Think Globally, Act Locally: a Local Strategy for Bayesian Optimization

Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, Svetha Venkatesh
Centre of Pattern Recognition and Data Analytic (PraDa), Deakin University
Email: {v.nguyen,sunil.gupta,santu.rana,cheng.l,svetha.venkatesh}@deakin.edu.au

Abstract

Bayesian optimization (BO) is a sample-efficient method for improving the performance of machine learning algorithms and laboratory experiments. We exploit the local property in BO to develop a new acquisition function, the expected local improvement (ELI) as an alternative to Expected Improvement (EI), aiming to address two underlying issues. First, we reduce the flatland issue in high dimension and second we allow greater explorative choices for batch BO unlike the existing strategies. We derive the convergence analysis using simple regret bound. We further demonstrate that the proposed strategy gains substantial performance improvement over the state-of-the-art baselines using the benchmark functions and real experiments on sequential and batch BO.

1 Introduction

Bayesian optimization (BO) offers an elegant alternative to optimize expensive black box functions by selecting the next experimental setting sequentially. The field is receiving increasing interest motivated by its diverse applicability [16, 15, 19]. BO uses a Gaussian Process [13] to express a "belief" over all possible objective functions. As data is observed, the posterior is updated and is then used to determine the next experimental setting to evaluate. 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. The advantage is that the acquisition function can be easily evaluated over the search space as opposed to the original expensive objective function. Alternative to the *sequential BO* which recommends one setting per iteration, the *batch BO* approaches [4, 3, 12] also gain increasing attention that recommend multiple settings per iteration in situations where parallel experimentation is possible.

The crucial step of finding the global maximum of the acquisition functions, particularly when estimated through few observations, remains challenging. This is because the acquisition function generally has a few sharp peaks marooned in mostly flat regions, especially in high dimension functions. Such flatlands problem brings challenges to most optimizers [10, 7]. The failure of this step can seriously compromise BO.

This paper explores a new way to address the above problem and guide the choice of the next experimental setting. For robust estimation, any strategy to create local "bumps" in the flatlands of the search space will be useful as there is at least some information at the "bumps" that deserve to be evaluated, in contrast to no information in flatlands. One approach to creating "bumps" is to construct the surrogate function using local information. That is, instead of finding the best point globally, we find a point which is the best around its neighbors. This strategy creates local "bumps" at different locations in the acquisition function because there is a higher chance to find a point that is better in a local neighborhood than across the whole domain considered. By looking at the promising candidates locally, we encourage exploration at more locations. This intuition fits in a broader perspective of "think globally, act locally" [8, 5], a widely used strategy in planning, environment, mathematics and

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

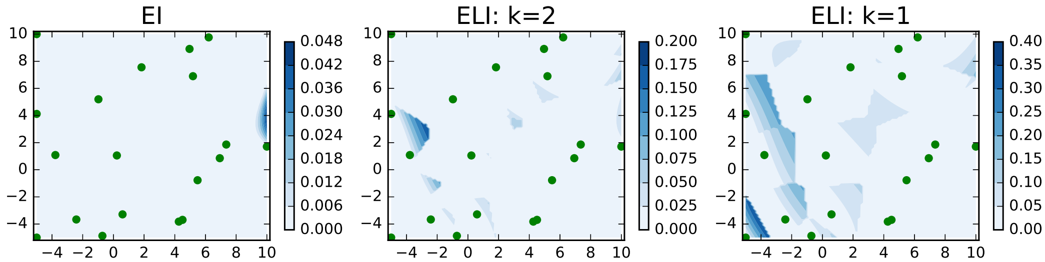

Figure 1: The acquisition functions of EI and ELI(s) on branin function. The dark regions indicate the peaks. The flatland effect can be seen in the case of EI (Left) with only one peak where most of the regions are flat with zero value. Although the flatland issue is not the case for 2D, it is serious in high dimensions. Decreasing the number of nearest neighbor k will result in more locality and more peaks that ELI$(k = 2)$ gets 8 peaks and ELI$(k = 1)$ gets 10 peaks at different locations. We may not find a point that satisfies global improvement (as in EI), but we can always find points that improve locally. This local property makes ELI greatly beneficial for batch BO which finds multiple settings at each iteration to evaluate in parallel.

business. A further point is that our method is particularly beneficial for batch BO that recommends multiple settings per iteration corresponding to B peaks of the acquisition function. That is because having many local "bumps" is likely to offer sufficiently many locally best candidates for parallel evaluations in batch BO. In contrast, existing acquisition functions may fall short of offering B peaks.

