OAK RIDGE National Laboratory

Gaussian Processes and Bayesian Optimization: An Introduction

Rama Vasudevan
Center for Nanophase Materials Sciences,

AI for atoms: How to machine learn STEM
December 9th, 2020

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Gaussian Process Regression

- Splines, polynomials, other functional fits
  - Problem: Uncertainty depends on the model. Also, which model is the correct model?
- Bayesian approach:
  - Use Gaussian Processes
- A Gaussian process is a non-parametric Bayesian approach to regression, that finds a distribution over functions f(x) that are consistent with the observed data
  - The similarity between points is defined by a covariance matrix
  - The covariance is determined by a kernel function

Oak Ridge National Laboratory

Gaussian Process Regression

- Why bother with multivariate normal distributions? Two reasons:
  - (1) Easy to marginalize: when you have many, if you want to limit yourself to some subset, it is simple to integrate out the variables you don't want
  - (2) Easy to condition: Can write analytical solutions for conditioning. The conditioned distribution is also normal.

Multivariate Gaussian Distribution

$$y \sim N(\mu, \Sigma)$$ Mean vector, covariance matrix

Generalizes to a Gaussian Process

$$f \sim GP(m(x), \mathrm{k}(\mathrm{x}, \mathrm{x}'))$$ Mean function, covariance function

Covariance Function returns the Covariance Matrix

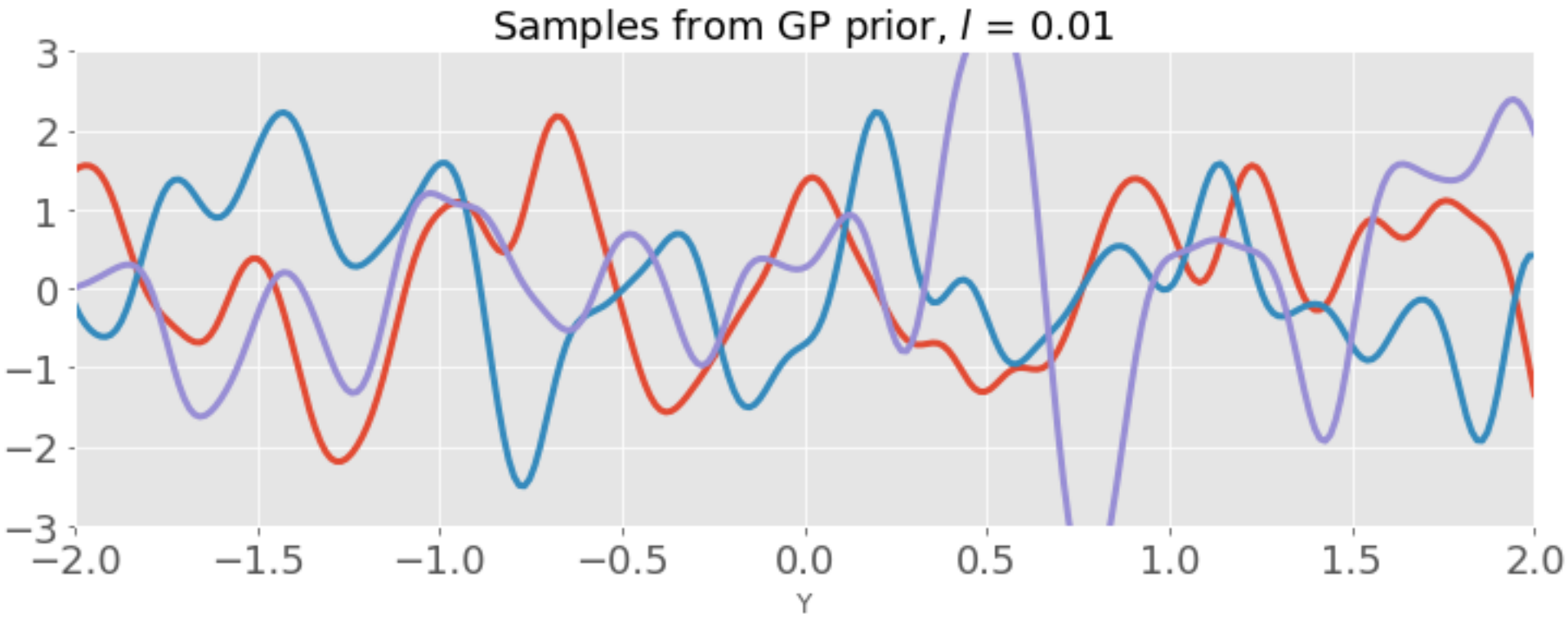
$$\Sigma \sim k(x, x' | \Phi)$$ Pass function values, return Cov. Mat.

