Predictive Variance Reduction Search

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, Svetha Venkatesh
Centre of Pattern Recognition and Data Analytics (PRaDA), Deakin University
Email: v.nguyen@deakin.edu.au

Abstract

Predictive Entropy Search (PES) is popular and successful Bayesian optimization (BO) strategy. It finds a point to maximize the information gained about the optima of an unknown function. However, PES is computationally expensive and thus is not scalable to large-scale experiment when the number of observations and dimensions are large. We propose a new scheme - the Predictive Variance Reduction Search (PVRS) - to find the best "informative" point which reduces the predictive variance of the Gaussian process model at the optimum locations. We draw a connection between our PVRS to the existing PES. Our novel modification will be beneficial in three ways. First, PVRS directly reduces the uncertainty at the optimum representative points, like the PES. Second, PVRS can be computed cheaply in closed-form, unlike the approximations made in PES. Third, the PVRS is simple and easy to implement. As a result, the proposed PVRS gains huge speed up for scalable BO whilst showing comparable optimization efficiency.

1 Introduction

Bayesian optimization (BO) offers an elegant approach to optimize expensive black box functions by selecting the next experimental setting sequentially. BO approaches are receiving increasingly interest motivated by their diverse applicabilities [17, 16, 1, 12, 14, 6, 10]. Our goal is to find the global maximizer $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ over the bounded domain $\mathcal{X} \subset \mathcal{R}^d$. At iteration t, we select a point x_t and observe a possibly noisy function evaluation $y_t = f(\boldsymbol{x}_t) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ are i.i.d. Gaussian variables.

Since the form of f is unknown, we use Gaussian processes [15] to express a "belief" over the latent function. As data is observed, the posterior is updated. The selection process for the next point is guided by a surrogate function - also called the acquisition function - which is built from the posterior distribution of the GP. The advantage is that the acquisition function can be easily evaluated over the search space as opposed to the original expensive objective function.

There are existing different types of acquisition functions including EI [9, 5, 11], GP-UCB [18]. They balance the exploitation and exploration property by directly using the predictive mean and predictive variance of the GP model. Alternatively, entropy search (ES) [2] and predictive entropy search (PES) [4] aim to "know" more about the global optimum locations by evaluating the queried point. Although the information-theoretic approaches of ES and PES are successful recently in offering competitive performances, their computation is expensive and hindering their scalability. This scalability is important when we can collect data in large-scale (both in number of observations and the dimensions) or when the function evaluation is not so expensive. Scaling the PES is still an open research direction.

In this paper, we propose the Predictive Variance Reduction Search, an efficient view to the Predictive Entropy Search. We present a novel view for gaining the information about the global optimum locations through the Gaussian process predictive variance (i.e., the smaller the predictive variance at $\boldsymbol{x}^*$ is, the more we learn about $\boldsymbol{x}^*$). This view offers for the closed-form computation of the

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

predictive variance, instead of requiring a lot of computation for approximations in PES. As a result of closed-form and exact solution, our PVRS gains an order of magnitude faster and favourable performances against the PES.

2 Predictive Entropy Search

Motivated by the information-theoretic method [7], Entropy Search (ES) [2] is proposed to learn the information at the locations $\boldsymbol{x}^*$ by selecting the point that is expected to cause the largest reduction in entropy of the distribution $p\left(\boldsymbol{x}^* \mid \mathcal{D}_t\right)$ [19]. ES measures the expected information gain from querying an arbitrary point x and selects the point that offers the most information about the unknown $\boldsymbol{x}^*$. Let $\boldsymbol{x}^*$ be the maximum point, Entropy Search uses the information gain defined as follows:

$$\alpha^{\text{ES}}\left(x\right) = I(\{\boldsymbol{x}, y\}, \boldsymbol{x}^* \mid \mathcal{D}_t) = H\left[p(\boldsymbol{x}_* \mid \mathcal{D}_t)\right] - \mathop{\mathbb{E}}_{p(y|\mathcal{D}_t, \boldsymbol{x})}\left[H\left(p(\boldsymbol{x}_* \mid \mathcal{D}_t, \boldsymbol{x}, y)\right)\right]$$