Our paper is the first to investigate the idea of *think globally, act locally* into the Bayesian optimization framework. We materialize this by presenting the expected local improvement (ELI). In contrast to the expected improvement (EI) [11] that finds improvements over the *global* best found value so far, ELI finds the points that get the highest *local* improvement over its neighbors. We devise algorithms for both the batch and sequential settings based on this new acquisition function. We derive an upper bound on the simple regret to ensure the convergence property of the proposed strategy. Finally, we conduct an extensive set of experiments in both sequential and batch Bayesian optimization settings to highlight the advantages of the local strategy.

2 Limitation of the existing acquisition functions

Most of the existing acquisition functions [9, 11, 17, 6, 18] look the optimization at the global perspective. For example, POI and EI improve over the current global best value y^{best}. Similarly, the PES finds the location that greatly reduces the (predictive) entropy of the function globally. GP-UCB is also balancing the predictive mean and variance globally.

We focus on the expected improvement (EI) [11] which has been widely used as the default choice in popular BO packages, such as Spearmint [16]. There has been a concern about flatland effects in optimizing the acquisition functions which tend to produce a few peaks in mostly flat regions. This effect may result in inaccurate and unstable estimation. In addition, since the function is often multi-modal that we may not find a suitable point that satisfies global improvement. This issue is specifically critical for the batch BO setting where we are seeking a batch of B promising points at each iteration for parallel evaluations. These points will be often selected from the peaks of the acquisition function [4, 2]. However, the global perspective used in the existing acquisition functions can not find enough B points to satisfy the criteria globally (cf. Fig. 1). As such, the greedy sequential peak suppression for batch BO will recommend either redundant points around the local optimum or dummy points if the real peaks in the acquisition function is less than the required batch size of B that we are looking for. This may waste time and resources to evaluate at these unnecessary points that violates the fundamental of Bayesian optimization to keep the number of evaluations as low as possible.

3 Think globally, act locally

To materialize the strategy of *think globally, act locally*, we present the expected local improvement (ELI), an alternative acquisition function to the well-known expected improvement (EI) [11]. Next, we present an upper bound on the simple regret to provide a guarantee on the convergence property. Finally, we present the batch BO setting using ELI as the underlying acquisition function.

3.1 Expected Local Improvement (ELI)

We present the *expected local improvement* (ELI) strategy to find a point that has the highest improvement over its local neighbors. Although the existing view of expected improvement (EI) [11] considers improvement globally, we suggest optimization should be treated locally to avoid the saddle point effects and to identify the promising candidates at multiple locations for batch Bayesian optimization. Let $\mathcal{D}_t = \left\{x_i \in \mathcal{R}^D, y_i \in \mathcal{R}\right\}_{i=1}^{t}$ be the observation set including the feature x_i and the outcome $y_i = f\left(x_i\right) + \epsilon_i$ where $f(.)$ is the black-box function. Let us denote the neighboring observations to x defined by a radius v as $[x] \triangleq \{x_i \in \mathcal{D}_t \mid ||x_i - x|| \leq v\}$, we define the local improvement function $I_t^{\text{ELI}}\left(x\right) = \max\left\{0, f(x) - f^{+}\left([x]\right)\right\}$ where $f^{+}\left([x]\right) \triangleq \max_{x_i \in [x]} f\left(x_i\right)$. Motivated by the EI, the expected local improvement (ELI) is then defined as $\alpha_t^{\text{ELI}}\left(x\right) = \mathbb{E}\left[I_t^{\text{ELI}}\left(x\right)\right]$. Explicitly, let denote $z = \frac{\mu_{t-1}(x) - f^{+}([x])}{\sigma_{t-1}(x)}$, we obtain the acquisition function as follows (refers the supplement for the derivation)

$$\alpha_t^{\text{ELI}}\left(x\right) = \sigma_{t-1}\left(x\right)\phi\left(z\right) + \left[\mu_{t-1}\left(x\right) - f^{+}\left([x]\right)\right]\Phi\left(z\right) \tag{1}$$

where ϕ and Φ are the standard normal pdf and cdf.