Oak Ridge National Laboratory

Gaussian Process Regression

- Covariance matrix determines what type of functions we will allow.

$$\mathrm{k}(\mathrm{x}, \mathrm{x}') = \exp\left(-\frac{1}{2l}(x - x')^2\right)$$

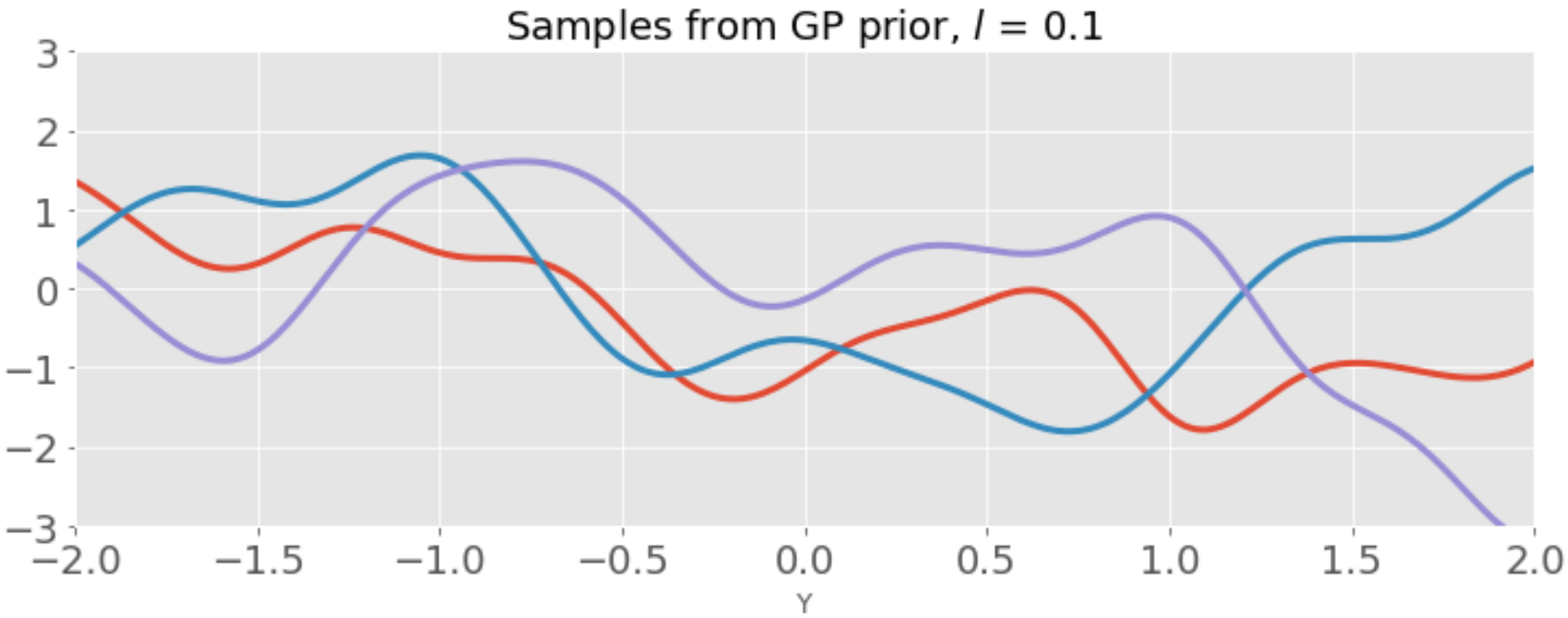


OAK RIDGE National Laboratory

Gaussian Process Regression

- Covariance matrix determines what type of functions we will allow.

$$\mathrm{k}(\mathrm{x}, \mathrm{x}') = \exp\left(-\frac{1}{2l}(x - x')^2\right)$$

