Bayesian Optimization over Sets

Jungtaek Kim[1], Michael McCourt[2], Tackgeun You[1], Saehoon Kim[3], and Seungjin Choi[1]

[1]Pohang University of Science and Technology, [2]SigOpt, [3] AITRICS

Introduction

- Classic BO assumes that a search region $\mathcal{X} \subset \mathbb{R}^d$ is defined and that the function f can only produce scalar output:
$$y = f(\mathbf{x}) + \epsilon \quad \text{for} \quad \mathbf{x} \in \mathcal{X}.$$
- Unlike this, assume that our search region is $\mathcal{X}_{\text{set}} = \{\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \mid \mathbf{x}_i \in \mathbb{R}^d\}$ for a fixed positive integer m.
- For $\mathbf{X} \in \mathcal{X}_{\text{set}}$, f would take in a set containing m elements, all of length d, and return a noisy function value y:
$$y = f\left(\mathbf{X}\right) + \epsilon.$$

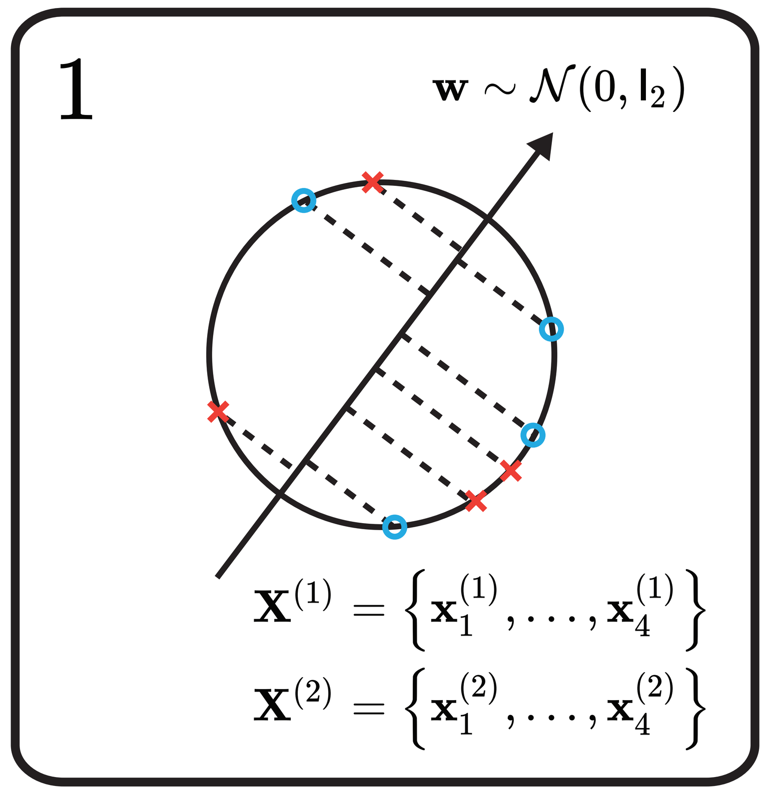
Motivating Example

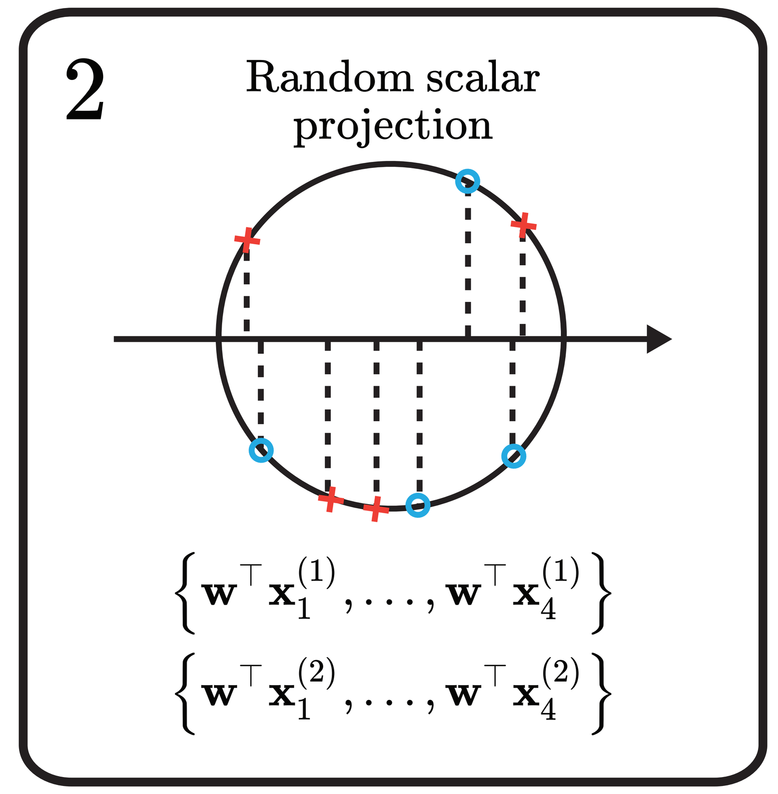
- The soft k-means clustering algorithm over a dataset $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$.
- We want to find the optimal initialization of such an algorithm.
- The function of k-means clustering is the converged clustering residual
$$f(\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}) = \sum_{i=1}^{N} \sum_{j=1}^{k} w_{ij} \|\mathbf{p}_i - \mathbf{c}_j\|_2^2.$$

Background

Bayesian Optimization

- A method to find global optimum for black-box function expensive to evaluate.