Although our formulation is a slight modification of that introduced by [11], the idea of making use of the local strategy for global optimization is novel - to the best of our knowledge.

Bound on simple regret for ELI

Our theoretical analysis uses the simple regret to bound the convergence, instead of the cumulative regret commonly used in literature, due to the explorative property of the proposed acquisition function that tends to have high cumulative regret. We assume that the noise process ϵ_t is sub-Gaussian, and the function f is smooth according to the reproducing kernel Hilbert space (RKHS) associated with the GP kernel. We follow [17] to define the maximum information gain γ_t. We refer the interested reader to the supplement for the theoretical derivation.

Theorem 1. *Given the maximum information gain γ_t, a Lipschitz constant L, assuming $\beta_t = 2||f||_k^2 + 300\gamma_t \ln^3\left(\frac{t}{\delta}\right)$ and a constant $Q = \frac{\tau\left(\sqrt{\beta_t}\right)}{\tau\left(-\sqrt{\beta_t}\right)}$, with probability at least $1-\delta$, the simple regret obeys the following rate $s_t \leq Q \times \tau\left(\sqrt{\beta_t} + \frac{1}{L \times t^2}\right)$.*

We obtain the smaller simple regret s_t with increasing t. The radius v_t plays a critical role in defining the neighborhood. The small v_t can make the cell empty (no data point lies within the neighborhood defined by v_t). On the contrary, the large v_t can diminish the idea of locality. In practice, we observe that some regions may have no observation, i.e., $[x] = \emptyset$. Therefore, instead of defining a fixed radius v_t, we utilize the kNN algorithm to find $k \in [1, N]$ closest neighbors $[x]$ from $\mathcal{D}_t$, then we define them as the k neighbors to x. This heuristic way ensures every location will have k observations as the neighbors and we found that it works well in practice. By fixing k, the radius v_t is non-increasing and tends to decrease at every iteration since we add more data points to $\mathcal{D}_t$, thus enable the convergence of the simple regret.

3.2 ELI for Batch Bayesian Optimization

Next, we consider batch BO setting using the proposed ELI where parallel evaluations are available. Formally, we identify a batch of B points at each iteration $\boldsymbol{X}_t = [\boldsymbol{x}_{t,1}, ..., \boldsymbol{x}_{t,B}] = \underset{\boldsymbol{x} \in \mathcal{X}}{\text{argmax}}\, \alpha_t^{\text{ELI}}\left(\boldsymbol{x}\right)$.

We aim to highlight the usefulness of ELI that can offer multiple local peaks at different locations, beneficial for batch BO. Due to the simplicity and robustness, we select to use the greedy peak

Approaches	POI	EST	UCB	PES	EI	**ELI**
Sincos 1D	**-8.93**±**2**	-8.3±2	-7.87±2	-7.59±2	-8.41±1.9	-8.89±2
Branin 2D	1.322±1	1.1±.6	2.98±2.3	3.17±2	1.42±.9	**0.92**±**.6**
Hartman 3D	-3.60±.3	-3.5±.4	-3.62±.3	-3.33±.3	-3.62±3	**-3.71**±**.2**
Ackley 5D	19.37±1	19±1	15.3±3.8	**11.11**±**2**	11.30±4	12.02±5
Alpine2 5D	-40.1±15	-13±9	-23.48±7	**-62**±**28**	-30.75±18	-43.8±25
Hartman 6D	-2.84±.3	-2.5±.3	-2.61±.2	-2.85±.1	**-2.91**±**.1**	**-2.91**±**.1**
Alpine2 10D	-1.8k±77	-139±74	-572±431	-858±1k	-447±410	**-3486**±**1k**
gSobol 10D	12k±5k	9k±8k	550±387	3k±562	297±234	**154**±**122**

Table 1: Best-found-value comparison on the benchmark functions.