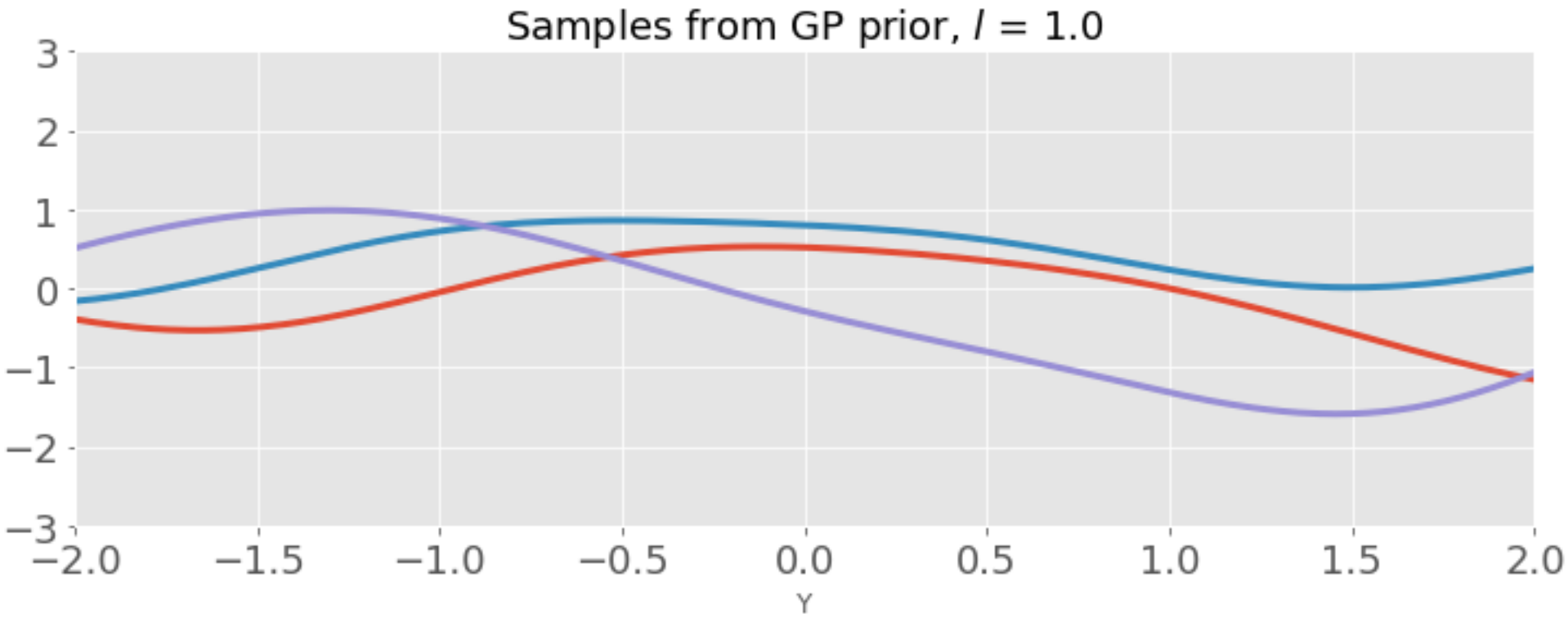


OAK RIDGE National Laboratory

Gaussian Process Regression

- Covariance matrix (kernel) determines what type of functions we will allow.

$$\mathrm{k}(\mathrm{x}, \mathrm{x}') = \exp\left(-\frac{1}{2l}(x - x')^2\right)$$



L controls the lengthscale – sort of how far points should be to make them independent of each other.

OAK RIDGE National Laboratory

Gaussian Process Regression

- Consider a (gaussian) joint probability distribution with two variables, as follows

Similarity of x_1 values to x_2 values

Variance in x_1

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

Variance in x_2

Nonzero entries off-diagonal indicate the correlations

OAK RIDGE National Laboratory

Gaussian Process Regression

- Generalize this to the function space
- When we observe the data, we now have to calculate the posterior over the functions. This posterior is a joint distribution over function values observed and those not observed.

