THE UNIVERSITY *of* TENNESSEE KNOXVILLE

Lecture 20: Gaussian Processes and Bayesian Optimization

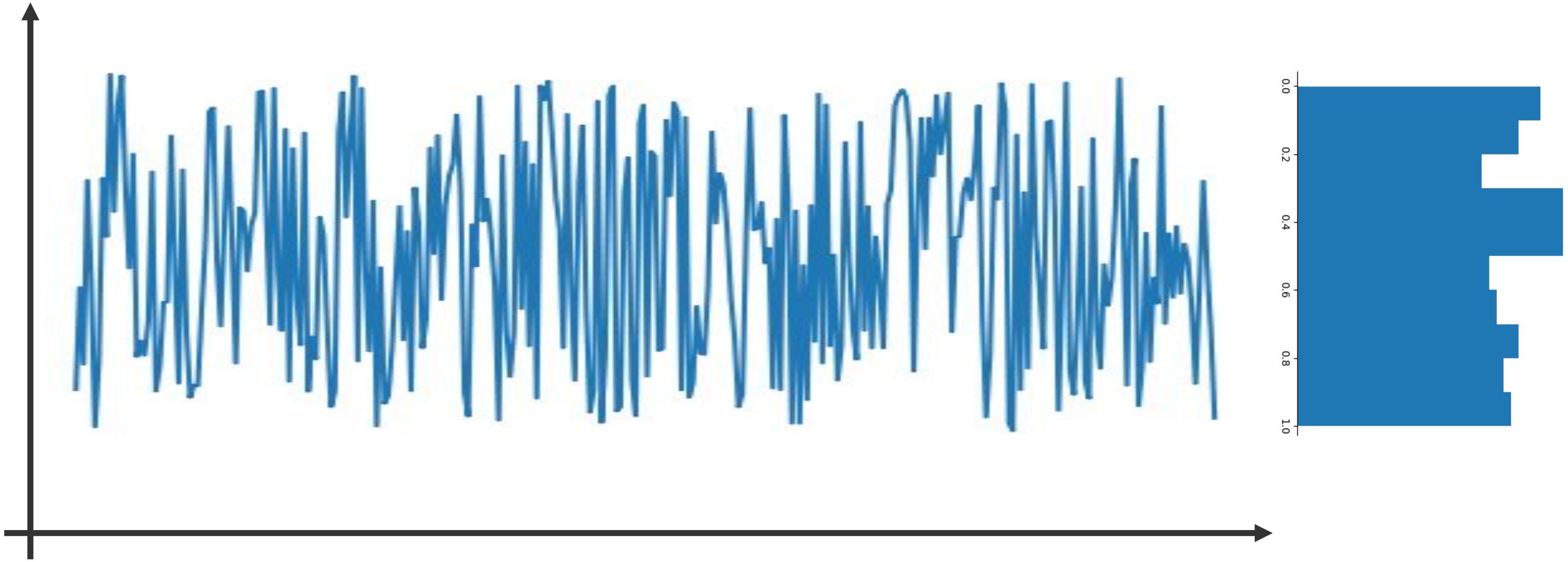
Instructor: Sergei V. Kalinin

shutterstock.com · 397158622

“Well! I've often seen a cat without a grin,' thought Alice *'but a grin without a cat! It's the most curious thing i ever saw in my life!”*

– Lewis Carroll, **Alice in Wonderland**

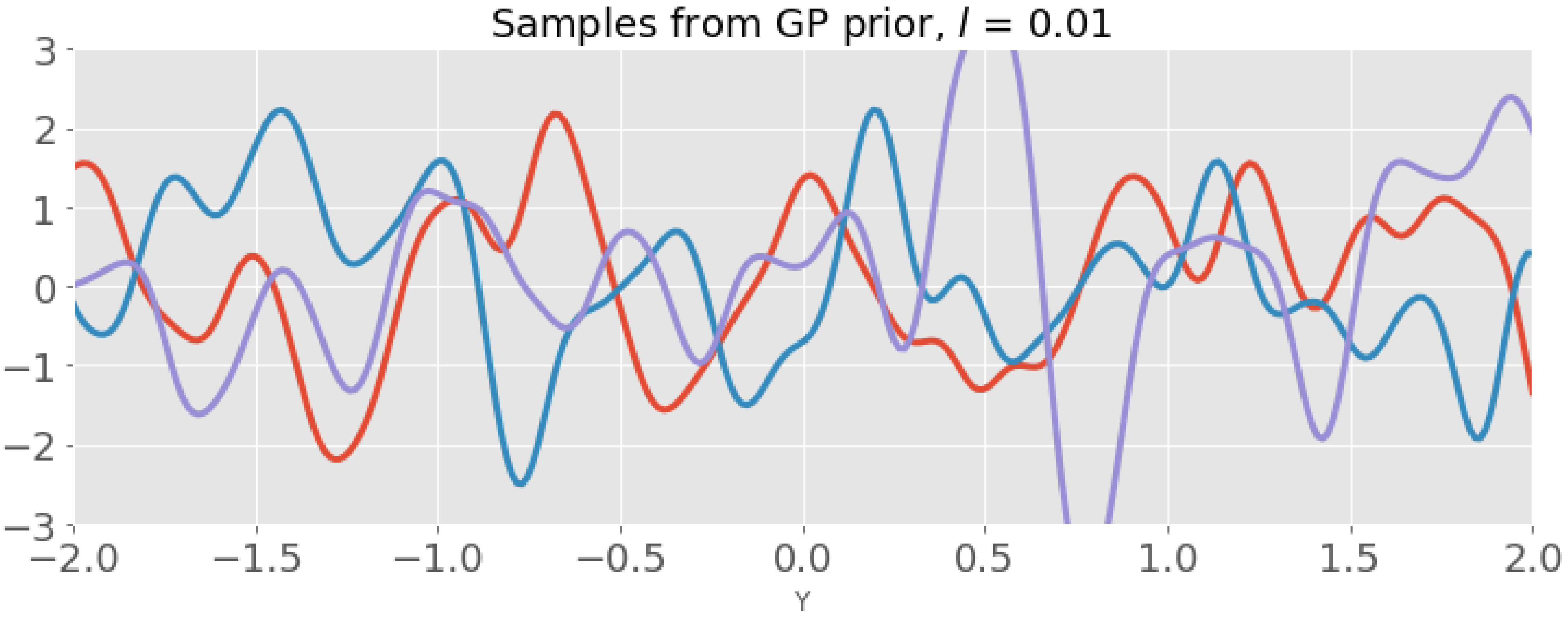
What do we know if we do not know anything?



