Bayesian Optimization

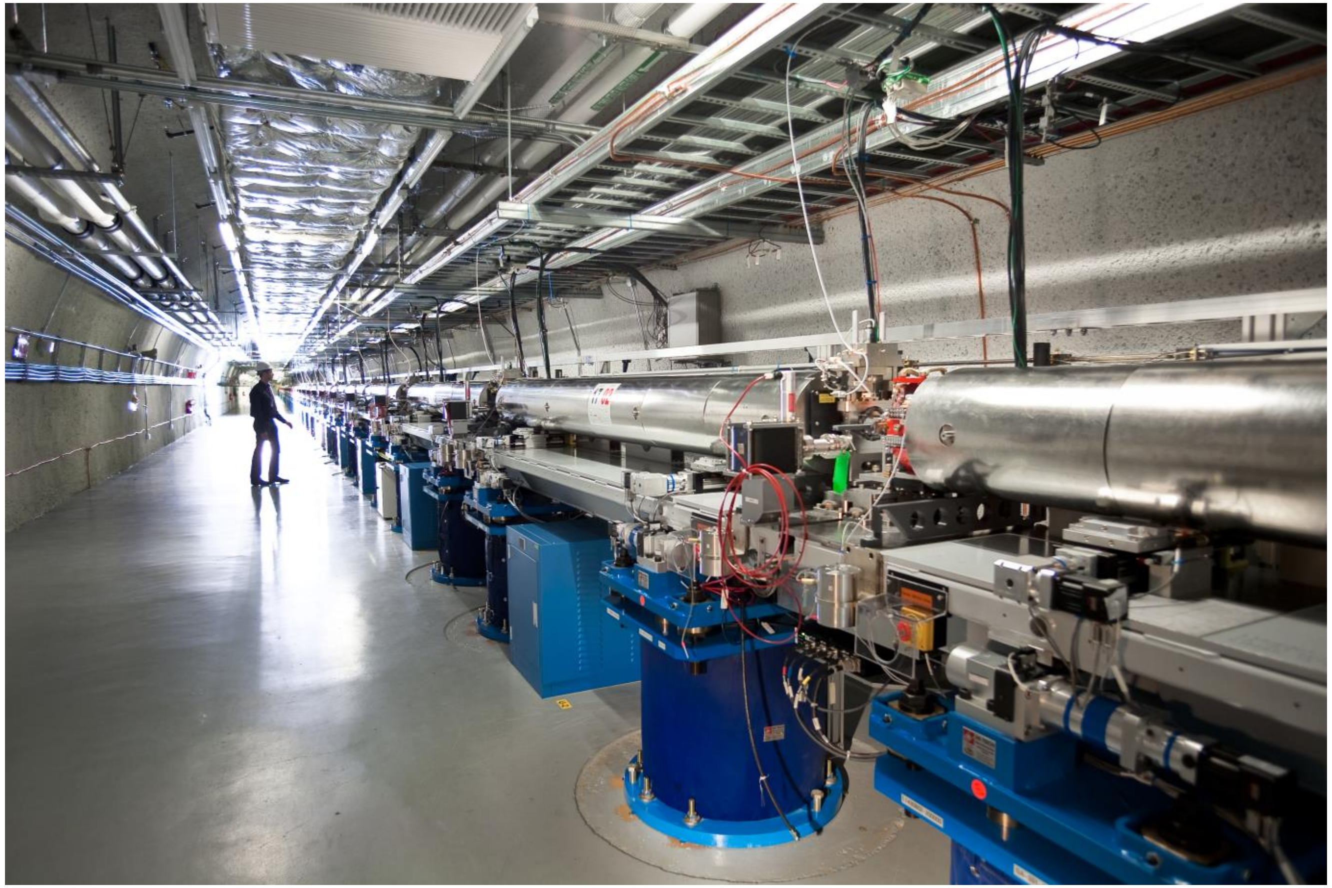
Presenter: Adi Hanuka

Day 5

Outline

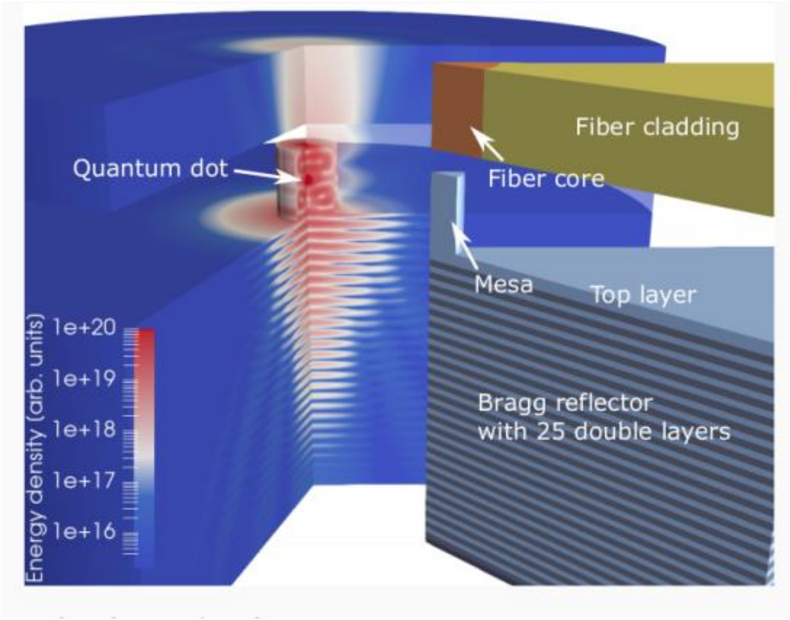
- Motivation
- Model-based vs model-free optimizers
- Bayesian Optimization (BO)
  - Overview
  - Acquisition functions
  - Accounting for constraints
  - Proximal optimization
- Applications
- Summary of optimization methods

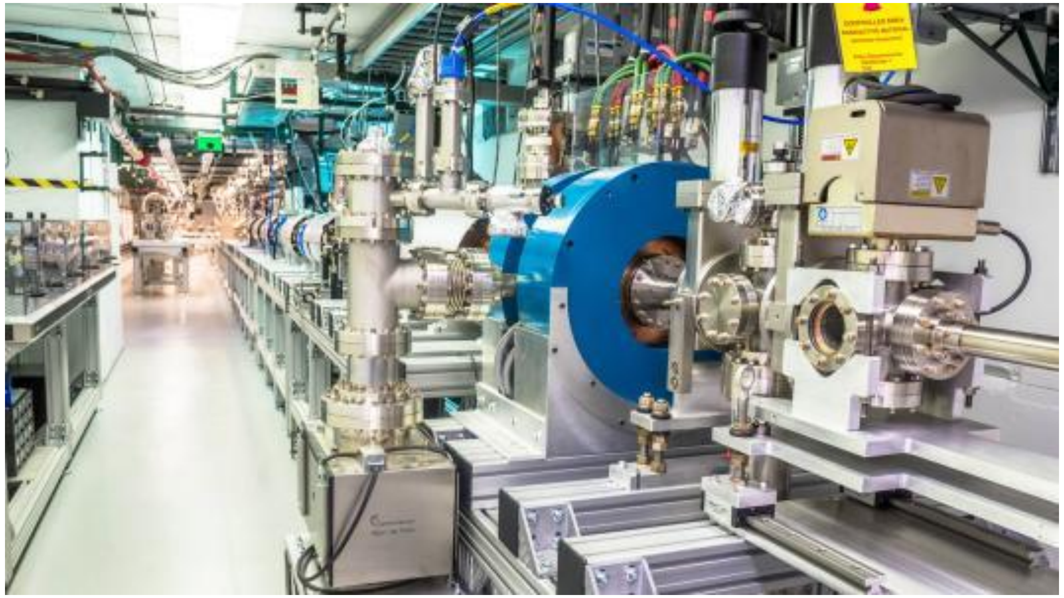
Motivating Application: Parameter Tuning of Accelerator



Optimize operations: maximize X-ray energy, minimize emittance, ….