where $H\left[p\left(x\right)\right] = -\int p\left(x\right) \log p\left(x\right) dx$ indicates the differential entropy and the expectation is over the distribution of the random variable $y \sim \mathcal{N}\left(\mu_n(\boldsymbol{x}), \sigma_n^2(\boldsymbol{x}) + \sigma_\epsilon^2\right)$. This function is not tractable for continuous search spaces $\mathcal{X}$ so approximations must be made. Typically, the approximation is done using discretization [19, 2]. This Entropy Search [2] is unfortunately $\mathcal{O}\left(M^4\right)$ where M is the number of discrete so-called representer points.

To overcome the problem of discretization above, the PES [4] modifies the ES by utilizing the symmetric function of the mutual information $I(\boldsymbol{x}, y; \boldsymbol{x}^*) = I(\boldsymbol{x}, \boldsymbol{x}^*; y)$ to obtain

$$\alpha^{\text{PES}}\left(x\right) = I(\{\boldsymbol{x}, \boldsymbol{x}^*\}, y \mid \mathcal{D}_t) = H\left[p(y \mid \mathcal{D}_t, \boldsymbol{x}^*)\right] - \mathop{\mathbb{E}}_{p(\boldsymbol{x}^*|\mathcal{D}_t)}\left[H\left(p(y \mid \mathcal{D}_t, \boldsymbol{x}, \boldsymbol{x}^*)\right)\right]$$

where $p(y \mid .)$ is the posterior predictive distribution given the observed data $\mathcal{D}_t$ and the location of the global maximizer $\boldsymbol{x}^*$ of f. The Predictive Entropy Search (PES) removes the need for a discretization. The expectation can be approximated via Monte Carlo with Thompson samples; and three simplifying assumptions are made to compute $H\left(p(y \mid D_t, \boldsymbol{x}, \boldsymbol{x}^*)\right)$. These assumptions are presented as three constraints including (1) $\nabla f(\boldsymbol{x}^*) = 0$, non-diagonal elements$\left[\nabla^2 f(\boldsymbol{x}^*)\right] = 0$, (2) $f\left(\boldsymbol{x}^*\right) \geq f(\boldsymbol{x}_i), \forall \boldsymbol{x}_i \in \mathcal{D}_t$ and (3) $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$. Then, [4] incorporates these constraints into a probabilistic form. Next, they use Expectation propagation (EP) [8] to approximate each non-Gaussian factor (from the constraint) with a Gaussian distribution. However, the EP approximation does not guarantee to converge and often leads to numerical instabilities, as pointed out in [3]. Although the PES has been shown to perform as well as or better than the discretized version without the unappealing quartic term, the complexity of PES is still high $\mathcal{O}\left(M\left[N + d + d(d-1)/2\right]^3\right)$ for each iteration given that M samples of $\boldsymbol{x}^*$ are precomputed with an additional complexity of $\mathcal{O}(MNV^2)$ where V is the random feature size.

Our new method, PVRS does not suffer from these pathologies of PES approximation. As a result, no ad-hoc approximations to the acquisition function and the individual factors are required. In addition, our PVRS is easier to implement than the ES and PES counterpart since computing the GP predictive variance is straight-forward [15].

3 Predictive Variance Reduction Search for Scalable Bayesian Optimization

We propose the Predictive Variance Reduction Search (PVRS) for Bayesian optimization. Given a collection of global optimum locations $\boldsymbol{x}^*$, PVRS finds the most "informative" point $\boldsymbol{x}$ such that by querying $\boldsymbol{x}$ we minimize the uncertainty at $\boldsymbol{x}^*$. For reducing the uncertainty at $\boldsymbol{x}^*$, we minimize the GP predictive variance at $\boldsymbol{x}^*$. By that way, we maximize our knowledge (a proxy to the information gain) about the global optimum locations.