Function values training points

Similarity to observed points

Similarity of training values to test values

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix} \right)$$

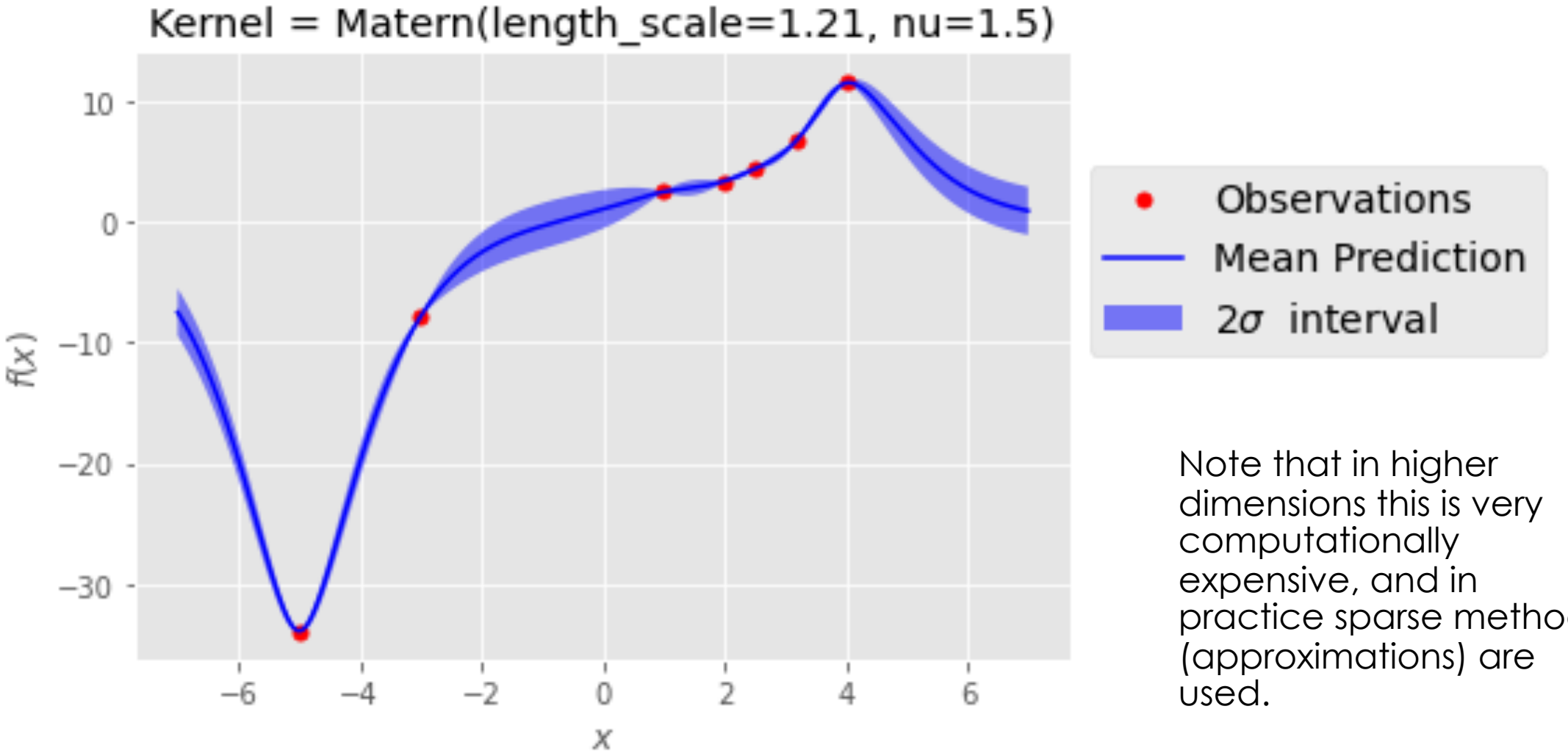
Function values test points

Similarity of test values to each other

Read Rasmussen and Williams for all the math
http://www.gaussianprocess.org/gpml/

OAK RIDGE National Laboratory

Gaussian Process Regression



Note that in higher dimensions this is very computationally expensive, and in practice sparse methods (approximations) are used.

