



UNIVERSITY OF  
CAMBRIDGE

# Machine Learning and the Physical World

## Lecture 6 : Sequential Decision Making - Bayesian Optimisation

---

Carl Henrik Ek - [che29@cam.ac.uk](mailto:che29@cam.ac.uk)

26th of October, 2023

<http://carlhenrik.com>

## Re-Cap

- Uncertainties to parametrise our ignorance/beliefs

## Re-Cap

- Uncertainties to parametrise our ignorance/beliefs
- Statistical Inference to update our knowledge from experiment

## Re-Cap

- Uncertainties to parametrise our ignorance/beliefs
- Statistical Inference to update our knowledge from experiment
- Emulators to parametrise our interest

# Active Learning

---

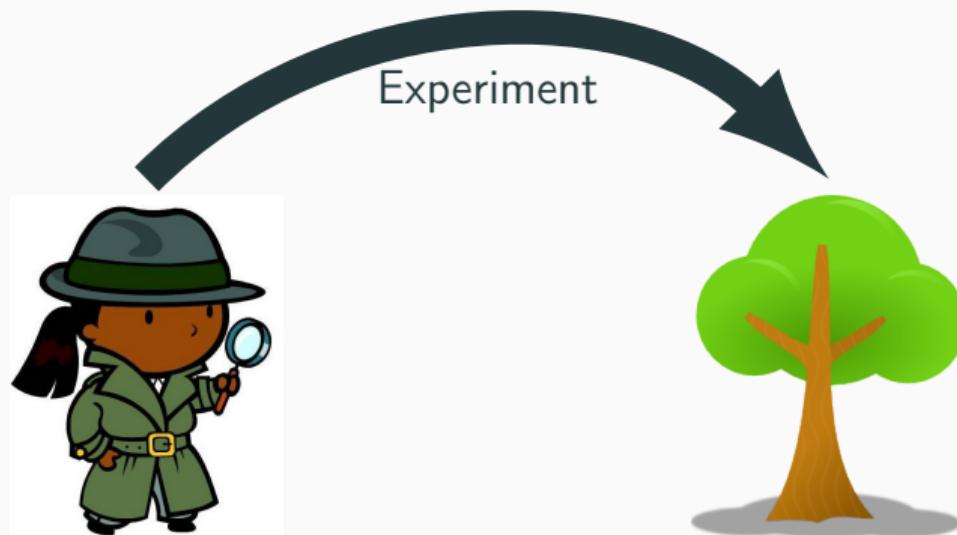


# Active Learning

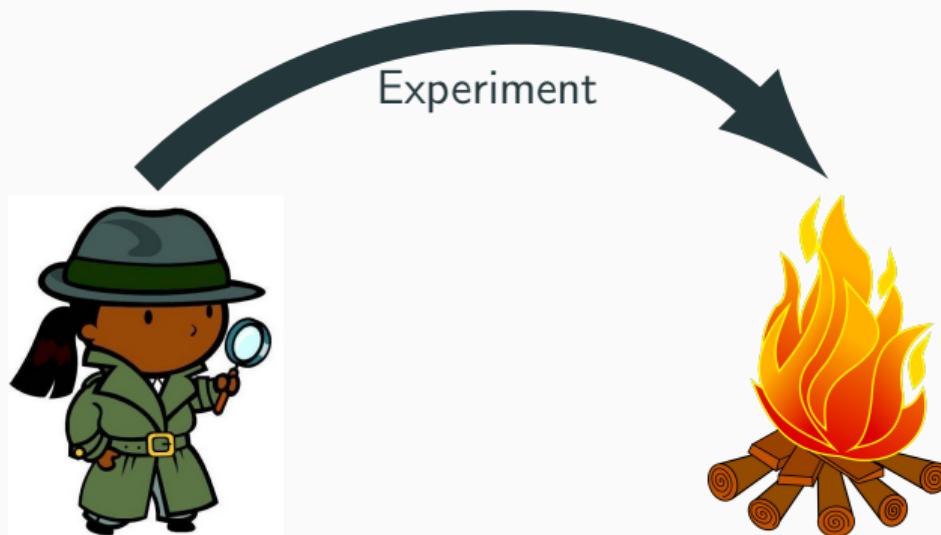
---



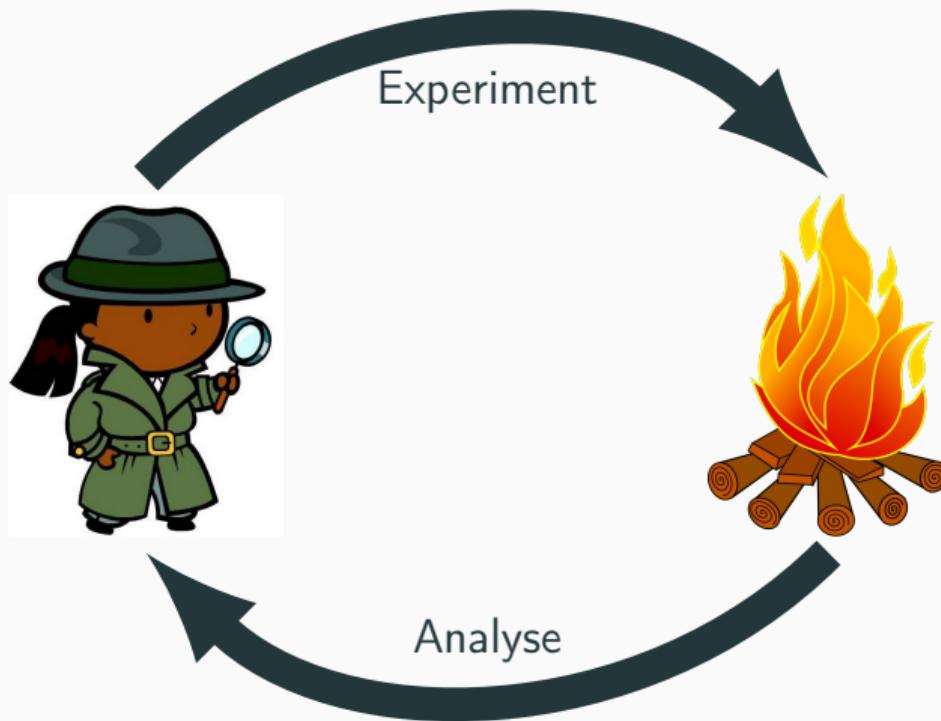
# Active Learning



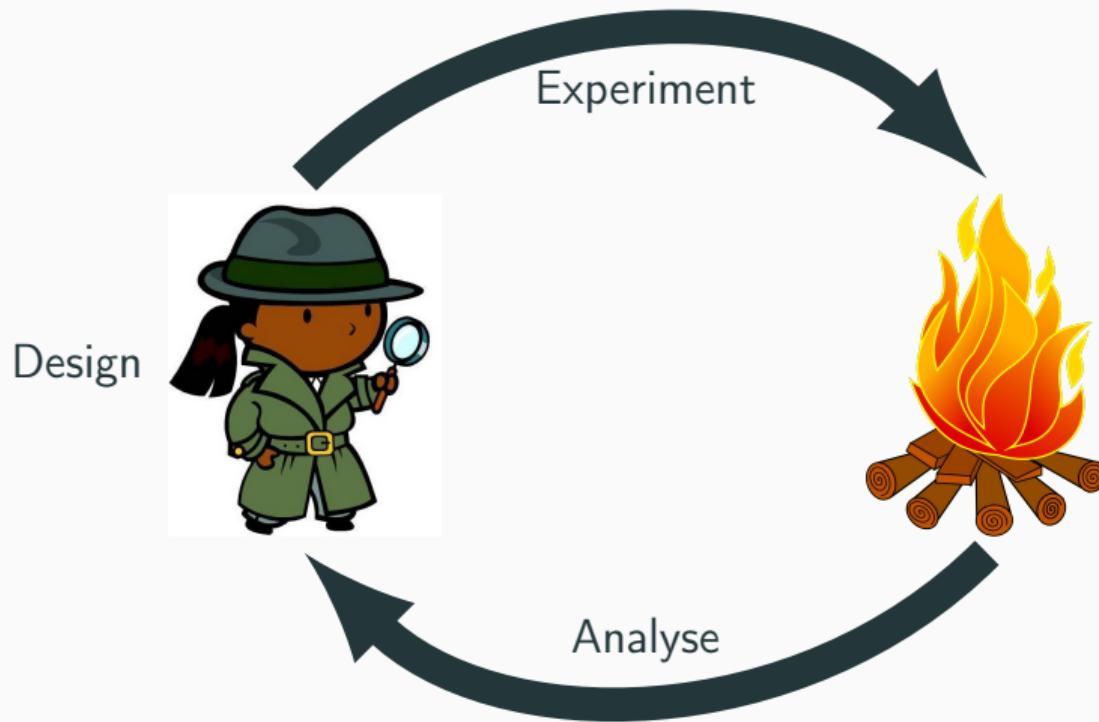
# Active Learning



# Active Learning

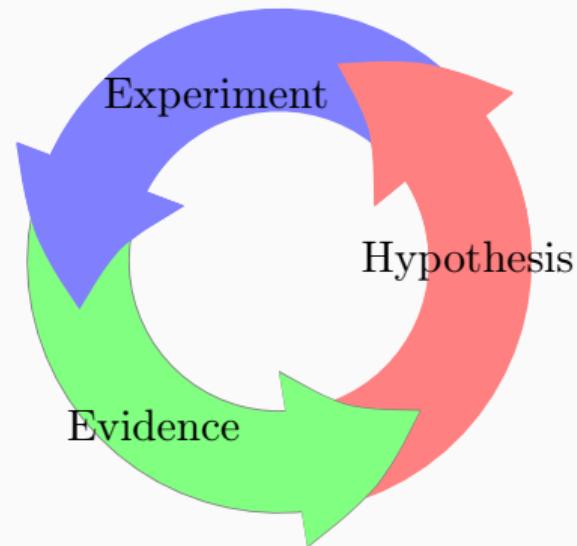


# Active Learning

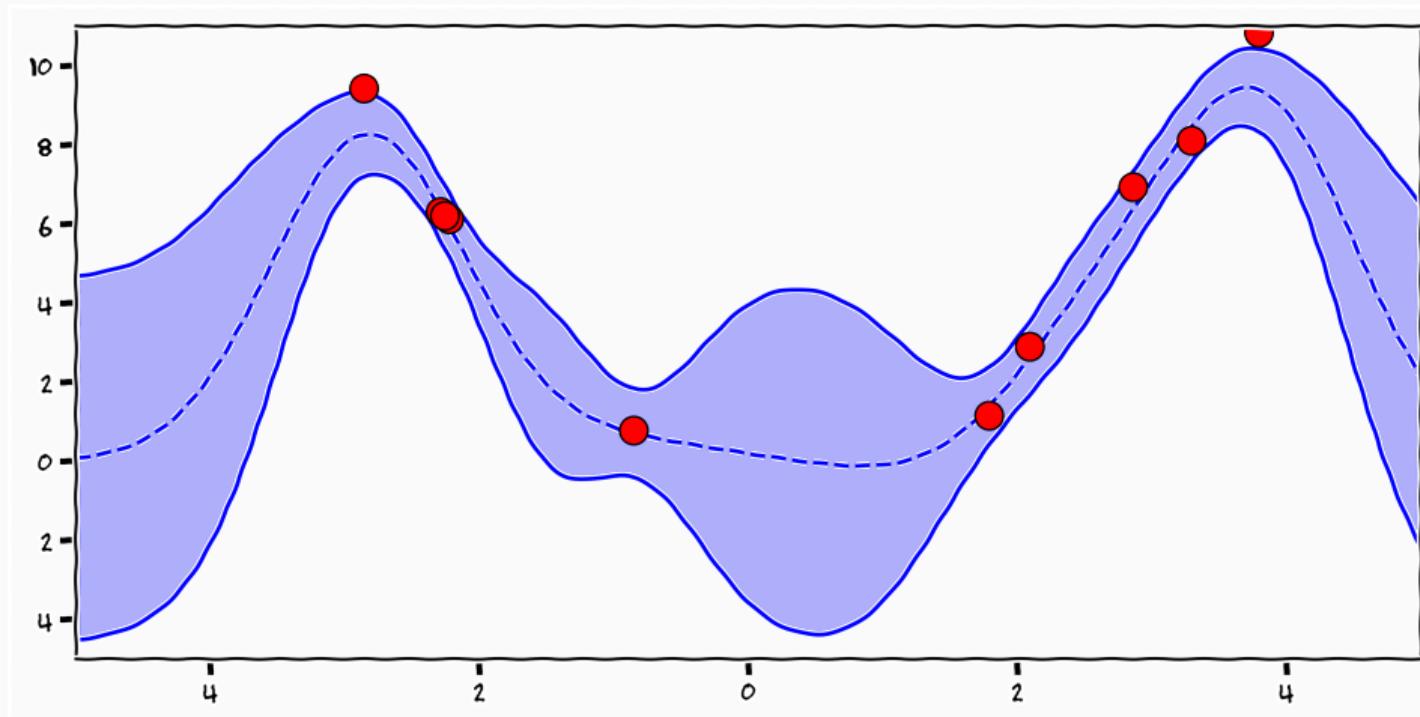


# The Role Uncertainty

---



## Functions



**Black-Box Optimisation** how can we find the minima of an explicitly unknown function?

**Black-Box Optimisation** how can we find the minima of an explicitly unknown function?

**Surrogate Models** how can we build a model as a surrogate for the unknown function?

**Black-Box Optimisation** how can we find the minima of an explicitly unknown function?

**Surrogate Models** how can we build a model as a surrogate for the unknown function?

**Sequential decision making** how can we come up with a strategy for sequentially exploring the function?

## Bayesian Optimisation

---

$$x^{(*)} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$$

- $\mathcal{X}$  is a bounded domain

# Black-box Optimisation

---

$$x^{(*)} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$$

- $\mathcal{X}$  is a bounded domain
- $f$  is explicitly unknown

$$x^{(*)} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$$

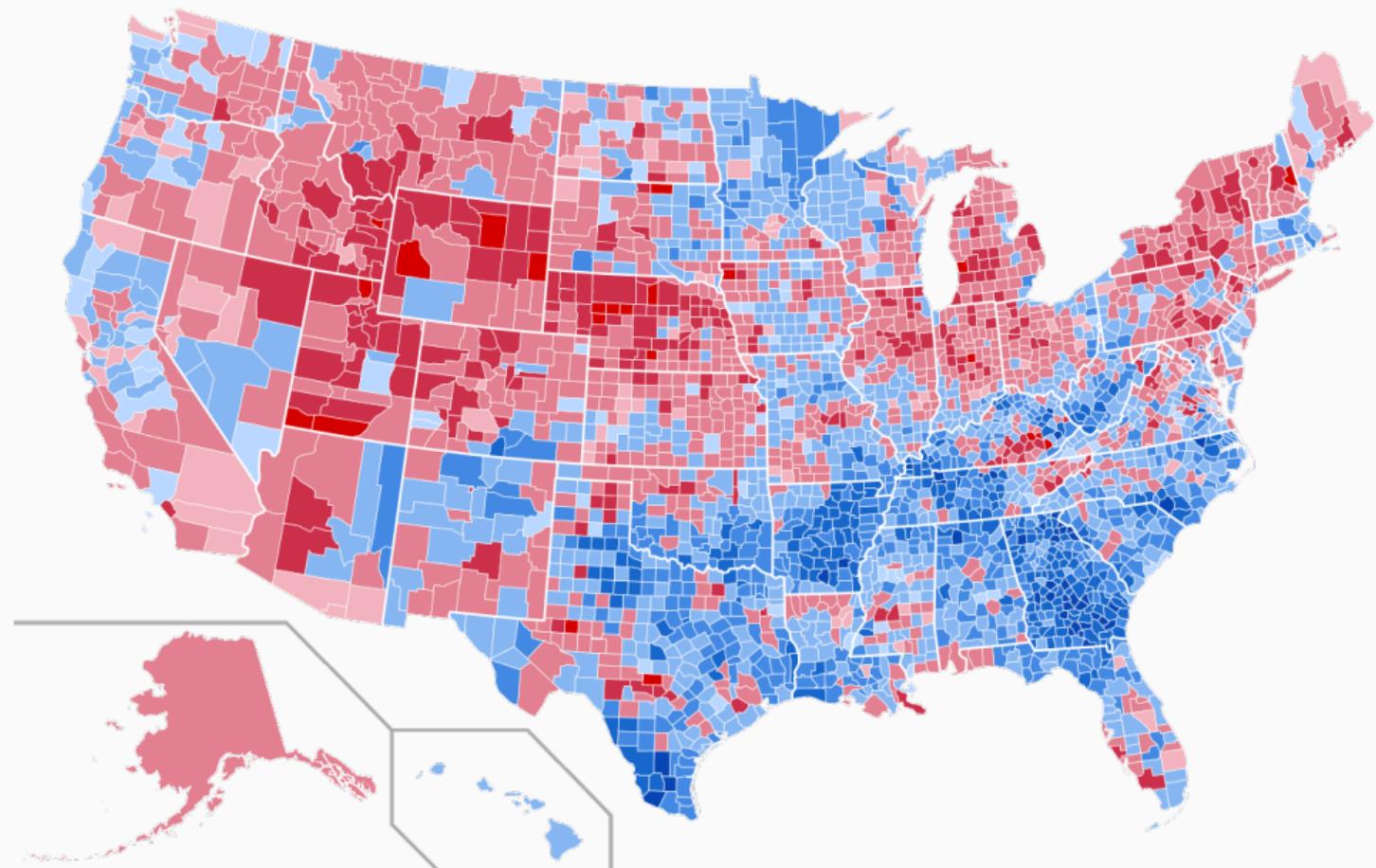
- $\mathcal{X}$  is a bounded domain
- $f$  is explicitly unknown
- Evaluations of  $f$  may be noisy

# Black-box Optimisation

---

$$x^{(*)} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$$

- $\mathcal{X}$  is a bounded domain
- $f$  is explicitly unknown
- Evaluations of  $f$  may be noisy
- Evaluations of  $f$  is expensive



- Random Search

$$f(x^{(-)}) \leq f(x^{(*)}) - \epsilon$$

## Random Search [Brochu et al., 2010]

---

- Random Search

$$f(x^{(-)}) \leq f(x^{(*)}) - \epsilon$$

- Lipschitz Continuity

$$\|f(x_1) - f(x_2)\| \leq C\|x_1 - x_2\|$$

## Random Search [Brochu et al., 2010]

---

- Random Search

$$f(x^{(-)}) \leq f(x^{(*)}) - \epsilon$$

- Lipschitz Continuity

$$\|f(x_1) - f(x_2)\| \leq C\|x_1 - x_2\|$$

- Requires  $\left(\frac{C}{2\epsilon}\right)^d$  evaluations on a  $d$ -dimensional hypercube

## Random Search [Brochu et al., 2010]

---

- Random Search

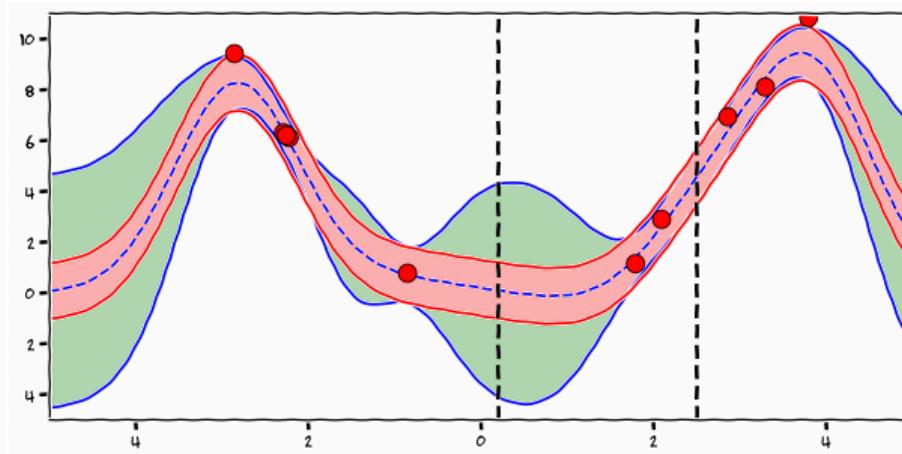
$$f(x^{(-)}) \leq f(x^{(*)}) - \epsilon$$

- Lipschitz Continuity

$$\|f(x_1) - f(x_2)\| \leq C\|x_1 - x_2\|$$

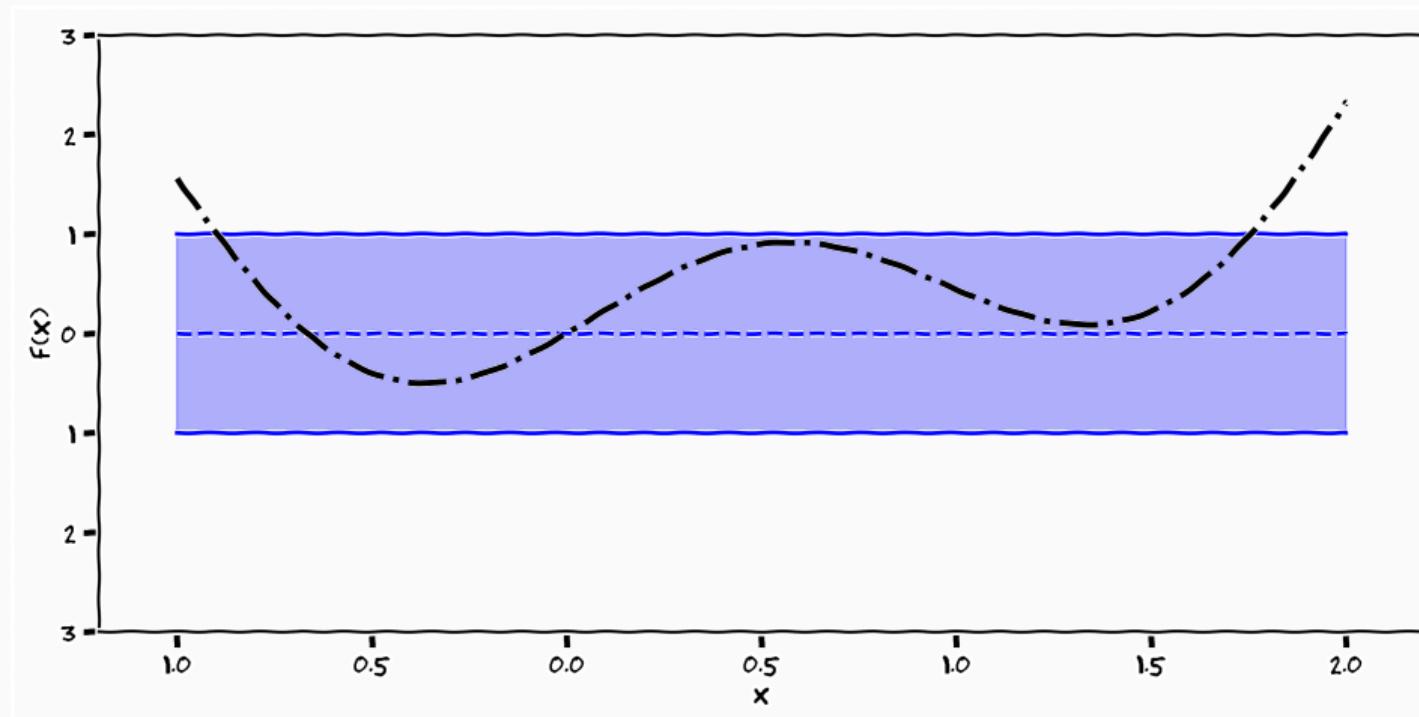
- Requires  $\left(\frac{C}{2\epsilon}\right)^d$  evaluations on a  $d$ -dimensional hypercube
- Surrogate model  $p(f)$

