Bayesian Optimization Tutorial

Module 2: Quantifying the Value of Information

Joel Paulson
Assistant Professor, Department of Chemical and Biomolecular Engineering, The Ohio State University

Great Lakes PSE Student Workshop, 2023

For copies of slides & code, see
https://github.com/joelpaulson/Great_Lakes_PSE_Workshop_2023

Bird’s-eye View of Bayesian Optimization

while {budget not exhausted}

Fit a Bayesian machine learning model
(usually Gaussian process regression)
to observations {x, f(x)}

Find x that maximizes acquisition(x, posterior)

Sample x & then observe f(x)

end

Assume our goal is to minimize f(x)
[same idea as maximize, just replace with -f(x)]

How to Define an Acquisition Function α_n?

- When properly selected, the value of $\alpha_n(x)$ at any $x \in \Omega$ should be a good measure of the (expected) benefit of querying f at that point in future
  - Must depend on the posterior distribution of $f|y_{1:n}$

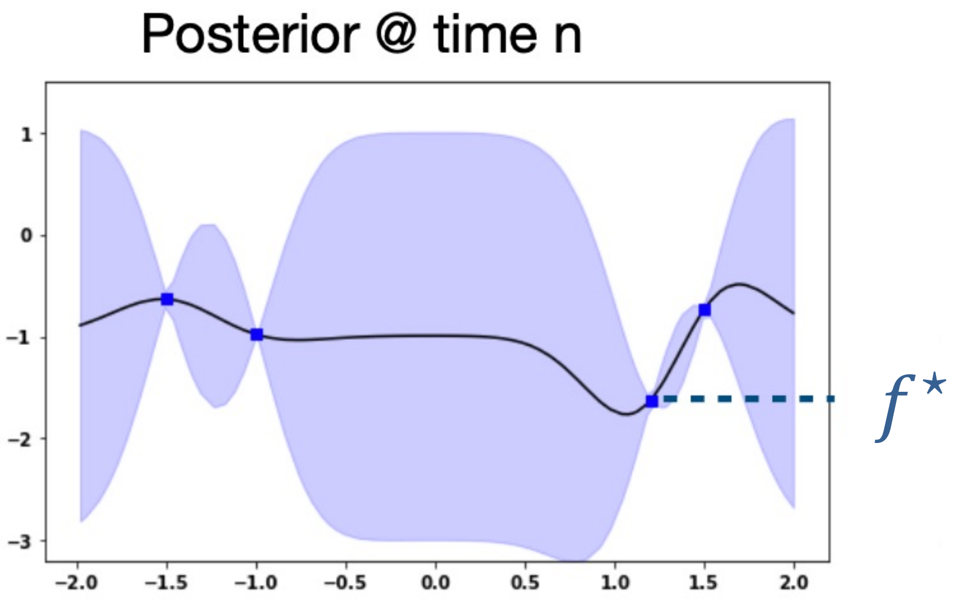
- This implies we should like to preferentially sample at the point that produces the highest possible value of the acquisition function:

$$x_{n+1} \in \mathrm{argmax}_{x \in \Omega}\, \alpha_n(x)$$

- It is expected for this problem to be much cheaper to solve since, unlike f, we have some equation-based form for α_n

Let’s Start with Expected Improvement (EI) Acquisition Function

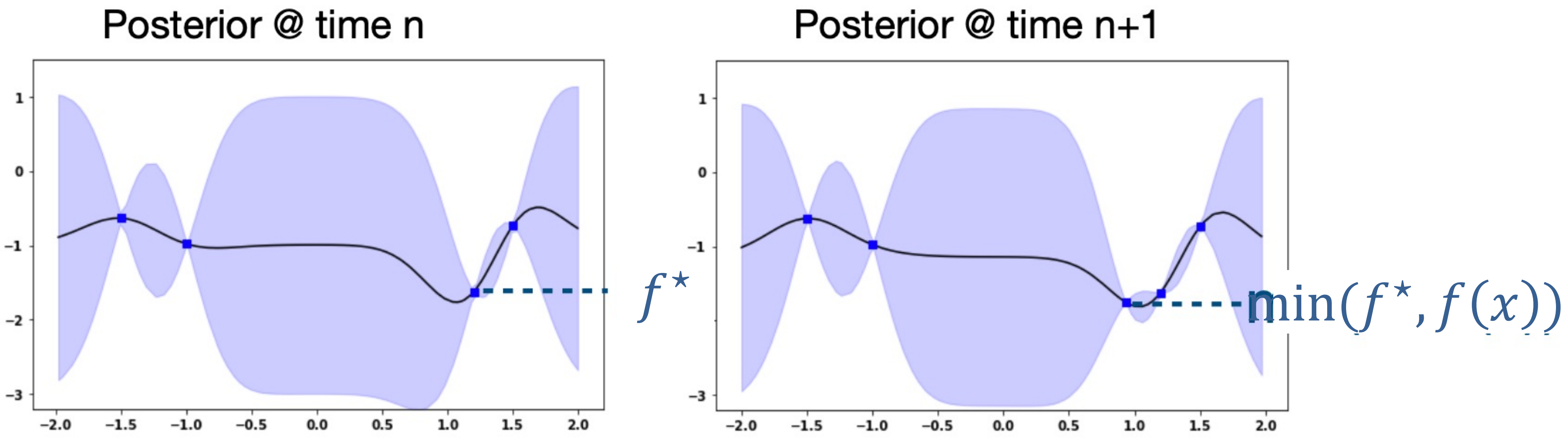
[Mockus 1989; Jones, Schonlau, and Welch 1998]



- Loss if we stop now: $f^\star$

Let's Start with Expected Improvement (EI) Acquisition Function

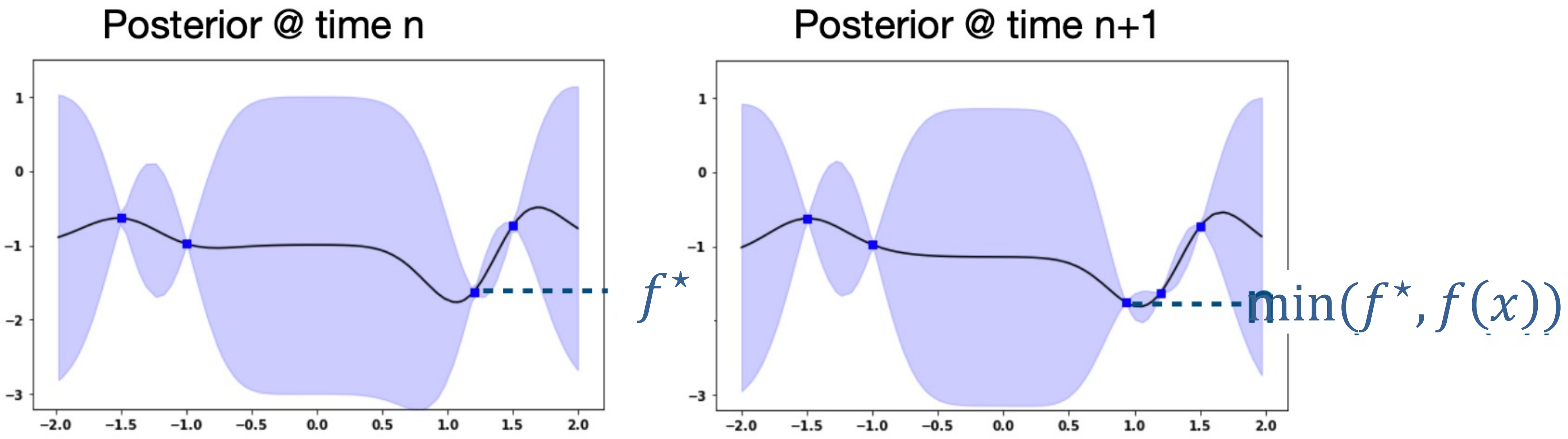
[Mockus 1989; Jones, Schonlau, and Welch 1998]



