Max-value Entropy Search for Efficient Bayesian Optimization (Appendix)

Zi Wang [1] **Stefanie Jegelka** [1]

1. Related work

Our work is largely inspired by the entropy search (ES) methods (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014), which established the information-theoretic view of Bayesian optimization by evaluating the inputs that are most informative to the $\arg\max$ of the function we are optimizing.

Our work is also closely related to probability of improvement (PI) (Kushner, 1964), expected improvement (EI) (Močkus, 1974), and the BO algorithms using upper confidence bound to direct the search (Auer, 2002; Kawaguchi et al., 2015; 2016), such as GP-UCB (Srinivas et al., 2010). In (Wang et al., 2016), it was pointed out that GP-UCB and PI are closely related by exchanging the parameters. Indeed, all these algorithms build in the heuristic that the next evaluation point needs to be likely to achieve the maximum function value or have high probability of improving the current evaluations, which in turn, may also give more information on the function optima like how ES methods queries. These connections become clear as stated in Section 3.1 of our paper.

Finding these points that may have good values in high dimensional space is, however, very challenging. In the past, high dimensional BO algorithms were developed under various assumptions such as the existence of a lower dimensional function structure (Djolonga et al., 2013; Wang et al., 2013), or an additive function structure where each component is only active on a lower manifold of the space (Li et al., 2016; Kandasamy et al., 2015). In this work, we show that our method also works well in high dimensions with the additive assumption made in (Kandasamy et al., 2015).

2. Using the Gumbel distribution to sample y_*

To sample the function maximum y_*, our first approach is to approximate the distribution for y^* and then sample from that distribution. We use independent Gaussians to approximate the correlated $f(\boldsymbol{x}), \forall \boldsymbol{x} \in \hat{\mathfrak{X}}$ where $\hat{\mathfrak{X}}$ is a discretization of the input search space $\mathfrak{X}$ (unless $\mathfrak{X}$ is discrete, in which case $\hat{\mathfrak{X}} = \mathfrak{X}$). A similar approach was adopted in (Wang et al., 2016). We can show that by assuming $\{f(\boldsymbol{x})\}_{\boldsymbol{x}\in\hat{\mathfrak{X}}}$, our approximated distribution gives a distribution for an upperbound on $f(\boldsymbol{x})$.

Lemma 2.1 (Slepian's Comparison Lemma (Slepian, 1962; Massart, 2007))**.** *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ be two multivariate Gaussian random vectors with the same mean and variance, such that*

$$\mathbb{E}[\boldsymbol{v}_i \boldsymbol{v}_j] \leq \mathbb{E}[\boldsymbol{u}_i \boldsymbol{u}_j], \forall i, j.$$

Then for every y

$$\Pr[\sup_{i\in[1,n]} \boldsymbol{v}_i \leq y] \leq \Pr[\sup_{i\in[1,n]} \boldsymbol{u}_i \leq y].$$

By the Slepian's lemma, if the covariance $k_t(\boldsymbol{x}, \boldsymbol{x}') \geq 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \hat{\mathfrak{X}}$, using the independent assumption with give us a distribution on the upperbound $\hat{y}_*$ of $f(\boldsymbol{x})$, $\Pr[\hat{y}_* < y] = \prod_{\boldsymbol{x}\in\hat{\mathfrak{X}}} \Psi(\gamma_y(\boldsymbol{x})))$.

We then use the Gumbel distribution to approximate the distribution for the maximum of the function values for $\hat{\mathfrak{X}}$, $\Pr[\hat{y}_* < y] = \prod_{\boldsymbol{x}\in\hat{\mathfrak{X}}} \Psi(\gamma_y(\boldsymbol{x})))$. If for all $\boldsymbol{x} \in \hat{\mathfrak{X}}$, $f(\boldsymbol{x})$ have the same mean and variance, the Gumbel approximation is in fact asymptotically correct by the Fisher-Tippett-Gnedenko theorem (Fisher, 1930).

