# Cost-aware Bayesian Optimization via the Pandora's Box Gittins Index

To appear in NeurIPS'24

Qian Xie (Cornell ORIE)

Joint work with Raul Astudillo, Peter Frazier, Ziv Scully, and Alexander Terenin

INFORMS'24 Data Mining Best General Paper Competition

# Coauthors

Raul Astudillo

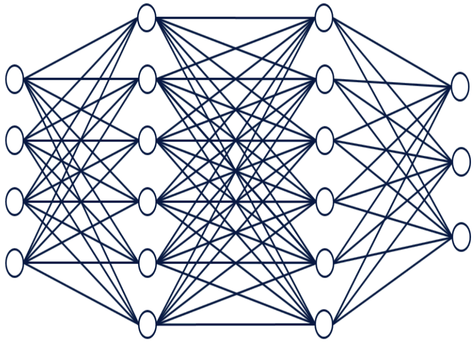
Peter Frazier

Ziv Scully

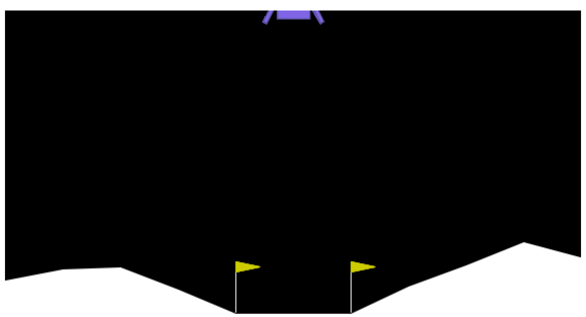
Alexander Terenin

# World of Parameter Optimization

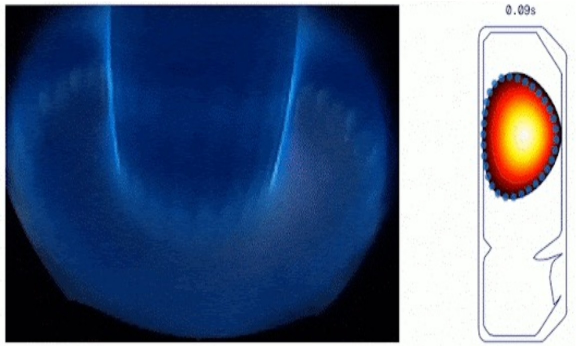
Hyperparameter tuning:

Training parameters →  → Accuracy

Control optimization:

Control parameters →  → Reward

Plasma physics:

Reactor parameters →  → Stability

# World of Parameter Optimization

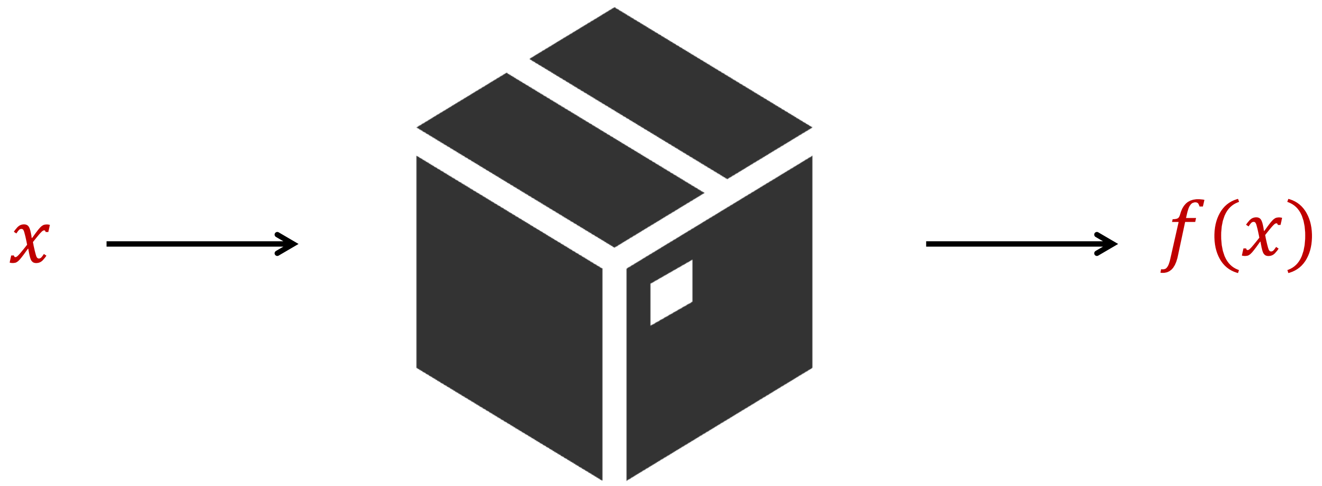
Black-box optimization:

Input parameters →

→ Performance metrics

# World of Parameter Optimization

Black-box optimization:



# Optimizing Black-box Functions

$x \longrightarrow$  $\longrightarrow f(x)$

Goal: $\mathbf{max}_{x\in\mathcal{X}}\, f(x)$

# Optimizing Black-box Functions

$x \longrightarrow$  $\longrightarrow f(x)$

Goal: $\mathbf{max}_{x \in \mathcal{X}}\, f(x)$

$f \sim$Stochastic Process

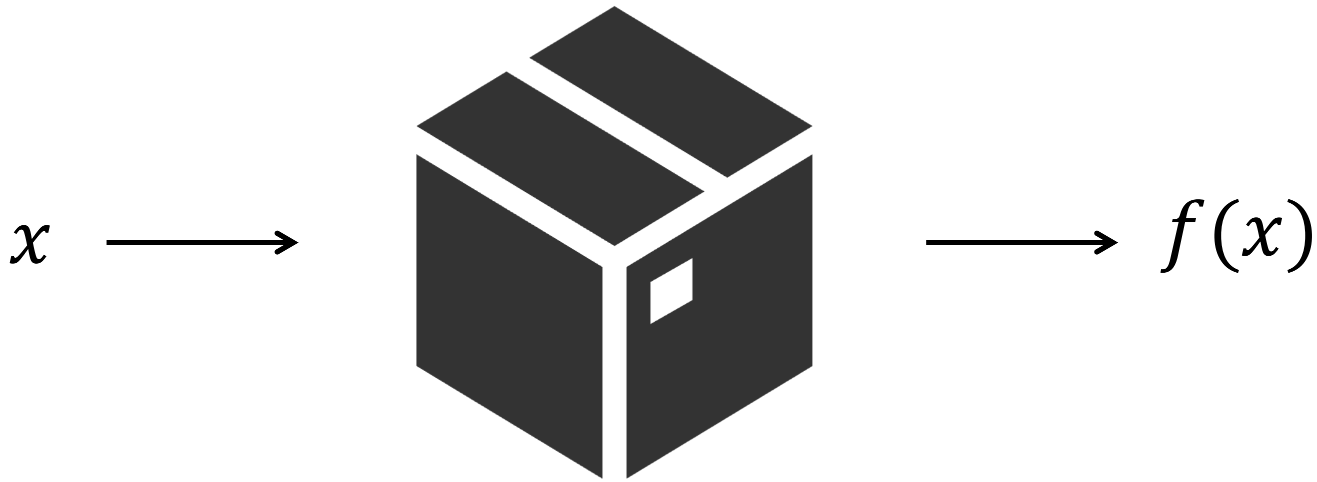
# Optimizing Black-box Functions

$x \longrightarrow$  $\longrightarrow f(x)$

expensive-to-evaluate

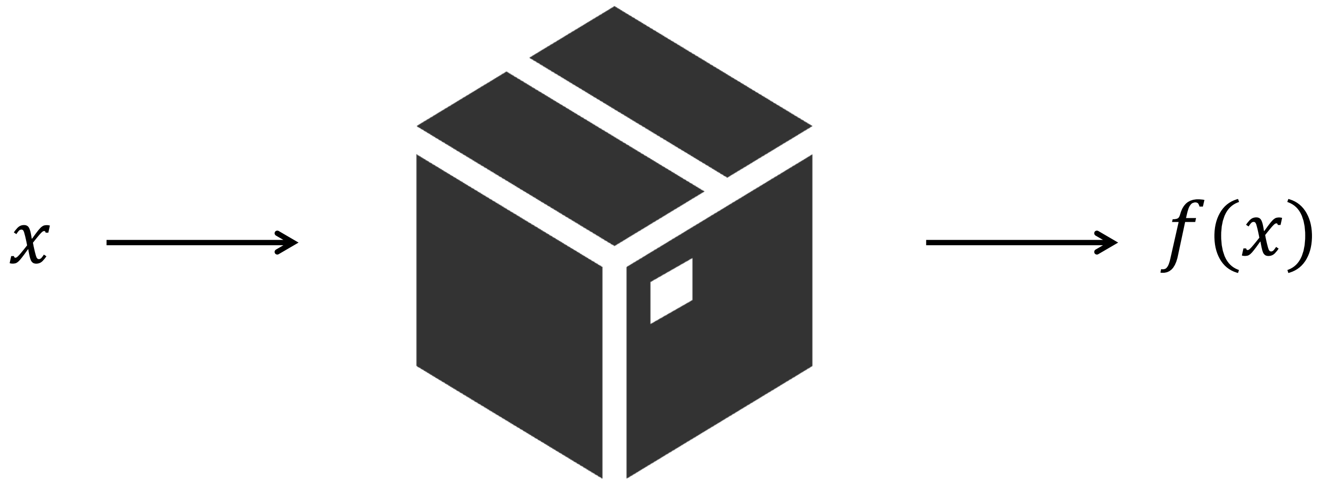
Goal: $\mathbf{max}_{x \in \mathcal{X}} f(x)$