Motivating Application: Experimental Design

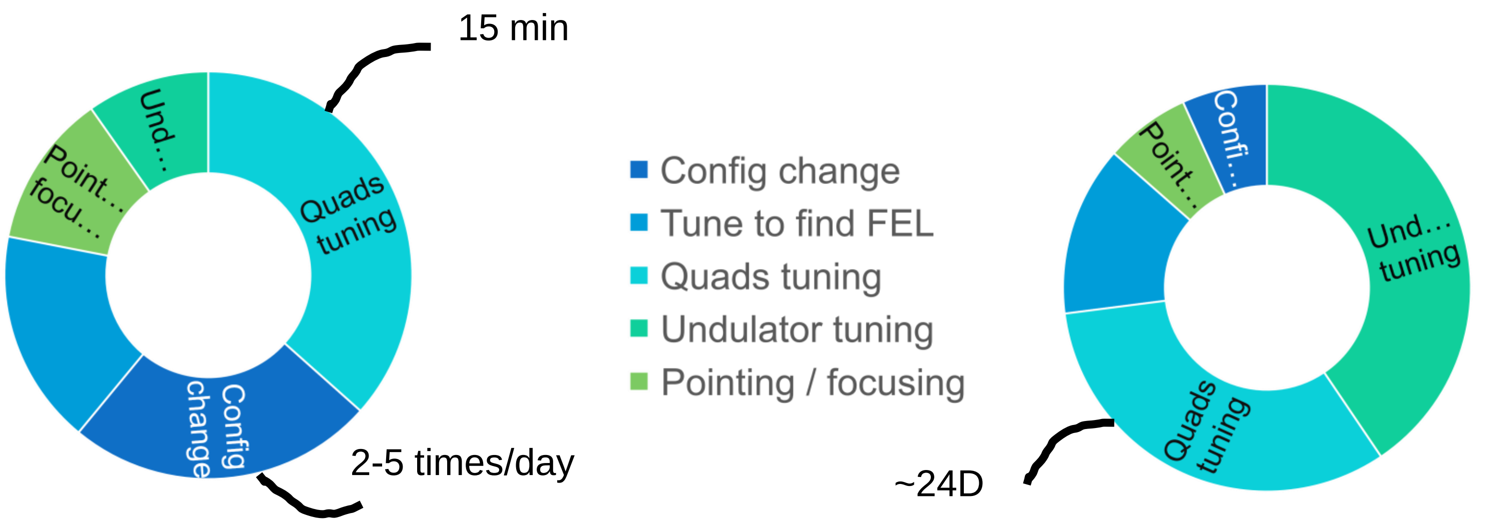


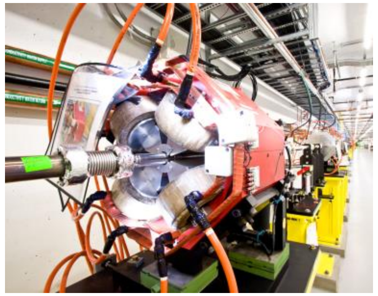


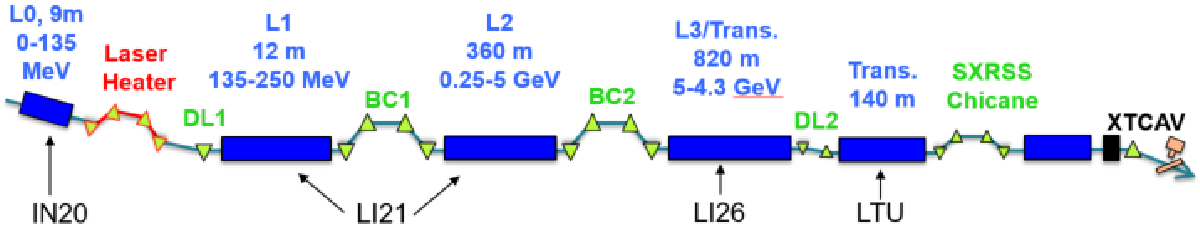
Optimize design parameters

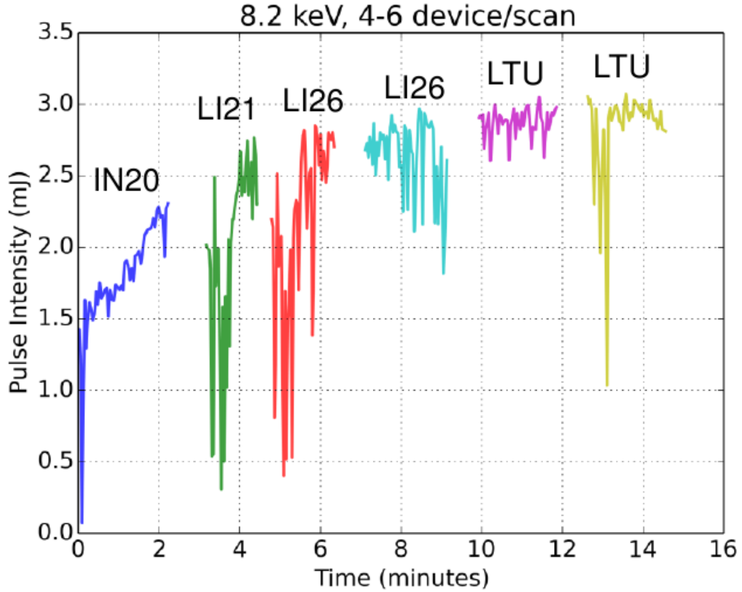
Online optimization of quadrupole magnets @LCLS, SLAC

Setup time [min] **Search space [n-D]**









Quadrupoles provide focusing
→ maintain small beam size
→ Higher X-ray pulse energy!

Outline

- Motivation
- **Model-based vs model-free optimizers**
- Bayesian Optimization (BO)
  - Overview
  - Acquisition functions
  - Accounting for constraints
  - Proximal optimization
- Applications
- Summary of optimization methods

Comparison of Optimizers

Model-based optimization

Deterministic
- Fit model to data

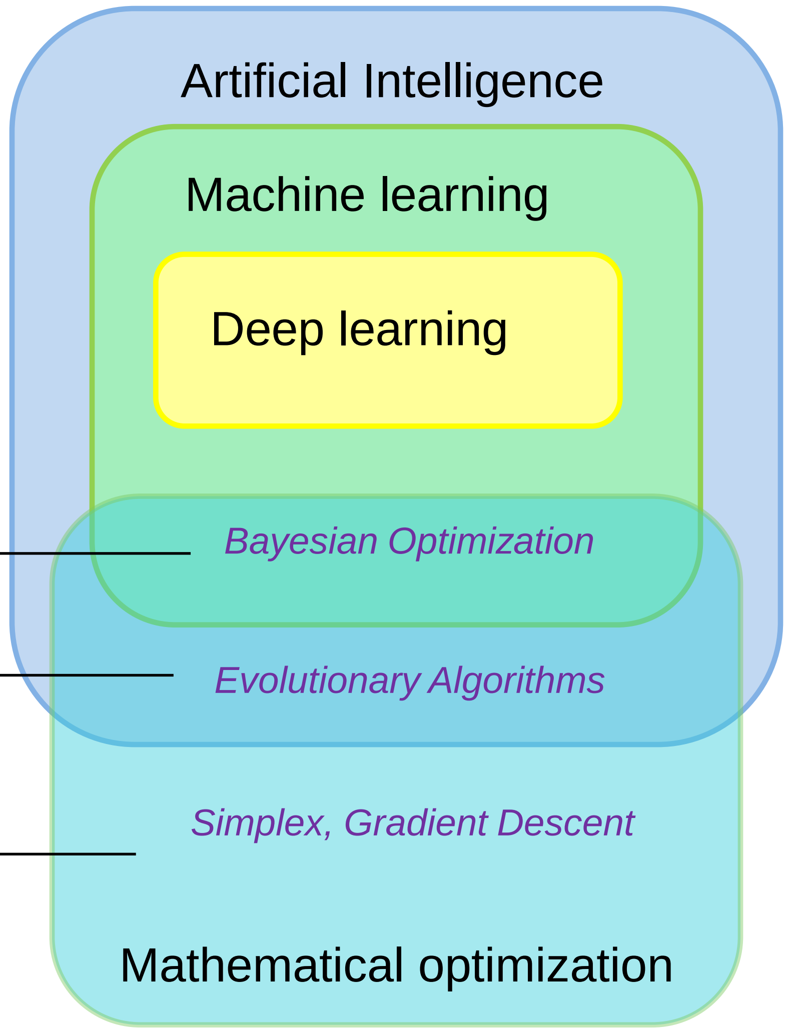
Bayesian
- Calculate probability over functions given data

Model-free optimization

Global
- Simulated annealing
- Particle swarm
- Genetic algorithms

Local
- Gradient descent
- Nelder-Mead simplex
- Extremum seeking
- RCDS

AI ⊃ ML ⊃ DL



Human optimization

Why Bayesian optimization?

Human optimization ≠ Numerical optimization

Human optimization

- Life-long learning
- Experience
- Mental modes
- (relatively) Slow decisions
- Limited working memory

Numerical optimization

- Bulk learning
- Cannot estimate uncertainty
- Juggle many things at once
- Fast decisions

Model-based Bayesian optimization
combines the complementary strengths of both approaches

"A good regulator of the system is a good model of that system."

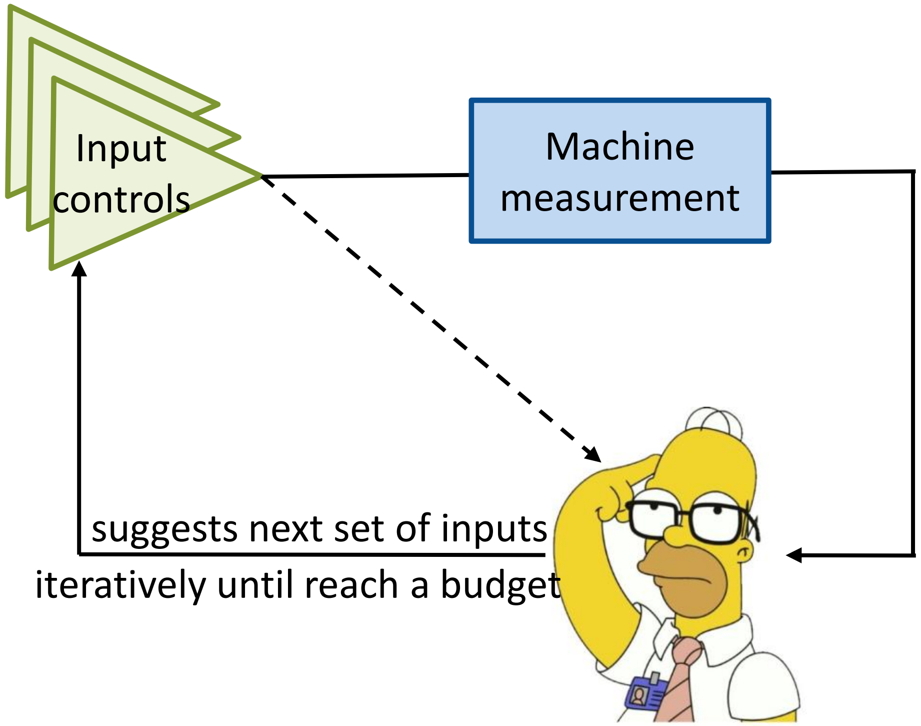
ROGER C. CONANT & W. ROSS ASHBY (1970) Science, 1:2, 89-97

Outline

- Motivation
- Model-based vs model-free optimizers
- **Bayesian Optimization (BO)**
  - Overview
  - Acquisition functions
  - Accounting for constraints
  - Proximal optimization
- Applications
- Summary of optimization methods

Bayesian Optimization: Overview

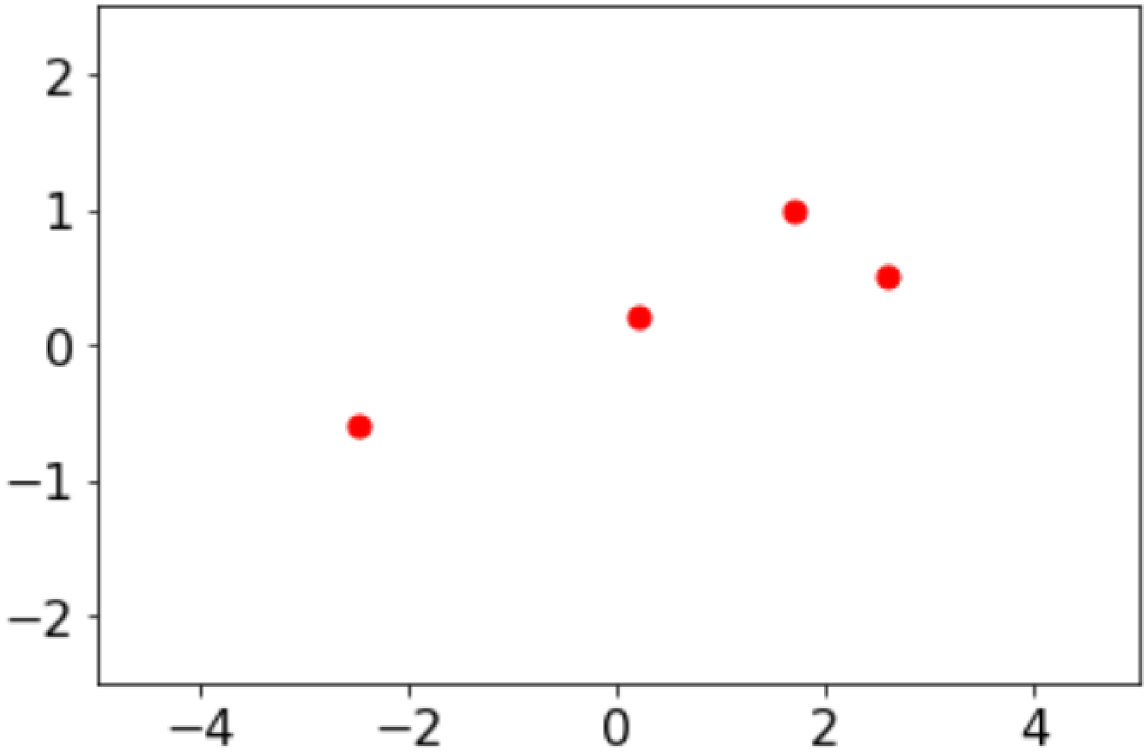
- Gradient-free
- Learns by experience



Acquisition function (utility function) that tells us where to query the system next.

Bayesian Optimization: Overview

Let's get some intuition… Where is the maximum of f ?



Question: Where should we take the next evaluation?

Probabilistic surrogate model for the values our function takes on unseen points.