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Massachusetts, USA. Correspondence to: Zi Wang <ziw@csail.mit.edu>, Stefanie Jegelka <stefje@csail.mit.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

Theorem 2.2 (The Fisher-Tippett-Gnedenko Theorem (Fisher, 1930)). *Let $\{v_i\}_{i=1}^{\infty}$ be a sequence of independent and identically-distributed random variables, and $M_n = \max_{1<i<n} v_i$. If there exist constants $a_n > 0, b_n \in \mathbb{R}$ and a non degenerate distribution function F such that $\lim_{n\to\infty} \Pr(\frac{M_n - b_n}{a_n} \leq x) = F(x)$, then the limit distribution F belongs to either the Gumbel, the Fréchet or the Weibull family.*

In particular, for i.i.d. Gaussians, the limit distribution of the maximum of them belongs to the Gumbel distribution (Von Mises, 1936). Though the Fisher-Tippett-Gnedenko theorem does not hold for independent and differently distributed Gaussians, in practice we still find it useful in approximating $\Pr[\hat{y}_* < y]$. In Figure 1, we show an example of the result of the approximation for the distribution of the maximum of $f(\boldsymbol{x}) \sim GP(\mu_t, k_t) \forall \boldsymbol{x} \in \hat{\mathfrak{X}}$ given 50 observed data points randomly selected from a function sample from a GP with 0 mean and Gaussian kernel.

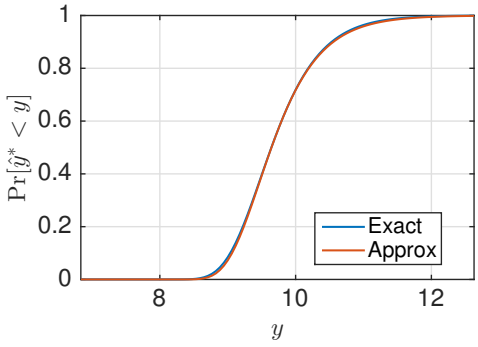

Figure 1. An example of approximating the cumulative probability of the maximum of independent differently distributed Gaussians $\Pr[\hat{y}_* < y]$ (Exact) with a Gumbel distribution $\mathcal{G}(a, b)$ (Approx) via percentile matching.

3. Regret bounds

Based on the connection of MES to EST, we show the bound on the learning regret for MES with a point estimate for $\alpha(x)$.

Theorem 3.1. *Let F be the cumulative probability distribution for the maximum of any function f sampled from $GP(\mu, k)$ over the compact search space $\mathfrak{X} \subset R^d$, where $k(\boldsymbol{x}, \boldsymbol{x}') \leq 1, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathfrak{X}$. Let $f_* = \max_{\boldsymbol{x}\in\mathfrak{X}} f(\boldsymbol{x})$ and $w = F(f_*) \in (0, 1)$, and assume the observation noise is iid $\mathcal{N}(0, \sigma)$. If in each iteration t, the query point is chosen as $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathfrak{X}} \gamma_{y_*^t}(\boldsymbol{x}) \frac{\psi(\gamma_{y_*^t}(\boldsymbol{x}))}{2\Psi(\gamma_{y_*^t}(\boldsymbol{x}))} - \log(\Psi(\gamma_{y_*^t}(\boldsymbol{x})))$, where $\gamma_{y_*^t}(\boldsymbol{x}) = \frac{y_*^t - \mu_t(\boldsymbol{x})}{\sigma_t(\boldsymbol{x})}$ and y_*^t is drawn from F, then with probability at least $1 - \delta$, in $T' = \sum_{i=1}^{T} \log_w \frac{\delta}{2\pi_i}$ number of iterations, the simple regret satisfies*

$$r_{T'} \leq \sqrt{\frac{C\rho_T}{T}}(\nu_{t^*} + \zeta_T) \tag{1}$$

where $C = 2/\log(1 + \sigma^{-2})$ and $\zeta_T = (2\log(\frac{\pi_T}{\delta}))^{\frac{1}{2}}$; π satisfies $\sum_{i=1}^{T} \pi_i^{-1} \leq 1$ and $\pi_t > 0$, and $t^* = \arg\max_t \nu_t$ with $\nu_t \triangleq \min_{\boldsymbol{x}\in\mathfrak{X}, y_*^t > f_*} \gamma_{y_*^t}(\boldsymbol{x})$, and ρ_T is the maximum information gain of at most T selected points.

