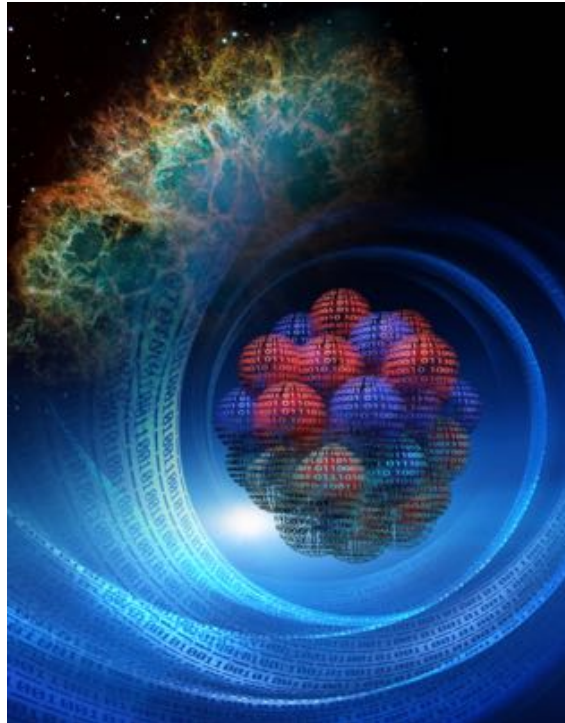


# HOW DO WE HANDLE COMPUTATIONALLY EXPENSIVE OBSERVABLES WHEN FITTING AN INTERACTION ?

ANDREAS EKSTRÖM



# NEW IDEAS IN CONSTRAINING NUCLEAR FORCES



## a precise description of the strong nuclear interaction.

To what extent can nuclei be described in effective field theories of quantum chromo dynamics ?

Several talks this week !

“Fitting Interactions”

How to estimate the uncertainties in theoretical predictions from effective field theory ?

Several talks this week !

**A possible scenario:** we need to optimize a set of LECs in an EFT by minimizing some objective function that includes experimental data for a class of observables that is computationally expensive: e.g. 3N scattering data, and/or  $A > 4$  many-body observable(s), and/or nuclear matter, ...

- Budget: ~50-100 function evaluations ~ (10-20-...) dimensional parameter space.
- Some type of simulation/surrogate computation should be invoked.
- Ab initio methods can run with reduced fidelity.
- Exploiting/extracting derivatives is most likely out of the question.
- In this talk I will focus on a possible method to handle expensive objective functions.

# AN OPTIMIZATION PROBLEM

*(finding global optimizers is generally intractable)*

A typical objective function is a non-linear least squares function, or some other measure of the goodness of fit.

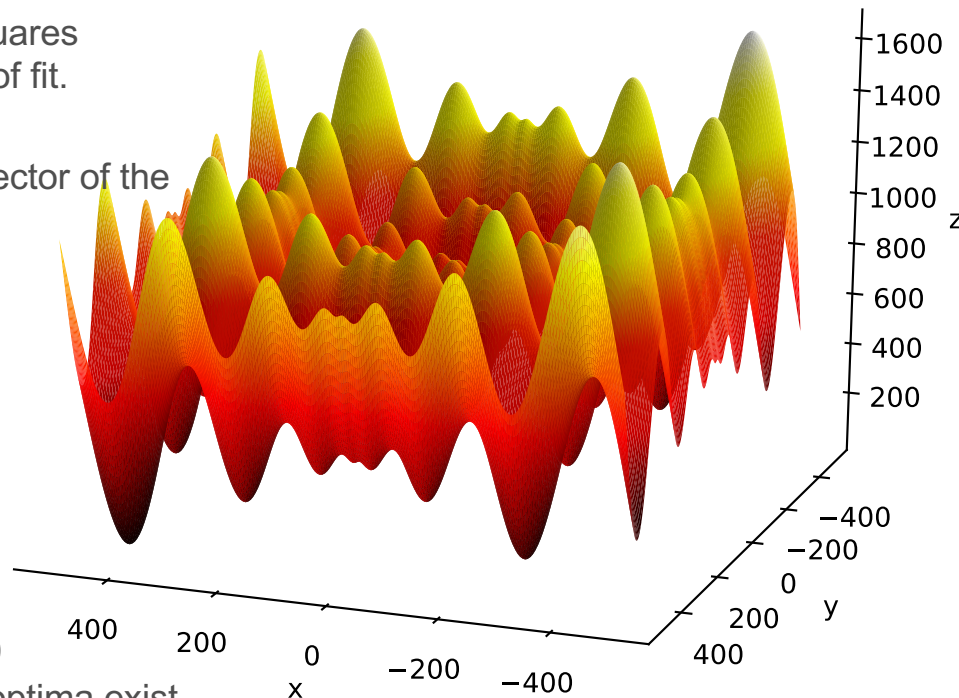
We are after the point estimate of the parameter vector of the nuclear interaction model.

Goal: find 'optimal' parameter vector(s)

$$\theta_{\star} = \arg \min_{\theta \in \mathcal{X} \subset \mathbb{R}^d} F(\theta)$$

*Additional challenges in a realistic application to nuclear forces:*

- high-dimensional parameter domain,  $10 \lesssim d \lesssim 30$
- often several local optima, cases with  $\sim 100$  local optima exist






# BAYESIAN INFERENCE AND PARAMETER ESTIMATION

Bayes formulation of statistics offers a convenient method to guard against overfitting, to incorporating prior knowledge (naturalness, the EFT error scaling), and to quantitatively compare models to each other!

$$P(\mathbf{a}|D, I) = \frac{P(D|\mathbf{a}, I) P(\mathbf{a}|I)}{P(D|I)}$$

Likelihood      Prior  
Posterior

Evidence (normalization)



**Toolbox:**

**Marginalization**

$$P(a_1|D, I) = \int da_2 \dots da_k P(\mathbf{a}|D, I)$$

$$P(D|\mathbf{a}, I) = \int dc_{\bar{\nu}+1} \dots dc_{\nu_{\max}} P(D|c_{\bar{\nu}+1} \dots c_{\nu_{\max}}, \mathbf{a}, I) \\ \times P(c_{\bar{\nu}+1} \dots c_{\nu_{\max}}|I)$$

If you can do Markov Chain Monte Carlo sampling then you can afford to do Bayes !

There exists “discounted” versions where you approximate your Likelihood and therefore obtain an approximate description of your posterior.

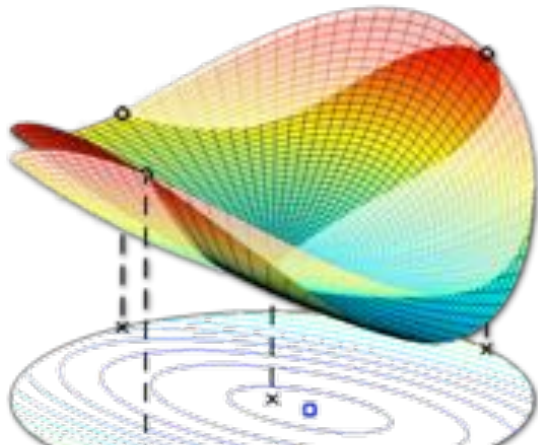
**Model comparison (Bayes factors)**

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{\int P(D|\mathbf{a}, M_1) P(\mathbf{a}|M_1) d\mathbf{a}}{\int P(D|\mathbf{b}, M_2) P(\mathbf{b}|M_2) d\mathbf{b}}$$

Tomorrow: Christian Forssen, Sarah Wesolowski

For more on Bayes and EFT:  
S. Wesolowski et al, J Phys G 43, 074001 (2016)  
M. Schindler, D. Phillips, Ann. Phys. 324, 682 (2009)

# POUNDERS: MODEL-BASED OPTIMIZATION



POUNDERS exploits the known 'squared-sum' structure of the objective function!

$$\min \left\{ f(x) = \frac{1}{2} \|F(x)\|_2^2 = \frac{1}{2} \sum_{i=1}^p F_i(x)^2 : x \in \mathcal{X} \subset \mathbb{R}^n \right\}$$

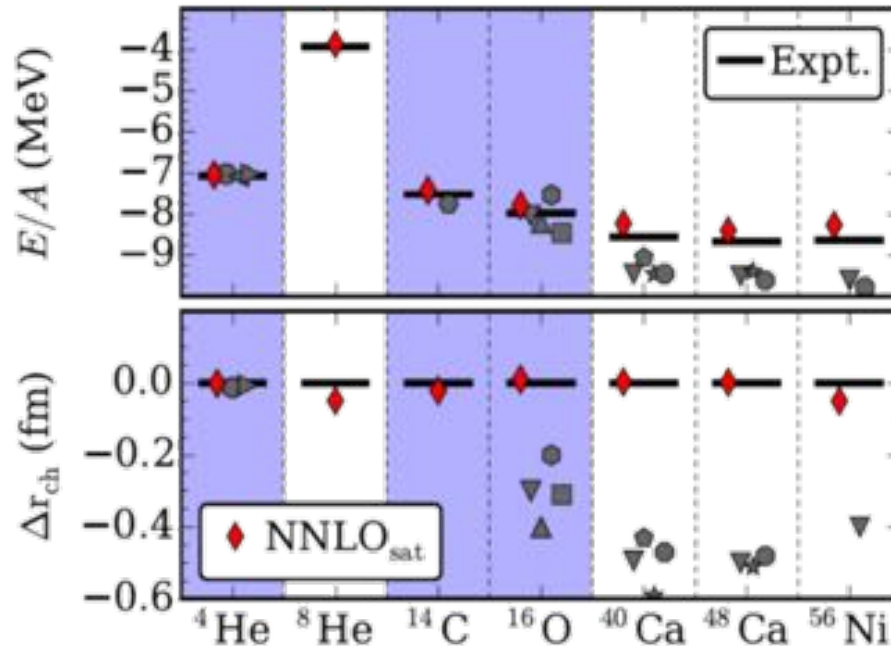
Exploit known structure (sum-of-squares) and setup interpolating *quadratic* model of each residual  $F_i(x)$  centered around  $x^k$ .

$$q_k^{(i)} = F_i(x_k) + (x - x_k)^T g_k^{(i)} + \frac{1}{2} (x - x_k)^T H_k^{(i)} (x - x_k)$$

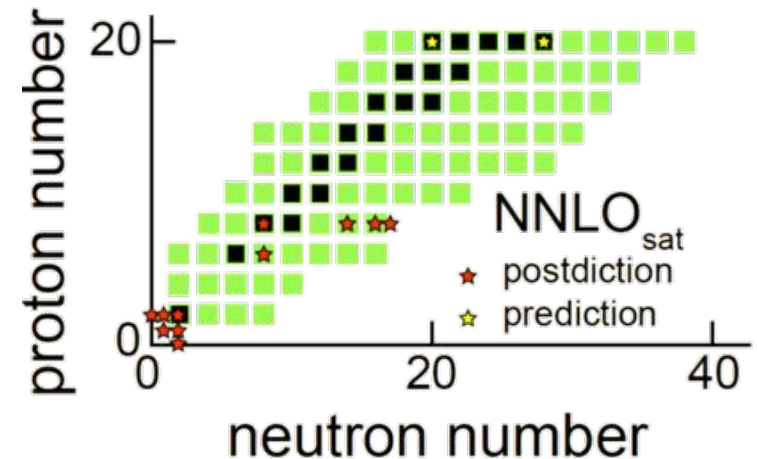
Solve for  $g_k$  and  $H_k$  by interpolating to subset of common function values. A master model uses second order information

$$m_k(x_k + \delta) = f(x_k) + \delta^T \sum_{i=1}^p F_i(x_k) g^{(i)} + \frac{1}{2} \delta^T \sum_{i=1}^p \left( g^{(i)} [g^{(i)}]^T + F_i(x_k) H^{(i)} \right) \delta$$

