DeepMind

Batched High-dimensional Bayesian Optimization via Structural Kernel Learning

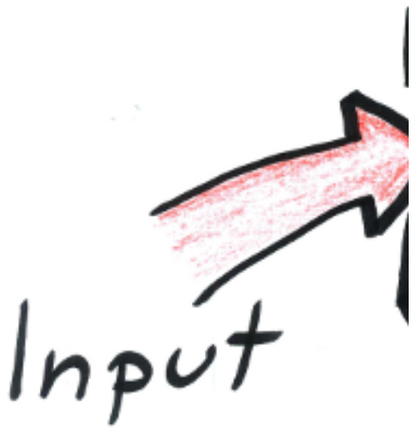
Zi Wang* Chengtao Li*

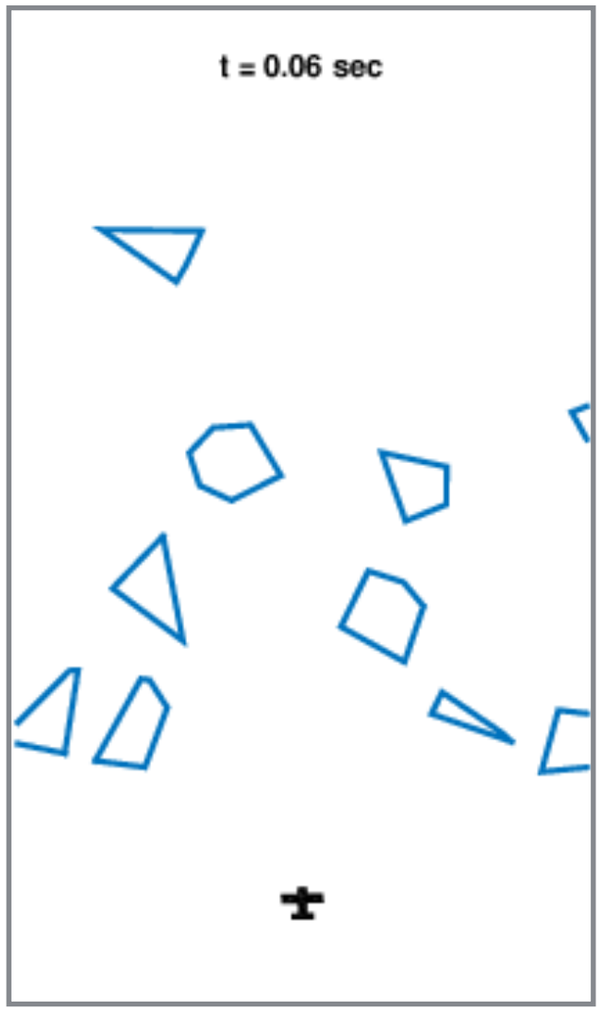
Stefanie Jegelka Pushmeet Kohli

Optimization: the Fundamental Pillar of ML

Many breakthroughs in ML (and many other fields) relies on the development of optimization.

- ☑ Cor
- ? Glo





Expensive

(Wang et al., 2015; Gonzalez et al., 2015)

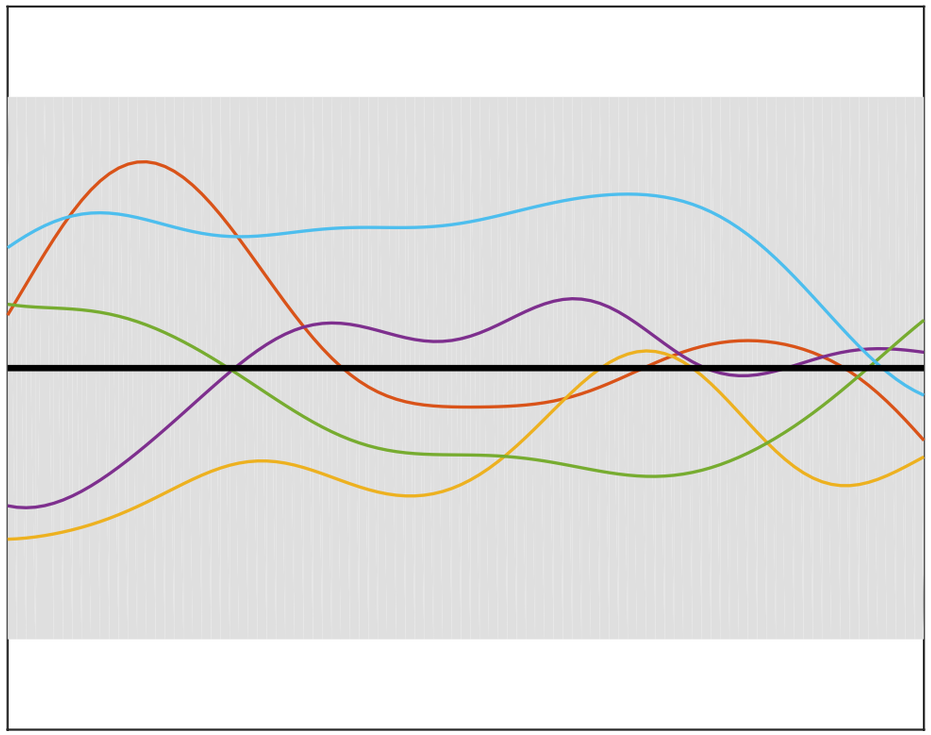
GP-UCB: an example of Bayesian Optimization