# Optimizing Black-box Functions




Goal: $\mathbf{max}\ \mathbb{E} \max_{t=1,2,\ldots,T} f(x_t)$

# Optimizing Black-box Functions




Goal: $\mathbf{max}\ \mathbb{E} \max_{t=1,2,\dots,T} f(x_t)$

$f \sim$ Gaussian Process

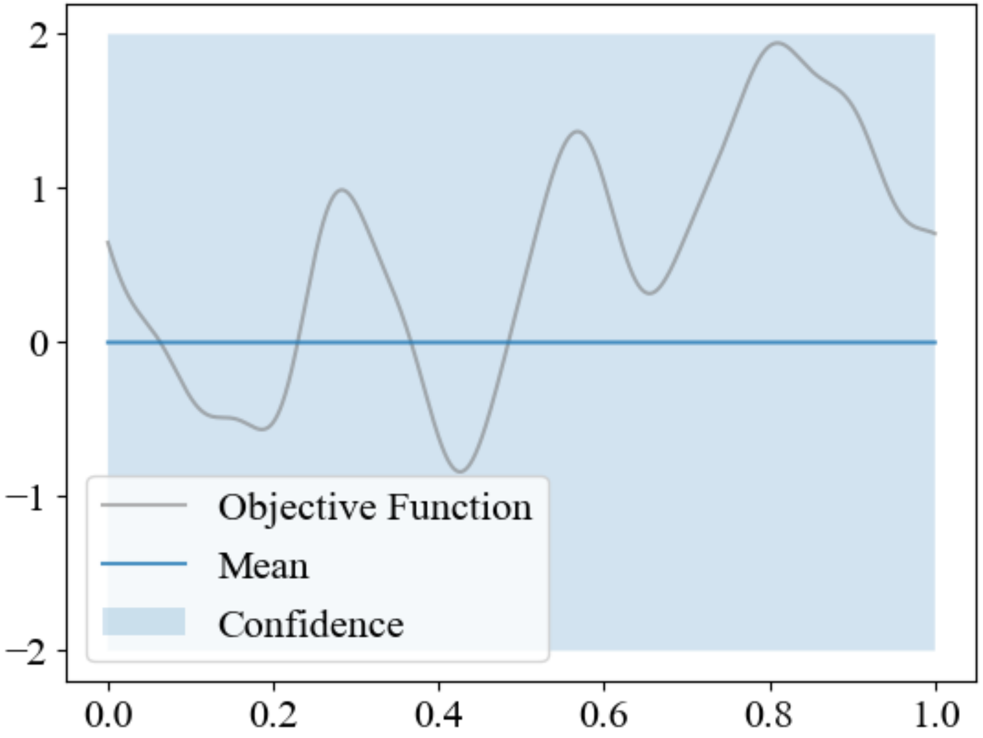
# Bayesian Optimization

$x \longrightarrow$  $\longrightarrow f(x)$

Goal: $\mathbf{max}\ \mathbb{E} \max_{t=1,2,\ldots,T} f(x_t)$

$f \sim$Gaussian Process

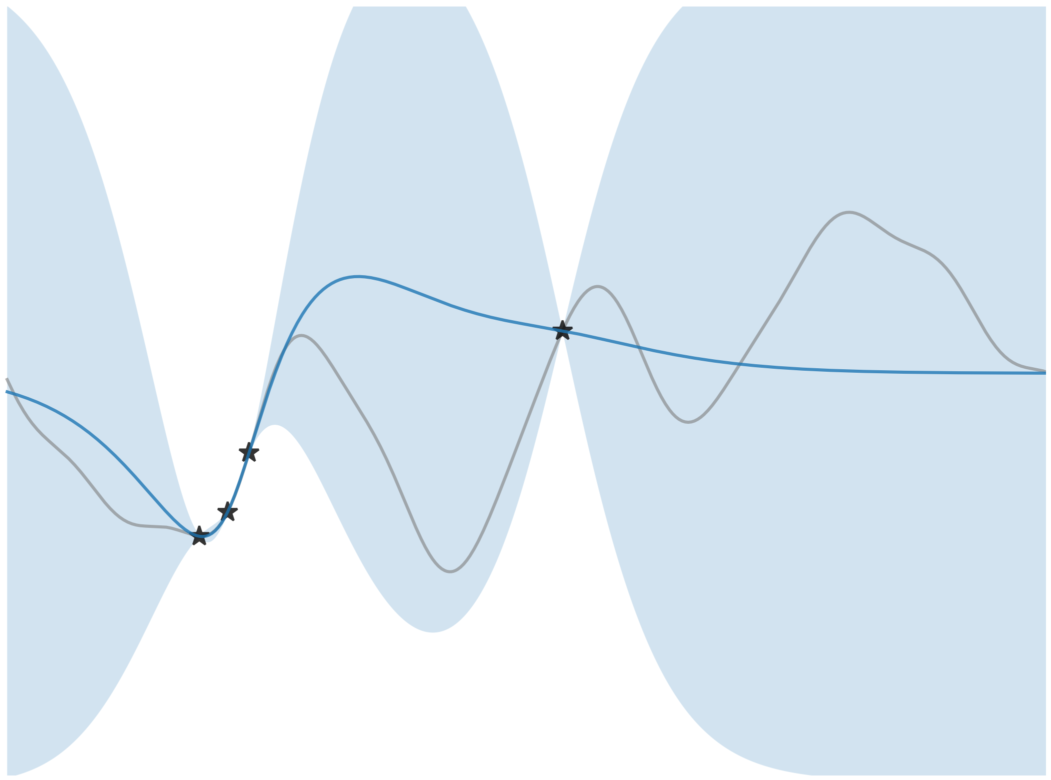
