Discussion: "Bayesian Optimal Design of Experiments for Inferring the Statistical Expectation of Expensive Black-Box Functions" (Pandita, P., Bilionis, I., and Panchal, J., 2019, ASME J. Mech. Des., 141(10), p. 101404)

Xianliang Gong
Department of Naval Architecture and Marine Engineering,
University of Michigan,
Ann Arbor, MI 48109
e-mail: xlgong@umich.edu

Yulin Pan
Department of Naval Architecture and Marine Engineering,
University of Michigan,
Ann Arbor, MI 48109
e-mail: yulinpan@umich.edu

1 Introduction

In Ref. [1], the authors developed a sequential Bayesian optimal design framework to estimate the statistical expectation of a black-box function $f(\mathbf{x}): \mathbb{R}^d \to \mathbb{R}$. Let $\mathbf{x} \sim p(\mathbf{x})$ with $p(\mathbf{x})$ the probability distribution of the input $\mathbf{x}$, the statistical expectation is then defined as follows:

$$q = \int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \tag{1}$$

The function $f(\mathbf{x})$ is not known a priori but can be evaluated at arbitrary $\mathbf{x}$ with Gaussian noise of variance σ^2:

$$y = f(\mathbf{x}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

Based on the Gaussian process surrogate learned from the available samples $\mathbf{D}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$, i.e., $f(\mathbf{x})|\mathbf{D}_n \sim \mathcal{GP}(m_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}'))$, the next-best sample is chosen by maximizing the information-based acquisition $G(\tilde{\mathbf{x}})$:

$$\mathbf{x}_{n+1} = \mathrm{argmax}_{\tilde{\mathbf{x}}}\; G(\tilde{\mathbf{x}}) \tag{3}$$

where $G(\tilde{\mathbf{x}})$ computes the information gain of adding a sample $\tilde{y}$ at $\tilde{\mathbf{x}}$, i.e., the expected KL divergence between the current estimation $p(q|\mathbf{D}_n)$ and the hypothetical next-step estimation $p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})$:

$$\begin{aligned} G(\tilde{\mathbf{x}}) &= \mathbb{E}_{\tilde{y}}\left[\mathrm{KL}\left(p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) | p(q|\mathbf{D}_n)\right)\right] \\ &= \iint p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) \log \frac{p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})}{p(q|\mathbf{D}_n)}\, dq\; p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n)\, d\tilde{y} \end{aligned} \tag{4}$$

In Eq. (4), $\tilde{y}$ is chosen based on the surrogate $f(\mathbf{x})|\mathbf{D}_n$ following a distribution of $\mathcal{N}(m_n(\tilde{\mathbf{x}}), k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2)$); q is considered as a random variable with uncertainties coming from the (current and next-step) surrogates. It is noted that $G(\tilde{\mathbf{x}})$ also depends on the hyperparameter θ in the learned Gaussian process $f(\mathbf{x})|\mathbf{D}_n$. We neglect this dependence for simplicity, which does not affect the main derivation.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received June 23, 2021; final manuscript received November 13, 2021; published online December 17, 2021. Assoc. Editor: Pingfeng Wang.

As a major contribution of the discussed paper, the authors simplified the information-based acquisition as Eq. (30) in Ref. [1]:

$$G(\tilde{\mathbf{x}}) = \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{1}{2}\frac{\sigma_2^2(\tilde{\mathbf{x}})}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2}\frac{v(\tilde{\mathbf{x}})^2}{\sigma_1^2(\sigma_n^2(\tilde{\mathbf{x}}) + \sigma^2)} \tag{5}$$

where σ_1^2 and σ_2^2 are, respectively, the variances of current estimation and hypothetical next-step estimation of q; $v(\tilde{\mathbf{x}}) = \int k_n(\tilde{\mathbf{x}}, \mathbf{x}) p(\mathbf{x})\, d\mathbf{x}$ and $\sigma_n^2(\tilde{\mathbf{x}}) = k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})$. Furthermore, for numerical computation of Eq. (5), the authors developed analytical formulas for each involved quantity (important for high-dimensional computation) under uniform distribution of $\mathbf{x}$.

The purpose of our discussion is to show the following two critical points:

(1) The last three terms of Eq. (5) always add up to zero, leaving a concise form with a much more intuitive interpretation of the acquisition.
(2) The analytical computation of Eq. (5) can be generalized to arbitrary input distribution of $\mathbf{x}$, greatly broadening the application of the developed framework.

These two points are discussed, respectively, in Secs. 2 and 3.