Before we continue to the proof, notice that if the function upper bound $\hat{y}_*$ is sampled using the approach described in Section 3.1 and $k_t(\boldsymbol{x}, \boldsymbol{x}') \geq 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \hat{\mathfrak{X}}$, we may still get the regret guarantee by setting $y_* = \hat{y}_*$ (or $y_* = \hat{y}_* + \epsilon L$ if $\mathfrak{X}$ is continuous) since $\Pr[\max_{\hat{\mathfrak{X}}} \leq y] \geq \Pr[\hat{y}_* < y]$. Moreover, Theorem 3.1 assumes y_* is sampled from a universal maximum distribution of functions from $GP(\mu, k)$, but it is not hard to see that if we have a distribution of maximums adapted from $GP(\mu_t, k_t)$, we can still get the same regret bound by setting $T' = \sum_{i=1}^{T} \log_{w_i} \frac{\delta}{2\pi_i}$, where $w_i = F_i(f_*)$ and F_i corresponds to the maximum distribution at an iteration where $y_* > f_*$. Next we introduce a few lemmas and then prove Theorem 3.1.

Lemma 3.2 (Lemma 3.2 in (Wang et al., 2016)). *Pick $\delta \in (0, 1)$ and set $\zeta_t = (2\log(\frac{\pi_t}{2\delta}))^{\frac{1}{2}}$, where $\sum_{t=1}^{T} \pi_t^{-1} \leq 1$, $\pi_t > 0$. Then, it holds that $\Pr[\mu_{t-1}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t) \leq \zeta_t \sigma_{t-1}(\boldsymbol{x}_t), \forall t \in [1, T]] \geq 1 - \delta$.*

Lemma 3.3 (Lemma 3.3 in (Wang et al., 2016)). *If $\mu_{t-1}(\boldsymbol{x}_t) - f(\boldsymbol{x}_t) \leq \zeta_t \sigma_{t-1}(\boldsymbol{x}_t)$, the regret at time step t is upper bounded as $\tilde{r}_t \leq (\nu_t + \zeta_t)\sigma_{t-1}(\boldsymbol{x}_t)$, where $\nu_t \triangleq \min_{\boldsymbol{x}\in\mathfrak{X}} \frac{\hat{m}_t - \mu_{t-1}(\boldsymbol{x})}{\sigma_{t-1}(\boldsymbol{x})}$, and $\hat{m}_t \geq \max_{\boldsymbol{x}\in\mathfrak{X}} f(\boldsymbol{x})$, $\forall t \in [1, T]$.*

Lemma 3.4 (Lemma 5.3 in (Srinivas et al., 2010)). *The information gain for the points selected can be expressed in terms of the predictive variances. If $f_T = (f(\boldsymbol{x}_t)) \in \mathbb{R}^T$:*

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = \frac{1}{2}\sum_{t=1}^{T} \log(1 + \sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)).$$

Proof. (Theorem 3.1) By lemma 3.1 in our paper, we know that the theoretical results from EST (Wang et al., 2016) can be adapted to MES if $y_* \geq f_*$. The key question is when a sampled y_* that can satisfy this condition. Because the cumulative density $w = F(f_*) \in (0, 1)$ and y_*^t are independent samples from F, there exists at least one y_*^t that satisfies $y_*^t > f_*$ with probability at least $1 - w^{k_i}$ in k_i iterations.

Let $T' = \sum_{i=1}^{T} k_i$ be the total number of iterations. We split these iterations to T parts where each part have k_i iterations, $i = 1, \cdots, T$. By union bound, with probability at least $1 - \sum_{i=1}^{T} w^{k_i}$, in all the T parts of iterations, we have at least one iteration t_i which samples $y_*^{t_i}$ satisfying $y_*^{t_i} > f_*, \forall i = 1, \cdots, T$.