# Gaussian Process Surrogate

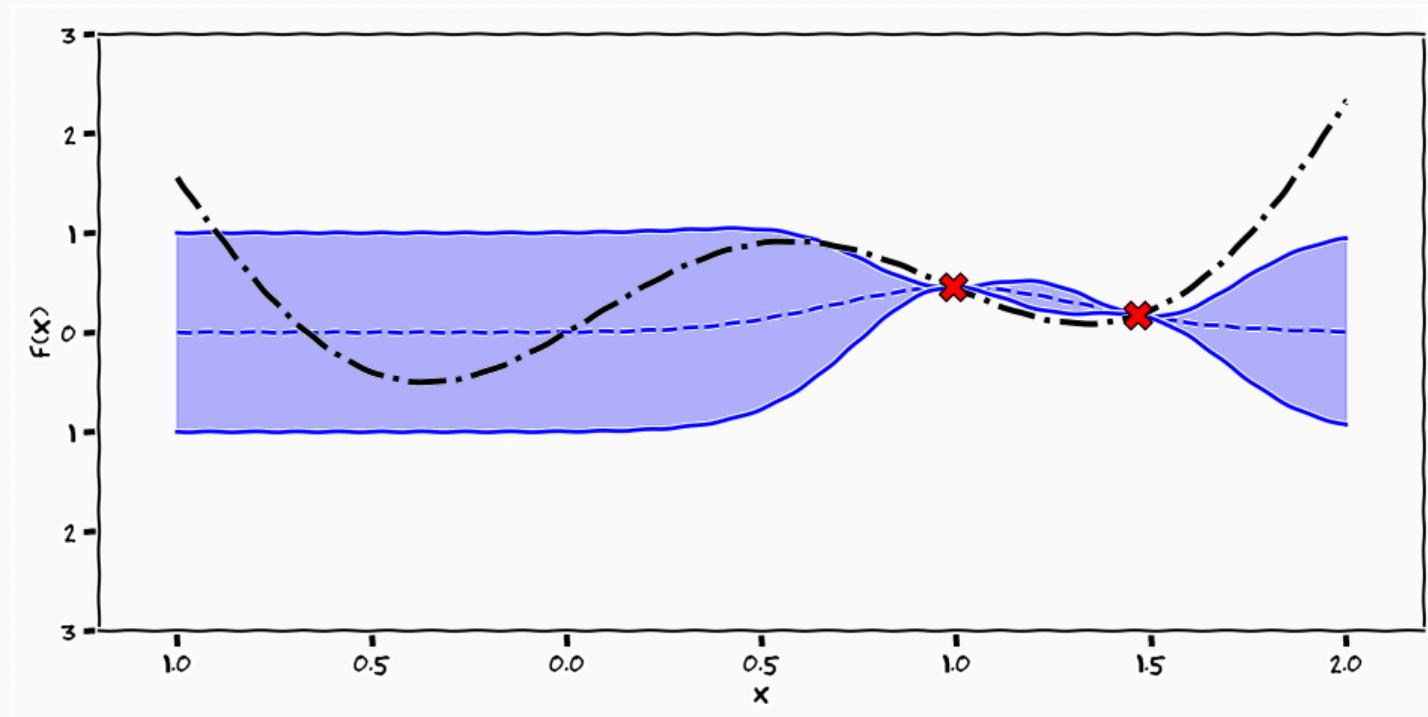


- allows for principled priors and **narrow** priors
- provides belief over the whole domain