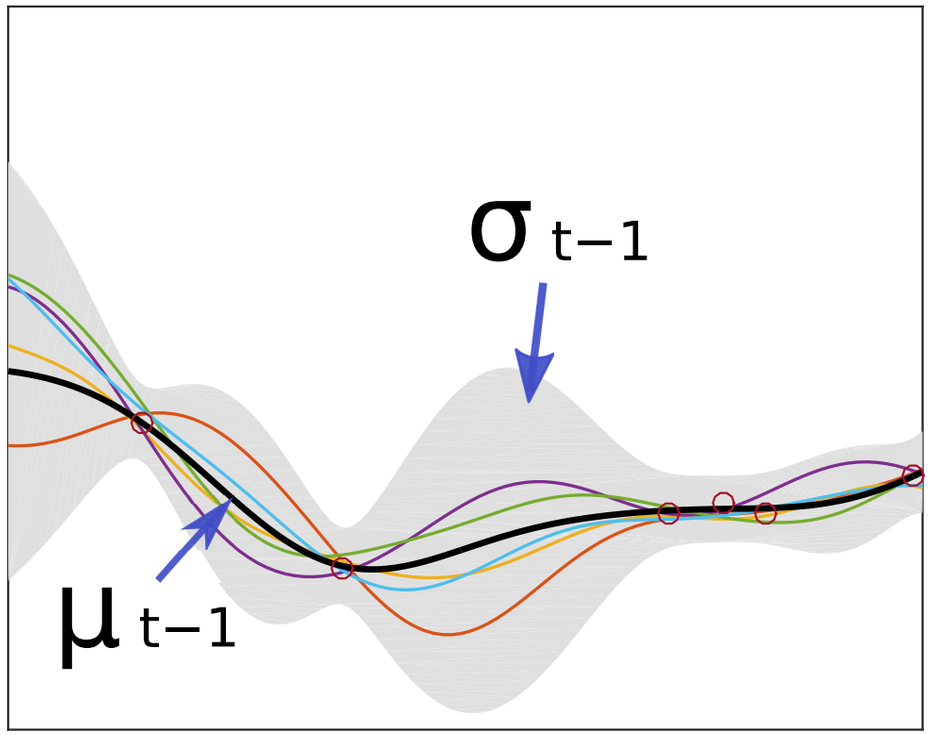
Prior: $f \sim GP(\mu, k)$

At iteration t,

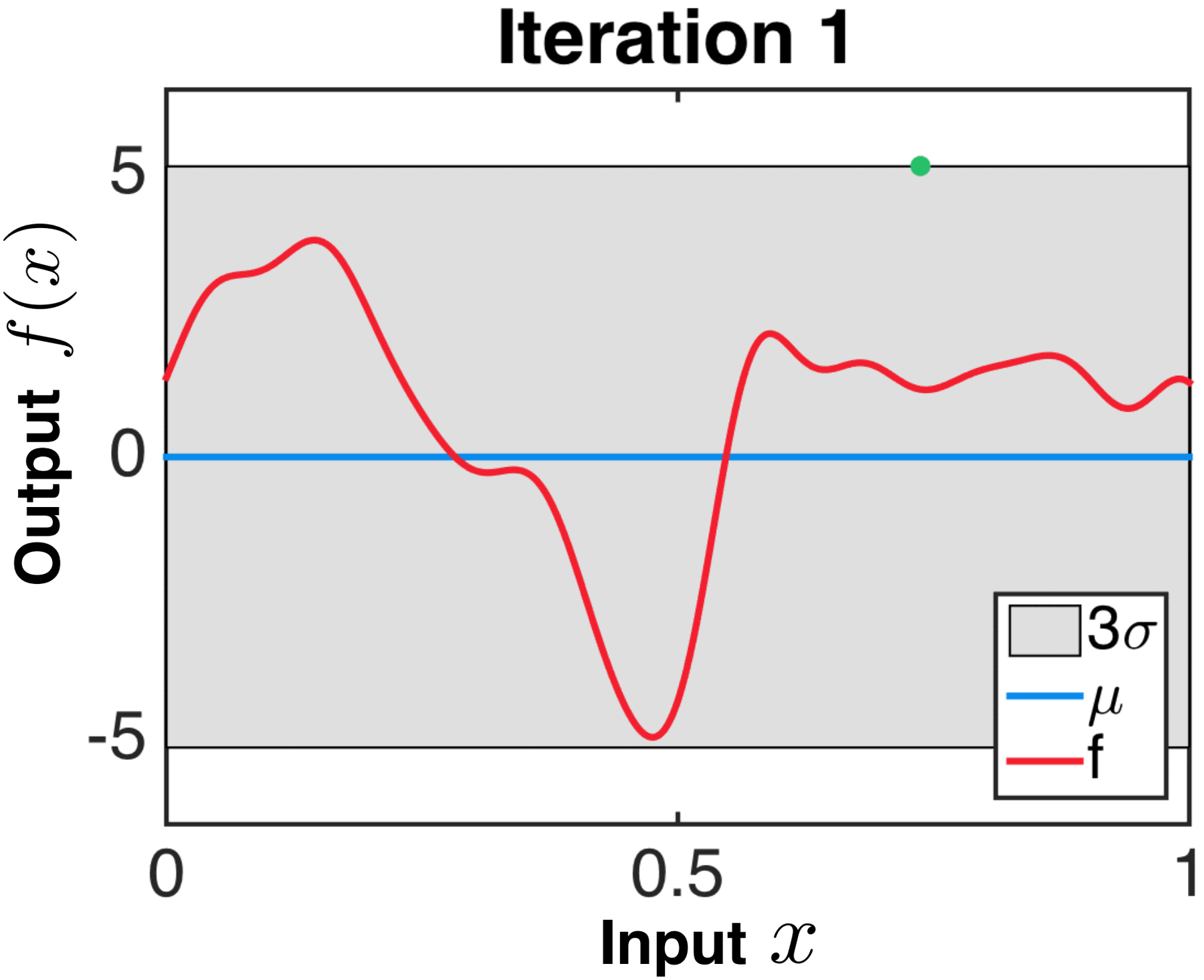
- predict the posterior $\mu_{t-1}(x)$ and $\sigma^2_{t-1}(x)$
- pick an input by optimizing the acquisition function