# NNLO<sub>SAT</sub> – POUNDERS & AB INITIO IN A SMALL SPACE



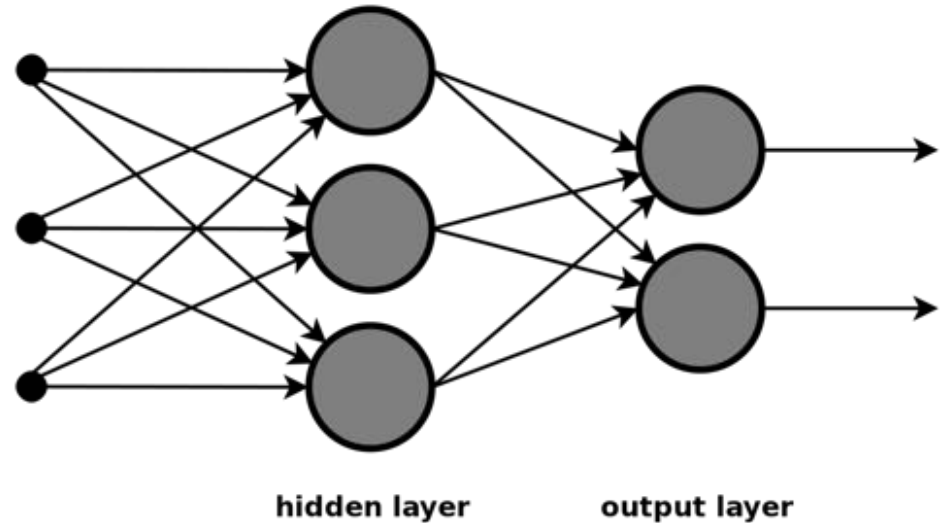
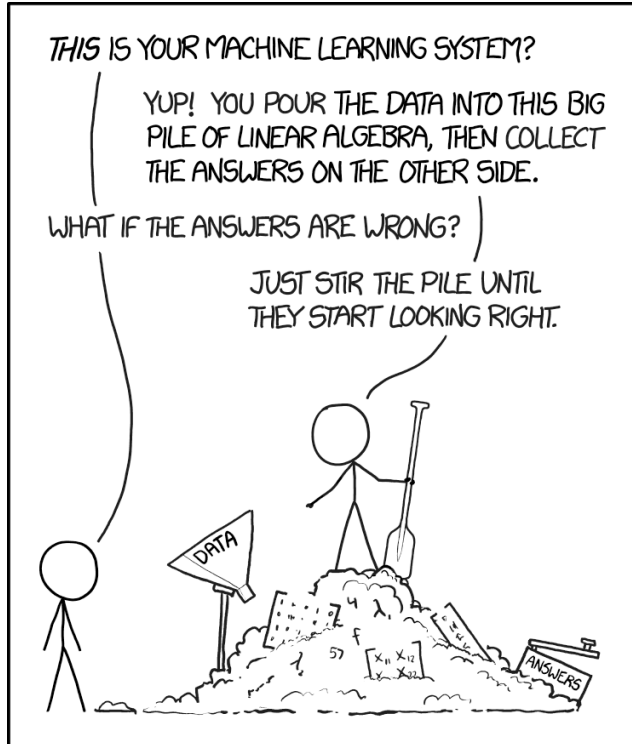
Simultaneous optimization of the chiral NN+NNN interaction at NNLO to charge radii and binding energies of  ${}^3\text{H}$ ,  ${}^3\text{He}$ ,  ${}^{14}\text{C}$ ,  ${}^{16}\text{O}$  and binding energies of  ${}^{22,24,25}\text{O}$  and NN-data ( $T_{\text{Lab}} < 35$  MeV).



# MACHINE LEARNING

*field of study that gives computers the ability to learn without being explicitly programmed.*

- Arthur Samuel, 1959

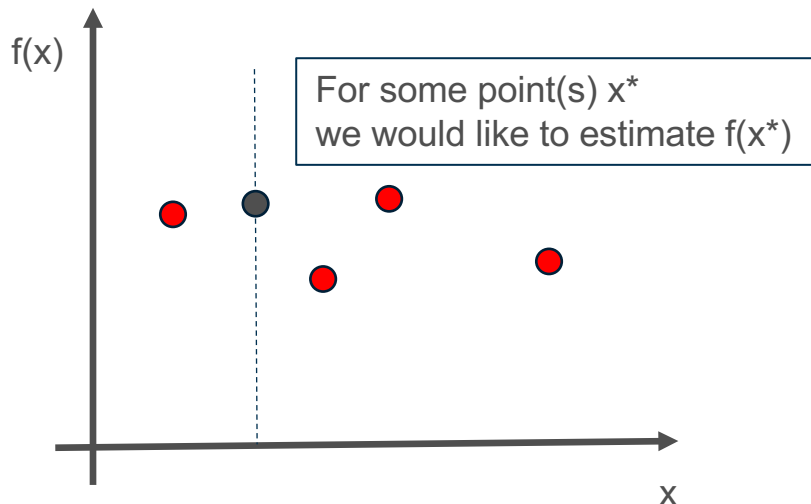


**Typically requires huge amounts of training data. Which we do not have in the scenario that I am discussing here today.**

# MACHINE LEARNING WITH GAUSSIAN PROCESSES

Gaussian process: a collection of random variables, any finite number of which have joint Gaussian distributions.  
(*Rasmussen & Williams, Gaussian Processes for Machine Learning*)

A GP is a distribution of possible functions  $f(x)$  that are consistent with observed data. In a Bayesian view, it begins with a prior, and updates a posterior as new data comes in.

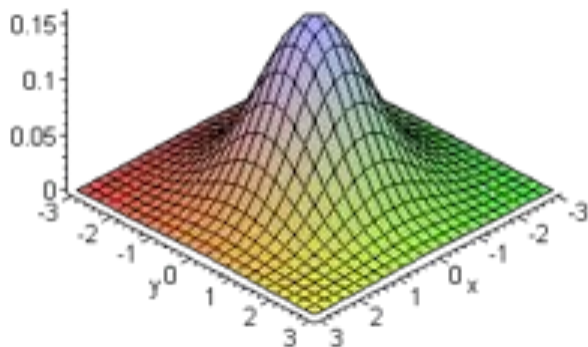


For some  $x$ , we have observed the outcome  $f(x)$ .

We want the conditional probability of  $f(x^*)$

$$f(x^*) | x^*, x, f(x)$$





$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right]$$

Easy to get the conditional probability of one of the variables, given the other.

"Gaussian Process"

$$\begin{pmatrix} f \\ f^* \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{pmatrix} K & K^* \\ (K^*)^T & K^{**} \end{pmatrix} \right]$$

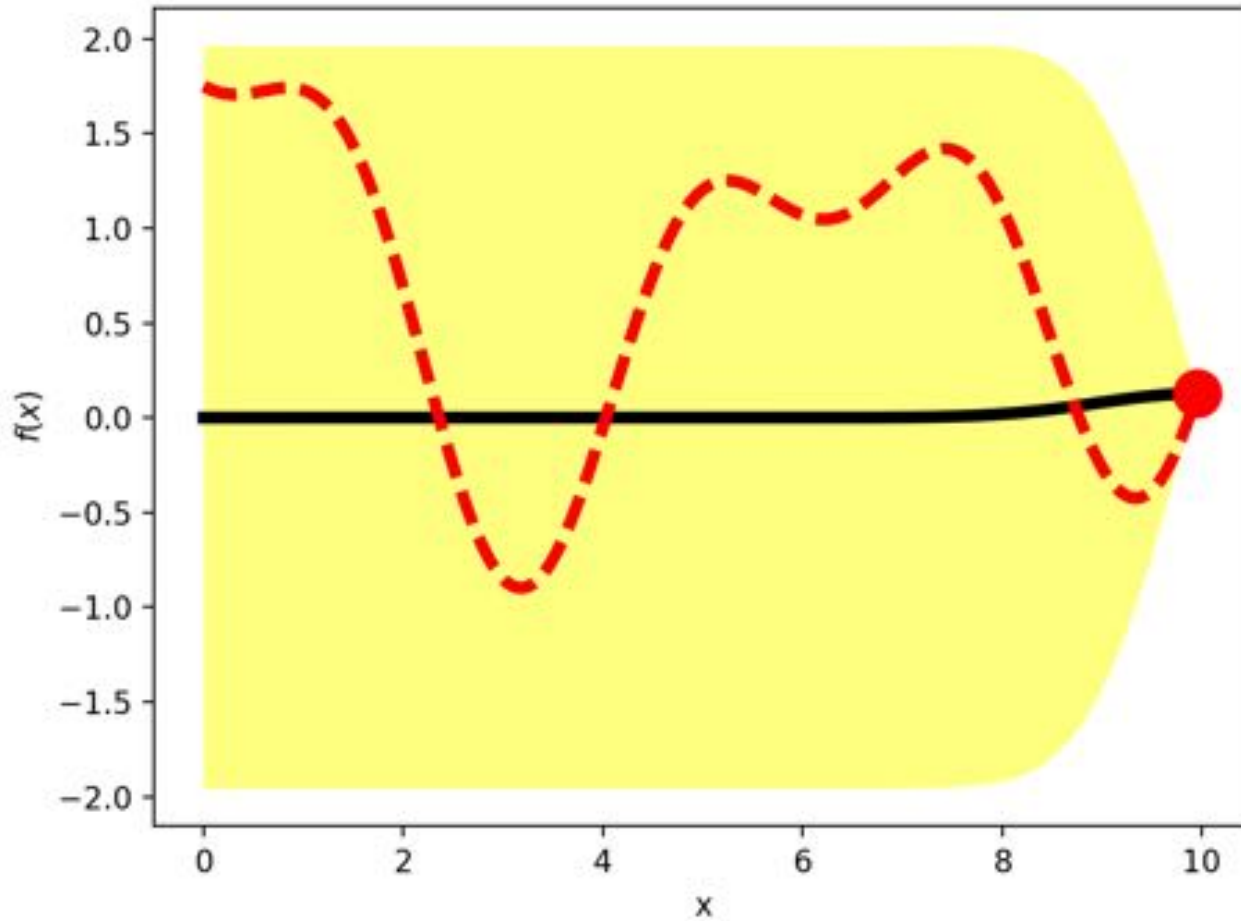
We express the covariance of the output in terms of the input.