Like the ES and PES, our PVRS aims to gain more information about the global optimum. For this information-theoretic purpose, our search is based on the predictive variance reduction at the global optimum location $\boldsymbol{x}^*$. Using GP predictive variance, our PVRS is much simpler to compute than the PES. PVRS can be computed in (exact) closed-form and run faster. The exact computation will ensure the robustness of the whole Bayesian optimization process and prevent from numerical instability and possible failure caused by approximations. They are our key advantages against ES and PES.

Algorithm 1 Bayesian Optimization using Predictive Variance Reduction Search.

Input: $\mathcal{D}_0 = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n_0}$, #iter T

1: **for** $t = 1$ to T **do**

2: Estimate the best GP hyper-parameter θ_t by maximizing GP marginal likelihood from $\mathcal{D}_{t-1}$.

3: Get a collection of points $\boldsymbol{S} = [\boldsymbol{s}_1, ...\boldsymbol{s}_M] \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}^*)$ from Thompson Sampling under θ_t.

4: Obtain $\boldsymbol{x}_t = \arg\min_{x \in \mathcal{X}} \sum_{m=1}^{M} \sigma_{t-1}(\boldsymbol{s}_m \mid \theta_t, \mathcal{D}_{t-1} \cup x)$.

5: Evaluate the function $y_t = f(\boldsymbol{x}_t)$ and augment the data $\mathcal{D}_t = \mathcal{D}_{t-1} \cup (\boldsymbol{x}_t, y_t)$.

6: **end for**

Output: $\boldsymbol{x}_{\max}, y_{\max}$

Since the global optimum location $\boldsymbol{x}^*$ is unknown, we make use of Thompson sampling (with Random Fourier features [13]) to draw samples $\boldsymbol{S} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_M] \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}^*)$ which express our belief about the location of $\boldsymbol{x}^*$. This Thompson sampling step is cheaper $\mathcal{O}(MNV^2)$ and is essential for all previous information-theoretic approaches [2, 4].

At the iteration t, we have the observations including $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}]$ and $Y = [y_1, ..., y_{t-1}]$. Given a collection of global optimum locations $\boldsymbol{S} = [\boldsymbol{s}_1, ...\boldsymbol{s}_M]$ where $\boldsymbol{s}_m \in \mathcal{R}^d$, we find the most "informative point" $\boldsymbol{x}_t$ so that if we evaluate this point, we learn as much as possible the information of the global optimum locations $\boldsymbol{S}$ without evaluating the function f.

$$\boldsymbol{x}_t = \underset{x \in \mathcal{X}}{\text{argmin}} \sum_{m=1}^{M} \sigma(\boldsymbol{s}_m \mid X \cup \boldsymbol{x}) \propto \underset{x \in \mathcal{X}}{\text{argmax}} \sum_{m=1}^{M} k(\boldsymbol{s}_m, X \cup \boldsymbol{x}) \mathbf{K}(X \cup \boldsymbol{x}, X \cup \boldsymbol{x})^{-1} k(\boldsymbol{s}_m, X \cup \boldsymbol{x})^T$$

The complexity for our PVRS is $\mathcal{O}(MN^2)$ and the sampling $[\boldsymbol{s}_1, ...\boldsymbol{s}_M]$ is $\mathcal{O}(MNV^2)$.

Connection to the information-theoretic approaches of PES.

We map our Predictive Variance Reduction Search view to the existing PES by using mutual information gain between the selected point and the global optimum location $I(\boldsymbol{x}, \boldsymbol{s}_m)$ conditioning on the existing observations D_t.

$$\begin{aligned}
\alpha_t^{PVRS}(\boldsymbol{x}) &= \min_{x \in X} \sum_{m=1}^{M} \sigma(\boldsymbol{s}_m \mid \mathcal{D}_t \cup \boldsymbol{x}) = \min_{x \in X} \frac{1}{M} \sum_{m=1}^{M} H\left[p(\boldsymbol{s}_m \mid \mathcal{D}_t \cup \boldsymbol{x})\right] \\
&= \max_{x \in X} \frac{1}{M} \sum_{m=1}^{M} \left\{ \underbrace{H\left[p(\boldsymbol{s}_m \mid \mathcal{D}_t)\right]}_{\text{const}} - H\left[p(\boldsymbol{s}_m \mid \mathcal{D}_t \cup \boldsymbol{x})\right] \right\} \\
&= \max_{x \in X} \frac{1}{M} \sum_{m=1}^{M} I(\boldsymbol{x}; \boldsymbol{s}_m \mid \mathcal{D}_t) \approx \max_{x \in X} \mathop{\mathbb{E}}_{\boldsymbol{s}_m} \left[I(\boldsymbol{x}; \boldsymbol{s}_m \mid \mathcal{D}_t)\right]
\end{aligned}$$