$$x_t = \arg\max \mu_{t-1}(x) + \beta\sigma_{t-1}(x)$$





GP-UCB: an example of Bayesian Optimization



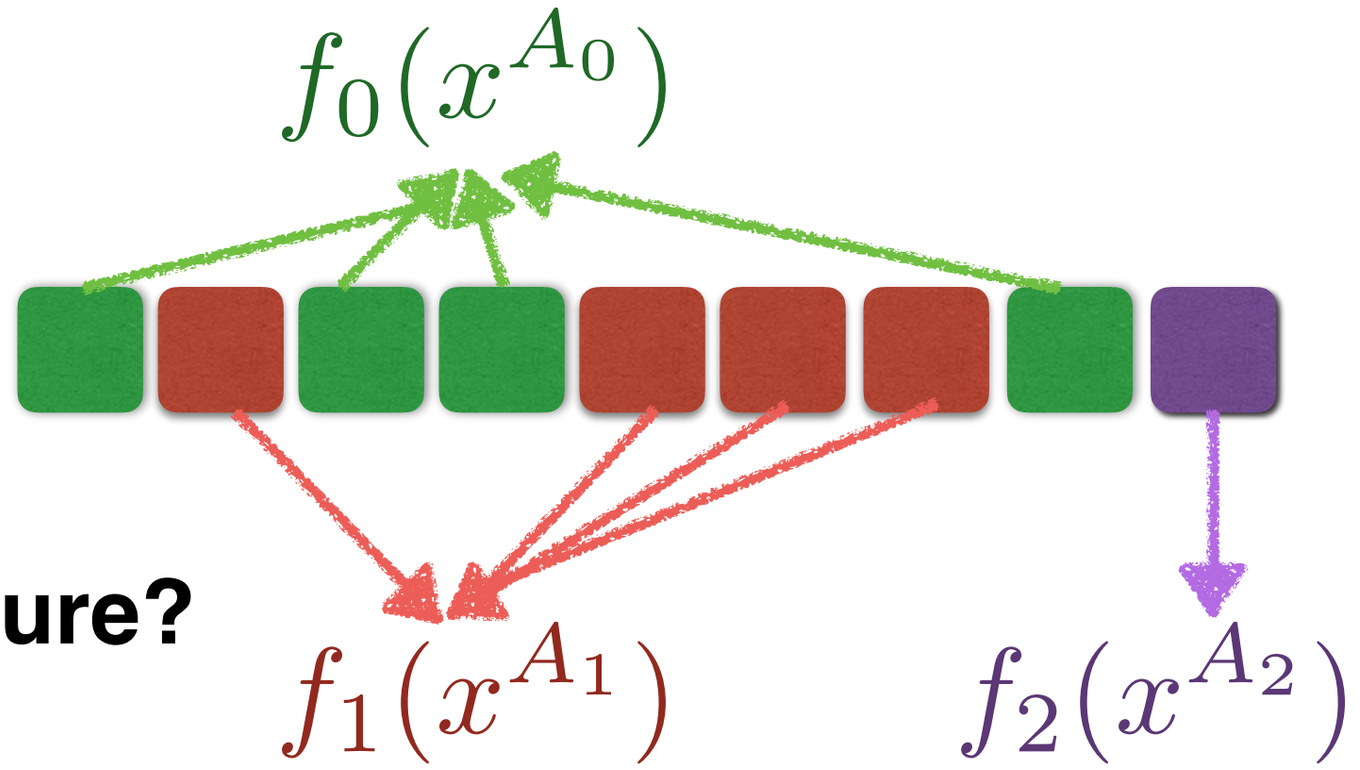
Challenges in high-dimensional BO

- Statistical challenges in high-dimensional function estimation;
- Computational challenges of high-dimensional acquisition function optimization.

(Hastie&Tibshirani, 1990; Kandasamy et al., 2015)

Possible solution: additive Gaussian processes?

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



What is the additive structure?

Our contributions

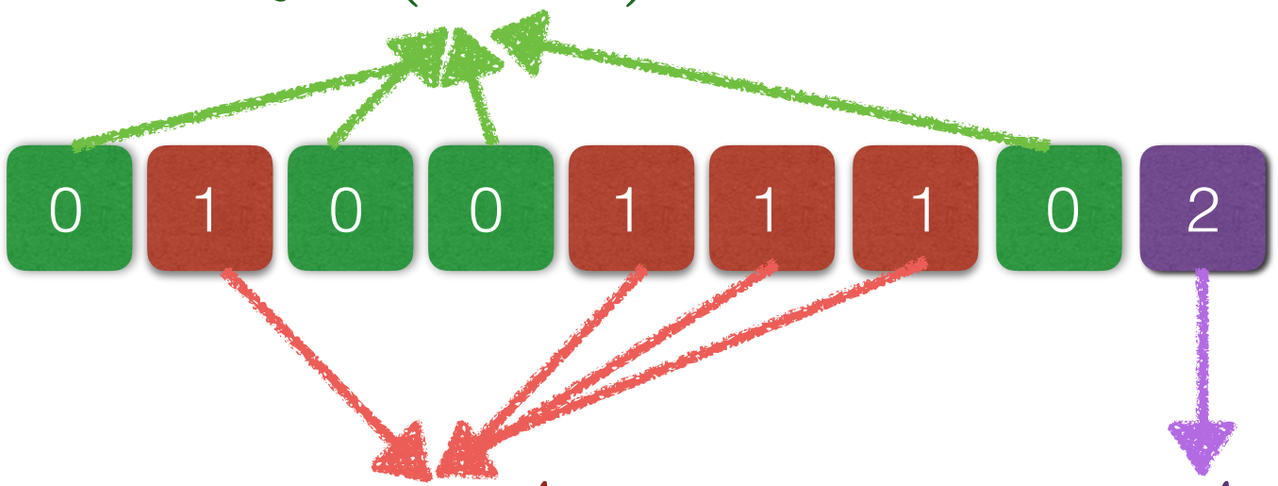
- Structural kernel learning: a novel approach to learn the additive structure in high-dimensional BO;

- Batched BO via Determinantal point process sampling.

