Imperial College
Business School

Imperial means
Intelligent Business

Bayesian Optimisation

Office Hours with Yu Qian Ang

Date: Jan 4, 2023

- What is Bayesian Optimisation?
- Intuition behind Bayesian Optimisation
- Steps Involved In Bayesian Optimisation
- Questions

Bayesian Optimisation

Bayesian Optimisation is a sequential iterative process that provides a method of discovering parameters that optimise the output(s) of a black-box function or model.

Expensive to Evaluate Functions

Bayesian Optimisation is commonly utilised to solve expensive-to-evaluate functions; for functions that fall under this category discovering the optimal input that provides the desired output will be costly in terms of time, expense and compute resources, if a naive approach is taken. A naive approach could be randomly guessing the input variables.

Bayesian Optimization

You have a function: *f(x)*

- It's expensive to evaluate
- You don't know its derivative
- It's not exactly an analytic or closed form expression

Easy? Difficult?

More difficult than other optimization problems within machine learning

In **Gradient Descent**, for example, you have access to the function's derivative

Adapted from Andre Ye

Bayesian Optimization

Limited in optimization by:

1. **Expensive to calculate:** (Ideally we can query the function enough to replicate, but not possible so our optimization method must work with limited sampling of inputs)

2. **Derivative unknown**: Knowing derivative give the optimizer a sense of direction - we may not have this.

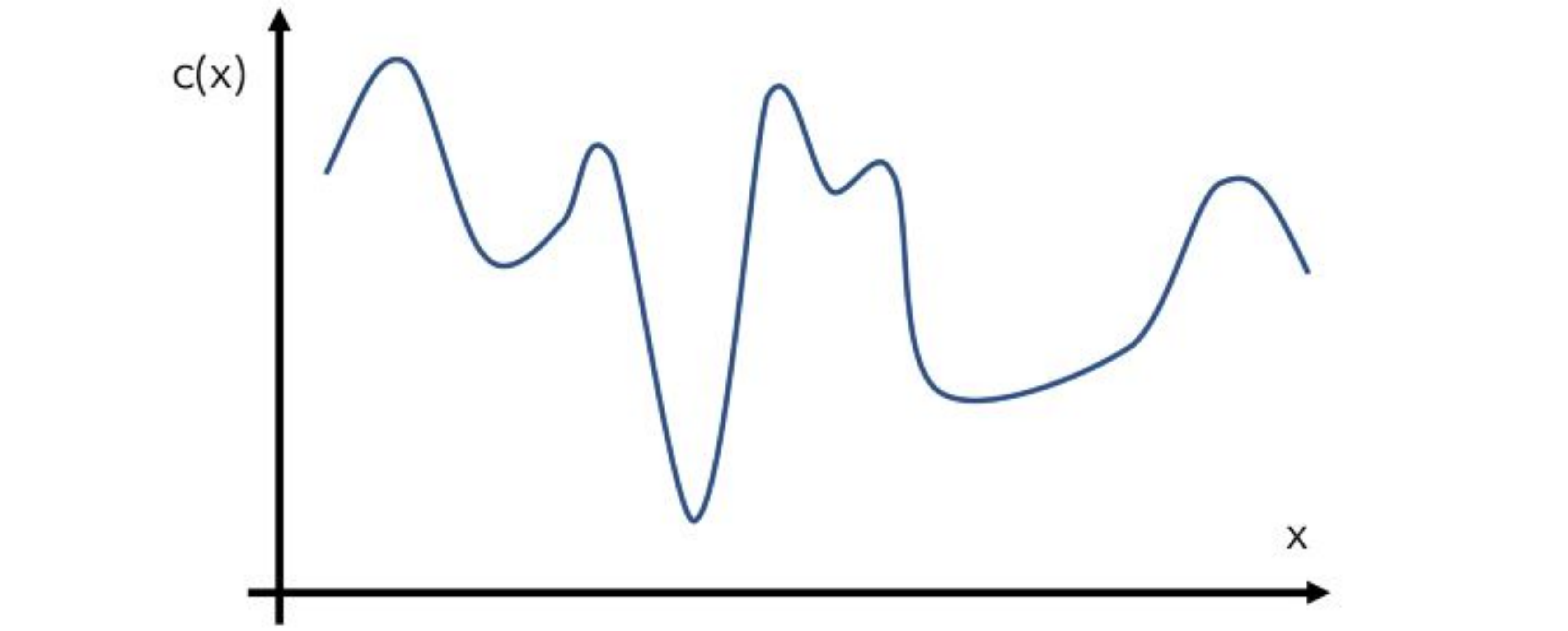
3. **Need to find global minima**: Difficult task even with gradient descent.

