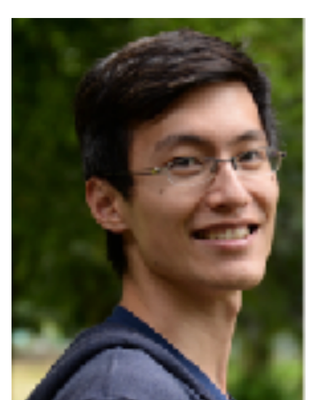
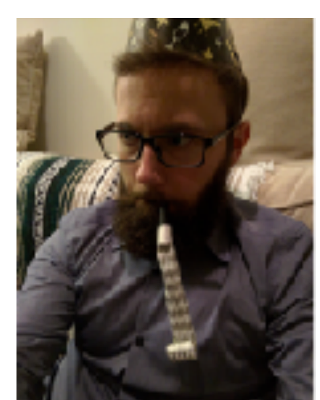
Massachusetts Institute of Technology

Scaling Bayesian Optimization in High Dimensions

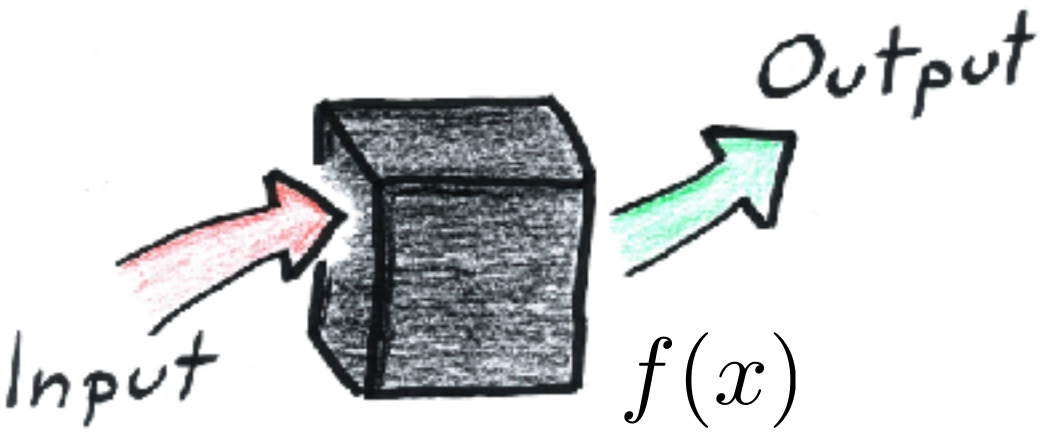
Stefanie Jegelka, MIT

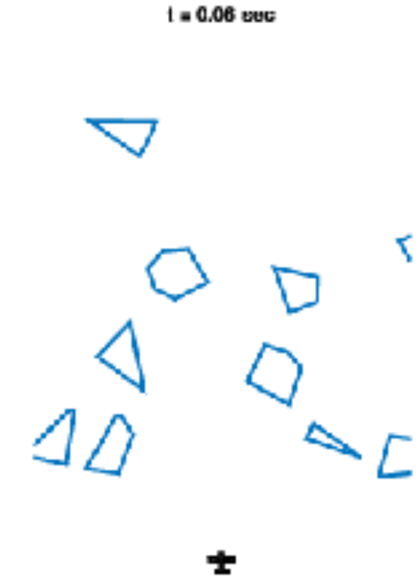
BayesOpt Workshop 2017

joint work with *Zi Wang*, Chengtao Li, Clement Gehring (MIT) and Pushmeet Kohli (DeepMind)


Bayesian Optimization with GPs





BO: sequentially build model of *f*

for *t*=*1*, … *T*:

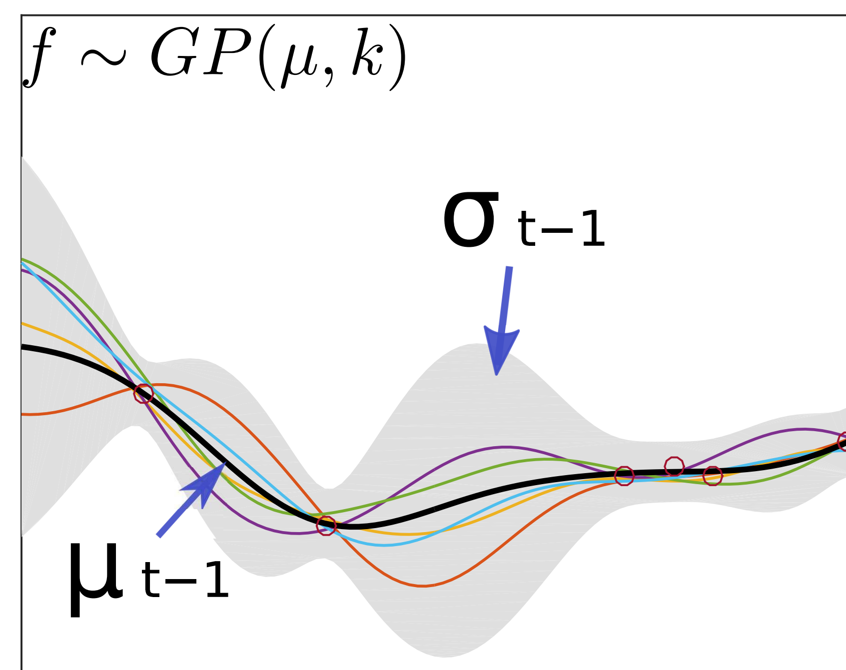
- select new query point(s) *x*

  selection criterion: acquisition function

  $$\arg\max_{x \in \mathfrak{X}} \ \alpha_t(x)$$

- observe *f*(*x*)
- update model & repeat

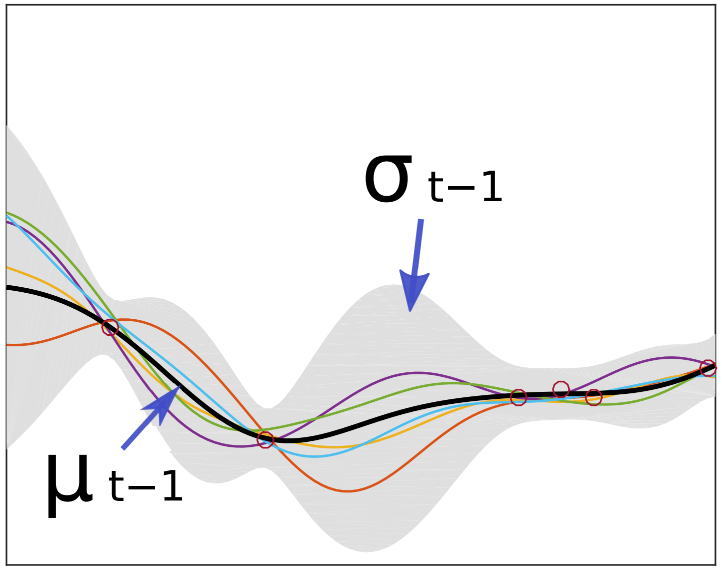
Gaussian process:
closed form expressions for posterior mean and variance (uncertainty)



Challenges in high dimensions

statistical & computational complexity:

- estimating & optimizing acquisition function
- function estimation in high dimensions
- many observations (data points): huge matrix in GP
- parallelization



(Predictive) Entropy Search

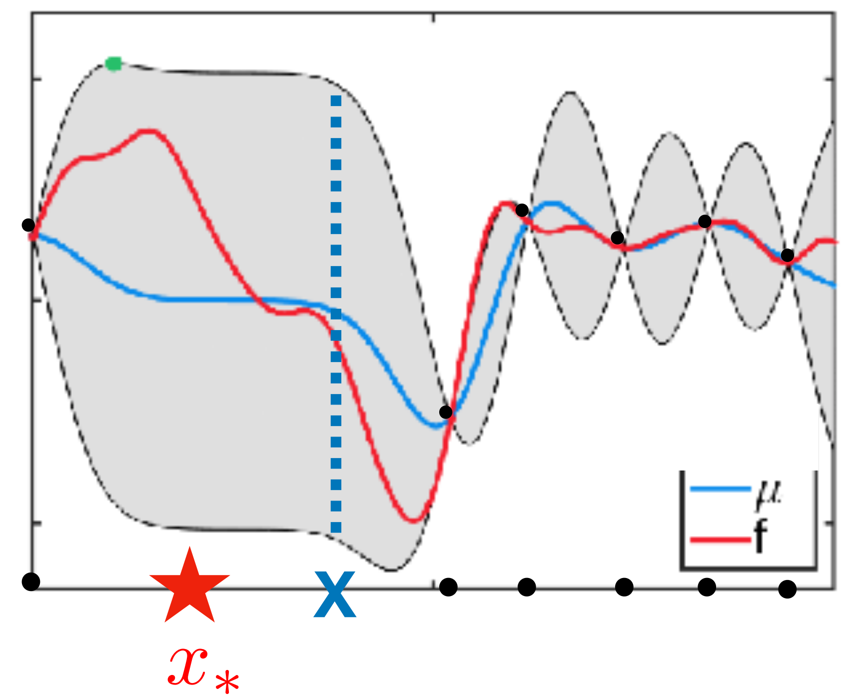
new query point: $\arg\max_{x \in \mathfrak{X}} \; \alpha_t(x)$

Point to query

Location of global optimum

Observed Data

$$\alpha_t(x) = I(\{x, y\}; x_* \mid D_t)$$



$$= H\left(p(x_* \mid D_t)\right) - \mathbb{E}\left[H(p(x_* \mid D_t \cup \{x, y\}))\right] \quad \text{ES}$$

$$= H(p(y \mid D_t, x)) - \mathbb{E}\left[H(p(y \mid D_t, x, x_*))\right] \quad \text{PES}$$

$$I(a; b) = H(a) - H(a|b)$$
$$= H(b) - H(b|a)$$

if x^* is high-dimensional: $\alpha_t(x)$ costly to estimate!