Bayesian optimization: Overview

quad
X-ray energy

unknown objective function $f(x)$; $[x_0, f(x_0)]$

```
for each step t = 1, 2, 3, . . . ,T:
    1. Build probabilistic model
```

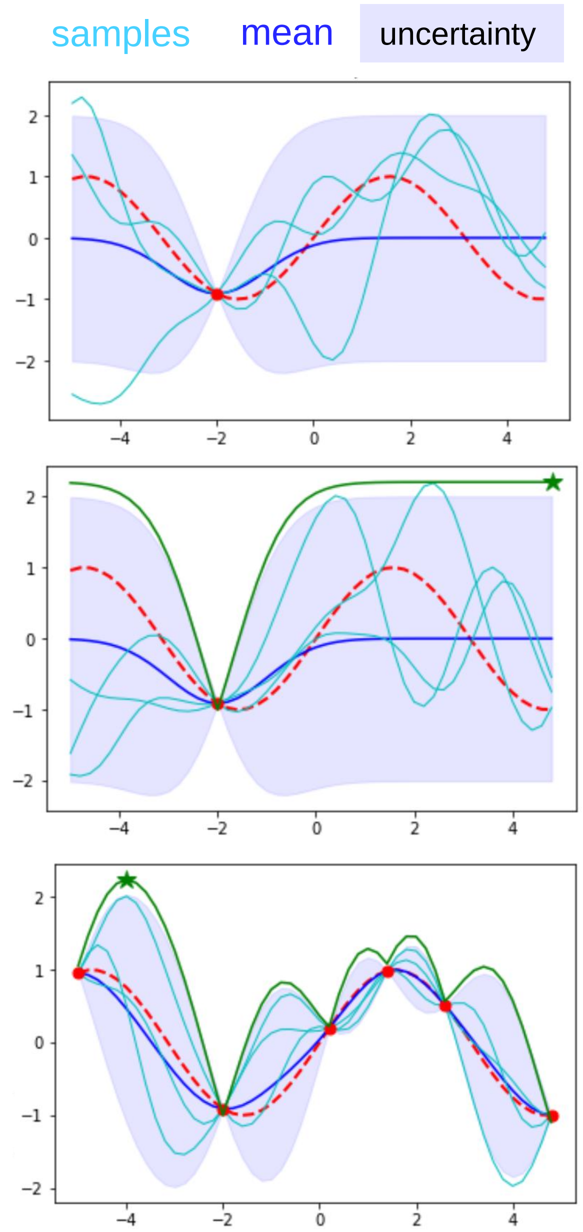
→ $\hat{f}_{t-1}(x)$ Gaussian process

```
    2. Choose next point to simultaneously increase
       objective & decrease model uncertainty
```

→ x_t = argmax$\left(\mathrm{UCB}(x|\hat{f}_{t-1})\right)$ $\mathrm{UCB}(x) = \mathrm{E}[\hat{f}(x)] + \beta\hat{\sigma}(x)$

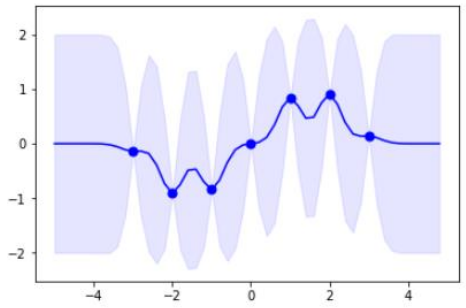
```
    3. Sample new (noisy) point
```

→ $f(x_t)$



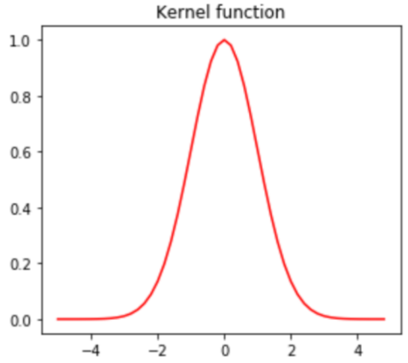
Surrogate model: Gaussian process

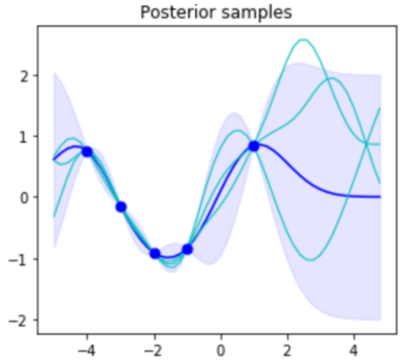
- Give a reliable estimate of their own uncertainty



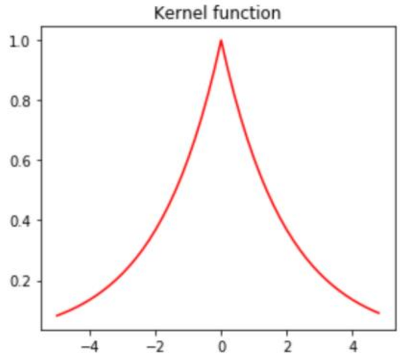
- Shape our prior belief via the choice of kernel $k(x, x')$

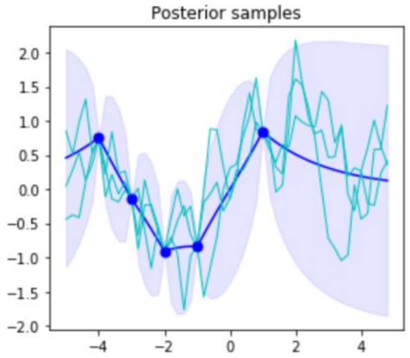
$$k_{\mathrm{RBF}}(x, x') = \exp\left(-\frac{(x - x')^2}{l^2}\right)$$



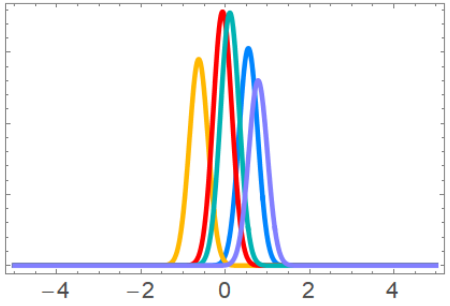


$$k_{\mathrm{exponential}}(x, x') = \exp\left(-\frac{\|x - x'\|}{l^2}\right)$$





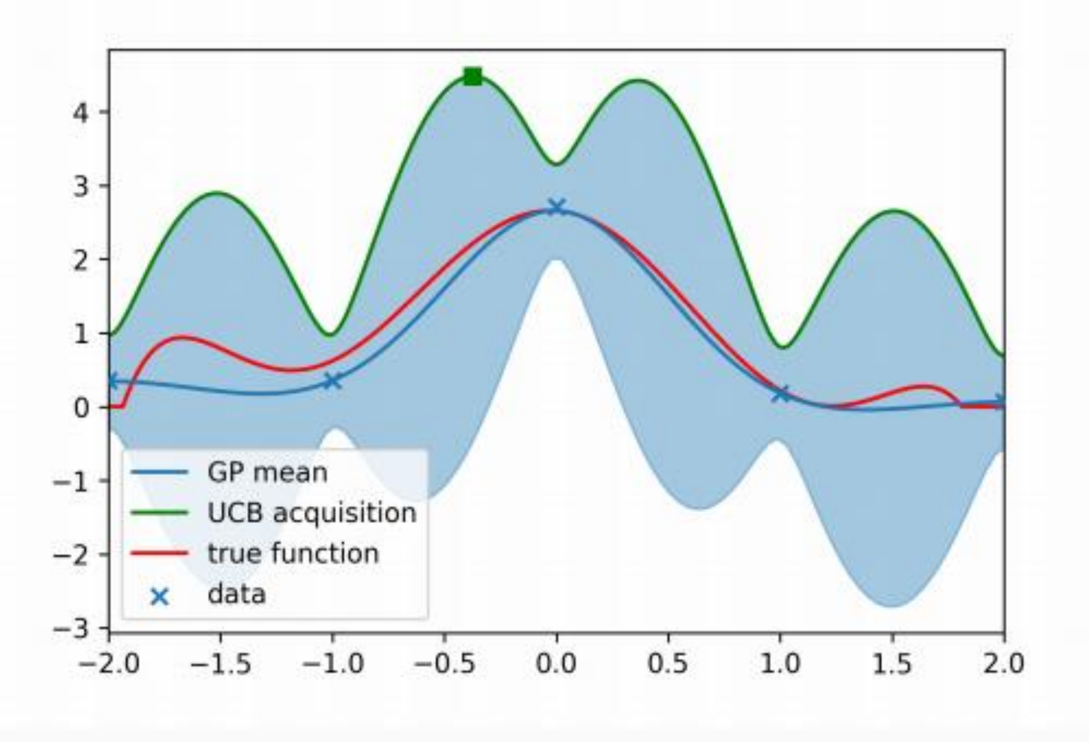
- Latent variables changing day to day
  - → **optimum moves**
  - → Kernel captures **shape**



Acquisition functions: Upper Confidence Bound (UCB)

$$\mathrm{UCB_t}(x) = \mu_t + \beta_t \sigma_t(x)$$

- μ_t - posterior mean after seeing t points.
- σ_t - posterior standard deviation after seeing t points.

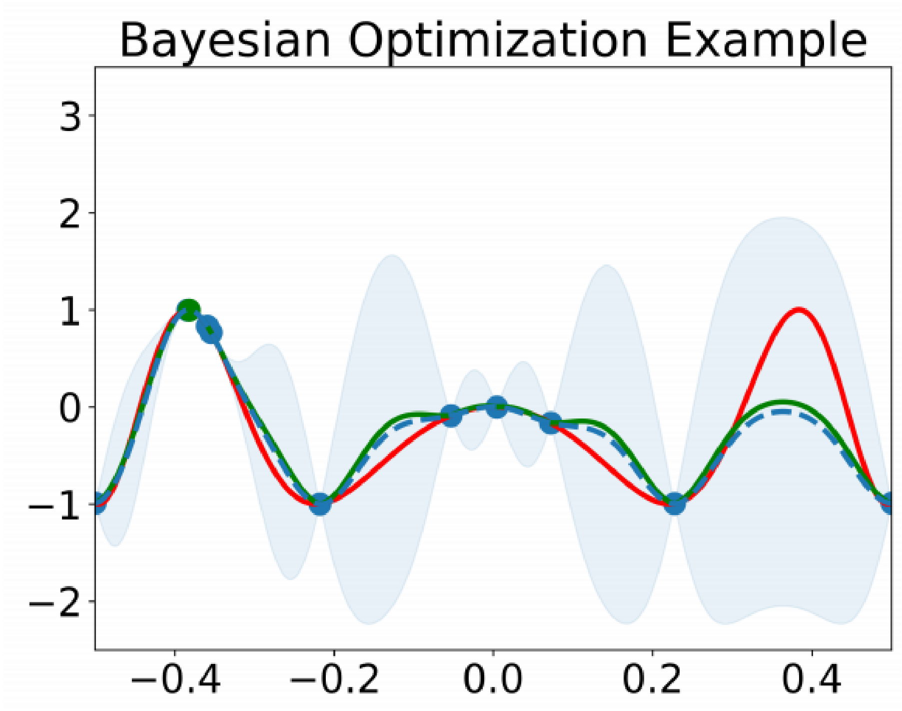


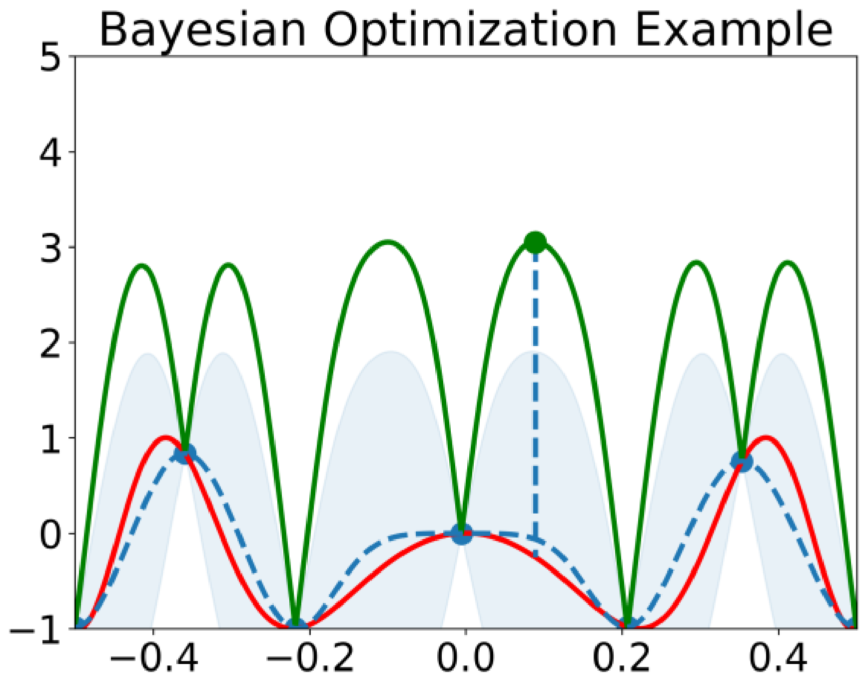
What is β_t?