- Loss if we stop now: $f^\star$
- Loss if we stop after sampling at $f(x)$: $\min(f^\star, f(x))$

Let's Start with Expected Improvement (EI) Acquisition Function

[Mockus 1989; Jones, Schonlau, and Welch 1998]



- Loss if we stop now: $f^\star$
- Loss if we stop after sampling at $f(x)$: $\min(f^\star, f(x))$
- <u>Expected</u> reduction in loss due to sampling: $\mathbb{E}_n[f^\star - \min(f^\star, f(x))]$

Let's Start with Expected Improvement (EI) Acquisition Function

[Mockus 1989; Jones, Schonlau, and Welch 1998]

$$\mathrm{EI}_n(x) = \mathbb{E}_n\{f^\star - \min(f^\star, f(x))\}$$

$$= \mathbb{E}_n\{\max\{f^\star - f(x), 0\}\}$$

$$= \underbrace{\mathbb{E}_Z\{\max\{f^\star - \mu_n(x) - \sigma_n(x)Z\ , 0\}\}}_{\text{Integral can be carried out analytically using integration by parts}}$$

Closed-form Expression Expected Improvement

Standard normal cumulative distribution function (CDF)

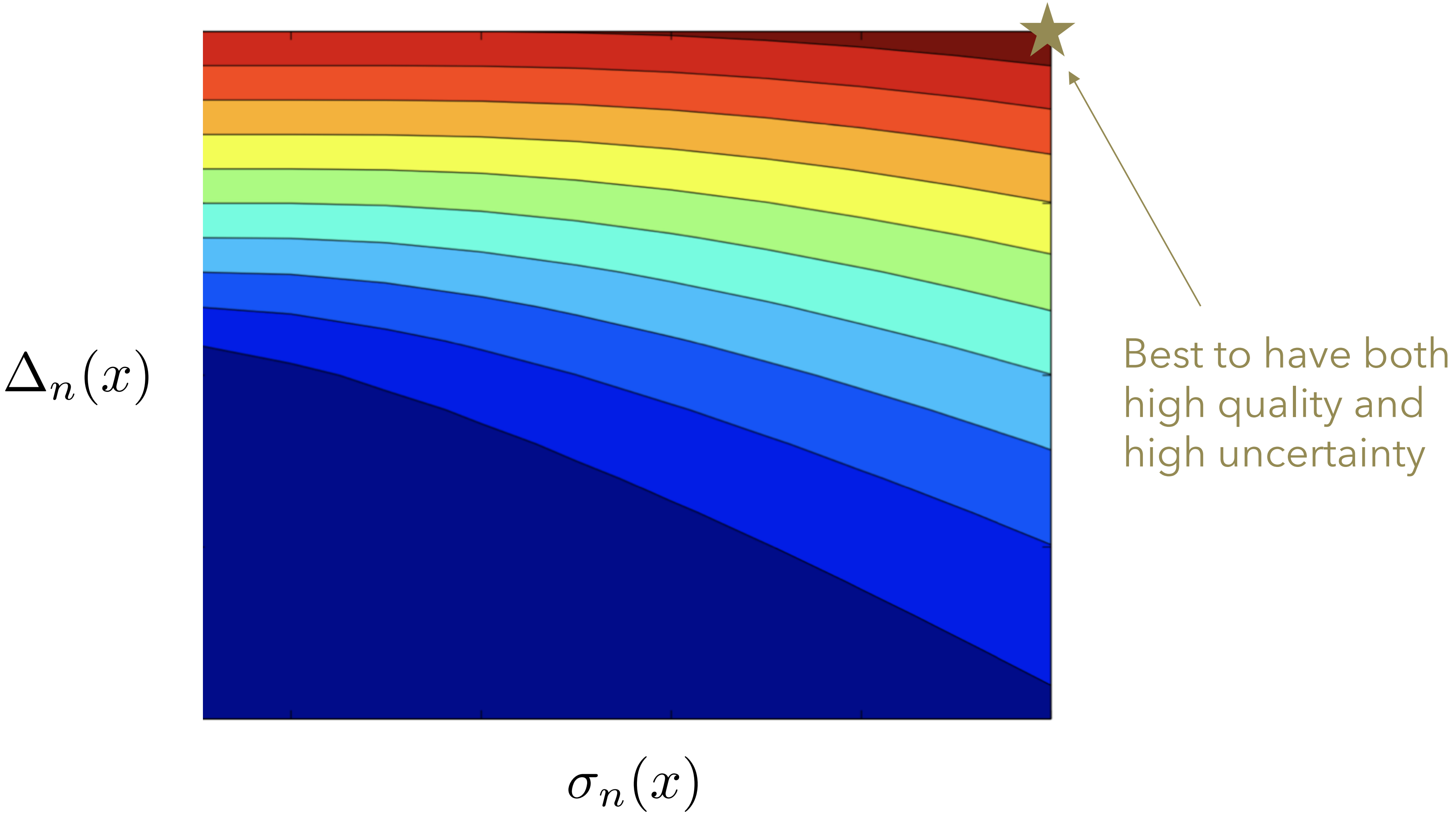
Standard normal probability density function (PDF)

$$\mathrm{EI}_n(x) = \Delta_n(x)\Phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) + \sigma_n(x)\phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right)$$