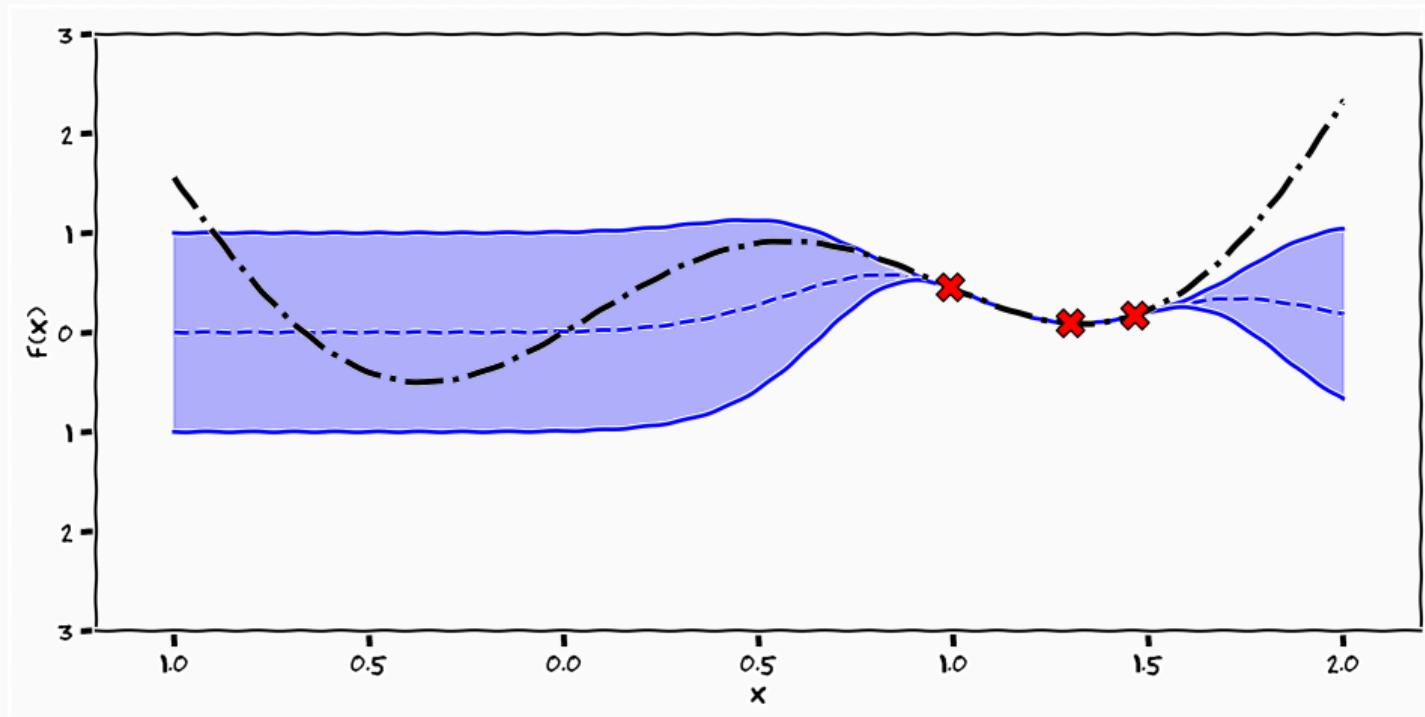
## Posterior Search: Random



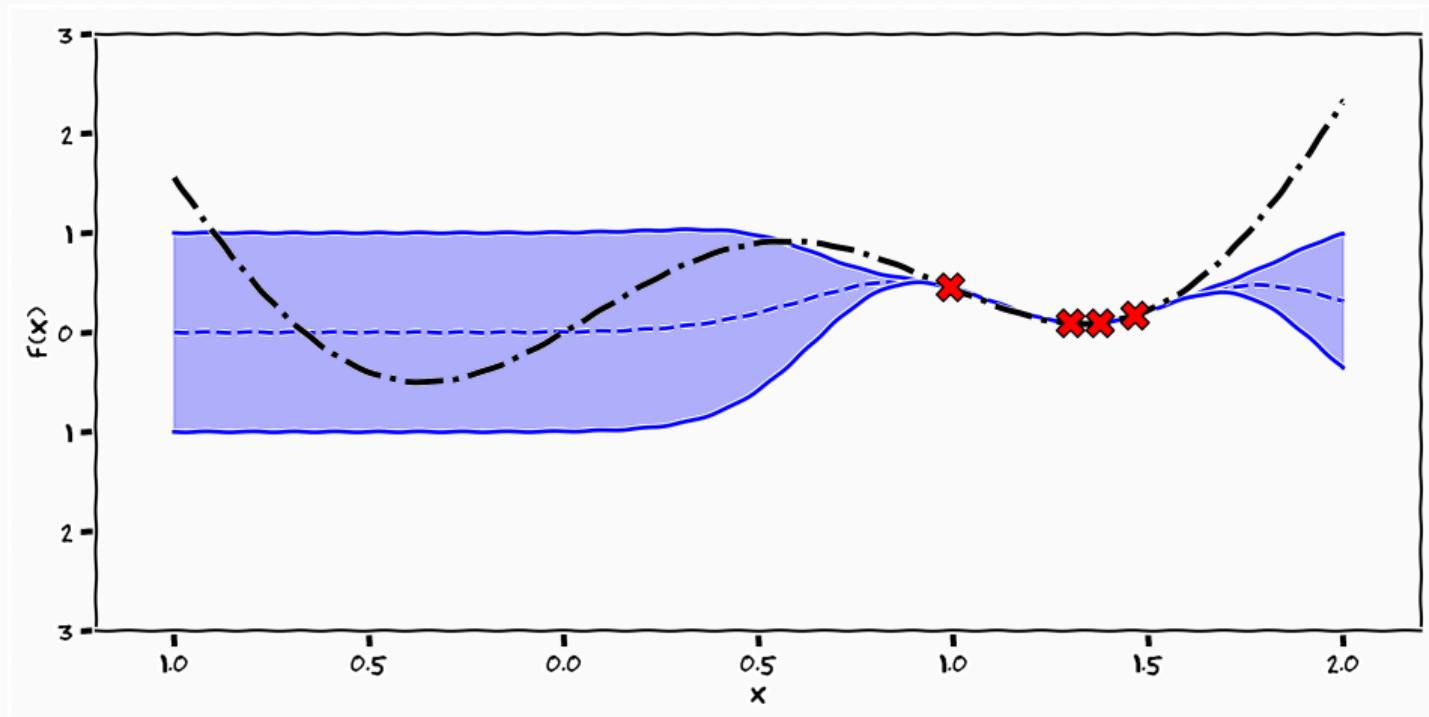
## Posterior Search: Random



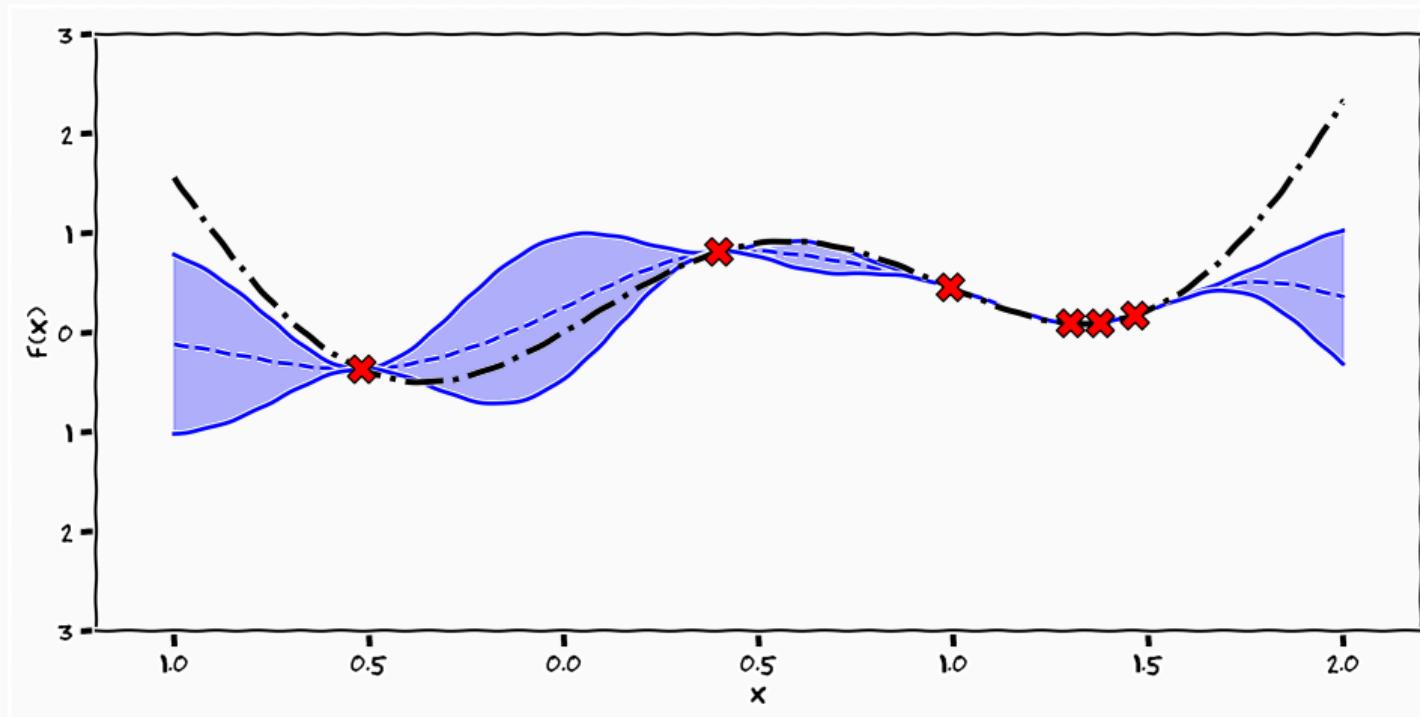
## Posterior Search: Random



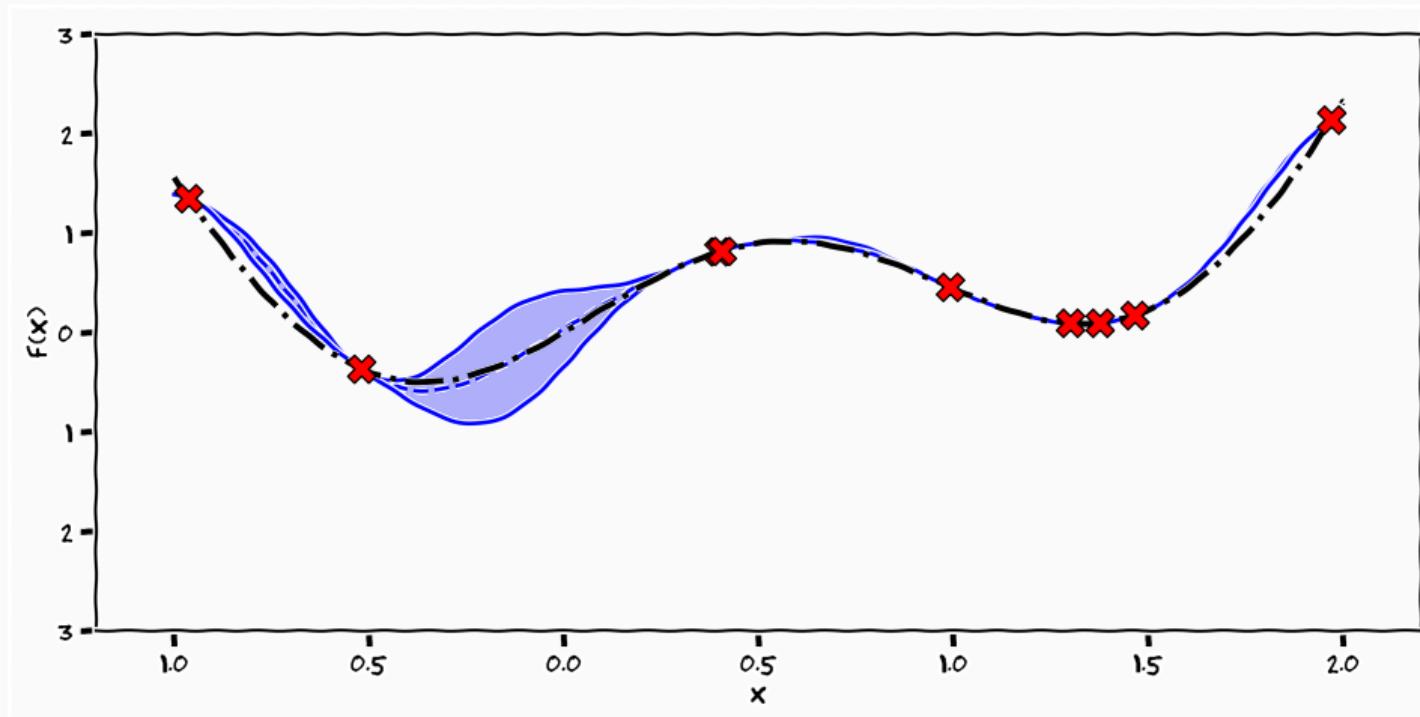
## Posterior Search: Random



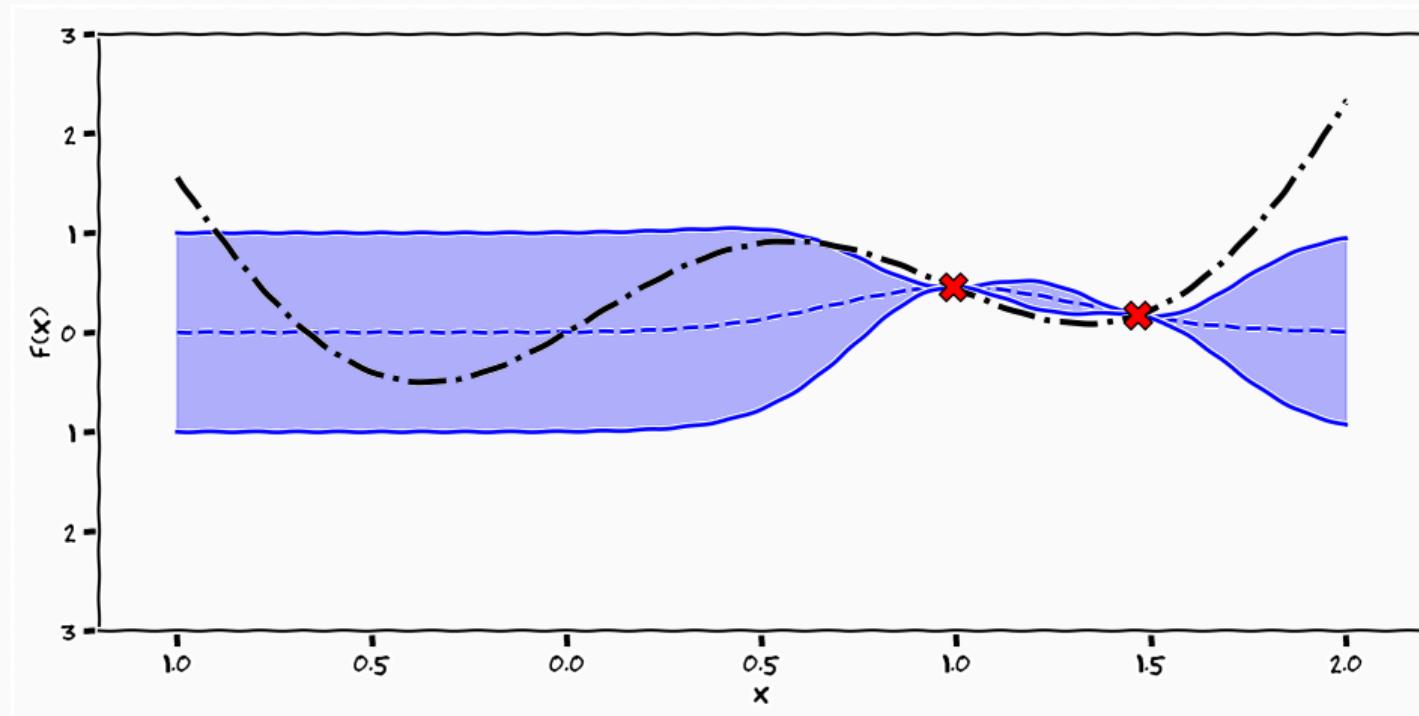
## Posterior Search: Random



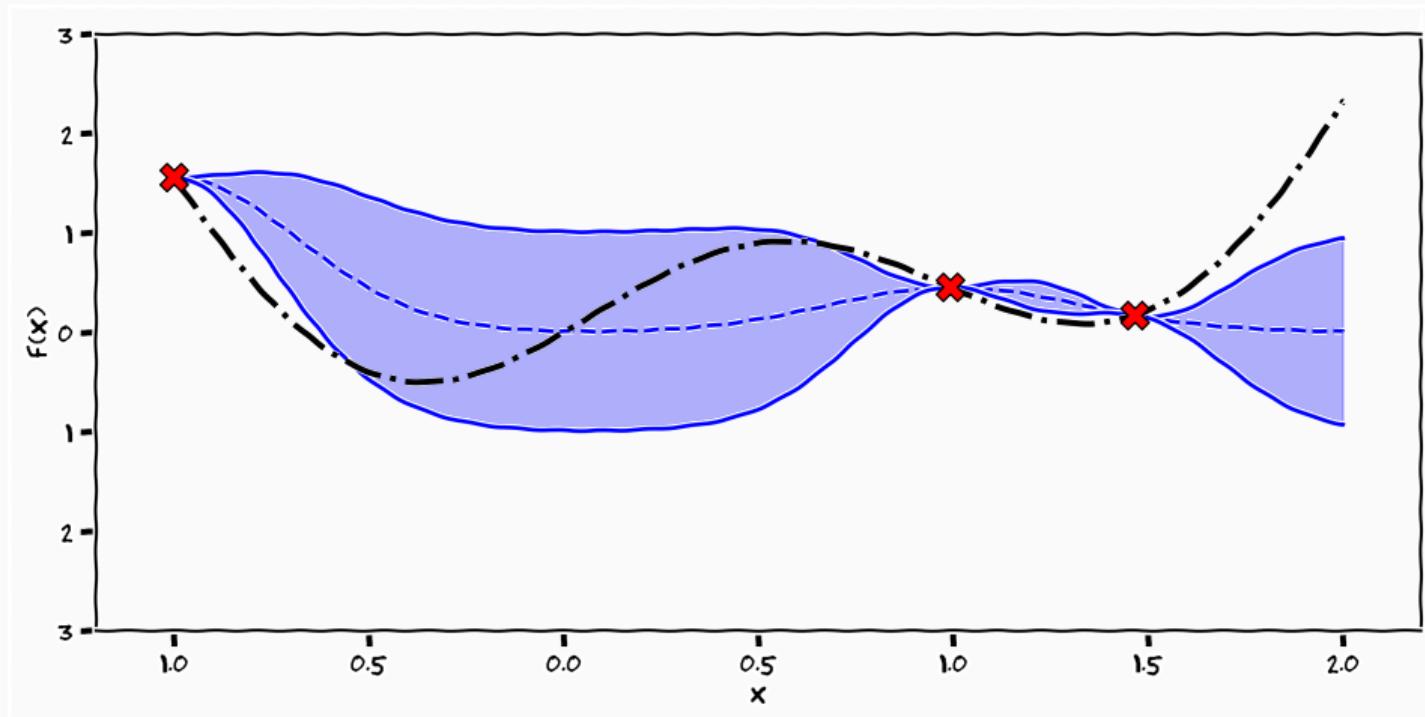
## Posterior Search: Random



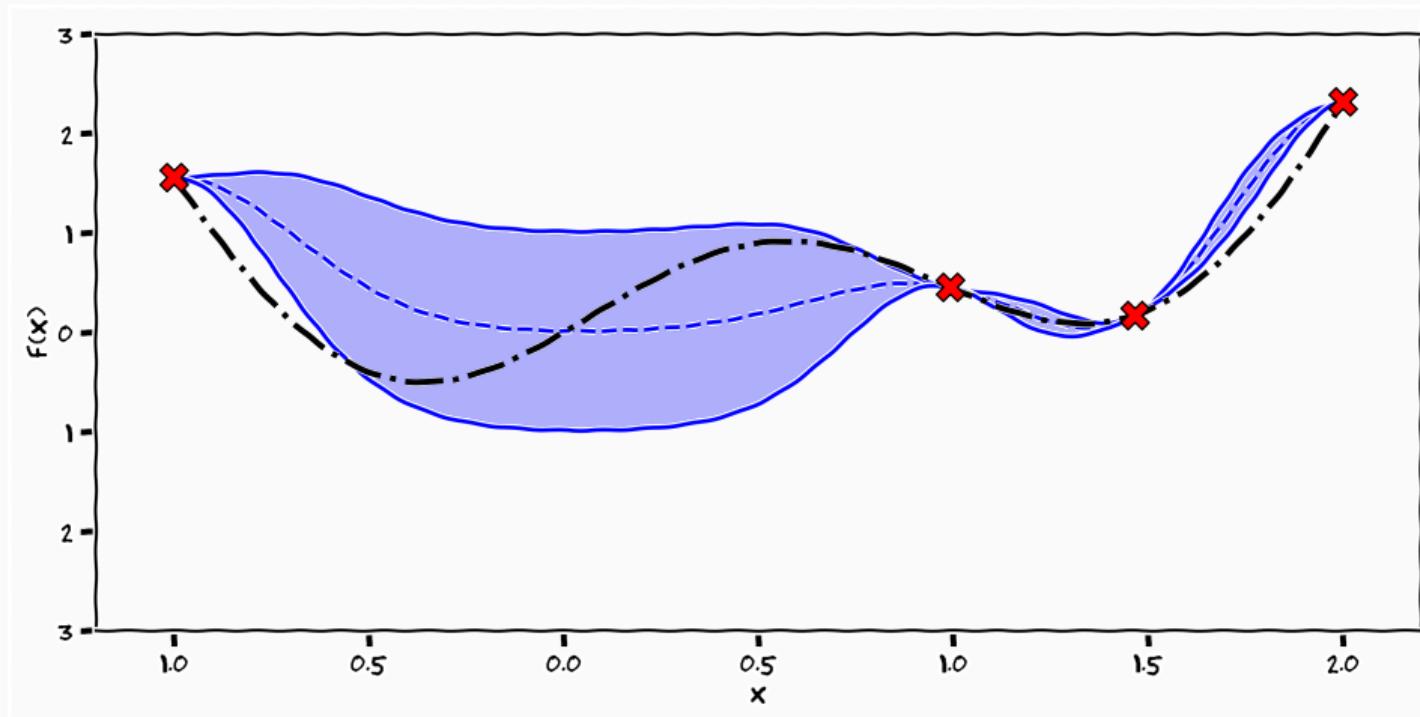
## Posterior Search: Min



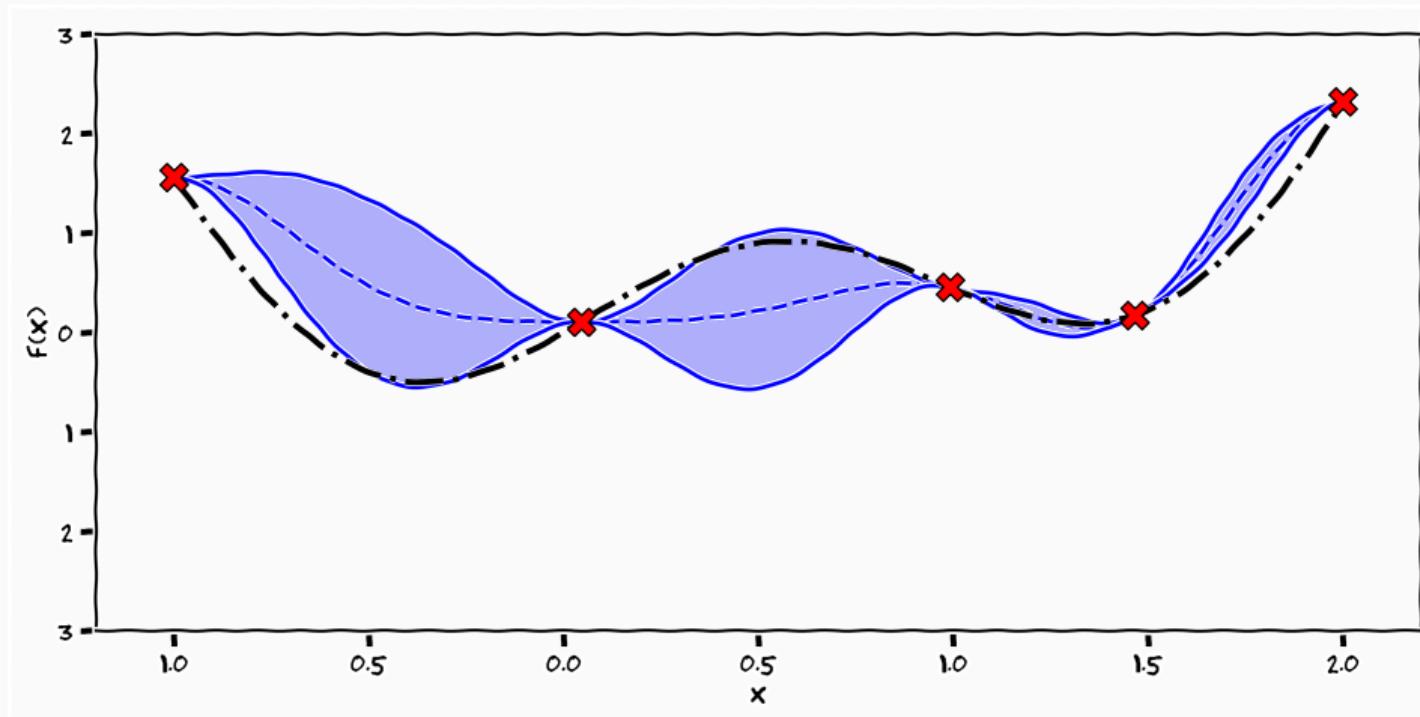
## Posterior Search: Min



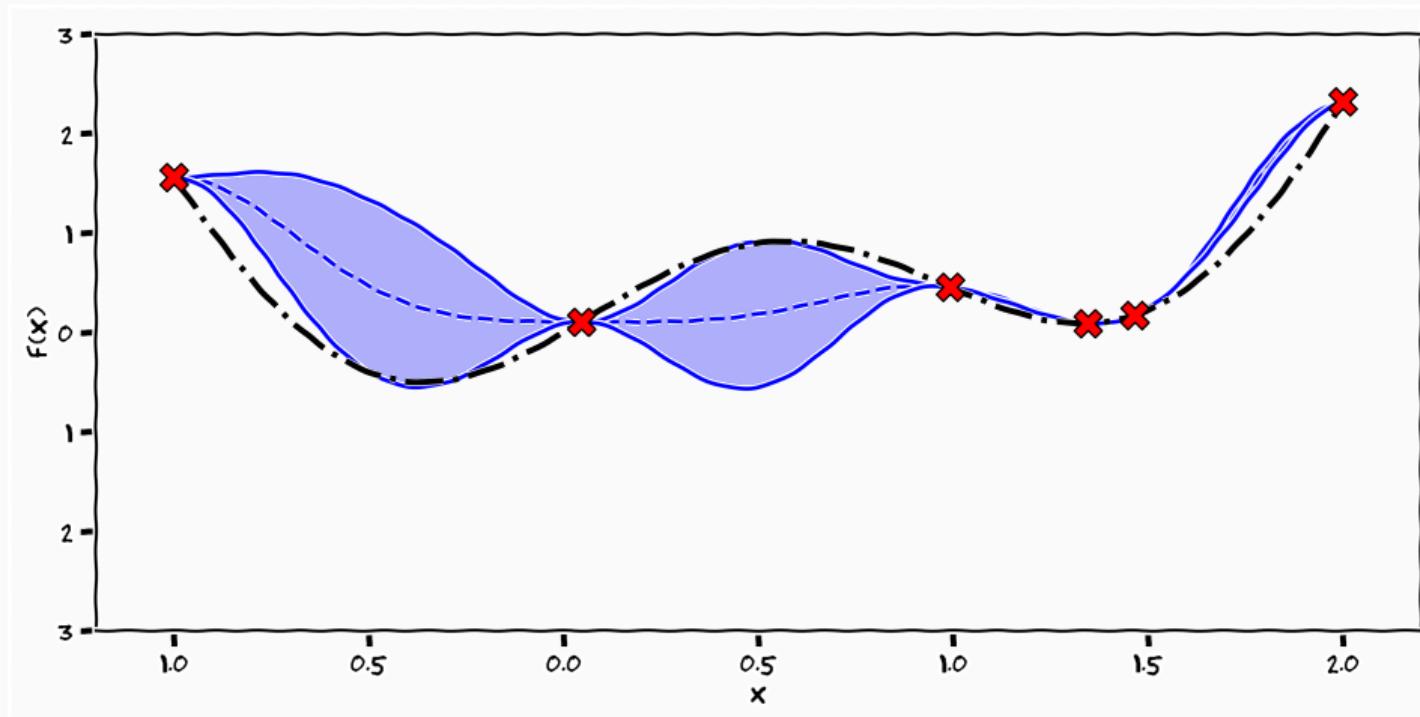
## Posterior Search: Min



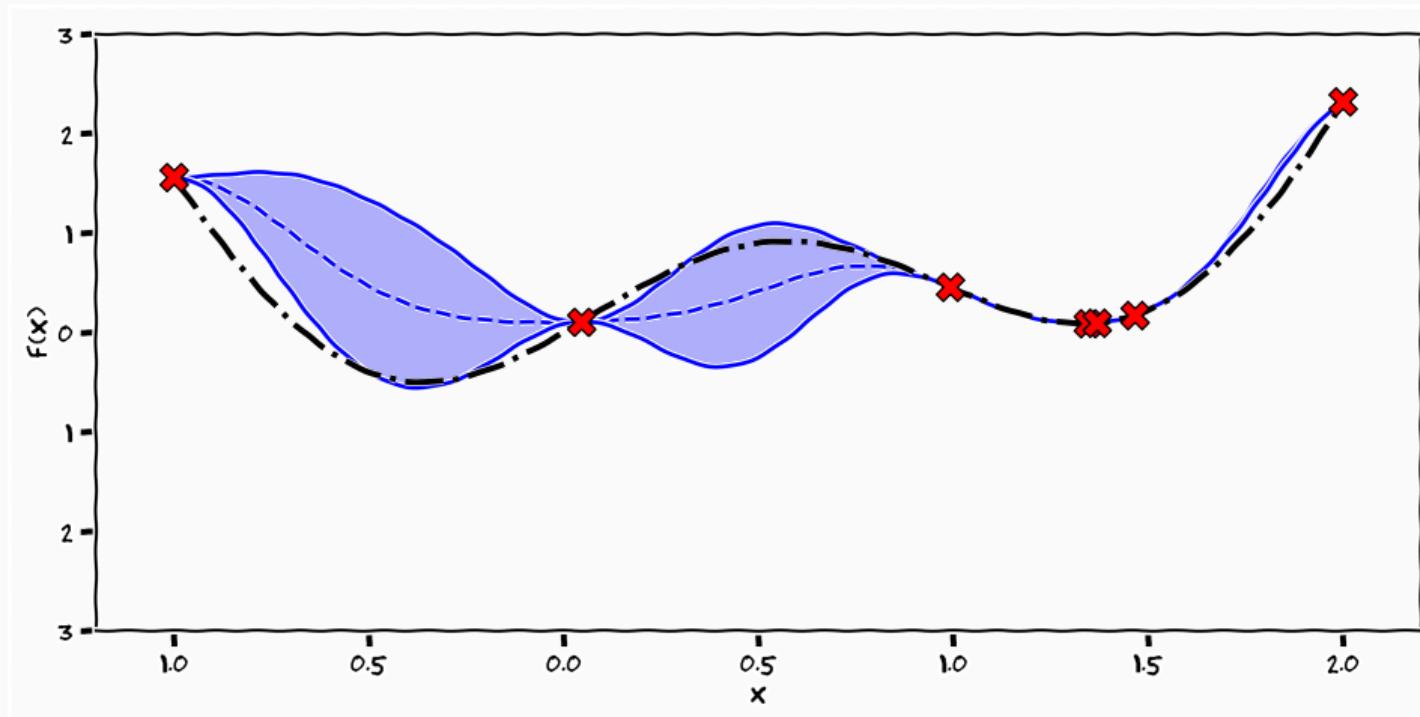
## Posterior Search: Min



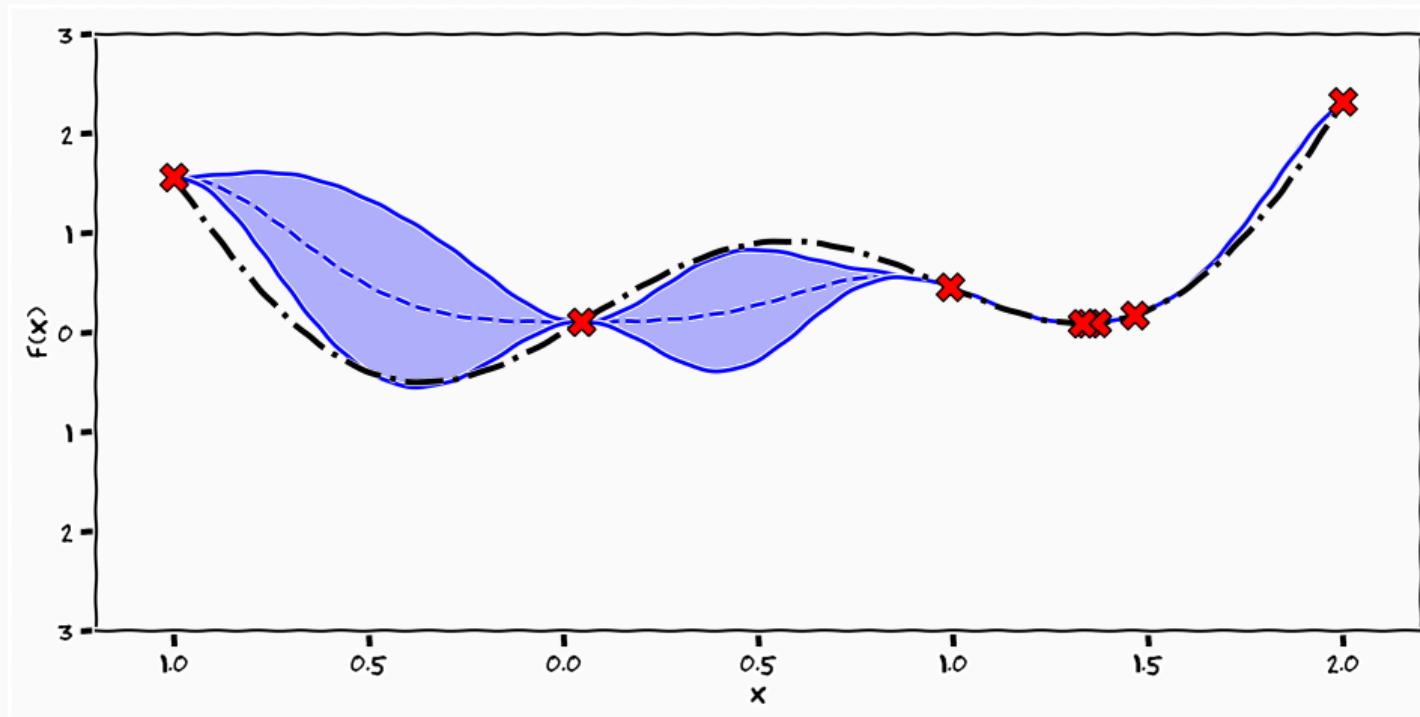
## Posterior Search: Min



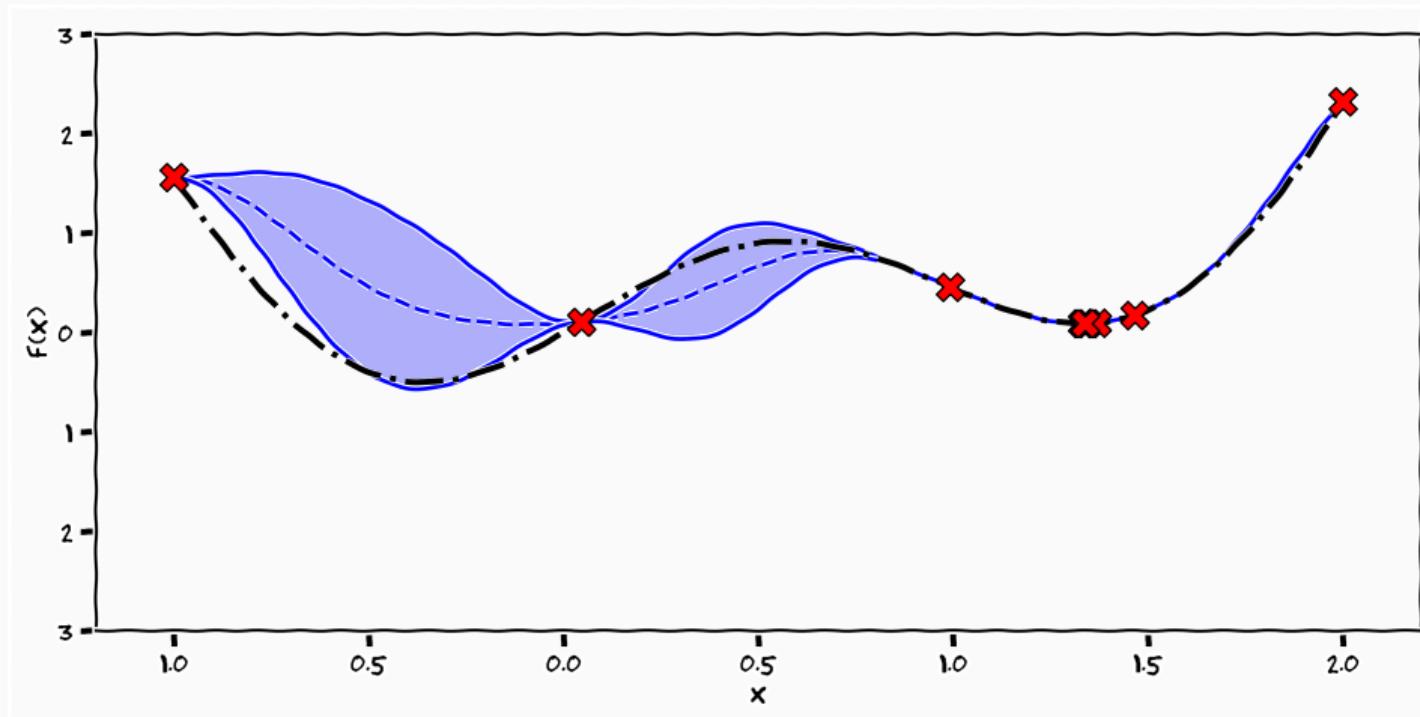
## Posterior Search: Min



## Posterior Search: Min



## Posterior Search: Min

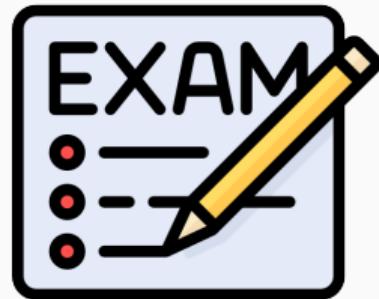








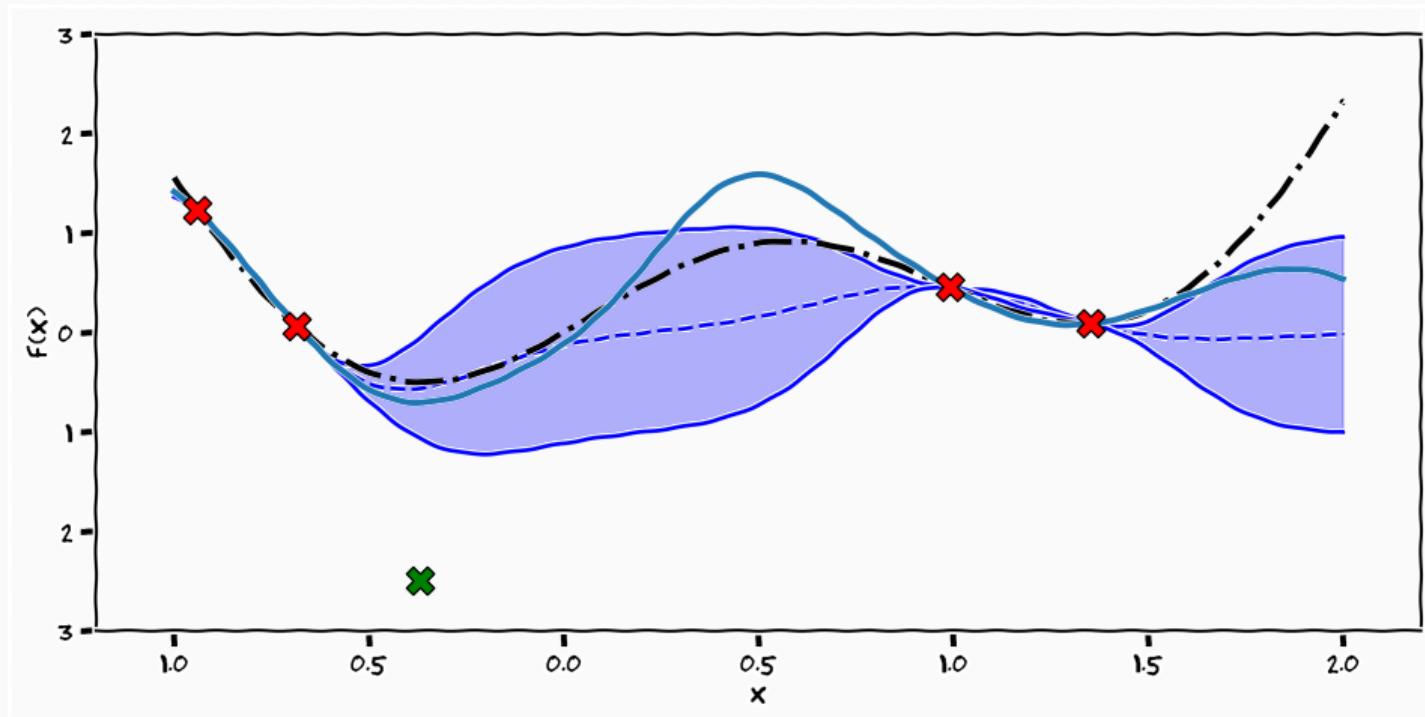




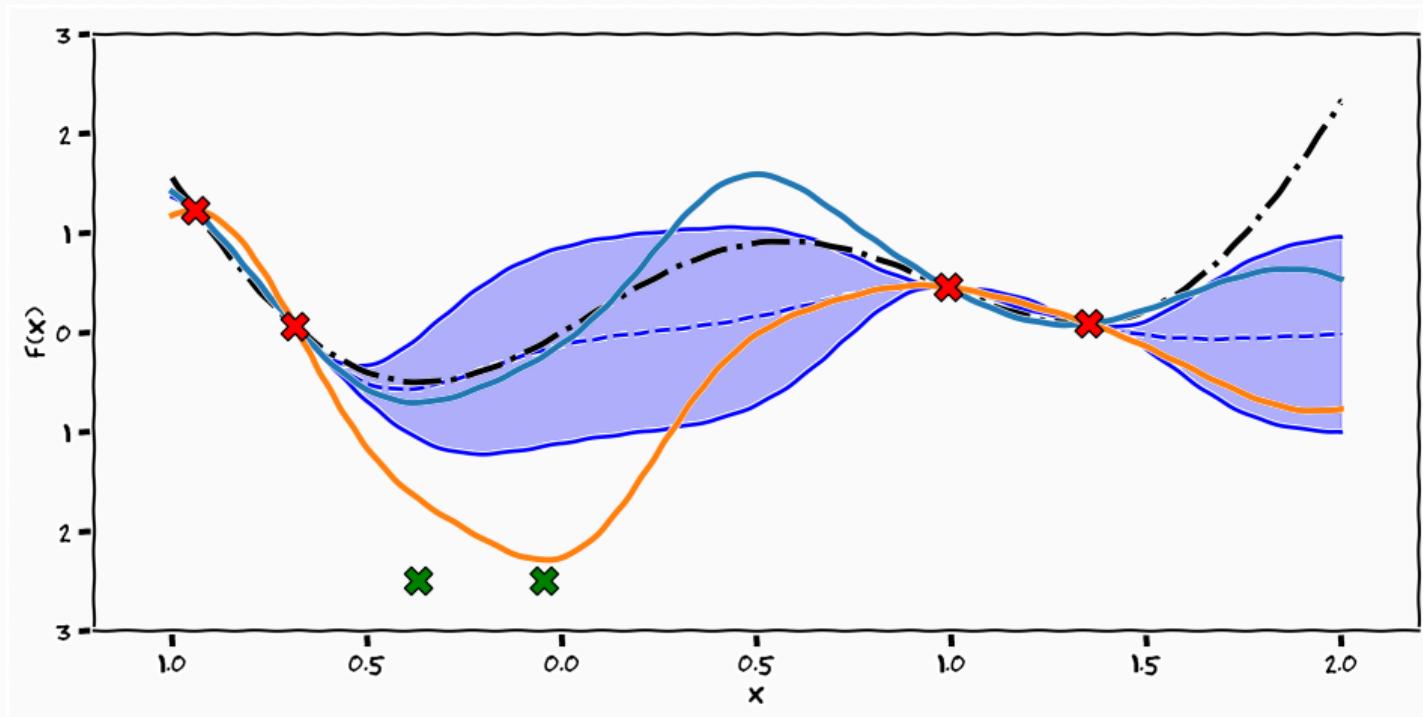
**Exploitation** use the knowledge that we currently have