- trades exploration and exploitation.
  - too small ⇒ gets stuck/hill climbing.
  - too high ⇒ incremental grid search.
- Common heuristic approach: $\beta \approx 2$.
- β_t may increase with time to trade exploration as the optimization progresses.

Acquisition functions: Upper Confidence Bound (UCB)

Question: Which of the examples below is a better optimization process?

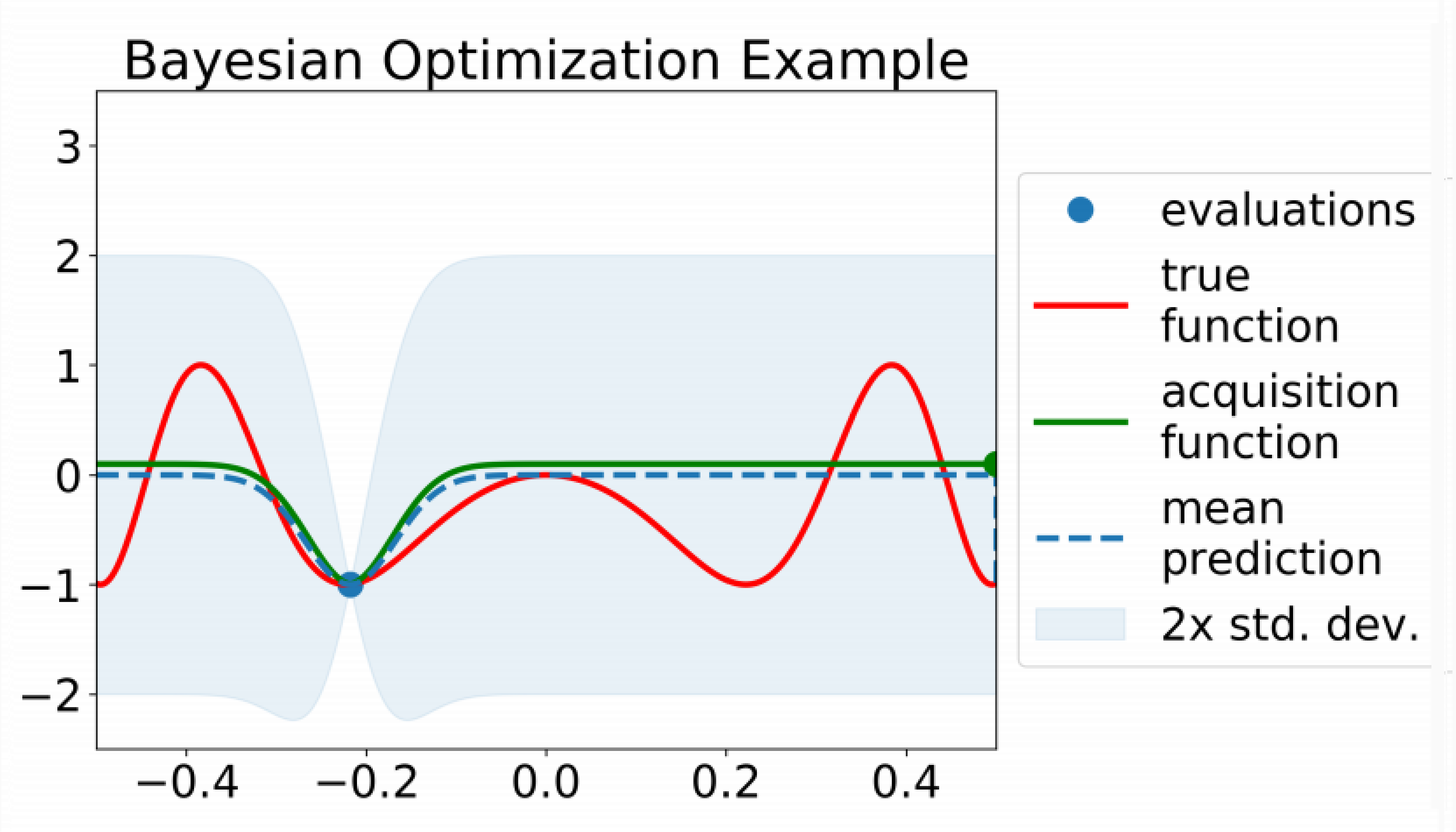




*Adapted from the 2nd ICFA workshop

Acquisition functions: Upper Confidence Bound (UCB)

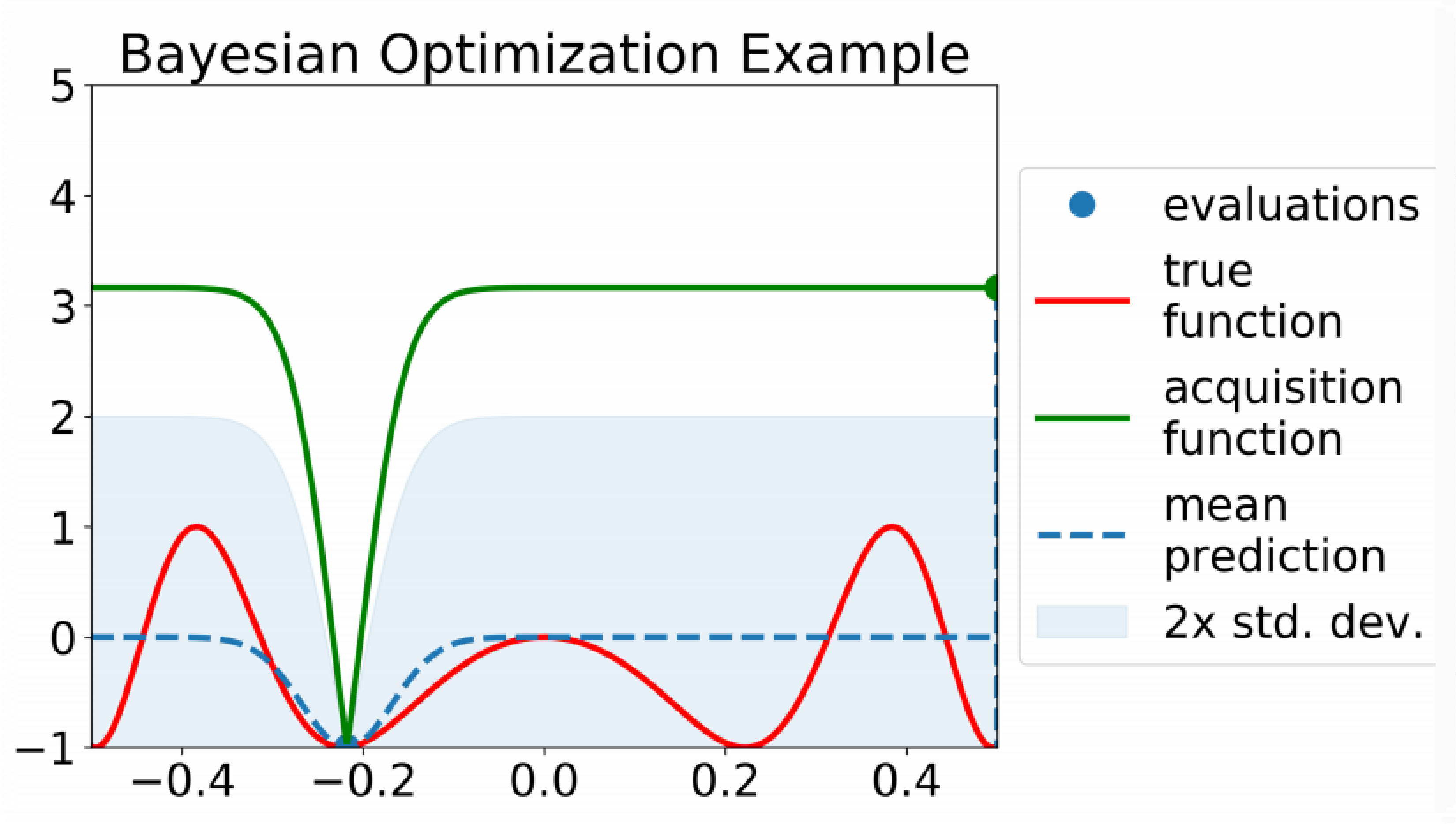
β small - hill climbing



*Adapted from the 2nd ICFA workshop

Acquisition functions: Upper Confidence Bound (UCB)