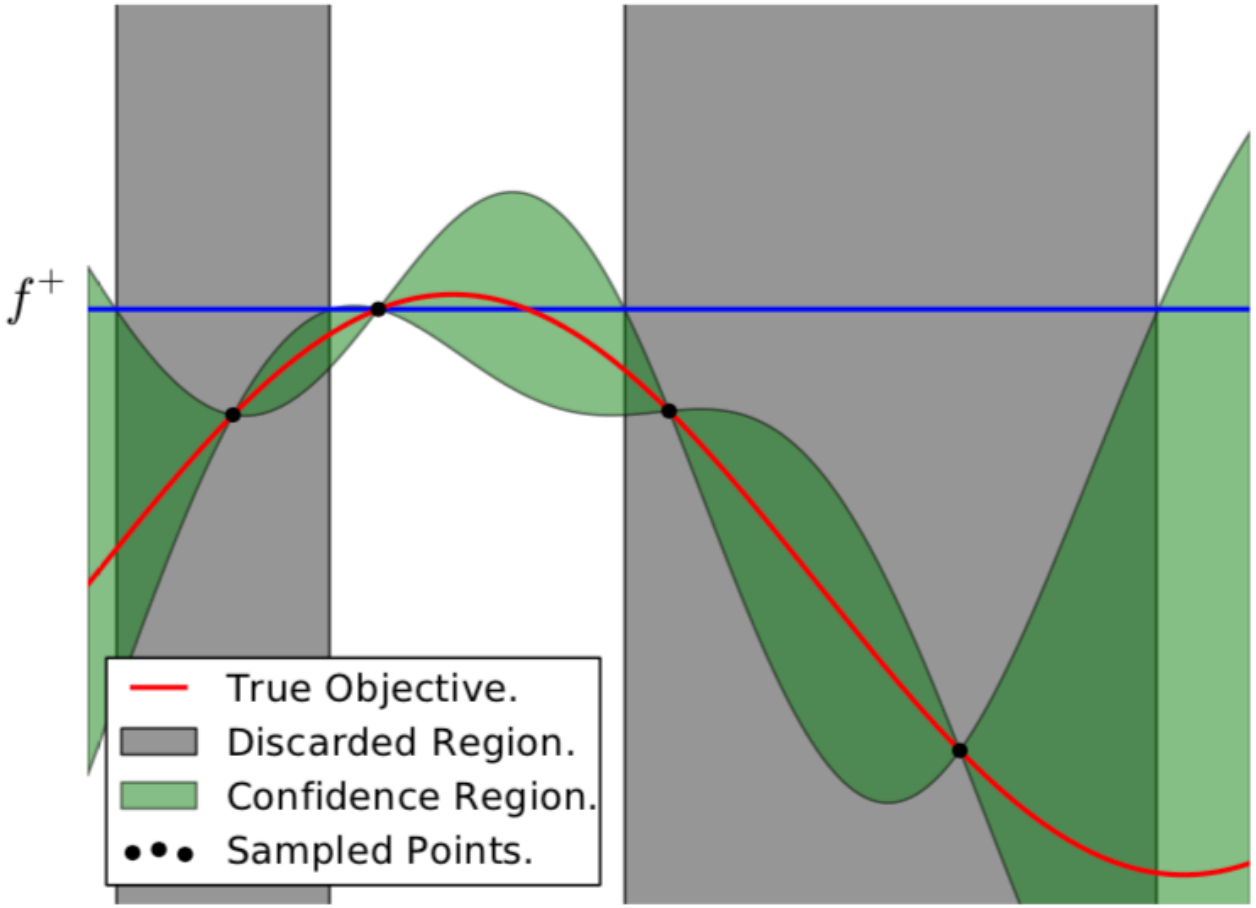
OAK RIDGE National Laboratory

Gaussian Process Regression

Open (01_Gaussian_Processes.ipynb)