**Exploration** try to gain new knowledge by trying new things

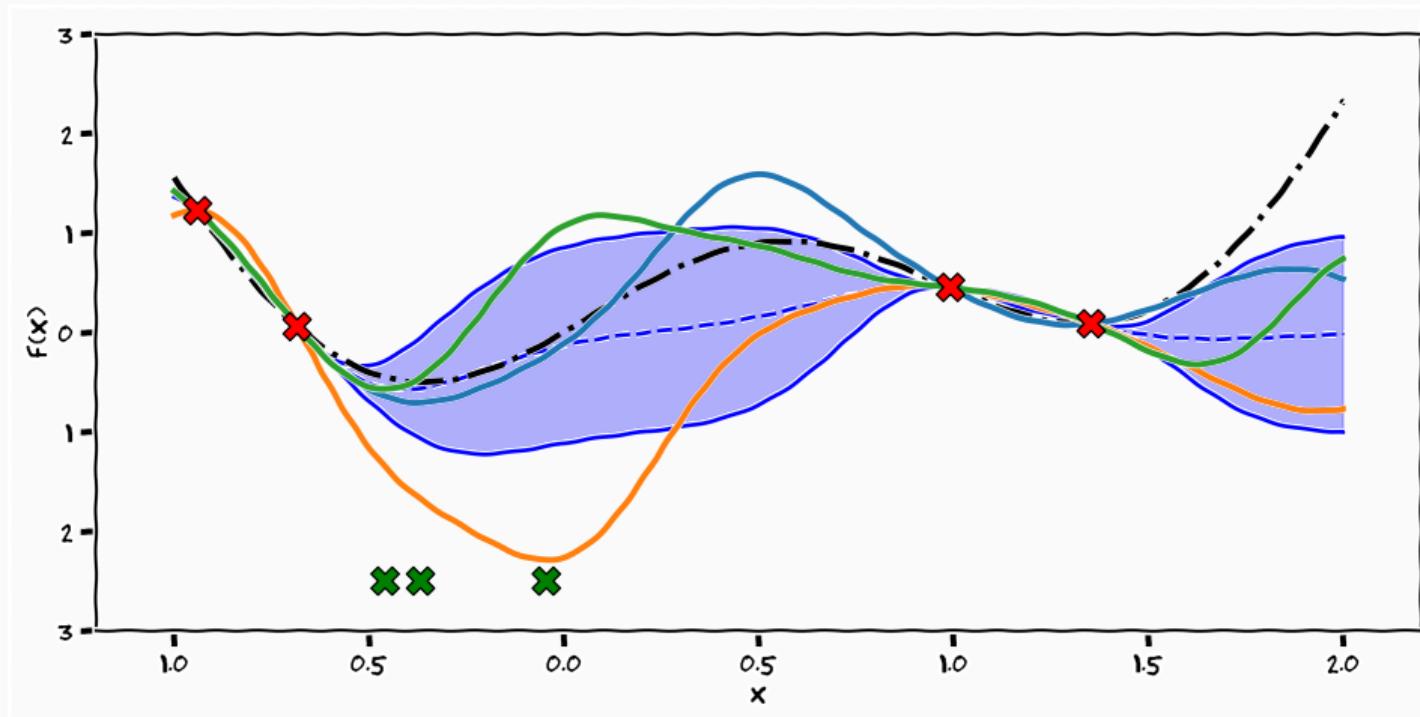
## Surrogate Uncertainty



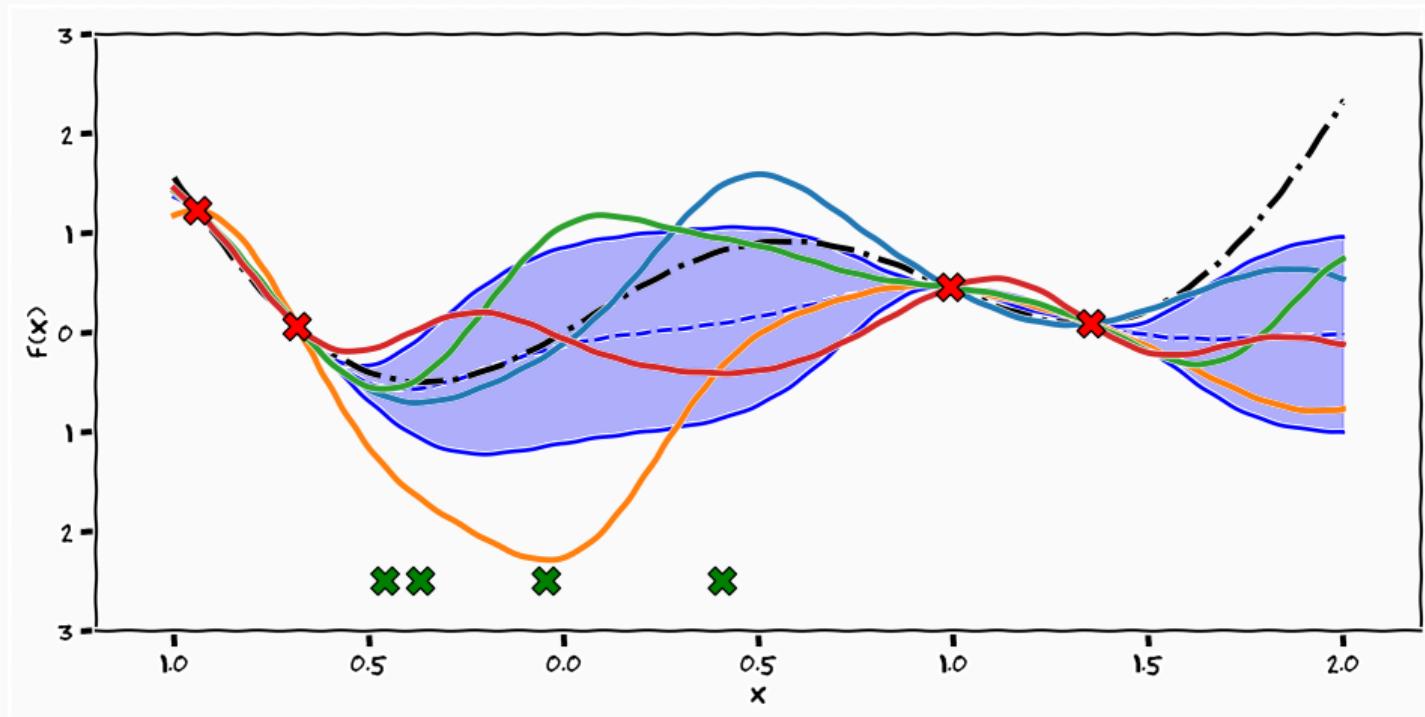
## Surrogate Uncertainty



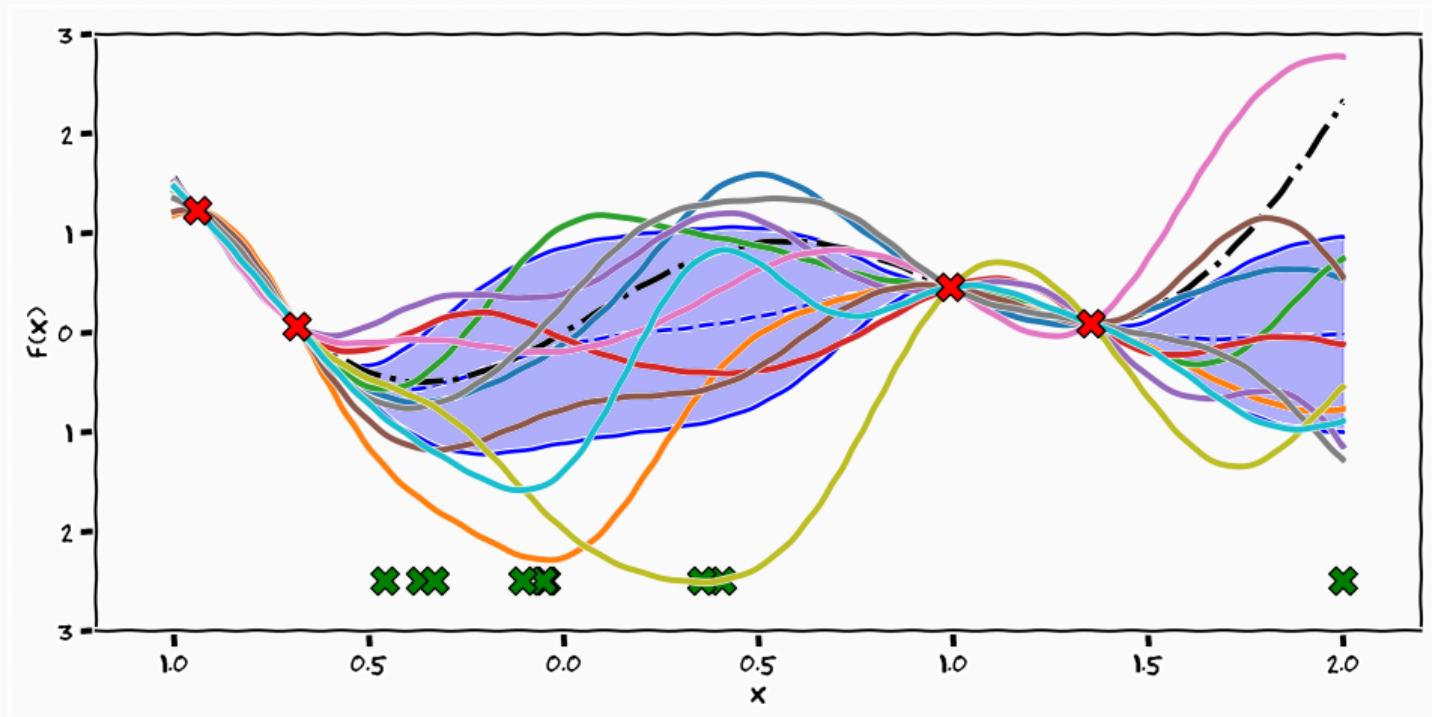
## Surrogate Uncertainty



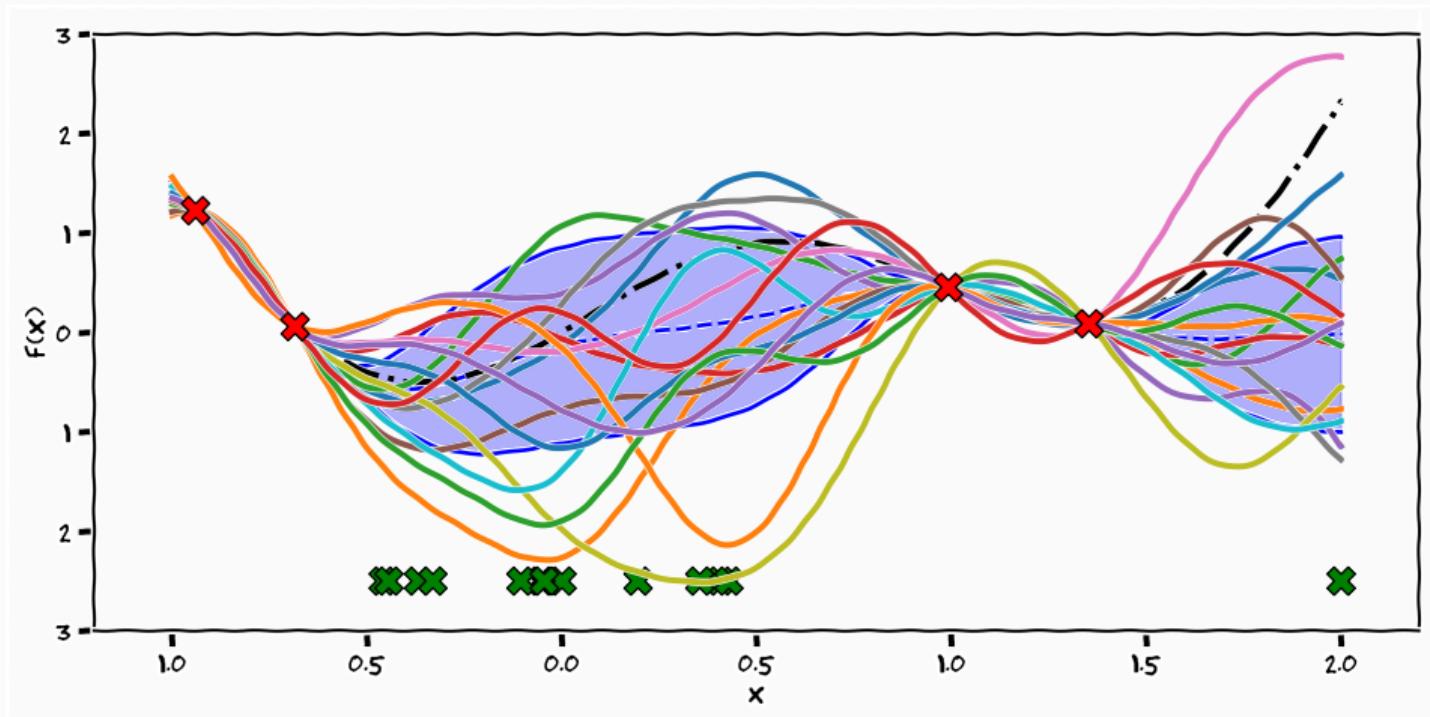
## Surrogate Uncertainty



## Surrogate Uncertainty



## Surrogate Uncertainty

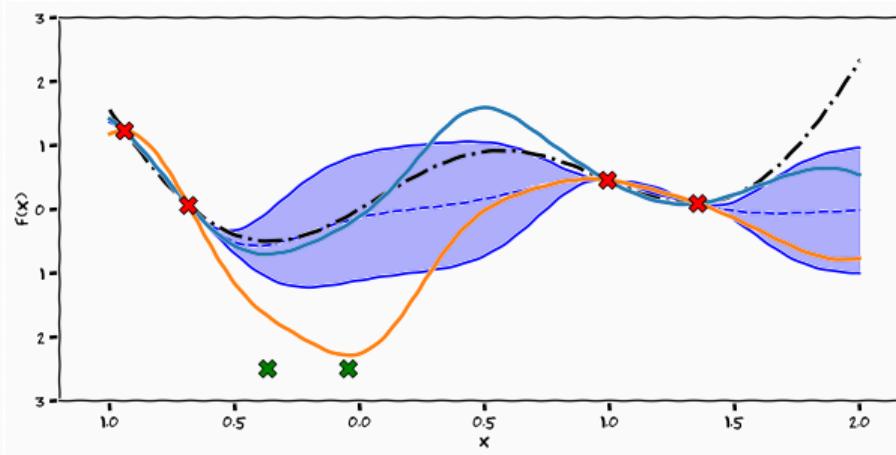


## Acquisition Function

$$x_{n+1} = \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x; \{x_i, y_i\}_{i=1}^n, \mathcal{M}_n)$$

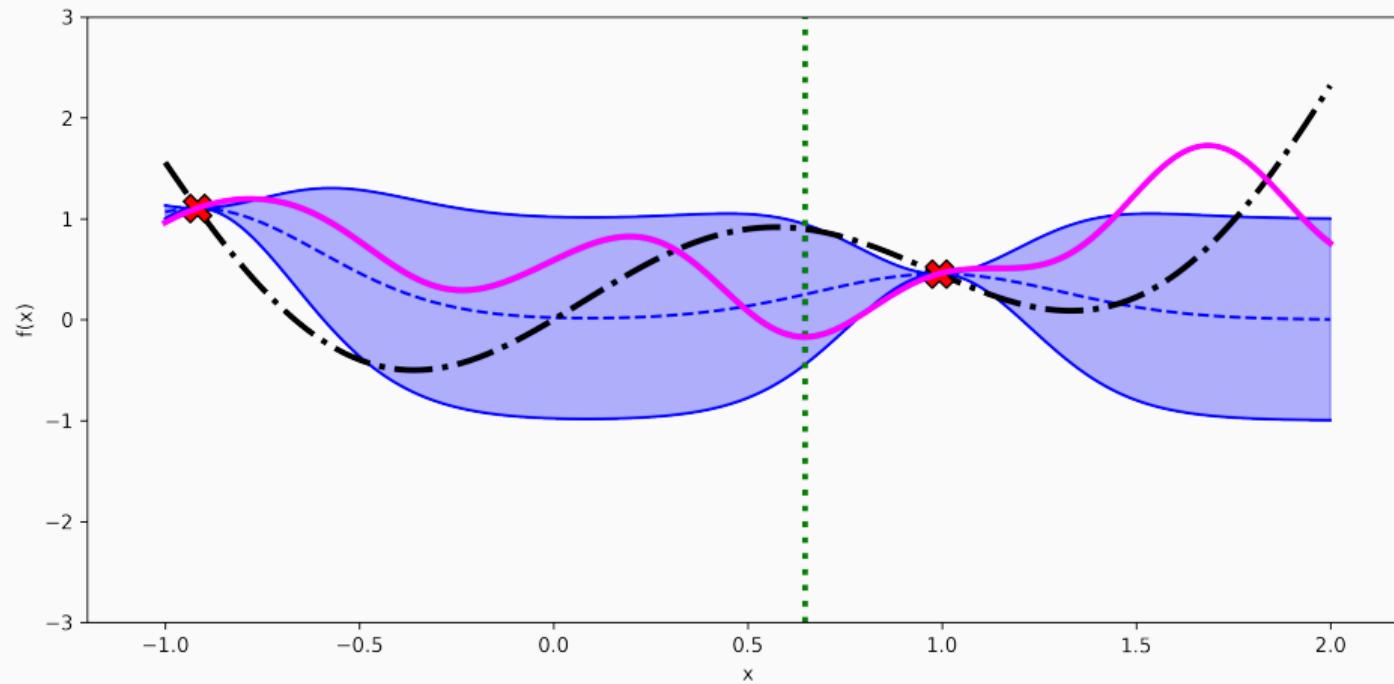
- Formulate a sequential decision problem
- This will work well if  $\alpha(x)$ 
  - is cheap to compute
  - balances *exploration* and *exploitation*

## Thompson Sampling [Thompson, 1933]

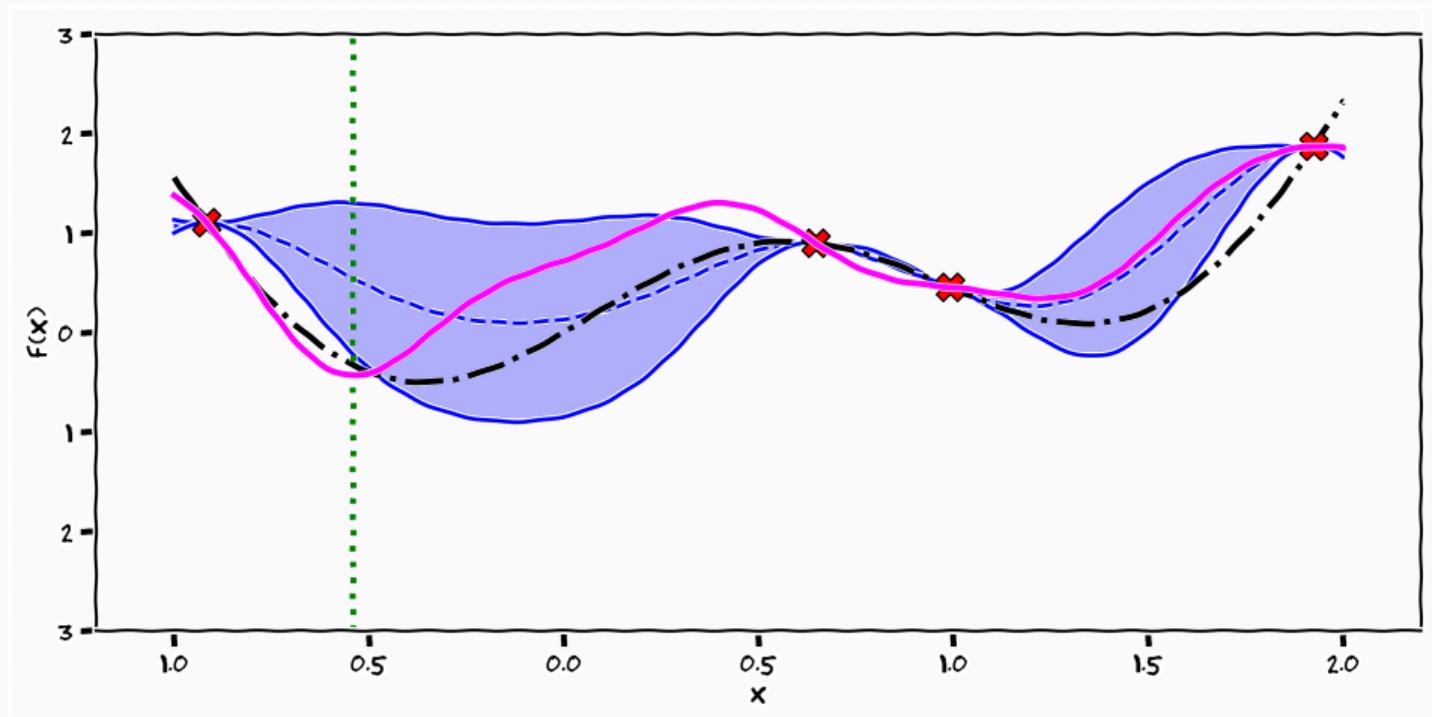


$$-\alpha(x; \{x_i, y_i\}_{i=1}^n, \mathcal{M}_n) \sim p(f \mid \{x_i, y_i\}_{i=1}^n)$$