- It improves the current best solution as iterating the steps: modeling a surrogate function and acquiring a next best point.
- It optimizes an acquisition function instead of an original target function.





3

$$\tilde{\mathbf{X}}^{(1)} = \left\{\mathbf{x}_1^{(1)}, \mathbf{x}_3^{(1)}\right\}$$
$$\tilde{\mathbf{X}}^{(2)} = \left\{\mathbf{x}_2^{(2)}, \mathbf{x}_3^{(2)}\right\}$$
$$\tilde{k}_{\text{set}}\left(\tilde{\mathbf{X}}^{(1)}, \tilde{\mathbf{X}}^{(2)}\right) = \frac{1}{4}\big(k(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(2)}) + k(\mathbf{x}_1^{(1)}, \mathbf{x}_3^{(2)}) + k(\mathbf{x}_3^{(1)}, \mathbf{x}_2^{(2)}) + k(\mathbf{x}_3^{(1)}, \mathbf{x}_3^{(2)})\big)$$

Set Kernel

- Denote that a set of m vectors is $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.
- A set kernel is defined as
$$k_{\text{set}}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = \frac{1}{|\mathbf{X}^{(1)}||\mathbf{X}^{(2)}|} \sum_{i=1}^{|\mathbf{X}^{(1)}|} \sum_{j=1}^{|\mathbf{X}^{(2)}|} k\left(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(2)}\right).$$

Proposed Method

Lemma 1 *Suppose we have a list $\mathfrak{X}$ which contains distinct sets $\mathbf{X}^{(i)}$ for $1 \leq i \leq n$. We define the matrix $\mathsf{K} \in \mathbb{R}^{n \times n}$ as*
$$(\mathsf{K})_{ij} = k_{set}\left(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}\right), \qquad (1)$$
for k_{set} defined with a chosen inner kernel k. Then, K is a symmetric positive-semidefinite matrix if k is a symmetric positive-definite kernel.

