A warped kernel improving robustness in Bayesian optimization via random embeddings

Mickaël Binois, David Ginsbourger, Olivier Roustant

A warped kernel improving robustness in Bayesian optimization via random embeddings

Mickaël Binois[1,2] David Ginsbourger[3] Olivier Roustant[1]

October 27, 2014

[1] Mines Saint-Étienne, UMR CNRS 6158, LIMOS, F-42023 Saint-Étienne, France
[2] Renault S.A.S., 78084 Guyancourt, France
[3] University of Bern, Department of Mathematics and Statistics, Alpeneggstrasse 22, CH-3012 Bern, Switzerland

1 Introduction

The scope of Bayesian Optimization methods is usually limited to moderate-dimensional problems [1]. [2] recently proposed to extend the applicability of these methods to up to billions of variables, when only few of them are actually influential, through the so-called Random EMbedding Bayesian Optimization (REMBO) approach. In REMBO, optimization is conducted in a low-dimensional domain $\mathcal{Y}$, randomly embedded in the high-dimensional source space $\mathcal{X}$. New points are chosen by maximizing the Expected Improvement (EI) criterion [3] with Gaussian process (GP) models incorporating the considered embeddings via two kinds of covariance kernels proposed in [2]. A first one, $k_{\mathcal{X}}$, relies on Euclidean distances in $\mathcal{X}$. It delivers good performance in moderate dimension, albeit its main drawback is to remain high-dimensional so that the benefits of the method are limited. A second one, $k_{\mathcal{Y}}$, is defined directly over $\mathcal{Y}$ and is therefore independent from the dimension of $\mathcal{X}$. However, it has been shown [2] to possess artifacts that may lead EI algorithms to spend many iterations exploring equivalent points. Here we propose a new kernel with a warping (see e.g. [4]) inspired by simple geometrical ideas, that retains key advantages of $k_{\mathcal{X}}$ while remaining of low dimension as $k_{\mathcal{Y}}$.

2 Background on the REMBO method and related issues

The considered minimization problem is to find $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$, with $f : \mathcal{X} \subseteq \mathbb{R}^D \mapsto \mathbb{R}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^D$, assumed to be $[-1, 1]^D$ for simplicity. From [2], one main hypothesis about f is that its effective dimensionality is $d_e < D$: there exists a linear subspace $\mathcal{T} \subset \mathbb{R}^D$ of dimension d_e such that $f(\mathbf{x}) = f(\mathbf{x}_\top + \mathbf{x}_\perp) = f(\mathbf{x}_\top)$, $\mathbf{x}_\top \in \mathcal{T}$ and $\mathbf{x}_\perp \in \mathcal{T}^\perp \subset \mathbb{R}^D$ ([2], Definition 1). Given a random matrix $\mathbf{A} \in \mathbb{R}^{D\times d}$, $d \geq d_e$ with components sampled independently from $\mathcal{N}(0, 1)$, for any optimizer $\mathbf{x}^* \in \mathbb{R}^D$, there exists at least a point $\mathbf{y}^* \in \mathbb{R}^d$ such that $f(\mathbf{x}^*) = f(\mathbf{A}\mathbf{y}^*)$ with probability 1 ([2], Theorem 2.). To respect box constraints, f is evaluated at $p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$, the convex projection of $\mathbf{A}\mathbf{y}$ onto $\mathcal{X}$. The low dimensional function to optimize is then $g : \mathbb{R}^d \mapsto \mathbb{R}$, $g(\mathbf{y}) = f(p_{\mathcal{X}}(\mathbf{A}\mathbf{y}))$.

Optimizing g is carried out using Bayesian Optimization, e.g, with the EGO algorithm [3]. It bases on Gaussian Process Regression [5], also known as Kriging [6], to create a surrogate of g. Supposing that g is a sample from a GP with known mean (zero here to simplify notations) and covariance kernel $k(.,.)$, conditioning it on n observations $f(\mathbf{x}_{1:n}) = g(\mathbf{y}_{1:n})$, denoted $\mathbf{Z}$, provides a GP $Z(.)$ with mean $m(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{Z}$ and kernel $c(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}')$, where $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_i))_{1\leq i\leq n}$ and $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{1\leq i,j\leq n}$. The choice of k is preponderant, since it reflects a number of beliefs about the function at hand. Among the most commonly used are the "squared exponential" (SE) and "Matérn" stationary kernels, with hyperparameters such as length scales or degree of smoothness [7, 8]. For REMBO, [2] proposed two versions of the SE kernel, namely the low-dimensional $k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \exp\left(-\|\mathbf{y} - \mathbf{y}'\|_d^2/2l_{\mathcal{Y}}^2\right)$ and the high-dimensional $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\left\|p_{\mathcal{X}}(\mathbf{A}\mathbf{y}) - p_{\mathcal{X}}(\mathbf{A}\mathbf{y}')\right\|_D^2 /2l_{\mathcal{X}}^2)$ $(\mathbf{y}, \mathbf{y}' \in \mathcal{Y})$.