2 Derivation of the Simplified Acquisition $G(\tilde{\mathbf{x}})$

To simplify Eq. (4), we first notice that $q|\mathbf{D}_n$ follows a Gaussian distribution with mean μ_1 and variance σ_1^2:

$$p(q|\mathbf{D}_n) = \mathcal{N}(q; \mu_1, \sigma_1^2)$$

$$\mu_1 = \mathbb{E}\left[\int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} | \mathbf{D}_n\right] \tag{6}$$

$$= \int m_n(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \tag{7}$$

$$\begin{aligned} \sigma_1^2 &= \mathbb{E}\left[\left(\int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}\right)^2 |\mathbf{D}_n\right] - \left(\mathbb{E}\left[\left(\int f_n(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}\right) |\mathbf{D}_n\right]\right)^2 \\ &= \iint k_n(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) p(\mathbf{x}')\, d\mathbf{x}'\, d\mathbf{x} \end{aligned} \tag{8}$$

After adding one hypothetical sample $\{\tilde{\mathbf{x}}, \tilde{y}\}$, the function follows an updated surrogate $f(x)|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y} \sim \mathcal{GP}(m_{n+1}(\mathbf{x}), k_{n+1}(\mathbf{x}, \mathbf{x}'))$ with

$$m_{n+1}(\mathbf{x}) = m_n(\mathbf{x}) + \frac{k_n(\tilde{\mathbf{x}}, \mathbf{x})(\tilde{y} - m_n(\tilde{\mathbf{x}}))}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \tag{9}$$

$$k_{n+1}(\mathbf{x}, \mathbf{x}') = k_n(\mathbf{x}, \mathbf{x}') - \frac{k_n(\tilde{\mathbf{x}}, \mathbf{x}) k_n(\mathbf{x}', \tilde{\mathbf{x}})}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \tag{10}$$

The quantity $q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}$ can then be represented by another Gaussian with mean μ_2 and variance σ_2^2:

$$p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) = \mathcal{N}(q; \mu_2(\tilde{\mathbf{x}}, \tilde{y}), \sigma_2^2(\tilde{\mathbf{x}}))\,, \tag{11}$$

$$\begin{aligned} \mu_2(\tilde{\mathbf{x}}, \tilde{y}) &= \mathbb{E}\left[\int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} | \mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right] \\ &= \int m_{n+1}(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \\ &= \mu_1 + \frac{\int k_n(\tilde{\mathbf{x}}, \mathbf{x}) p(\mathbf{x})\, d\mathbf{x}}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2}(\tilde{y} - m_n(\tilde{\mathbf{x}})) \end{aligned} \tag{12}$$

$$\sigma_2^2(\tilde{\mathbf{x}}) = \mathbb{E}\left[\left(\int f(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^2 |\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right] - \left(\mathbb{E}\left[\left(\int f_n(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}\right]\right)^2$$
$$= \iint k_{n+1}(\mathbf{x}, \mathbf{x}')p(\mathbf{x})p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\,\mathrm{d}\mathbf{x}$$
$$= \sigma_1^2 - \frac{(\int k_n(\tilde{\mathbf{x}}, \mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x})^2}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} \tag{13}$$

We note that Eqs. (7), (8), (12), and (13) are, respectively, intermediate steps of Eqs. (19), (21), (26), and (28) in the discussed paper. Substitute Eq. (6) and Eq. (11) into Eq. (4), one can obtain:

$$G(\tilde{\mathbf{x}}) = \iint p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y}) \log \frac{p(q|\mathbf{D}_n, \tilde{\mathbf{x}}, \tilde{y})}{p(q|\mathbf{D}_n)}\,\mathrm{d}q\; p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n))\,\mathrm{d}\tilde{y}$$
$$= \int\left(\log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{\sigma_2^2(\tilde{\mathbf{x}})}{2\sigma_1^2} + \frac{(\mu_2(\tilde{\mathbf{x}}, \tilde{y}) - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2}\right)p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n)\,\mathrm{d}\tilde{y}$$
$$= \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{1}{2\sigma_1^2}\left(\int(\mu_2(\tilde{\mathbf{x}}, \tilde{y}) - \mu_1)^2 p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}_n)\,\mathrm{d}\tilde{y} + \sigma_2^2(\tilde{\mathbf{x}}) - \sigma_1^2\right)$$
$$= \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) + \frac{1}{2\sigma_1^2}\left(\frac{(\int k_n(\tilde{\mathbf{x}}, \mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x})^2}{k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2} + \sigma_2^2(\tilde{\mathbf{x}}) - \sigma_1^2\right) \tag{14}$$

$$= \log\left(\frac{\sigma_1}{\sigma_2(\tilde{\mathbf{x}})}\right) \tag{15}$$

where Eq. (14) is exactly Eq. (5) (or Eq. (30) in discussed paper). The fact that the last three terms of Eq. (14) sum up to zero is a direct result of Eq. (13).