Structural kernel learning for additive functions

$$f = f_0 + f_1 + f_2$$

$f_0(x^{A_0})$



$f_1(x^{A_1})$

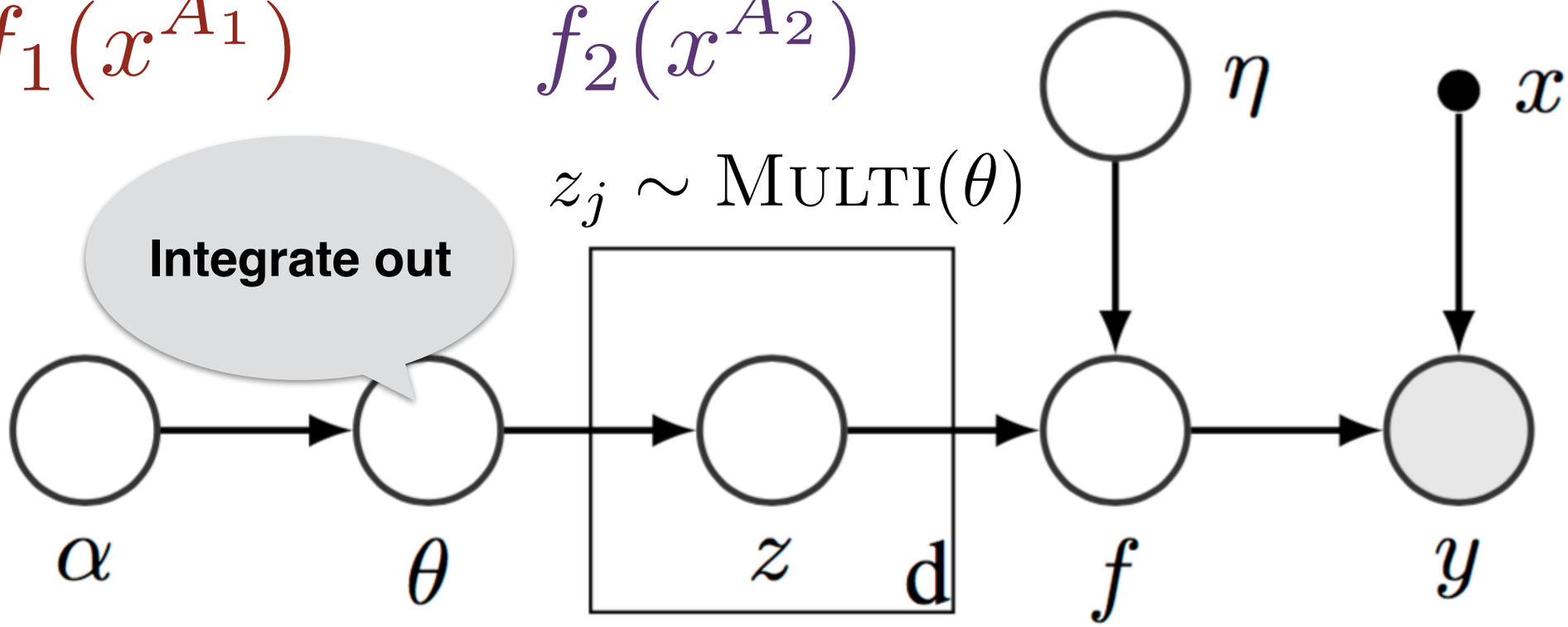
$f_2(x^{A_2})$

Example:

z = [0 1 0 0 1 1 1 0 2]

Key idea:

Put a Dirichlet prior on z



Scaling up the Dimensions via Structural Kernel Learning

Posterior of the group indicator z:

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

$$\phi_m = -\frac{1}{2}\boldsymbol{y}^\top (\boldsymbol{K}_n^{(z_j=m)} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{y} - \frac{1}{2}\log\det(\boldsymbol{K}_n^{(z_j=m)} + \sigma^2 \boldsymbol{I}) + \log(|A_m| + \alpha_m)$$