# Bayesian Optimization




Goal: $\mathbf{max}\ \mathbb{E} \max_{t=1,2,\ldots,T} f(x_t)$

$f$ ~Gaussian Process

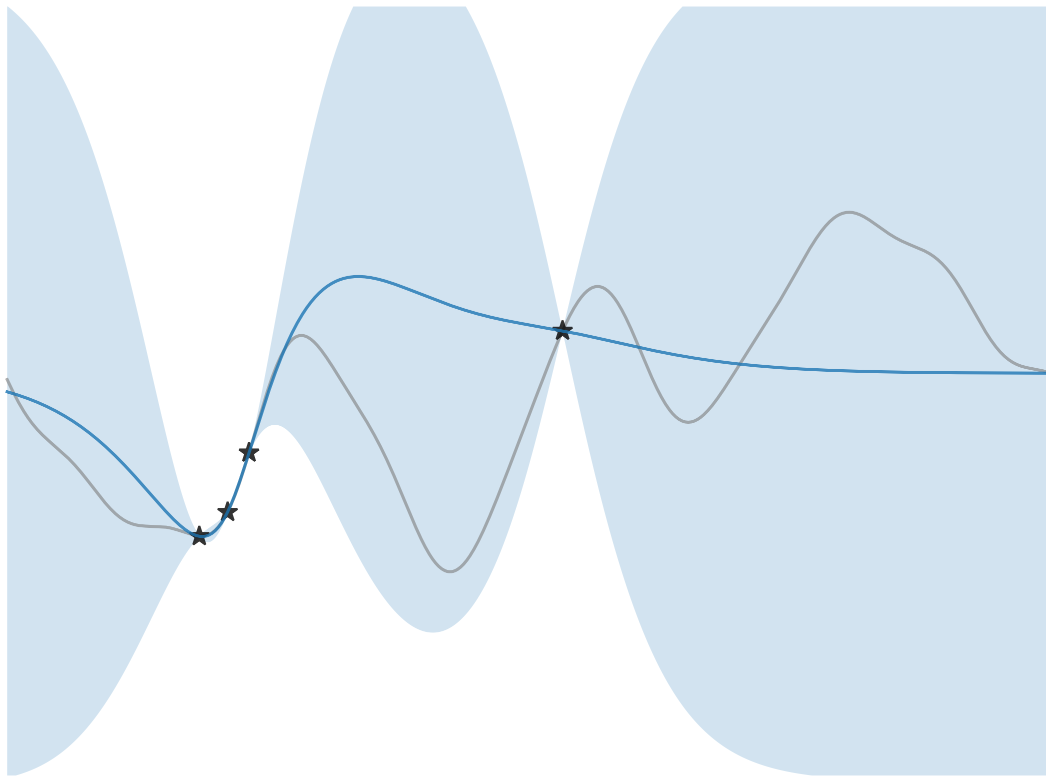
# Bayesian Optimization



Goal: $\mathbf{max}\ \mathbb{E} \max_{t=1,2,\ldots,T} f(x_t)$

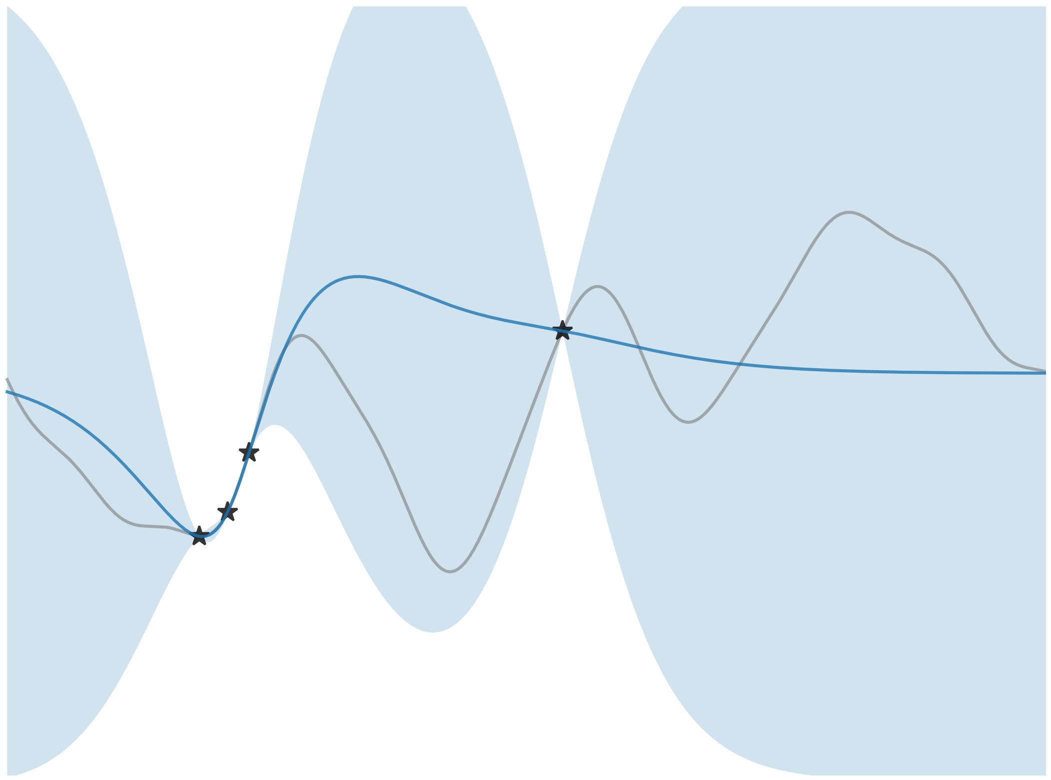
$f \sim$Gaussian Process

# Bayesian Optimization



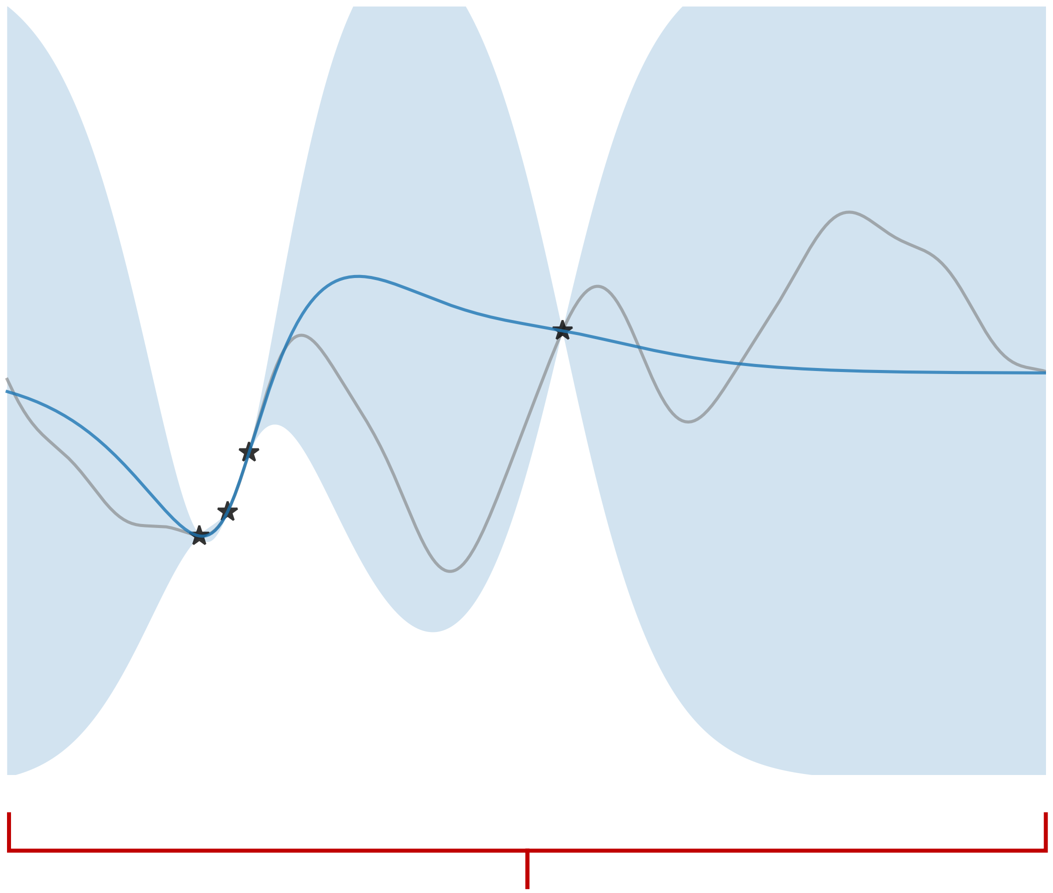
What to evaluate next?

# Bayesian Optimization



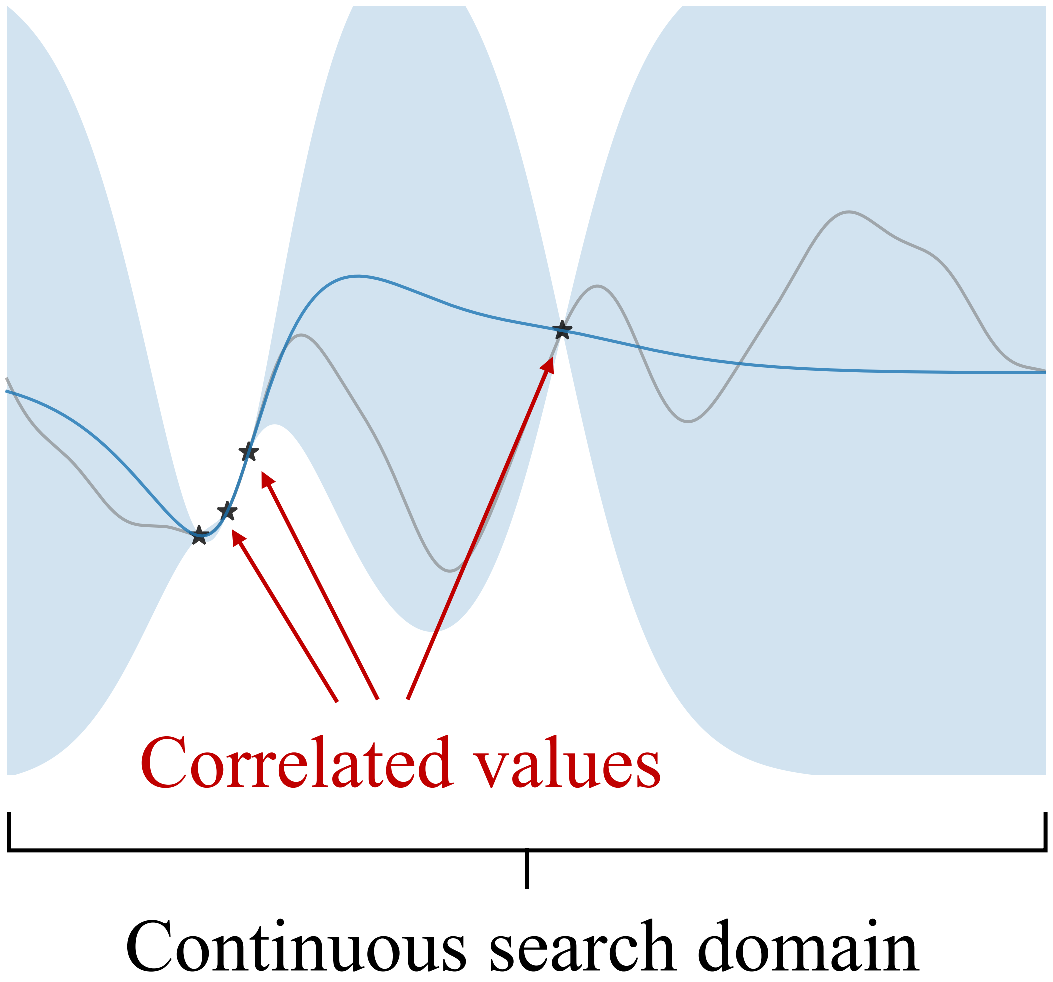
Optimal policy?