(Hennig & Schuler, 2012; Hernandez-Lobato, Hoffman & Ghahramani 2014)

Max-value Entropy Search

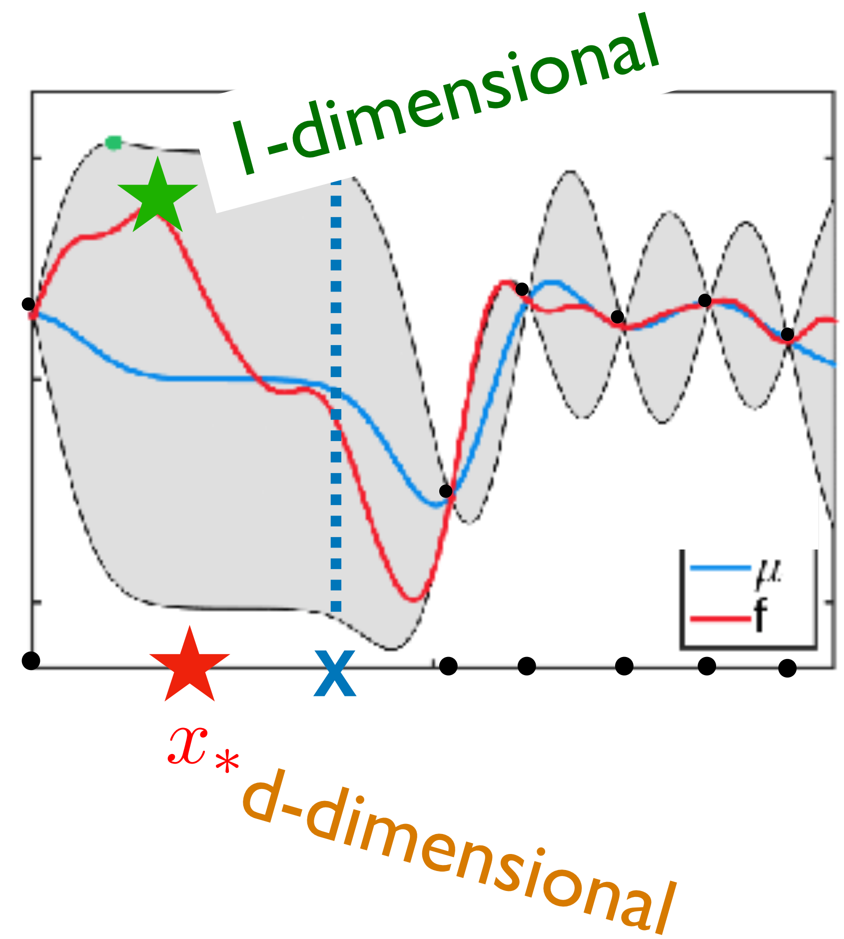
Query Point

Observed Data

$$\alpha_t(x) = I(\{x, y\}; x_* \mid D_t)$$ Input space

$$\alpha_t(x) = I(\{x; y\}; y_* \mid D_t)$$ **Output space**

$d \to 1$ **dimensions!**



$$\approx \frac{1}{K} \sum_{y_* \in Y_*} \left[\frac{\gamma_{y_*}(x)\psi(\gamma_{y_*}(x))}{2\Psi(\gamma_{y_*}(x))} - \log(\Psi(\gamma_{y_*}(x))) \right]$$

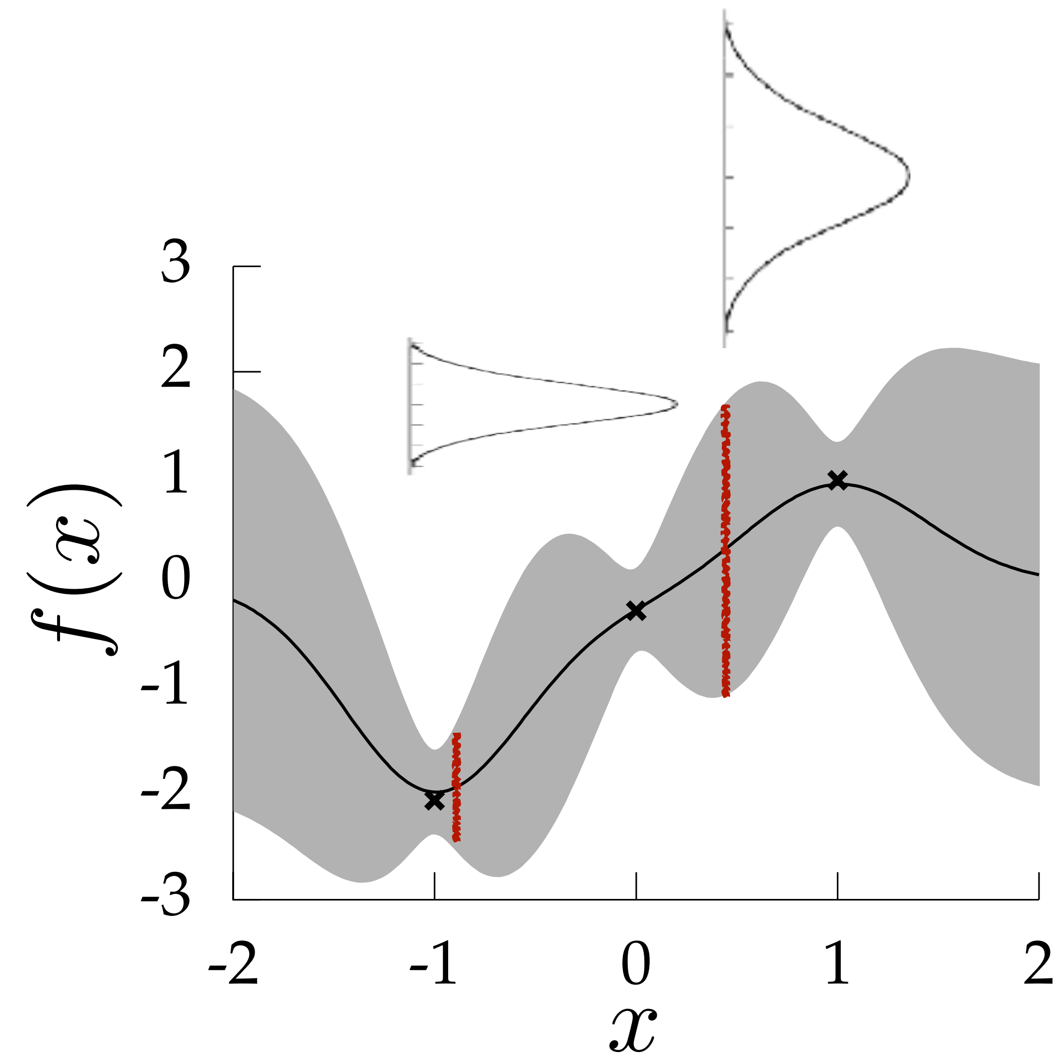
closed-form

Expectation over $p(y_*|D_t)$. How sample y_*?

Sampling y*: Idea 1

$p(f(x))$ is a 1D Gaussian

Fisher-Tippett-Gnedenko Theorem
The maximum of a set of i.i.d. Gaussian variables is asymptotically described by a Gumbel distribution.

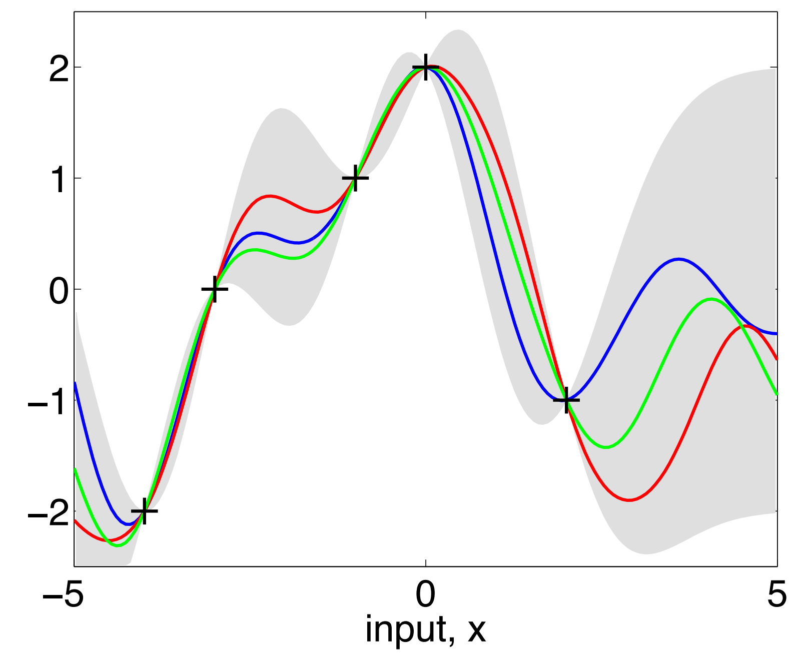


- sample representative points

- approximate max-value of the representative points by a Gumbel distribution

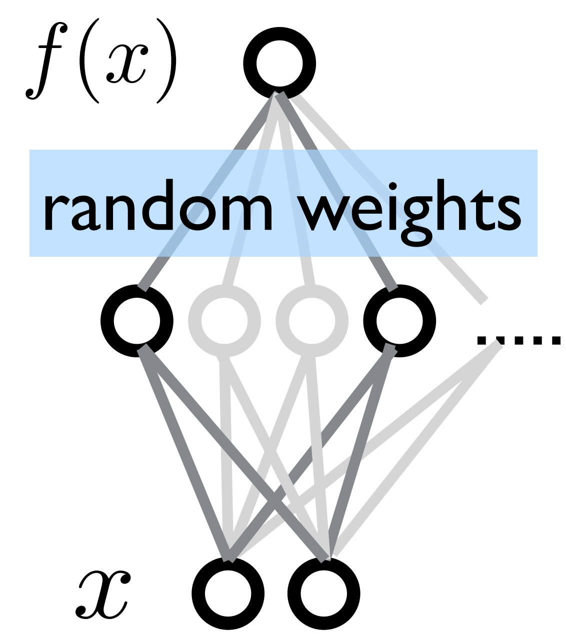
Sampling y*: Idea 2

draw functions from GP posterior
and maximize each. How?