Gaussian Process Regression

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

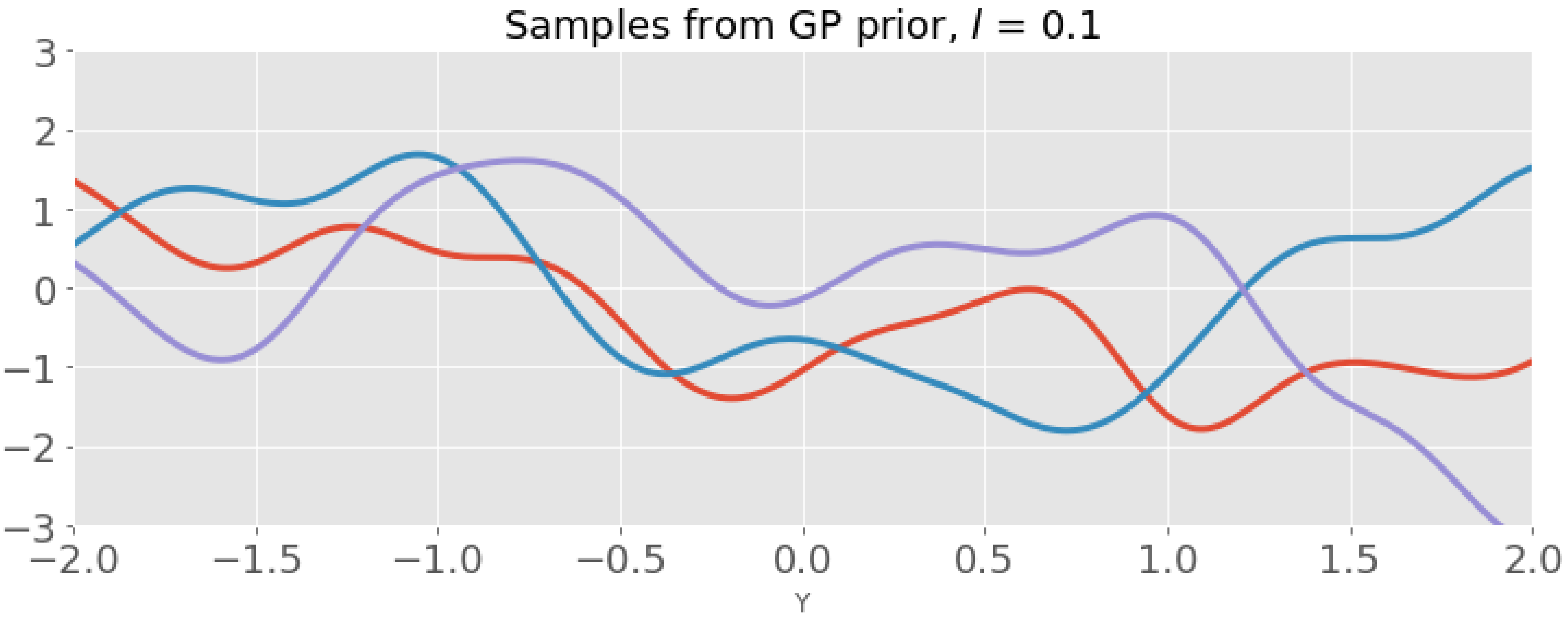
$$\mathrm{k}(\mathrm{x},\mathrm{x}') = \exp\left(-\frac{1}{2l}(x-x')^2\right)$$