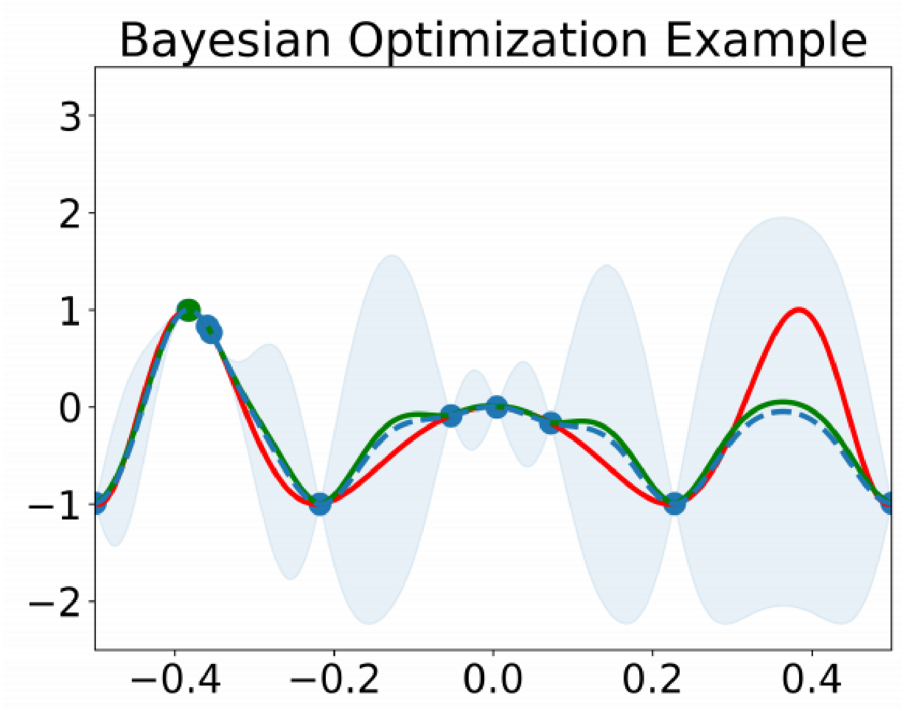
β high - incremental grid search



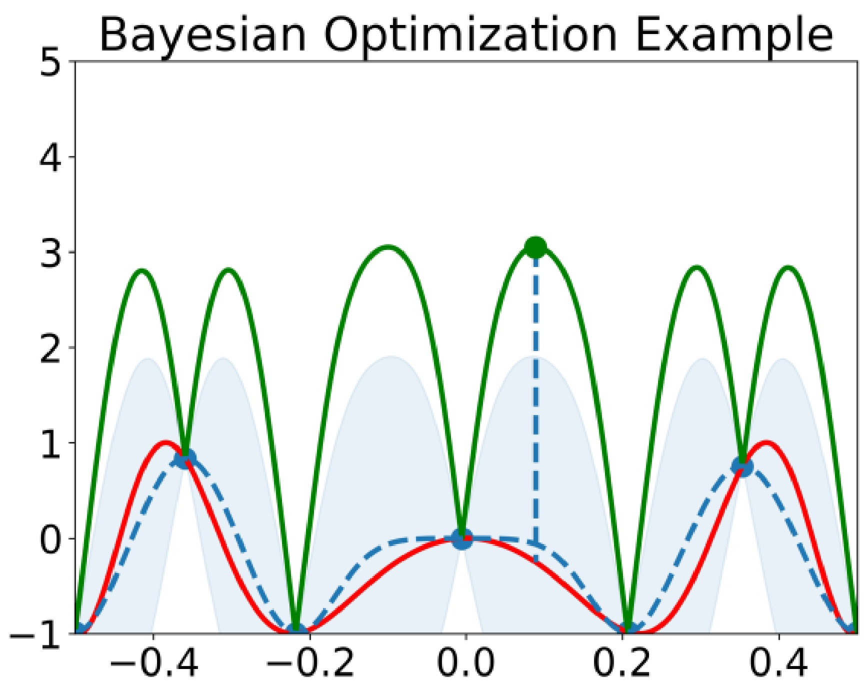
*Adapted from the 2nd ICFA workshop

Acquisition functions: Upper Confidence Bound (UCB)

β small - hill climbing



β high - incremental grid search



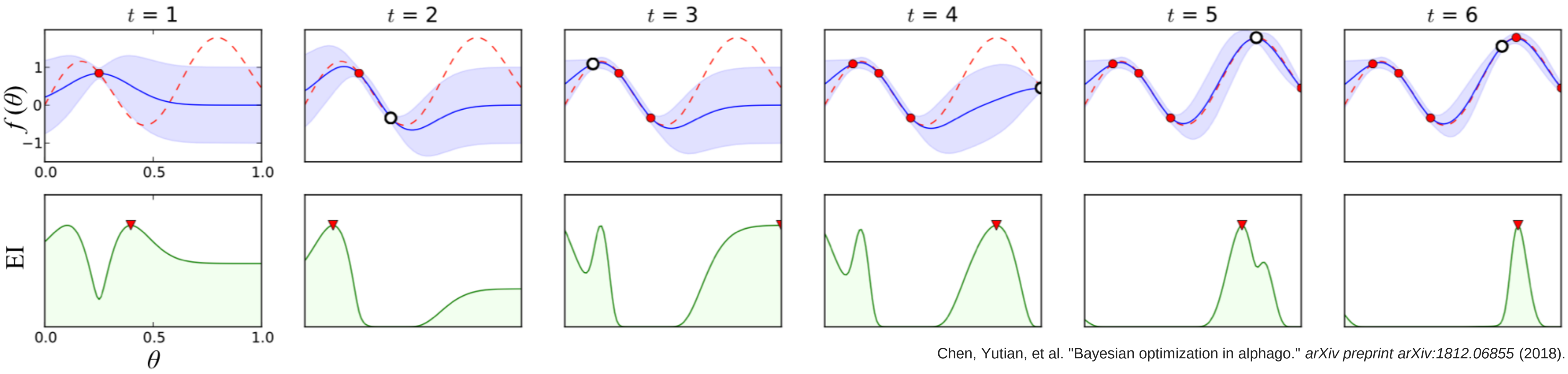
*Adapted from the 2nd ICFA workshop

Acquisition functions: Expected Improvement (EI)

$$\mathrm{EI}_t(\boldsymbol{x}) = \mathbf{E}[\max(0, \mathrm{f}(\boldsymbol{x}) - \mathrm{f}(\boldsymbol{x}^+))]$$

- Analytical solution: $(\mu_t(x) - \mu(x^+))\Phi(Z) + \sigma(x)\varphi(Z)$

  where $Z = \frac{\mu_t - \mu(x^+)}{\sigma_t(x)}$ and Φ, φ are cdf and pdf of standard normal.

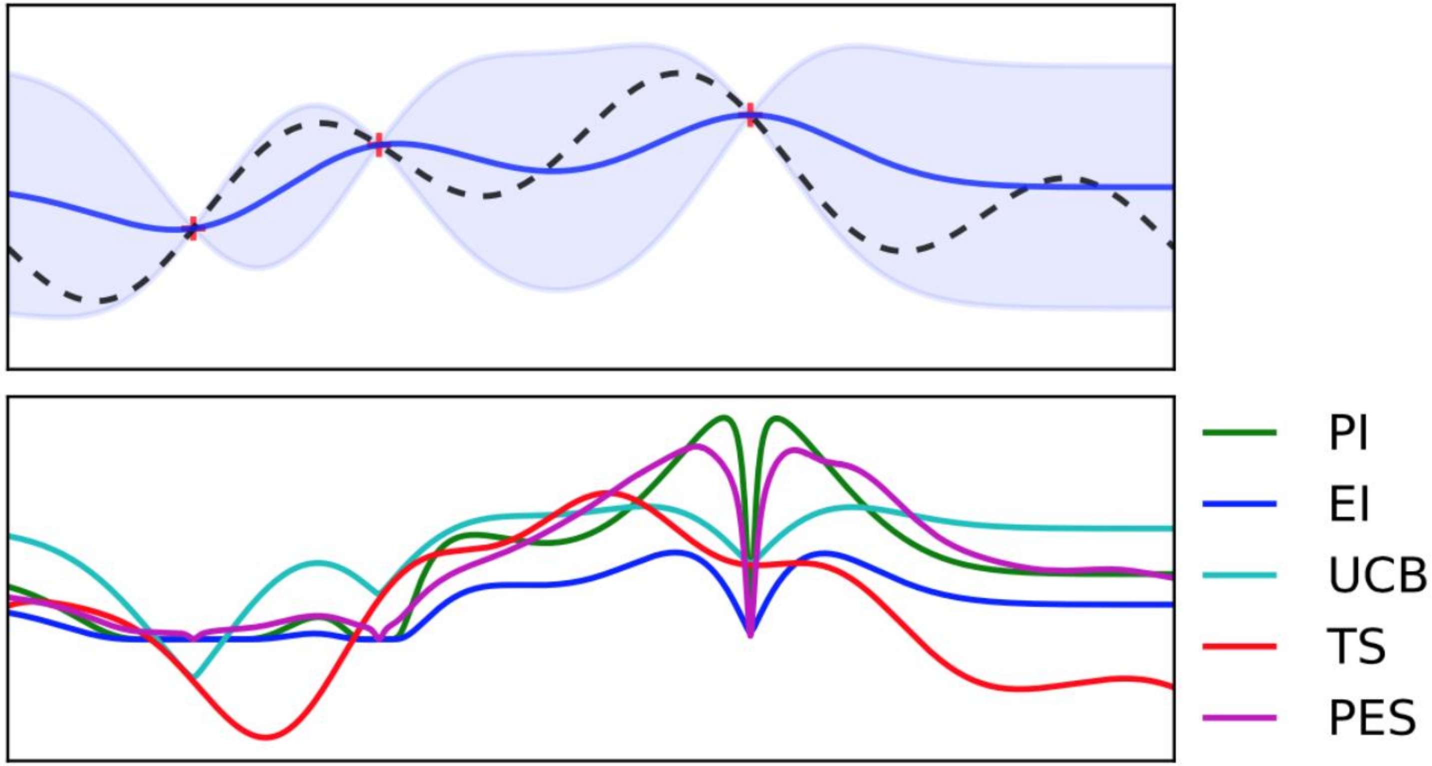


Chen, Yutian, et al. "Bayesian optimization in alphago." *arXiv preprint arXiv:1812.06855* (2018).

Acquisition functions: Other

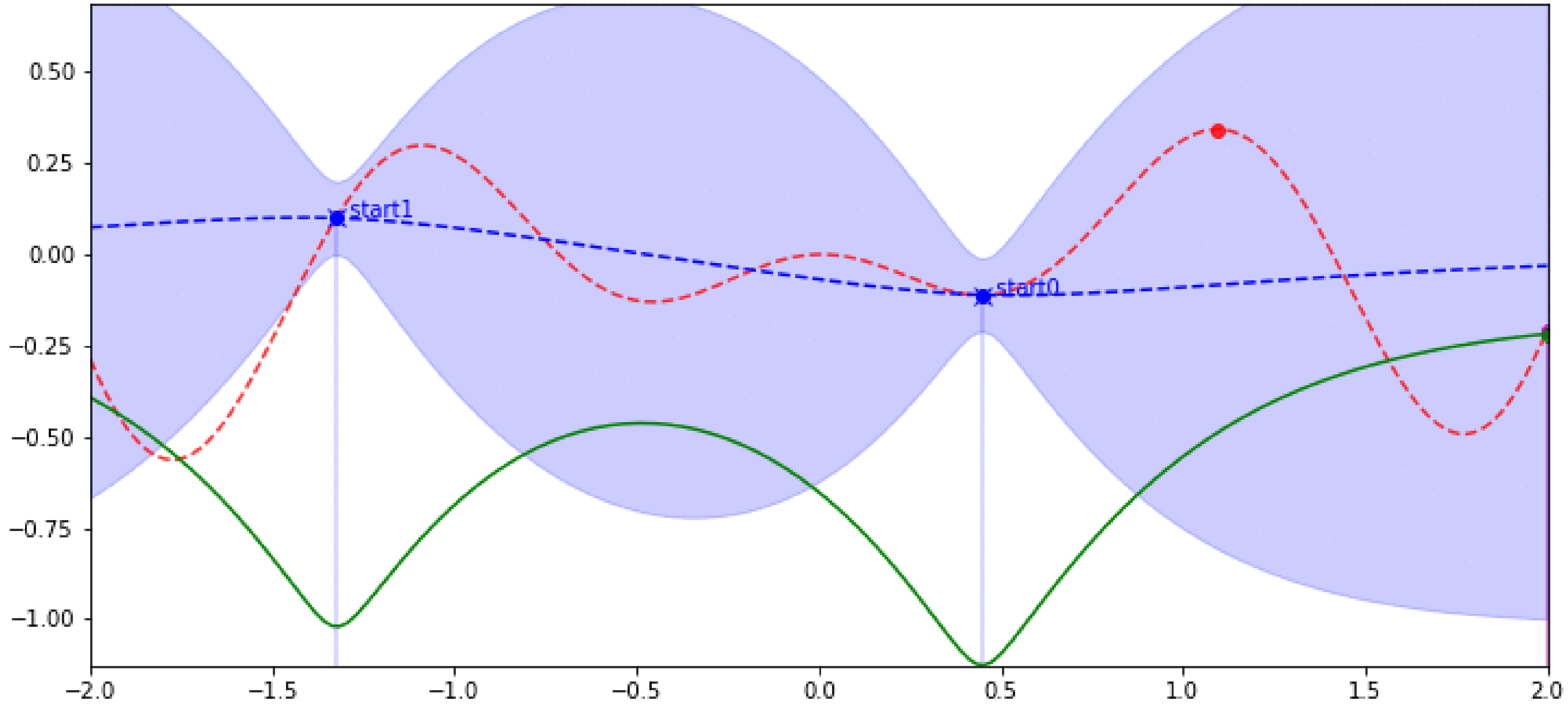
Other types of acquisition functions, each results in a different optimization process.