OAK RIDGE National Laboratory

Bayesian Optimization



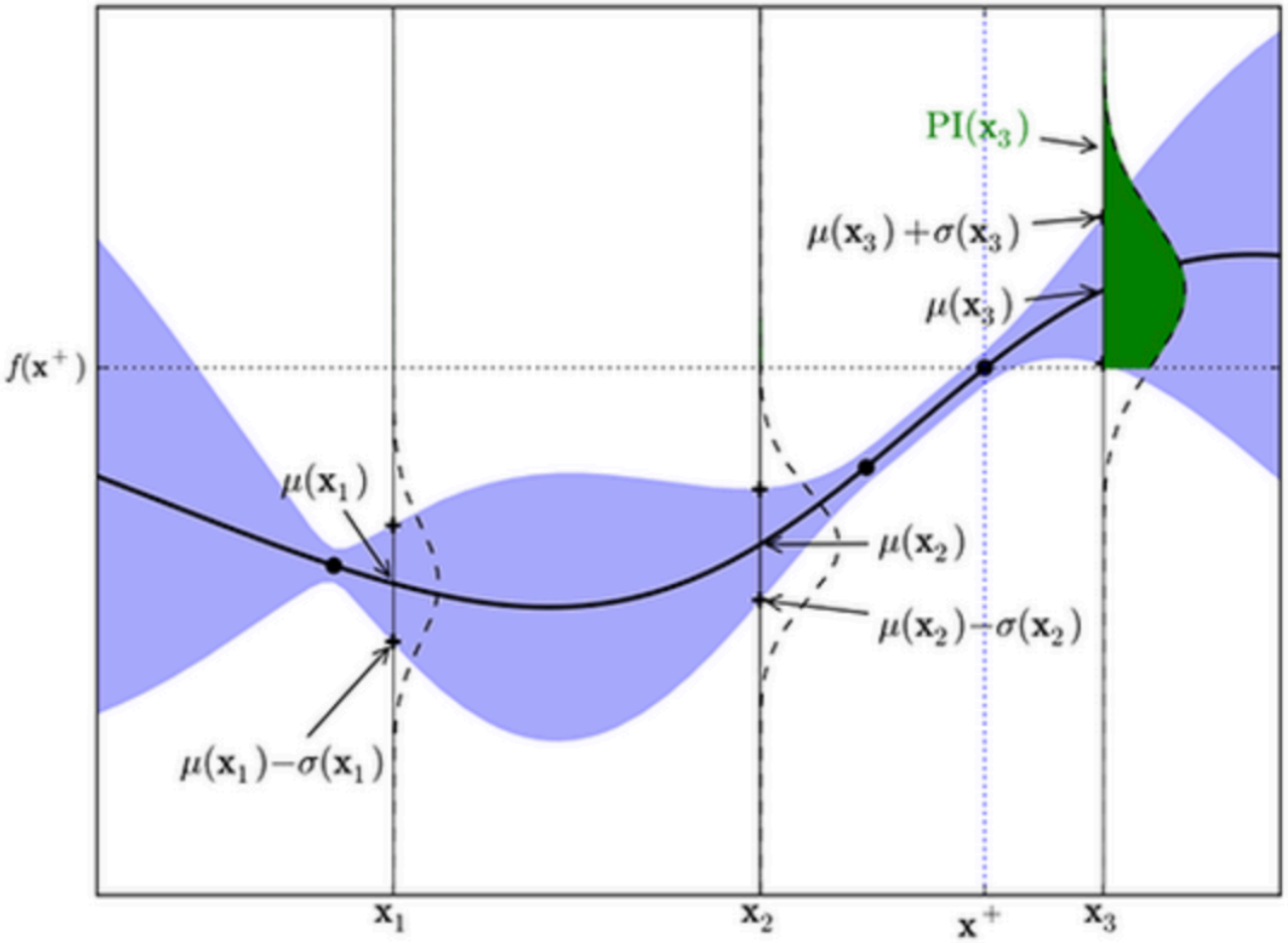
N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

- We have some measurements in space X, and we want to maximize some property f(X).
- How can we decide what point to measure next to best maximize f?
- We need to balance the exploration of the space with exploitation of regions near we have already know