Neal 1994:
GP $\equiv$ infinite 1-layer neural network with Gaussian weights.

- approximate GP as finite neural network
  (random features)
  & sample posterior weights
- maximize network output for each sample



(Hernández-Lobato, Hoffman & Ghahramani 2014)

Max-value Entropy Search

Query Point

Observed Data

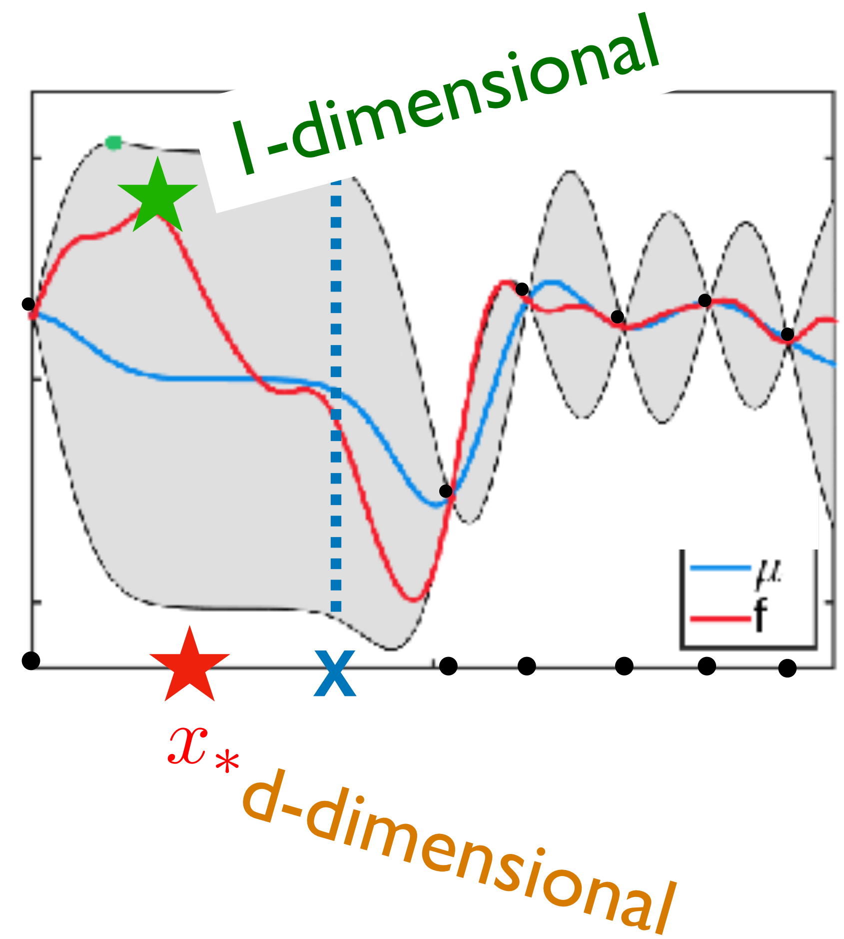
$$\alpha_t(x) = I(\{x, y\}; x_* \mid D_t)$$

Input space

$$\alpha_t(x) = I(\{x; y\}; y_* \mid D_t)$$

Output space

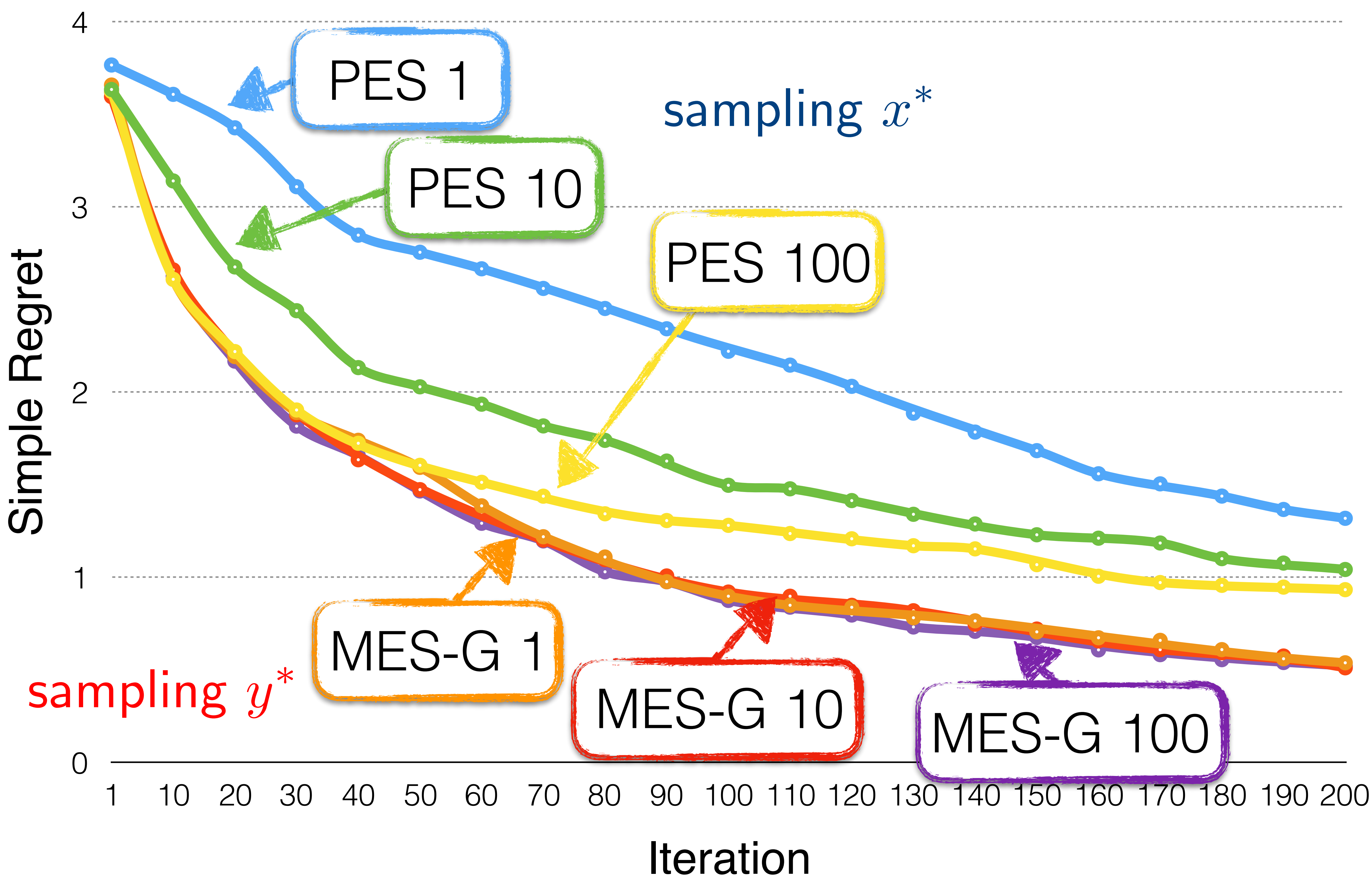
$d \rightarrow 1$ **dimensions!**



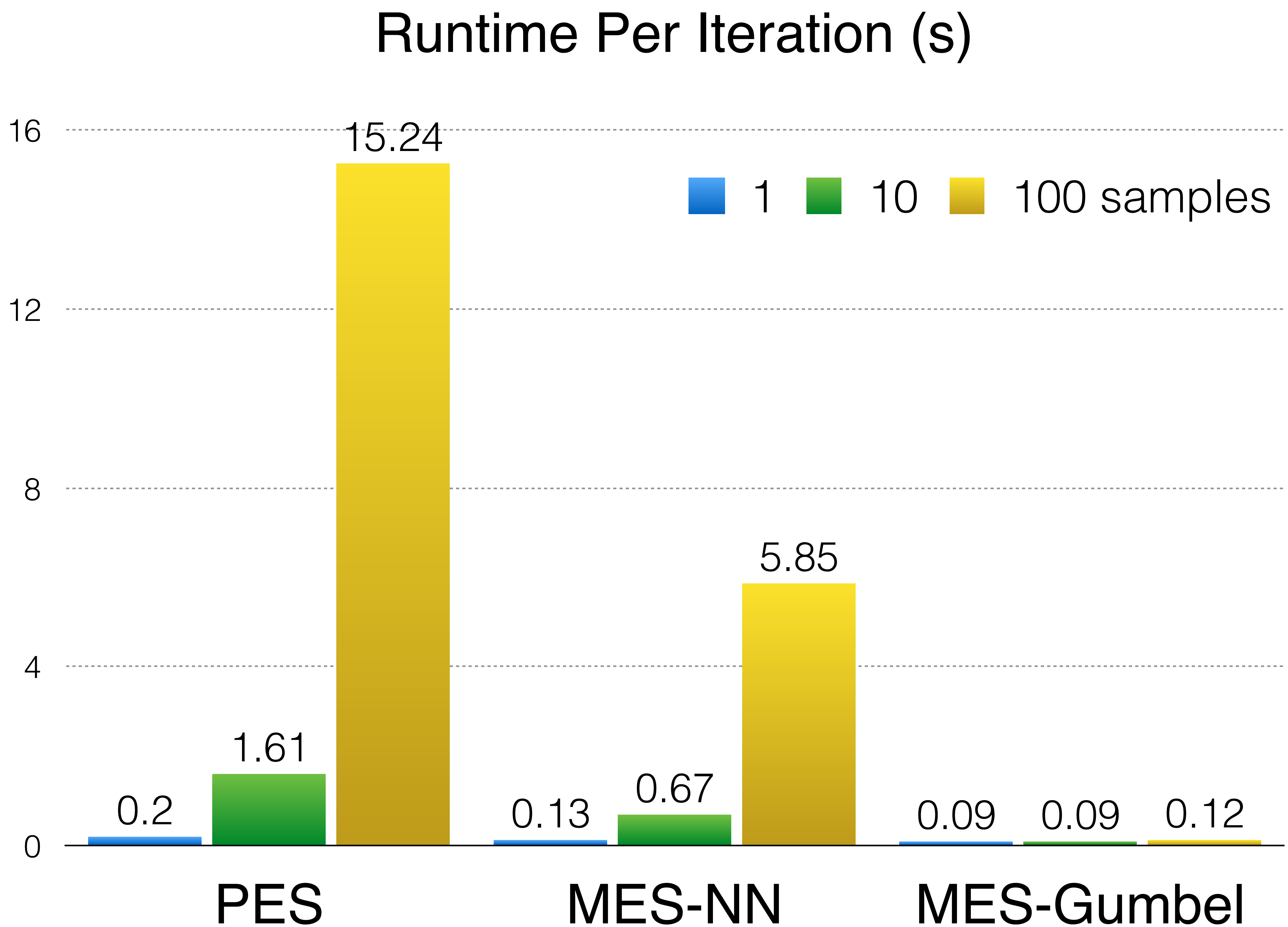
Expectation over $p(y_*|D_t)$. Can sample y_*!