Where $\Delta_n(x) = f_n^\star - \mu_n(x)$ is expected quality

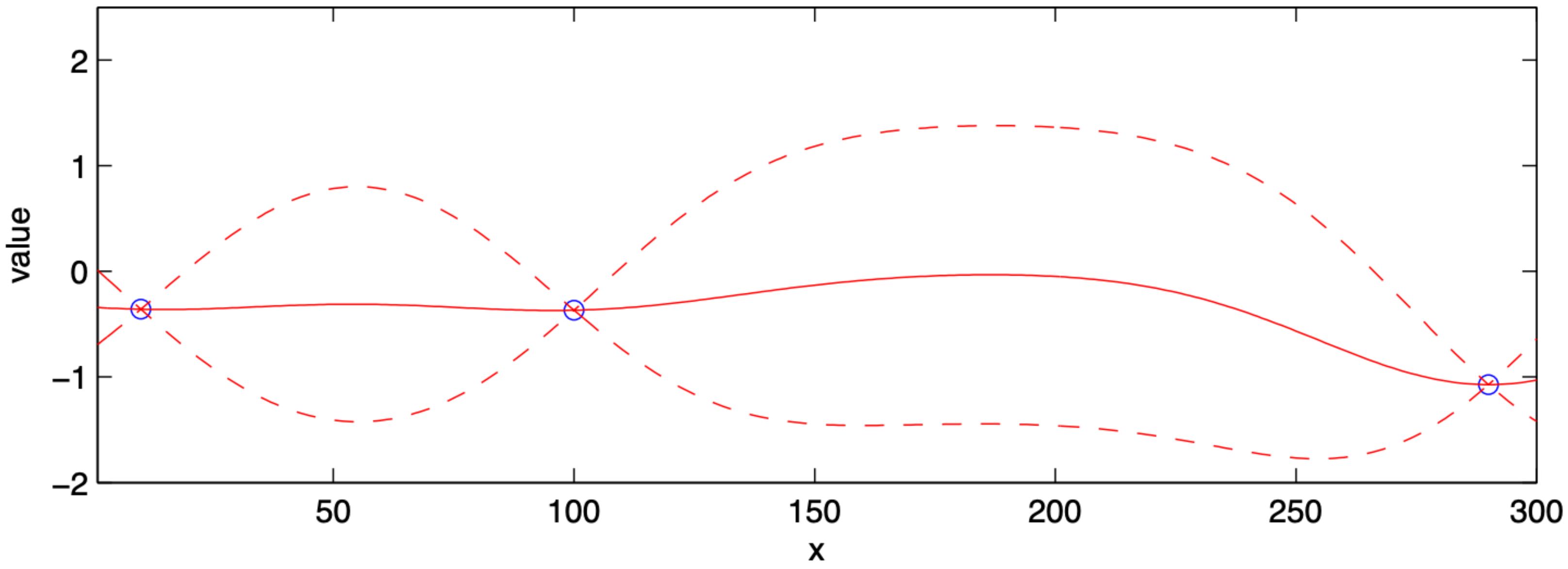
[Jones et al., 1998; Frazier, 2018]

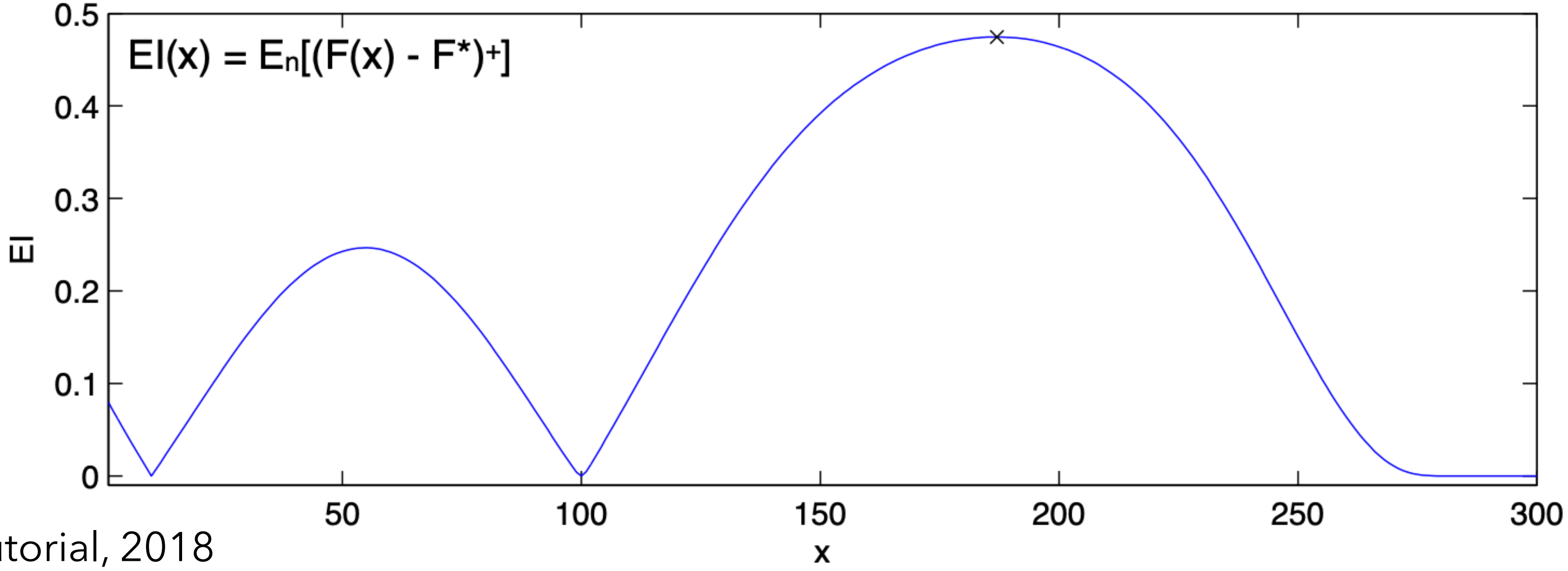
EI Tradeoffs Exploration ($\Delta_n(x)$) vs. Exploitation ($\sigma_n(x)$)



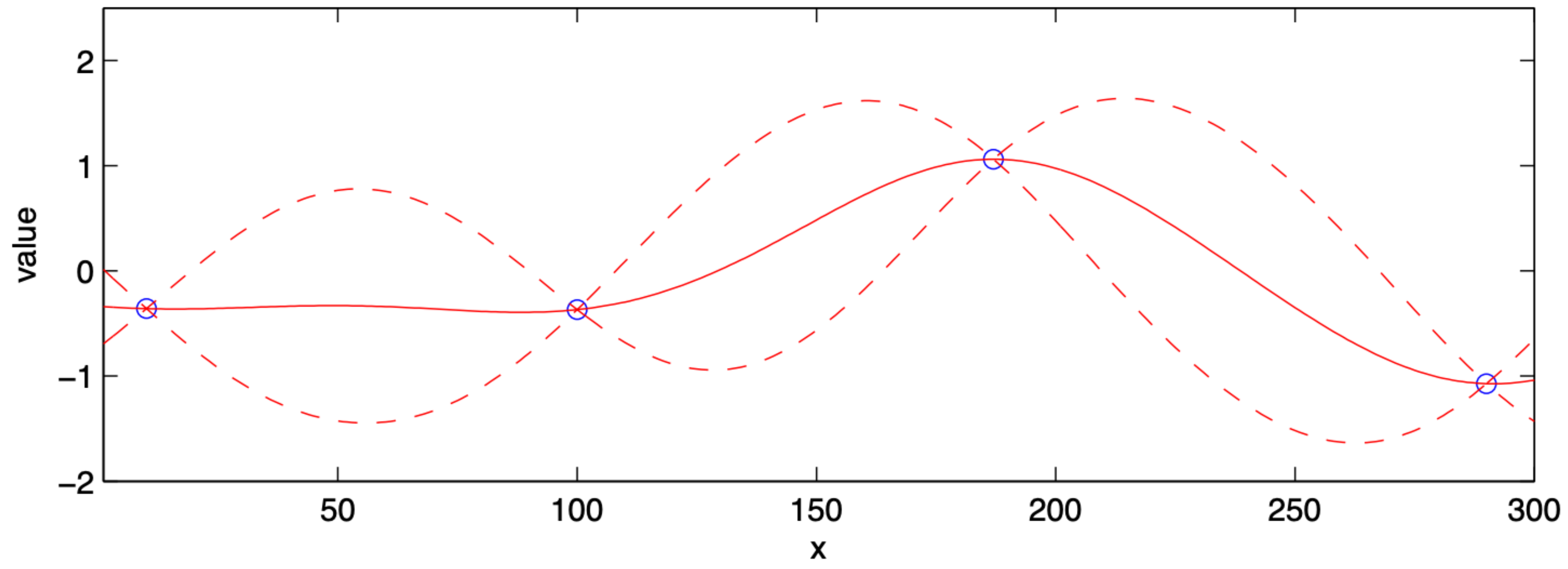
[Jones et al., 1998; Frazier, 2018]