Gaussian Process Regression

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

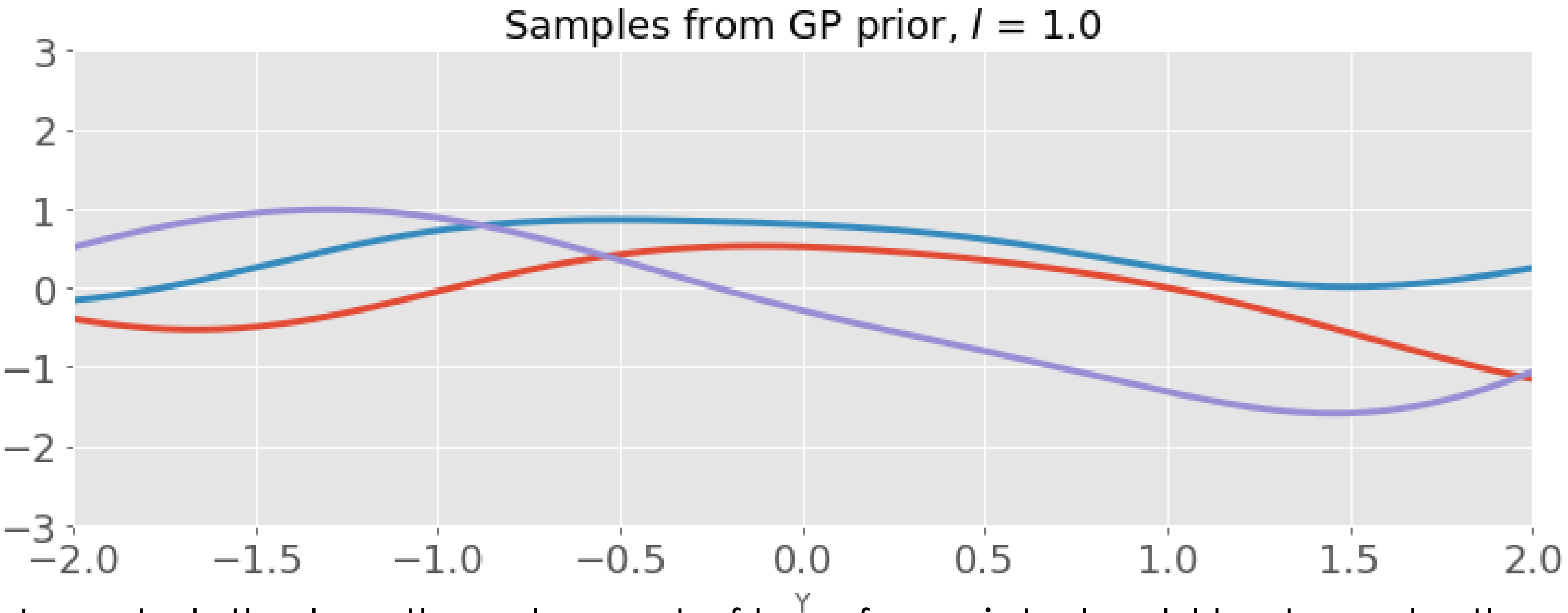
$$\mathrm{k}(\mathrm{x},\mathrm{x}') = \exp\left(-\frac{1}{2l}(x-x')^2\right)$$



Gaussian Process Regression

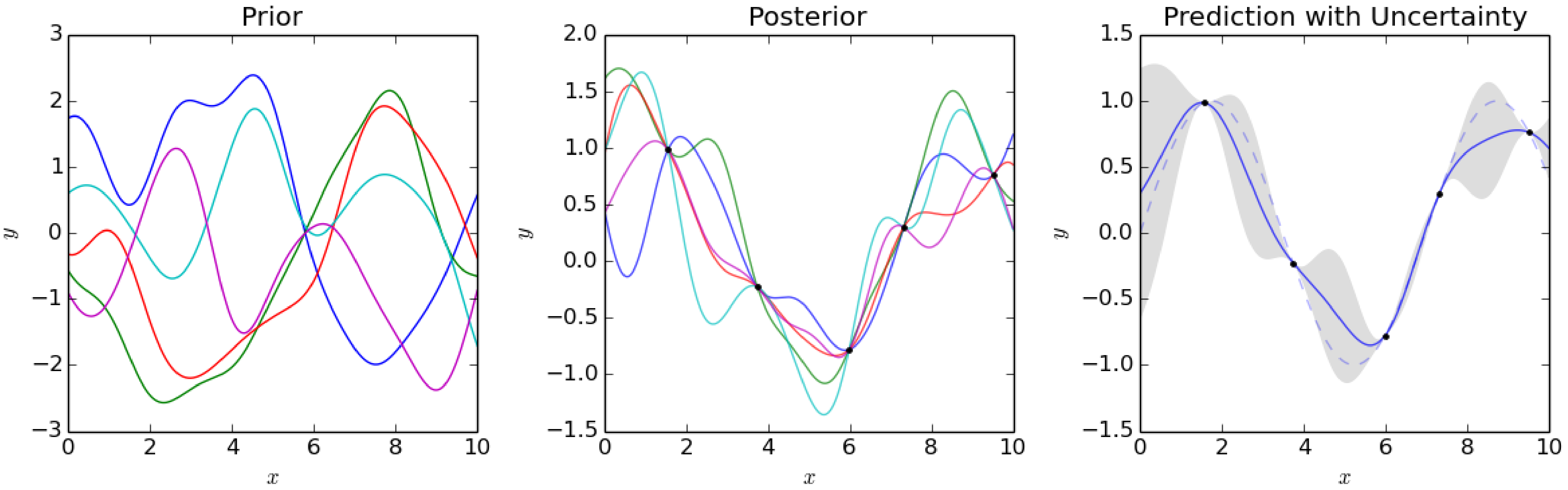
- Covariance matrix (kernel) determines what type of functions we will allow.

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



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

Gaussian Process Regression



Prior: What can the function be before the measurement
Data: Measurement
Posterior: What can the function be after measurement

Policy: How do we balance exploration and exploitation (acquisition function)

GP Vocabulary

- Gaussian Process
- Kernel and kernel parameters
- Kernel Priors
- Noise Priors
- Posteriors

Kernels in Molecular Spaces

Graph Kernels: Since chemical molecules can be represented as graphs with atoms as nodes and bonds as edges, graph kernels are a natural choice. Graph kernels measure the similarity between graphs and can be used in GPs to model molecular properties. Examples include the Random Walk Kernel, Weisfeiler-Lehman (WL) Kernel, and Graphlet Kernel.