- PI: Probability of improvement
- TS: Thompson sampling
- PES: Predictive entropy search



https://towardsdatascience.com/shallow-understanding-on-bayesian-optimization-324b6c1f7083

Exploration vs Exploitation



Unknown objective
$f(x)$

Acquisition function
$\mathrm{UCB}(x) = \mathrm{E}[\hat{f}(x)] + \beta\hat{\sigma}(x)$

– $\hat{f}(x)$ **esitmate**
$\hat{f}(x)$ **uncertainty**
X **Evaluation points**

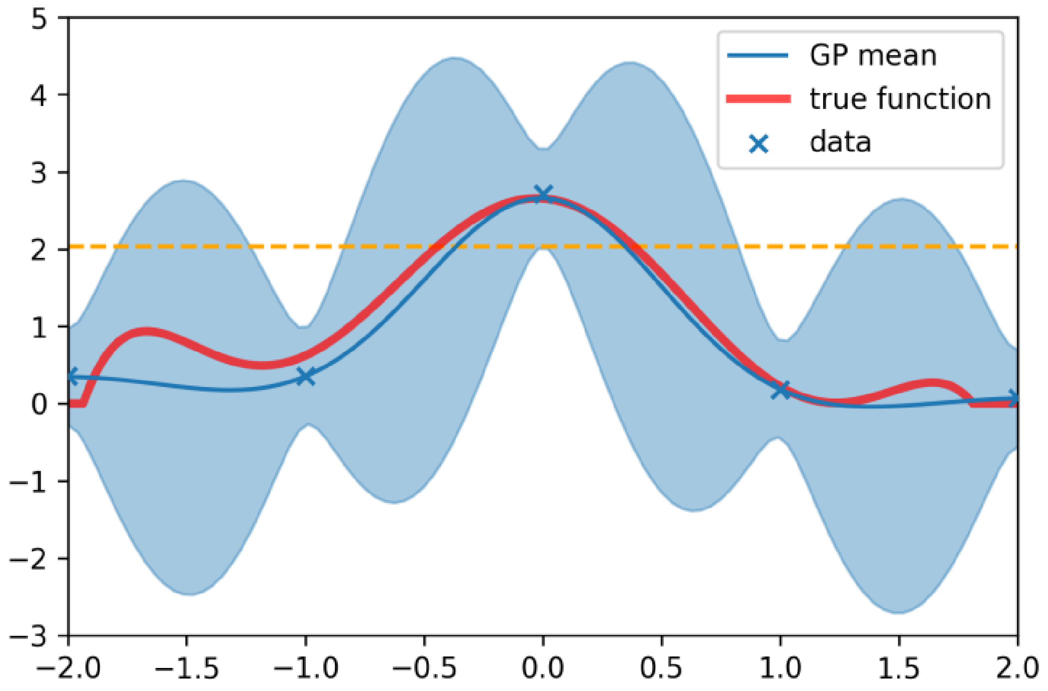
Video credit: cnc-selfbuild blogspot

Outline

- Motivation
- Model-based vs model-free optimizers
- **Bayesian Optimization (BO)**
  - Overview
  - Acquisition functions
  - **Accounting for constraints**
  - Proximal optimization
- Applications
- Summary of optimization methods

Accounting for constraints

Constrain the acquisition function search:
- Avoid unnecessary evaluations.
- Safe BO – not to harm the system.



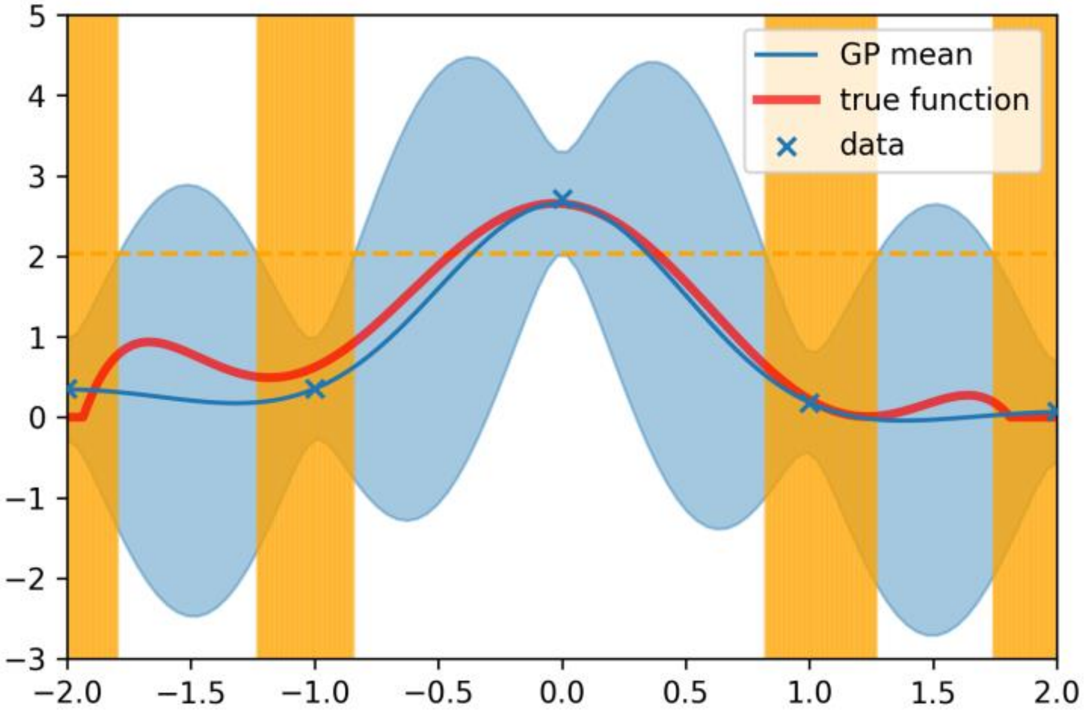
*Adapted from the 2nd ICFA workshop

Accounting for constraints

Constrain the acquisition function search:

- Avoid unnecessary evaluations.
- Safe BO – not to harm the system.

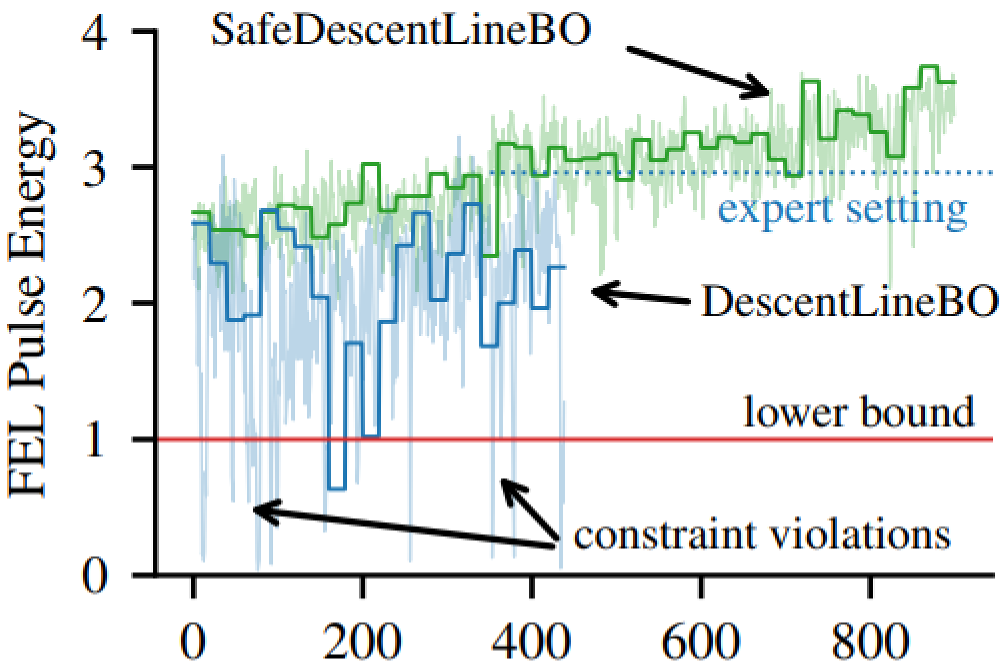
Orange regions to be avoided

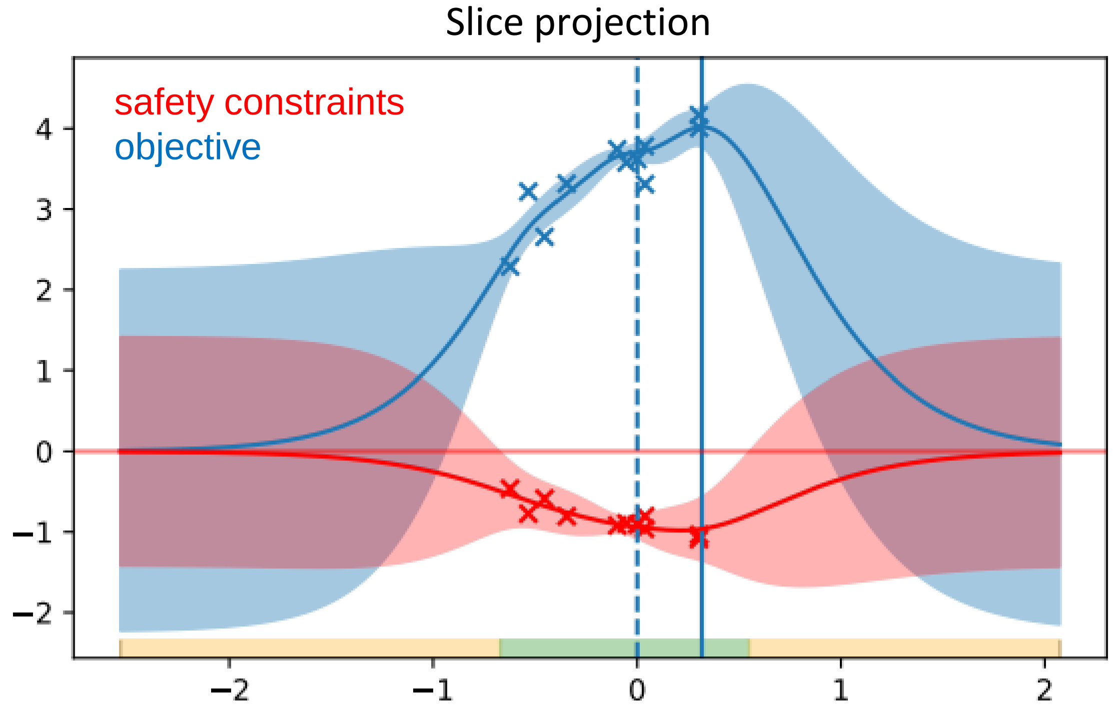


*Adapted from the 2nd ICFA workshop

Accounting for constraints - Example

Maximize FEL energy at SwissFEL using 24 parameters with constrains (lower bound on intensity).