The advantage of having a simplified form (15) is that the optimization (3) yields a much more intuitive physical interpretation. Since σ_1 does not depend on $\tilde{\mathbf{x}}$, Eq. (3) can be reformulated as

$$\mathbf{x}_{n+1} = \mathrm{argmin}_{\tilde{\mathbf{x}}}\; \sigma_2^2(\tilde{\mathbf{x}}) \tag{16}$$

which selects the next-best sample minimizing the expected variance of q. Similar optimization criterion is also used in Refs. [2,3] for the purpose of computing the extreme-event probability.

Another alternative interpretation can be obtained by writing (3) as

$$\mathbf{x}_{n+1} \equiv \mathrm{argmax}_{\tilde{\mathbf{x}}}\; \sigma_1^2 - \sigma_2^2(\tilde{\mathbf{x}})$$
$$\equiv \mathrm{argmax}_{\tilde{\mathbf{x}}}\left(\frac{\int k_n(\tilde{\mathbf{x}}, \mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}}{(k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2)^{1/2}}\right)^2 \tag{17}$$

$$\equiv \mathrm{argmax}_{\tilde{\mathbf{x}}}\left(\int \rho_{y|\mathbf{D}_n}(\tilde{\mathbf{x}}, \mathbf{x})(k_n(\mathbf{x}, \mathbf{x}) + \sigma^2)^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^2, \tag{18}$$

where Eq. (17) is a result of Eq. (13), and $\rho_{y|\mathbf{D}_n}(\tilde{\mathbf{x}}, \mathbf{x}) = k_n(\tilde{\mathbf{x}}, \mathbf{x})/\left((k_n(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \sigma^2)(k_n(\mathbf{x}, \mathbf{x}) + \sigma^2)\right)^{1/2}$ is the correlation of y for two inputs $\tilde{\mathbf{x}}$ and $\mathbf{x}$. Equation (18) can be interpreted as to select the next sample which has overall most (weighted) correlation with all $\mathbf{x}$.

We finally remark that the above derivation is for given hyperparamter values θ in $f(\mathbf{x})|\mathbf{D}_n$. This is consistent with the Bayesian approach where the optimal values of θ are chosen from maximizing the likelihood function. However, the discussed paper used a different approach by sampling a distribution of θ and computed $G(\tilde{\mathbf{x}})$ as an average of the sampling. In the latter case, the above analysis should be likewise considered in a slightly different way, i.e., Eq. (16) should be considered as maximization of the multiplication of σ_2^2 from all samples of $\{\theta^{(i)}\}_{i=1}^s$:

$$\mathbf{x}_{n+1} = \mathrm{argmax}_{\tilde{\mathbf{x}}}\; \frac{1}{s}\sum_{i=1}^{s} G(\tilde{\mathbf{x}}, \theta^{(i)})$$
$$\equiv \mathrm{argmax}_{\tilde{\mathbf{x}}}\; \frac{1}{s}\sum_{i=1}^{s} \log\left(\frac{\sigma_1(\theta^{(i)})}{\sigma_2(\tilde{\mathbf{x}}, \theta^{(i)})}\right)$$
$$\equiv \mathrm{argmin}_{\tilde{\mathbf{x}}}\; \frac{1}{s}\sum_{i=1}^{s} \log\sigma_2(\tilde{\mathbf{x}}, \theta^{(i)}) \tag{19}$$
$$\equiv \mathrm{argmin}_{\tilde{\mathbf{x}}}\; \prod_{i=1}^{s} \sigma_2^2(\tilde{\mathbf{x}}, \theta^{(i)})$$

By using the Github code from the authors of Ref. [1] for their test cases, we have confirmed that the new results based on Eq. (15) are the same as the original results based on Eq. (5).

3 Analytical Computation of $G(\mathbf{x})$ for Arbitrary Input Distribution $p(\mathbf{x})$

In the computation of $G(\mathbf{x})$ in the form of Eq. (17), the most heavy computation involved is the integral $\int k_n(\tilde{\mathbf{x}}, \mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$ (which is prohibitive in high-dimensional problem if direct integration is performed). Following the discussed paper, the integral can be reformulated as

$$\int k_n(\tilde{\mathbf{x}}, \mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathcal{K}(\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}}, \mathbf{X}_n)\left(\mathbf{K}(\mathbf{X}_n, \mathbf{X}_n) + \sigma^2\mathbf{I}_n\right)^{-1}\mathcal{K}(\mathbf{X}_n) \tag{20}$$

where

$$\mathcal{K}(\mathbf{x}) = \int k(\mathbf{x}, \mathbf{x}')p(\mathbf{x}')\,\mathrm{d}\mathbf{x}' \tag{21}$$

$$k(\mathbf{x}, \mathbf{x}') = s^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T\Lambda^{-1}(\mathbf{x} - \mathbf{x}')\right) \tag{22}$$

with s and Λ involving hyperparameters of the kernel function (with either optimized values from training or selected values as in Ref. [1]).