Does it work?

Empirically: max-value enough? sample-efficiency?



Empirically: faster than PES



Connections & Theory

zoo of acquisition functions: EI *(Mockus, 1974)*, PI *(Kushner, 1964)*, GP-UCB *(Auer, 2002; Srinivas et al., 2010)*, GP-MI *(Contal et al., 2014)*, ES *(Hennig & Schuler, 2012)*, PES *(Hernández-Lobato et al., 2014)*, EST *(Wang et al., 2016)*, GLASSES *(González et al., 2016)*, SMAC *(Hutter et al., 2010)*, ROAR *(Hutter et al., 2010)*, … MES

Lemma *(Wang-J17)* *Equivalent* acquisition functions:

- MES with a single sample of y_* per step
- UCB (upper confidence bound, *Srinivas et al., 2010*)
- PI (probability of improvement, *Kushner, 1964*)

with specific, *adaptive* parameter setting

Theorem: Regret bound *(Wang-J 17)*

With probability $1-\delta$, within $T' = O(T \log \delta)$ iterations:

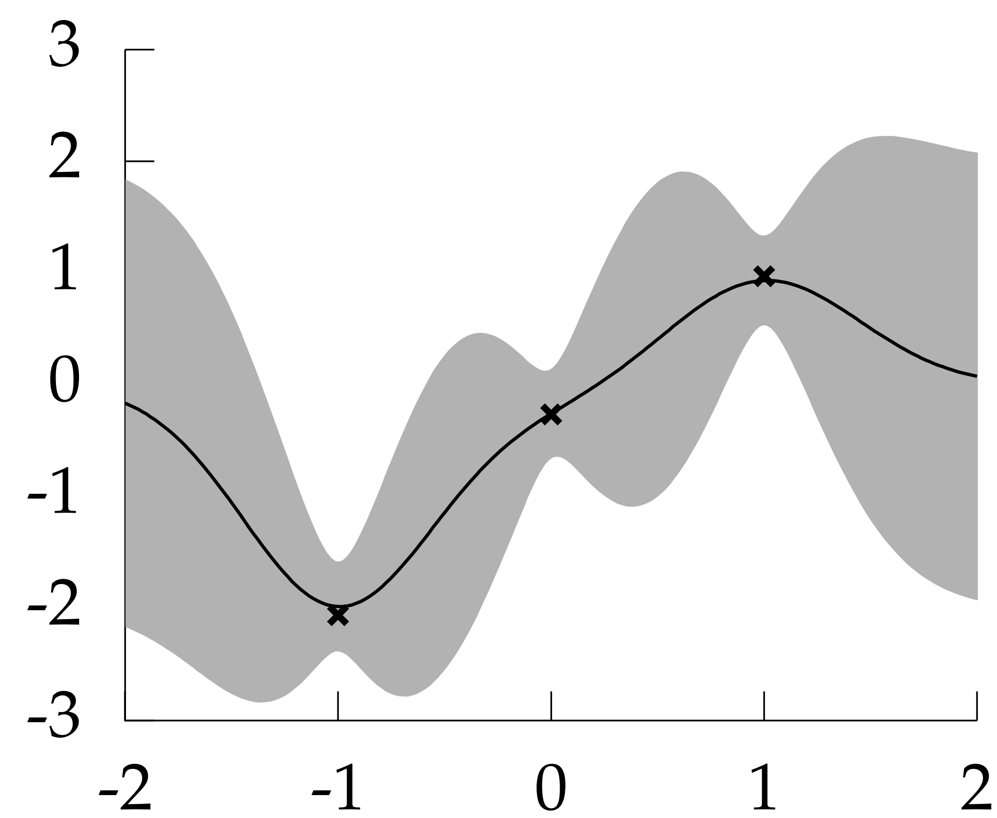
$$f^* - \max_{t \in [1,T']} f(x_t) = O\Big(\sqrt{\frac{(\log T)^{d+2}}{T}}\Big)$$

Gaussian Processes in high dimensions

- estimating a nonlinear function in high input dimensions:
  statistically challenging

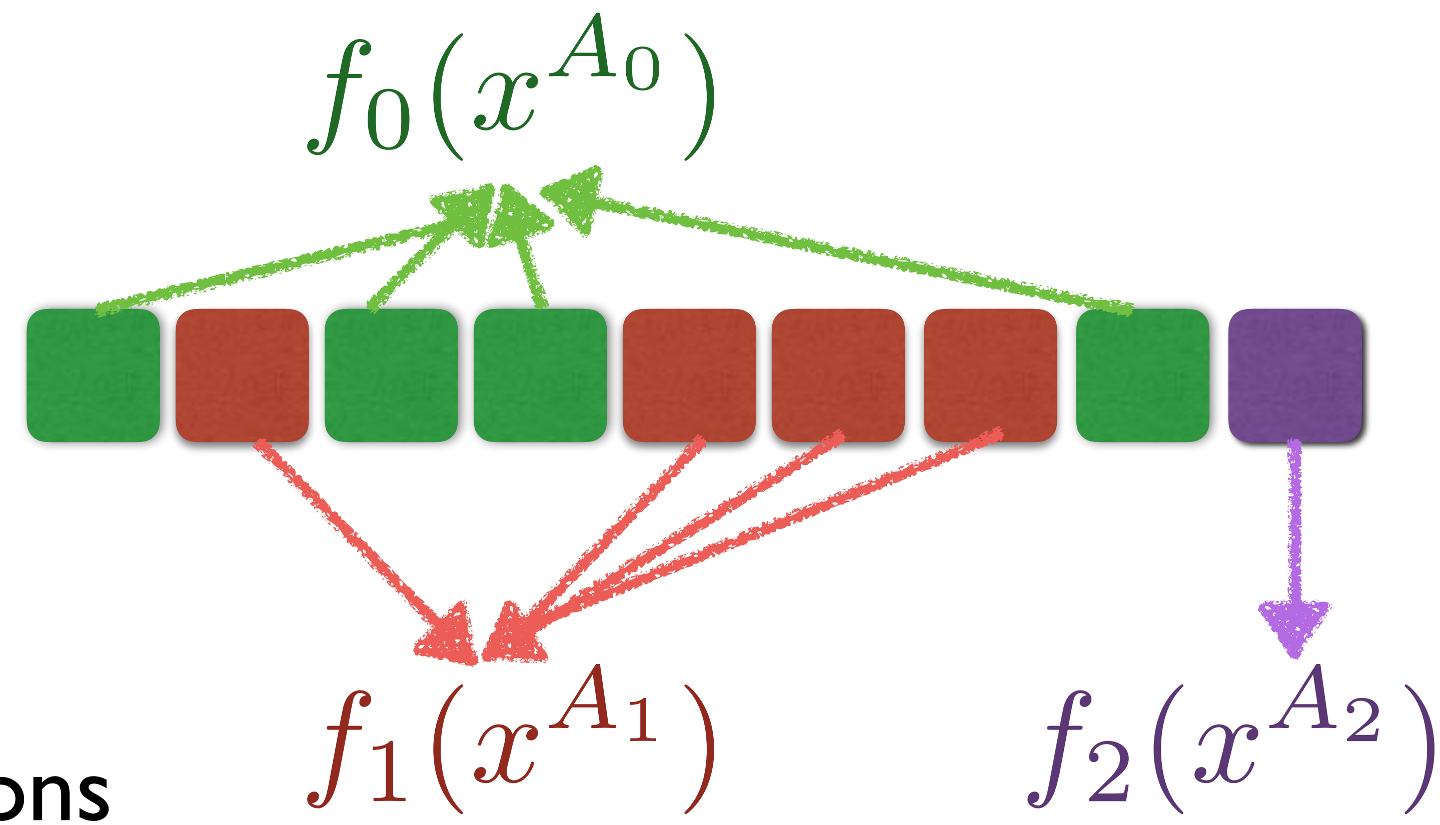
- optimizing nonconvex acquisition function in high dimensions
  computationally challenging