# Challenges of Bayesian Optimization

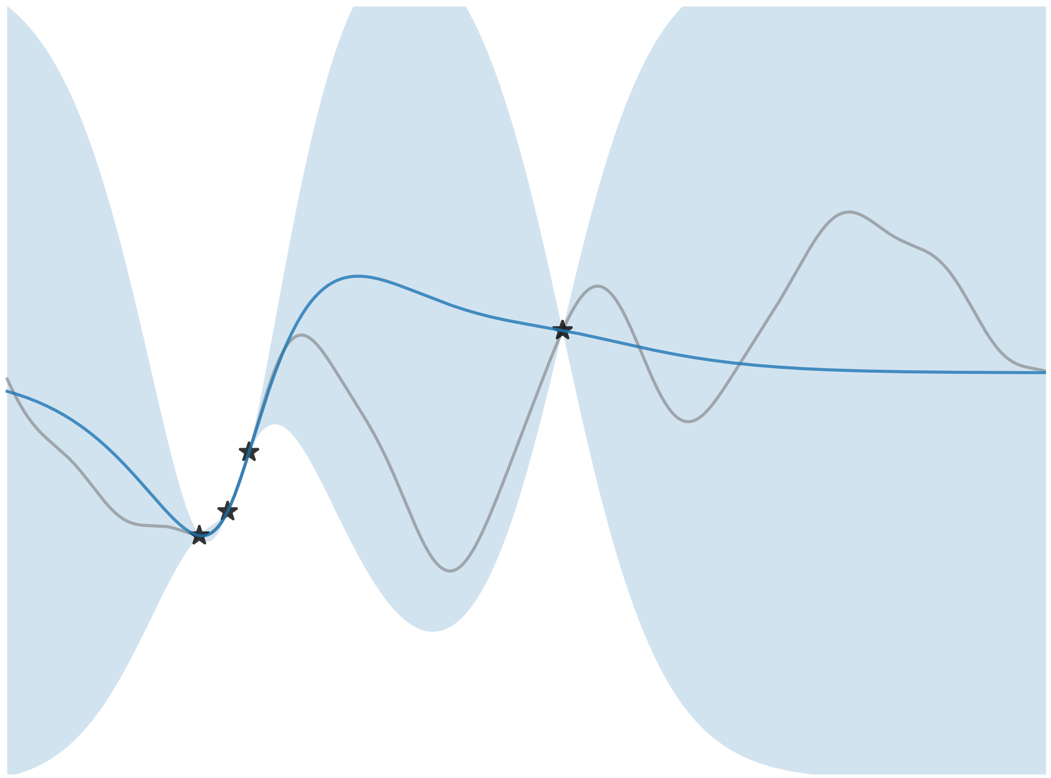


Continuous search domain

# Challenges of Bayesian Optimization

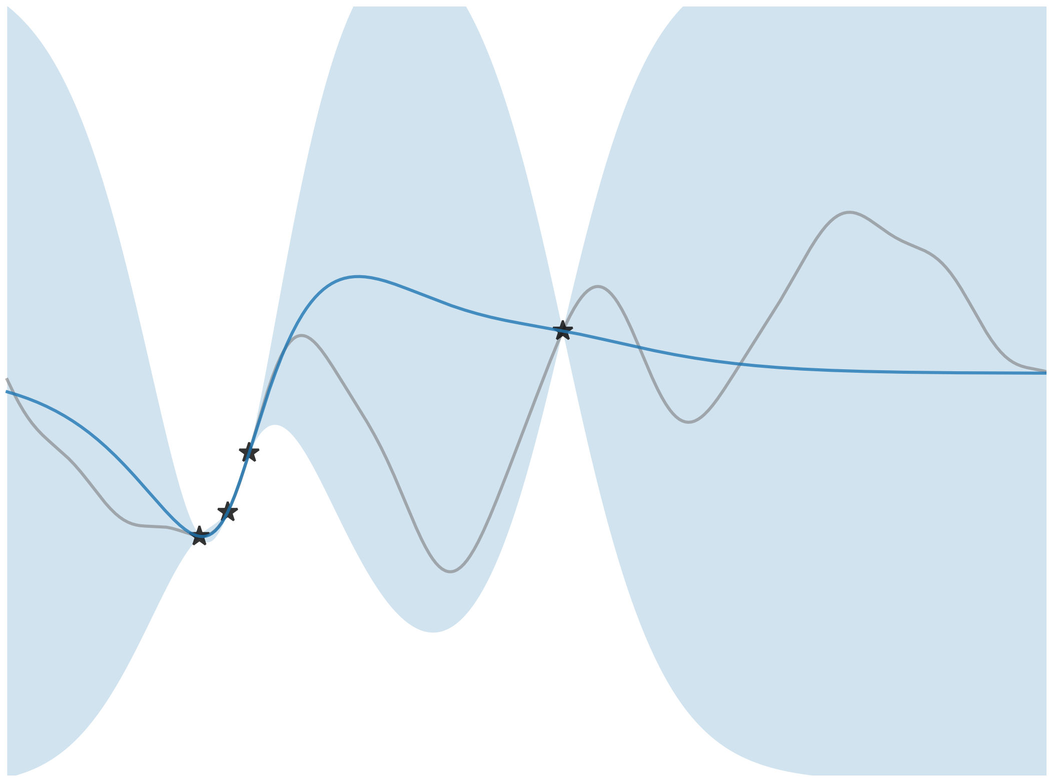


# Challenges of Bayesian Optimization



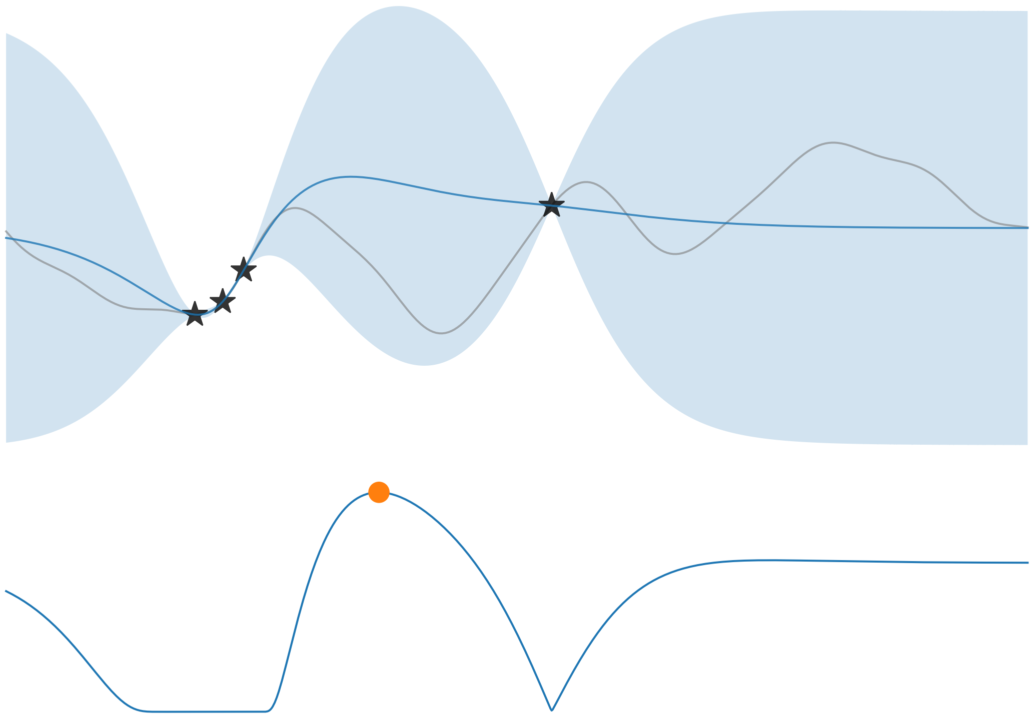
Correlation & continuity $\Longrightarrow$ Intractable MDP