$A_m = \{j : z_j = m\}$ $\mathcal{D}_n$:observations

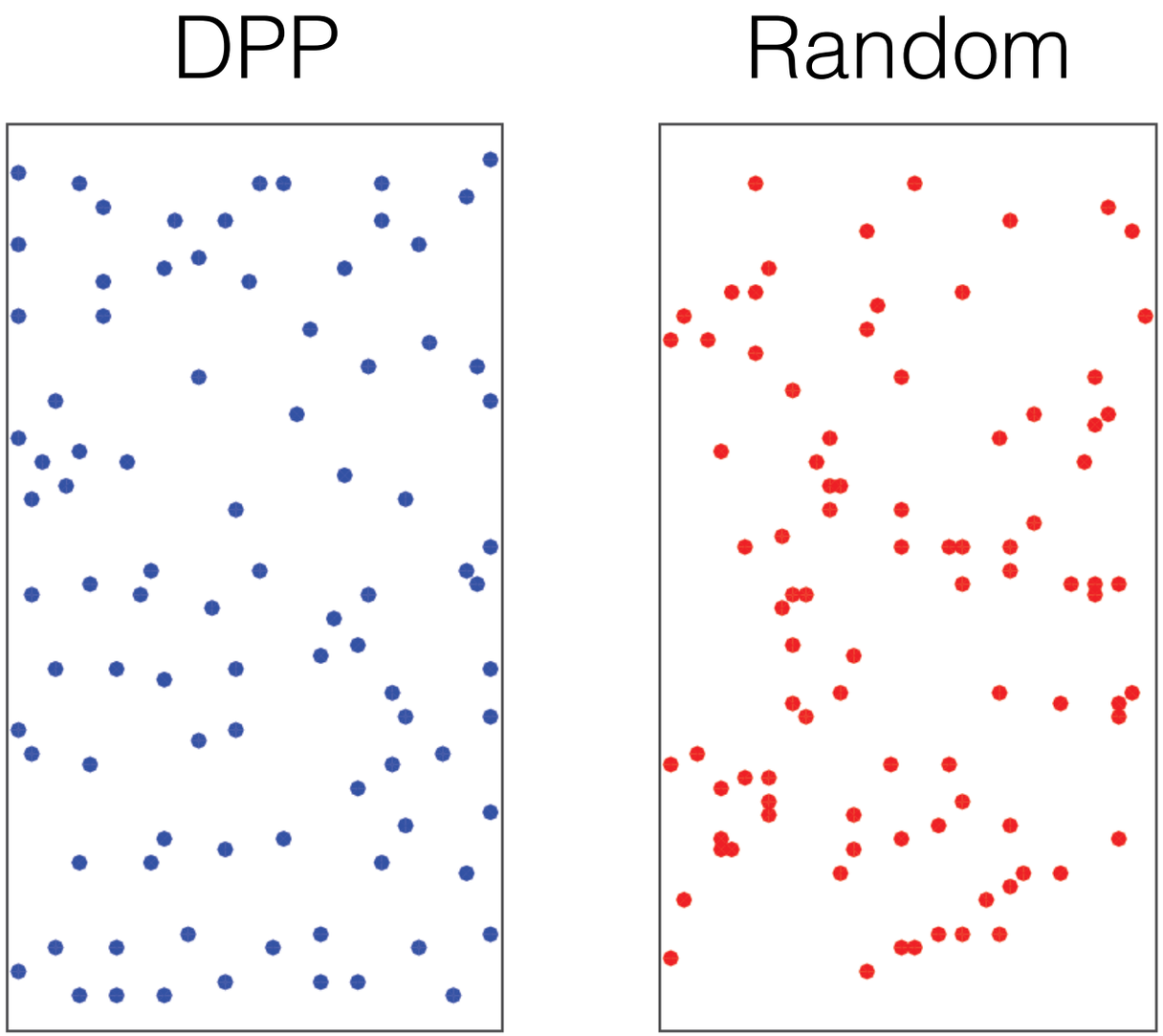
Diverse Batch Sampling for Parallel Queries

Sequential Batch Bayesian Optimization

- Select with GP-UCB
- Diverse Determinental Point Process (DPP) selections in high-valued areas