Adapted from Andre Ye

Bayesian Optimization

Let's look at a hypothetical function *h(x)*

True shape of objective function hidden from optimizer

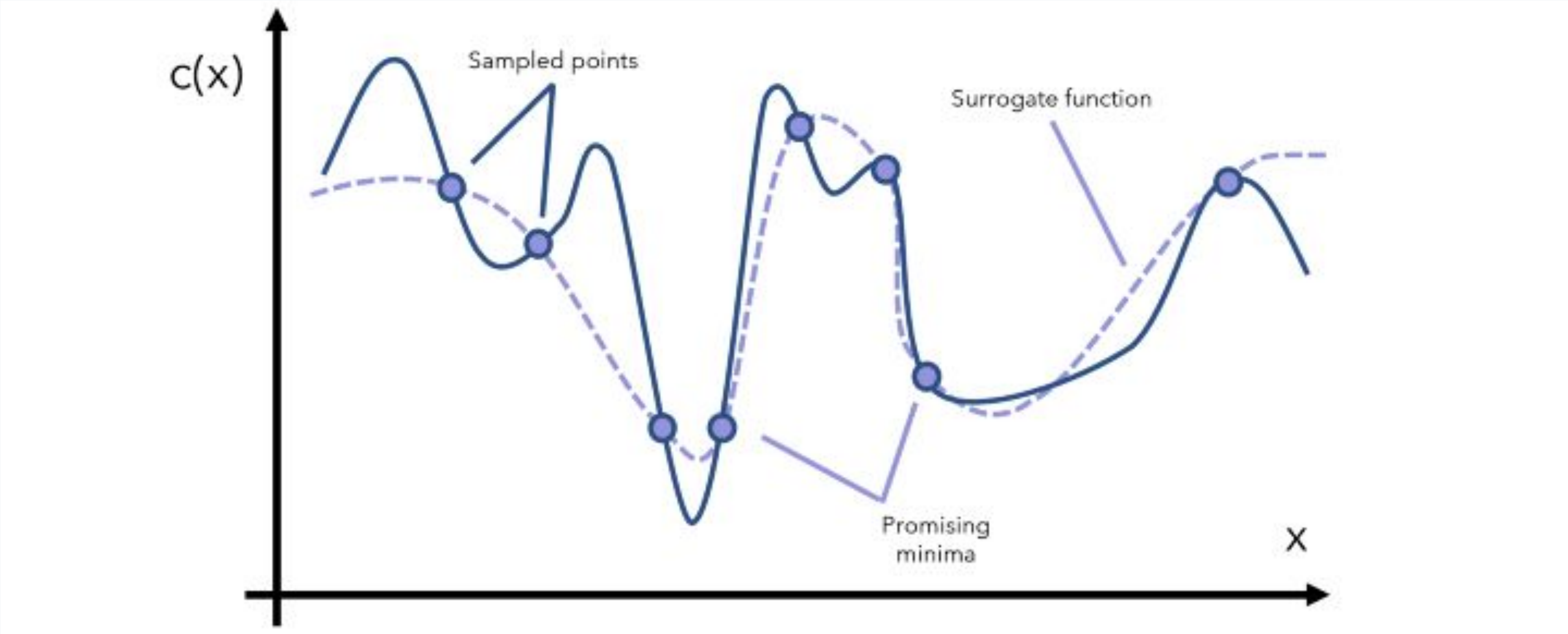


Adapted from Andre Ye

Bayesian Optimization

Bayesian optimization approaches this task through a method known as surrogate optimization

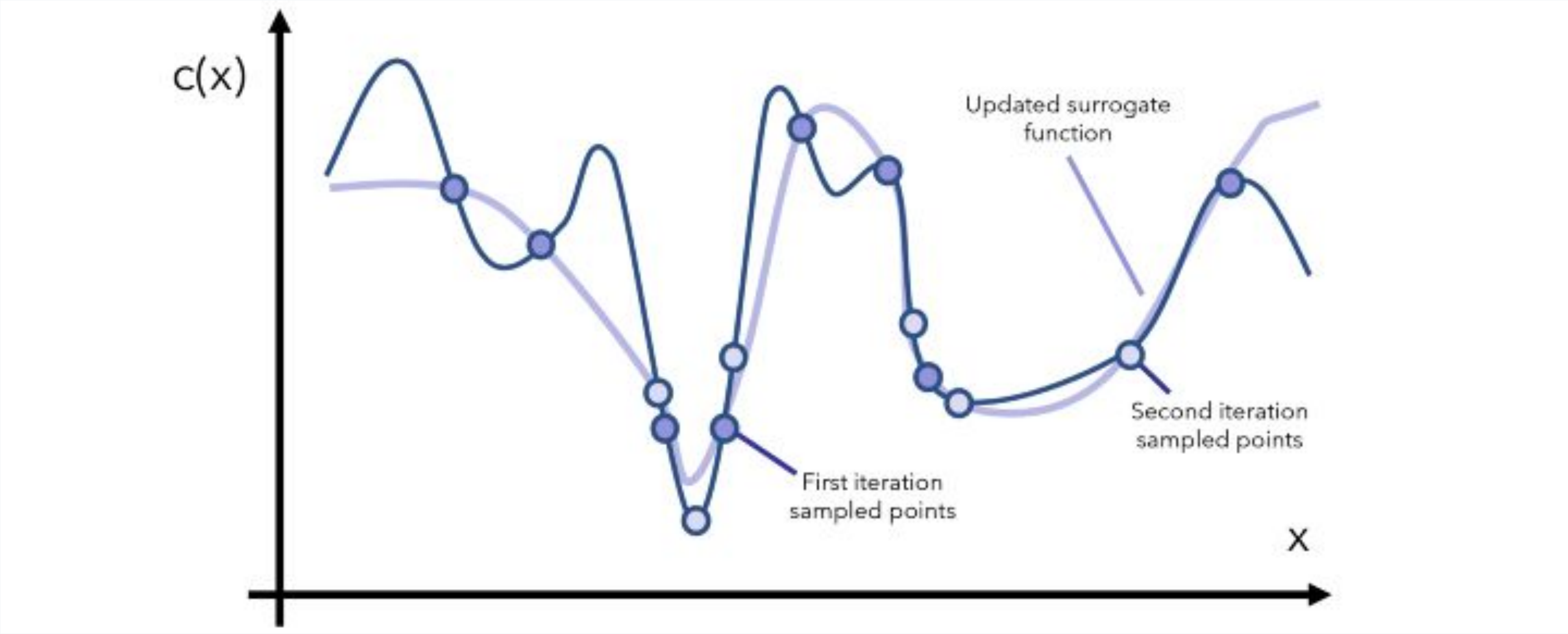
Surrogate function is an approximation of the objective function