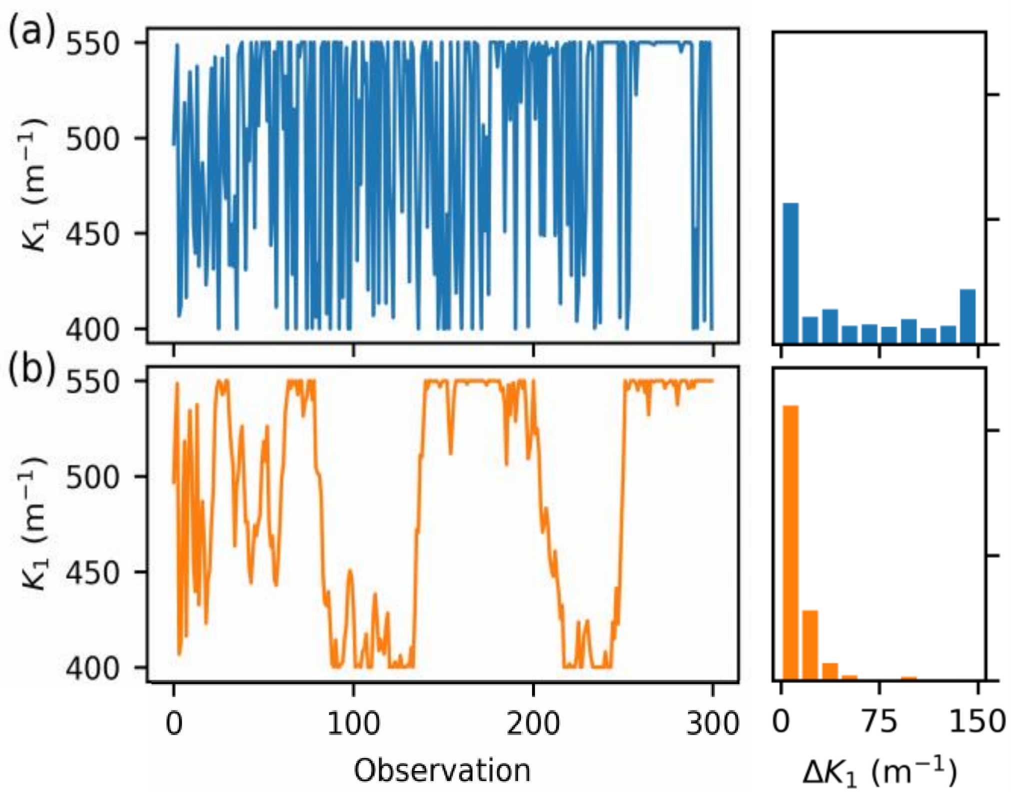
Kirschner, arxiv: 1902.03229

Proximal Optimization

Problem: making large changes in machine input parameters (magnetic field strengths, cavity phases) frequently is undesirable or infeasible.

Solution: Prioritize points in input space that are near the current or most recently observed parameter setting.

Done by penalizing the acquisition function (i.e. by multiplying a multivariate Gaussian distribution).



Curse of Dimensionality

Surrogate model:

GP regression - $O(n^3)$

- Speed: Sparse GP.
- Accuracy: correlated kernels, non-zero prior.

Acquisition function optimization:

Also called "BO's inner optimization problem"; wealth of diverse methods were proposed.

- Speed: local optimization (LineBO), parallelism, constrains
- Safety: constrains.

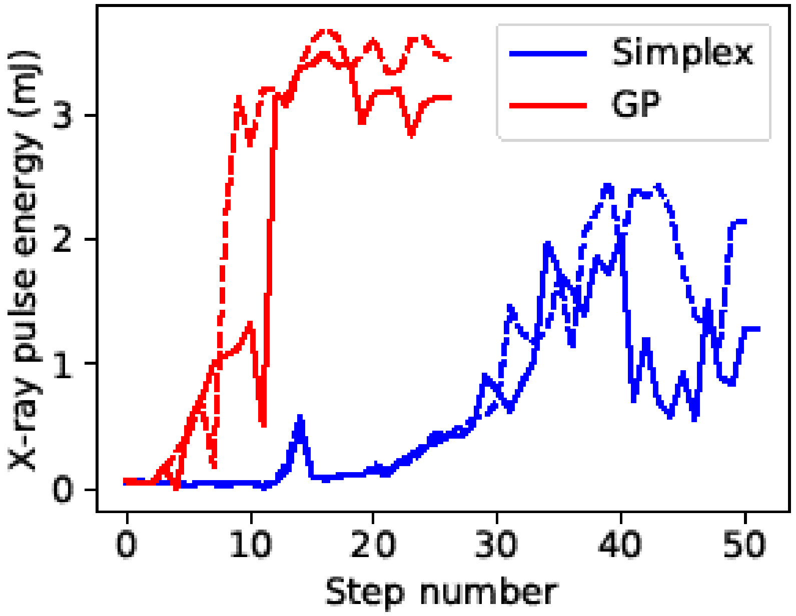
Outline

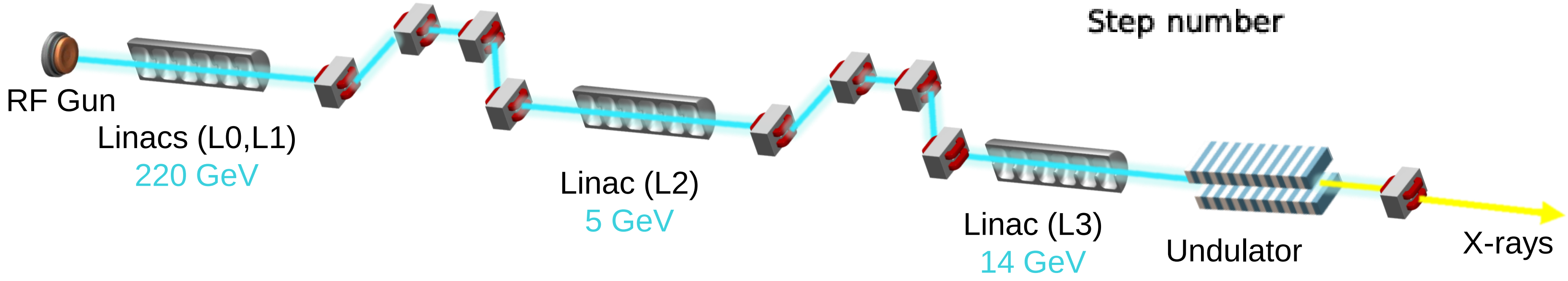
- Motivation
- Model-based vs model-free optimizers
- Bayesian Optimization (BO)
    - Overview
    - Acquisition functions
    - Accounting for constraints
    - Proximal optimization
- **Applications**
- Summary of optimization methods

Application: Online FEL maximization

Maximize X-ray pulse energy simultaneously on 12 quadrupoles with diagonal kernel.

- GP reaches higher optimum
- GP is 4 times faster



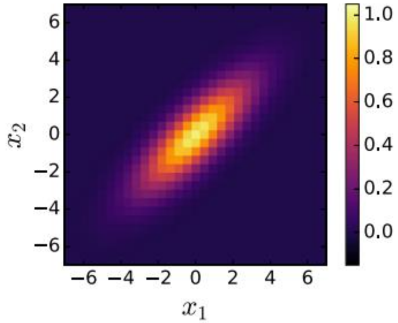


J. Duris, PRL, 2020

Side note: Faster BO with Correlated Kernels

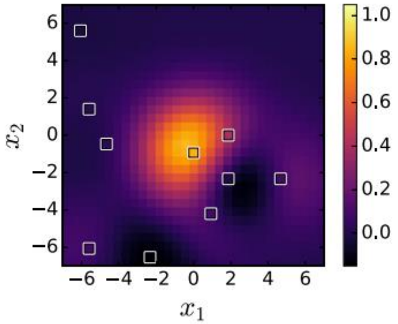
Learn correlations based on physics/ historical data.

$$k_{\mathrm{RBF}}(\boldsymbol{x},\boldsymbol{x}') = \sigma_f^2 \exp(-(\boldsymbol{x}-\boldsymbol{x}')^T \boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{x}'))$$



(a) Ground truth

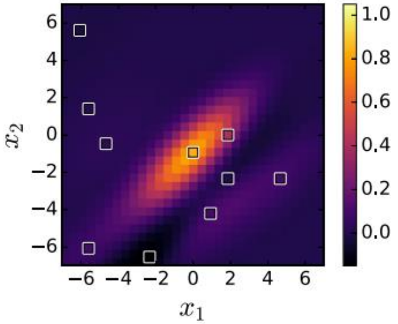
$$\Sigma = \begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix}$$