where in the first line we have utilized the property that $H(p(\boldsymbol{x})) = \frac{1}{2} \log\left(2\pi e \sigma^2(\boldsymbol{x})\right)$ for Gaussian distribution. Although $p(\boldsymbol{x}^*)$ may not Gaussian, we follow [4] to assume the Gaussian distribution to show the connection. We discuss the key difference from our information theoretic view to the existing ES, PES as follows. The mutual information of ES and PES $I(\boldsymbol{x}, y; \boldsymbol{s}_m \mid \mathcal{D}_t)$ [4, 2, 20] include the outcome y associated with the location of selection $\boldsymbol{x}$. Unlike the PES, our view is motivated by reducing the predictive GP variance which only depends on the location $\boldsymbol{x}$, not the outcome y (see [15]). This predictive variance property is crucial to get rid of many approximation made by ES and PES for scalability without sacrificing the optimization quality.

4 Experiments

We use squared exponential kernel with its kernel parameters optimized by maximizing the marginal likelihood. We compare the performance of PES [4] using Spearmint package. We use a fixed number of iterations $T = 20d$ and the number of initial point $n_0 = 3d$ where d is the dimension.

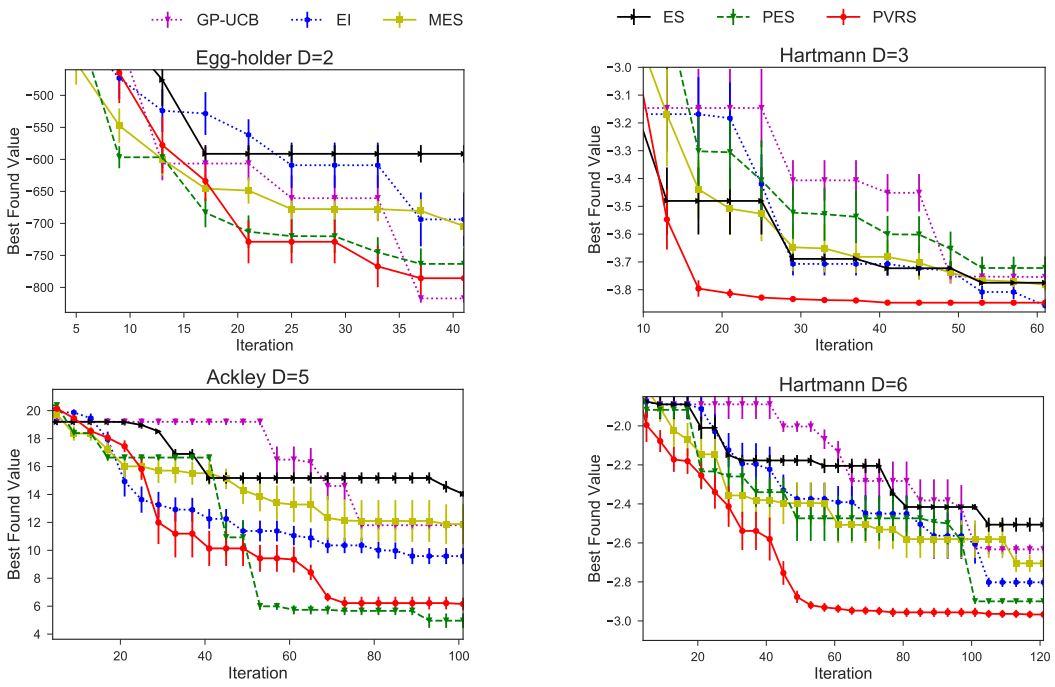

Figure 1: Best-found-value comparison on benchmark functions.