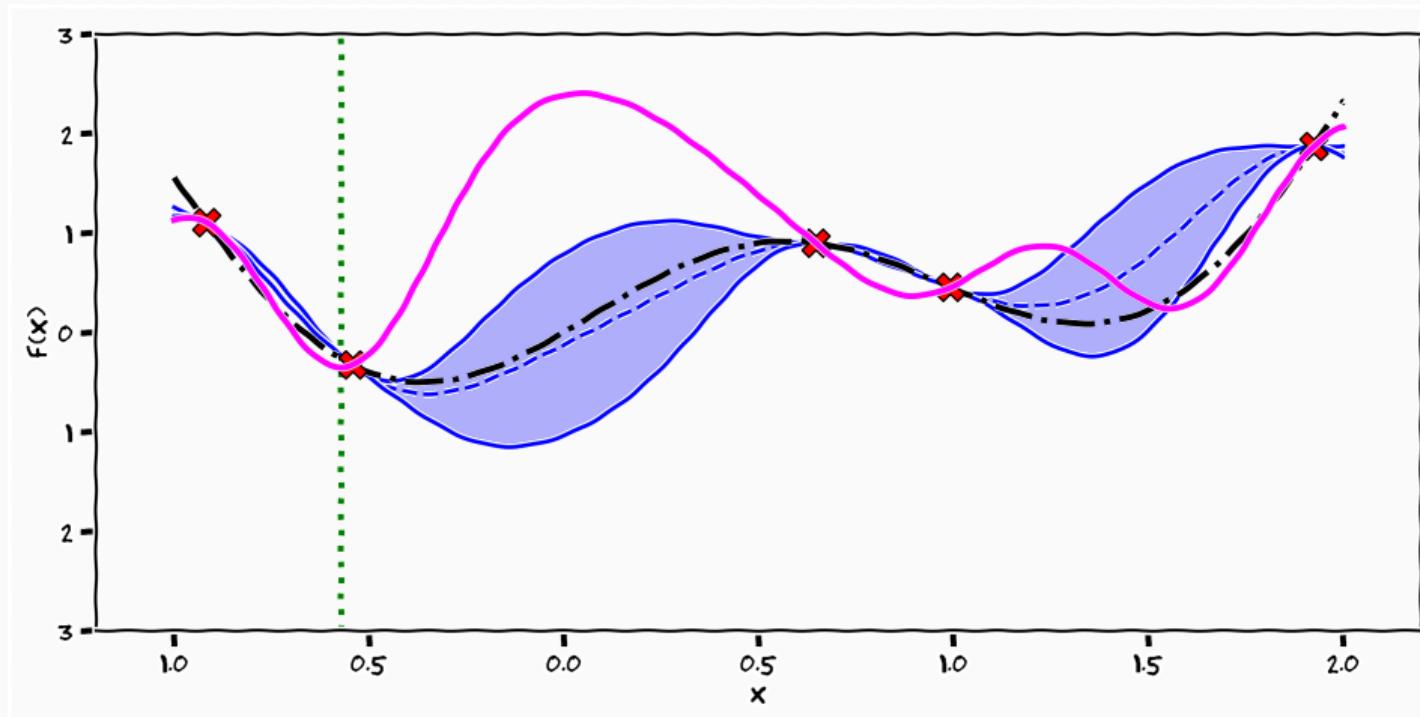
# Thompson Sampling



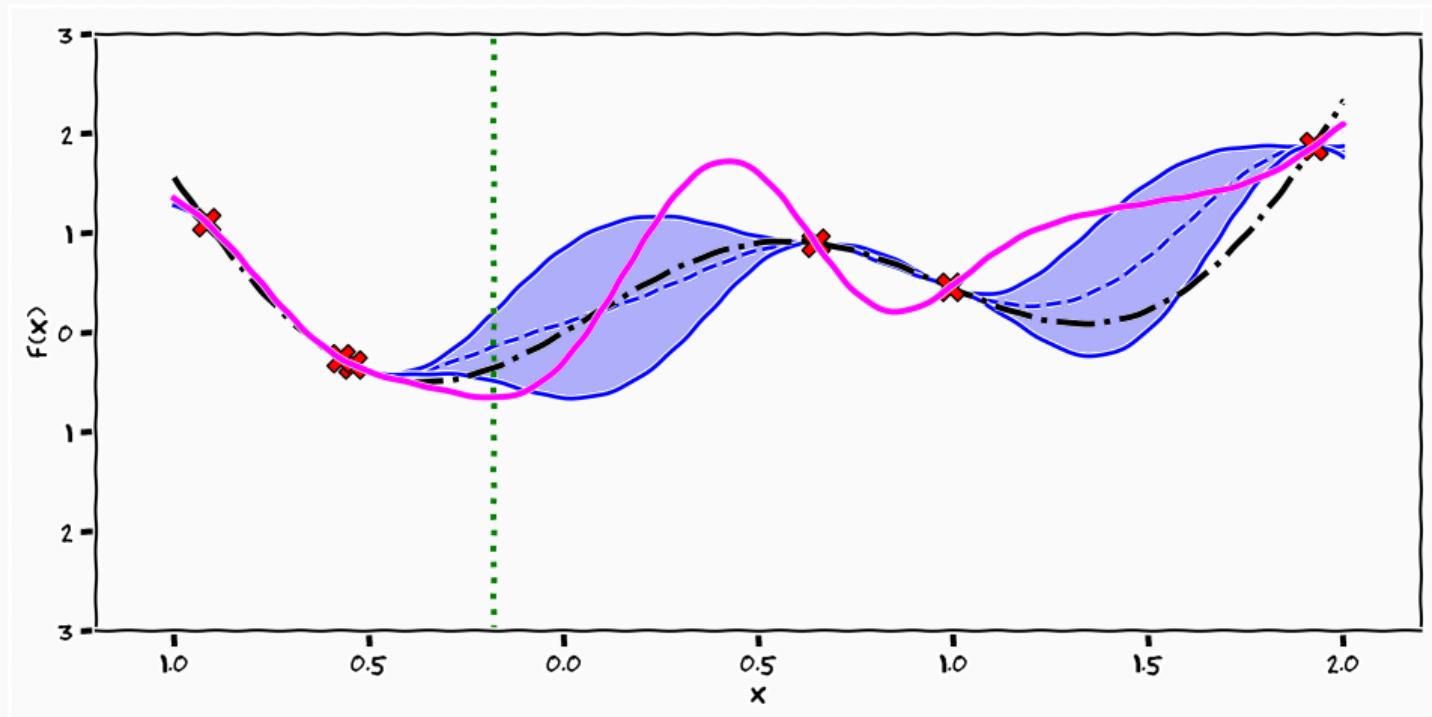
# Thompson Sampling



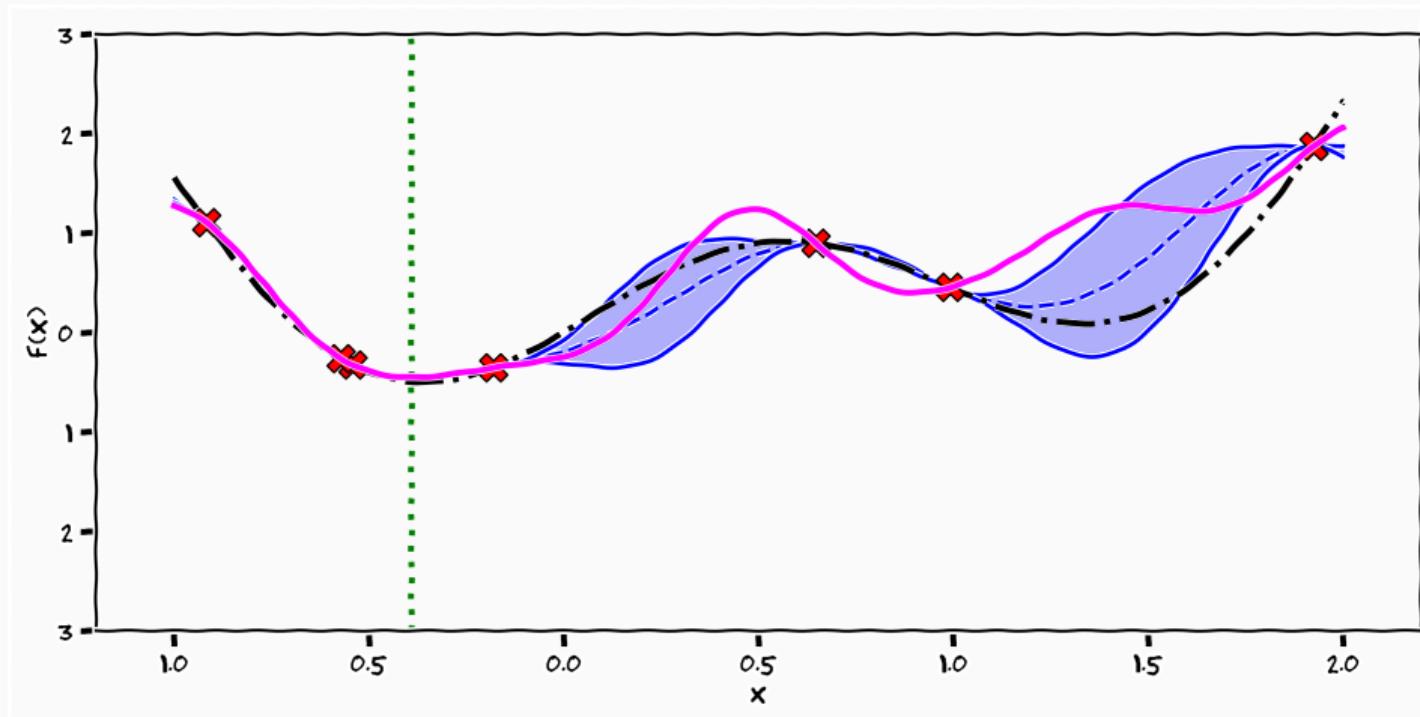
# Thompson Sampling



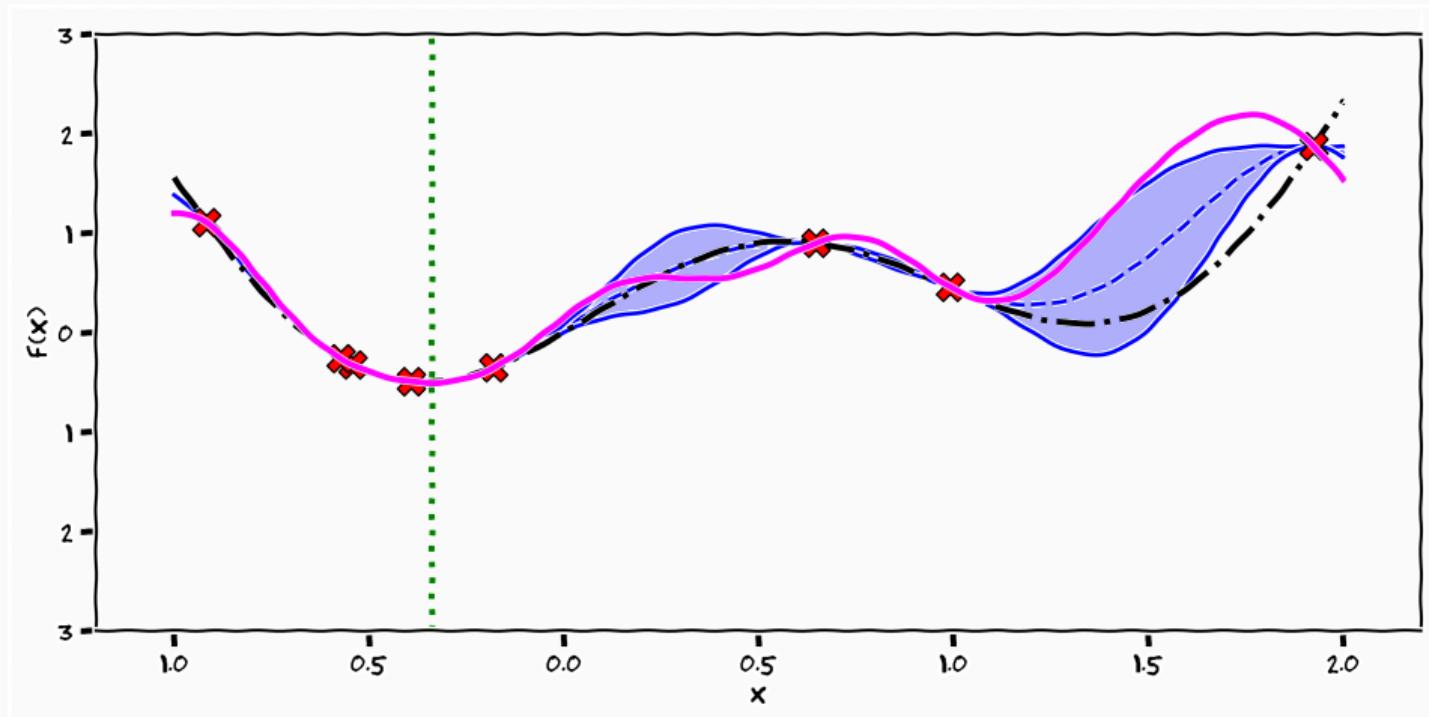
# Thompson Sampling



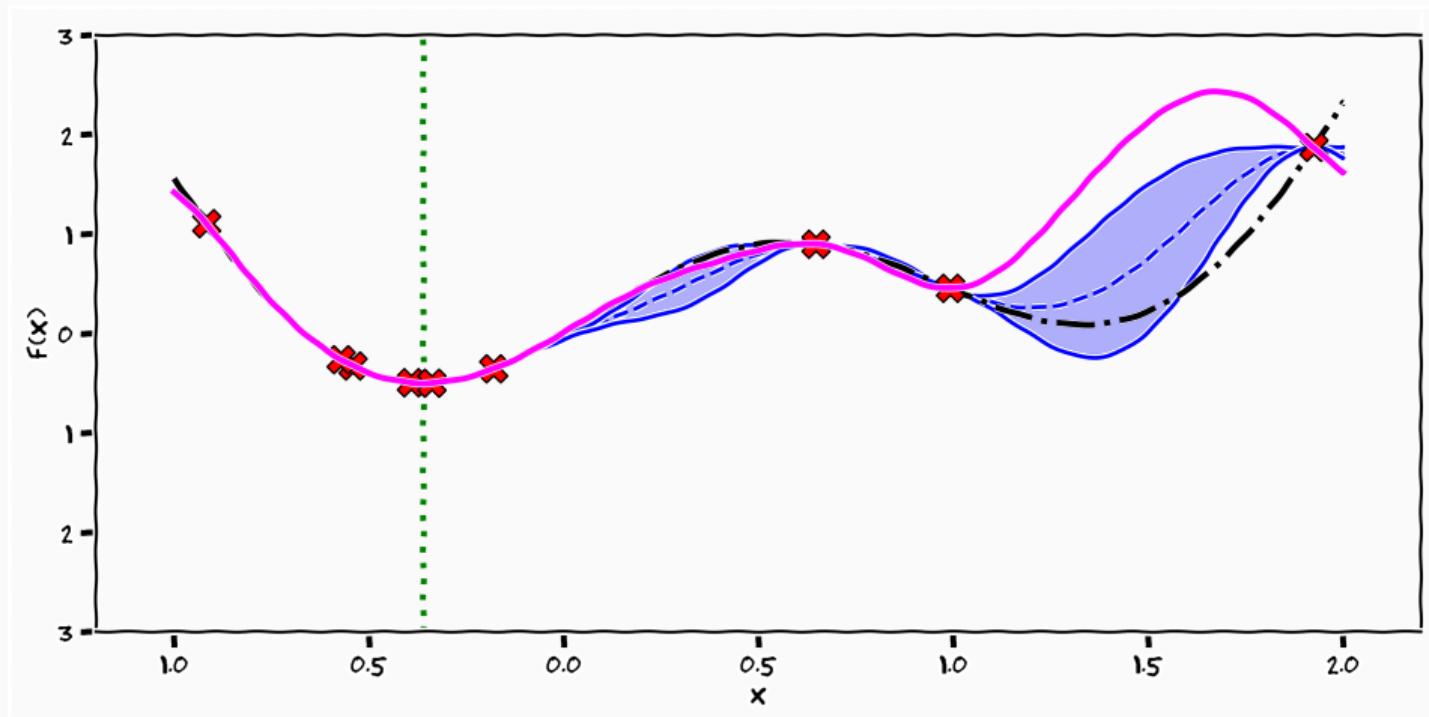
# Thompson Sampling



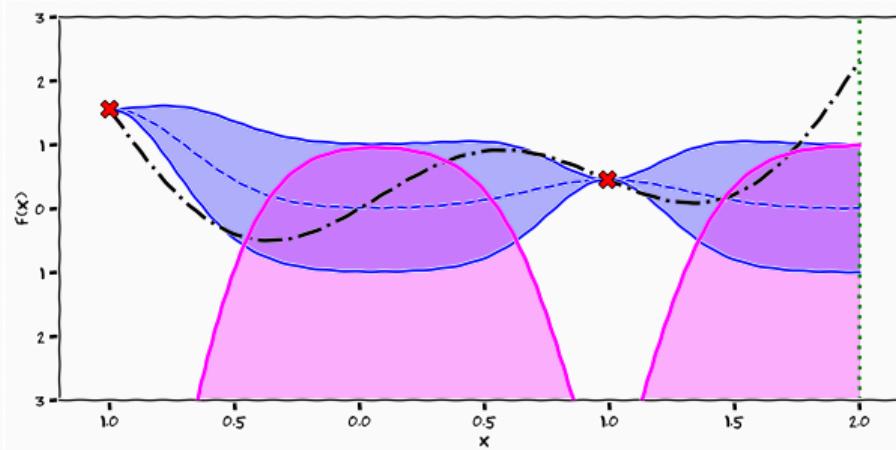
# Thompson Sampling



# Thompson Sampling



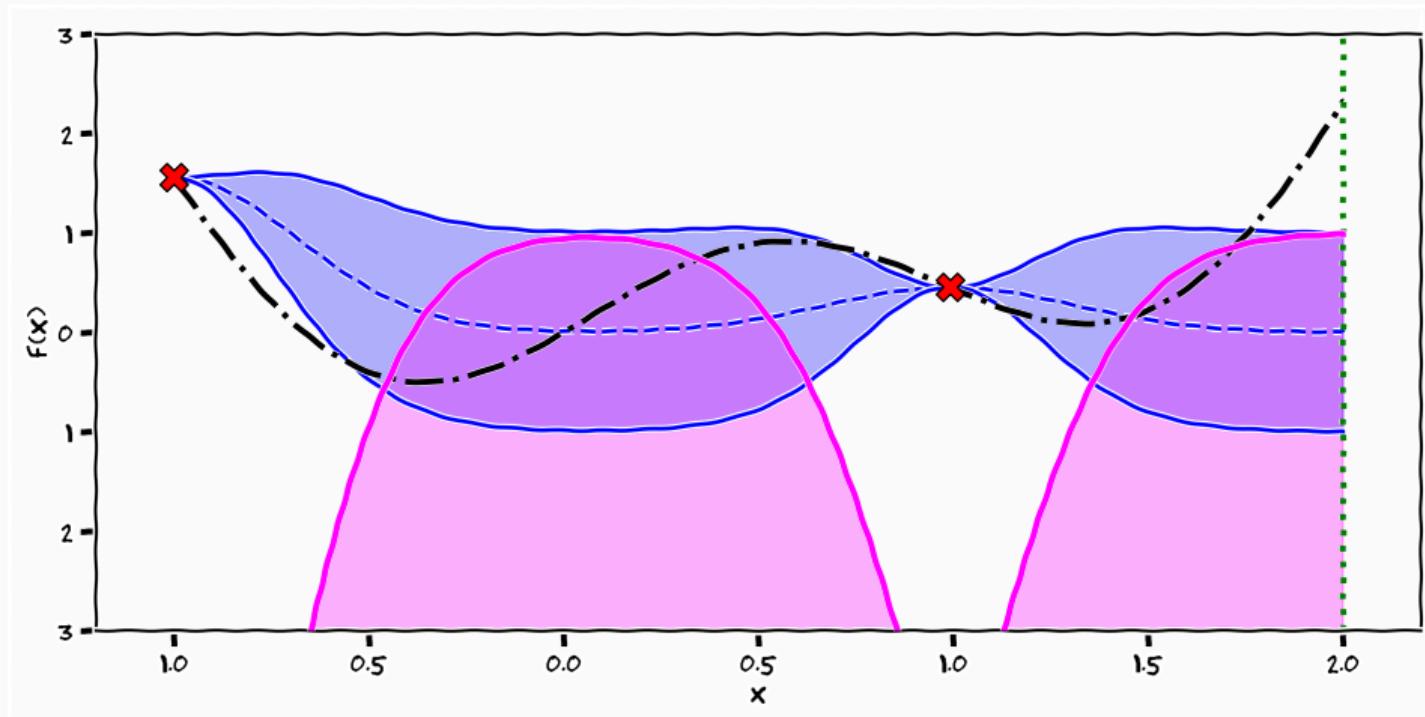
## Upper Confidence Bound [Cox et al., 1997]



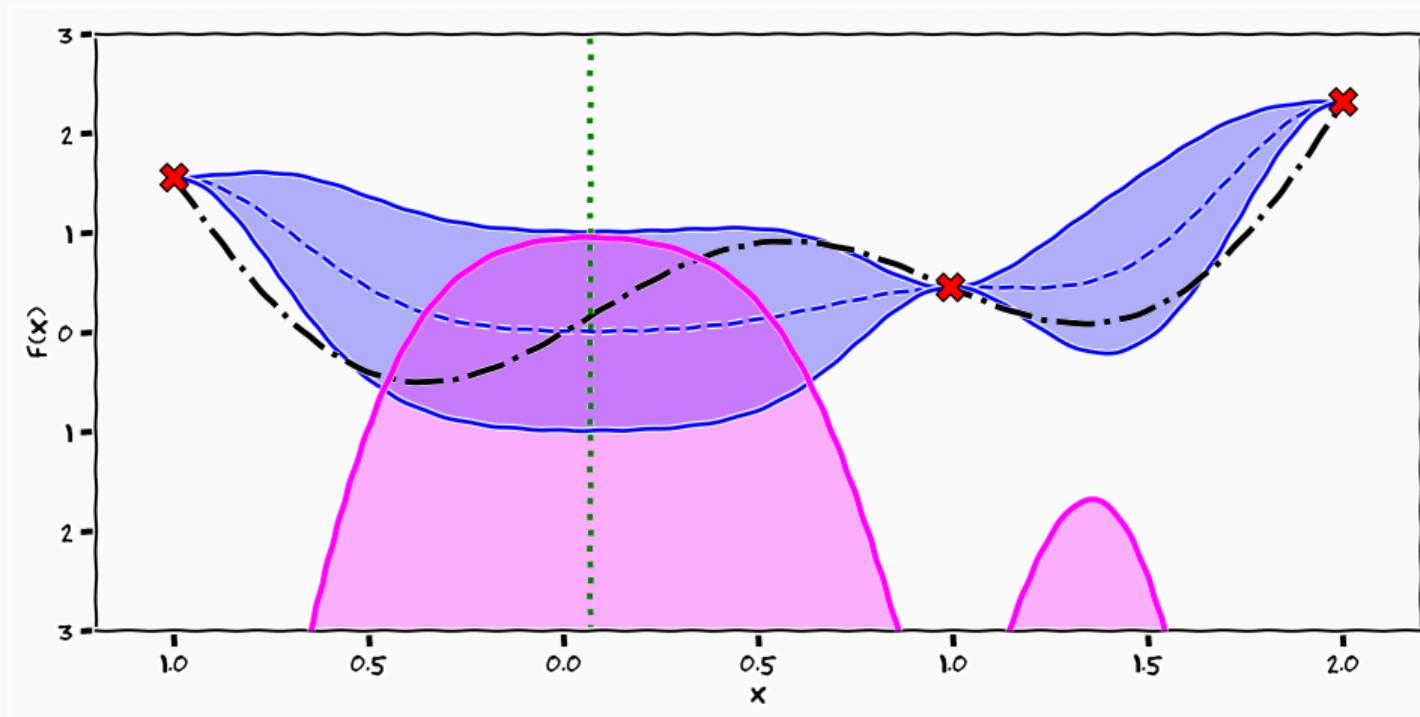
- Acquisition Function

$$\alpha(x; \{x_i, y_i\}_{i=1}^n, \mathcal{M}_n) = -\mu(x; \{x_i, y_i\}_{i=1}^n) + \beta\sigma(x; \{x_i, y_i\}_{i=1}^n)$$