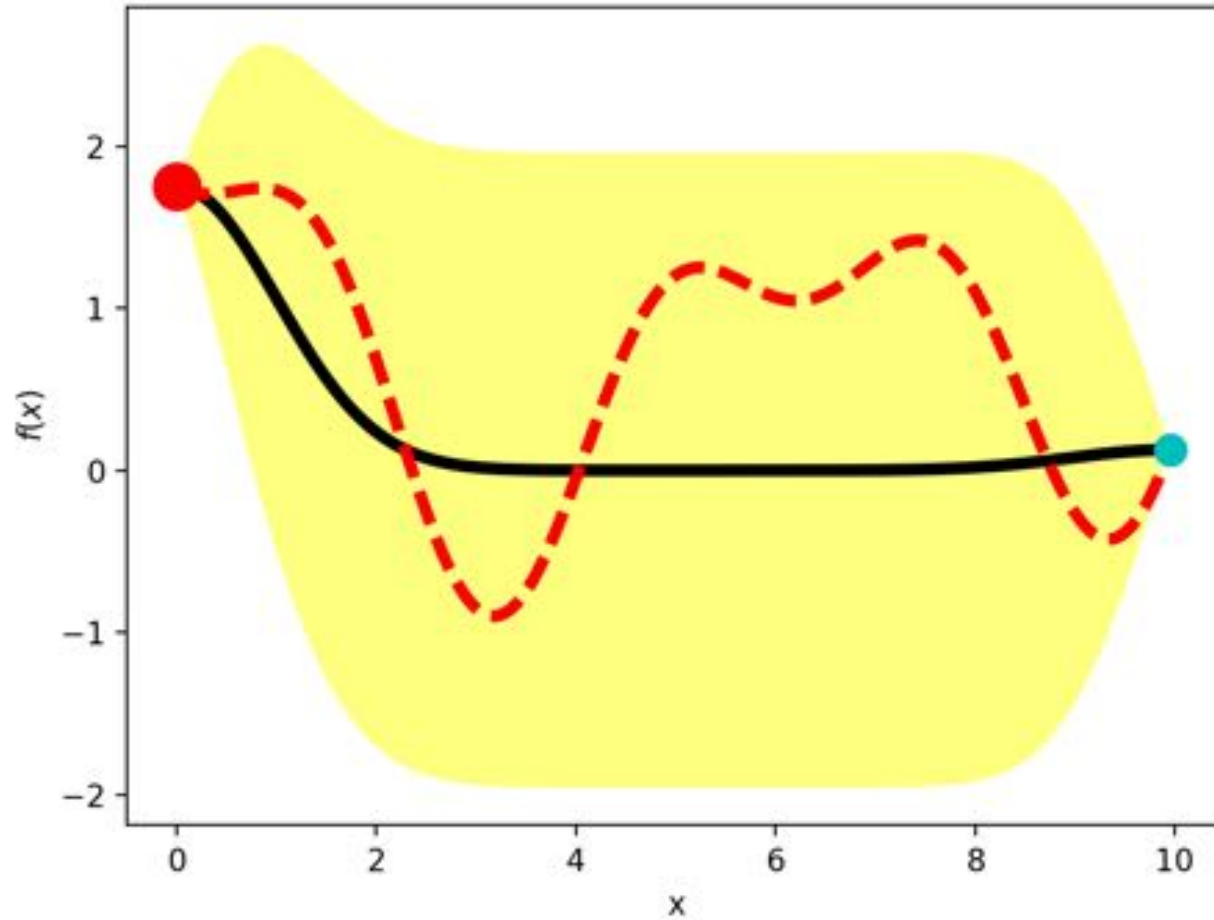
$$\text{cov}(f(x_p), f(x_q)) = k(x_p, x_q) = \exp \left[ -\frac{1}{2} |x_p - x_q|^2 \right]$$

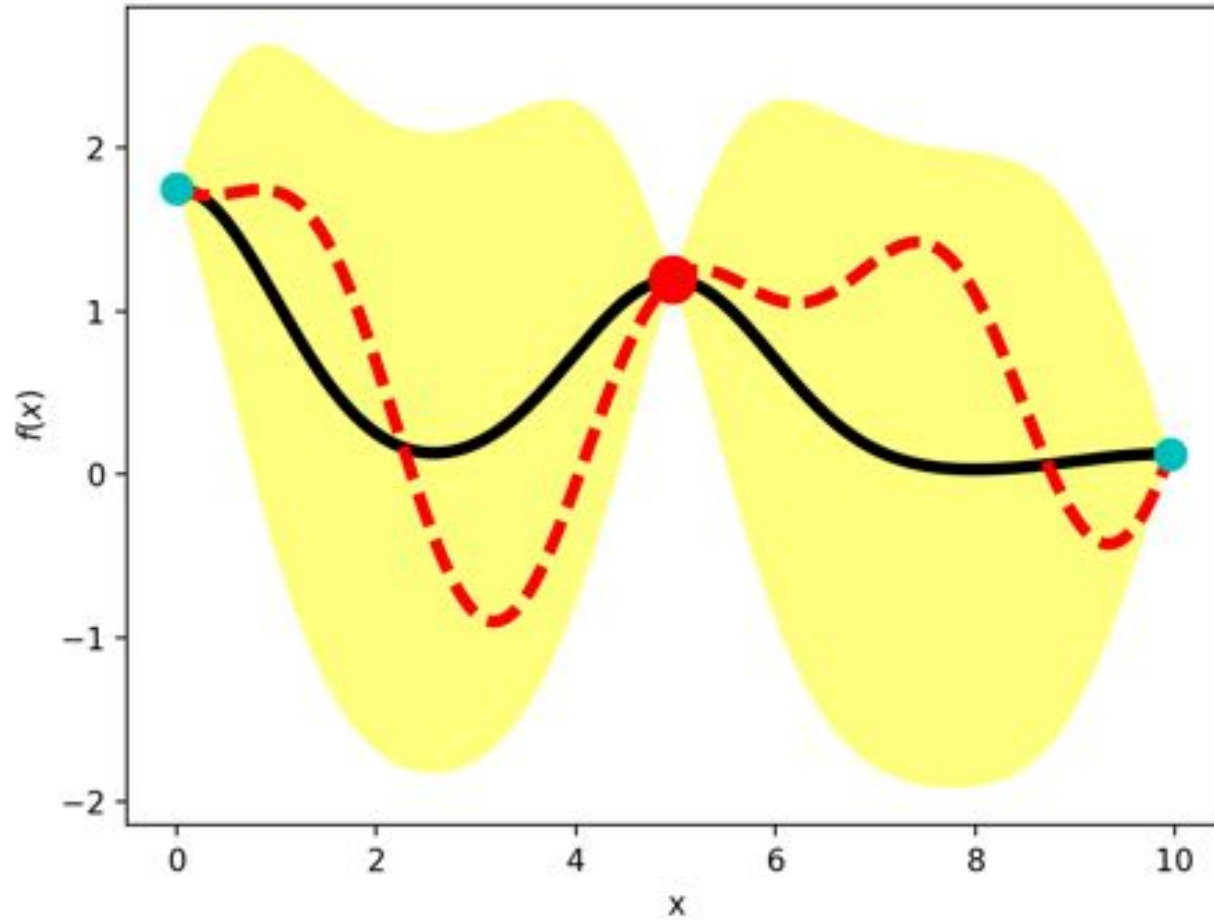
The prediction  $f^*$ , conditioned on the observations, given a covariance function, is given by:

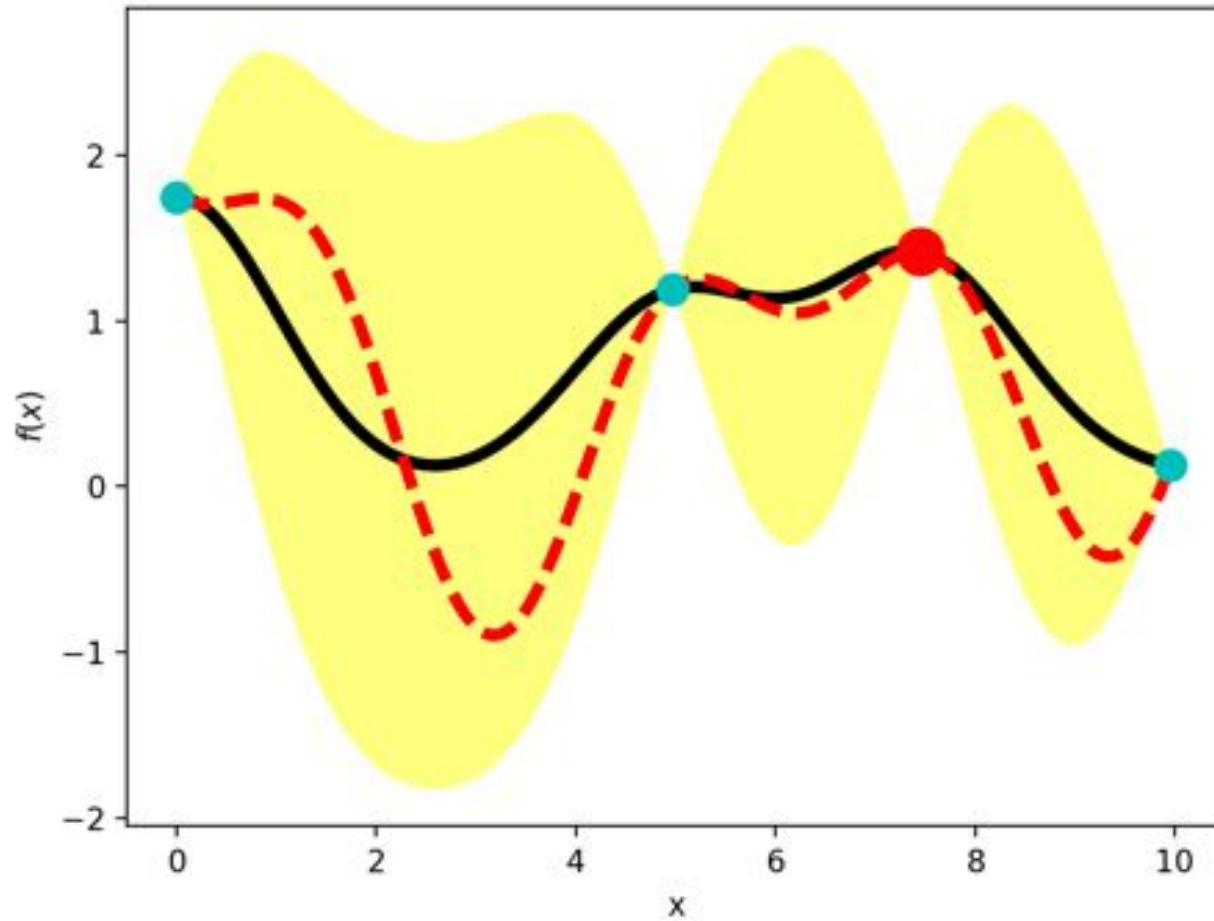
$$\mathbf{f}^* | X^*, X, \mathbf{f} \sim \mathcal{N} [K(X^*, X)K(X, X)^{-1}\mathbf{f}, K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)]$$