# Challenges of Bayesian Optimization



Intractable MDP $\Longrightarrow$ Optimal policy unknown

# Popular Policy: Expected Improvement

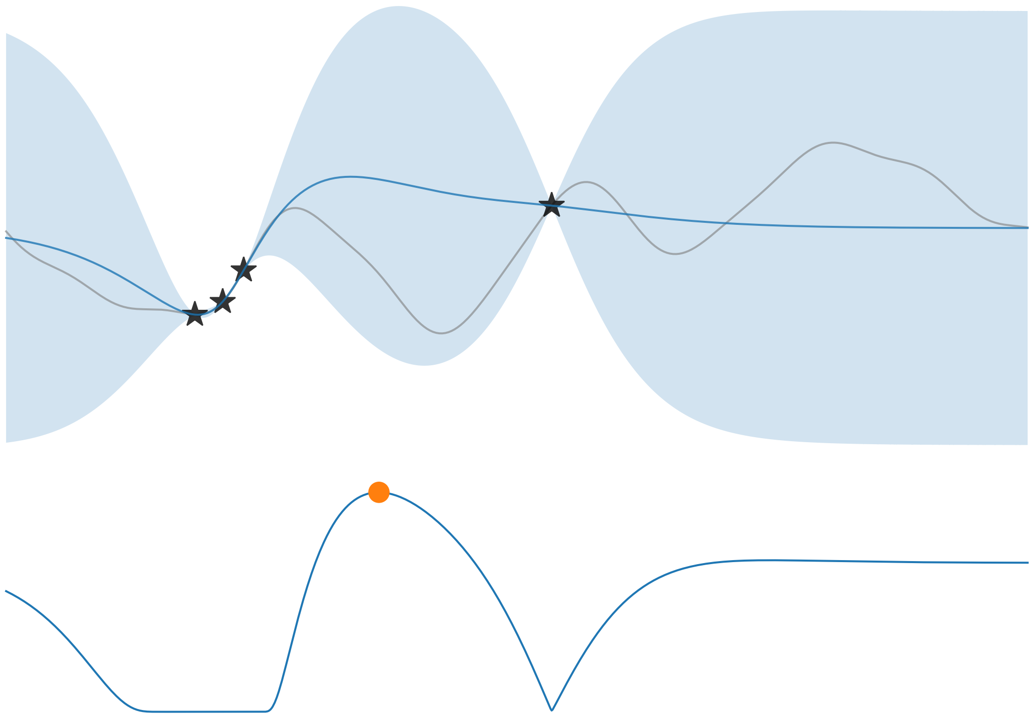


current best observed

data

$$\mathrm{EI}(x) = \mathbb{E}[\underbrace{\max(f(x) - y_{\text{best}}, 0)}_{\text{“improvement”}} \mid D]$$

# Popular Policy: Expected Improvement

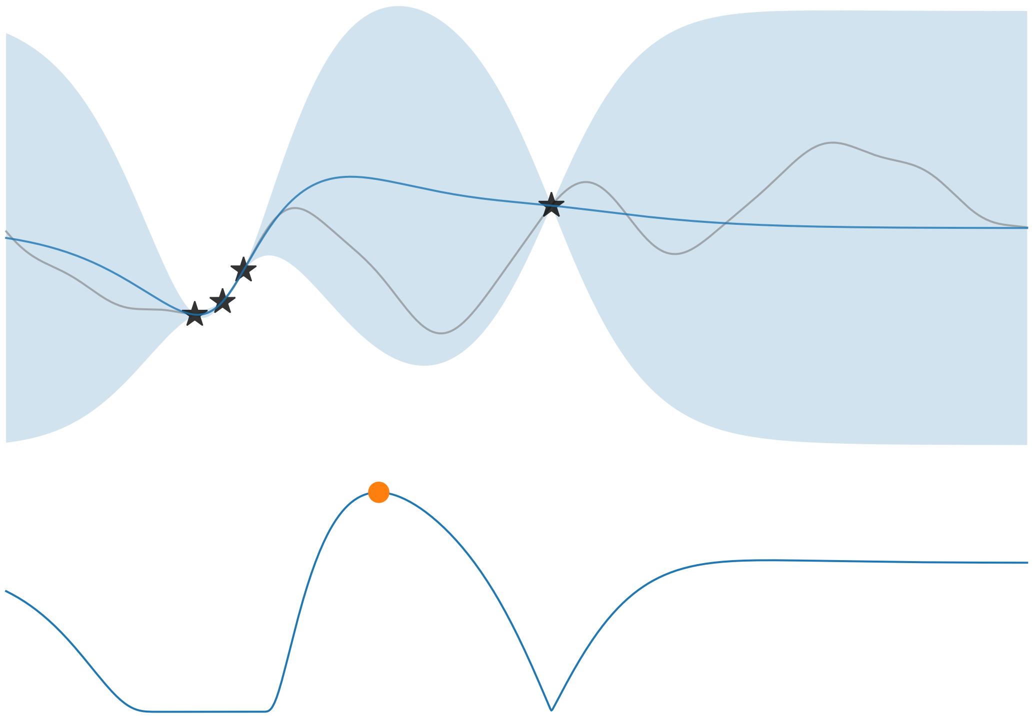


current best observed    data

$$\mathrm{EI}(x) = \mathbb{E}[\underbrace{\max(f(x) - y_{\text{best}}, 0)}_{\text{“improvement”}} \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