Adapted from Andre Ye

Bayesian Optimization

At each iteration, we continue to **look at the current surrogate function**, **learn more about the areas of interest by sampling**, and **update the function**



Adapted from Andre Ye

Bayesian Optimization

Take a moment to think about the beauty of this approach

1. Doesn't make any assumptions about the function
2. Doesn't require information on derivatives
3. Able to use common-sense reasoning and data through continuous updating

Makes evaluating of originally expensive objective function feasible

Adapted from Andre Ye

Bayesian Optimization

So why is it Bayesian

Updating of prior belief in light of new information

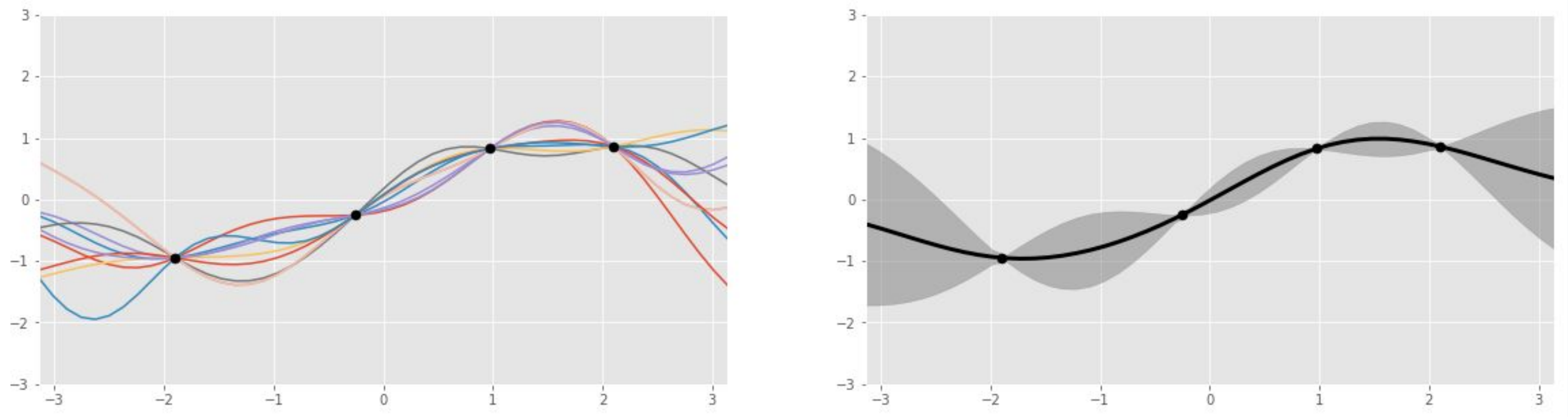
Product updated posterior belief

Surrogate function usually represented by Gaussian Processes

Adapted from Andre Ye

Bayesian Optimization

Surrogate function usually represented by Gaussian Processes



Adapted from Andre Ye

Bayesian Optimization

The surrogate function is updated with an ‘acquisition function’, responsible for driving the proposition of new points to test, in a exploration and exploitable trade-off.

Exploitation: seeks to sample where surrogate model predicts a good objective

Exploration: seeks to sample in locations where uncertainty is high

Must consider both exploitation and exploration. Common acquisition functions include expected improvement and maximum probability of improvement, all of which measures the probability that a specific input may pay off in the future, given information about the prior (GP)