Molecular Fingerprint Kernels: Molecular fingerprints are binary vectors representing the presence or absence of certain substructures within a molecule. Kernels can be defined on these fingerprints, such as the Tanimoto Kernel (also known as the Jaccard Index), which measures the similarity between two fingerprint vectors. This approach is particularly popular in cheminformatics for its interpretability and efficiency.

Spectral Kernels: Spectral methods can be used to define kernels on the spectrum (eigenvalues) of the Laplacian matrix of molecular graphs. These kernels can capture the topological and spectral properties of molecules, offering a powerful way to model molecular similarities.

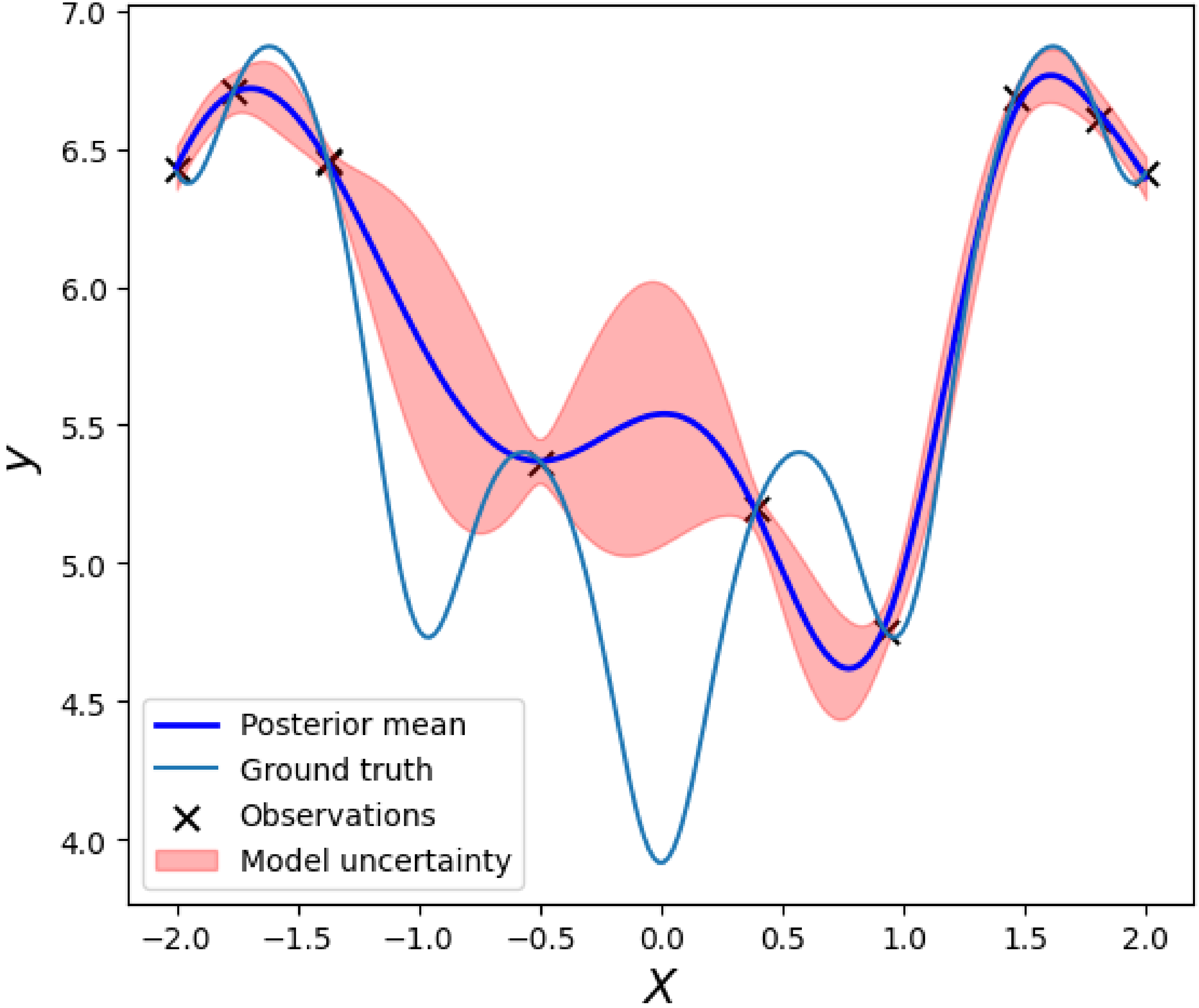
Kernels in Molecular Spaces

Gaussian Molecular Kernels: This type of kernel measures the similarity between molecules based on Gaussian-shaped functions centered on atoms. The overlap between these Gaussian functions for different molecules can be used to compute a kernel value, incorporating both structural and spatial information.