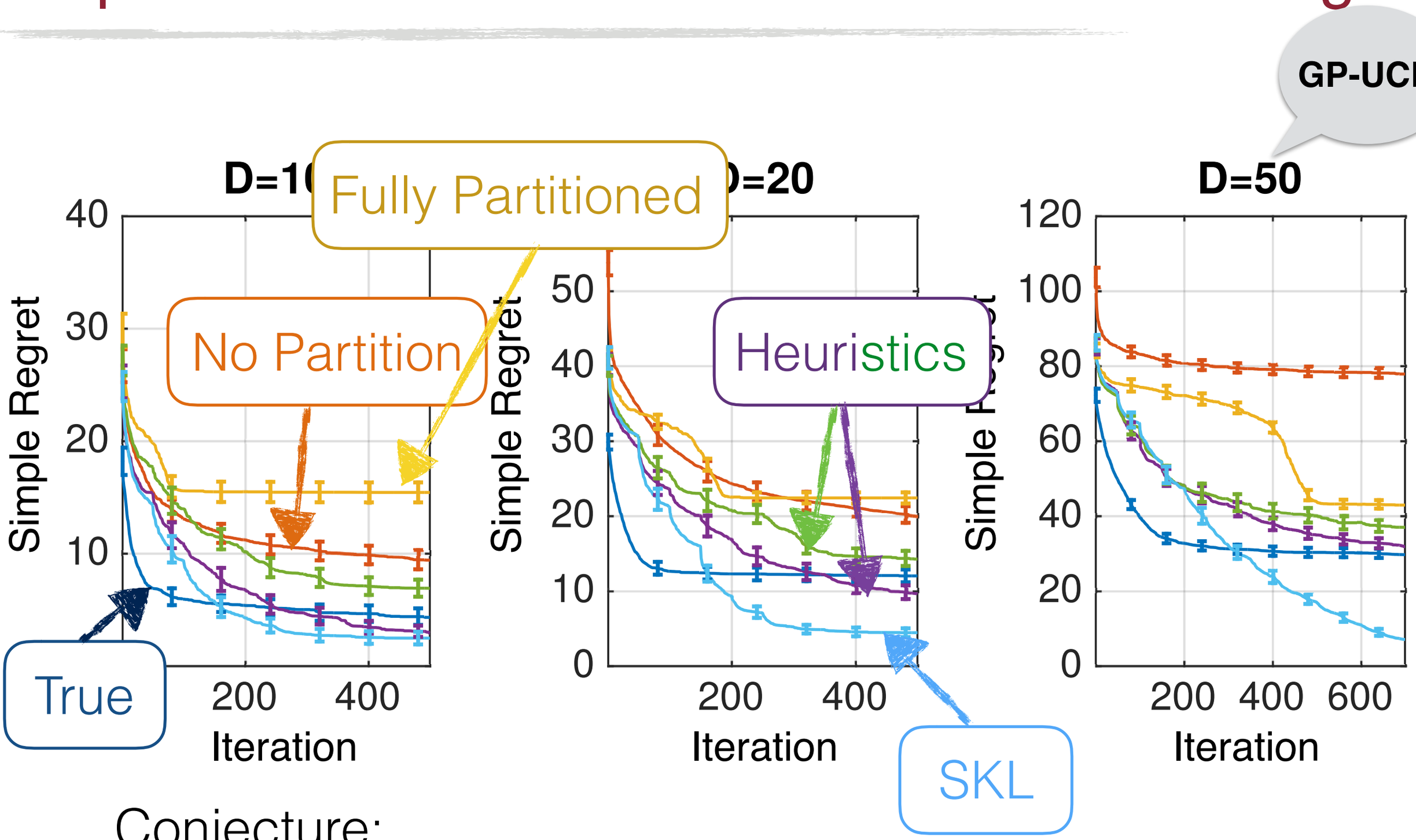
☑ Quality

☑ Diversity



(Kathuria et al., 2016)