Cost:  $O(N^3)$

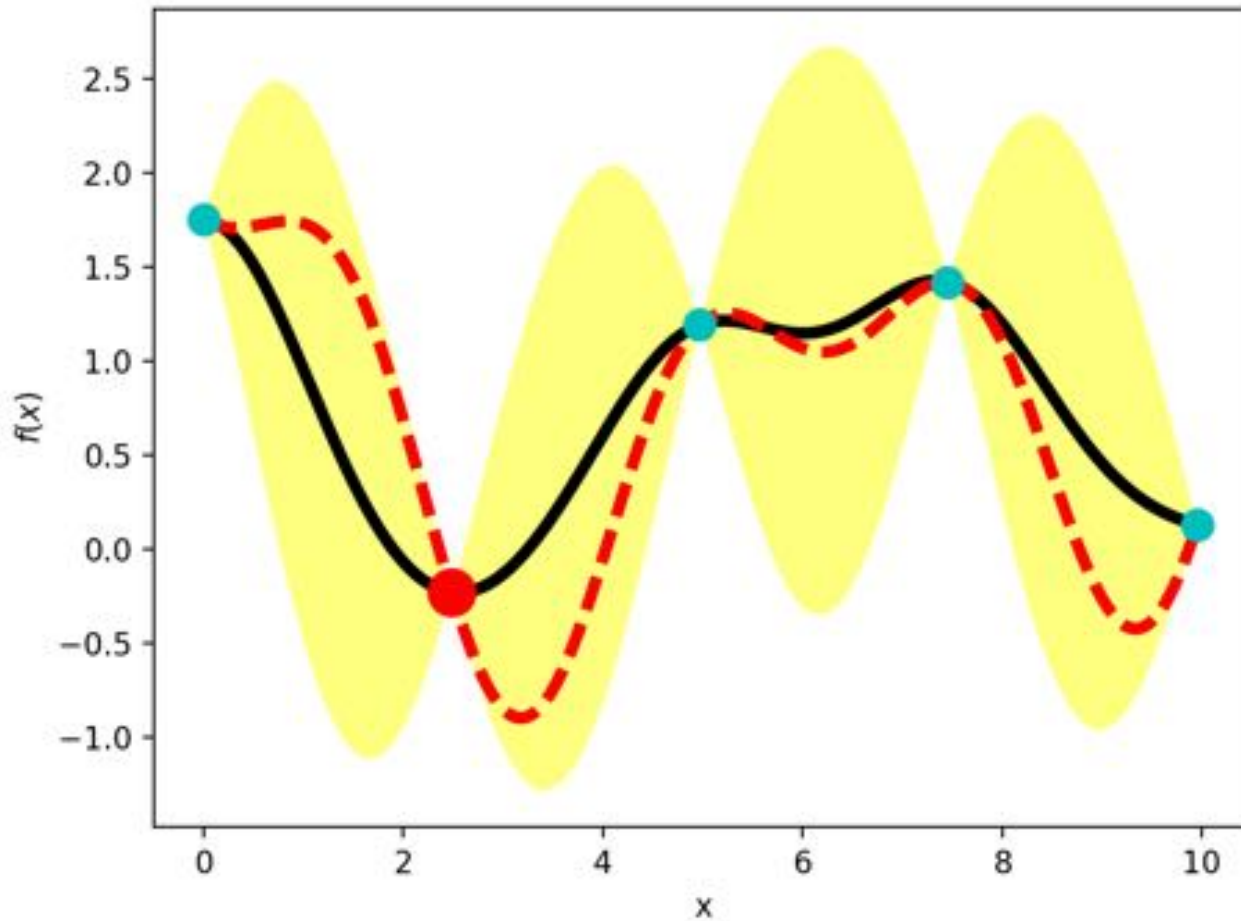


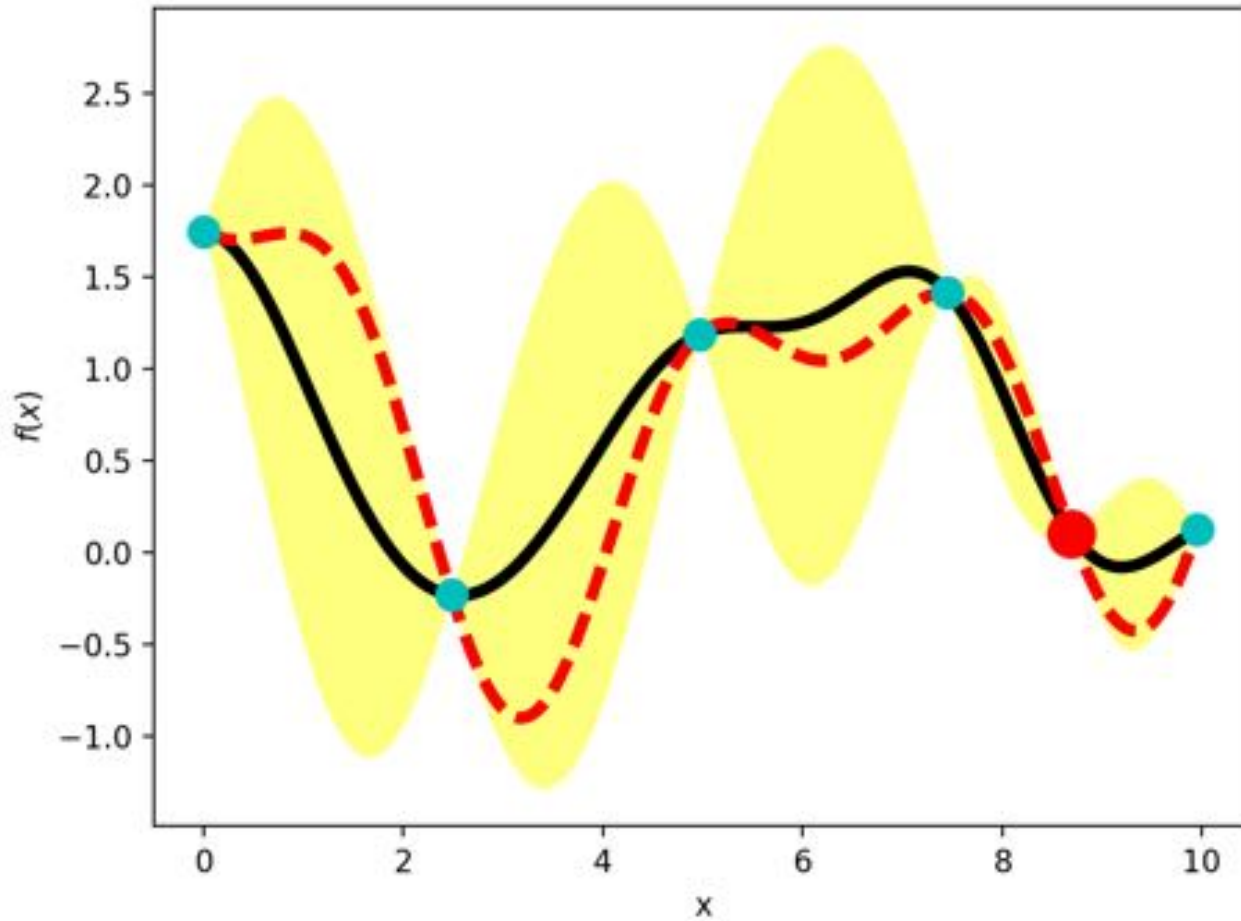


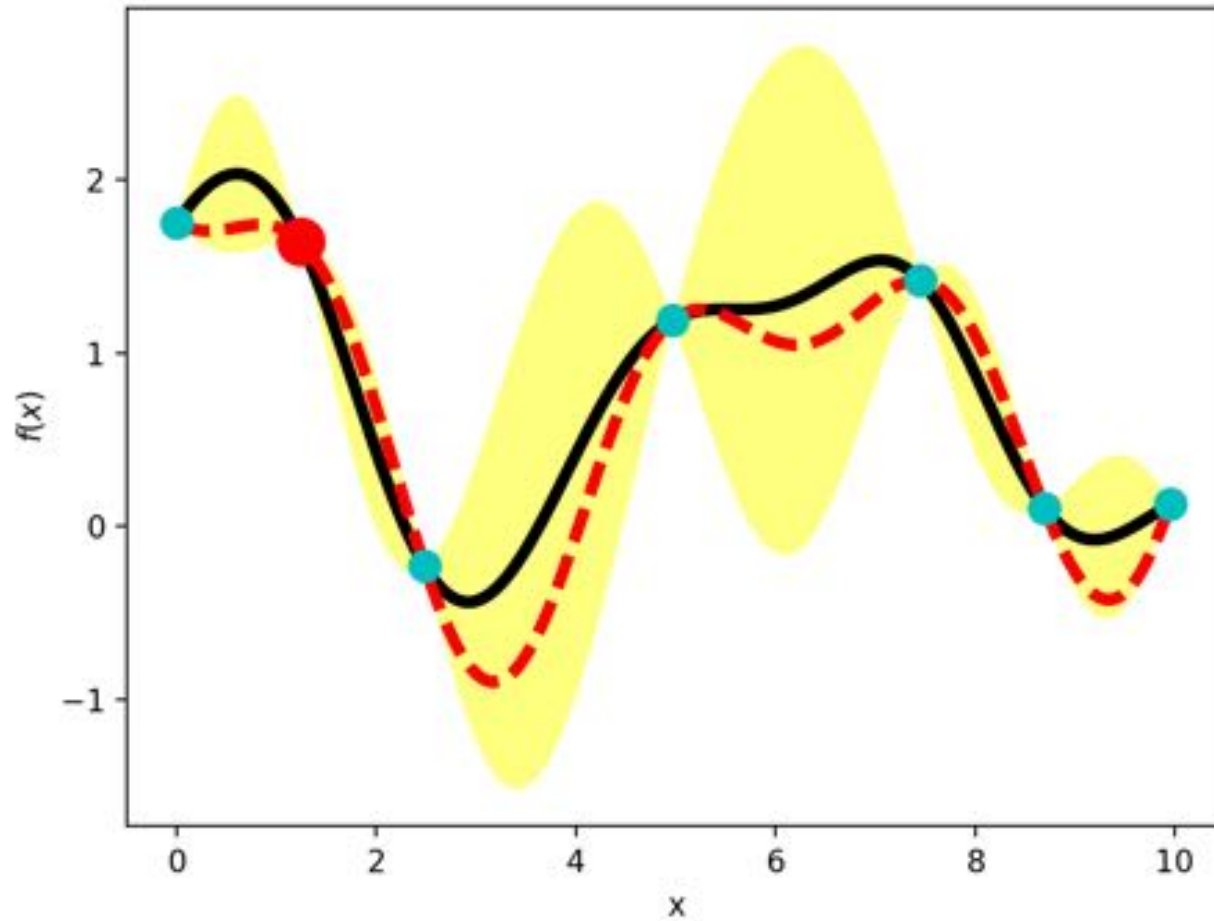


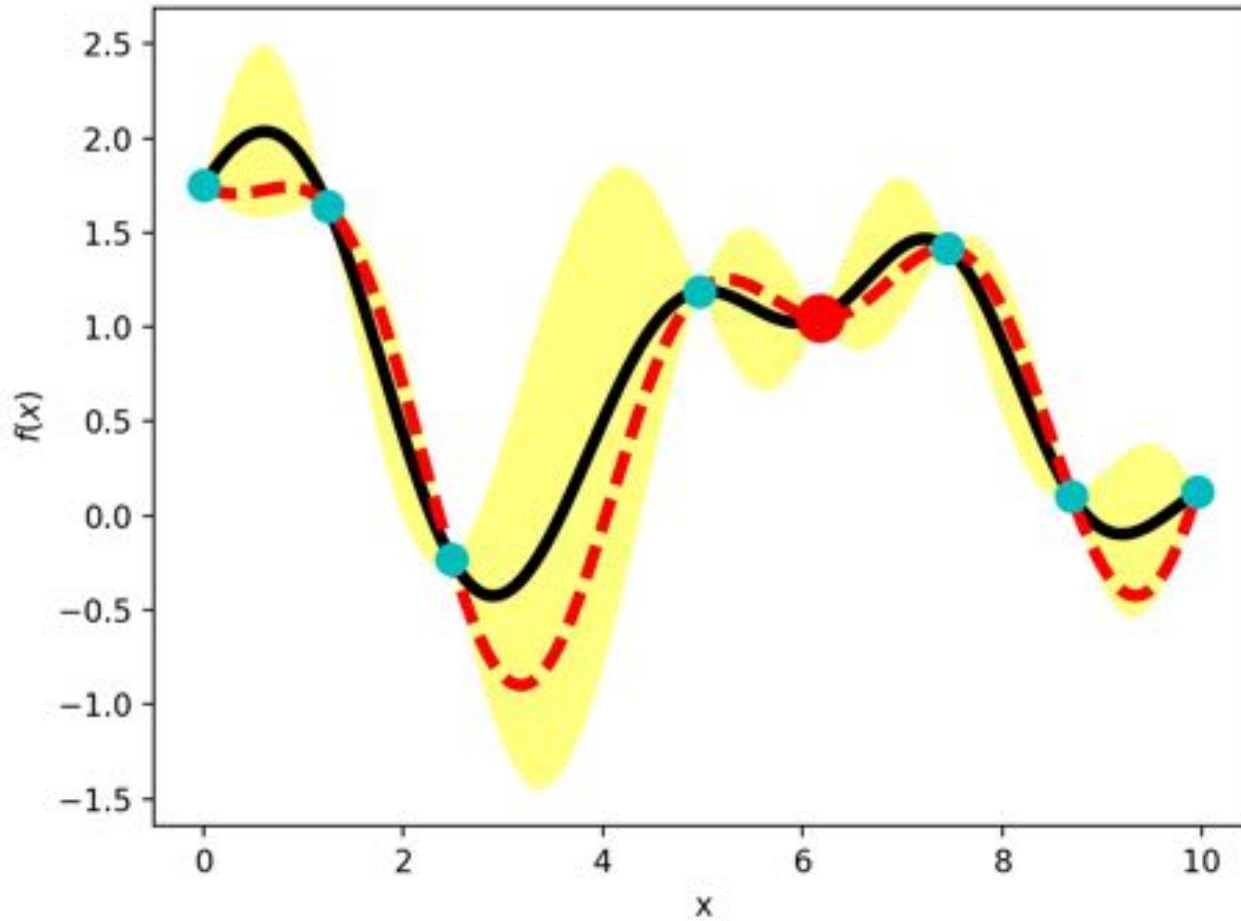


