Batched Large-scale Bayesian Optimization in High-dimensional Spaces

Z. Wang, C. Gehring, P. Kohli, and S. Jegelka

A Paper Review by:
Luca Choteborsky

Background Context

Bayesian Optimization

- Optimize black-box function
- **Surrogate Model:** Uses GP to approximate the function
- **Acquisition Function:** Guides where to sample next
- **Iterative Process:** Updated model with each new sample from acquisition

Applications to High-Dimensional Functions

- Limited to low-dimensional problems due to its computational and statistical challenges

Background Context

Addressed by assuming a simpler underlying structure

- Djolonga et al. (2013) assume a low-dimensional effective subspace
- Kandasamy et al. (2015) assume additive structure of the function, constituent functions operate on disjoint low-dimensional subspaces
- **Fully optimising the decomposition is intractable**

Adapting the decomposition

- Maximise the GP marginal likelihood every certain number of iterations
- However, this maximisation is **computationally intractable** due to combinatorial nature of the partitions of the feature space
- Instead used randomized search heuristics

Background Context

Changes over the past years

- There has been an increased interest modelling functions with a large number of parameters
- Movement to more parallel architectures: multi-core, GPUs, clusters

Problem Tackled

Tackle BO on high-dimensional black-box functions

- Assume a latent additive structure in the function and infer it properly for more efficient and effective BO
- Perform multiple evaluations in parallel to reduce the number of iterations required by the method

Problem Tackled

Additive BO

- We want to find $f(x^*) = \max_{x \in \mathcal{X}} f(x).$

- Assume a latent decomposition of the feature dimensions into disjoint subspaces. Further f can be decomposed into the following additive form

$$f(x) = \sum_{m \in [M]} f_m(x^{A_m}).$$

- Assume each function is drawn independently from $\mathcal{GP}(0, k^{(m)})$

Problem Tackled

Additive BO

- The log data likelihood for $\mathcal{D}_n$

$$\log p(\mathcal{D}_n | \{k^{(m)}, A_m\}_{m \in [M]}) \tag{2.1}$$
$$= -\frac{1}{2}(\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y} + \log|\boldsymbol{K}_n + \sigma^2 \boldsymbol{I}| + n \log 2\pi)$$

- Can then infer the posterior mean and covariance function of the subfunction

$$\mu_n^{(m)}(x^{A_m}) = \boldsymbol{k}_n^{(m)}(x^{A_m})^{\mathrm{T}}(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y},$$
$$k_n^{(m)}(x^{A_m}, x'^{A_m}) = k^{(m)}(x^{A_m}, x'^{A_m})$$
$$- \boldsymbol{k}_n^{(m)}(x^{A_m})^{\mathrm{T}}(\boldsymbol{K}_n + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{k}_n^{(m)}(x'^{A_m}),$$

where $\boldsymbol{k}_n^{(m)}(x^{A_m}) = [k^{(m)}(x_t^{A_m}, x^{A_m})]_{t \leq n}$.

Problem Tackled

Additive BO

- The authors use regret to evaluate the BO algorithms in the sequential and the batch selection case
- For the sequential case

$$\tilde{r}_t = \max_{x \in \mathcal{X}} f(x) - f(x_t)$$

- For the batch selection case

$$\tilde{r}_t = \max_{x \in \mathcal{X}, b \in [B]} f(x) - f(x_{t,b})$$

- The authors then looked at averaged accumulative regret and simple regret

$$R_T = \frac{1}{T} \sum_t \tilde{r}_t \qquad\qquad r_T = \min_{t \leq T} \tilde{r}_t$$

Key Solution

Learning Additive Structure

- Takes a Bayesian view on the task of learning the latent structure of the GP kernel
- The decomposition of the input space is learnt **simultaneously** with optimization
- Decomposition is sampled using $\theta \sim \textsc{Dir}(\alpha)$ $z_j \sim \textsc{Multi}(\theta)$