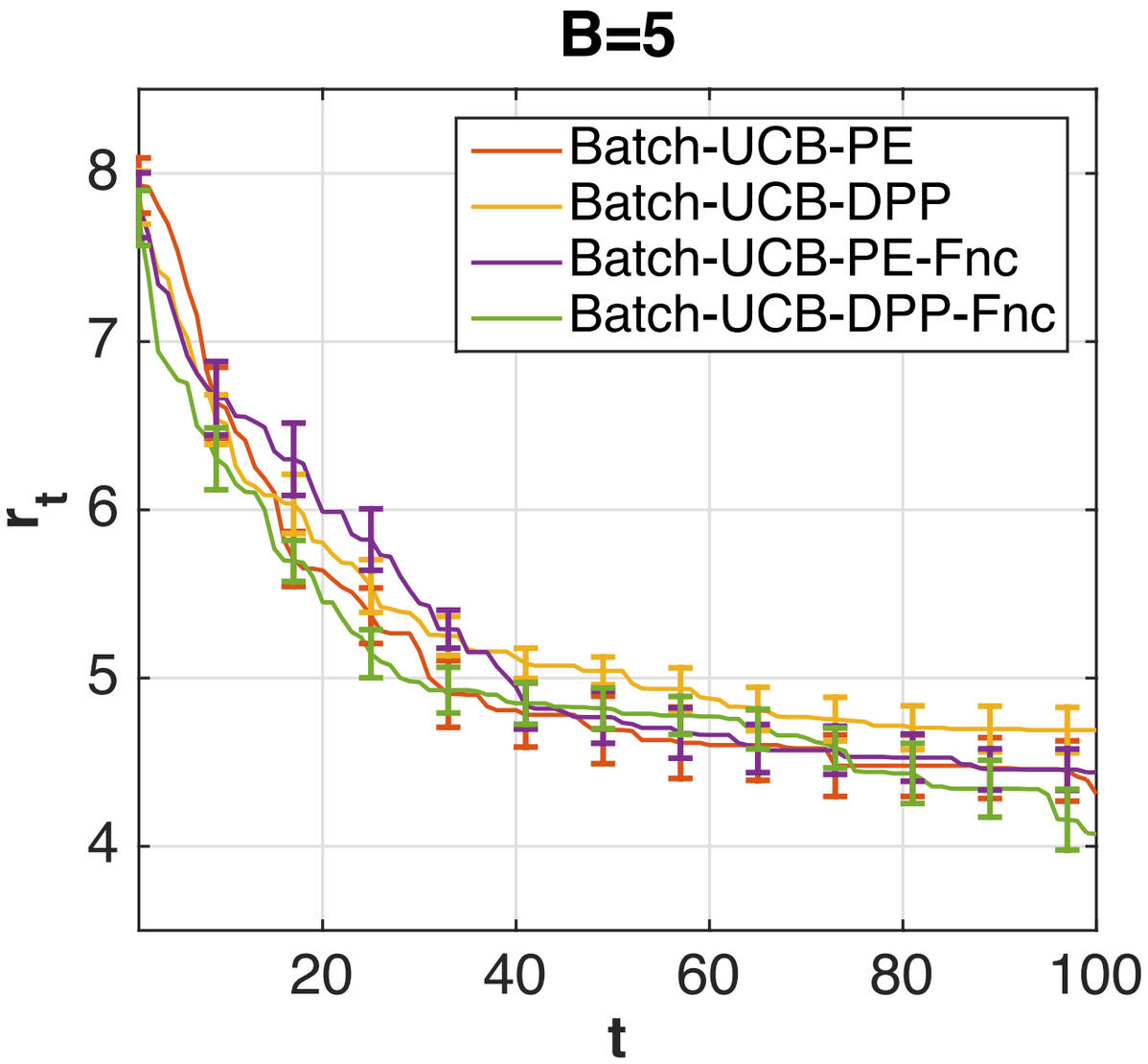
Empirical Results of Structural Kernel Learning