OAK RIDGE National Laboratory

Acquisition Functions

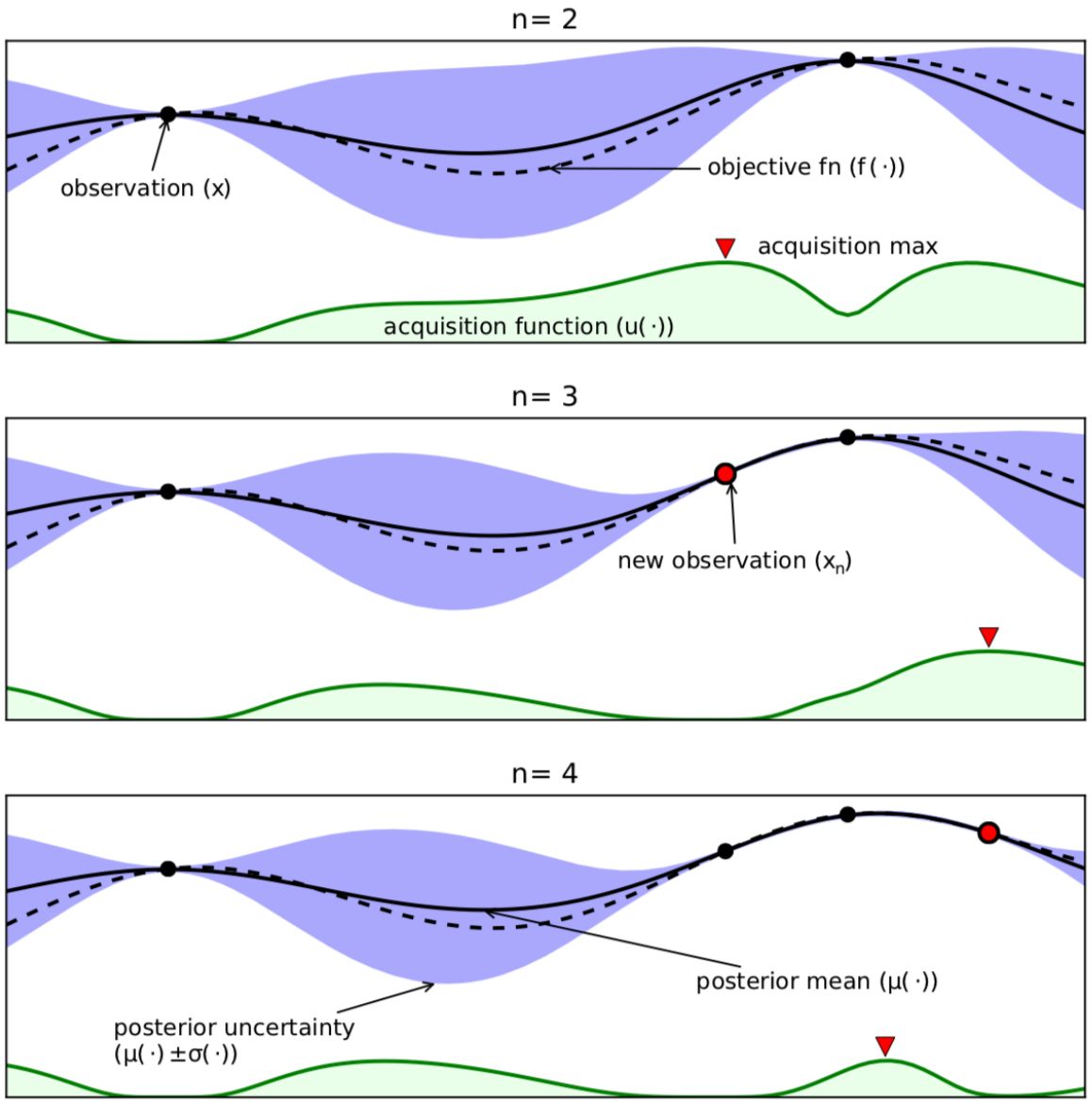
Probability of Improvement Acquisition Function



1. **Confidence bound:** simplest possible - just take the upper confidence bound from the prediction
2. **Probability of Improvement:** Integral from current functional maximum to upper limit of distribution as test point
3. **Expected Improvement:** Instead of probability of improvement, we want to maximize the expected increase in the function value
4. **There are (always) more…**

OAK RIDGE National Laboratory

"Active Learning"



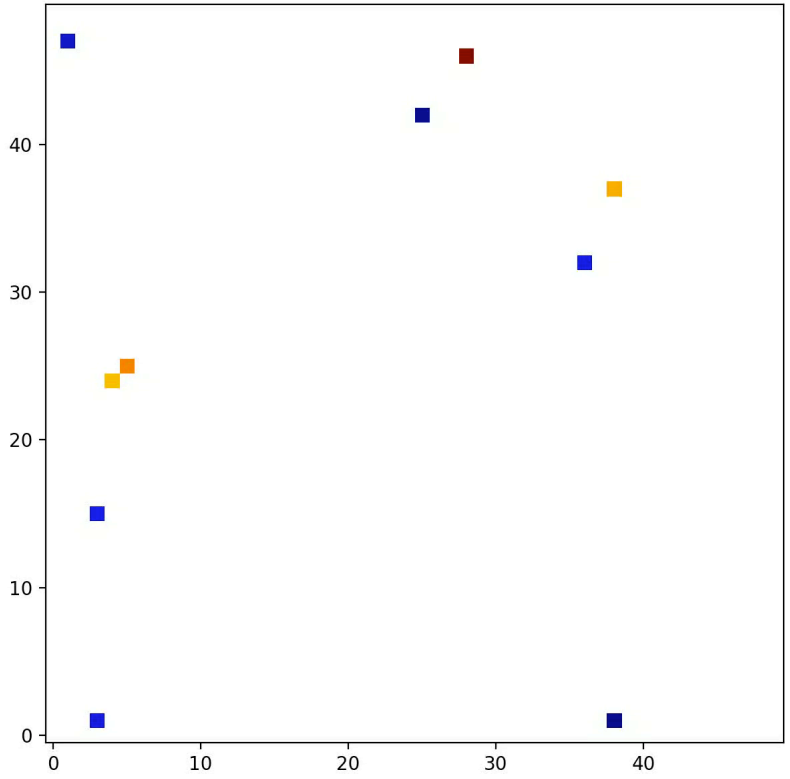
- Shown here are three iterations of Bayesian optimization
- We can do this for dictating how to collect new observations in our experiment