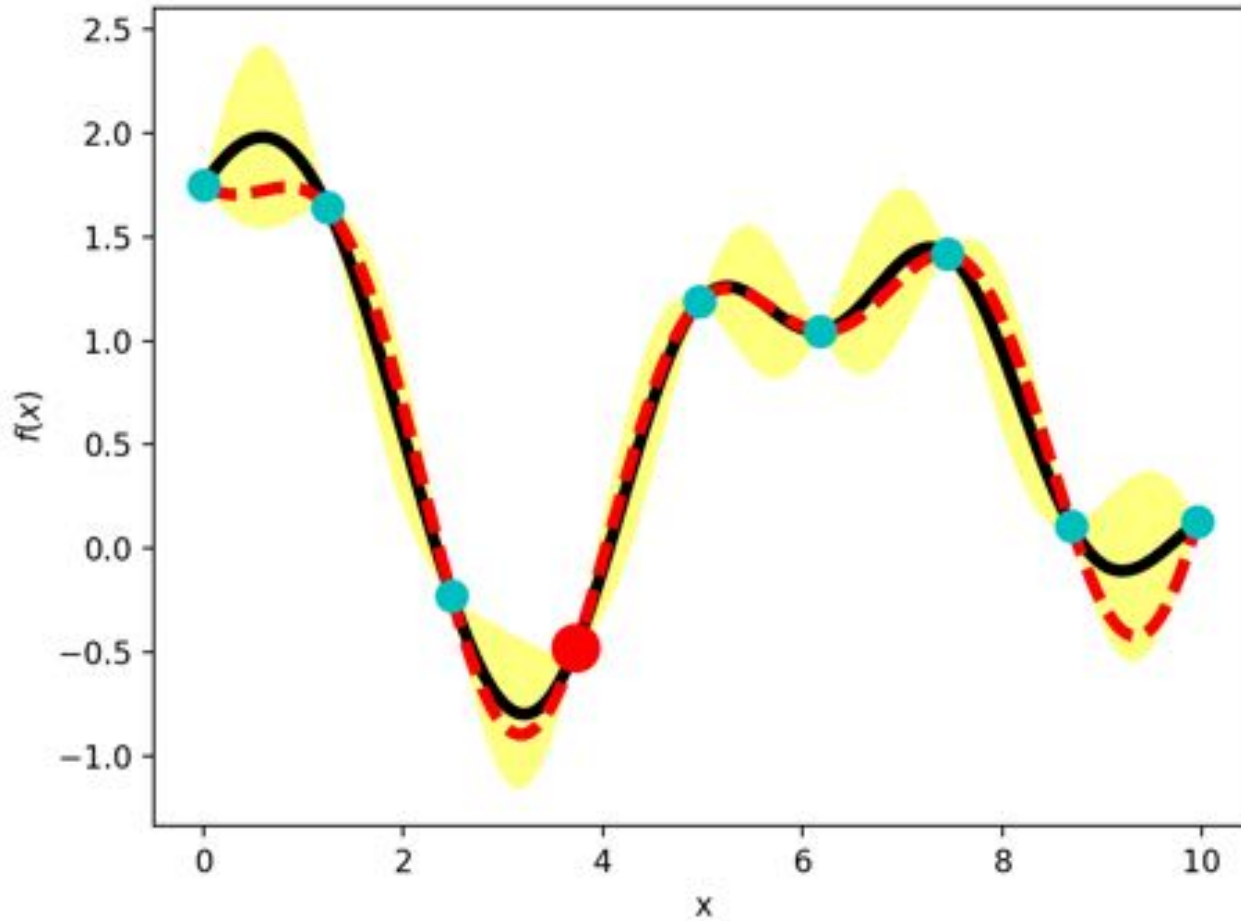




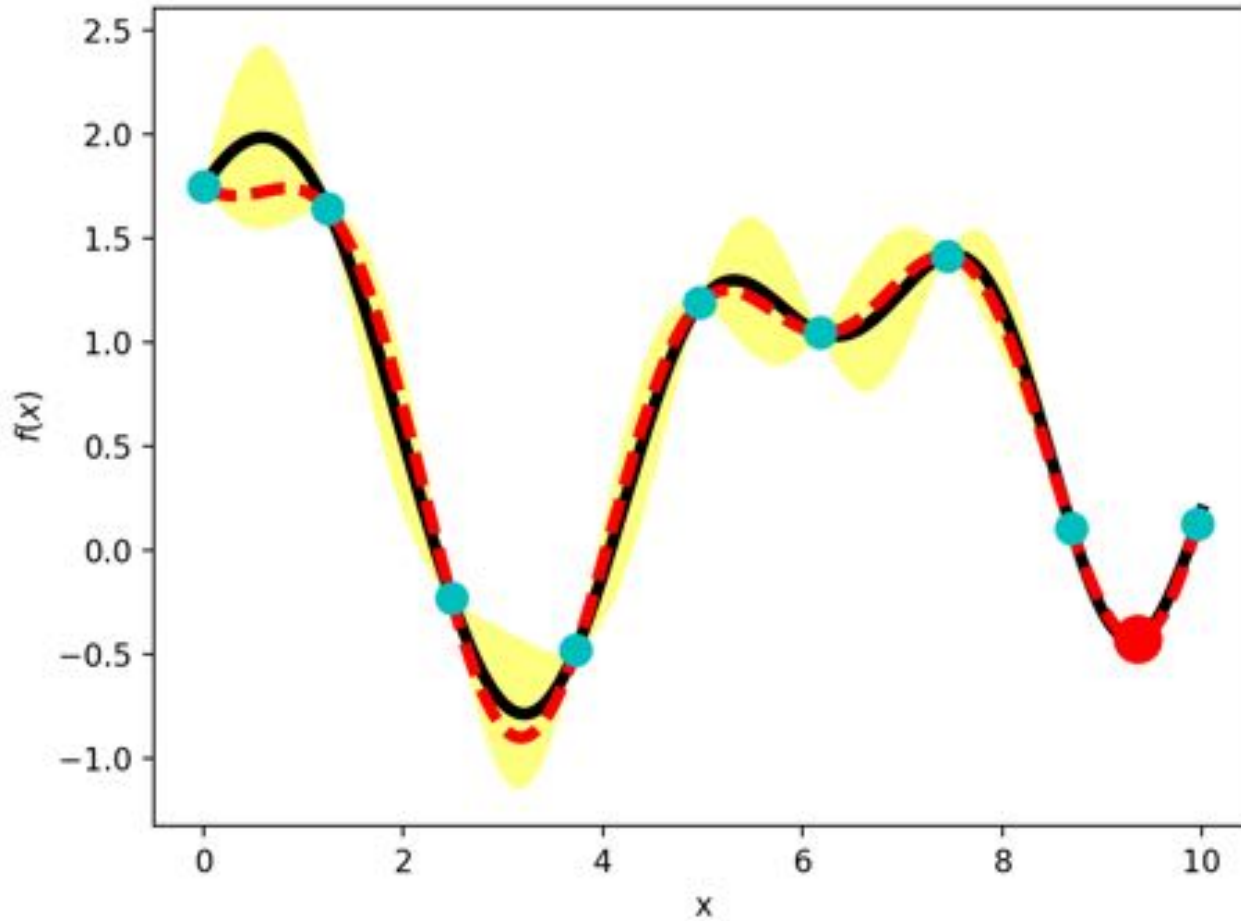


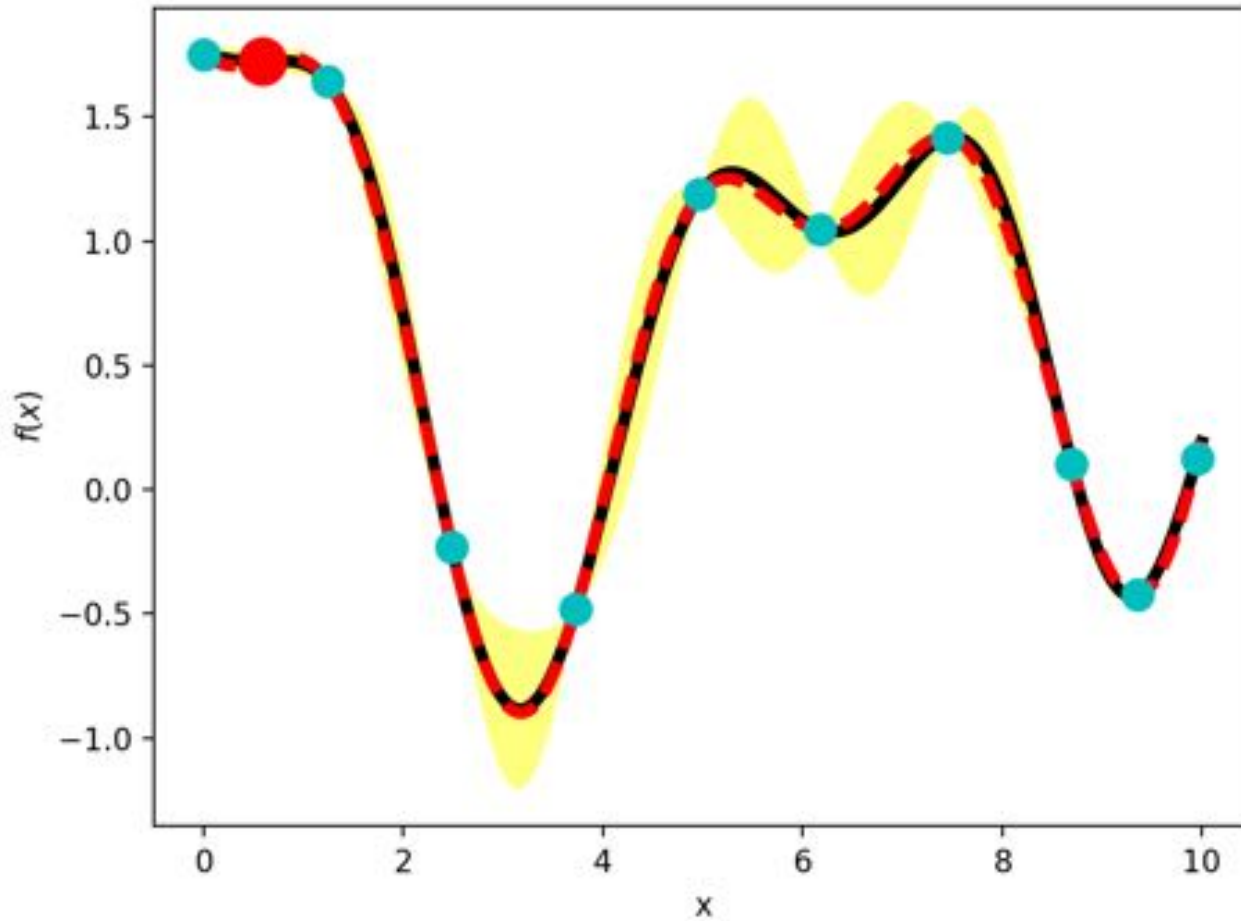


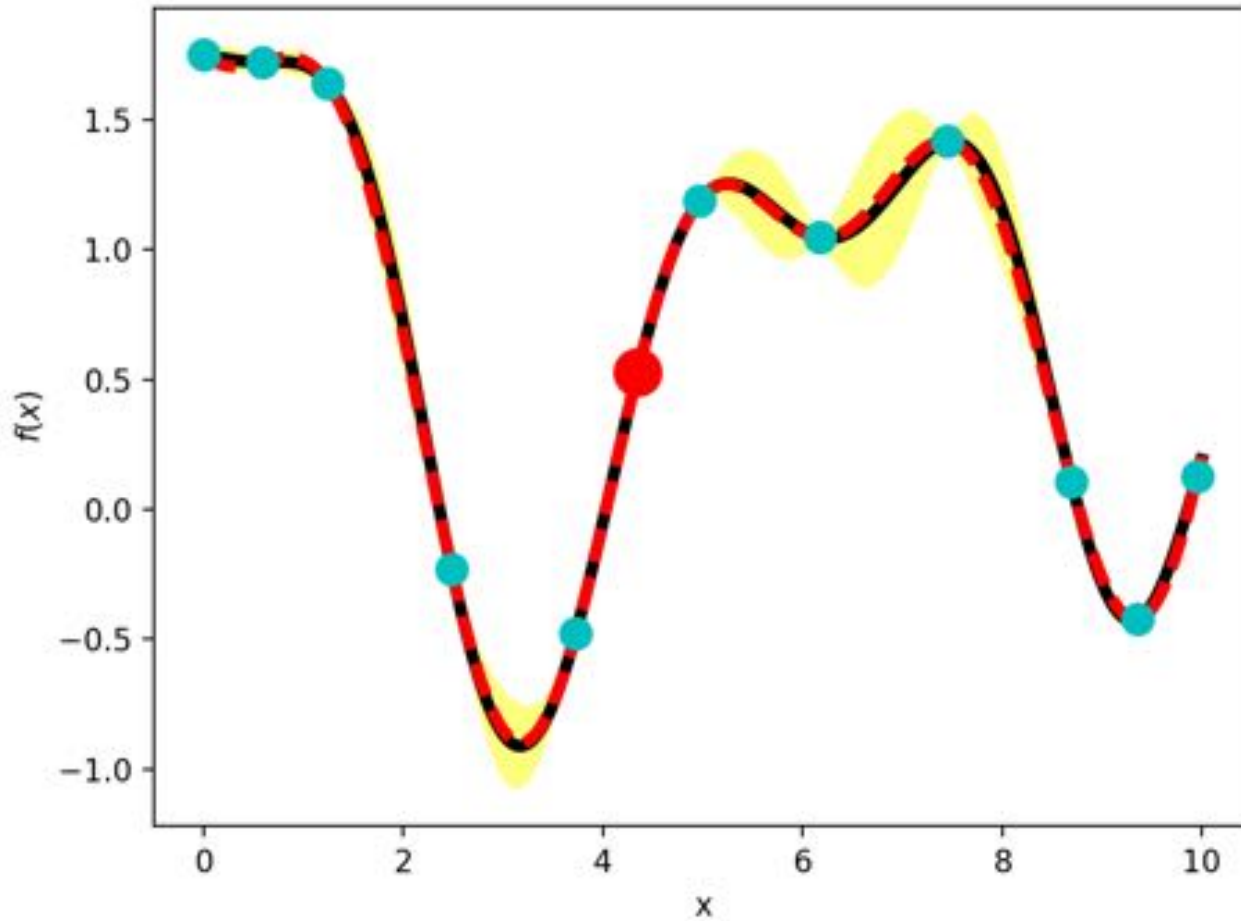


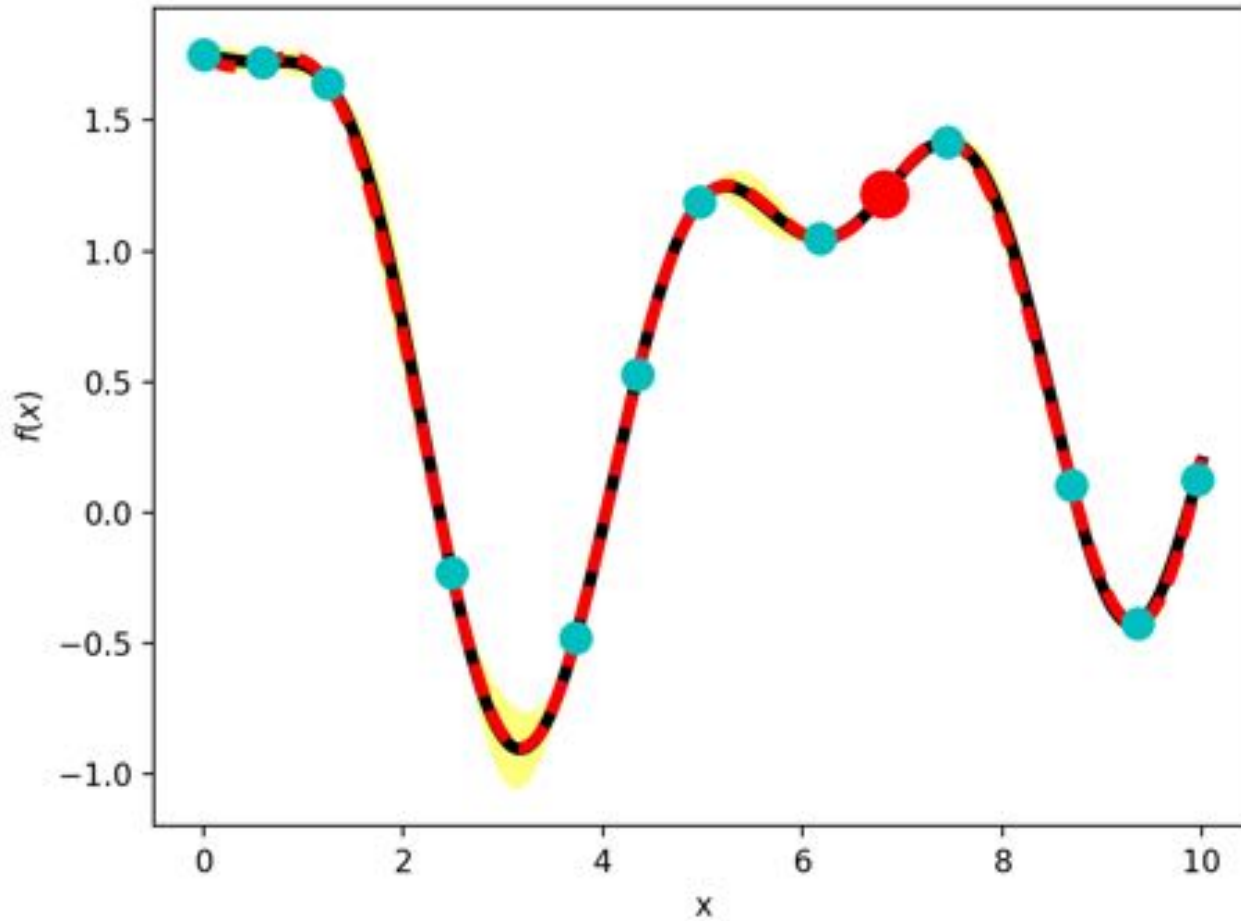