Adapted from Andre Ye

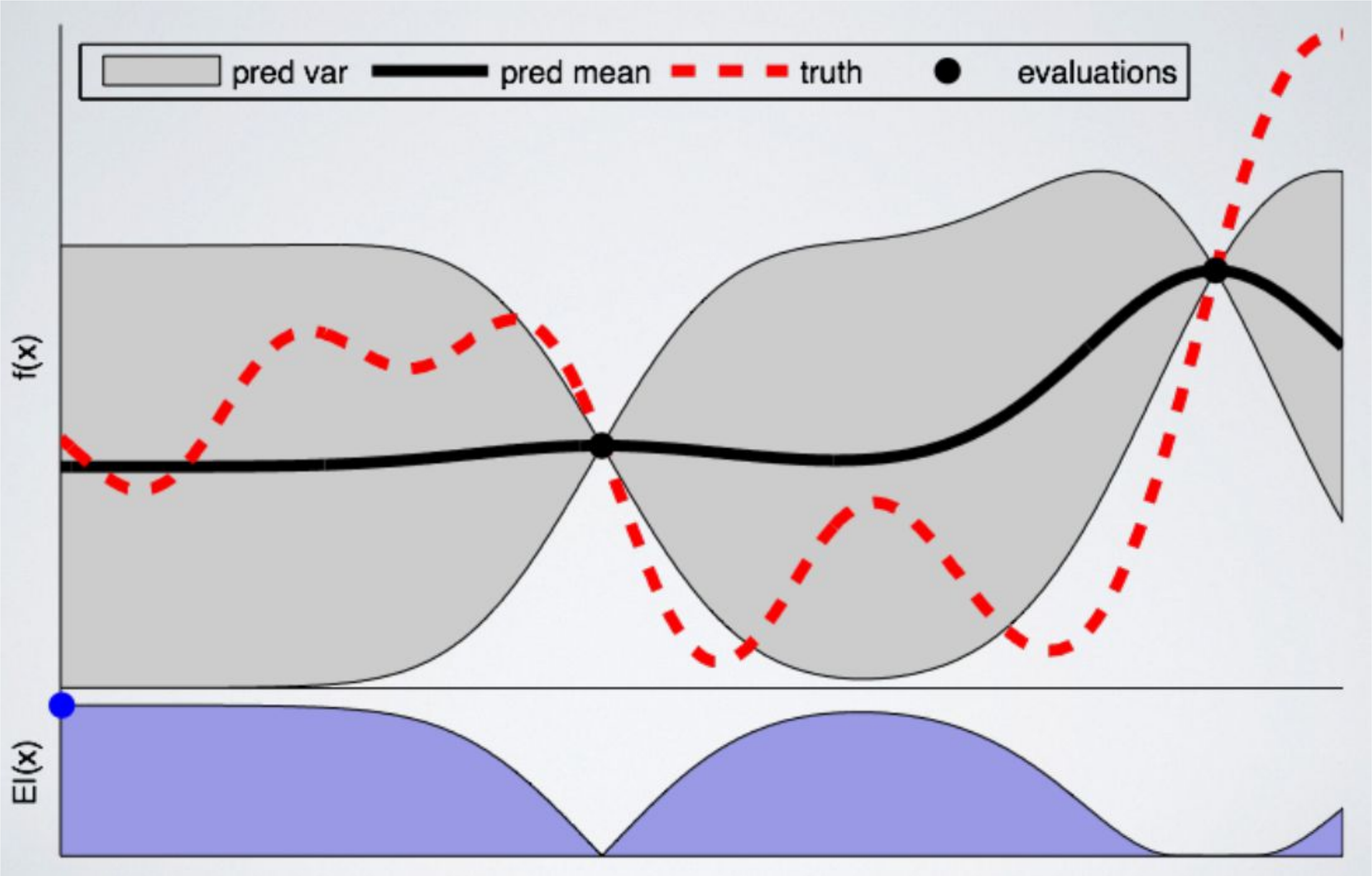
Bayesian Optimization

Let's put everything together!

1. Initialize a GP 'surrogate function' prior distribution
2. Choose a few data points *x* such as the acquisition function *a(x)* operating on the current prior distribution is maximized
3. Evaluate date points *x* in the objective cost function *c(x)* and obtain the results, *y*
4. Update the GP prior distribution with new data to produce a posterior (which will become the prior in the next step)
5. Repeat steps 2 to 5 for several iterations
6. Interpret the GP distribution to find global minima