- many observations: huge matrices
  computationally challenging



Additive Gaussian Processes

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



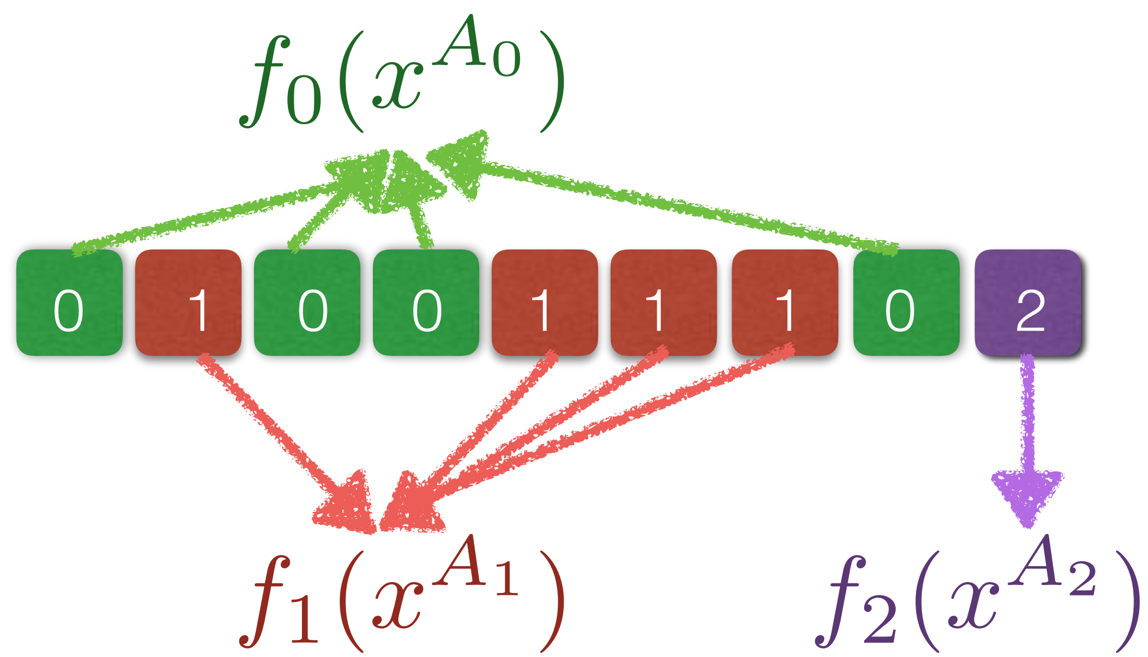
- lower-complexity functions
  statistical efficiency
- optimize acquisition function block-wise
  computational efficiency

What is the partition?

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

Structural Kernel Learning

$$f = f_0 + f_1 + f_2$$

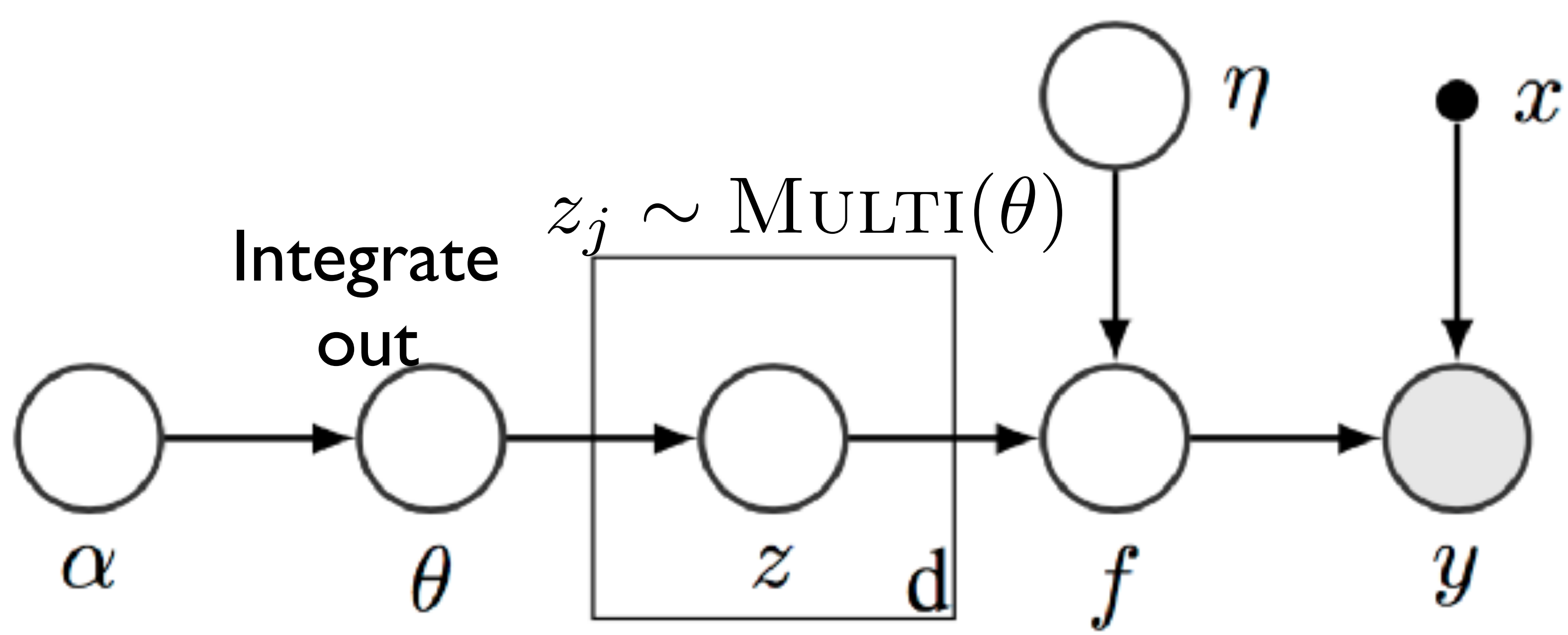


z = [0 1 0 0 1 1 1 0 2]

Learn the assignment!

Key idea:

Dirichlet prior on z



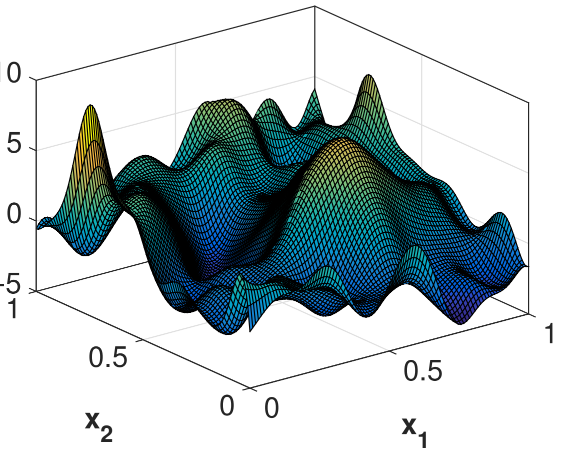
Posterior

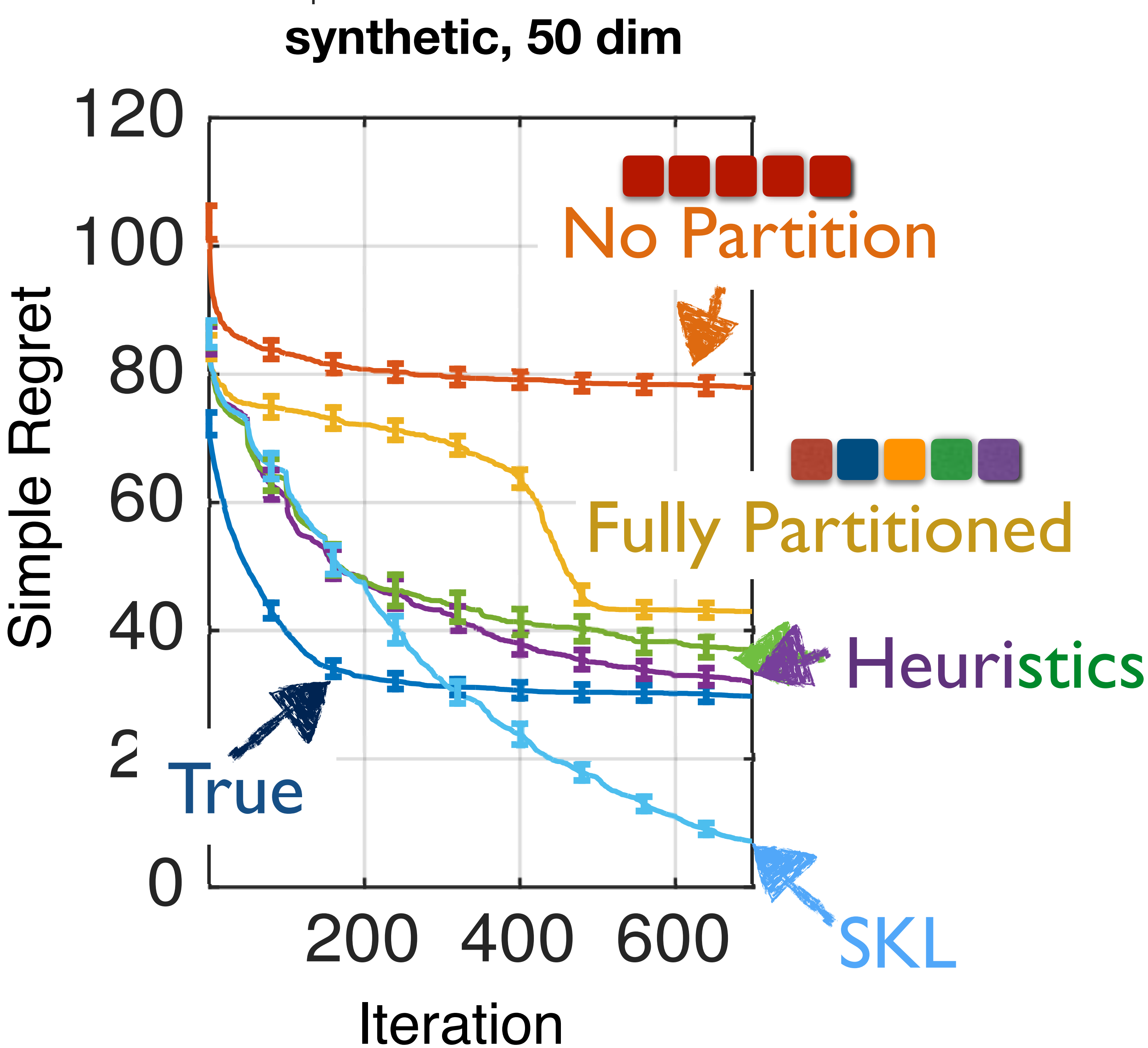
$$p(z \mid D_n; \alpha)$$

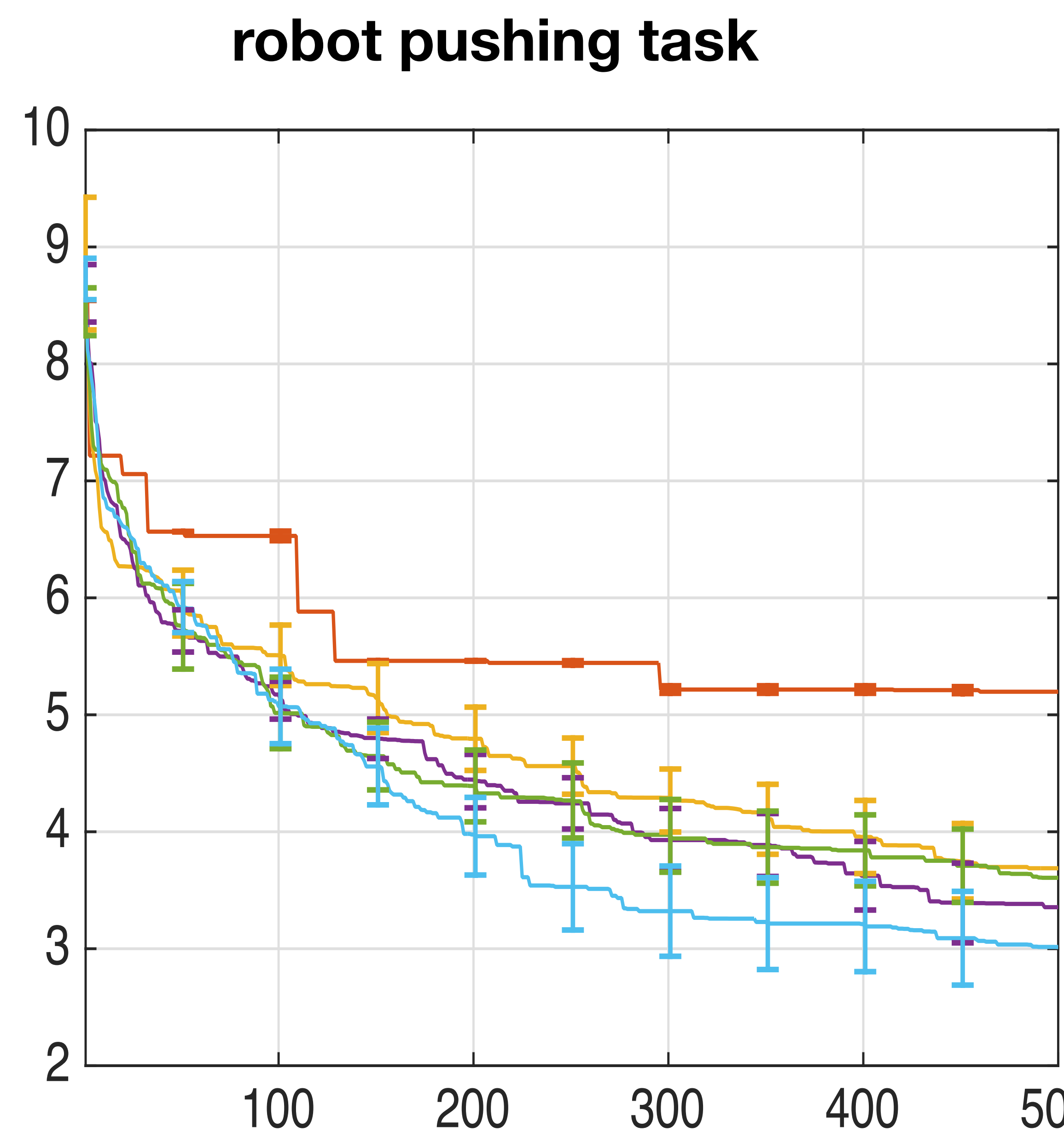
via Gibbs sampling.

easy updates

Empirical Results

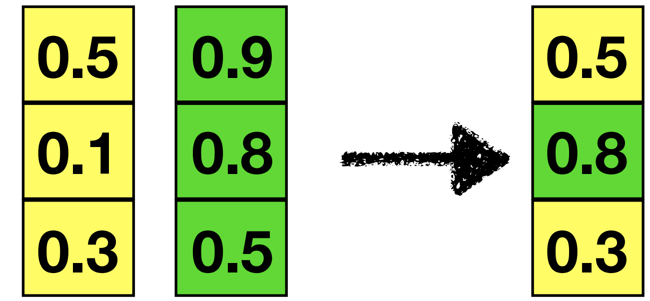


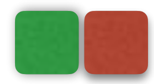


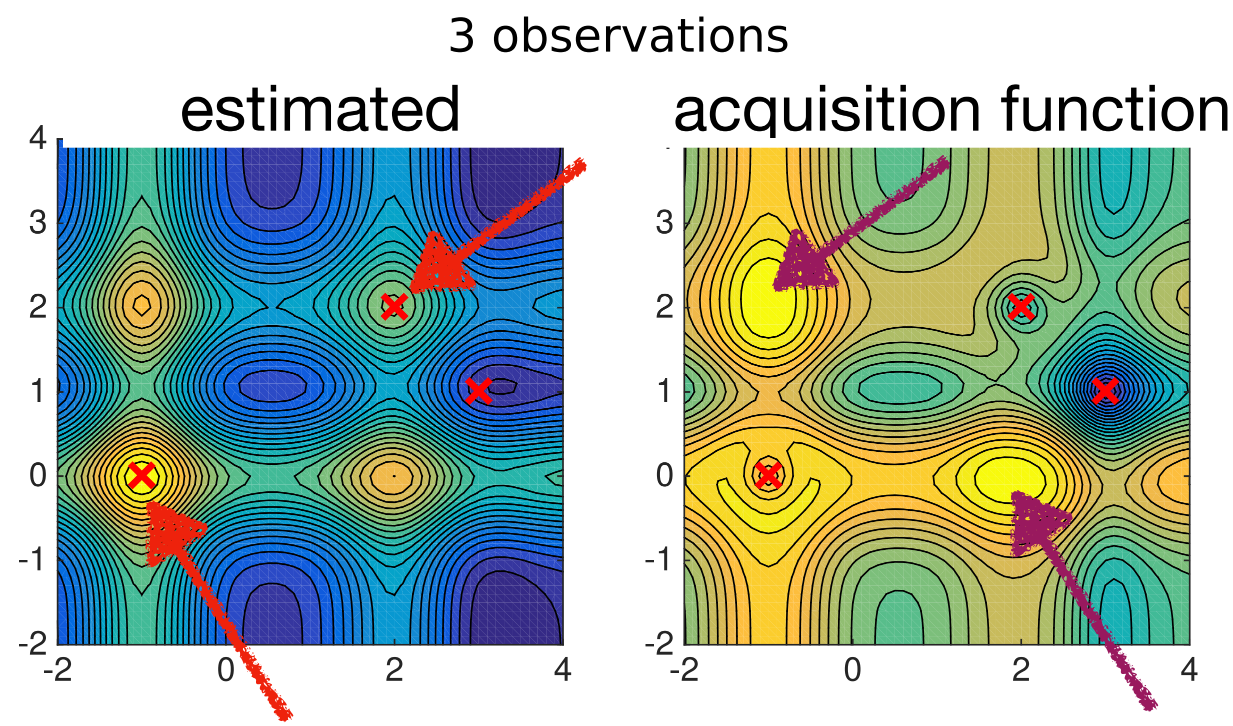


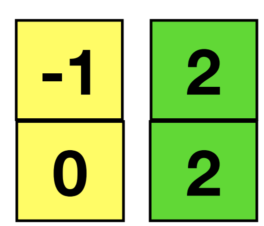
Curious connections

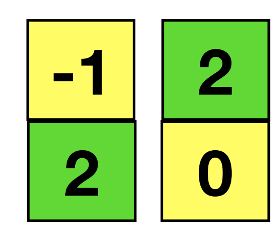
- crossover in evolutionary algorithms:



- BO with additive GP: 