Approximation of the Set Kernel

- (1) requires a complexity $\mathcal{O}(n^2 m^2 d)$.
- To alleviate this cost, we propose to approximate the set kernel with
$$\tilde{k}_{\text{set}}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}; \pi, \mathbf{w}, L\right) = k_{\text{set}}\left(\tilde{\mathbf{X}}^{(1)}, \tilde{\mathbf{X}}^{(2)}\right)$$
where $\pi : [1, \ldots, m] \to [1, \ldots, m]$, $\mathbf{w} \in \mathbb{R}^d$, $L \in \mathbb{Z}_+$, $\tilde{\mathbf{X}}^{(i)}$ is a subset of $\mathbf{X}^{(i)}$.

Theorem 1 *Suppose that we are given two sets $\mathbf{X}, \mathbf{Y} \in \mathcal{X}_{set}$ and $L \in \mathbb{Z}_+$. Suppose, furthermore, that $\mathbf{w}$ and π can be generated randomly to form subsets $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. The value of $\tilde{k}_{set}(\mathbf{X}, \mathbf{Y}; \mathbf{w}, \pi, L)$ is an unbiased estimator of the value of $k_{set}(\mathbf{X}, \mathbf{Y})$.*

Theorem 2 *Suppose the same conditions as in Theorem 1. Suppose, furthermore, that $k(\mathbf{x}, \mathbf{x}') \geq 0$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The variance of $\tilde{k}_{set}(\mathbf{X}, \mathbf{Y}; \mathbf{w}, \pi, L)$ is bounded by a function of m, L and $k_{set}(\mathbf{X}, \mathbf{Y})$:*
$$\mathit{Var}\left[\tilde{k}_{set}(\mathbf{X}, \mathbf{Y}; \mathbf{w}, \pi, L)\right] \leq \left(\frac{m^4}{L^4} - 1\right) k_{set}(\mathbf{X}, \mathbf{Y})^2.$$

- The complexity of ours is $\mathcal{O}(n^2 L^2 d)$.
- We use either the set kernel or the approximated kernel in modeling Gaussian process regression.
- It can suggest the best candidate set.

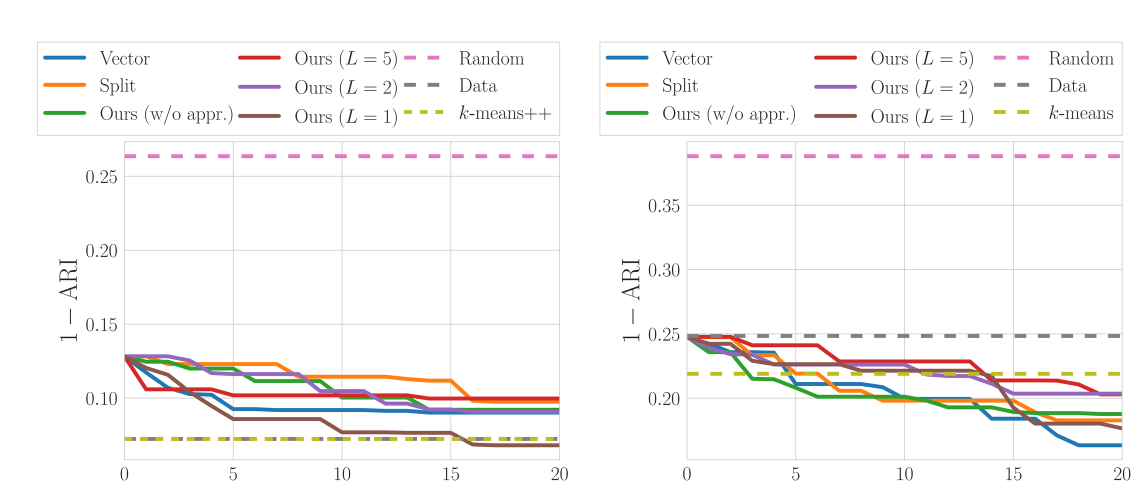


Figure 1: Results on k-means clustering and MoG.