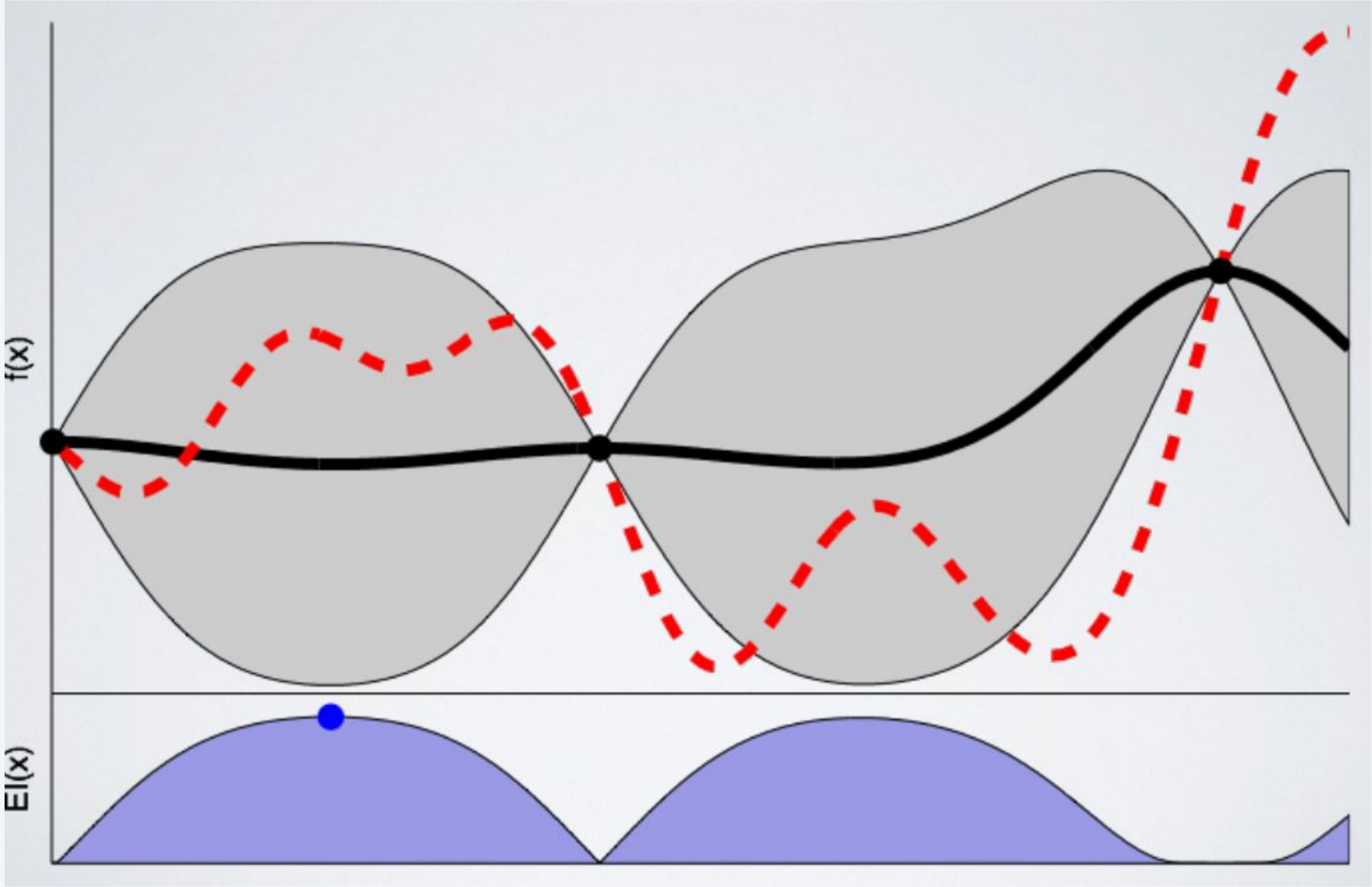
Adapted from Andre Ye

Illustrating Bayesian Optimization



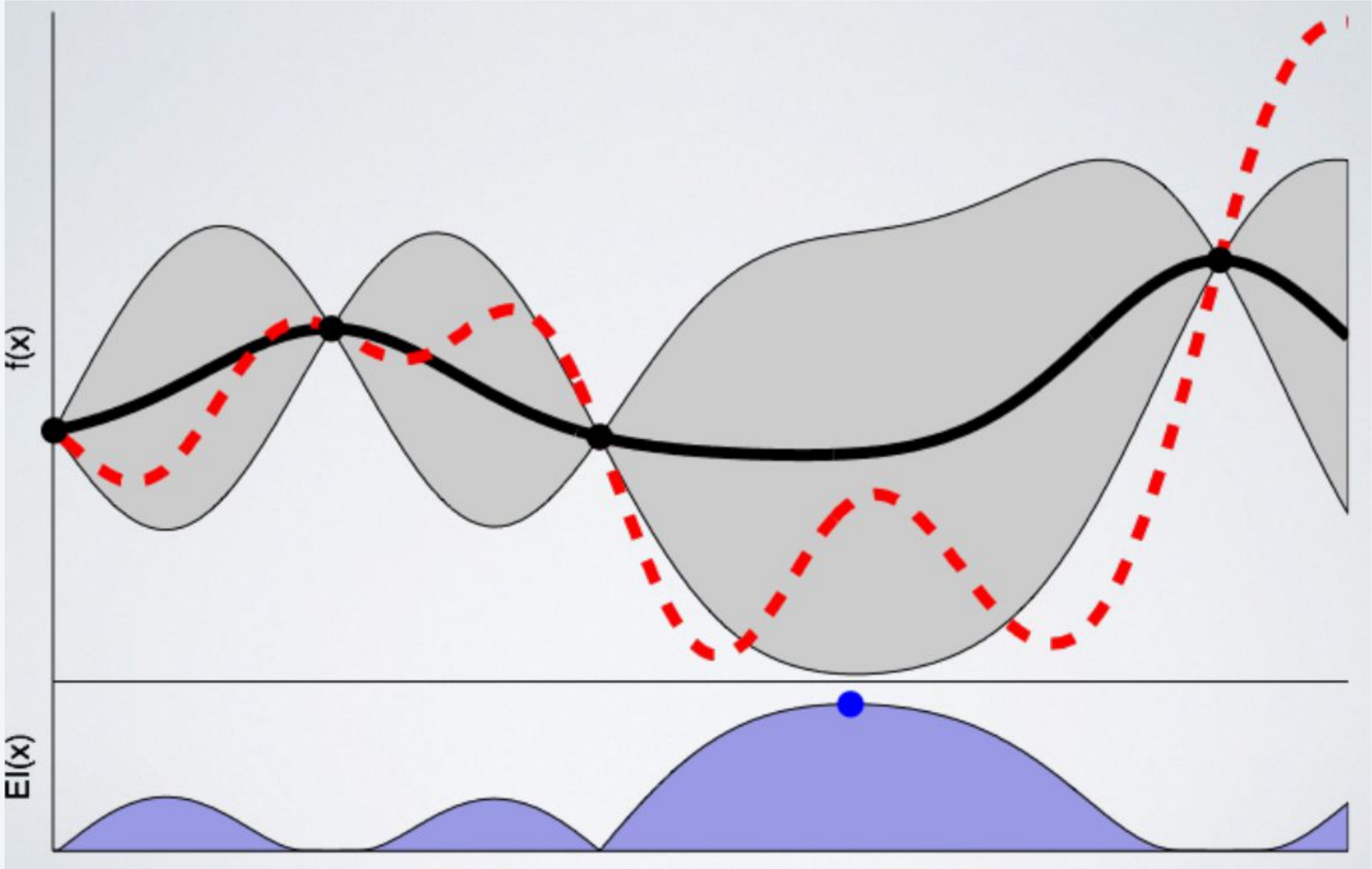
Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