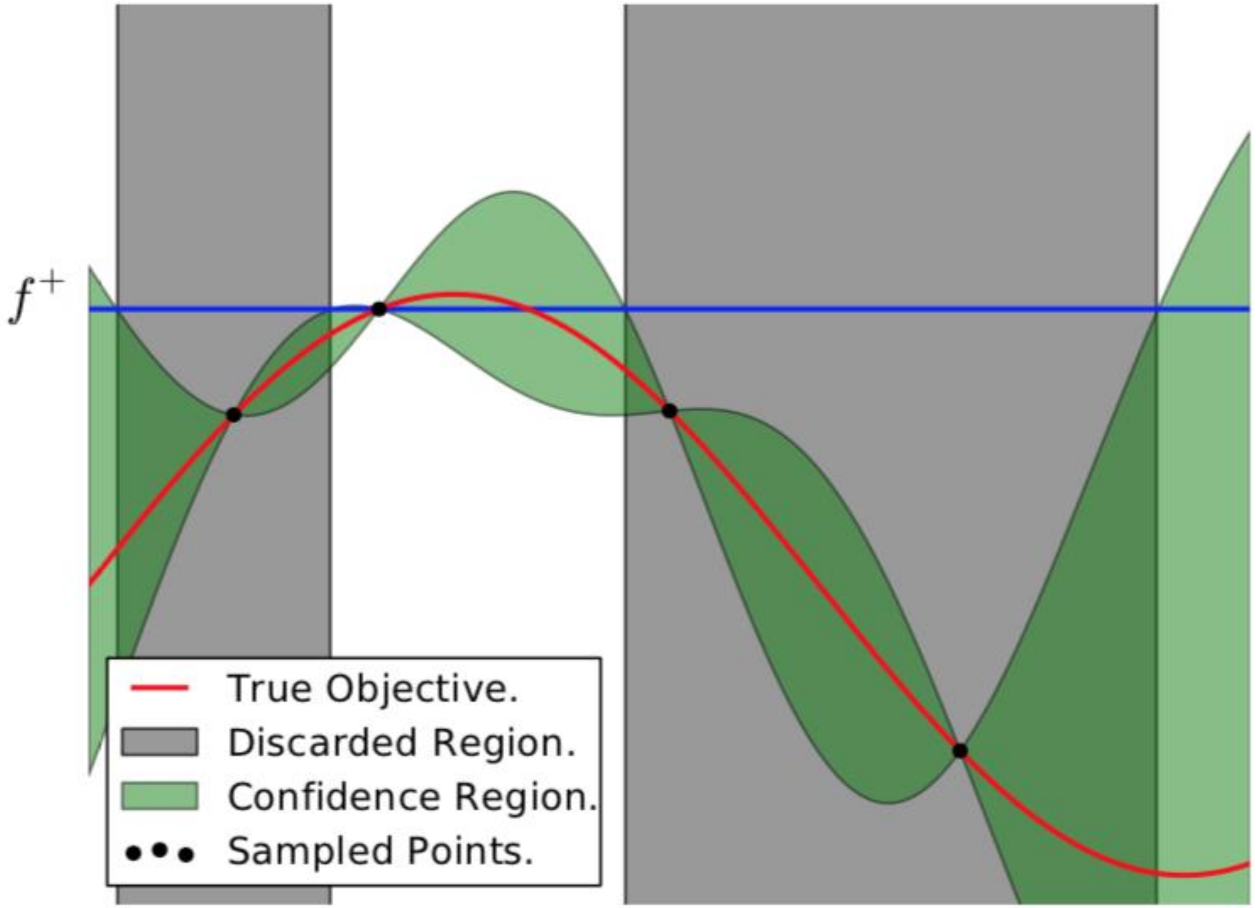
Physicochemical Property Kernels: Kernels can be designed based on the physicochemical properties of molecules, such as hydrophobicity, charge distribution, or molecular weight. By defining a similarity measure based on the distance between these property vectors of different molecules, one can construct a kernel that reflects the similarities in physical and chemical properties.

Deep Learning-Based Kernels: With the rise of deep learning, kernels can also be derived from the embeddings or latent spaces generated by neural networks trained on molecular data. For instance, a neural network can be trained to generate molecular embeddings, and a kernel can be defined as the dot product or cosine similarity between these embeddings.

Colab


7.0
6.5
6.0
5.5
5.0
4.5
4.0
y
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X
Posterior mean
Ground truth
Observations
Model uncertainty