Selecting the domain $\mathcal{Y} \subset \mathbb{R}^d$ is a major difficulty of the method: if too small, the optimum may not be reachable while a too large domain renders optimizing harder, in particular since $p_{\mathcal{X}}$ is far from being injective. Distant points in $\mathcal{Y}$ may coincide in $\mathcal{X}$, especially far from the center, so that using $k_{\mathcal{Y}}$ leads to sample useless new points in $\mathcal{Y}$ corresponding to the same location in $\mathcal{X}$. On the other hand, $k_{\mathcal{X}}$ suffers from the curse of dimensionality when $\mathcal{Y}$ is large enough: whereas embedded points $p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ lie in a d dimensional subspace when they are inside of $\mathcal{X}$, they belong to a D-dimensional domain when they are projected onto the faces of $\mathcal{X}$. To alleviate these shortcomings, after showing that with probability $1-\epsilon$ the optimum is contained in the centered ball of radius d_e/ϵ (Theorem 3), the authors of [2] then suggest to set $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$. In practice, they split the evaluation budget over several random embeddings or set $d > d_e$ to increase the probability for the optimum to actually be inside $\mathcal{Y}$, slowing down the convergence.

3 Proposed kernel and experimental results

Both $k_{\mathcal{Y}}$ and $k_{\mathcal{X}}$ suffering from limitations, it is desirable to have a kernel that retains as much as possible of the actual high dimensional distances between points while remaining of low dimension. This can be achieved by first projecting orthogonally points on the faces of the hypercube to the subspace spanned by $\mathbf{A}$: $\mathrm{Ran}(\mathbf{A})$, with $p_{\mathbf{A}} : \mathcal{X} \mapsto \mathbb{R}^D$, $p_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$. Note that these back-projections from the hypercube can be outside of $\mathcal{X}$. The calculation of the projection matrix is done only once, inverting a $d \times d$ matrix. This solves the problem of adding already evaluated points: their back-projections coincide. Nevertheless, distant points on the sides of $\mathcal{X}$ from the convex projection can be back-projected close to each other, which may cause troubles with the stationary kernels classically used.

The next step is to respect as much as possible distances on the border of $\mathcal{X}$. Unfolding and parametrizing the manifold corresponding to the convex projection of the embedding of $\mathcal{Y}$ with $\mathbf{A}$ would be best but unfortunately it seems intractable with high D. Alternatively, we propose to distort the back-projections outside of $\mathcal{X}$, coming from points on the sides of $\mathcal{X}$. From the back-projection of the initial mapping with $p_{\mathcal{X}}$, a pivot point is selected as the intersection between the frontier of

$\mathcal{X}$ and the line $(O; p_{\mathbf{A}}(p_{\mathcal{X}}(\mathbf{A}y))$. Then the back-projection is stretched out such that the distance between the pivot point and the initial convex projection are equal. It results in respecting the distance *on the embedding* between the center O and the initial convex projection. The resulting warping, denoted Ψ, is detailed in Algorithm 1 and illustrated in Figure 1. Based on this, any positive definite kernel k on $\mathcal{Y}$ can be used. For example, the resulting SE kernel is $k_\Psi(\mathbf{y}, \mathbf{y'}) = \exp\left(-\|\Psi(\mathbf{y}) - \Psi(\mathbf{y'})\|_D^2 / 2l_\Psi^2\right)$.

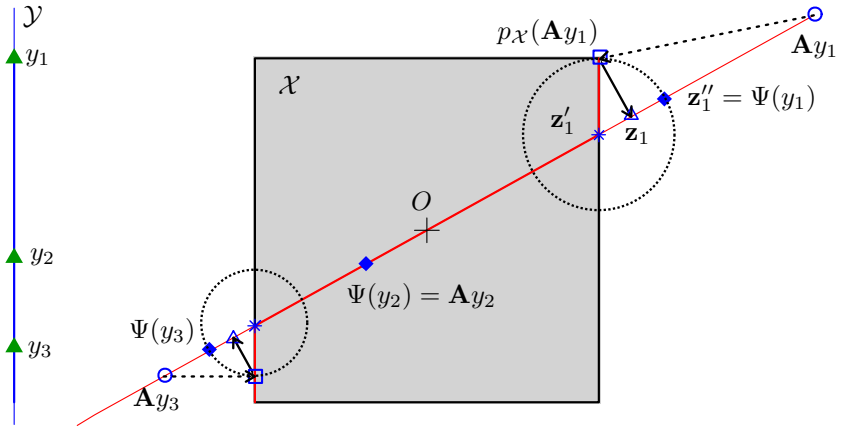

Figure 1: New warping with Ψ, $d = 1$ and $D = 2$, from triangles in $\mathcal{Y}$ to diamonds in $\mathcal{X}$. Steps of the construction of $\Psi(y_1)$ are detailed.