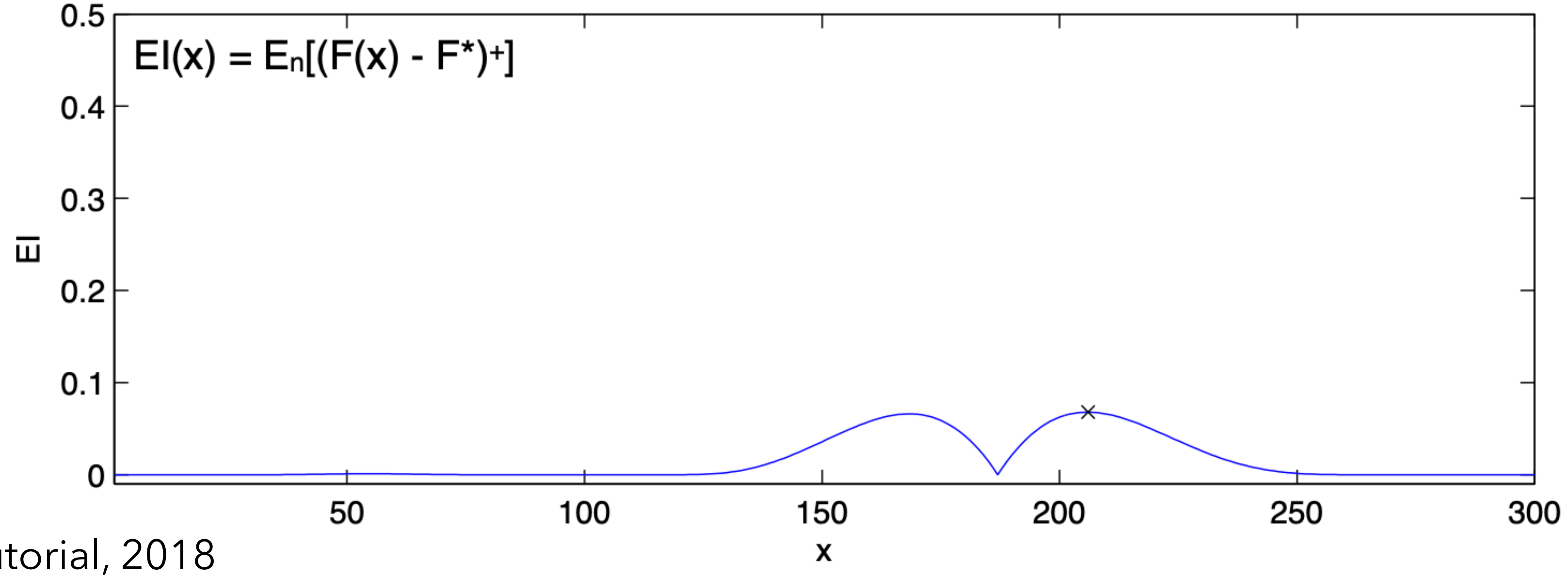
Quick Example of EI for Maximizing 1-Dimensional Objective



