.

☐

3 GPs for Hyperparameter Optimization

In the context of hyperparameter optimization with two continuous fidelities, i.e. the number of training iterations ($s_{(1)}$) and the amount of training data ($s_{(2)}$), we set the kernel function of the GP as

$$K_0(z,\tilde{z}) = K(x,\tilde{x}) \times K_1(s_{(1)},\tilde{s}_{(1)}) \times K_2(s_{(2)},,\tilde{s}_{(2)}),$$

where $K(\cdot,\cdot)$ is a square-exponential kernel. If we assume that the learning curve looks like

$$g(x,s) = h(x) \times \left(\beta_0 + \beta_1\exp\left(-\lambda s_{(1)}\right)\right) \times l(s_{(2)}),\tag{3.1}$$

then inspired by [8], we set the kernel $K_1(\cdot,\cdot)$ as

$$K_1(s_{(1)},\tilde{s}_{(1)}) = \left(w + \frac{\beta^\alpha}{(s_{(1)} + \tilde{s}_{(1)} + \beta^\alpha)}\right),$$

where $w,\beta,\alpha > 0$ are hyperparameters. We add an intercept w compared to the kernel in [8] to model the fact that the loss will not diminish. We assume that the kernel $K_2(\cdot,\cdot)$ has the form

$$K_2(s_{(2)},\tilde{s}_{(2)}) = \left(c + (1 - s_{(2)})^{(1+\delta)}(1 - \tilde{s}_{(2)})^{(1+\delta)}\right),$$

where $c,\delta > 0$ are hyperparameters.

All the hyperparameters can be treated in a Bayesian way as proposed in Snoek et al. [7].

4 Additional experimental details

4.1 Synthetic experiments

Here we define in detail the synthetic test functions on which we perform numerical experiments The test functions are:

$$\begin{aligned}&\text{augmented-Branin}(\mathbf{x}, \mathbf{s})\\&= \left(x_2 - \left(\frac{5.1}{4\pi^2} - 0.1 * (1 - s_1)\right) x_1^2 + \frac{5}{\pi} x_1 - 6\right)^2\\&+ 10 * \left(1 - \frac{1}{8\pi}\right) \cos(x_1) + 10\end{aligned}$$

$$\begin{aligned}&\text{augmented-Hartmann}(\mathbf{x}, \mathbf{s})\\&= (\alpha_1 - 0.1 * (1 - s_1)) \exp\left(-\sum_{j=1}^{d} A_{ij}(x_j - P_{1j})^2\right)\\&+ \sum_{i=2}^{4} \alpha_i \exp\left(-\sum_{j=1}^{d} A_{ij}(x_j - P_{ij})^2\right)\end{aligned}$$

$$\begin{aligned}&\text{augmented-Rosenbrock}(\mathbf{x}, \mathbf{s})\\&= \sum_{i=1}^{2} \big(100 * (x_{i+1} - x_i^2 + 0.1 * (1 - s_1))^2\\&+ \left(x_i - 1 + 0.1 * (1 - s_2)^2\right)^2\big).\end{aligned}$$

4.2 Real-world experiments

The range of search domain for feedforward NN experiments: the learning rate in $[10^{-6}, 10^0]$, dropout rate in $[0, 1]$, batch size in $[2^5, 2^{10}]$ and the number of units at each layer in $[100, 1000]$.

The range of search domain for CNN experiments: the learning rate in $[10^{-6}, 1.0]$, batch size $[2^5, 2^{10}]$, and number of filters in each convolutional block in $[2^5, 2^9]$.

References

[1] R. G. Bartle. *The elements of integration*. John Wiley & Sons, 1966.

[2] A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial Intelligence and Statistics*, 2017.

[3] H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003.

[4] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[5] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN ISBN 0-262-18253-X.

[6] S. P. Smith. Differentiation of the cholesky algorithm. *Journal of Computational and Graphical Statistics*, 4(2):134 – 147, 1995.

[7] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

[8] K. Swersky, J. Snoek, and R. P. Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.