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

Yu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, Svetla Venkatesh
Centre of Pattern Recognition and Data Analytics (PraDa), Deakin University
Email: {v.nguyen,sunil.gupta,santu.rana,cheng.l,svetla.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
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) \[^1\] 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 \[^9\] \[^1\] \[^17\] \[^6\] \[^18\] look the optimization at the global perspective. For example, POI and EI improve over the current global best value $y_{\text{best}}$. 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) \[^1\] 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). 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) 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 \( D_t = \{ x_i \in \mathcal{R}^D, y_i \in \mathcal{R} \}_{i=1}^t \) be the observation set including the feature \( x_i \) and the outcome \( y_i = f(x_i) + \epsilon_i \) where \( f(.) \) is the black-box function. Let us denote the neighboring observations to \( x \) defined by a radius \( v \) as \( [x] \overset{\Delta}{=} \{ x_i \in D_t \mid ||x_i - x|| < v \} \), we define the local improvement function \( I_t^{ELI}(x) = \max \{ 0, f(x) - f^+([x]) \} \) where \( f^+([x]) \overset{\Delta}{=} \max_{x_i \in [x]} f(x_i) \).

Motivated by the EI, the expected local improvement (ELI) is then defined as \( \alpha_t^{ELI}(x) = \mathbb{E} \left[ I_t^{ELI}(x) \right] \).

Explicitly, let \( z = \frac{\mu_{t-1}(x) - f^+([x])}{\sigma_{t-1}(x)} \), we obtain the acquisition function as follows (refers the supplement for the derivation)

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

where \( \phi \) and \( \Phi \) 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 \( \epsilon_i \) 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 \( \gamma_t \). We refer the interested reader to the supplement for the theoretical derivation.

**Theorem 1.** Given the maximum information gain \( \gamma_t \), a Lipschitz constant \( L \), assuming \( \beta_t = 2||f||^2_2 + 300\gamma_t \ln^3 \left( \frac{E}{\delta} \right) \) and a constant \( Q = \frac{\tau(\sqrt{\tau})}{\tau(-\sqrt{\tau})} \), with probability at least \( 1 - \delta \), the simple regret obeys the following rate \( s_t \leq Q \times \tau \left( \sqrt{\frac{1}{t}} + \frac{1}{t} \right) \).

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 \( D_t \), 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 \( D_t \), 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 \( X_t = [x_{t,1}, ..., x_{t,B}] = \arg\max_{x \in X} \alpha_t^{ELI}(x) \).

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
We compare the performance of the batch BO using the best-found-value on the benchmark functions. We further demonstrate the efficiency of the local principle for batch BO. In particular, we employ the greedy approach of peak suppression \[4,5,2\] for identifying the batch \(B\) of points sequentially. 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 \(k(x, x') = \exp(-l \times \|x - x'\|^2)\) 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 neighbors 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 \(\max_{x_i \in D_t} f(x_i)\), can be seen as the reverse version of the simple regret \(s_t = f(x^*) - \max_{x_i \in D_t} f(x_i)\). 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 \([13]\) 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 \([4,5,2]\) 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, cannot 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.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>POI</th>
<th>EST</th>
<th>UCB</th>
<th>PES</th>
<th>EI</th>
<th>ELI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sincos 1D</td>
<td>-8.93±2</td>
<td>-8.3±2</td>
<td>-7.87±2</td>
<td>-7.59±2</td>
<td>-8.41±1.9</td>
<td>-8.89±2</td>
</tr>
<tr>
<td>Brann 2D</td>
<td>1.32±1</td>
<td>1.1±6</td>
<td>2.98±2.3</td>
<td>3.17±2</td>
<td>1.42±2.9</td>
<td>0.92±2.6</td>
</tr>
<tr>
<td>Hartman 3D</td>
<td>-3.60±3</td>
<td>-3.54±3</td>
<td>-3.62±3</td>
<td>-3.33±3</td>
<td>-3.62±3</td>
<td>-3.71±2</td>
</tr>
<tr>
<td>Ackley 5D</td>
<td>19.37±1</td>
<td>19±1</td>
<td>15.3±3.8</td>
<td>11.1±2</td>
<td>11.3±10</td>
<td>12±0.5</td>
</tr>
<tr>
<td>Alpine 2D</td>
<td>-40.1±15</td>
<td>-13.9±9</td>
<td>-23.48±7</td>
<td>-62±28</td>
<td>-30.75±18</td>
<td>-43.8±25</td>
</tr>
<tr>
<td>Hartman 6D</td>
<td>-2.84±3</td>
<td>-2.5±3</td>
<td>-2.61±2</td>
<td>-2.85±1</td>
<td>-2.91±1</td>
<td>-2.91±1</td>
</tr>
<tr>
<td>Alpine 10D</td>
<td>-1.8±77</td>
<td>-139±74</td>
<td>-572±431</td>
<td>-858±1k</td>
<td>-447±410</td>
<td>-4386±1k</td>
</tr>
<tr>
<td>gSobol 10D</td>
<td>12k±5k</td>
<td>9k±8k</td>
<td>550±387</td>
<td>3k±562</td>
<td>297±234</td>
<td>154±122</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Approaches</th>
<th>BatchPOI</th>
<th>BatchUCB</th>
<th>BatchEI</th>
<th>Batch ELI(k=1)</th>
<th>Batch ELI(k=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ackley 5D</td>
<td>12.95±2.9</td>
<td>13.58±1.5</td>
<td>8.50±2.29</td>
<td>6.55±1.6</td>
<td>7.00±1.5</td>
</tr>
<tr>
<td>Alpine 5D</td>
<td>-74.99±27</td>
<td>-38.13±14</td>
<td>-39.4±17</td>
<td>-70.33±23</td>
<td>-77.7±32</td>
</tr>
<tr>
<td>Hartman 6D</td>
<td>-3.02±0.02</td>
<td>-2.74±0.09</td>
<td>-2.92±0.07</td>
<td>-3.02±0.04</td>
<td>-3.02±0.03</td>
</tr>
<tr>
<td>Alpine 10D</td>
<td>-4625±1k</td>
<td>-2722±1k</td>
<td>-2432±1k</td>
<td>-4907±933</td>
<td>-5792±1k</td>
</tr>
<tr>
<td>gSobol 10D</td>
<td>2509±2351</td>
<td>169.7±120</td>
<td>182±127</td>
<td>188.3±139</td>
<td>286±252</td>
</tr>
</tbody>
</table>

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