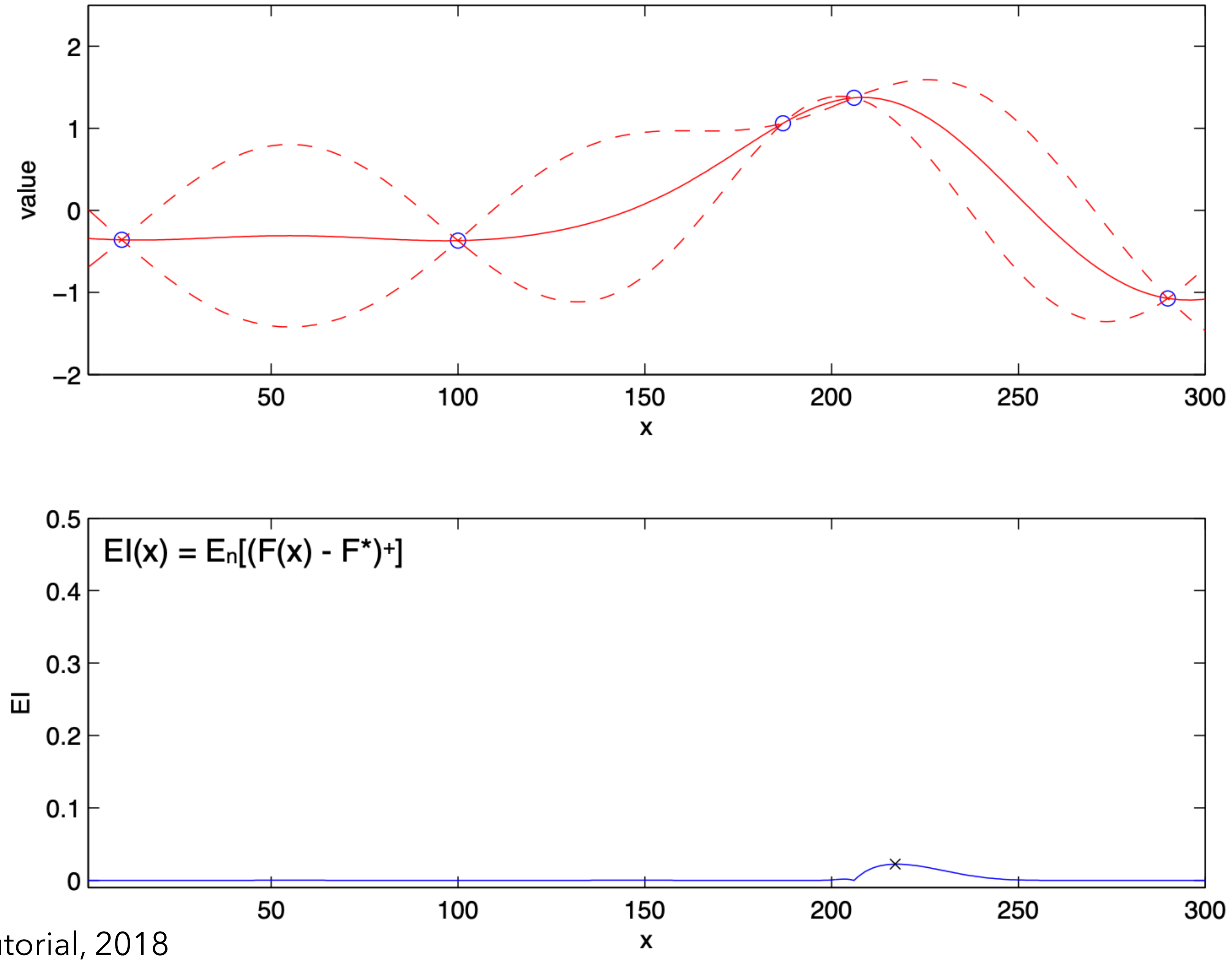


Quick Example of EI for Maximizing 1-Dimensional Objective

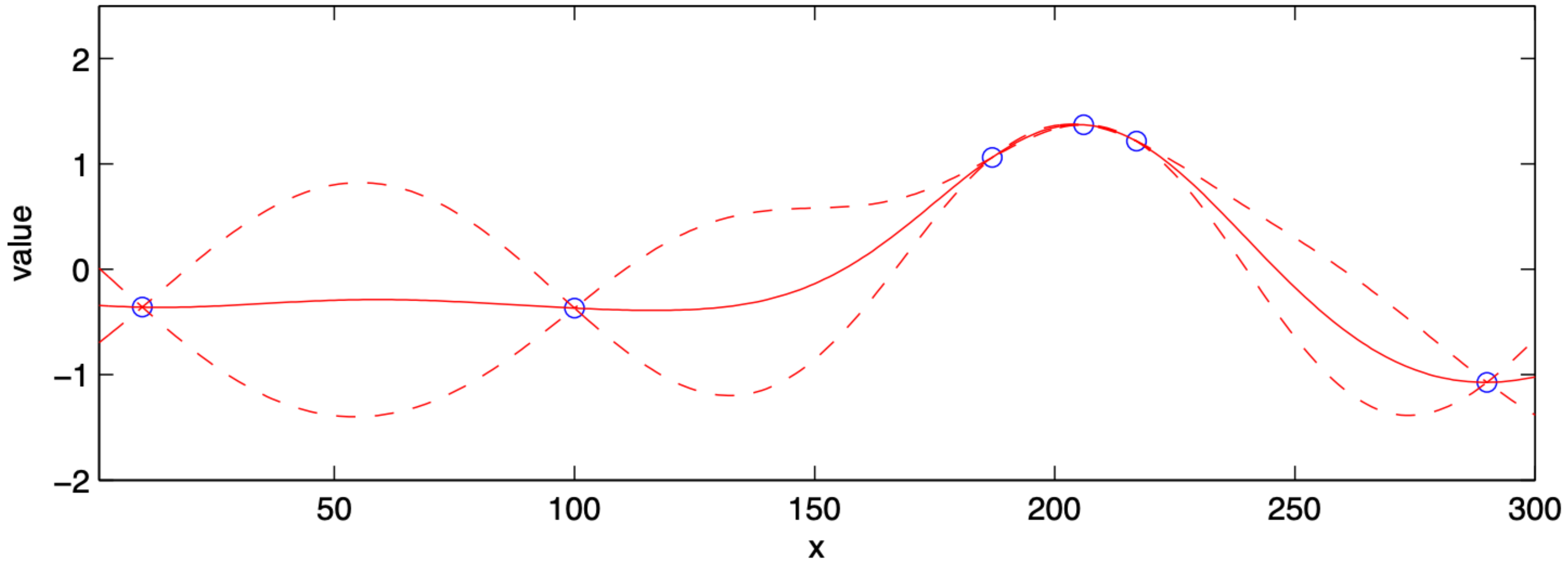


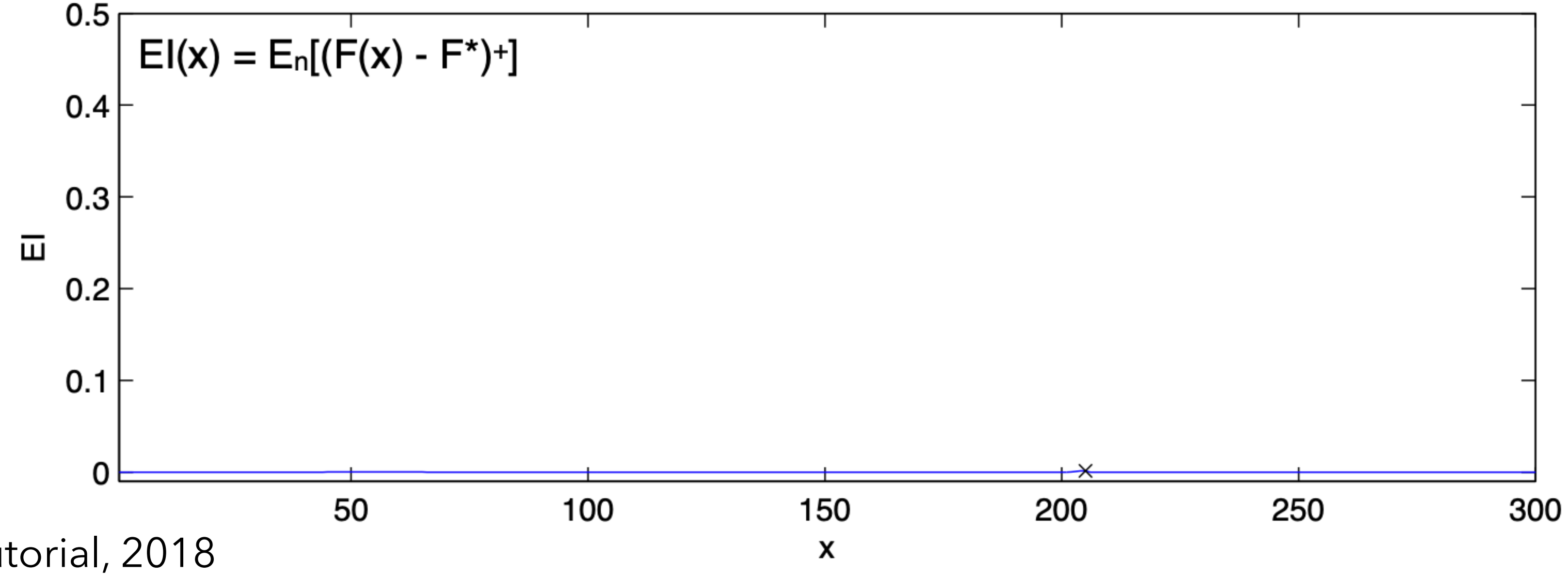


Quick Example of EI for Maximizing 1-Dimensional Objective



Quick Example of EI for Maximizing 1-Dimensional Objective





Thought Experiment

- What should expected improvement (EI) reduce to when the variance of the prediction is zero everywhere?
  - We no longer have uncertainty, so we no longer need to sequentially search (simply find the minimum of the mean function)
  - We can think of traditional optimization as placing a GP prior with perfectly known mean function $\mu_0(x) = f(x)$ (and zero variance/covariance), then running EI one step to find the "true" minimum
  - EI in some sense generalizes traditional "white-box" optimization to the unknown "black-box" setting → attempts to be information-optimal

Is EI Optimal in any Sense?