Key Solution

Learning Additive Structure

- The authors then use Gibbs sampling to learn the posterior distribution for z
- Choose the decomposition among the samples that achieves the highest data likelihood, then proceed with BO.
- Gibbs sampler draws z_j according to

$$\begin{aligned} p(z_j = m \mid z_{\neg j}, \mathcal{D}_n;\ \alpha) &\propto p(\mathcal{D}_n \mid z) p(z_j \mid z_{\neg j}) \\ &\propto p(\mathcal{D}_n \mid z)(|A_m| + \alpha_m) \propto e^{\phi_m}, \end{aligned}$$

Key Solution

Diverse Batch Sampling

- Selects a batch of *B* observations to be made in parallel, then the model is updated with all simultaneously
- Need an efficient strategy that encourages observations that are both informative and non-redundant
- Given a decomposition *z*, they define a separate Determinantal Point Process (DPP) on each group of *A*□ dimensions and sample a set of points in the subspace.
- As group sizes are upper-bounded by some constant, sampling from each such DPP gives an exponential speedup

Key Solution

Combining Samples

- Combines samples from each group *randomly without replacement*
- Or *greedily*, define a *quality function* for each group, and combine samples to maximise this function
- Then showed how the batched framework works with GP-UCB, by setting both the acquisition function and quality function to

$$(f_t^{(m)})^+(x) = \mu_{t-1}^{(m)}(x) + \beta_t^{1/2}\sigma_t^{(m)}(x)$$

- To ensure that points with high acquisition function values are selected, they define a relevance region for each group *m*

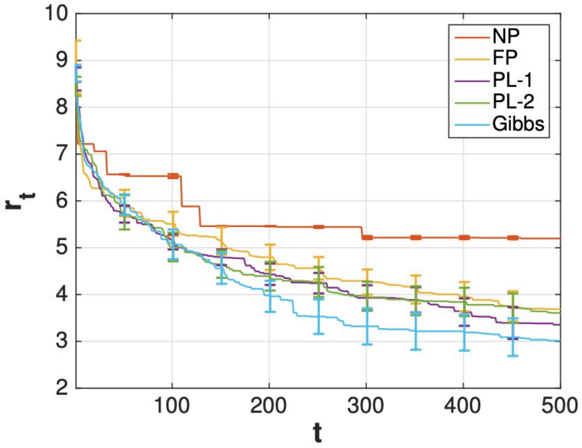
Evaluation

Table 1. Empirical posterior of any two dimensions correctly being grouped together by Gibbs sampling.

D \ N	50	150	250	450
5	0.81 ± 0.28	0.91 ± 0.19	1.00 ± 0.03	1.00 ± 0.00
10	0.21 ± 0.13	0.54 ± 0.25	0.68 ± 0.25	0.93 ± 0.15
20	0.06 ± 0.06	0.11 ± 0.08	0.20 ± 0.12	0.71 ± 0.22
50	0.02 ± 0.03	0.02 ± 0.02	0.03 ± 0.03	0.06 ± 0.04
100	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.02 ± 0.02