N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

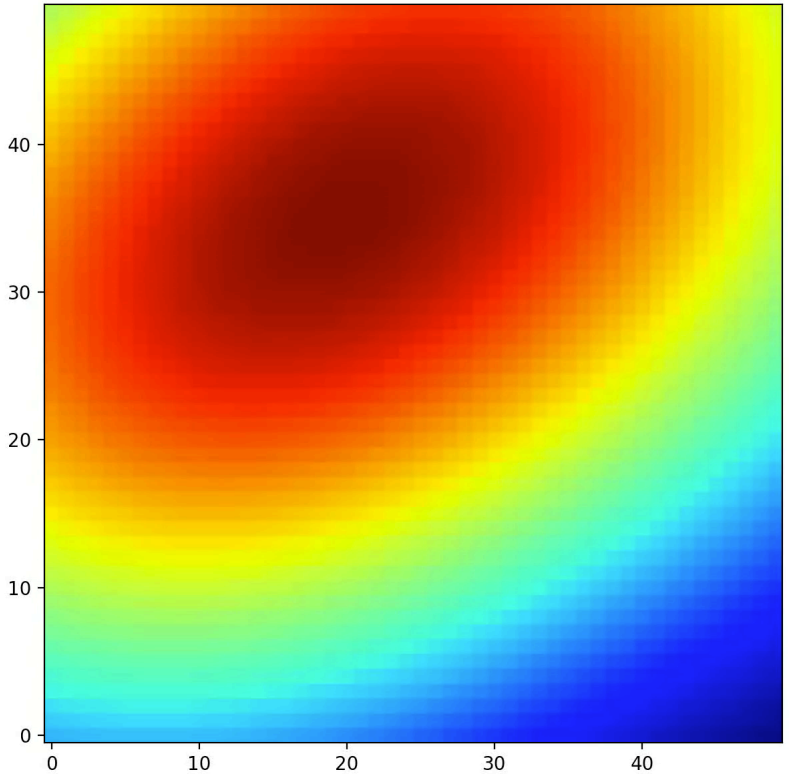
Example: Hysteresis Loops Area Maximization

Batch size = 10. 2.35s/pixel, and each iteration took ~6s on dgx. 410 points in total (40 GP iterations). 1200s in total. Total time for all pixels: 5875s (20% the time)

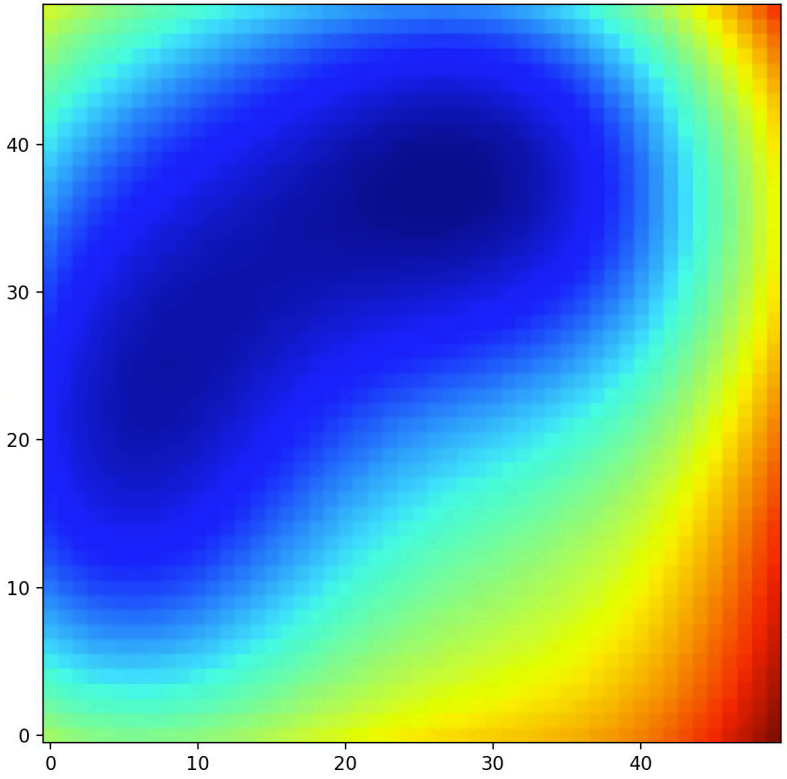
Measured Loop Areas



GP Prediction



GP Uncertainty



OAK RIDGE National Laboratory