- Yes, it turns out that EI is Bayes-optimal under some assumptions:
  - There is no noise in the observations of the objective function
  - We are only willing to select previously evaluated point as final solution
  - We are risk neutral (i.e., we value a random outcome according to its expected value, hence $\mathbb{E}[\text{Reduction in Loss}]$)
  - This is our last evaluation

Why is this assumption needed?

In general, we must solve a sequential decision-making problem

- The loss that we calculated previously is only a function of the next sample that we take; however, in general, we have a budget of N remaining samples $\{x_1, x_2, \ldots, x_N\}$
- Furthermore, every sample that we take yields more data, such that we have more information to make our next decision
- We can formulate this as a stochastic optimal control problem where our state is current data, action is next sample, and immediate reward is reduction in loss

Best (finite-budget) sampling strategy is policy that optimizes the value function (total loss reduction)

- Policy: $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_N\}$, $x_k = \pi_k(\mathcal{D}_{k-1})$

- Value function: $V_{\boldsymbol{\pi}}(\mathcal{D}_0) = \mathbb{E}\left[\sum_{k=1}^{N} r(\mathcal{D}_{k-1}, \mathcal{D}_k)\right]$, r(·) is loss reduction

- Optimal policy: $V^\star(\mathcal{D}_0) = V_{\boldsymbol{\pi}^\star}(\mathcal{D}_0) = \max_{\boldsymbol{\pi}\in\Pi} V_{\boldsymbol{\pi}}(\mathcal{D}_0)$

- Solution expressed using dynamic programing:

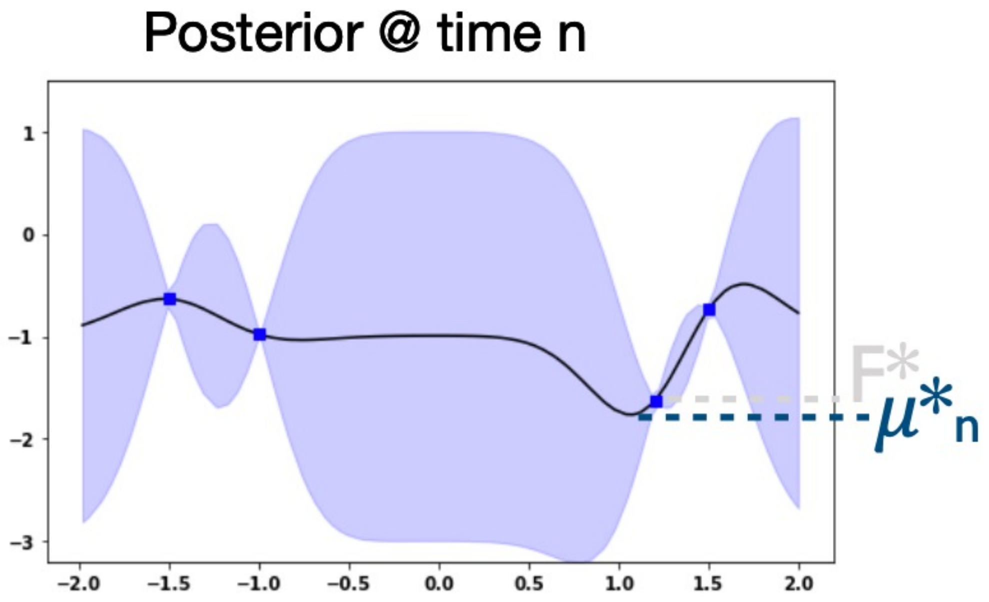
$$V_k(\mathcal{D}) = \max_{x\in\Omega} \mathbb{E}_{\mathcal{D}+}\left[r(\mathcal{D}, \mathcal{D}^+) + V_{k-1}(\mathcal{D}^+)\right] \quad , \quad \forall k = 1, \ldots, N$$

Let’s Drop Two of These Assumptions

- Yes, it turns out that EI is Bayes-optimal under some assumptions:
  - ~~There is no noise in the observations of the objective function~~
  - ~~We are only willing to select previously evaluated point as final solution~~
  - We are risk neutral (i.e., we value a random outcome according to its expected value, hence $\mathbb{E}[\text{Reduction in Loss}]$)
  - This is our last evaluation

Yields Knowledge Gradient (KG) acquisition function, what should be loss?

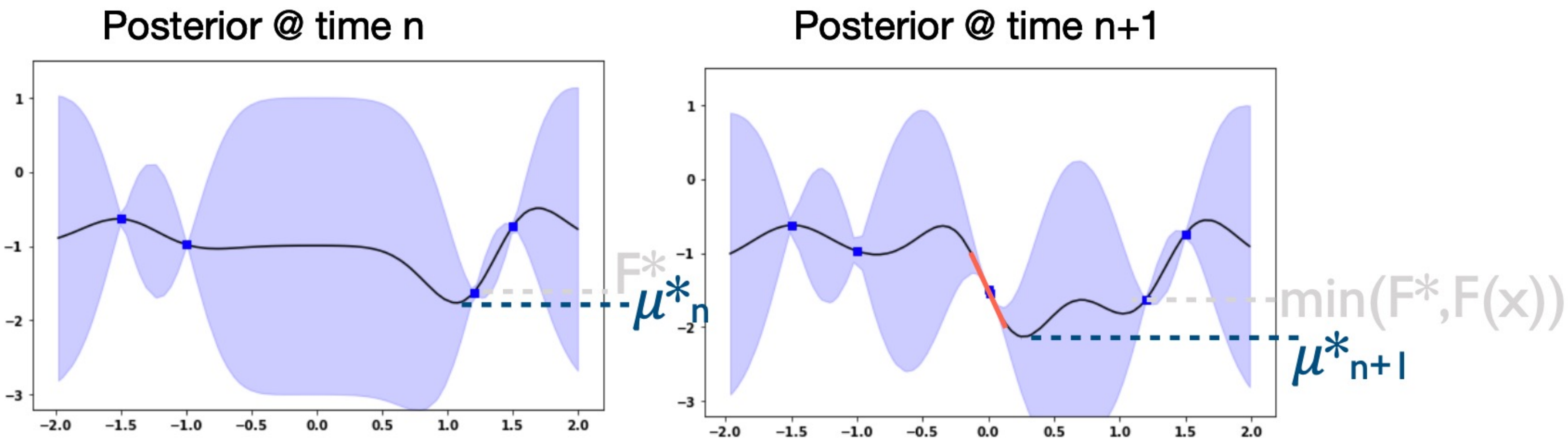
Knowledge Gradient (KG) Acquisition Function