Bayesian Optimization

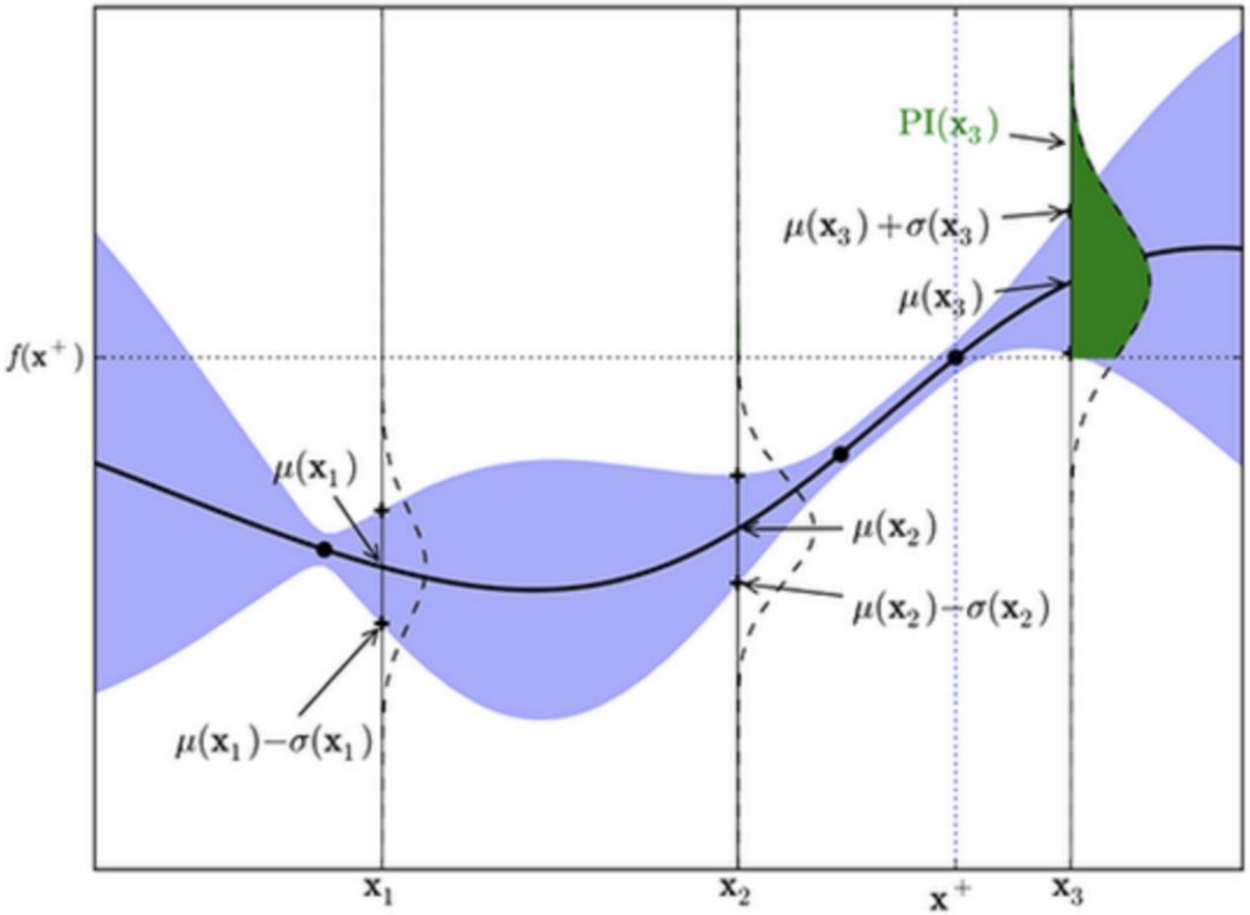


- We have some measurements in space X, and we want to maximize some property f(X).
- How can we decide what point to measure next to best maximize f?
- We need to balance the exploration of the space with exploitation of regions near we have already know

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

Acquisition Functions

Probability of Improvement
Acquisition Function



1. **Upper 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…**

Expected improvement is defined as

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E}\max(f(\mathbf{x}) - f(\mathbf{x}^+), 0) \qquad (1)$$

where $f(\mathbf{x}^+)$ is the value of the best sample so far and $\mathbf{x}^+$ is the location of that sample i.e. $\mathbf{x}^+ = \mathrm{argmax}_{\mathbf{x}_i \in \mathbf{x}_{1:t}} f(\mathbf{x}_i)$. The expected improvement can be evaluated analytically under the GP model[3]:

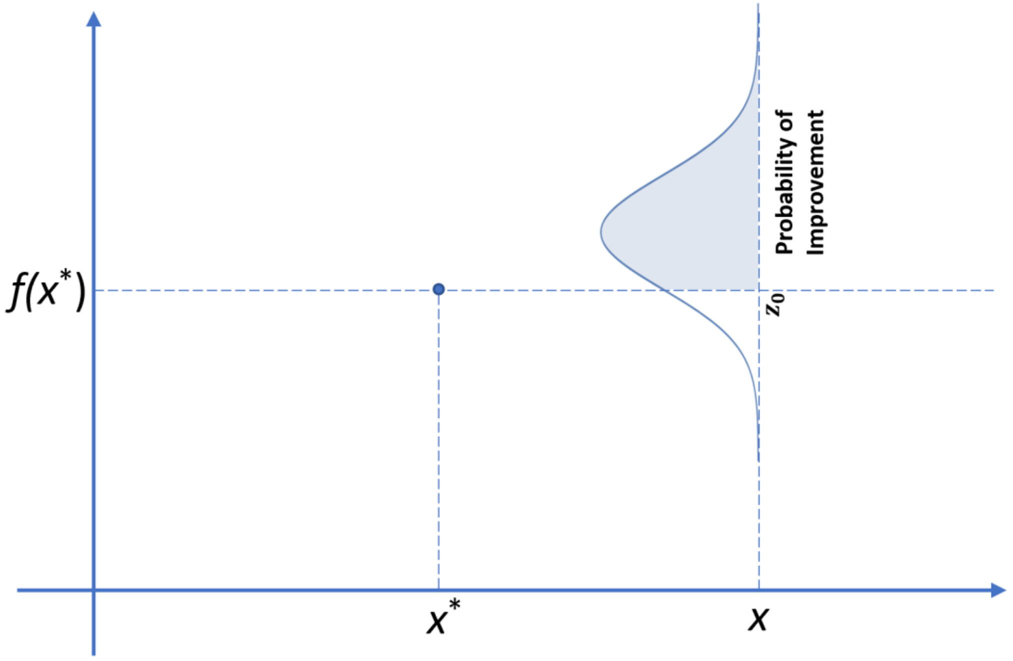
$$\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \qquad (2)$$