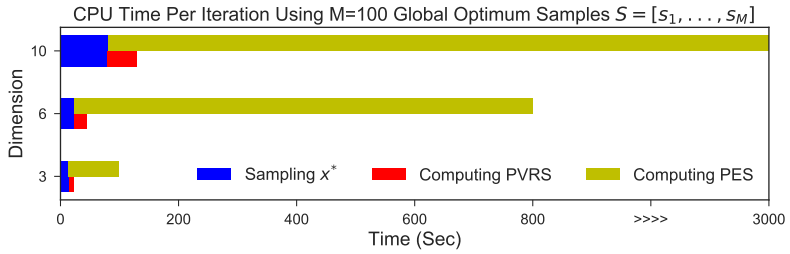

Figure 2: Computational time comparison with different dimensions. Both PVRS and PES require Thompson sampling to draw x^* samples. PVRS is significantly faster than PES in high dimension.

We demonstrate that our PVRS can reach closer to optimal values (minimum) using chosen benchmark functions. We report the best-found-value defined as $\min_{x_i \in \mathcal{D}_t} f(x_i)$ in Fig. 1. The optimization accuracy of PVRS is comparable or better than the PES. This is because the exact computation will ensure the robustness and prevent from numerical instability and possible failure caused by approximations - as in PES.

Different from our PVRS (as well as ES and PES), the Max-value Entropy Search (MES) [20] suggests to learn the information about the values y^*. However, we argue that learning the values y^* may not bring sufficient information for the optimal locations $\boldsymbol{x}^*$ (represented by $\boldsymbol{s}_m$) since our ultimate goal is to find the optimum location $\boldsymbol{x}^*$.

In addition, our PVRS gains significant speed up against the PES in Fig. 2. Let us consider the experiment when $d = 10$ dimensions, drawing $M = 100$ $\boldsymbol{x}^*$ samples from Thompson sampling takes 90 secs (blue bar). Then, PVRS takes 60 sec (red bar) while PES takes about 3000 sec (green bar) per iteration for optimization due to the complexity of $\mathcal{O}\left(M\left[N + d + d(d-1)/2\right]^3\right)$ [4]. The computational speed up is the key advantage of our approach due to a closed form solution and does not require many approximations as used in PES.

We are going to perform the real-world experimental designs and tuning of hyper-parameters for machine learning algorithm in the extended version of this paper.

References

[1] T. Dai Nguyen, S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, K. J. Deane, and P. G. Sanders. Cascade Bayesian optimization. In *Australasian Joint Conference on Artificial Intelligence*, pages 268–280. Springer, 2016.

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

[3] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International Conference on Machine Learning*, pages 1699–1707, 2015.

[4] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pages 918–926, 2014.

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

[6] C. Li, S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, and A. Shilton. High dimensional bayesian optimization using dropout. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2096–2102, 2017.

[7] D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[8] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc., 2002.

[9] J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

[10] V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. Bayesian optimization in weakly specified search space. In *IEEE 17th International Conference on Data Mining (ICDM)*, 2017.

[11] V. Nguyen, S. Gupta, S. Rana, C. Li, and S. Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Proceedings of The 9th Asian Conference on Machine Learning (ACML)*, pages 279–294, 2017.

[12] V. Nguyen, S. Rana, S. Gupta, C. Li, and S. Venkatesh. Budgeted batch Bayesian optimization. In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 1107–1112, 2016.

[13] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

[14] S. Rana, C. Li, S. Gupta, V. Nguyen, and S. Venkatesh. High dimensional Bayesian optimization with elastic gaussian process. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2883–2891, 2017.

[15] C. E. Rasmussen. Gaussian processes for machine learning. 2006.

[16] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

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

[18] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1015–1022, 2010.

[19] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

[20] Z. Wang and S. Jegelka. Max-value entropy search for efficient bayesian optimization. *arXiv preprint arXiv:1703.01968*, 2017.