- Loss if we stop now: $\mu_n^{\star} = \min_{x \in \Omega} \mu_n(x)$

[Frazier et al. , *SIAM Journal on Optimization and Control*, 2009]

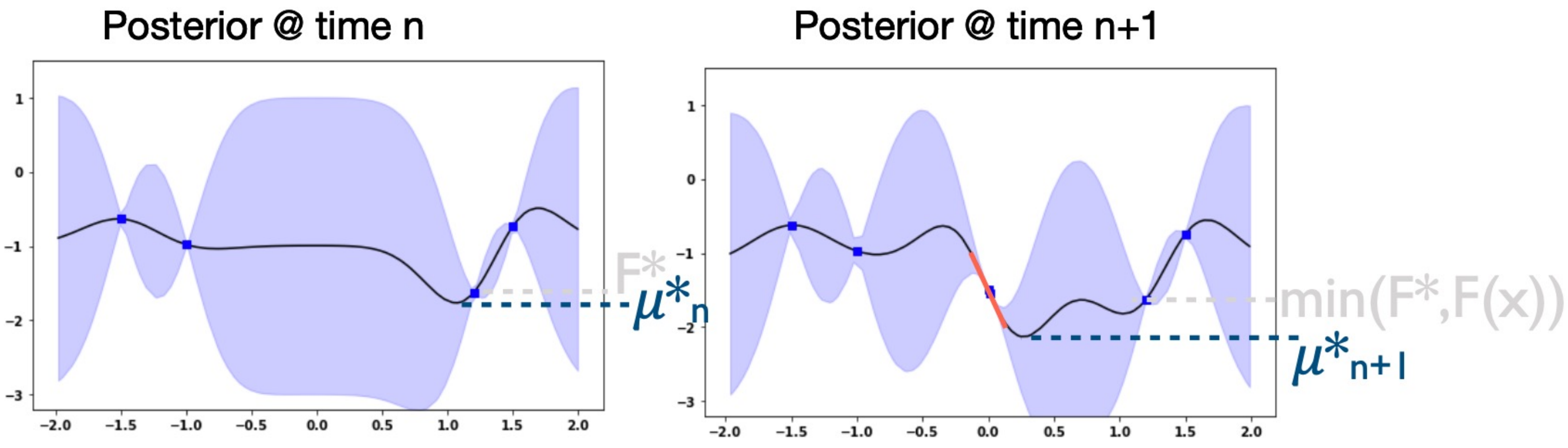
Knowledge Gradient (KG) Acquisition Function



- Loss if we stop now: $\mu_n^\star = \min_{x \in \Omega} \mu_n(x)$
- Loss if we stop after sampling $f(x)$: $\mu_{n+1}^\star = \min_{x \in \Omega} \mu_{n+1}(x)$

[Frazier et al. , *SIAM Journal on Optimization and Control*, 2009]

Knowledge Gradient (KG) Acquisition Function



- Loss if we stop now: $\mu_n^\star = \min_{x \in \Omega} \mu_n(x)$
- Loss if we stop after sampling $f(x)$: $\mu_{n+1}^\star = \min_{x \in \Omega} \mu_{n+1}(x)$
- Expected reduction in loss due to sampling: $\mathbb{E}_n[\mu_n^\star - \mu_{n+1}^\star \mid \text{sample } x]$

[Frazier et al. , *SIAM Journal on Optimization and Control*, 2009]

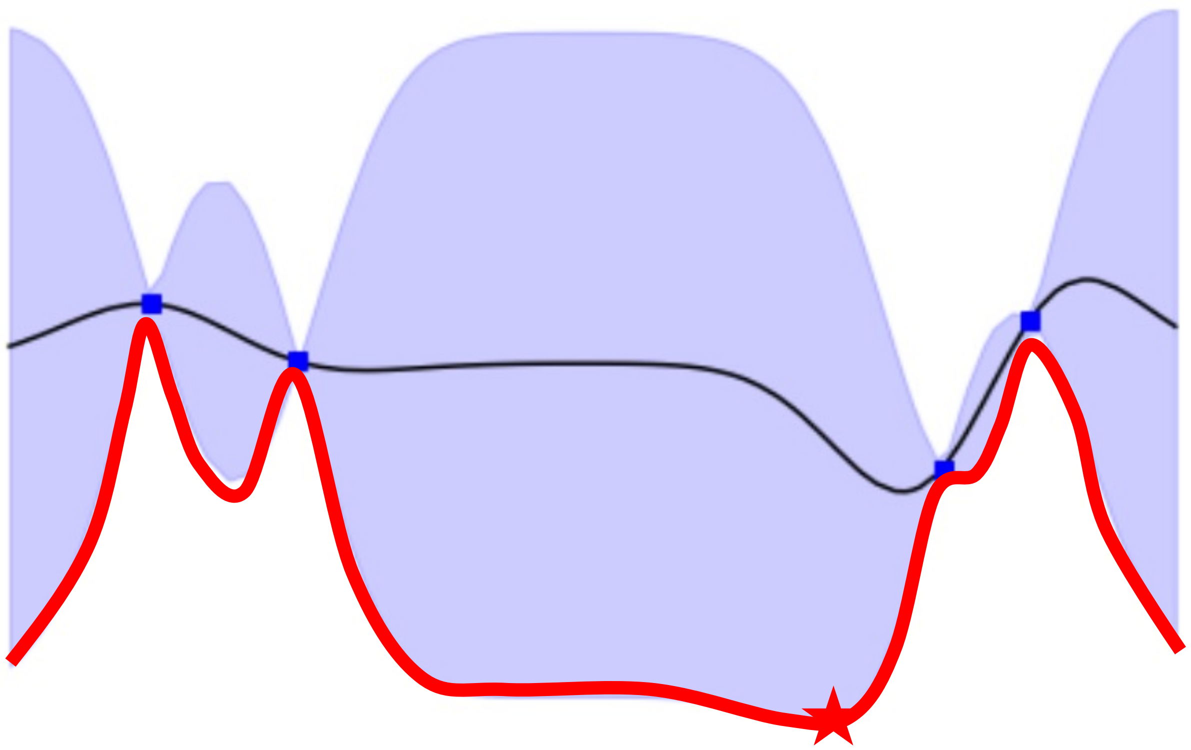
KG is significantly harder to maximize than EI due to two-stage optimization

- The main disadvantage of KG is that we lose the analytic formula that we were able to derive for EI