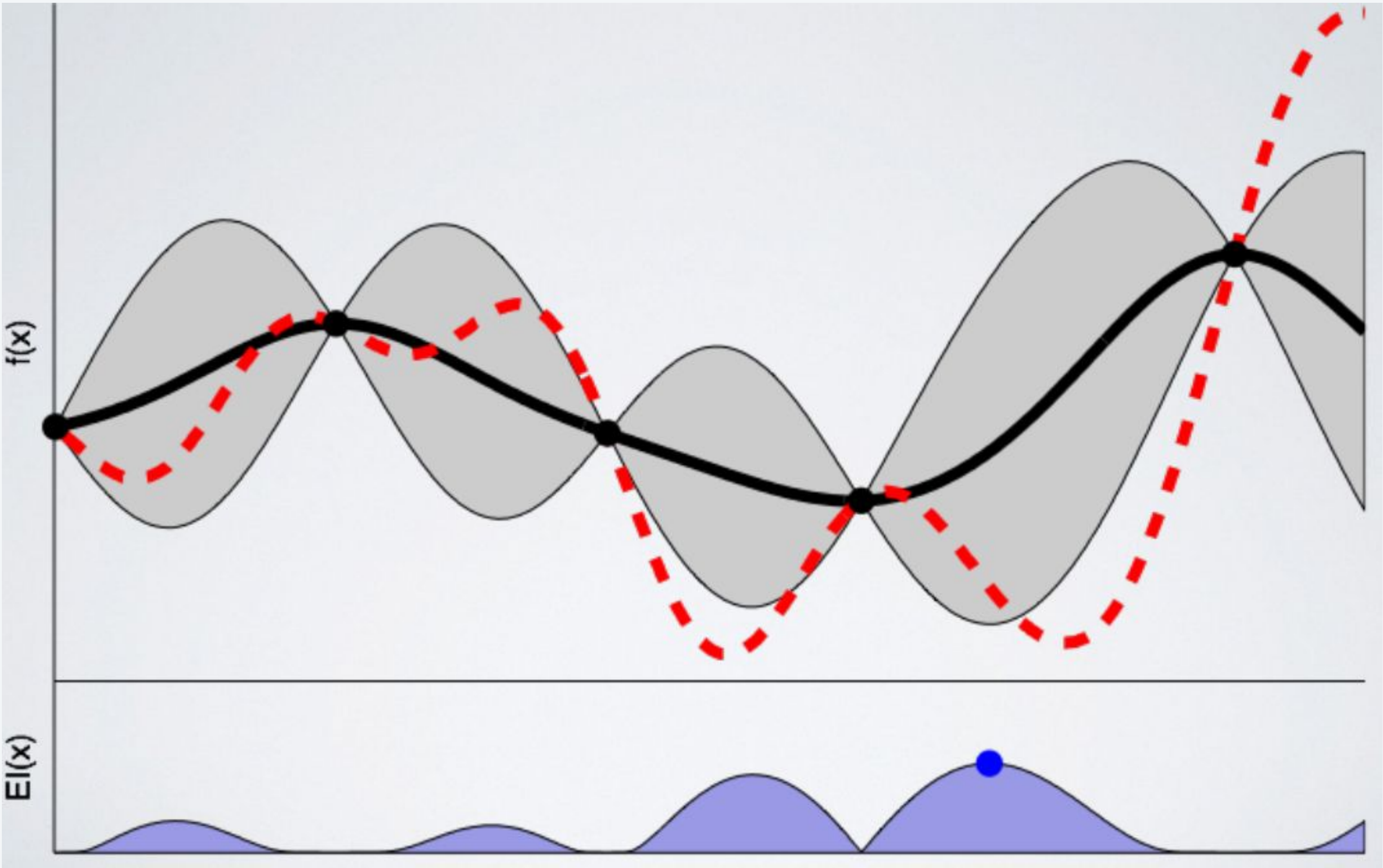
Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