where

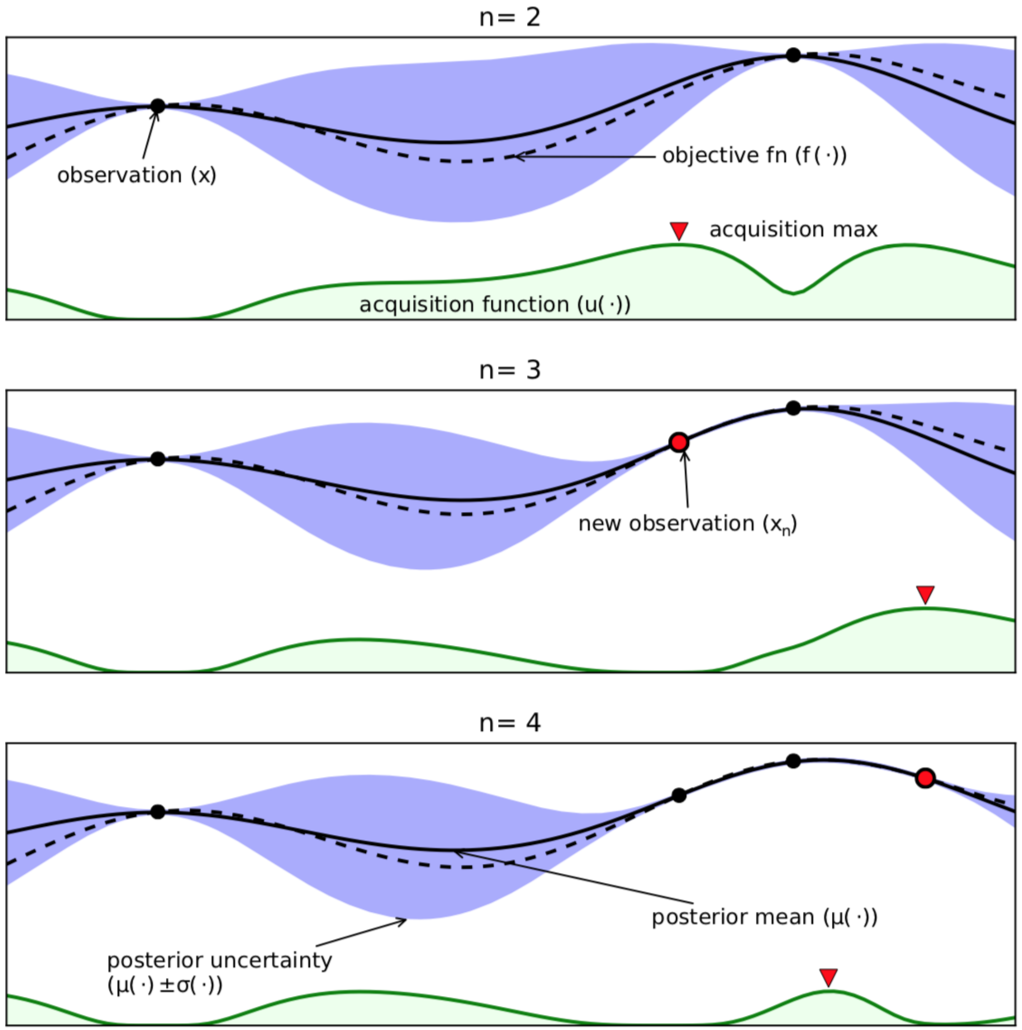
$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the mean and the standard deviation of the GP posterior predictive at $\mathbf{x}$, respectively. Φ and ϕ are the CDF and PDF of the standard normal distribution, respectively. The first summation term in Equation (2) is the exploitation term and second summation term is the exploration term.

Parameter ξ in Equation (2) determines the amount of exploration during optimization and higher ξ values lead to more exploration. In other words, with increasing ξ values, the importance of improvements predicted by the GP posterior mean $\mu(\mathbf{x})$ decreases relative to the importance of potential improvements in regions of high prediction uncertainty, represented by large $\sigma(\mathbf{x})$ values. A recommended default value for ξ is 0.01.