Approaches	BatchPOI	BatchUCB	BatchEI	Batch **ELI**(k=1)	Batch**ELI**(k=3)
Ackley 5D	12.95±2.9	13.58±1.5	8.50±2.29	**6.558**±**1.6**	***7.001***±***1.5***
Alpine2 5D	***-74.99***±***27***	-38.13±14	-39.4±17	-70.33±23	**-77.7**±**32**
Hartman 6D	**-3.02**±**0.02**	-2.74±0.09	-2.92±0.07	**-3.02**±**.04**	**-3.02**±**0.03**
Alpine2 10D	-4625±1k	-2722±1k	-2432±1k	***-4907***±***933***	**-5792**±**1k**
gSobol 10D	2509±2351	**169.7**±**120**	*182*±*127*	188.3±139	286±252

Table 2: Best-found-value comparison in batch Bayesian optimization setting.

suppression approach by sequentially visiting all the maxima of the acquisition function [3, 2, 4]. Our proposed ELI can also be applied to other existing batch BO methods (e.g., [14, 12]).

4 Experiments

Experimental Setting We use squared exponential kernels $k(x, x') = \exp\left(-l \times ||x - x'||^2\right)$ where l is set to the dimension size. The performance of the algorithms is compared for a fixed number of iterations $T = 10 \times D$ and the initialization point $n_0 = 3$. The number of nearest neighbor in our approach is set default as $k = 3$. The UCB parameter is set as $\beta_t = 2$ as used in [4]. We optimize the acquisition function using DIRECT. We use the Spearmint toolbox for PES [6].

We demonstrate that our acquisition functions can reach closer to optimal values (minimum) in both sequential and batch settings using chosen benchmark functions. Further experiments in real-world experimental designs are available in the supplement.

4.1 Sequential Bayesian optimization

We report the best-found-value (BFV) in Table 1. The BFV at iteration t, defined as $\max_{x_i \in \mathcal{D}_t} f(x_i)$, can be seen as the reverse version of the simple regret $s_t = f(x^*) - \max_{x_i \in \mathcal{D}_t} f(x_i)$. ELI is more robust in identifying the best settings (especially for high dimension functions) at each iteration because we can reduce flat region issues [1] happened in the existing acquisition functions. TS and POI have higher tendency to exploit aggressively on high dimension functions [15] and thus generally perform poorer than the others.

4.2 Batch Bayesian Optimization

We further demonstrate the efficiency of the local principle for batch BO. In particular, we employ the greedy approach of peak suppression [4, 3] for identifying the batch $B = 3$ of points sequentially. We compare the performance of the batch BO using the best-found-value on the benchmark functions in Table. 2. We show that our ELI is robust in outperforming the baselines of POI, GP-UCB and EI in batch BO w.r.t. different choices of $k = 1$ and 3. The existing acquisition functions, using global strategy, can not find enough B regions that satisfy the improvement globally. As a result, batch BO may start selecting redundant points after all of the real peaks are exhausted. In contrast, ELI can produce more peaks and thus is suitable for batch Bayesian optimization. In some situations, where batch BO seeks a large number of peaks B for evaluating parallelly, we can reduce the neighborhood parameter, e.g., $k = 1$ to encourage more number of peaks (see Fig. 1) while the existing acquisition functions are unable to do so.

References

[1] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[2] T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *The Journal of Machine Learning Research*, 15(1):3873–3923, 2014.

[3] D. Ginsbourger, R. Le Riche, and L. Carraro. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. 2008.

[4] J. González, Z. Dai, P. Hennig, and N. D. Lawrence. Batch bayesian optimization via local penalization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 648–657, 2016.

[5] L. Hennighausen and G. W. Robinson. Think globally, act locally: the making of a mouse mammary gland. *Genes & development*, 12(4):449–455, 1998.

[6] 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.

[7] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

[8] A. G. Kefalas. Think globally, act locally. *Thunderbird International Business Review*, 40(6):547–562, 1998.

[9] H. J. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

[10] K. I. McKinnon. Convergence of the nelder–mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9(1):148–158, 1998.

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

[12] V. Nguyen, S. Rana, S. K. Gupta, C. Li, and S. Venkatesh. Budgeted batch bayesian optimization. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Spain, 2016.

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

[14] A. Shah and Z. Ghahramani. Parallel predictive entropy search for batch global optimization of expensive objective functions. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015.

[15] 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.

[16] 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.

[17] 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-10)*, pages 1015–1022, 2010.

[18] Z. Wang, B. Zhou, and S. Jegelka. Optimization as estimation with gaussian processes in bandit settings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1022–1031, 2016.

[19] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, N. Freitas, et al. Bayesian optimization in high dimensions via random embeddings. AAAI Press/International Joint Conferences on Artificial Intelligence, 2013.