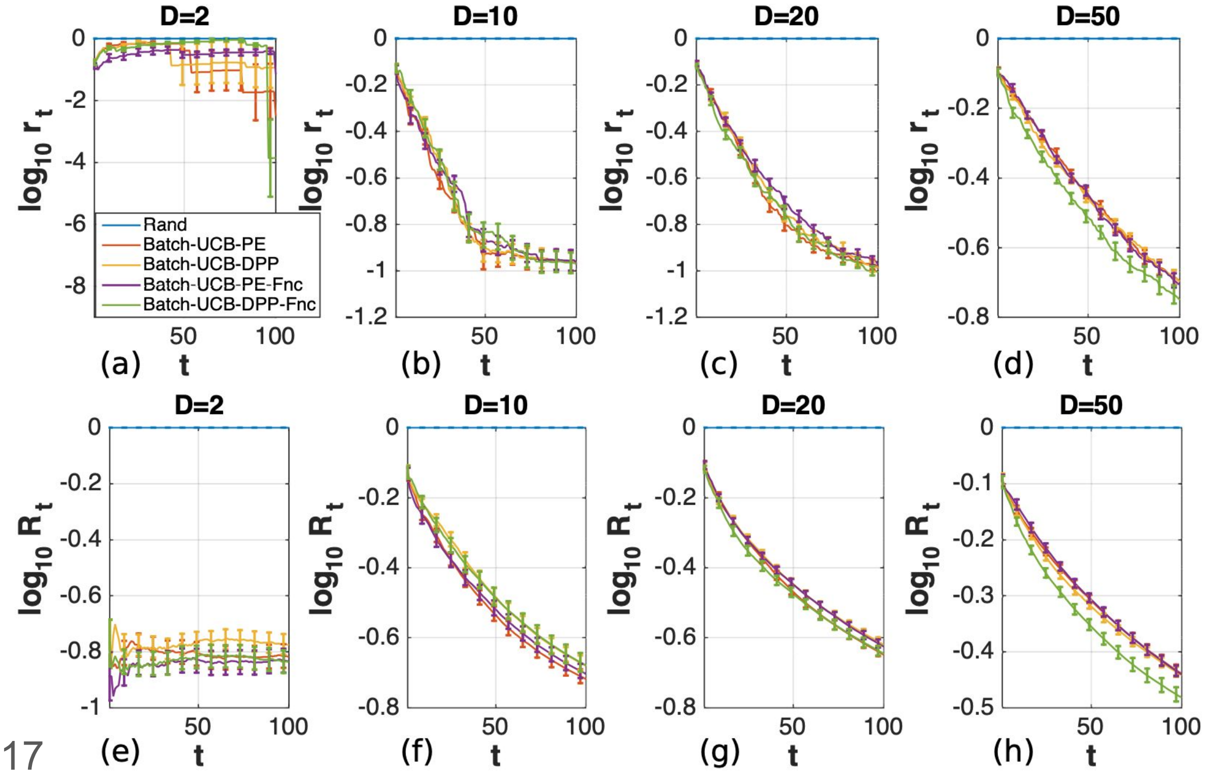
Table 2. Empirical posterior of any two dimensions correctly being separated by Gibbs sampling.

D \ N	50	150	250	450
2	0.30 ± 0.46	0.30 ± 0.46	0.90 ± 0.30	1.00 ± 0.00
5	0.87 ± 0.17	0.80 ± 0.27	0.60 ± 0.32	0.50 ± 0.34
10	0.88 ± 0.05	0.89 ± 0.06	0.89 ± 0.07	0.94 ± 0.07
20	0.94 ± 0.02	0.94 ± 0.02	0.94 ± 0.02	0.97 ± 0.02
50	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.01	0.98 ± 0.01
100	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00