Like $k_{\mathcal{X}}$, k_Ψ is not hindered by the non-injectivity brought by the convex projection $p_{\mathcal{X}}$. Furthermore, it can explore sides of the hypercube without spending too much budget since remaining on $\mathrm{Ran}(\mathbf{A})$. It is thus possible to extend the size of $\mathcal{Y}$ to avoid the risk of missing the optimum. In experiments with k_Ψ, we set $\mathcal{Y}$ to $[-\gamma, \gamma]^d$ with γ such that $\min_{j\in[1,D]} \sum_{i=1}^{d} |A_{j,i}|\gamma = 1$, ensuring to span $[-1, 1]$ for each of the D variables.

Algorithm 1 Calculation of Ψ.

1: Map $\mathbf{y} \in \mathcal{Y}$ to $\mathbf{Ay}$
2: **If** $\mathbf{Ay} \in \mathcal{X}$ **Then**
3: Define $\Psi(\mathbf{y}) = \mathbf{Ay}$
4: **Else**
5: Project onto $\mathcal{X}$ and back-project onto $\mathrm{Ran}(\mathbf{A})$: $\mathbf{z} = p_{\mathbf{A}}(p_{\mathcal{X}}(\mathbf{Ay}))$
6: Compute the intersection of $[O; \mathbf{z}]$ with $\mathcal{X}$: $\mathbf{z}' = (\max_{i=1,\dots,D} |z_i|)^{-1}\mathbf{z}$
7: Define $\Psi(\mathbf{y}) = \mathbf{z}' + \|p_{\mathcal{X}}(\mathbf{A}y) - \mathbf{z}'\|_D . \frac{\mathbf{z}'}{\|\mathbf{z}'\|_D}$
8: **EndIf**

We compare the usual REMBO method with $k_{\mathcal{Y}}$ and the proposed k_Ψ, with a unique embedding. Tests are conducted with the *DiceKriging* and *DiceOptim* packages [7]. We use the Matérn 5/2 kernel with hyperparameters estimated with Maximum Likelihood and we start optimization with space filling designs of size $10d$. Initial designs are modified such that no points are repeated in $\mathcal{X}$ for $k_{\mathcal{Y}}$. For k_Ψ, we apply Ψ to

bigger initial designs before selecting the right number of points, as distant as possible between each other. Experiments are repeated fifty times, taking the same random embeddings for both kernels. Results in Figure 2 show that the proposed kernel k_Ψ outperforms $k_\mathcal{Y}$ when $d = 6$, in particular since it avoids the problem of a too small domain $\mathcal{Y}$. Additional tests with $k_\mathcal{X}$ actually resulted in a worse performance. Studying the efficiency of splitting the evaluation budget between several random embeddings, compared to relying on a single one along with k_Ψ, would be the scope of future research.

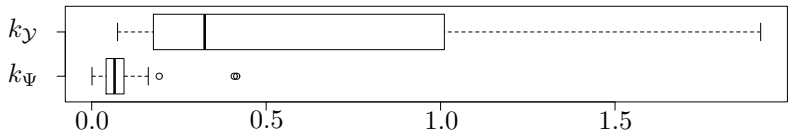

Figure 2: Boxplot of the optimality gap (best value found minus minimal function value) for kernels $k_\mathcal{Y}$ and k_Ψ on the Hartmann6 test function (see e.g. [3]) with 250 evaluations, $d = d_e = 6$, $D = 25$.

Acknowledgments

This work has been conducted within the frame of the ReDice Consortium, gathering industrial (CEA, EDF, IFPEN, IRSN, Renault) and academic (Ecole des Mines de Saint-Etienne, INRIA, and the University of Bern) partners around advanced methods for Computer Experiments.

References

[1] S. Koziel, D. E. Ciaurri, and L. Leifsson, "Surrogate-based methods," in *Computational Optimization, Methods and Algorithms*, pp. 33–59, Springer, 2011.

[2] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. de Freitas, "Bayesian optimization in a billion dimensions via random embeddings," *In IJCAI*, 2013.

[3] D. Jones, M. Schonlau, and W. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[4] J. Snoek, K. Swersky, R. S. Zemel, and R. P. Adams, "Input warping for Bayesian optimization of non-stationary functions," *In ICML*, 2014.

[5] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[6] G. Matheron, "Principles of geostatistics," *Economic geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[7] O. Roustant, D. Ginsbourger, and Y. Deville, "DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization," *Journal of Statistical Software*, vol. 51, no. 1, pp. 1–55, 2012.

[8] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer, 1999.