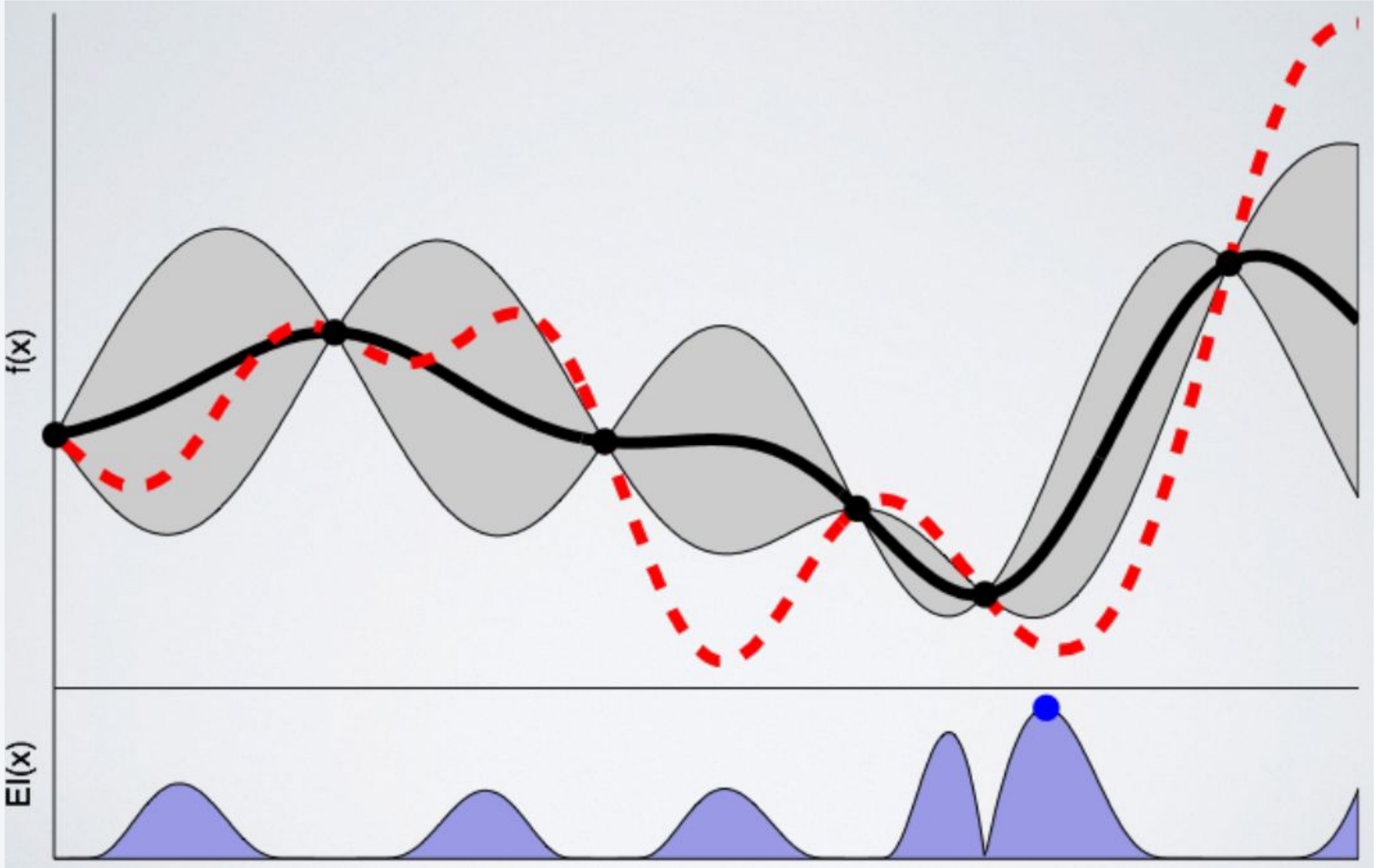
Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



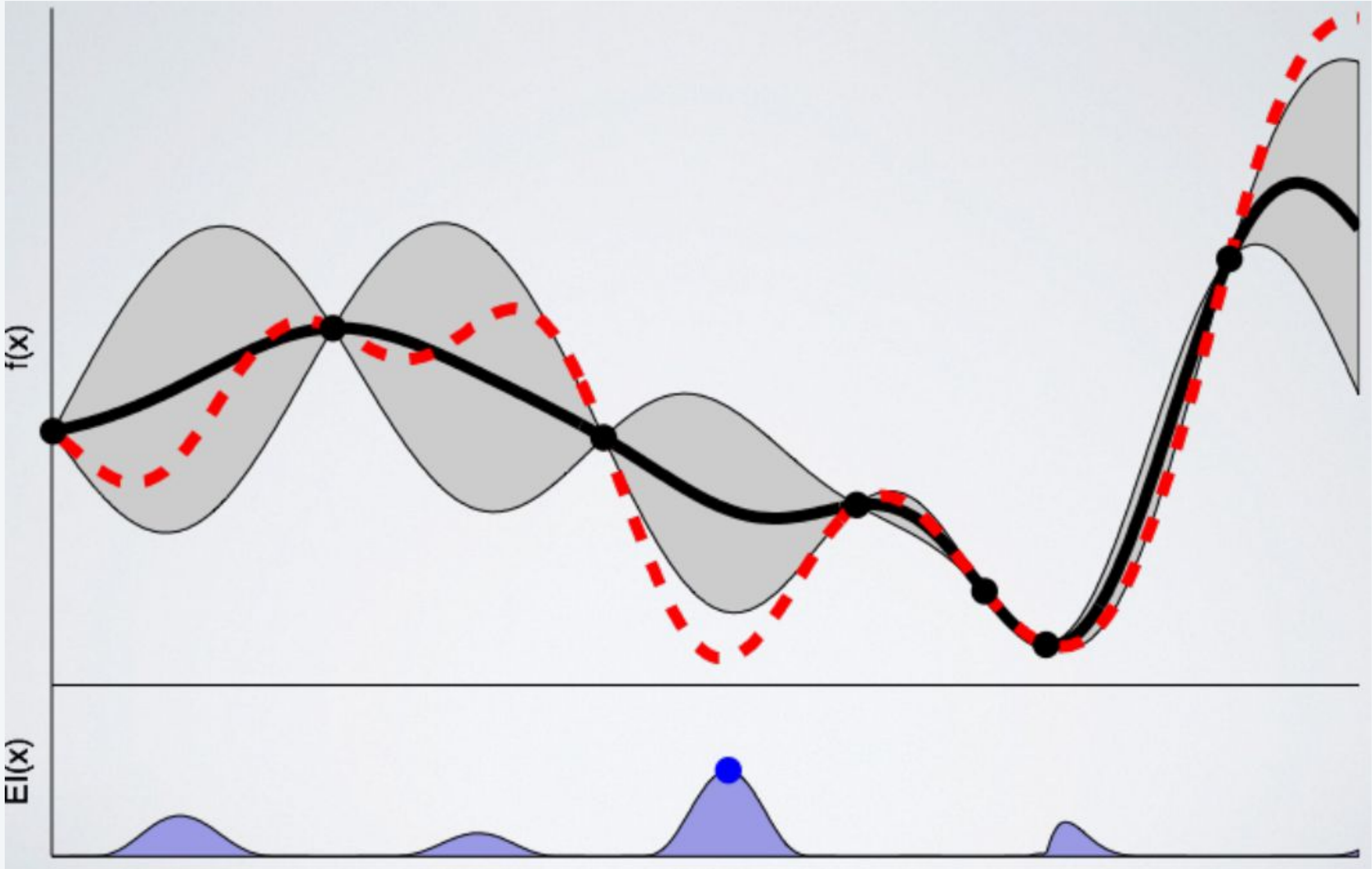
Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