The main computation is then Eq.(21), for which the authors of the discussed paper addressed the situation of uniform $p(\mathbf{x})$. To generalize the formulation to arbitrary $p(\mathbf{x})$, we can approximate $p(\mathbf{x})$ with the Gaussian mixture model (as an universal approximator of distributions [4]) :

$$p(\mathbf{x}) \approx \sum_{i=1}^{n_{GMM}} \alpha_i \mathcal{N}(\mathbf{x}; \mathbf{w}_i, \Sigma_i) \tag{23}$$

Equation (21) can then be formulated as:

$$\mathcal{K}(\mathbf{x}) \approx \sum_{i=1}^{n_{GMM}} \alpha_i \int k(\mathbf{x}, \mathbf{x}')\mathcal{N}(\mathbf{x}'; \mathbf{w}_i, \Sigma_i)\,\mathrm{d}\mathbf{x}'$$
$$= \sum_{i=1}^{n_{GMM}} \alpha_i |\Sigma_i\Lambda^{-1} + \mathrm{I}|^{-1/2} k(\mathbf{x}, \mathbf{w}_i; \Sigma_i + \Lambda) \tag{24}$$

which yields an analytical computation. In practice, the number of mixtures n_{GMM} is determined by the complexity of the input distributions, but any distribution of $p(\mathbf{x})$ can be approximated in such a way.

Finally, in computing $G(\tilde{\mathbf{x}})$ in the form of Eq. (19), computation of σ_1^2 in Eq. (8) is necessary. This can also be generalized for arbitrary $p(\mathbf{x})$ using the Gaussian mixture model as follows:

$$
\begin{aligned}
\sigma_1^2 &= \iint k_n(\mathbf{x}, \mathbf{x}')p(\mathbf{x})p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\,\mathrm{d}\mathbf{x} \\
&= \iint k(\mathbf{x}, \mathbf{x}')p(\mathbf{x})p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\,\mathrm{d}\mathbf{x} \\
&\quad - \int \mathbf{k}(\mathbf{x}, \mathbf{X}_n)p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\left(\mathbf{K}(\mathbf{X}_n, \mathbf{X}_n) + \sigma^2\mathbf{I}_n\right)^{-1}\int \mathbf{k}(\mathbf{X}_n, \mathbf{x}')p(\mathbf{x}')\,\mathrm{d}\mathbf{x}' \\
&= \sum_{i=1}^{n_{GMM}}\sum_{j=1}^{n_{GMM}} \alpha_i\alpha_j|\Lambda|^{1/2}|\Lambda + \Sigma_i + \Sigma_j|^{-1/2}k(\mathbf{w}_i, \mathbf{w}_j; \Lambda + \Sigma_i + \Sigma_j) \\
&\quad - \mathcal{K}(\mathbf{X}_n)^T\left(\mathbf{K}(\mathbf{X}_n, \mathbf{X}_n) + \sigma^2\mathbf{I}_n\right)^{-1}\mathcal{K}(\mathbf{X}_n)
\end{aligned}
\tag{25}
$$

Conflicts of Interest

There are no conflicts of interest.

Data Availability Statement

No data, models, or code were generated or used for this paper.

References

[1] Pandita, P., Bilionis, I., and Panchal, J., 2019, "Bayesian Optimal Design of Experiments for Inferring the Statistical Expectation of Expensive Black-Box Functions," ASME J. Mech. Des., **141**(10), p. 101404.

[2] Hu, Z., and Mahadevan, S., 2016, "Global Sensitivity Analysis-Enhanced Surrogate (gsas) Modeling for Reliability Analysis," Struct. Multidiscipl. Optim., **53**(3), pp. 501–521.

[3] Blanchard, A., and Sapsis, T., 2021, "Output-weighted Optimal Sampling for Bayesian Experimental Design and Uncertainty Quantification," SIAM/ASA J. Uncertainty Quantif., **9**(2), pp. 564–592.

[4] Goodfellow, I., Bengio, Y., and Courville, A., 2016, *Deep Learning*, MIT Press, Cambridge, MA.