Let $\sum_{i=1}^{T} w^{k_i} = \frac{\delta}{2}$, we can set $k_i = \log_w \frac{\delta}{2\pi_i}$ for any $\sum_{i=1}^{T}(\pi_i)^{-1} = 1$. A convenient choice for π_i is $\pi_i = \frac{\pi^2 i^2}{6}$. Hence with probability at least $1 - \frac{\delta}{2}$, there exist a sampled $y_*^{t_i}$ satisfying $y_*^{t_i} > f_*, \forall i = 1, \cdots, T$.

Now let $\zeta_{t_i} = (2\log\frac{\pi_{t_i}}{\delta})^{\frac{1}{2}}$. By Lemma 3.2 and Lemma 3.3, the immediate regret $r_{t_i} = f_* - f(\boldsymbol{x}_{t_i})$ can be bounded as

$$r_{t_i} \leq (\nu_{t_i} + \zeta_{t_i})\sigma_{t_i-1}(\boldsymbol{x}_{t_i}).$$

Note that by assumption $0 \leq \sigma_{t_i-1}^2(\boldsymbol{x}_{t_i}) \leq 1$, so we have $\sigma_{t_i-1}^2 \leq \frac{\log(1+\sigma^{-2}\sigma_{t_i-1}^2(\boldsymbol{x}_{t_i}))}{\log(1+\sigma^{-2})}$. Then by Lemma 3.4, we have $\sum_{i=1}^{T}\sigma_{t_i-1}^2(\boldsymbol{x}_{t_i}) \leq \frac{2}{\log(1+\sigma^{-2})} I(\boldsymbol{y}_T; \boldsymbol{f}_T)$ where $\boldsymbol{f}_T = (f(\boldsymbol{x}_{t_i}))_{i=1}^T \in \mathbb{R}^T, \boldsymbol{y}_T = (y_{t_i})_{i=1}^T \in \mathbb{R}^T$. From assumptions, we have $I(\boldsymbol{y}_T; \boldsymbol{f}_T) \leq \rho_T$. By Cauchy-Schwarz inequality, $\sum_{i=1}^{T}\sigma_{t_i-1}(\boldsymbol{x}_{t_i}) \leq \sqrt{T\sum_{i=1}^{T}\sigma_{t_i-1}^2(\boldsymbol{x}_{t_i})} \leq \sqrt{\frac{2T\rho_T}{\log(1+\sigma^{-2})}}$. It follows that with probability at least $1 - \delta$,

$$\sum_{i=1}^{T} r_{t_i} \leq (\nu_{t^*} + \zeta_T)\sqrt{\frac{2T\rho_T}{\log(1+\sigma^{-2})}}.$$

As a result, our learning regret is bounded as

$$r_{T'} \leq \frac{1}{T}\sum_{i=1}^{T} r_{t_i} \leq (\nu_{t^*} + \zeta_T)\sqrt{\frac{2\rho_T}{T\log(1+\sigma^{-2})}},$$

where $T' = \sum_{i=1}^{T} k_i = \sum_{i=1}^{T} \log_w \frac{\delta}{2\pi_i}$ is the total number of iterations. □

At first sight, it might seem like MES with a point estimate does not have a converging rate as good as EST or $GP{-}UCB$. However, notice that $\min_{\boldsymbol{x}\in\mathfrak{X}} \gamma_{y_1}(\boldsymbol{x}) < \min \boldsymbol{x} \in \mathfrak{X}\gamma_{y_2}(\boldsymbol{x})$ if $y_1 < y_2$, which decides the rate of convergence in Eq. 1. So if we use y_* that is too large, the regret bound could be worse. If we use y_* that is smaller than f_*, however, its value won't count towards the learning regret in our proof, so it is also bad for the regret upper bound. With no principled way of setting y_* since f_* is unknown. Our regret bound in Theorem 3.1 is a randomized trade-off between sampling large and small y_*.