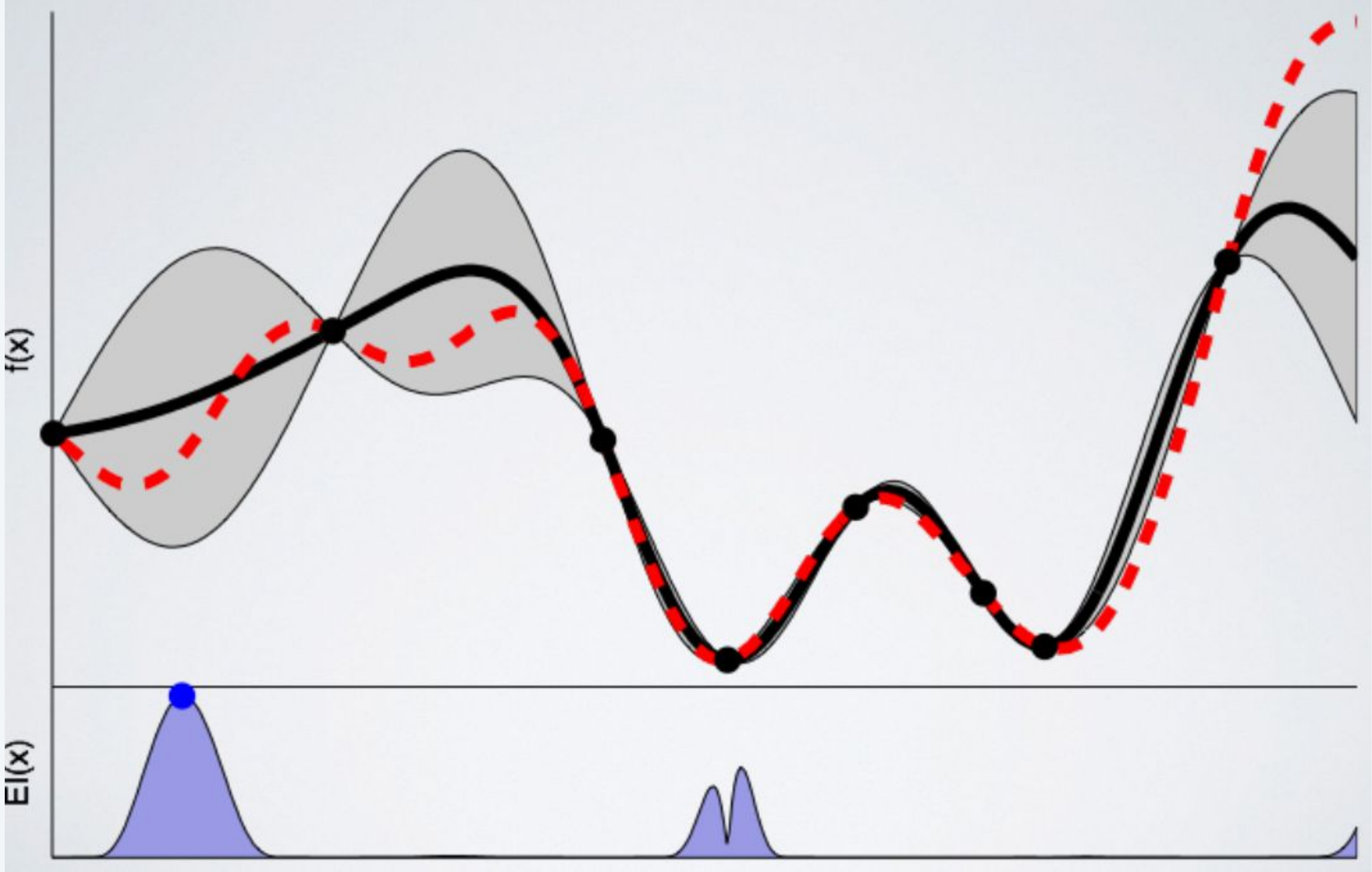
Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



Adapted from Ryan P. Adams (Harvard)

Illustrating Bayesian Optimization



Adapted from Ryan P. Adams (Harvard)

Bayesian Optimization

Why Doesn’t Everyone Use This?

- Potential fragility and poor default choice (getting function model wrong can be disastrous)
- No standard software available, tricky to build from scratch
- Experiments run sequentially
- Limited scalability in dimensions and evaluations

Adapted from Andre Ye

Bayesian Optimization

Summary

Surrogate optimize uses a surrogate (or approximation) function to estimate the objective function through sampling

Bayesian optimization puts surrogate optimization in a probabilistic framework by representing surrogate functions as probability distributions, which can be updated in light of new information

Acquisition functions are used to evaluate the probability that exploring a certain point in space will yield a ‘good’ return given what is currently known from the prior, balancing exploration and exploitation

Use Bayesian Optimization primarily when objective function is expensive to evaluate (e.g. in hyperparameter tuning). There are libraries like HyperOpt available for this purpose.

Adapted from Andre Ye


ANY
QUESTIONS
?