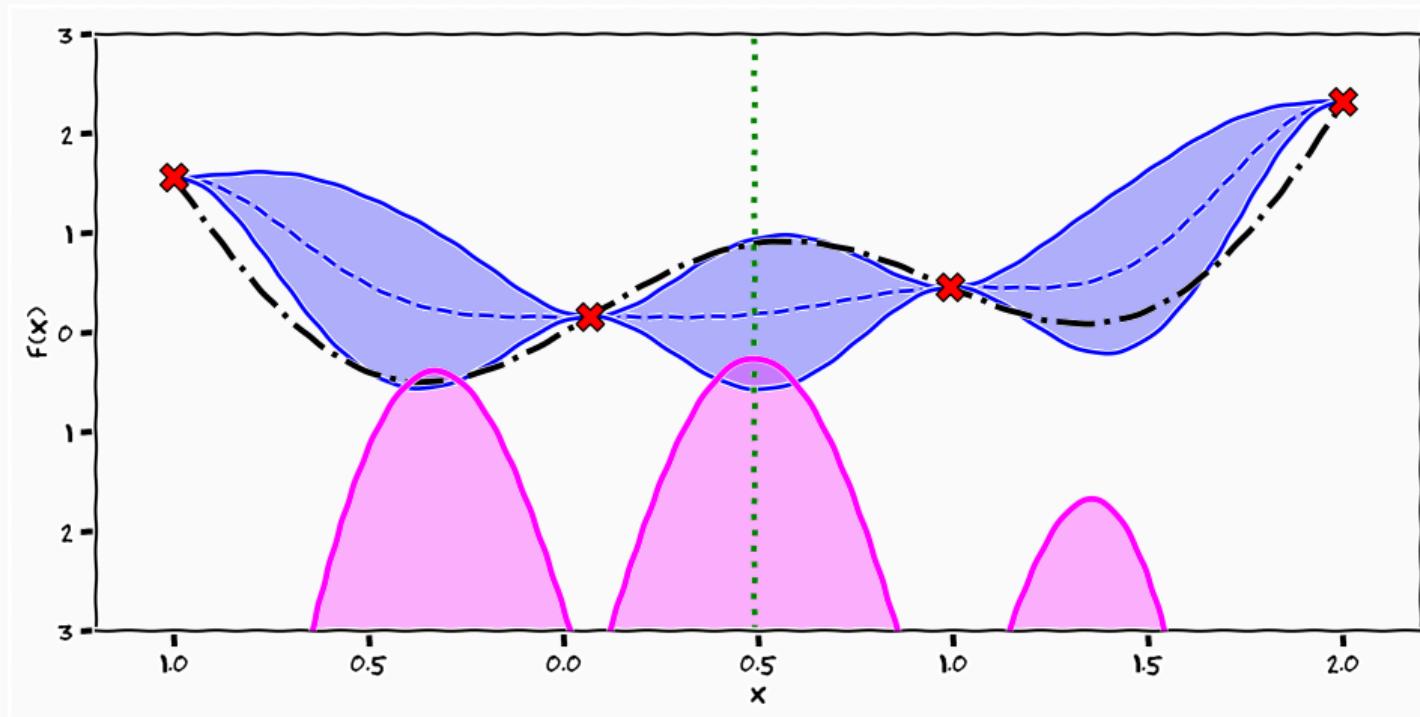
## Upper Confidence Bound



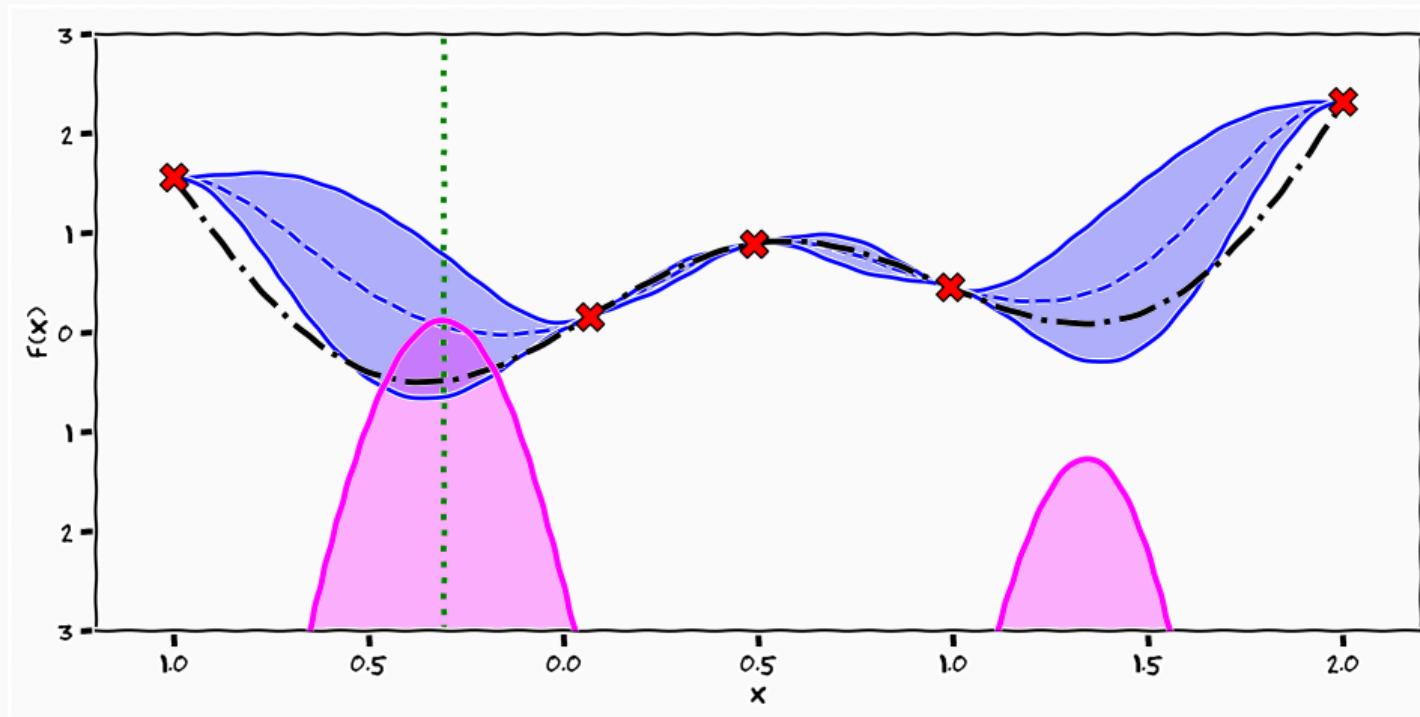
## Upper Confidence Bound



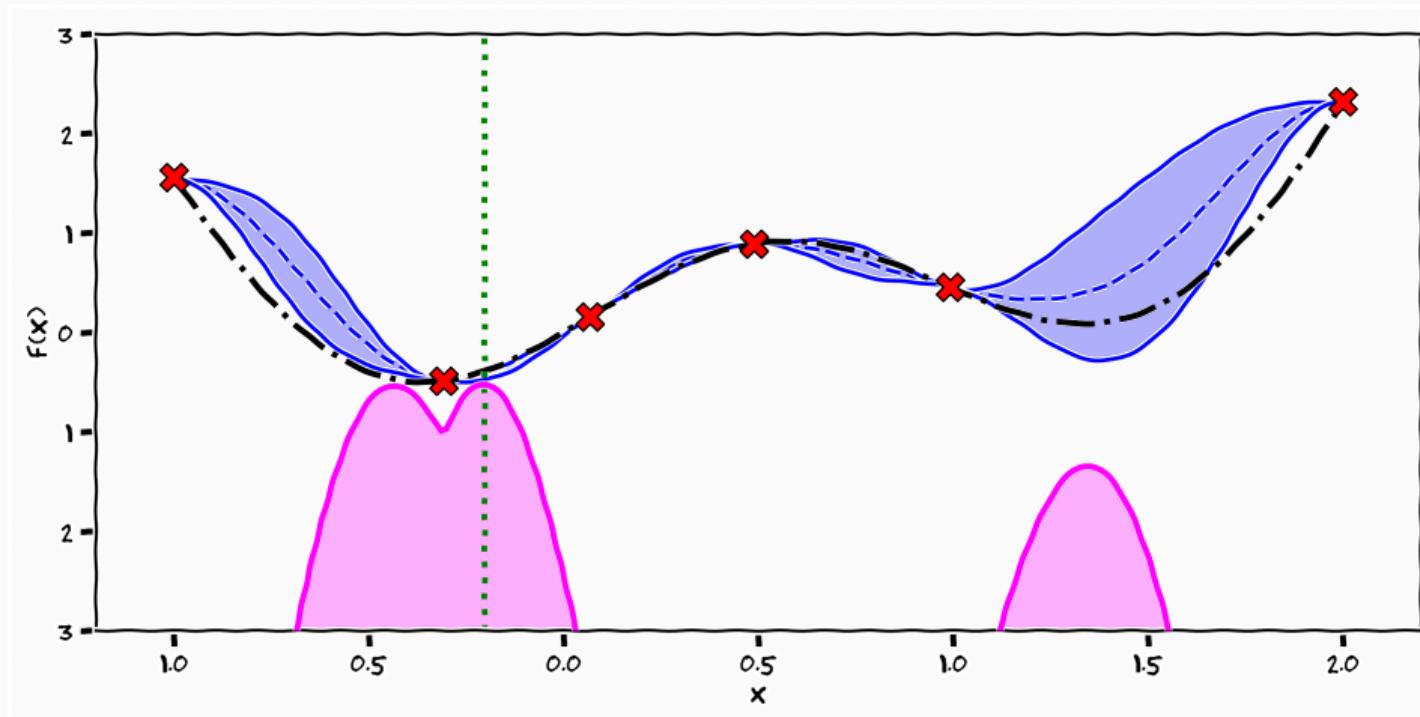
## Upper Confidence Bound



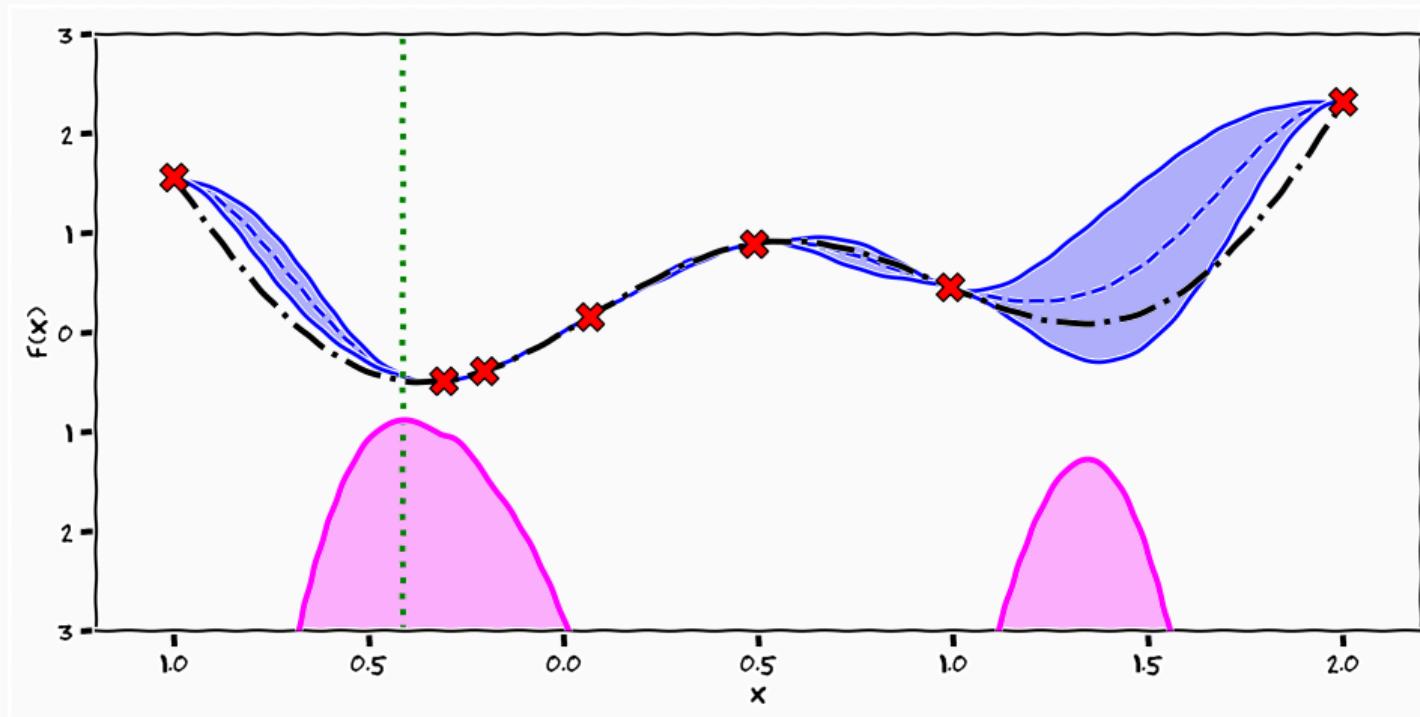
## Upper Confidence Bound



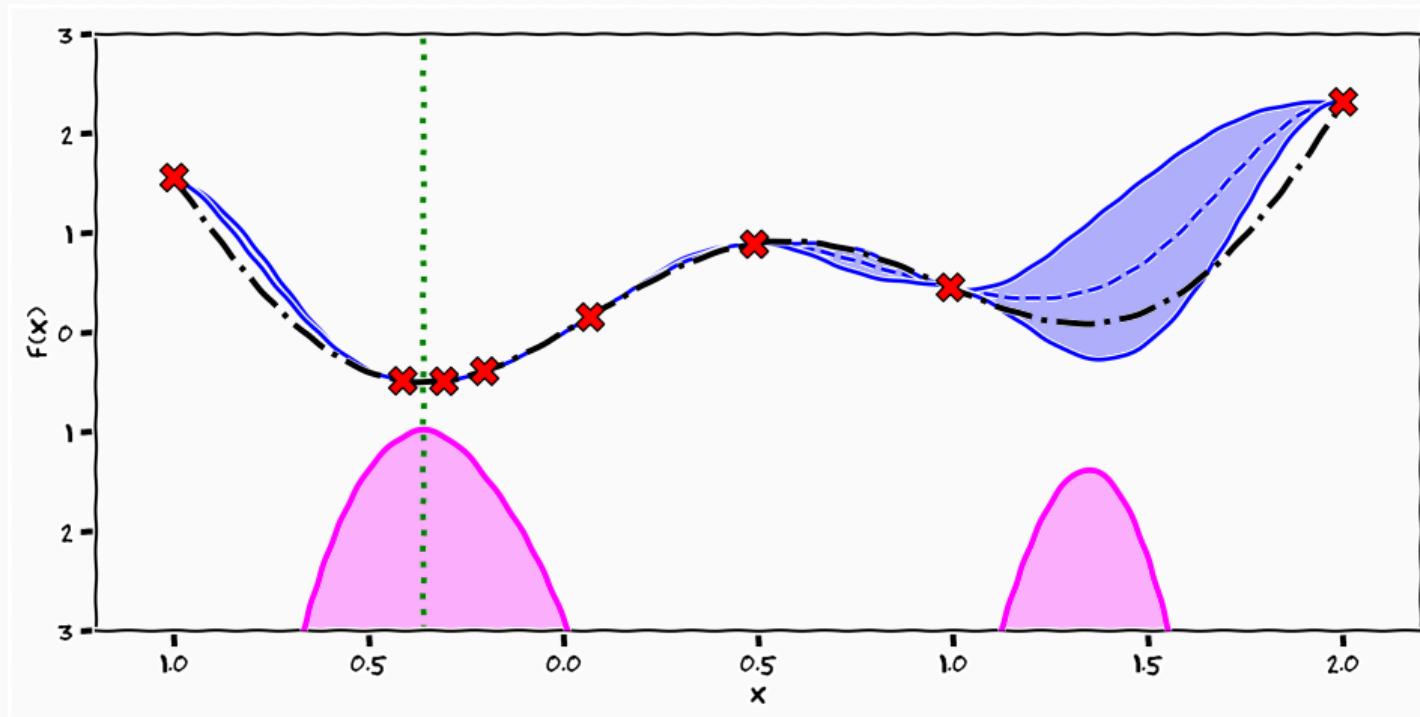
## Upper Confidence Bound



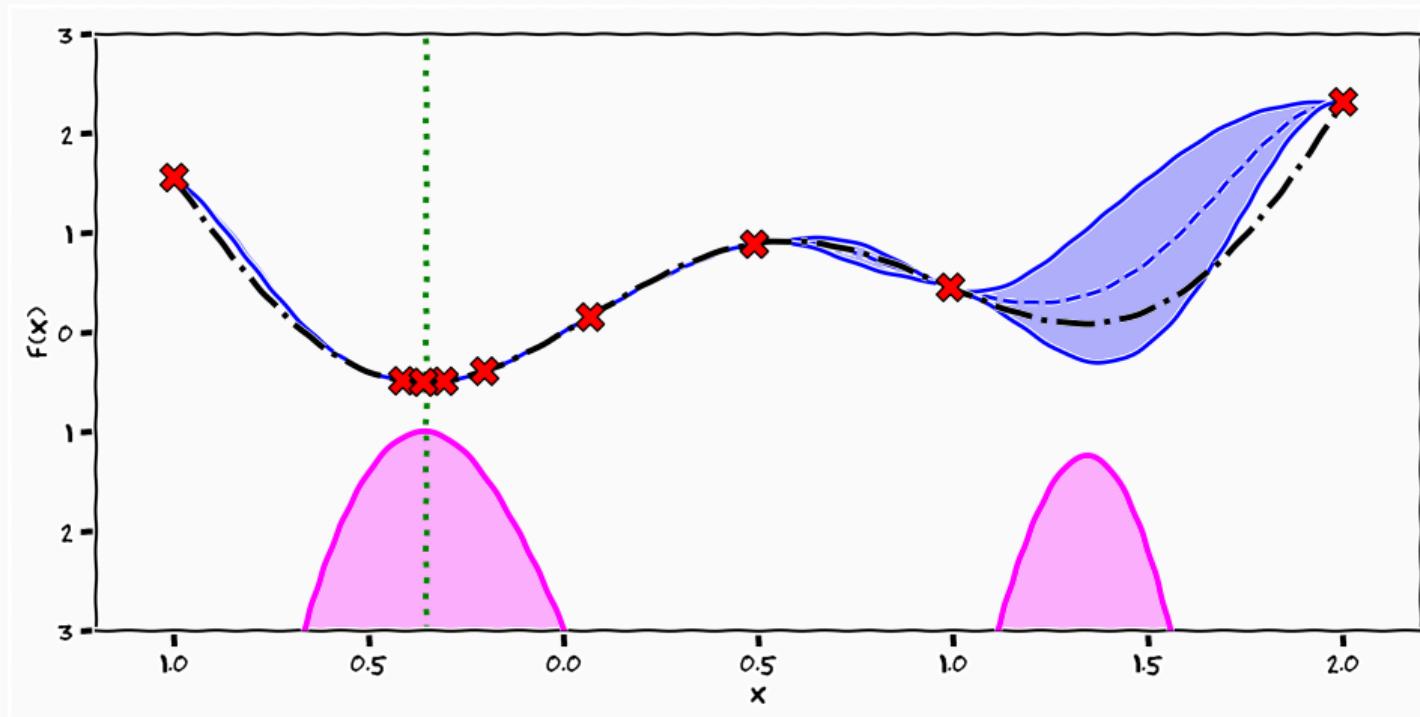
## Upper Confidence Bound



## Upper Confidence Bound



## Upper Confidence Bound



- We can come up with lots of heuristics of how to define acquisition functions

- We can come up with lots of heuristics of how to define acquisition functions
- Define a function that defines the **utility** of observing each location

$$u(x, f(x^{(*)}), \mathcal{M}_n)$$

- We can come up with lots of heuristics of how to define acquisition functions
- Define a function that defines the **utility** of observing each location

$$u(x, f(x^{(*)}), \mathcal{M}_n)$$

- Define the acquisition function as the expected utility

$$\begin{aligned}\alpha(x; \{x_i, y_i\}_{i=1}^n, \mathcal{M}_n) &= \mathbb{E}_{p(f)}[u(x)] \\ &= \int u(x, f(x^{(*)}), \mathcal{M}_n) p(f \mid \{x_i, y_i\}_{i=1}^n) df\end{aligned}$$

## Probability of Improvement [Kushner, 1963]

---

- Utility Function

$$u(x) = \begin{cases} 0 & f(x) > f(x^{(*)}) \\ 1 & f(x) \leq f(x^{(*)}) \end{cases}$$

## Probability of Improvement [Kushner, 1963]

---

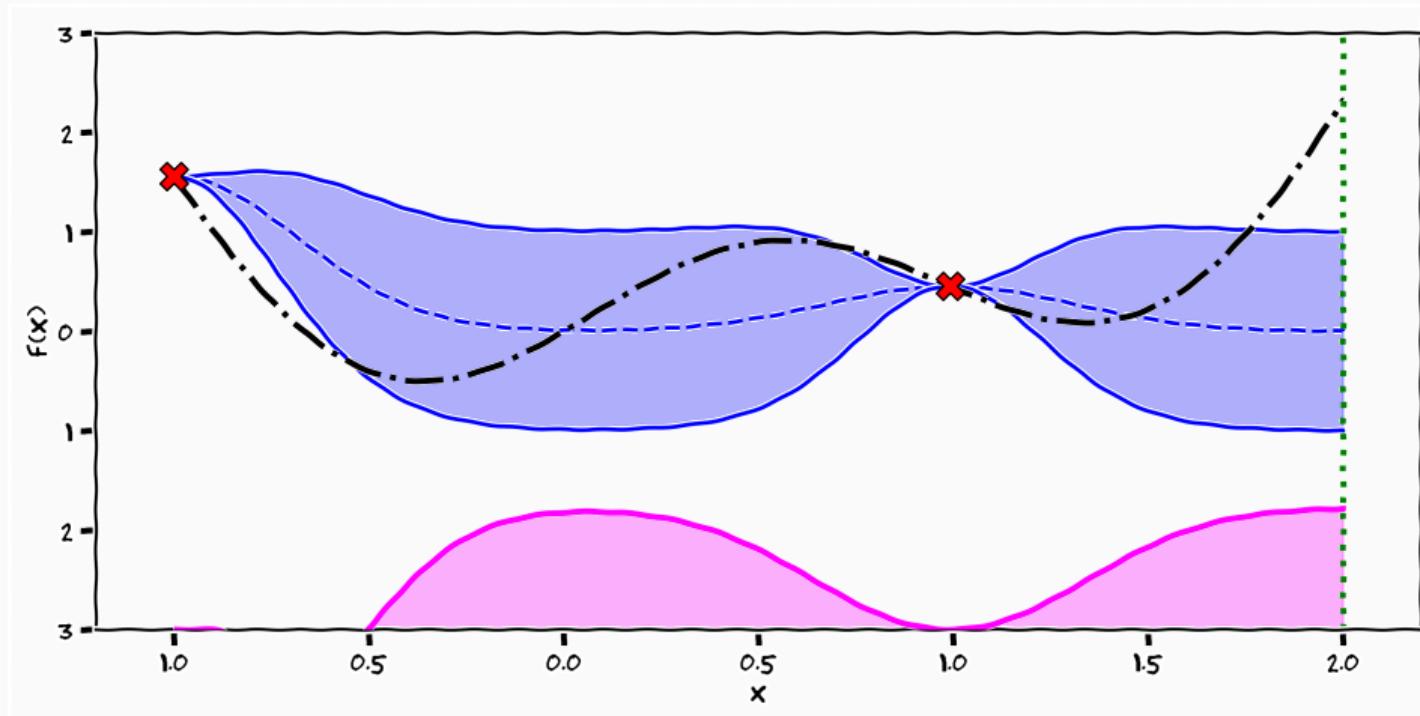
- Utility Function

$$u(x) = \begin{cases} 0 & f(x) > f(x^{(*)}) \\ 1 & f(x) \leq f(x^{(*)}) \end{cases}$$

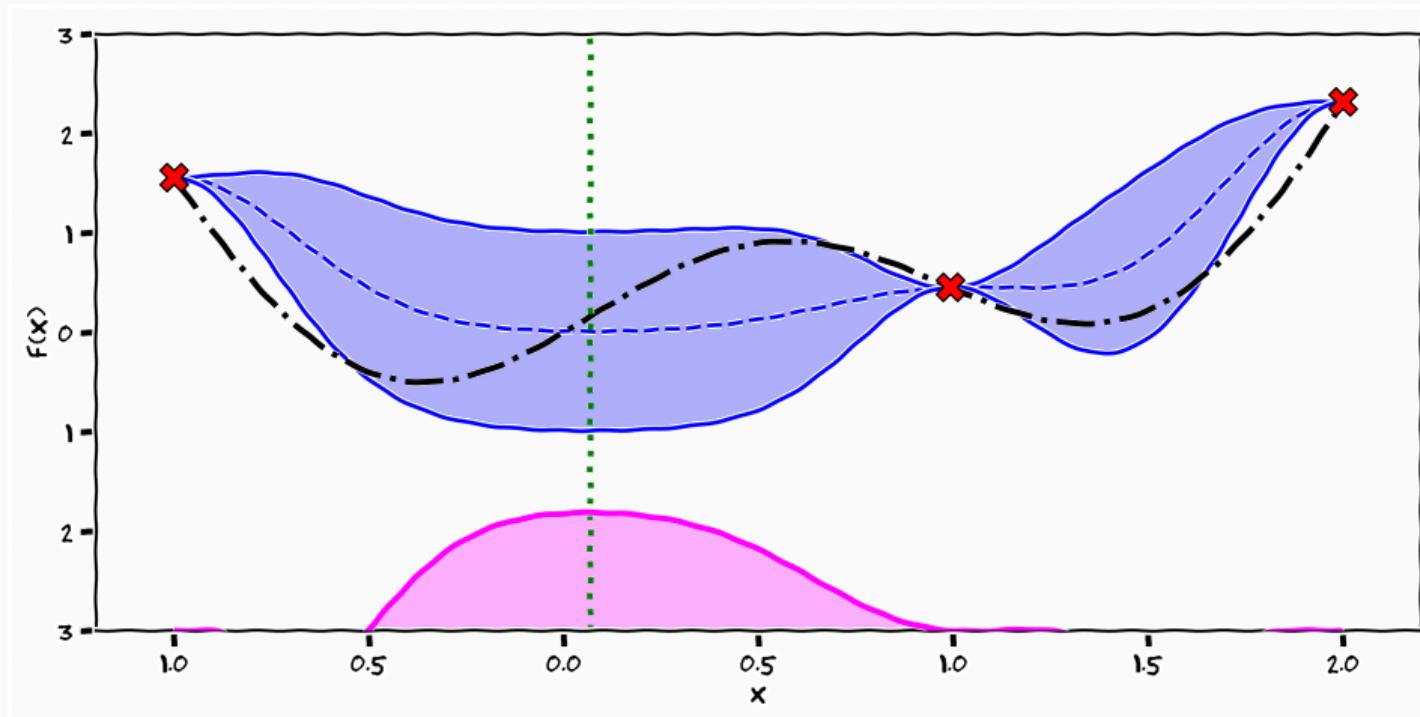
- Acquisition Function

$$\begin{aligned}\alpha(x; \{x_i, y_i\}_{i=1}^n, f(x^{(*)}), \mathcal{M}_n) &= \mathbb{E}[u(x)] = p(f(x) \leq f(x^{(*)})) \\ &= \int_{-\infty}^{f(x^{(*)})} \mathcal{N}(f \mid \mu(x), K(x, x)) df \\ &= \Phi(f(x^{(*)}) \mid \mu(x), K(x, x))\end{aligned}$$