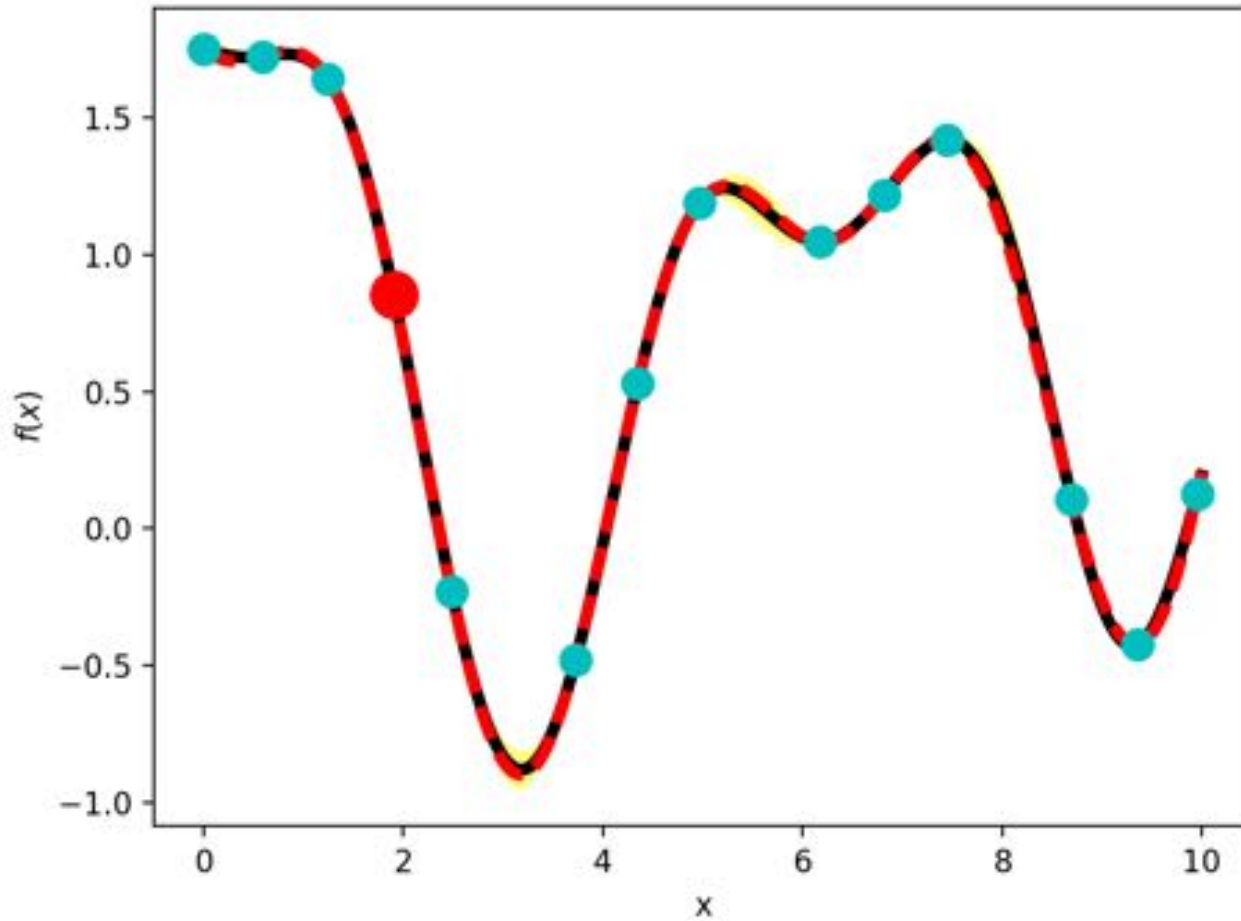










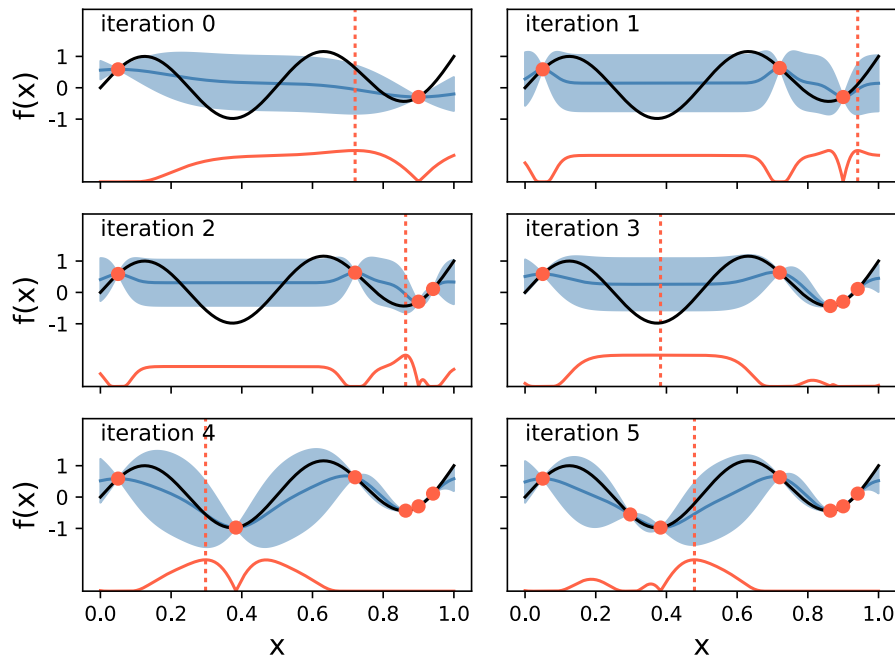




# BAYESIAN OPTIMIZATION

Jonas Mockus (2013). Bayesian approach to global optimization: theory and applications. Kluwer Academic.