For the regret bound in add-GP-MES, it should follow add-GP-UCB. However, because of some technical problems in the proofs of the regret bound for add-GP-UCB, we haven't been able to show a regret bound for add-GP-MES either. Nevertheless, from the experiments on high dimensional functions, the methods worked well in practice.

4. Experiments

In this section, we provide more details on our experiments.

Optimization test functions In Fig. 2, we show the simple regret comparing BO methods on the three challenging optimization test functions: the 2-D eggholder function, the 10-D Shekel function, and the 10-D Michalewicz function.

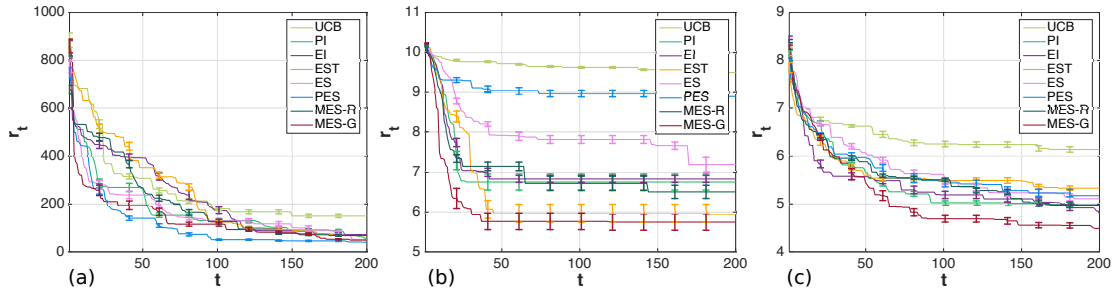

Figure 2. (a) 2-D eggholder function; (b) 10-D Shekel function; (c) 10-D Michalewicz function. PES achieves lower regret on the 2-d function while MES-G performed better than other methods on the two 10-d optimization test functions.

Choosing the additive decomposition We follow the approach in (Kandasamy et al., 2015), and sample 10000 random decompositions (at most 2 dimensions in each group) and pick the one with the best data likelihood based on 500 data points uniformly randomly sampled from the search space. The decomposition setting was fixed for all the 500 iterations of BO for a fair comparison.

References

Auer, Peter. Using confidence bounds for exploitation-exploration tradeoffs. *Journal of Machine Learning Research*, 3: 397–422, 2002.

Djolonga, Josip, Krause, Andreas, and Cevher, Volkan. High-dimensional Gaussian process bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Fisher, Ronald Aylmer. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

Hennig, Philipp and Schuler, Christian J. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

Hernández-Lobato, José Miguel, Hoffman, Matthew W, and Ghahramani, Zoubin. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Kandasamy, Kirthevasan, Schneider, Jeff, and Poczos, Barnabas. High dimensional Bayesian optimisation and bandits via additive models. In *International Conference on Machine Learning (ICML)*, 2015.

Kawaguchi, Kenji, Kaelbling, Leslie Pack, and Lozano-Pérez, Tomás. Bayesian optimization with exponential convergence. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Kawaguchi, Kenji, Maruyama, Yu, and Zheng, Xiaoyu. Global continuous optimization with error bound and fast convergence. *Journal of Artificial Intelligence Research*, 56(1):153–195, 2016.

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

Li, Chun-Liang, Kandasamy, Kirthevasan, Póczos, Barnabás, and Schneider, Jeff. High dimensional Bayesian optimization via restricted projection pursuit models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Massart, Pascal. *Concentration Inequalities and Model Selection*, volume 6. Springer, 2007.

Močkus, J. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, 1974.

Slepian, David. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, and Seeger, Matthias. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010.

Von Mises, Richard. La distribution de la plus grande de n valeurs. *Rev. math. Union interbalcanique*, 1936.

Wang, Zi, Zhou, Bolei, and Jegelka, Stefanie. Optimization as estimation with Gaussian processes in bandit settings. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Wang, Ziyu, Zoghi, Masrour, Hutter, Frank, Matheson, David, and De Freitas, Nando. Bayesian optimization in high dimensions via random embeddings. In *International Conference on Artificial Intelligence (IJCAI)*, 2013.