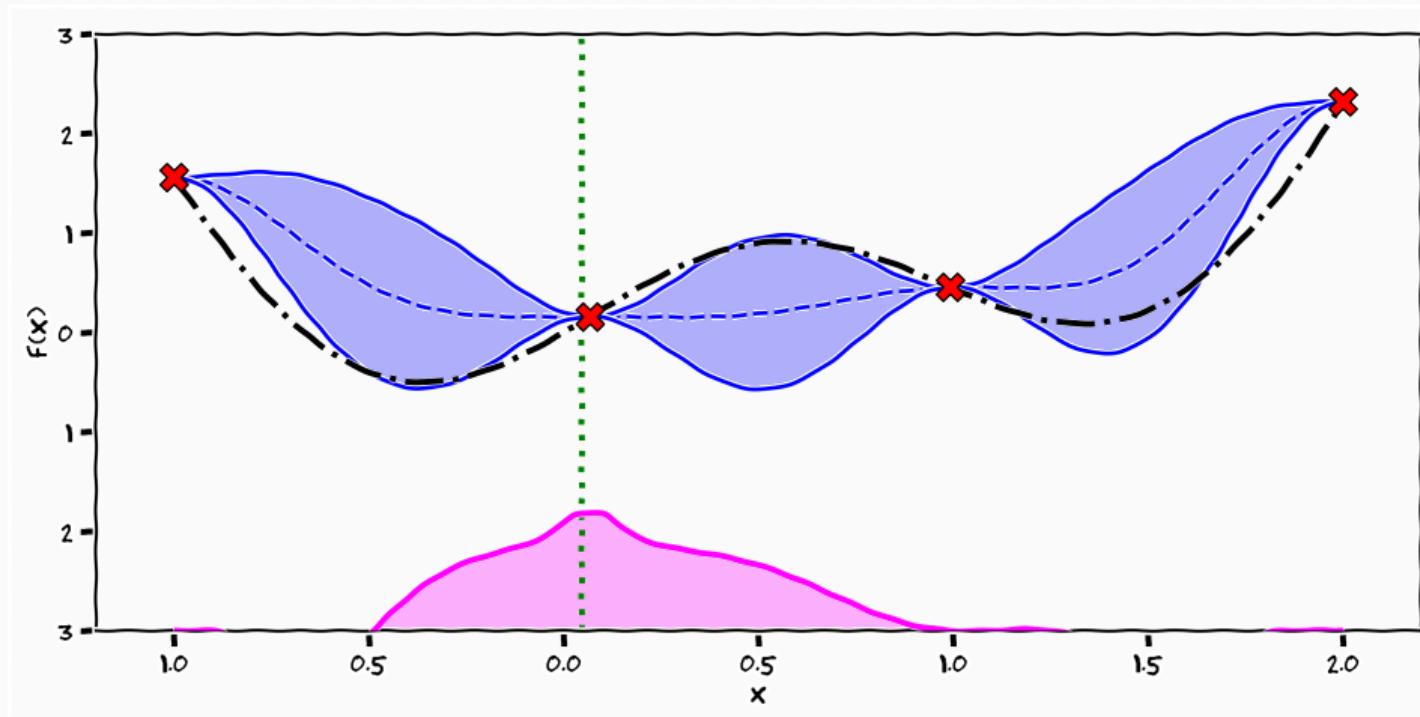
## Probability of Improvement



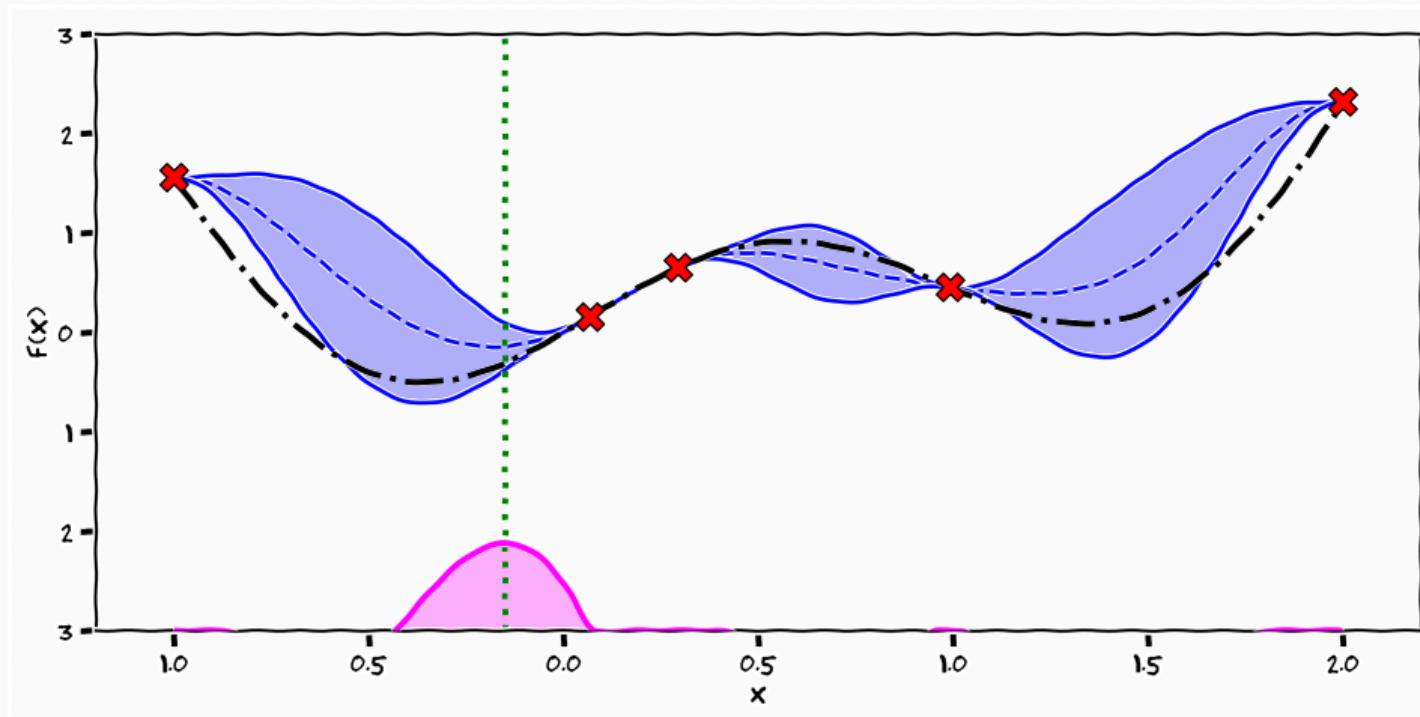
## Probability of Improvement



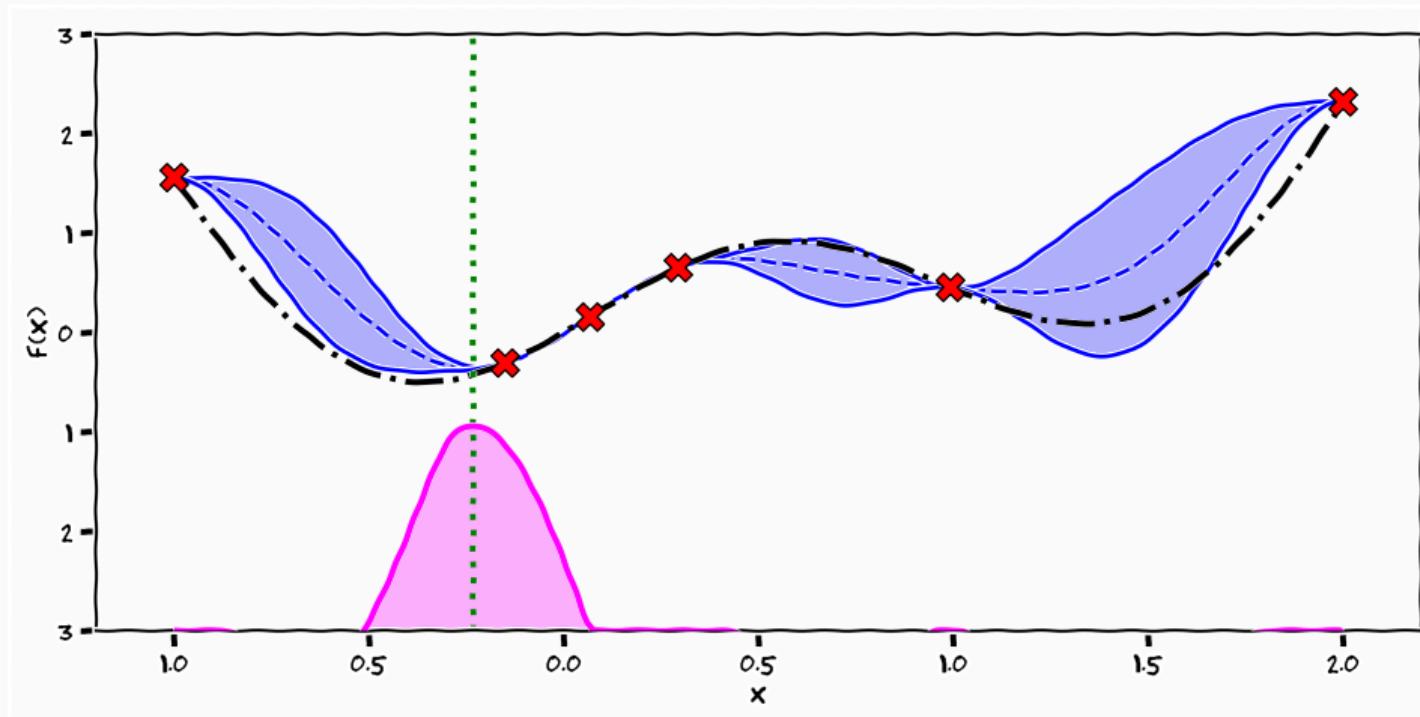
## Probability of Improvement



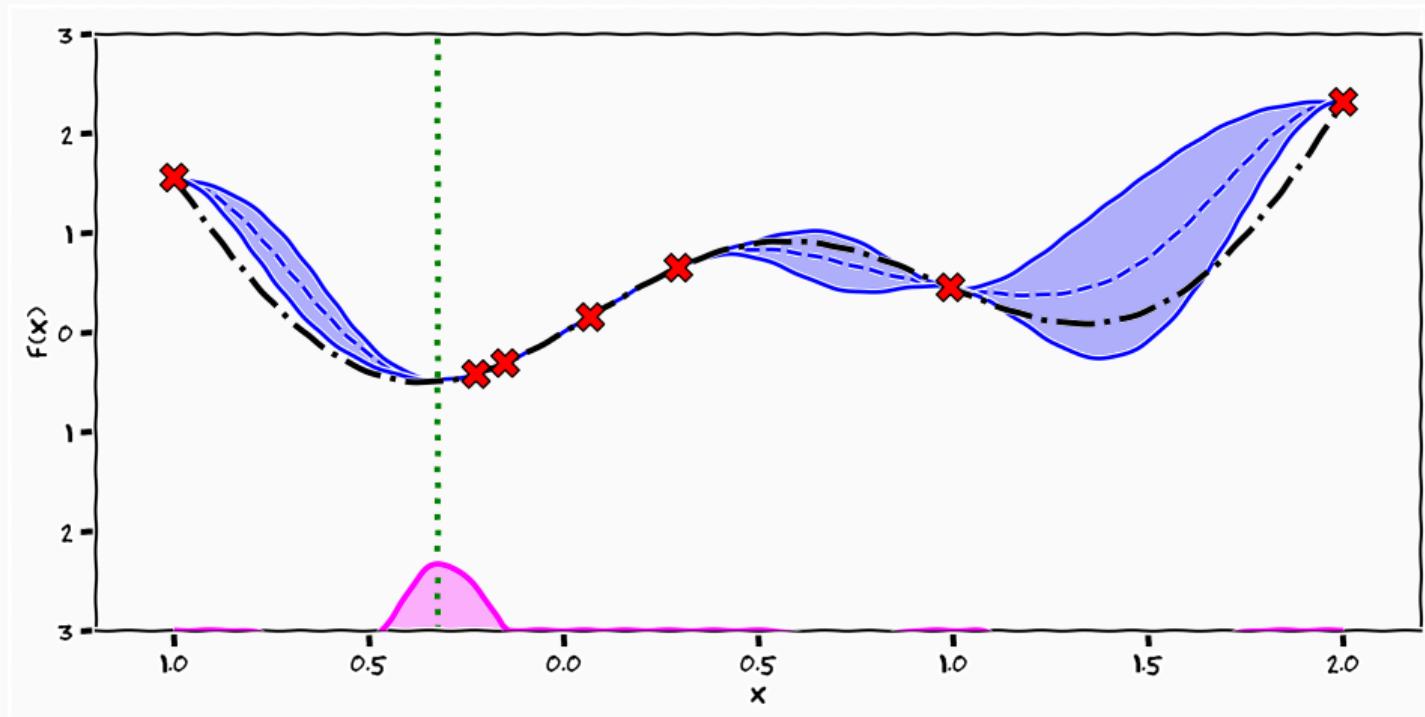
## Probability of Improvement



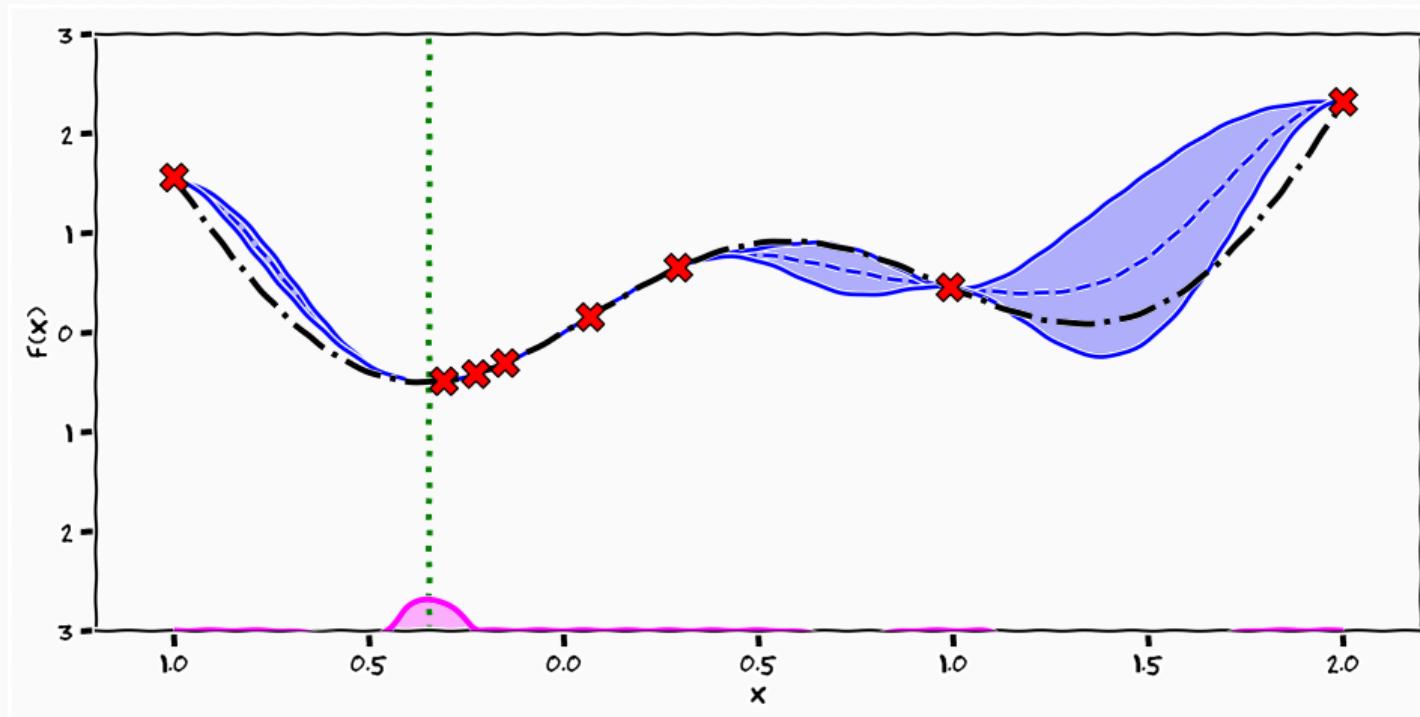
# Probability of Improvement



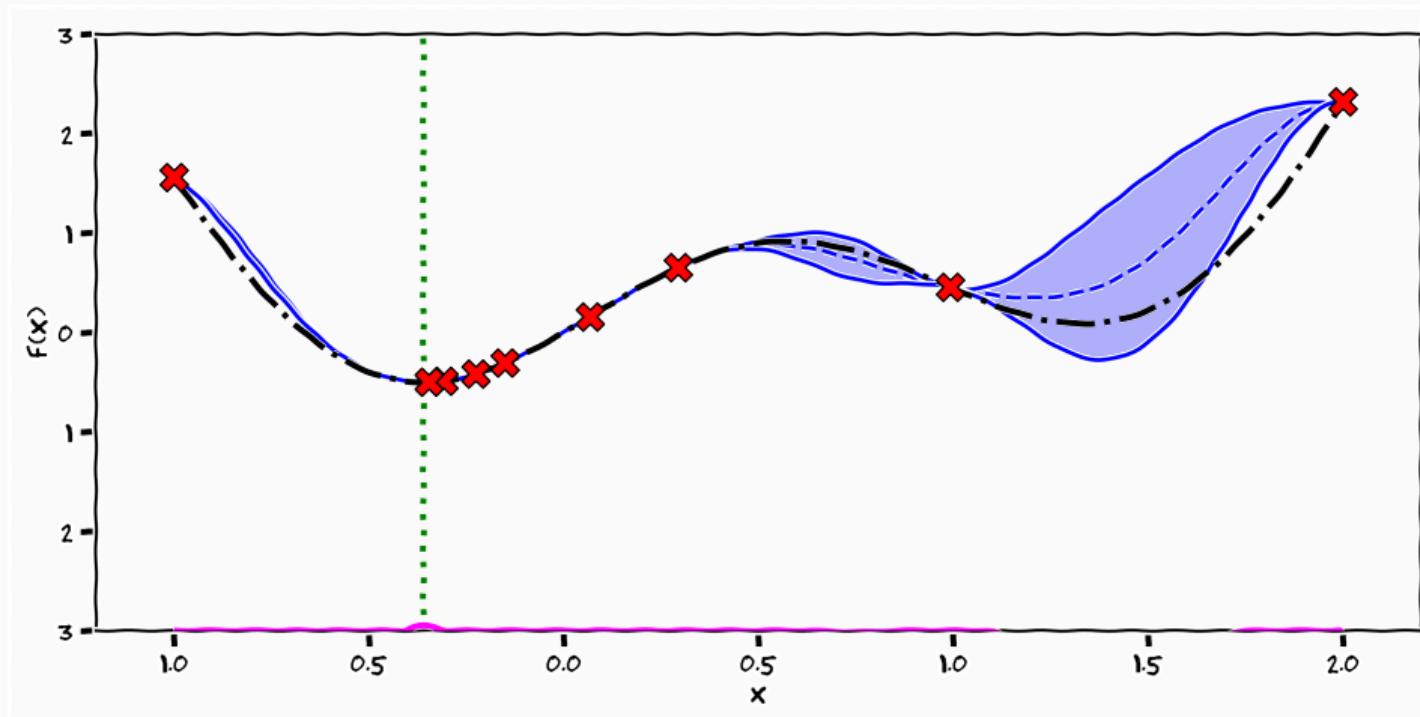
## Probability of Improvement



## Probability of Improvement



## Probability of Improvement



- Utility Function

$$u(x) = \max(0, f(x^{(*)}) - f(x))$$

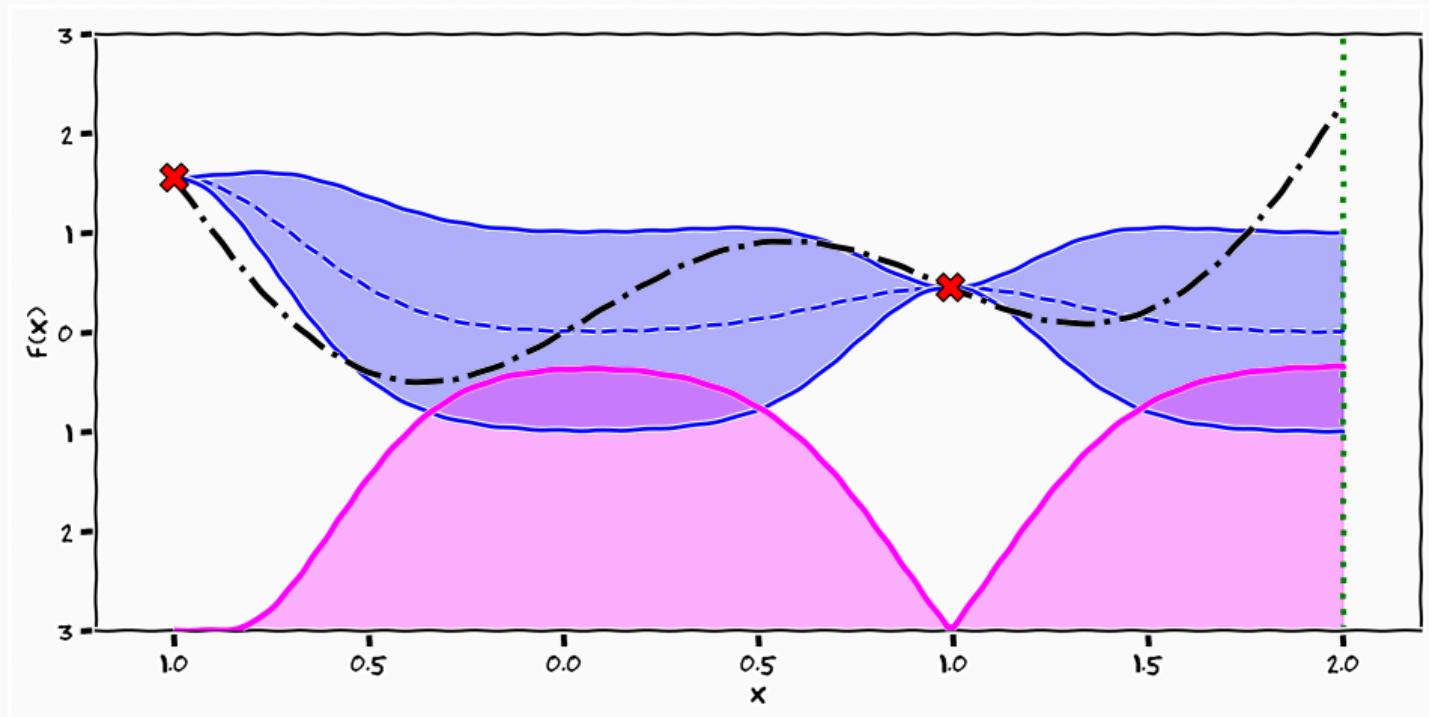
- Utility Function

$$u(x) = \max(0, f(x^{(*)}) - f(x))$$

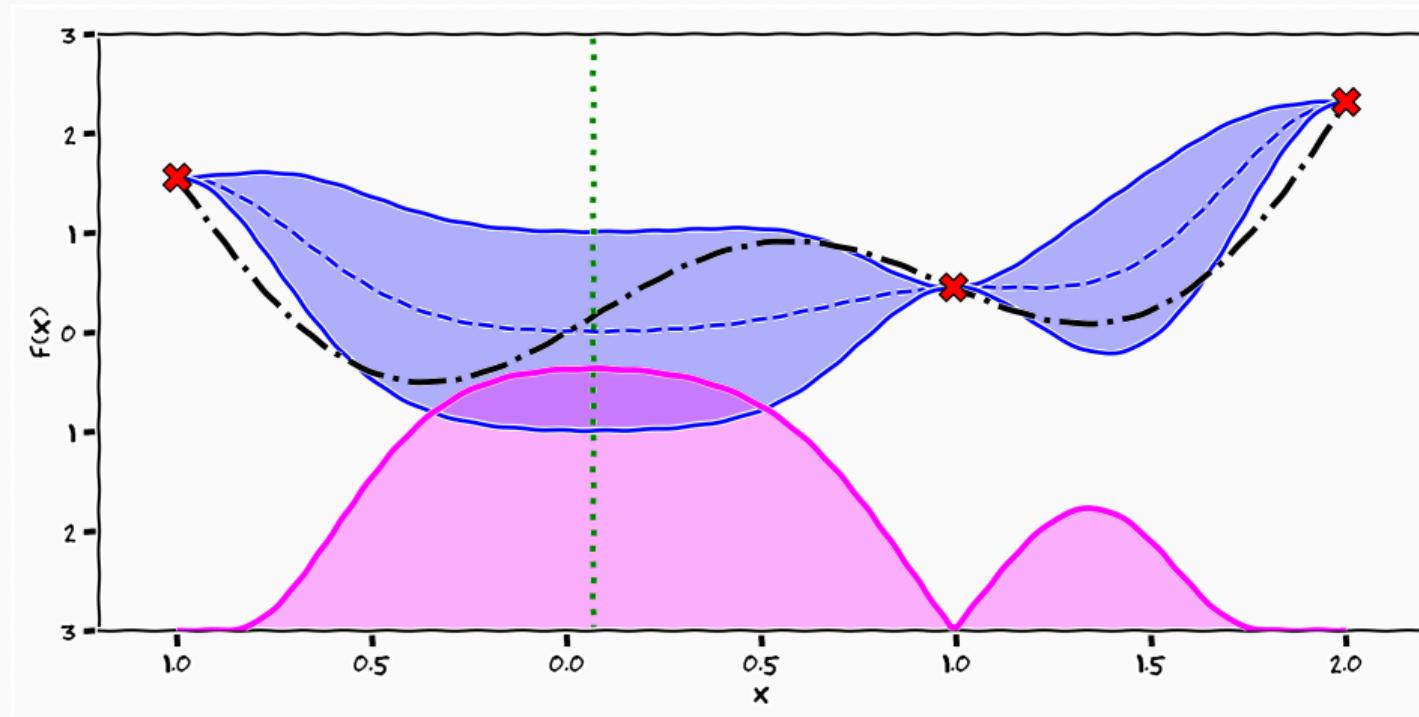
- Acquisition Function

$$\begin{aligned}\alpha(x; \{x_i, y_i\}_{i=1}^n, f(x^{(*)}), \mathcal{M}_n) &= \mathbb{E}[u(x)] \\ &= \int_{-\infty}^{f(x^{(*)})} (f(x^{(*)}) - f) \mathcal{N}(f \mid \mu(x), K(x, x)) df \\ &= (f(x^{(*)}) - \mu(x)) \Phi \left( f(x^{(*)}) \mid \mu(x), K(x, x) \right) \\ &\quad + K(x, x) \mathcal{N} \left( f(x^{(*)}) \mid \mu(x), K(x, x) \right)\end{aligned}$$