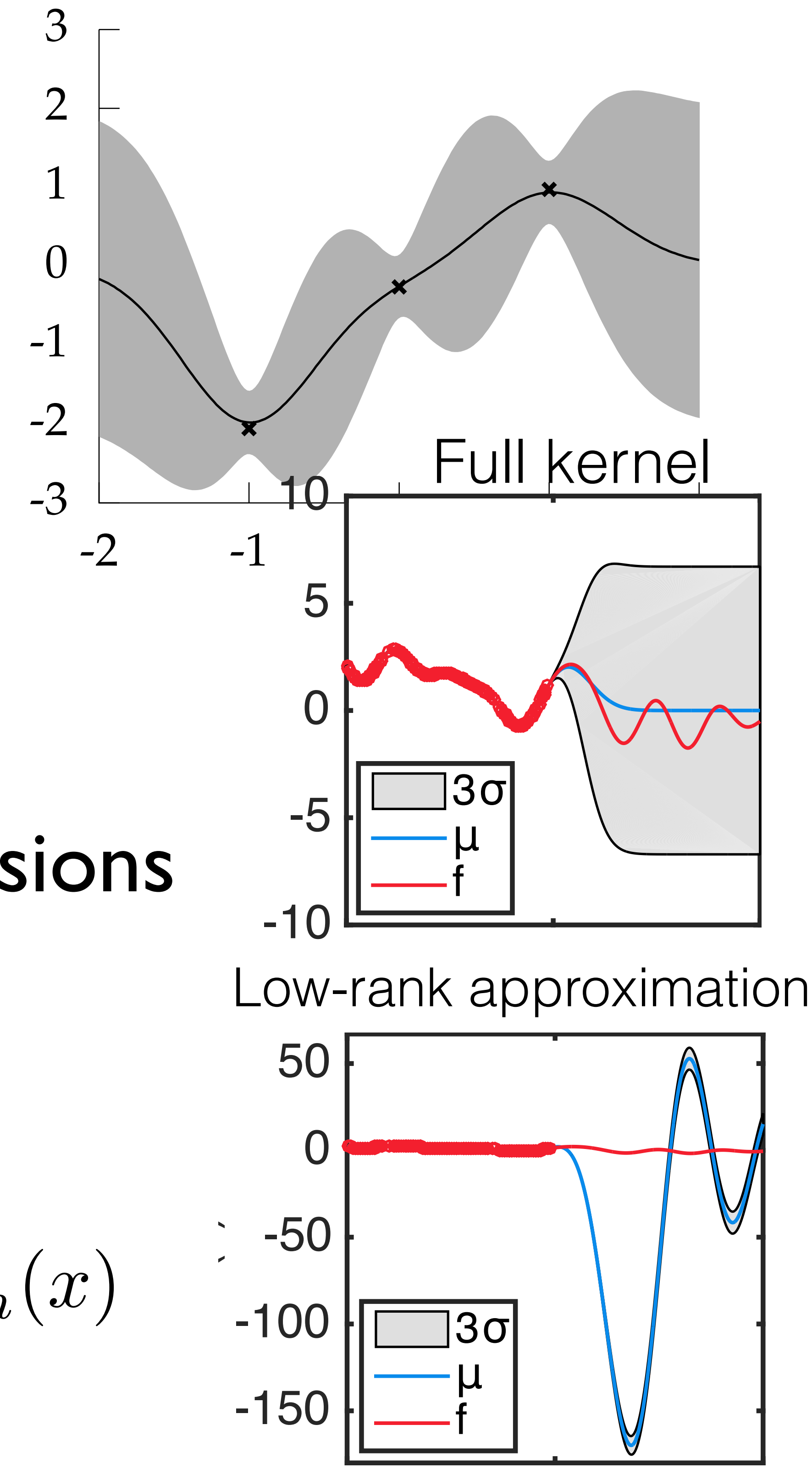
- observed good points: 

query points: 

learned instead of completely random coordinate partition

Gaussian Processes in high dimensions

- estimating nonlinear functions in high input dimensions: *statistically challenging*
- optimizing nonconvex acquisition function in high dimensions *computationally challenging*

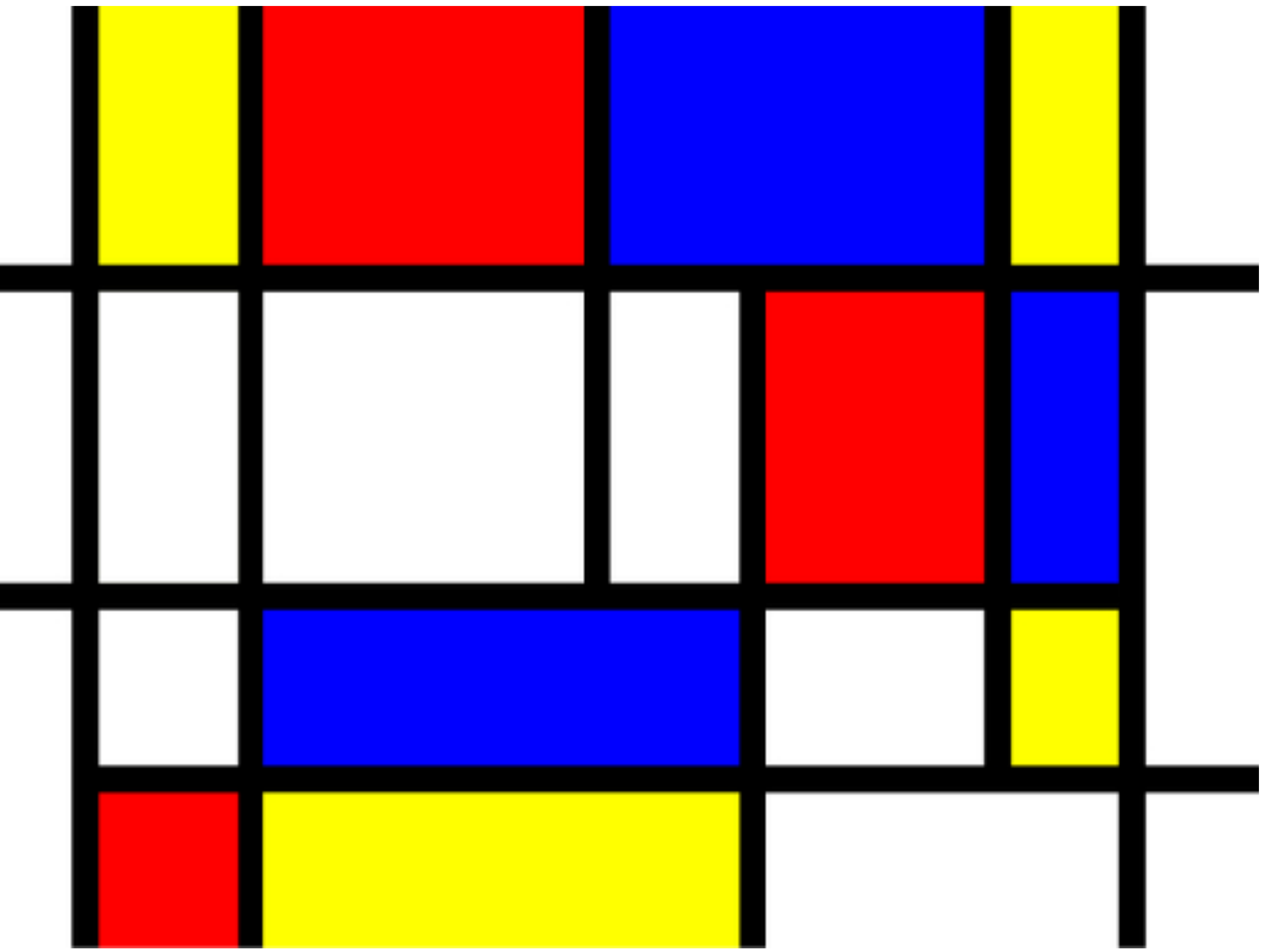


- **many observations:** huge matrix inversions *computationally challenging*

$$\mu(x) = \mathbf{k}_n(x)^\top (\mathbf{K}_n + \tau^2 \mathbf{I})^{-1} \mathbf{y}_t$$

$$\sigma^2(x) = k(x,x) - \mathbf{k}_n(x)^\top (\mathbf{K}_n + \tau^2 \mathbf{I})^{-1} \mathbf{k}_n(x)$$

Ensemble Bayesian Optimization

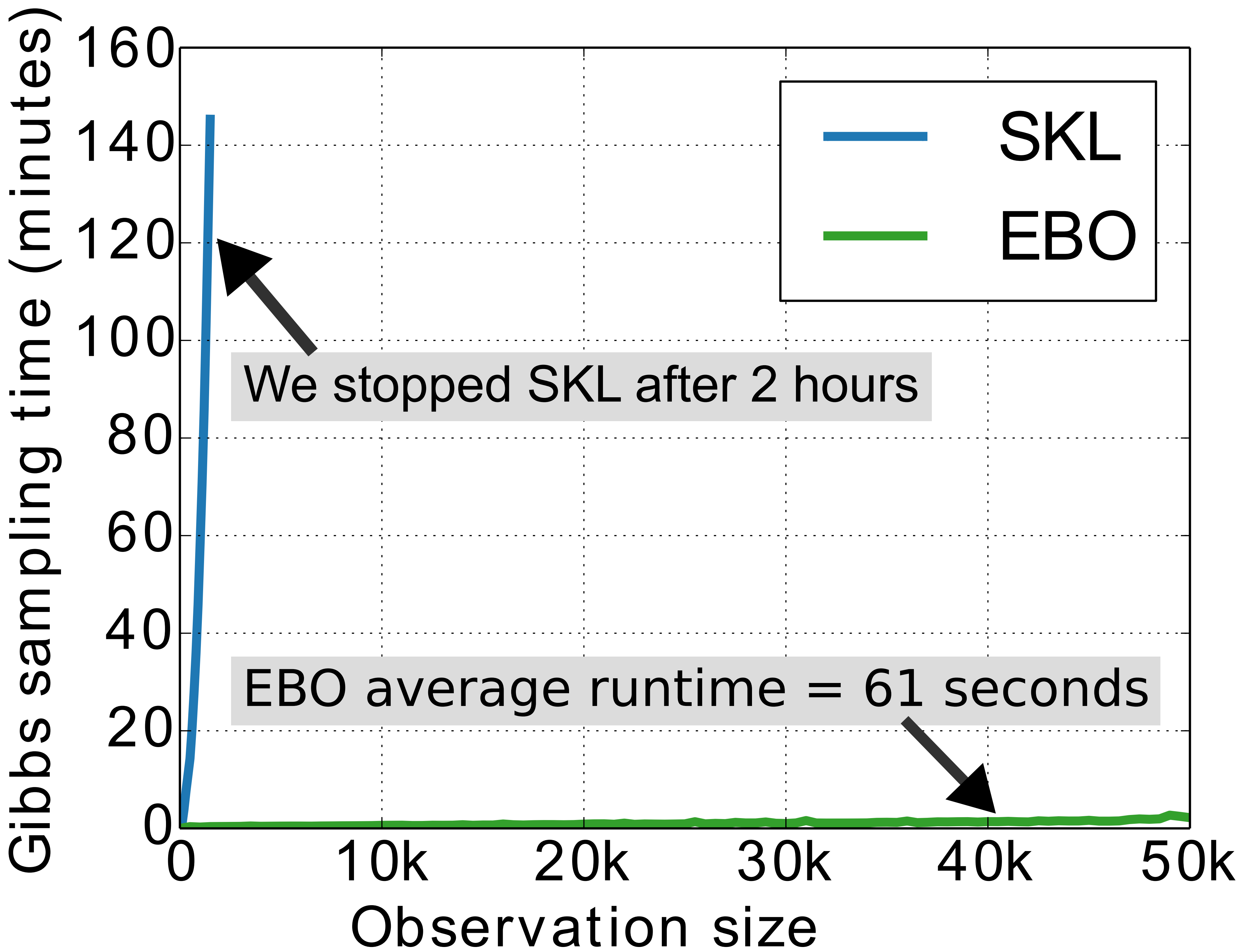


in each iteration:

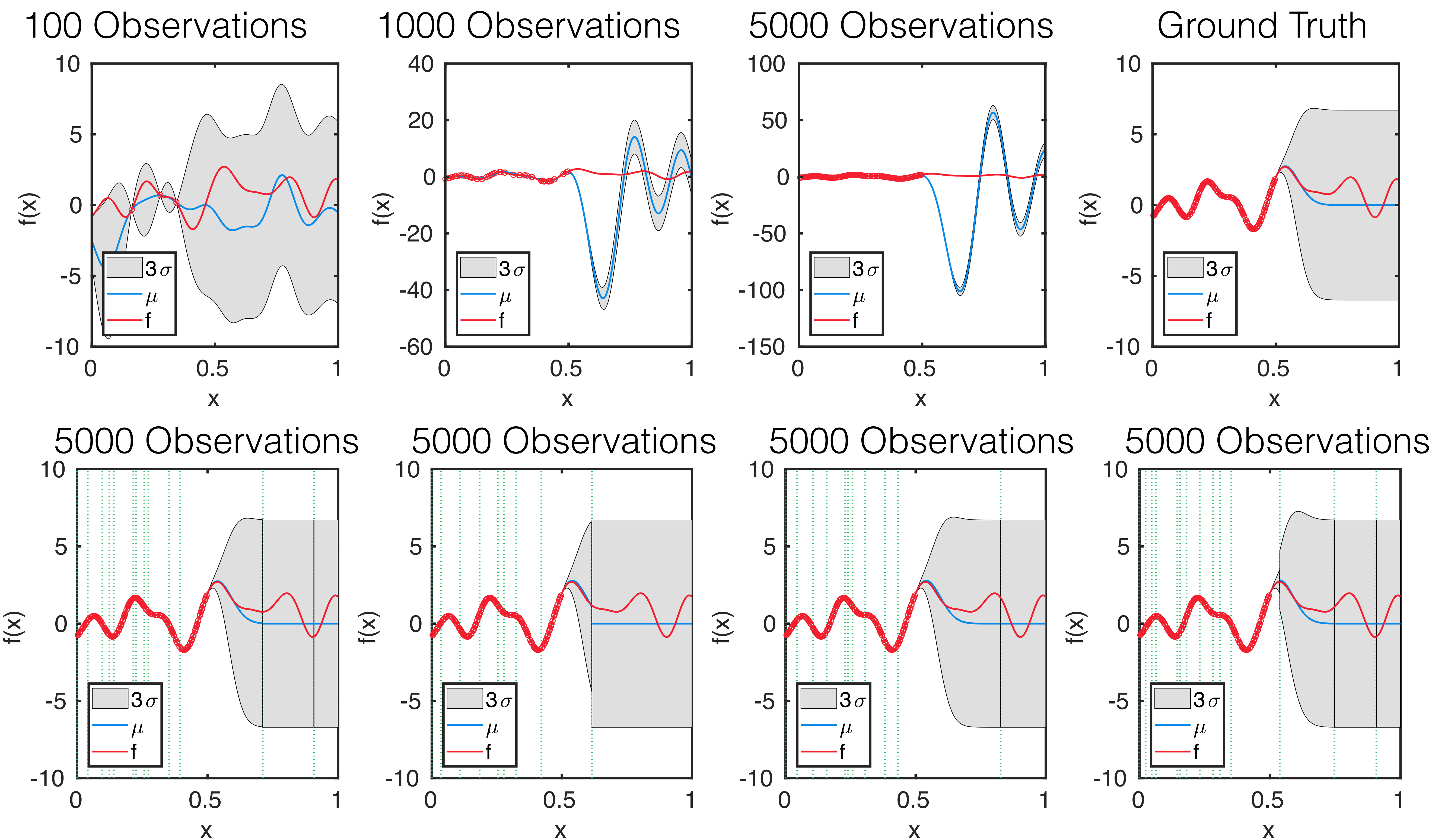
- partition data via Mondrian process
- fit GP in each part: structure learning + Tile Coding; synchronize
- select query points in parallel & filter

parallelization across parts
distribution over partitions — new draw in each iteration

Does it scale?

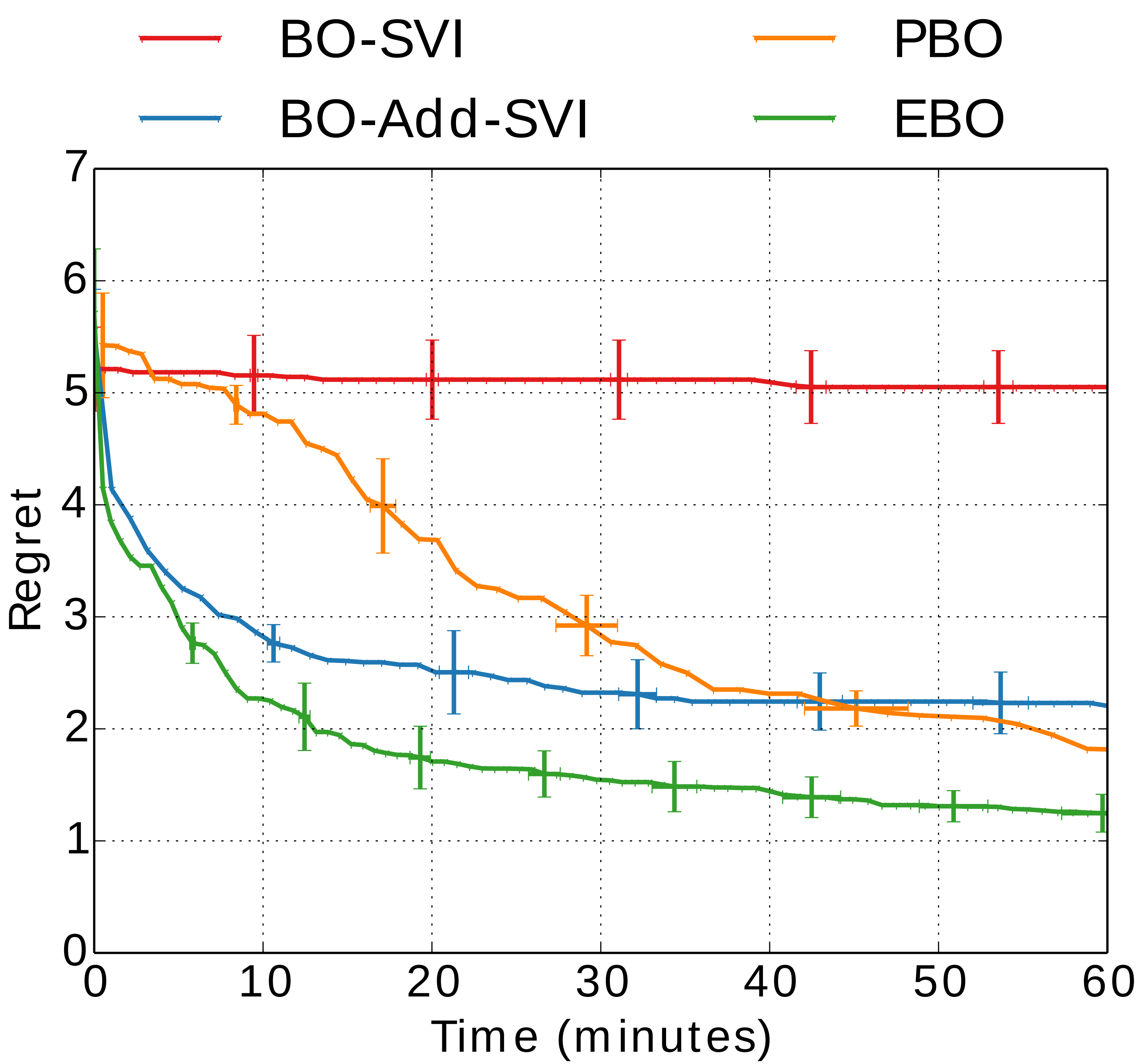


Variances



100 Observations
1000 Observations
5000 Observations
Ground Truth
5000 Observations
5000 Observations
5000 Observations
5000 Observations
f(x)
x
3σ
μ
f

Empirical Results



(Hensman et al., 2013, Wang et al., 2017)

Summary: GP-BO in high dimensions

Challenge: high dimensions, many observations
statistical & computational efficiency

- **Max-value Entropy Search**
  sample-efficient, effective acquisition function
  (Wang, Jegelka, ICML 2017)

- **Many dimensions: learning structured kernels**
  (Wang, Li, Jegelka, Kohli, ICML 2017)

- **Many observations & dimensions & parallelization: ensemble Bayesian Optimization**
  (Wang, Gehring, Kohli, Jegelka, BayesOpt 2017)

References

- Zi Wang, Stefanie Jegelka. Max-value entropy search for efficient Bayesian Optimization. ICML 2017.

- Zi Wang, Chengtao Li, Stefanie Jegelka, Pushmeet Kohli. Batched High-dimensional Bayesian Optimization via Structural Kernel Learning. ICML 2017.

- Zi Wang, Clement Gehring, Pushmeet Kohli, Stefanie Jegelka. Batched Large-scale Bayesian Optimization in High-dimensional Spaces. BayesOpt, 2017.