- We will discuss two practical strategies to maximize KG in Module 3

- Are there more practical strategies to handle noise?
  - Yes, there are other approximations that we can develop, but we may lose performance and/or theoretical properties…

Method 1: Lower Confidence Bound (LCB)



- Simple idea: Just directly minimize a lower bound on the function

$$\min_{x \in \Omega} \mu_n(x) + \sqrt{\beta_{n+1}}\sigma_n(x)$$

We can establish rigorous bounds on "regret" for LCB

- Lower confidence bound: $l_n(x) = \mu_{n-1}(x) - \sqrt{\beta_n}\sigma_{n-1}(x)$
- Upper confidence bound: $u_n(x) = \mu_{n-1}(x) + \sqrt{\beta_n}\sigma_{n-1}(x)$

- Assume that true function satisfies $f(x) \in [l_n(x), u_n(x)]$
  (can prove this holds with high probability for sufficiently large β_n)

- Performance measure: Regret r_n defined as distance to optimal solution:

$$r_n = f(x_n) - f(x^\star)$$

[Srinivas et al. , *IEEE Transactions on Information Theory,* 2012]

We can establish rigorous bounds on "regret" for LCB

- The following sequence of inequalities hold:

$$r_n = f(x_n) - \min_{x \in \Omega} f(x)$$ [Definition of regret]

$$\leq u_n(x_n) - \min_{x \in \Omega} f(x)$$ [Property of upper bound]

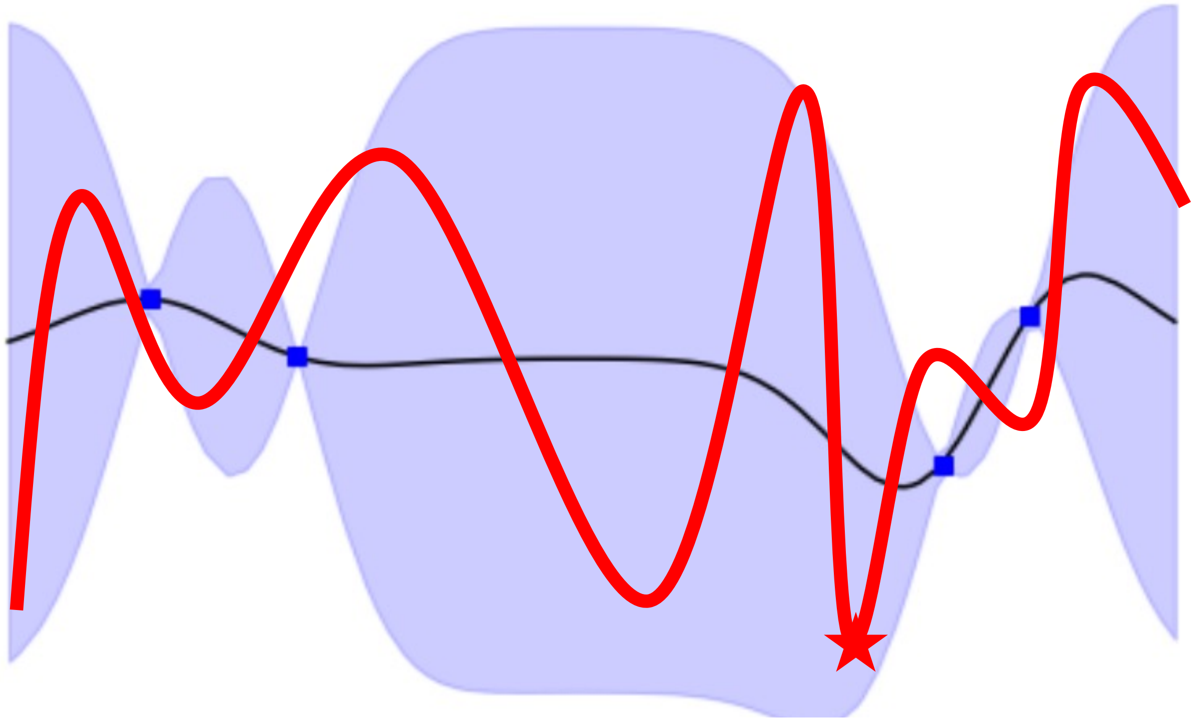
$$\leq u_n(x_n) - \min_{x \in \Omega} l_n(x)$$ [Property of lower bound]

$$= u_n(x_n) - l_n(x_n)$$ [Definition of our sample choice $x_n = \mathrm{argmin}_x\, l_n(x)$]

$$= 2\sqrt{\beta_n}\sigma_{n-1}(x_n)$$ [Difference between bounds given by standard deviation]

[Srinivas et al. , *IEEE Transactions on Information Theory,* 2012]

Method 2: Thompson Sampling (TS)



- Minimize random sample of the GP, i.e., $f^{(n)} \sim \mathcal{GP}(\mu_n(x), \sigma_n^2(x))$

$$\min_{x \in \Omega} f^{(n)}(x)$$

What about expensive black-box constraints?

Assume scalar for simplicity, but could easily be vector $c(x) = \max_{i=1,\ldots,n_c} c_i(x)$

How to handle black-box constraints?

- Short answer: Need another GP model for $c(x)$

$$\min_{x \in \Omega} \; f(x) \quad \text{s.t.} \quad c(x) \leq 0$$

Black-box constraints that are usually **coupled** with evaluation of the objective

What about expensive black-box constraints?

Feasibility indicator function
= 1 if $c(x) \leq 0$ and 0 otherwise

Improvement over our best incumbent value $f_n^* = \min_{i=1,\dots,n} f(x_i)$

$$\mathrm{EIC}(x) = \mathbb{E}_n\{\mathbf{1}_{\{c(x)\leq 0\}}(x)\max\{0, f_n^\star - f(x)\}\}$$

$$= \mathbb{E}_n\{\mathbf{1}_{\{c(x)\leq 0\}}(x)\}\mathbb{E}_n\{\max\{0, f_n^\star - f(x)\}\}$$

Conditional independence of objective and constraints*

$$= \underbrace{\Pr_n\{c(x) \leq 0\}}_{\Phi\left(-\frac{\mu_n^c(x)}{\sigma_n^c(x)}\right)}\underbrace{\mathrm{EI}_n(x)}$$

$$\Phi\left(-\frac{\mu_n^c(x)}{\sigma_n^c(x)}\right)$$

Standard expected improvement value that has analytic solution (shown previously)

Probability of feasibility;
analytic solution available for GP model using normal CDF

*Only holds when constraints / objective are independently modeled

Gardner et al., 2014.

CODE REVIEW

Workshop Schedule

9:00 – 9:20	Introduction: Why Go Beyond Traditional Optimization?
9:20 – 10:20	Module 1: Probabilistic Surrogate Modeling*
10:20 – 10:30	Break
10:30 – 11:20	Module 2: Quantifying the Value of Information*
11:20 – 12:20	Module 3: The BO Feedback Loop*
12:20 – 12:30	Break
12:30 – 1:00	Module 4: Beyond Bayesian Optimization

*module includes Python code review / exercises