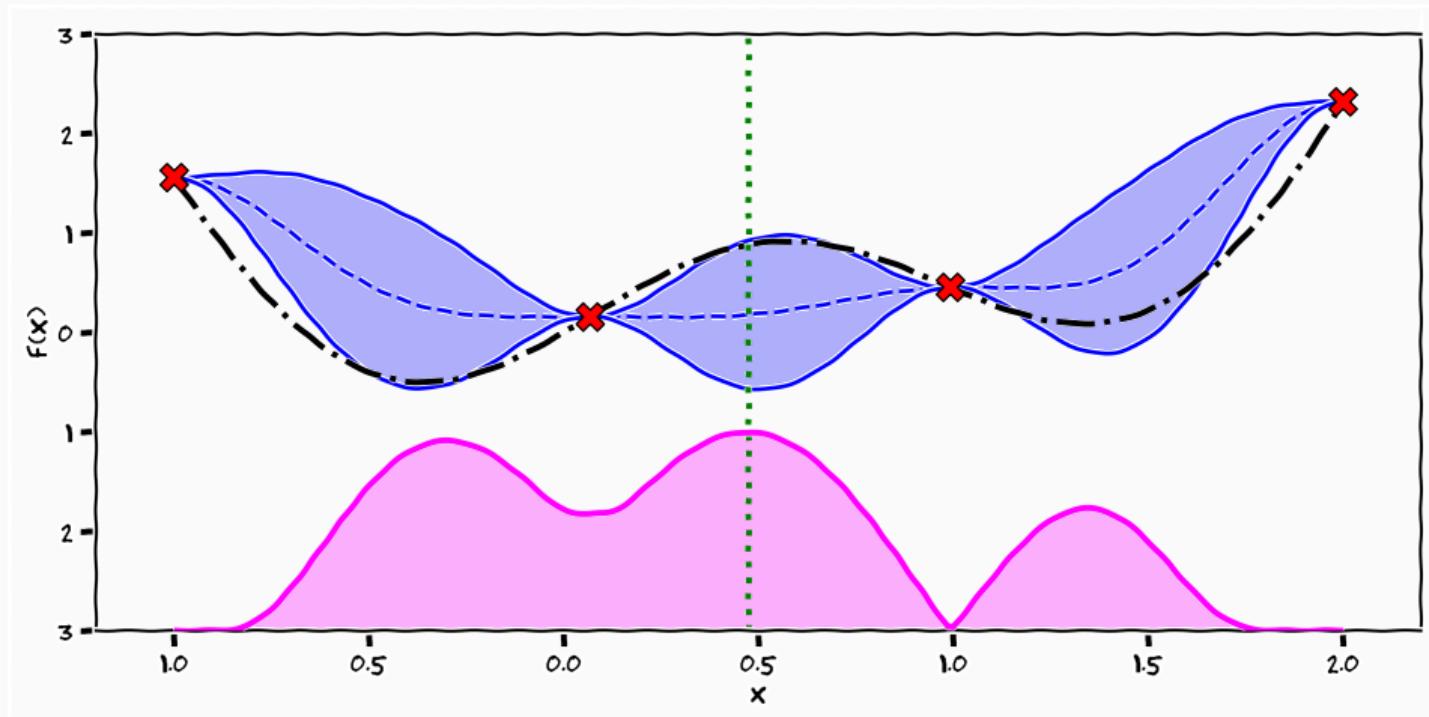
## Expected Improvement



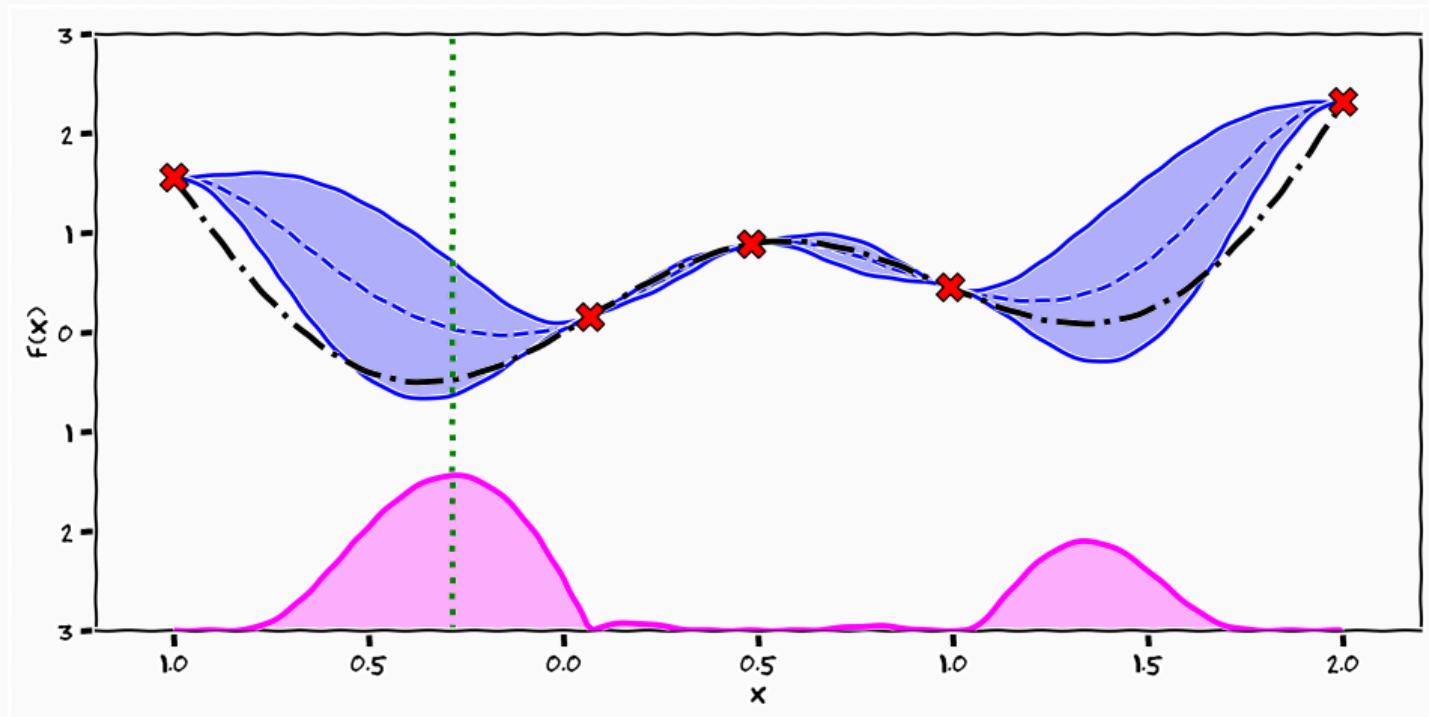
## Expected Improvement



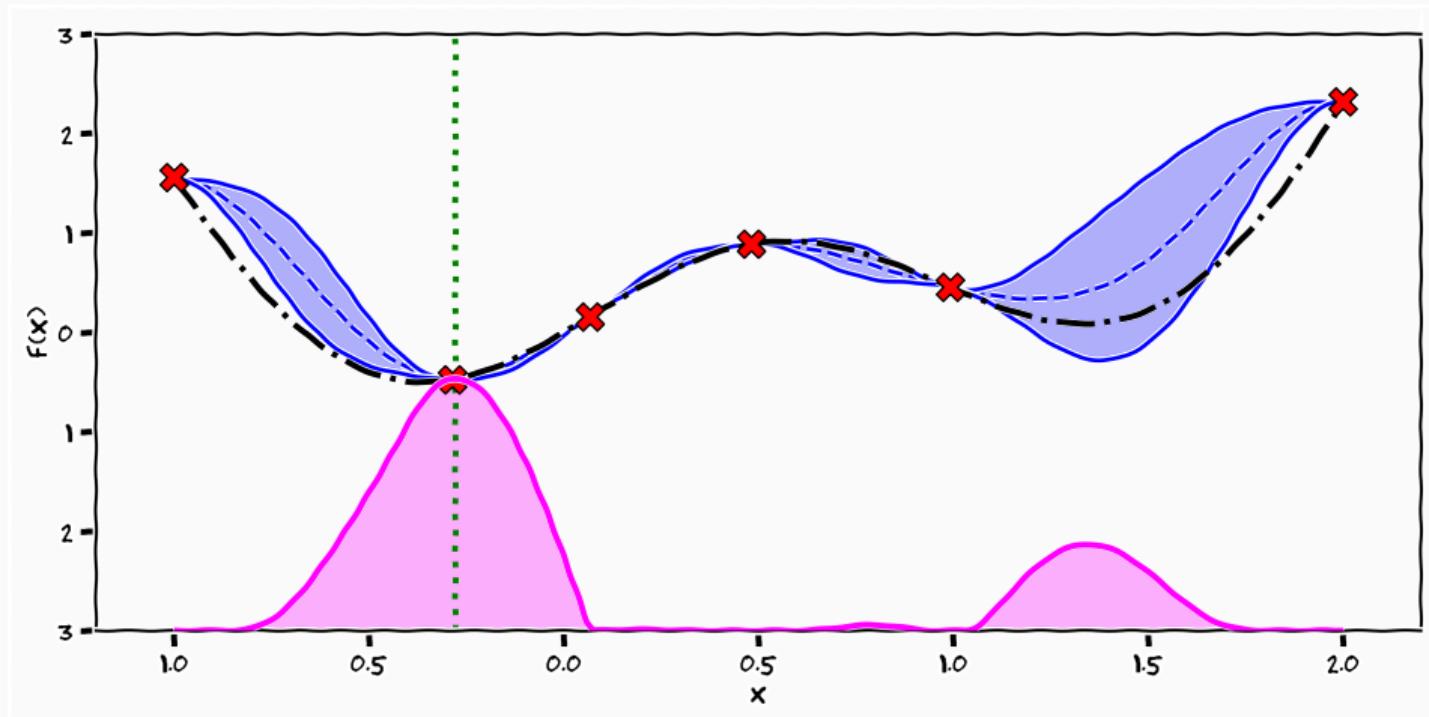
## Expected Improvement



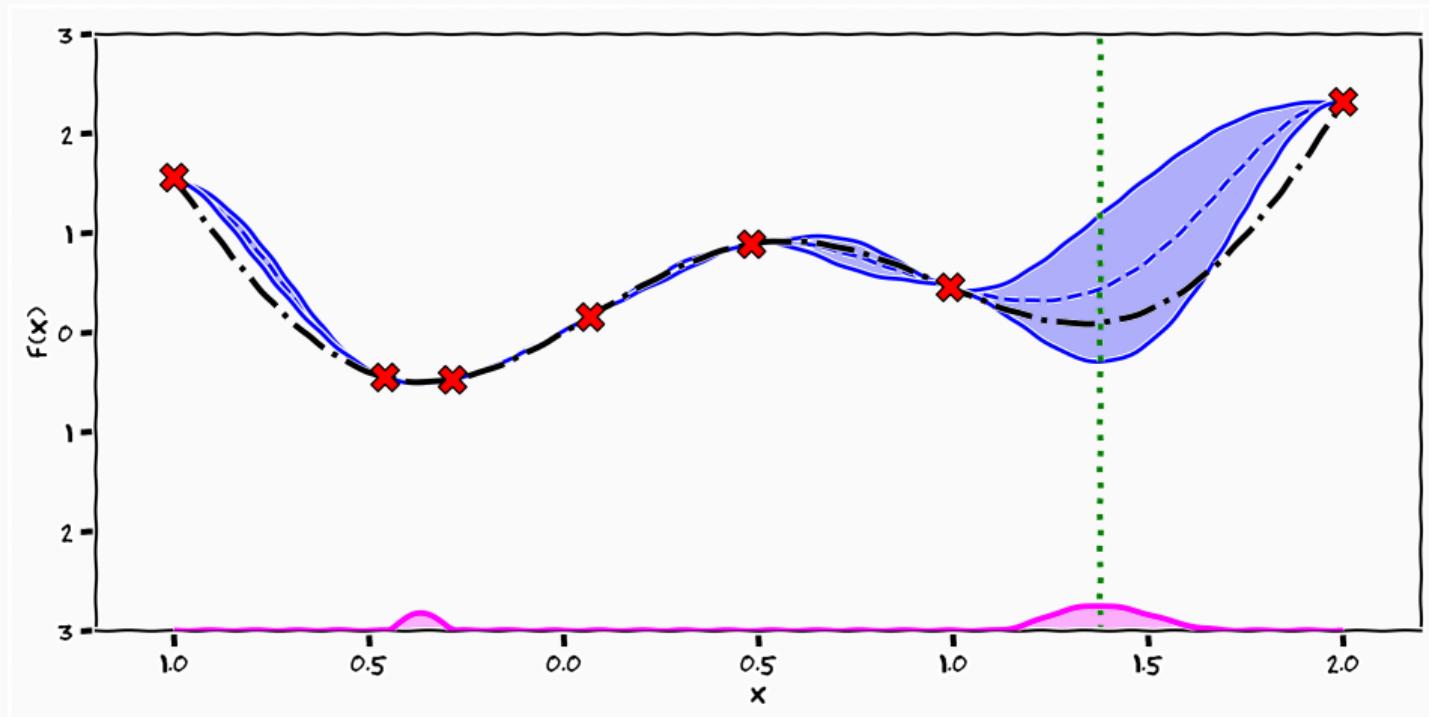
## Expected Improvement



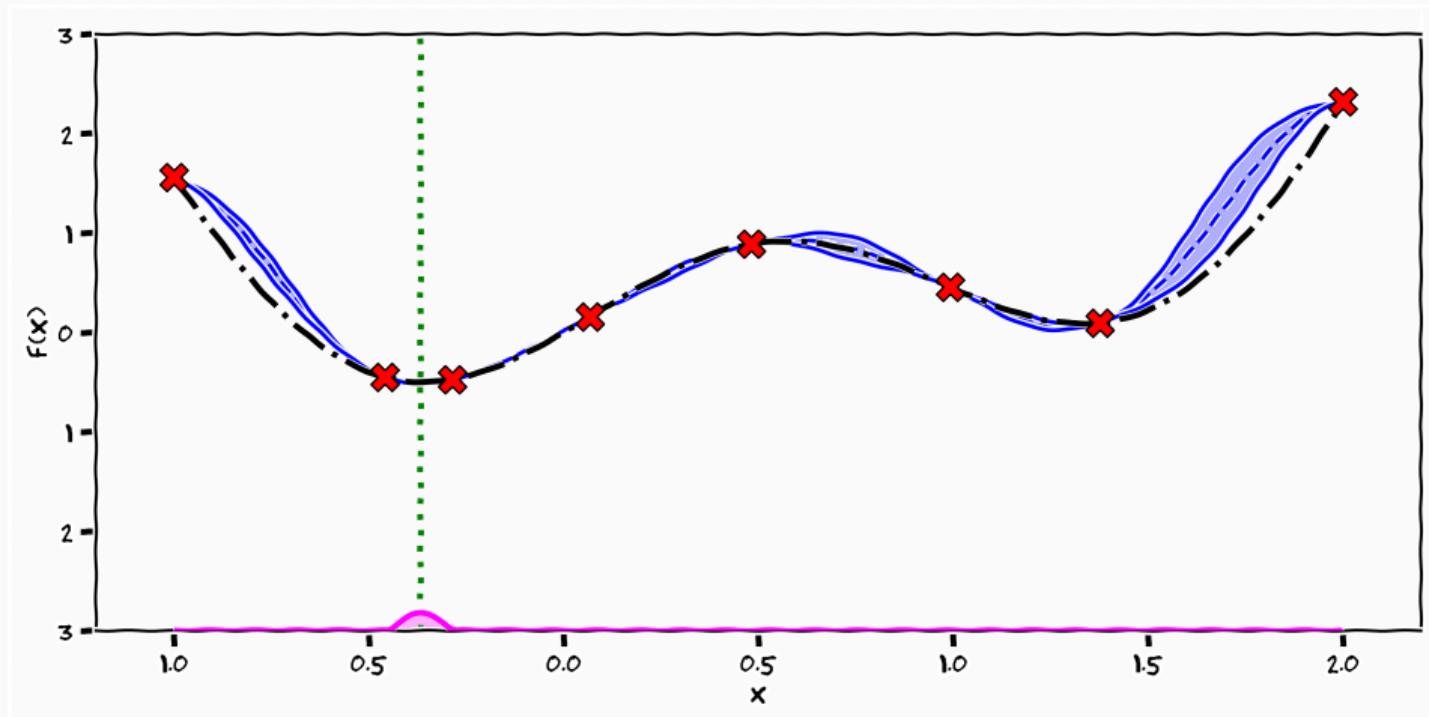
## Expected Improvement



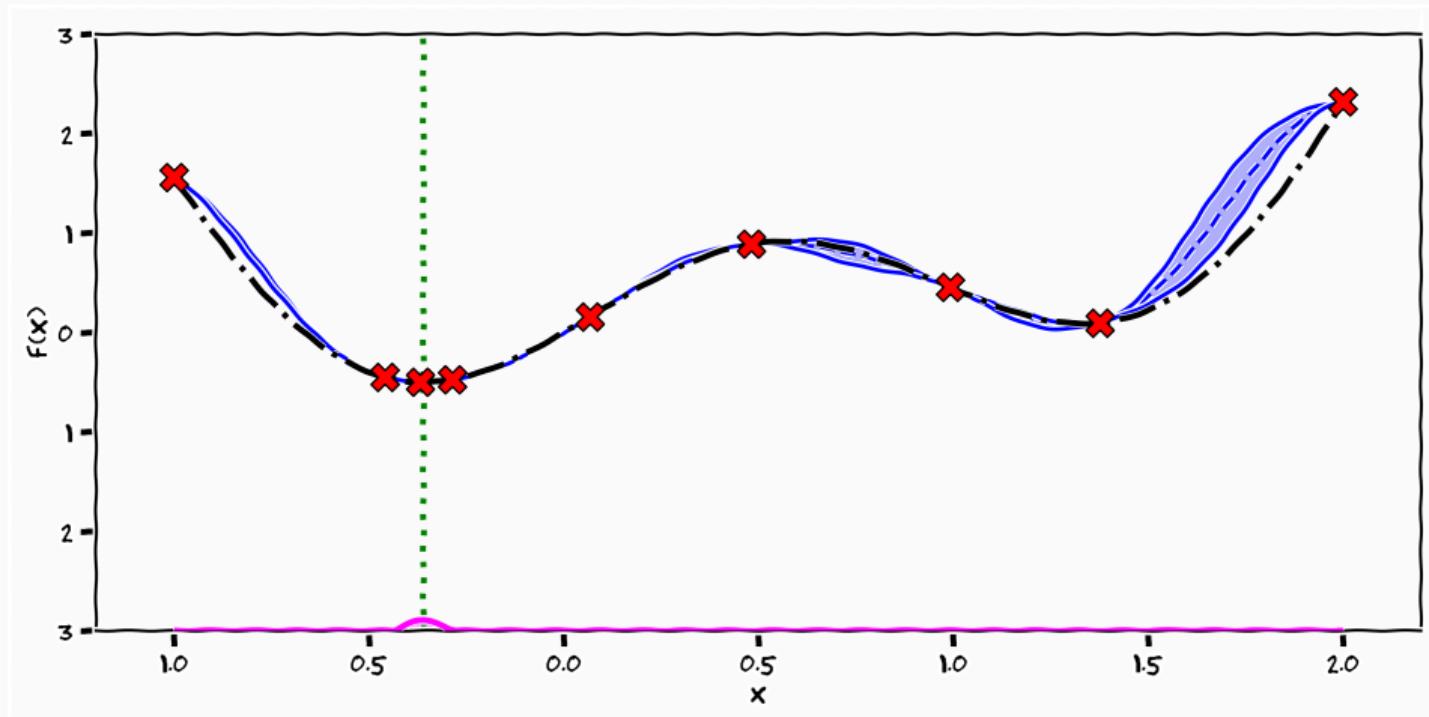
## Expected Improvement



## Expected Improvement



## Expected Improvement



Task 1 encode your knowledge about **the function** in the GP prior

---

<sup>1</sup>till they open the door to the exam.

Task 1 encode your knowledge about **the function** in the GP prior

Task 2 randomly sample some data

---

<sup>1</sup>till they open the door to the exam.

**Task 1** encode your knowledge about **the function** in the GP prior

**Task 2** randomly sample some data

**Task 3** specify your acquisition function

---

<sup>1</sup>till they open the door to the exam.

**Task 1** encode your knowledge about **the function** in the GP prior

**Task 2** randomly sample some data

**Task 3** specify your acquisition function

**Task 4** evaluate and maximise the acquisition function

---

<sup>1</sup>till they open the door to the exam.

**Task 1** encode your knowledge about **the function** in the GP prior

**Task 2** randomly sample some data

**Task 3** specify your acquisition function

**Task 4** evaluate and maximise the acquisition function

**Task 5** add new data to model and **re-estimate** hyperparameters

---

<sup>1</sup>till they open the door to the exam.

**Task 1** encode your knowledge about **the function** in the GP prior

**Task 2** randomly sample some data

**Task 3** specify your acquisition function

**Task 4** evaluate and maximise the acquisition function

**Task 5** add new data to model and **re-estimate** hyperparameters

**Loop 4-5** till budget is gone<sup>1</sup>

---

<sup>1</sup>till they open the door to the exam.

## Summary

---

## Summary

---

- GPs are quite useful surrogates!

## Summary

---

- GPs are quite useful surrogates!
- Degrees of beliefs are **really** useful

## Summary

---

- GPs are quite useful surrogates!
- Degrees of beliefs are **really** useful
- The uncertainty allows us to design rich strategies for how to acquire data

## Summary

---

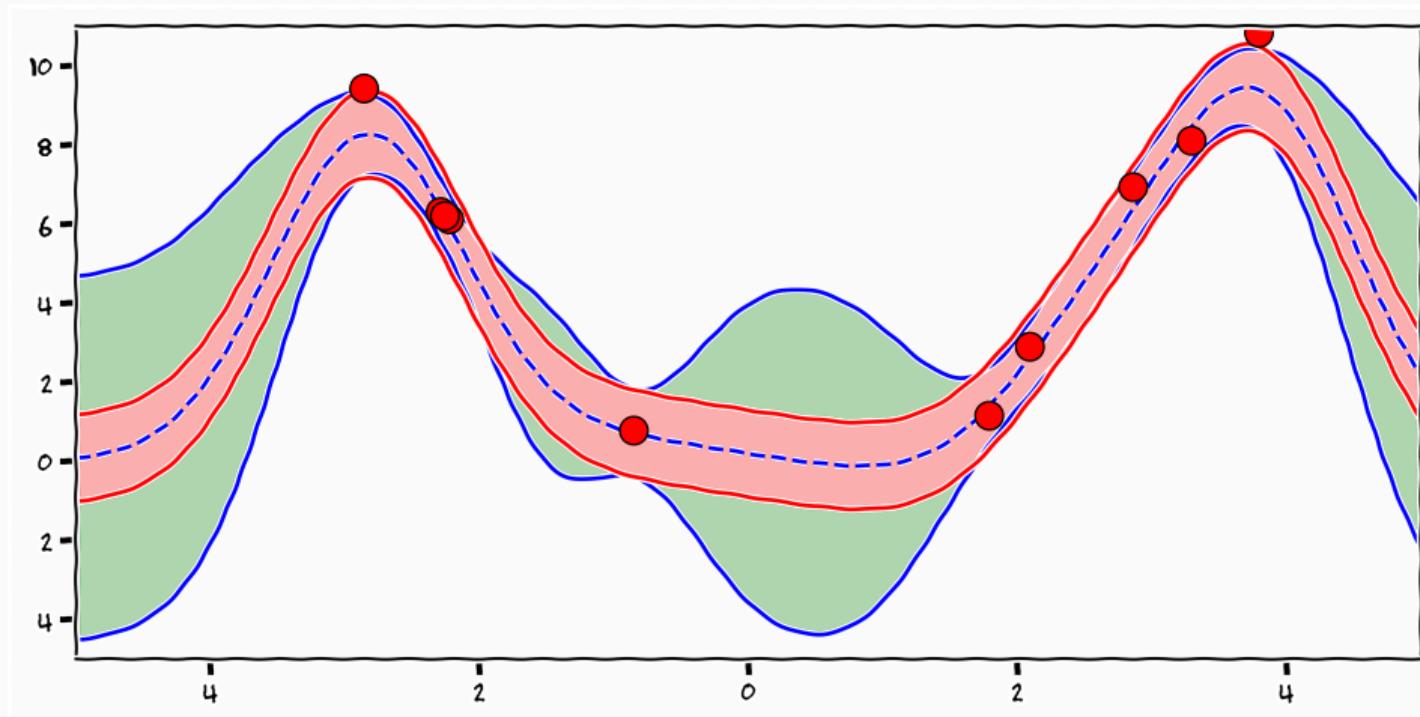
- GPs are quite useful surrogates!
- Degrees of beliefs are **really** useful
- The uncertainty allows us to design rich strategies for how to acquire data
- The factorisation of uncertainty allows us to describe search strategies in simple acquisition functions

## Individual Submission

---

- I have reluctantly made another Jupyter Notebook
- Similar to the material in the PDF
- Deadline Friday 3rd of November at 12:00

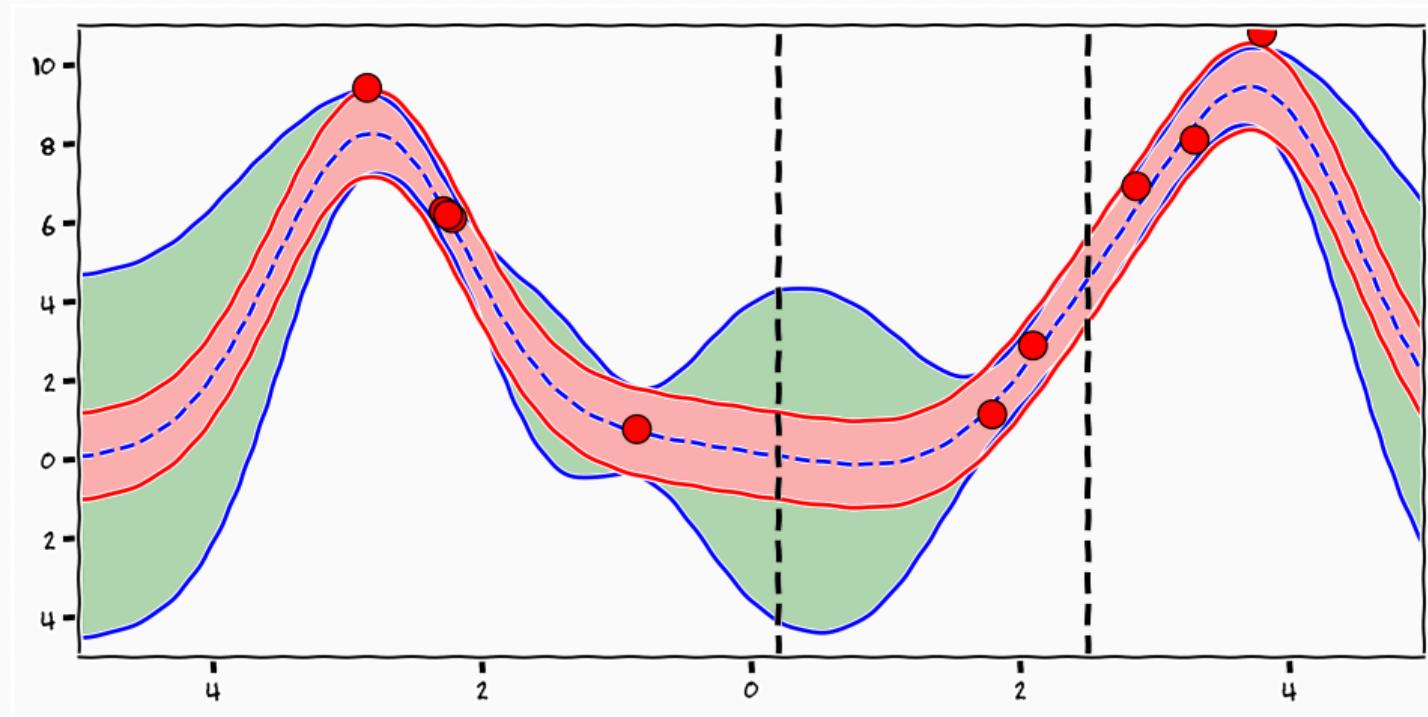
## Uncertainty Quantification/Factorisation



**Aleatoric/Stochastic** "Randomness" inherent in system, or noise in our measurement of system

**Epistemic** Uncertainty related to our ignorance of the underlying system

# Uncertainty for Decision Making



\*

eof

## References

---

- Brochu, Eric, Vlad M. Cora, and Nando de Freitas (2010). "A Tutorial on Bayesian Optimization of Expensive Cost Functions, With Application To Active User Modeling and Hierarchical Reinforcement Learning." In: *CoRR*.
- Cox, Dennis and Susan John (Mar. 1997). "SDO: A Statistical Method for Global Optimization." In: *Multidisciplinary Design Optimization: State of the Art*. Ed. by M. N. Alexandrov and M. Y. Hussaini, pp. 315–329.
- Kushner, Harold J. (1963). "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise." Undetermined. In: *Joint Automatic Control Conference 1*, pp. 69 –79.

- Mockus, J., Vytautas Tiesis, and Antanas Zilinskas (Sept. 1978). “The application of Bayesian methods for seeking the extremum.” In: vol. 2, pp. 117–129.
- Thompson, William R. (1933). “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples.” In: *Biometrika* 25.3/4, pp. 285–294.