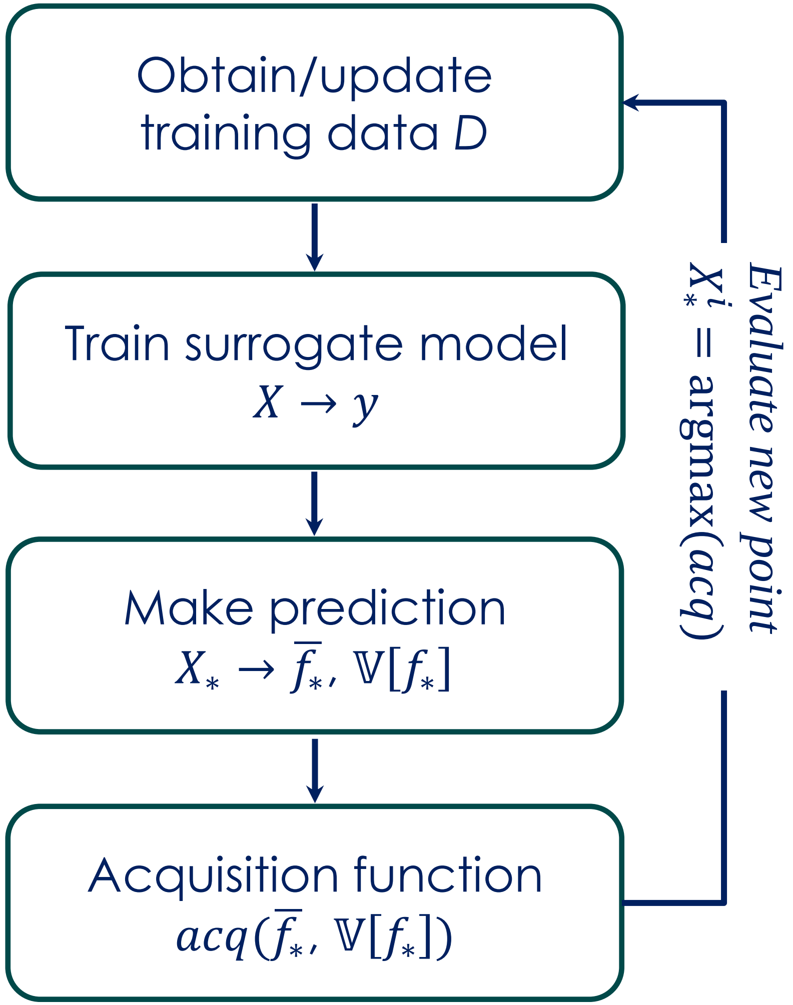


The basics: Bayesian Optimization



X, y: (sparse) Training data

X_*: New (not yet evaluated) points



Obtain/update training data *D*

Train surrogate model $X \rightarrow y$

Make prediction $X_* \rightarrow \bar{f}_*, \mathbb{V}[f_*]$

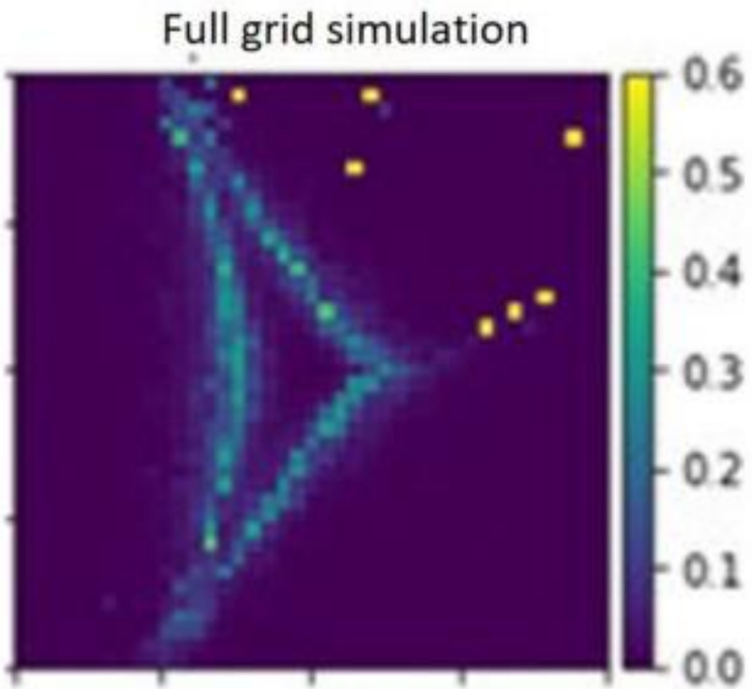
Acquisition function $acq(\bar{f}_*, \mathbb{V}[f_*])$

Evaluate new point $X_*^i = \mathrm{argmax}(acq)$

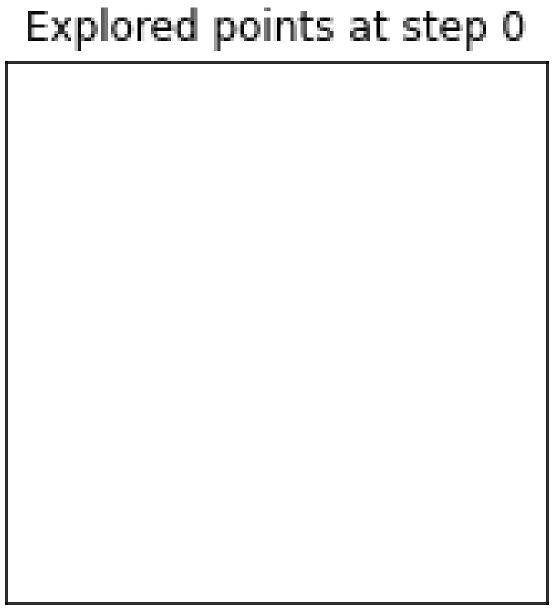
N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

Bayesian Optimization for physical discovery

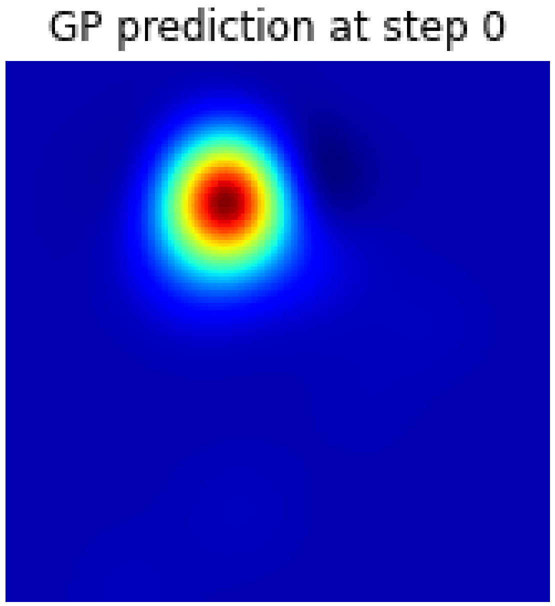
Discovering regions where heat capacity is maximized in NNN Ising model



Full grid simulation



Explored points at step 0



GP prediction at step 0