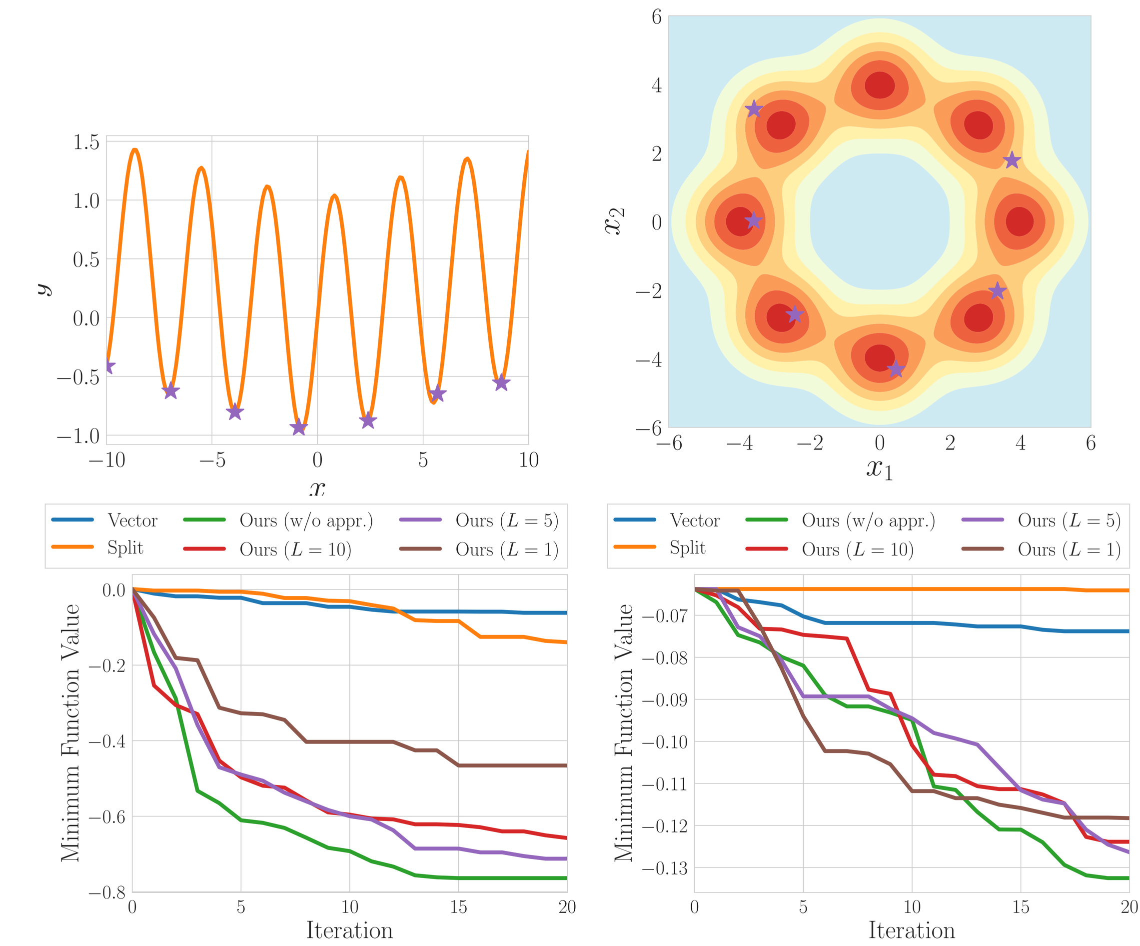


Figure 2: Results on synthetic functions.

Experiments

- We test our method in
  - two synthetic functions
  - initialization of two clustering methods.

Conclusion

- We propose a Bayesian optimization method over sets with set kernels.
- To reduce the complexity, we approximate the kernels to efficient kernels.
- We demonstrate that our method can be applied in some set-input examples.
- Our open repository: https://github.com/jungtaekkim/bayeso

Contact Information

- Homepage: http://mlg.postech.ac.kr/~jtkim
- Email: jtkim@postech.ac.kr