When your objective function is expensive to evaluate !



Setup prior belief about the function.

This is often a **Gaussian process**

$$p(f) = \mathcal{GP}(f; \mu, K)$$

Confront prior with some **data**

$$\mathcal{D}_{1:n} = [x_1, x_2, \dots, x_n; f_1, f_2, \dots, f_n]$$

Update the posterior description of the Unknown function

$$p(f|\mathcal{D}_{1:n}) = \mathcal{GP}(f; \mu_{f|\mathcal{D}_{1:n}}, K_{f|\mathcal{D}_{1:n}})$$

**Decide where to sample next**

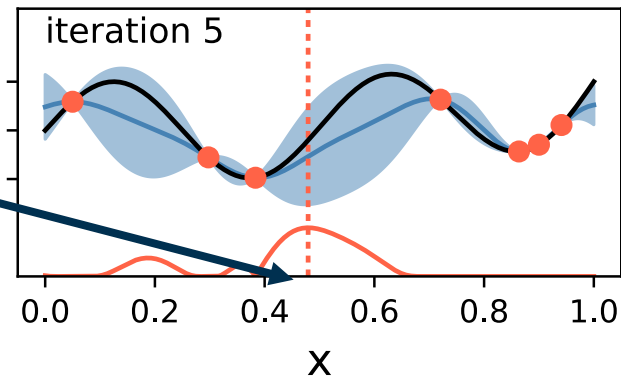
$$x_{n+1}$$

Augment the data and update the **Gaussian process**  $\mathcal{D}_{1:n+1}$

# ACQUISITION FUNCTION

Expected Improvement  $x_{n+1} = \operatorname{argmax} \mathcal{A}(x)$

It is not clear at the outset which acquisition function will work best. Problem specific, but EI is commonly used.



$f_{\min}$ : insofar lowest recorded function value.

$$\begin{aligned} \mathcal{A}(x) &= \langle u(x) \rangle = \int_{f(x)} \max(0, f_{\min} - f(x)) p(f(x) | \mathcal{D}_{1:n}) df && \text{Reward expected reduction in } f \\ &&& \text{in proportion to the reduction.} \\ &= (f_{\min} - \mu(x)_{\mathcal{D}}) \Phi \left( \frac{f_{\min} - \mu(x)_{\mathcal{D}}}{\sigma(x)_{\mathcal{D}}} \right) + \sigma(x)_{\mathcal{D}} \mathcal{N} \left( \frac{f_{\min} - \mu(x)_{\mathcal{D}}}{\sigma(x)_{\mathcal{D}}}; 0, 1 \right) \end{aligned}$$

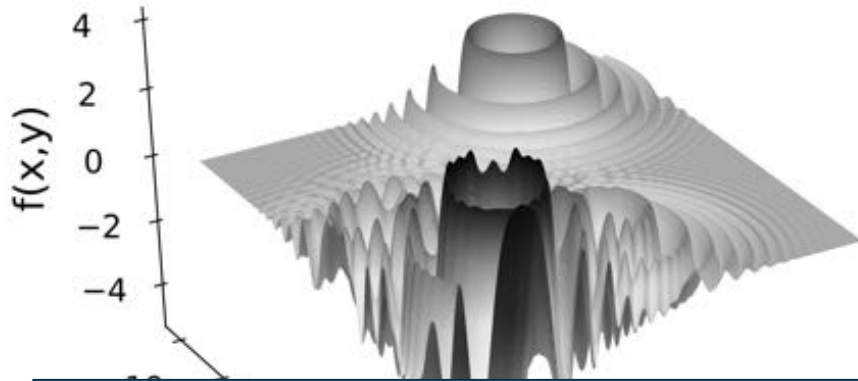
## Exploitation

Sampling areas of  
likely improvement

## Exploration

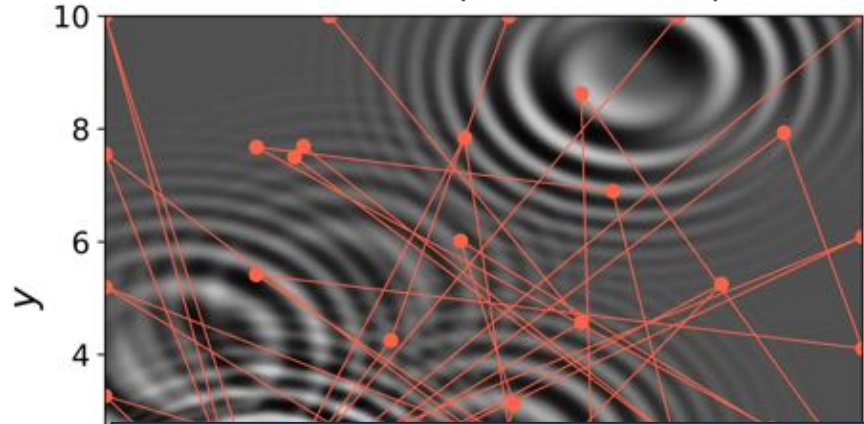
Sampling areas of  
high uncertainty

# BAYESIAN OPTIMIZATION



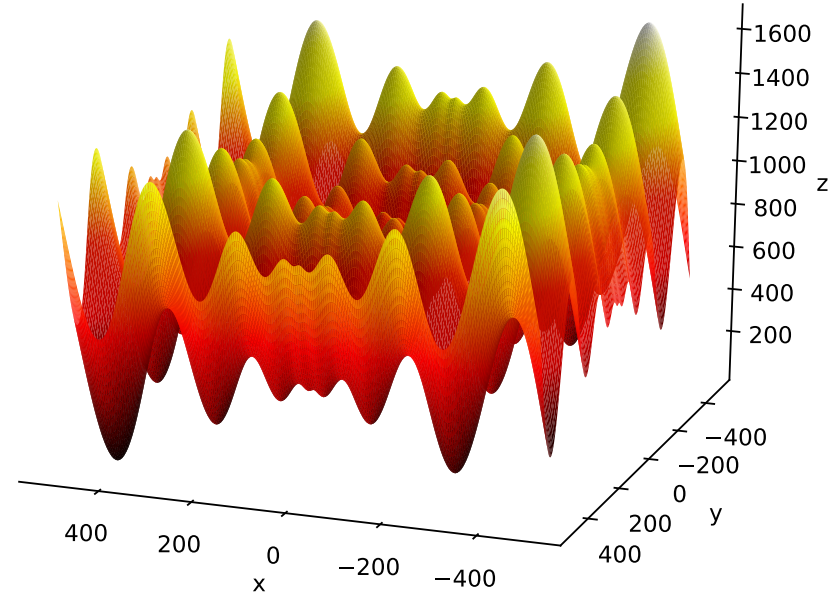
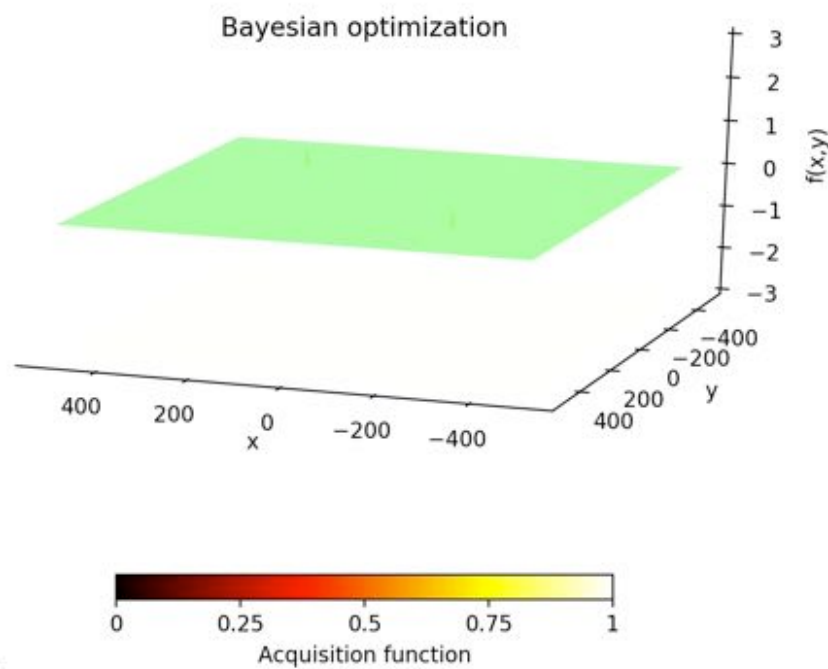
Bayesian optimization can easily avoid local minima and explore the parameter domain. It can exploit prior assumptions about the function, and utilize all previous function evaluations.

Exploitation vs. Exploration



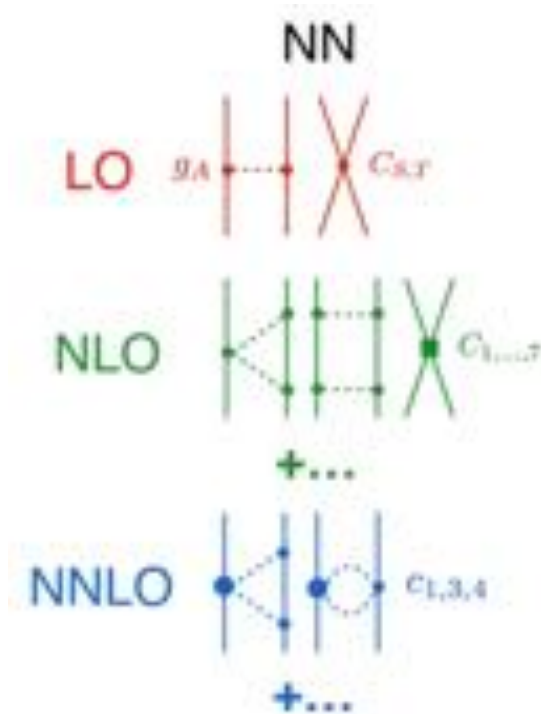
It can be challenging to scale Bayesian optimization to multi-dimensional parameter domains. So, how far can we take this approach in nuclear physics applications?

# BAYESIAN OPTIMIZATION



# NP SCATTERING, BELOW 75 MEV

R. Navarro-Perez et al, Phys. Rev. C **88**, 064002 (2013).



Simple enough to benchmark different BayesOpt algorithms. Complex enough provide a realistic setting.

**Theoretical model (N2LO chiral EFT) has 12 parameters.**

We have a model that we want to minimize. We then sample how well different solvers (optimization algorithms) perform by varying the starting points.

We use the same 24 quasi-random (Sobol) starting points for all solvers.