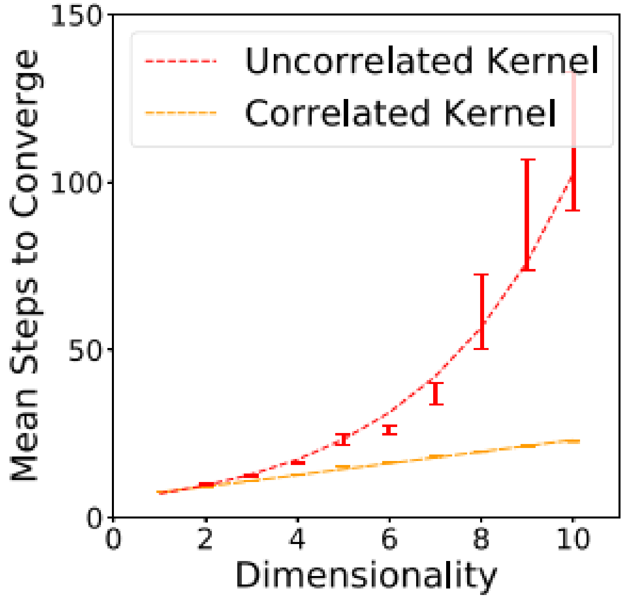
(b) Isotropic kernel

$$\Sigma = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}$$

$$\Sigma = -\frac{H_{ij}}{2}$$



(c) Correlated kernel



(d) Convergence tests

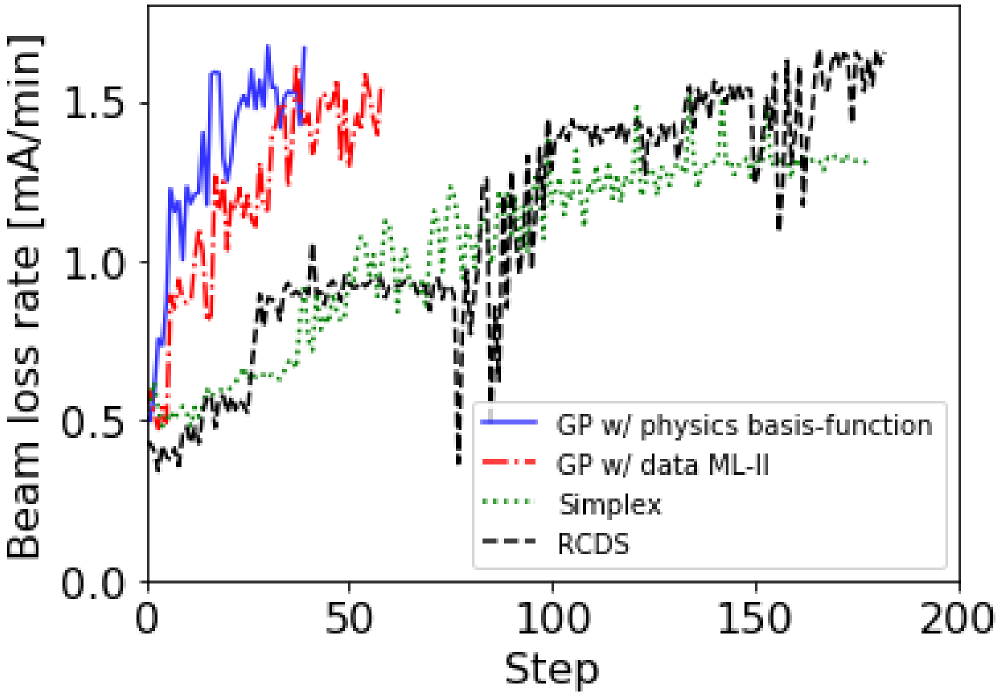
J. Duris, PRL, 2020

A. Hanuka, PRAB, 2021

Application: Online vertical emittance minimization

Minimize vertical emittance (= maximize beam loss rate)
with 13 skew quadrupole magnets

- GP 10x speedup.



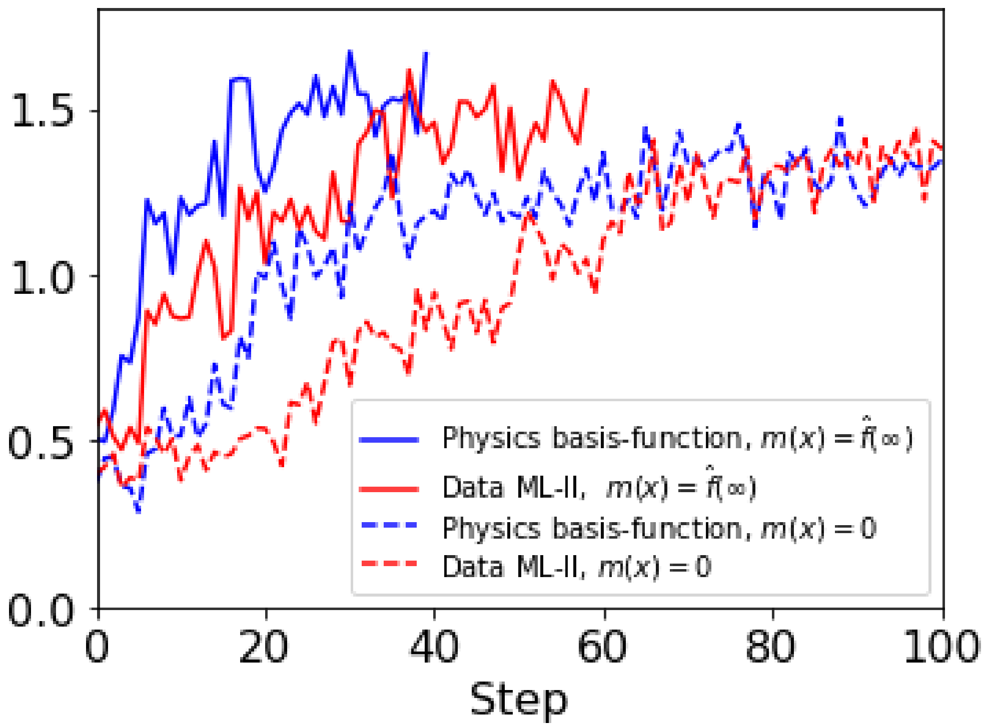
2-3 sec / step
RCDS: 6 sec / step

A. Hanuka, NeurIPS 2019
A. Hanuka, PRAB, 2021

Side: Effect of GP prior mean on the optimization

$y \sim GP(m(x), k(x, x'))$

mean function



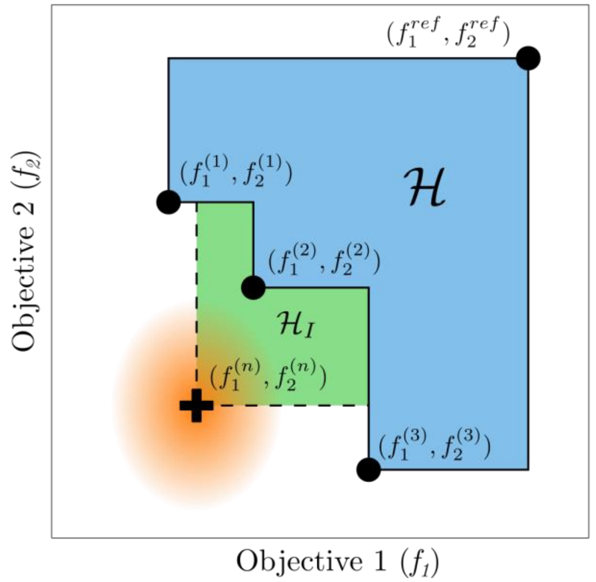
GP with prior mean $m(x) = 0$ (dashed) converged slower to a lower optimum.

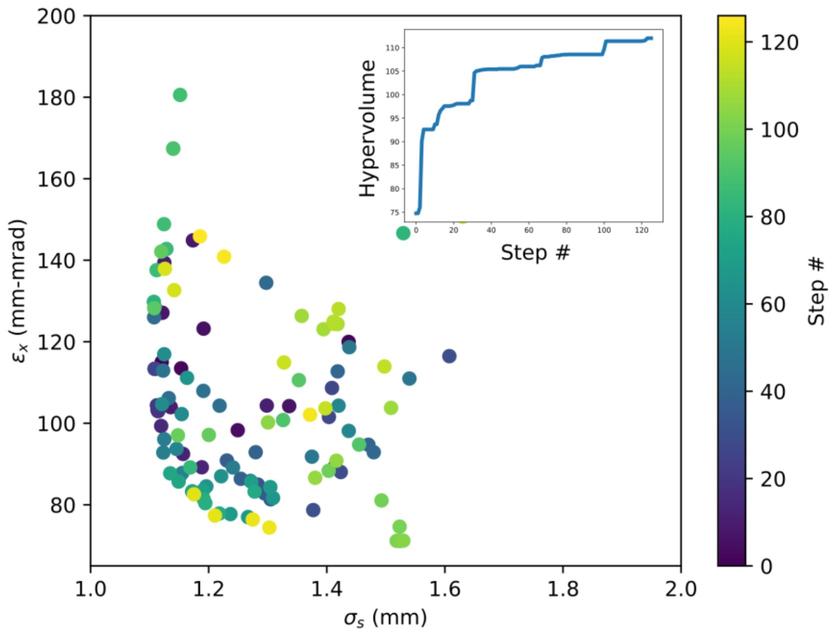
A. Hanuka, NeurIPS 2019
A. Hanuka, PRAB, 2021

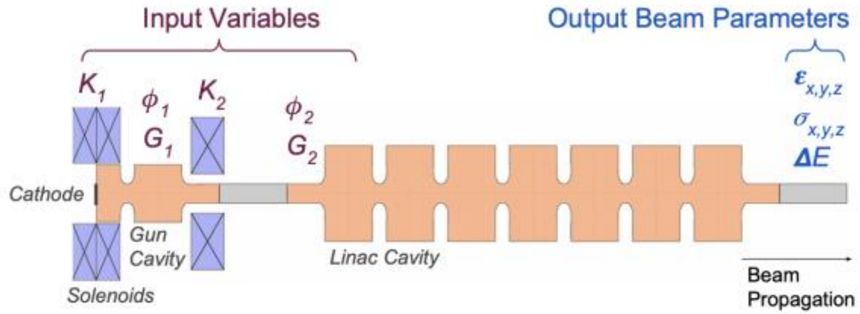
Application: Multi-Objective BO (MOBO)

MOBO - Find the set of Pareto-optimal points in objective space.

Simultaneously minimize transverse emittance & longitudinal bunch length in the AWA photoinjector.







R. Roussel, PRAB 2021

BO Summary

Advantages:

- Noise robust.
- Data efficient (statistical model).
- Global guarantees.
- Can handle safety constraints.

Caveats:

- Computational efficiency : Maximizing the acquisition function, GP regression.
- Curse of dimensionality.
- Practical: Hyperparameters, difficult to evaluate model fit.

Outline

- Motivation
- Model-based vs model-free optimizers
- Bayesian Optimization (BO)
  - Overview
  - Acquisition functions
  - Accounting for constraints
  - Proximal optimization
- Applications
- **Summary of optimization methods**

Summary of optimization methods

Instructions:

- We're going to split to breakout room.
- Each breakout room will fill in the table of comparison for one algorithm (room 1 → algo 1, etc)
- Table of comparison
- Optional answers – Low/Medium/High or Yes/No.
- Choose one presenter to present the table in the main room.

- Sample efficiency
- Computational cost of picking the next point
- Multi-objective
- Sensitivity to local minima
- Sensitivity to noise
- Requires to compute or estimate derivatives of f
- Evaluations of f inherently done in parallel
- Hyper-parameters

1. Nelder-Mead
2. Gradient descent
3. Powell / RCDS
4. L-BFGS
5. Genetic algorithm
6. Bayesian optimization

Summary of optimization methods

	Nelder-Mead	Gradient descent	Powell / RCDS	L-BFGS	Genetic algorithm	Bayesian optimization
Sample efficiency						
Computational cost of picking the next point						
Multi-objective						
Sensitivity to local minima						
Sensitivity to noise						

Let’s review the answers…

Summary of optimization methods

	Nelder-Mead	Gradient descent	Powell / RCDS	L-BFGS	Genetic algorithm	Bayesian optimization
Sample efficiency	Medium	Medium	Medium/high	Medium/high	Low	High
Computational cost of picking the next point	Low/Mediu m	Low	Low	Low	Medium (e.g. sorting)	High (esp. in high dimensions)
Multi-objective	No	No	No	No	Yes	Yes
	(but can use scalarization)					
Sensitivity to local minima	High	High	High	High	Low	Low (builds a **global** model of *f*)
	(but can use multi-start)					
Sensitivity to noise	High	High	High (Powell) Low (RCDS)	High	Medium	Low (can model noise itself)

Summary of optimization methods

	Nelder-Mead	Gradient descent	Powell / RCDS	L-BFGS	Genetic algorithm	Bayesian optimization
Requires to compute or estimate derivatives of f	No	Yes	No	Yes	No	No
Evaluations of f *inherently* done in parallel	No	No	No	No	Yes	No
Hyper-parameters	Initial simplex	Step size: α (+momentum: β)	# fit points Noise level	Accuracy of hessian estimate	• Population size • Mutation rate • Cross-over rate • Number of generations	• Kernel function • Kernel length scales, amplitude • Noise level • Acquisition function

Thank you for your attention!

1 For the weekend!

2 Lectures only! We still have lab afternoon