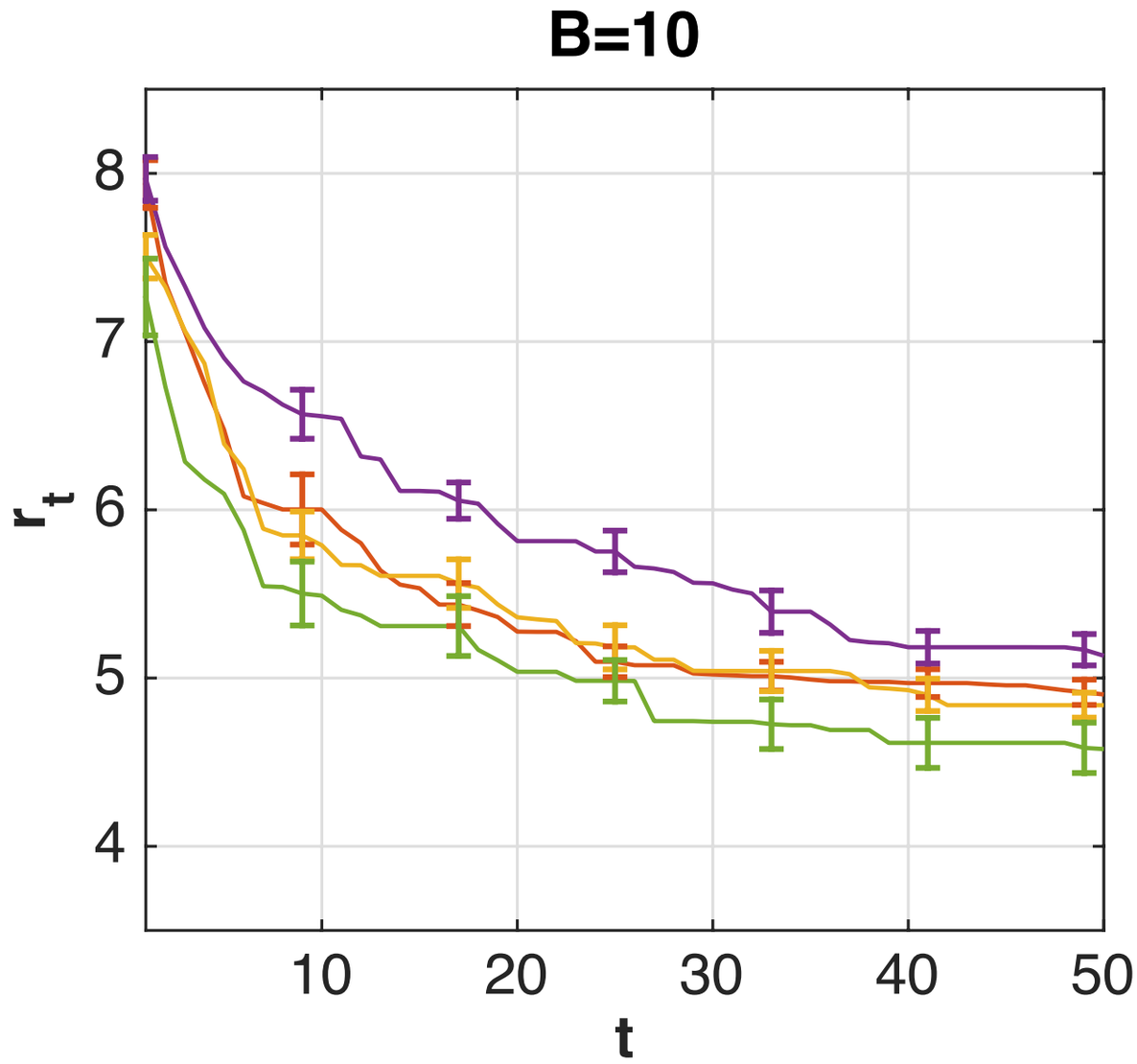


Conjecture:

SKL is better than True because of more exploration

Empirical Results of Batched BO with DPP





Summary

- High-dimensional BO with learned additive kernel structure
- Batched high-dimensional BO with DPP

Open Questions

- How to scale up the observations/speed up the inference?
- Why additive structures are useful?

http://zi-wang.com/#publications

More results @ Gallery #133