**Benchmarking 12 different BayesOpt solvers:**

- 3 Gaussian process kernels (RBF, Matern 3/2, Matern 5/2)
- 2 Acquisition functions (EI, LCB)
- With and without Automatic Relevance Determination (ARD)

# DATA PROFILES FOR BENCHMARKING

$$d_s(\alpha) = \frac{1}{|\mathcal{P}|} \text{size} \left\{ p \in \mathcal{P} : \frac{t_{p,s}}{d_p + 1} \leq \alpha \right\}$$

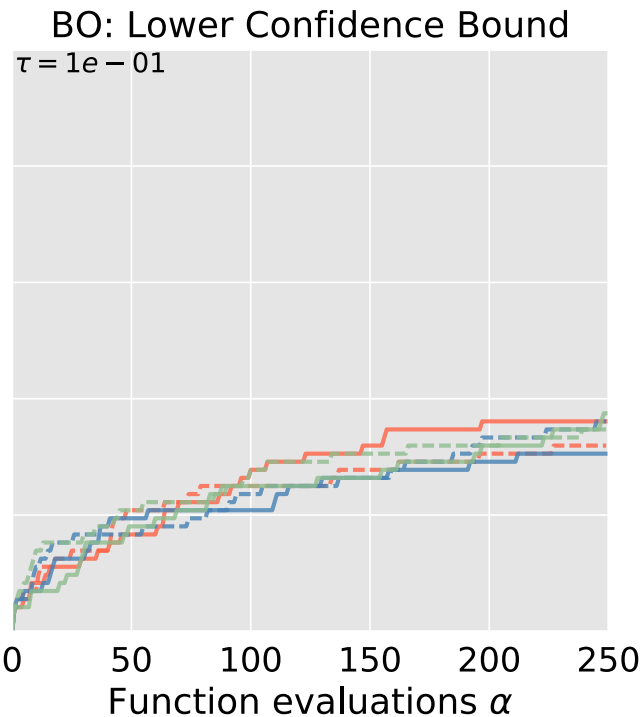
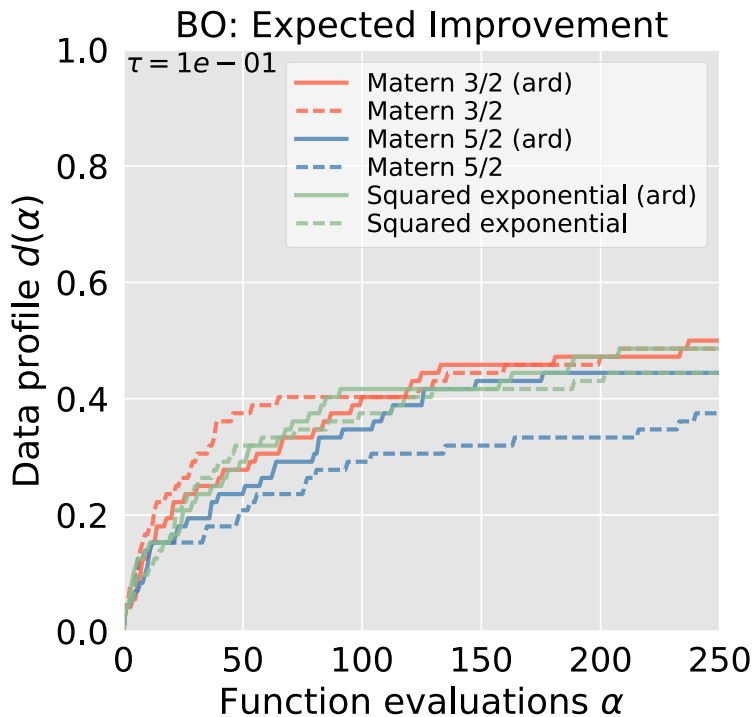
The data profile enables direct comparison between a set of optimization algorithms  $S$  all of which applied to a set of optimization problems  $P$ . For each  $(s, p) \in S \times P$ , the performance measure  $t_{s,p} > 0$  denotes the number of function evaluations that are required for optimization algorithm  $s$ , applied to a problem  $p$ , to satisfy some convergence criterion. **Thus,  $d_s(\alpha)$  is the fraction of problems that can be solved within  $\alpha$  function calls.** The performance measure can be normalized to  $d_p + 1$  (dimensional normalization) for comparing optimizers across different spaces.

J.J. Moré, S.M. Wild, Benchmarking Derivative-Free Optimization Algorithms, SIAM J. Optim. 20 (2009) 172–191. doi:10.1137/080724083.

# DATA PROFILES

*convergence criterion: 90% reduction*

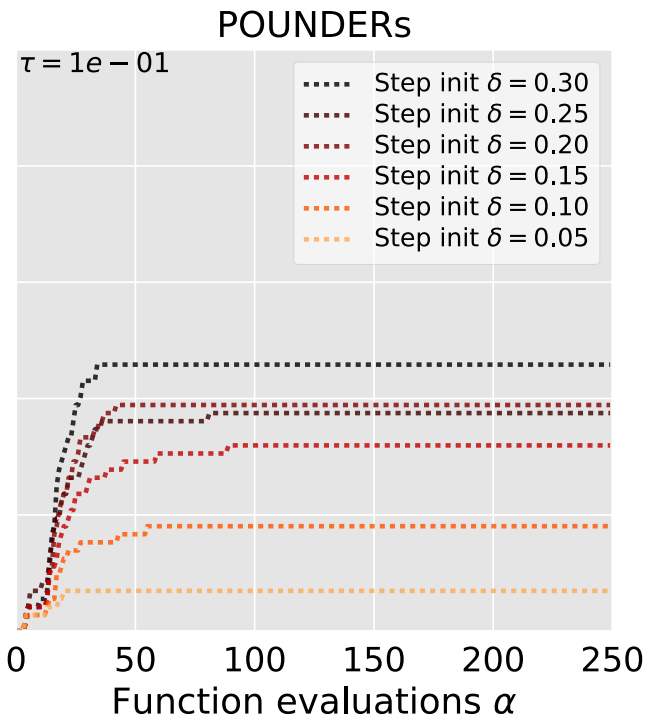
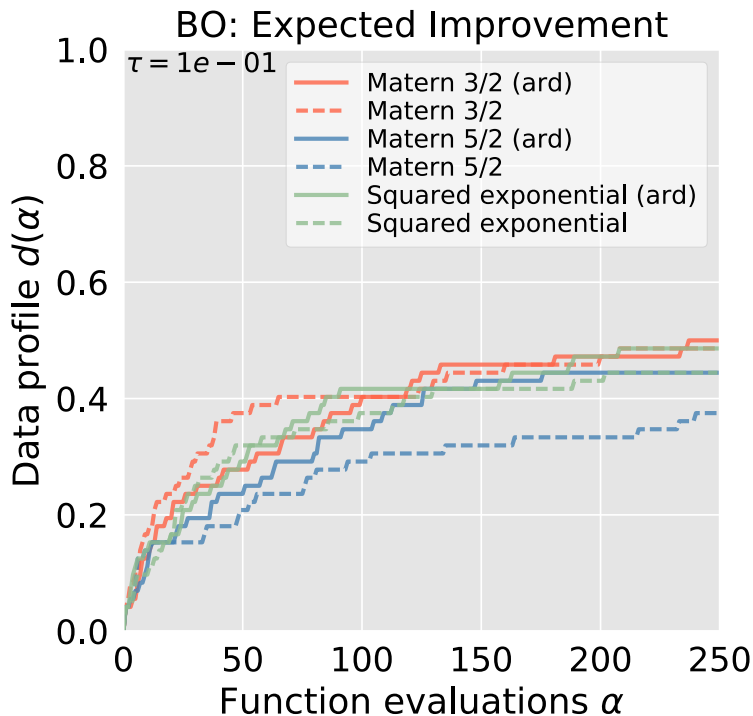
$$f(x_0) - f(x) \geq (1 - \tau)(f(x_0) - f_L).$$



# DATA PROFILES

*convergence criterion: 90% reduction*

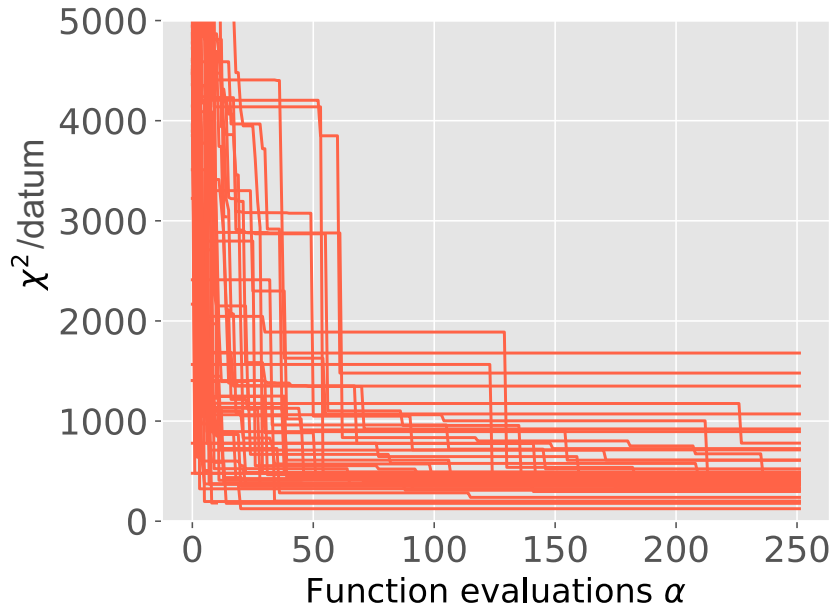
$$f(x_0) - f(x) \geq (1 - \tau)(f(x_0) - f_L).$$



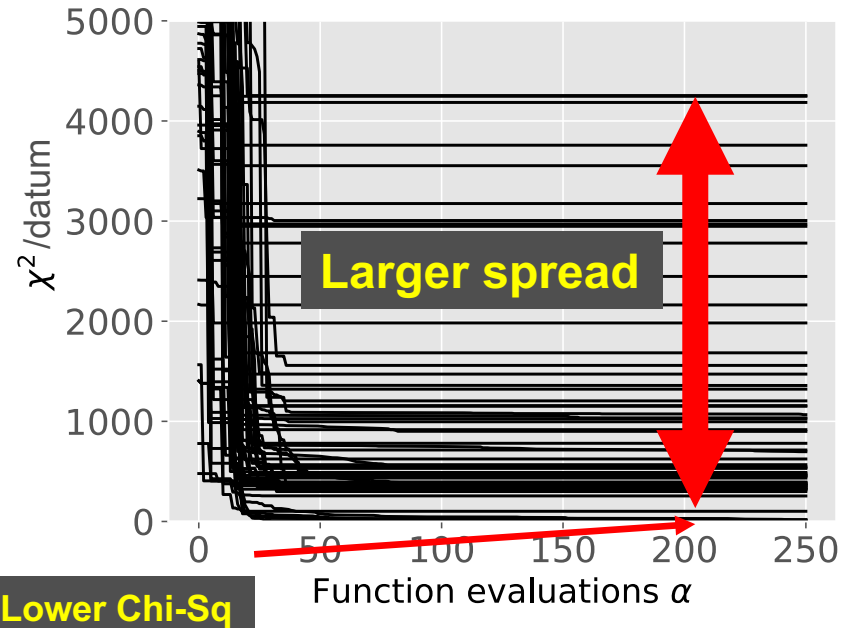


# TRACES (MINIMUM CHI-SQ VS ITERATION)

BayesOpt – EI – Matern 3/2



POUNDERs – Initial step length 0.3



# COMMENTS

- It is challenging to handle expensive objectives. Any additional information should be incorporated. Next steps:
  - Exploit that the objective function is sum of squared residuals. (ignored in BayesOpt now)
  - Exploit any additive structure in the objective function. (ignored in BayesOpt and POUNDERS now)
- BayesOpt can be infinitely tweaked and can balance exploration-exploitation strategies. It is not designed to locate the exact point of an optimum.
- POUNDERS is easy to use (!), can find a rather good minimizer fast, however only when launched with an optimal initial step length. Less explorative.
- For EFT: BayesOpt[El,Matern] exhibits overall good performance with **small spread**.
- Suggested strategy: Initiate optimization with BayesOpt. Refine with POUNDERS, and thereafter with possibly existing higher-order methods.

# ACKNOWLEDGEMENTS

*Devdatt Dubhasi*  
*Christian Forssén*  
*Andreas Johansson*  
*Håkan T. Johansson*  
*Hans Salomonsson*  
*Alexander Schliep*  
*Muhammed Azam Sheikh*



European Research Council  
Established by the European Commission

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758027 PrecisionNuclei).



Vetenskapsrådet





**CHALMERS**  
UNIVERSITY OF TECHNOLOGY