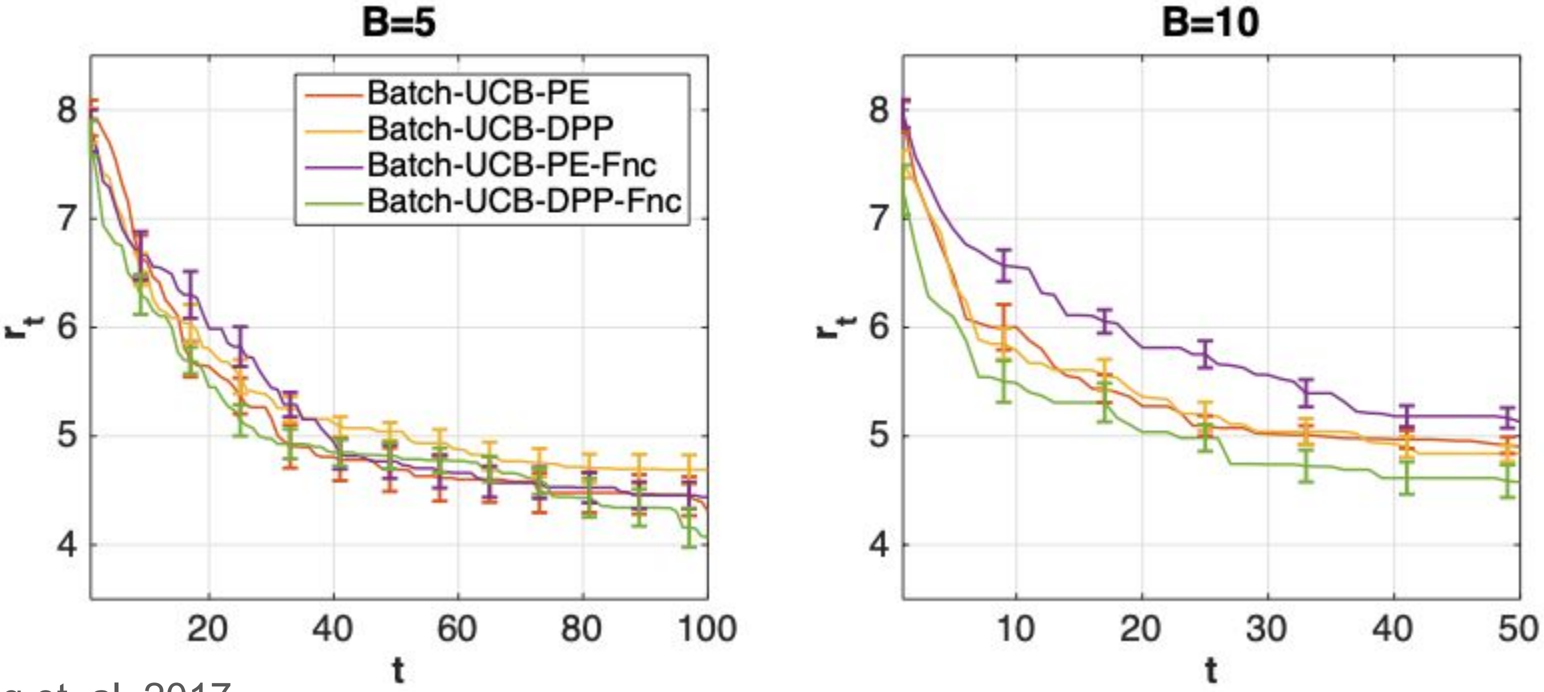
Wang et. al. 2017

Evaluation



Wang et. al. 2017

Evaluation



Wang et. al. 2017

Opinion of Paper

Agree

- The necessity of the paper
- The rational and techniques used
- The results of the paper and the subsequent explanations

Disagree

- Claims that performance increases with additional dimensions but graphs do not show that

Opinion of Paper

Strengths

- Extended application of BO to higher-dimensional problems
- Proposed a dimensional decomposition technique that can be applied in parallel to the optimisation
- Proposed a batch sampler for high-dimensions using subspace decompositions
- Evaluated decomposition and batch samplers on artificial functions and on a real-life function suspected to have a latent additive structure
- Well-written

Opinion of Paper

Weakness

- Only works on functions with latent (partial) additive structure
- Did not evaluate performance on non-additive functions
- Mostly theoretical so didn’t show real-world performance
- Did not compare performance with existing systems

Opinion of Paper

Key Takeaway

- Propose two solutions for high-dimensional BO: inferring latent structure, and combining it with batch BO
- Results of experiments demonstrate that proposed techniques are effective

Key Impact

- Has lead to paper expanding on scalability (Wang et al., 2018)
- Gibbs sampler learns also the kernel parameters
- Partitions input space for scalability using Mondrian forests
- Automatically generates batch queries