# Popular Policy: Expected Improvement



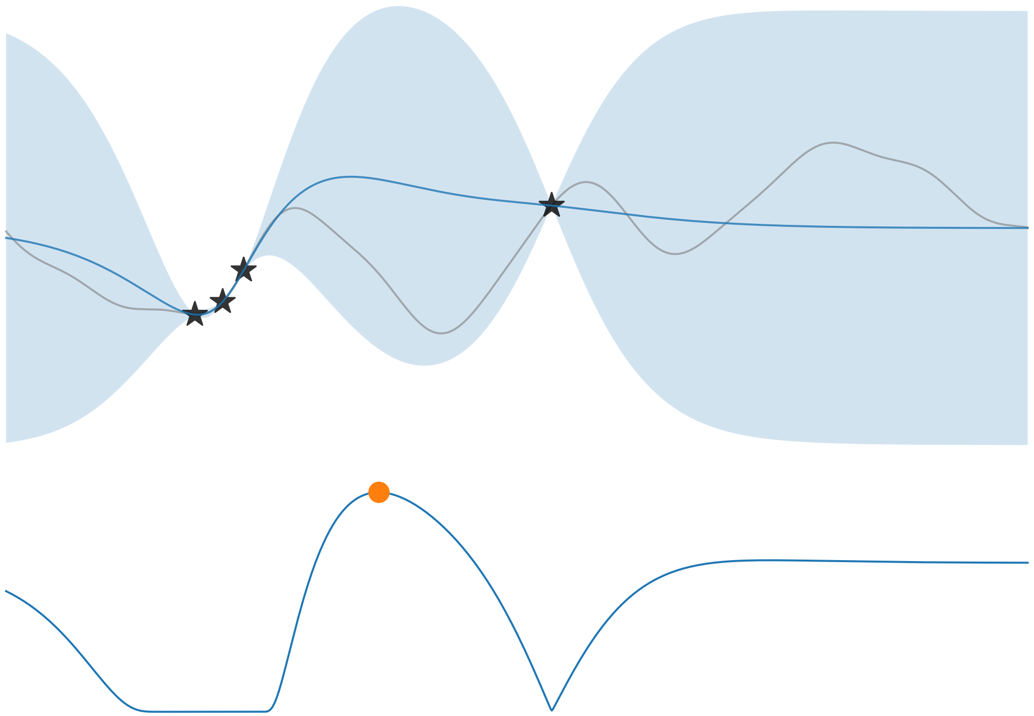
current best observed data

$$\mathrm{EI}(x) = \mathbb{E}[\underbrace{\max(f(x) - y_{\text{best}}, 0)}_{\text{“improvement”}} \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

One-step approximation to MDP

# Popular Policy: Expected Improvement



Other improvement-based policy:

- Probability of Improvement
- Knowledge Gradient
- Multi-step Lookahead EI
- …

# Approaches to Bayesian Optimization

- Improvement-based:
  - Expected Improvement
  - Probability of Improvement
  - Knowledge Gradient
  - Multi-step Lookahead EI

# Approaches to Bayesian Optimization

- Improvement-based:
  - Expected Improvement
  - Probability of Improvement
  - Knowledge Gradient
  - Multi-step Lookahead EI
- Entropy-based:
  - Max-value Entropy Search
  - Predictive Entropy Search

# Approaches to Bayesian Optimization

- Improvement-based:
  - Expected Improvement
  - Probability of Improvement
  - Knowledge Gradient
  - Multi-step Lookahead EI
- Entropy-based:
  - Max-value Entropy Search
  - Predictive Entropy Search
- Upper Confidence Bound

# Approaches to Bayesian Optimization

- Improvement-based:
  - Expected Improvement
  - Probability of Improvement
  - Knowledge Gradient
  - Multi-step Lookahead EI
- Entropy-based:
  - Max-value Entropy Search
  - Predictive Entropy Search
- Upper Confidence Bound
- Thompson Sampling

# Approaches to Bayesian Optimization

- Improvement-based
- Entropy-based
- Upper Confidence Bound
- Thompson Sampling

# Approaches to Bayesian Optimization

- Improvement-based
- Entropy-based
- Upper Confidence Bound
- Thompson Sampling
- Our work: Gittins Index

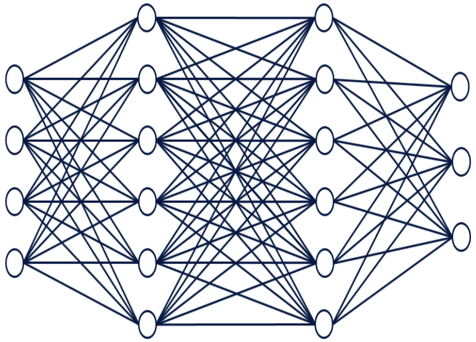
# Approaches to Bayesian Optimization

- Improvement-based
- Entropy-based
- Upper Confidence Bound
- Thompson Sampling
- Our work: Gittins Index

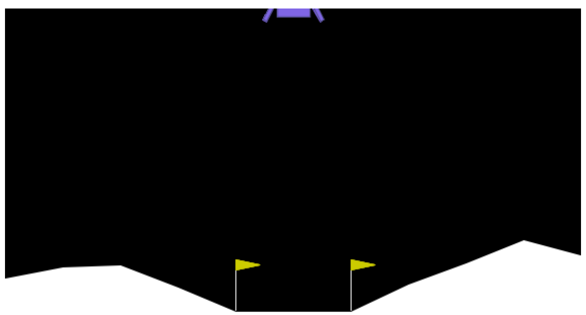
Why another approach?

# Another Challenge: Varying Evaluation Costs

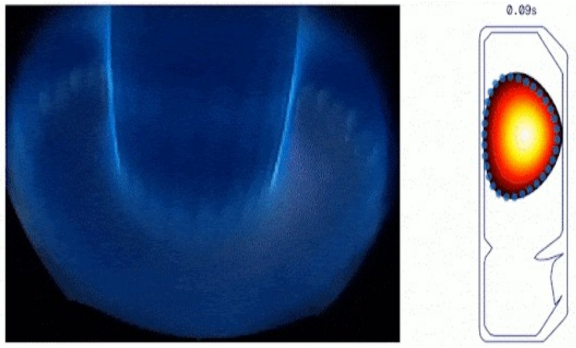
Hyperparameter tuning:

Training parameters →  → Accuracy

Control optimization:

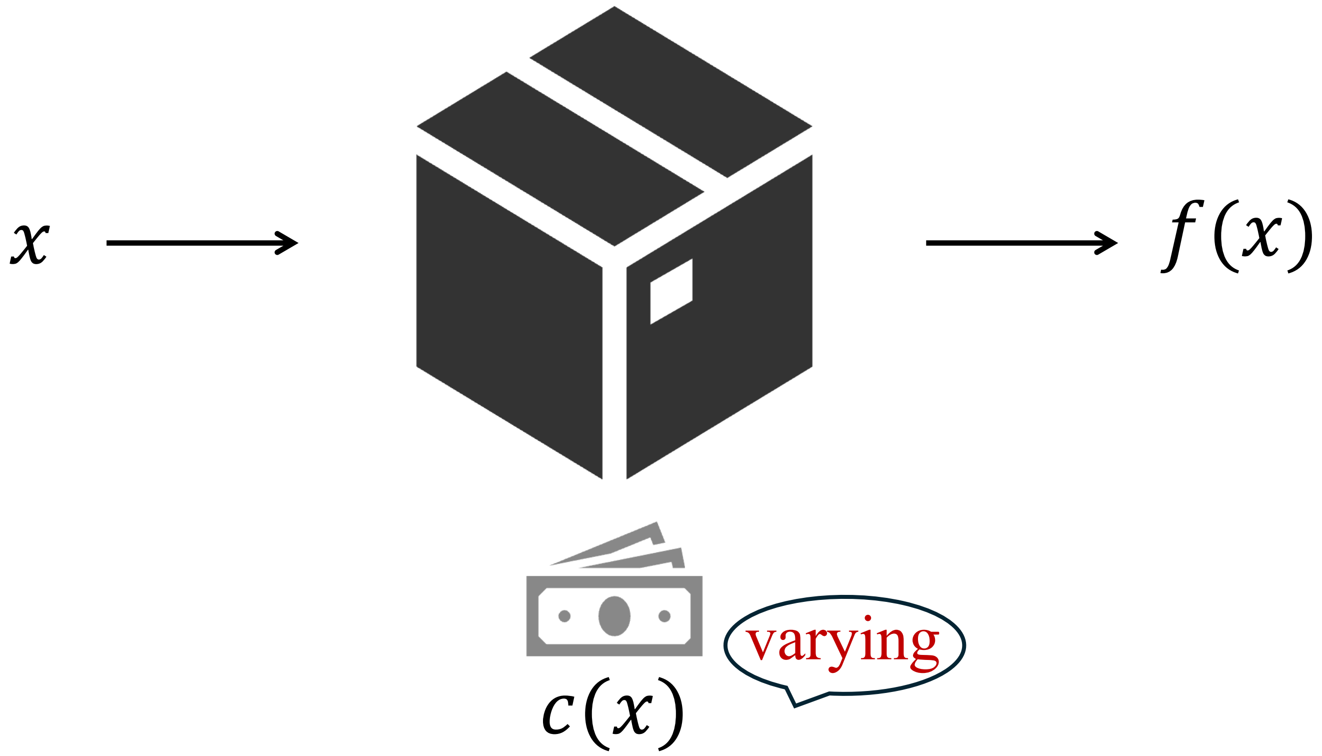
Control parameters →  → Reward

Plasma physics:

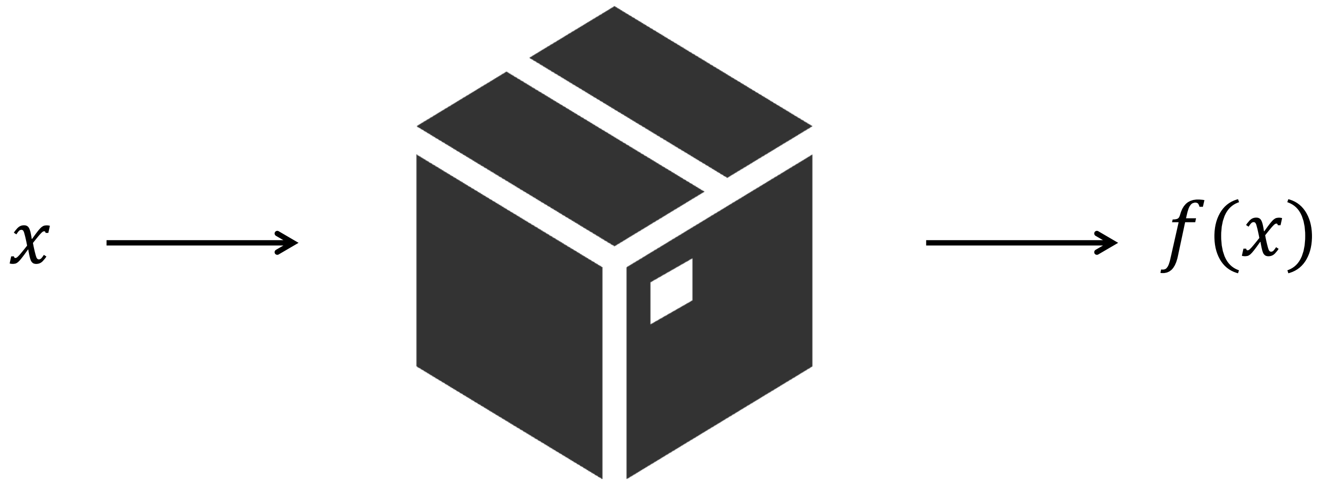
Reactor parameters → 

→ Stability

# Another Challenge: Varying Evaluation Costs



# Cost-aware Bayesian Optimization




$$\text{Goal: } \mathbf{max}\ \mathbb{E} \max_{t=1,2,\ldots,T} f(x_t)$$

$$\text{s.t. } \mathbb{E} \sum_{t=1}^{T} c(x_t) \le B$$

[Lee, Perrone, Archambeau, Seeger’21]

[Astudillo, Jiang, Balandat, Bakshy, Frazier’21]

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement<br>$\max_x \mathrm{EI}_{f\mid D}(x; y_{\text{best}})$ | |

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| | $\max_x \mathrm{EI}_{f\|D}(x; y_{\text{best}})$ | [Snoek, Larochelle, Adams'21] |

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| | $\max_x \mathrm{EI}_{f\|D}(x; y_{\mathrm{best}})$ | $\max_x \mathrm{EI}_{f\|D}(x; y_{\mathrm{best}})/c(x)$ |

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| | $\max_x \mathrm{EI}_{f\mid D}(x; y_{\text{best}})$ | $\max_x \mathrm{EI}_{f\mid D}(x; y_{\text{best}})/c(x)$ |

Why divide?

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
| --- | --- | --- |
| One-step | Expected improvement | Expected improvement per cost |
| | $\max_x \mathrm{EI}_{f\|D}(x; y_{\mathrm{best}})$ | $\max_x \mathrm{EI}_{f\|D}(x; y_{\mathrm{best}})/c(x)$ |

EI and EIPC policy can be arbitrarily bad under varying costs!

[Astudillo, Jiang, Balandat, Bakshy, Frazier'21]

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| Multi-step | Multi-step Lookahead EI | Budgeted Multi-step Lookahead EI |

slow

[Astudillo, Jiang, Balandat, Bakshy, Frazier'21]

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| Multi-step | Multi-step Lookahead EI | Budgeted Multi-step Lookahead EI |
| | Upper Confidence Bound | ? |
| | Thompson Sampling | ? |

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| Multi-step | Multi-step Lookahead EI | Budgeted Multi-step Lookahead EI |
| | Upper Confidence Bound | ? |
| | Thompson Sampling | ? |
| | ⋮ | ⋮ |

Our view: lack of a guidance to incorporate costs

# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| Multi-step | Multi-step Lookahead EI | Budgeted Multi-step Lookahead EI |
| | Upper Confidence Bound | ? |
| | Thompson Sampling | ? |
| | ⋮ | ⋮ |

New design principle: Gittins Index

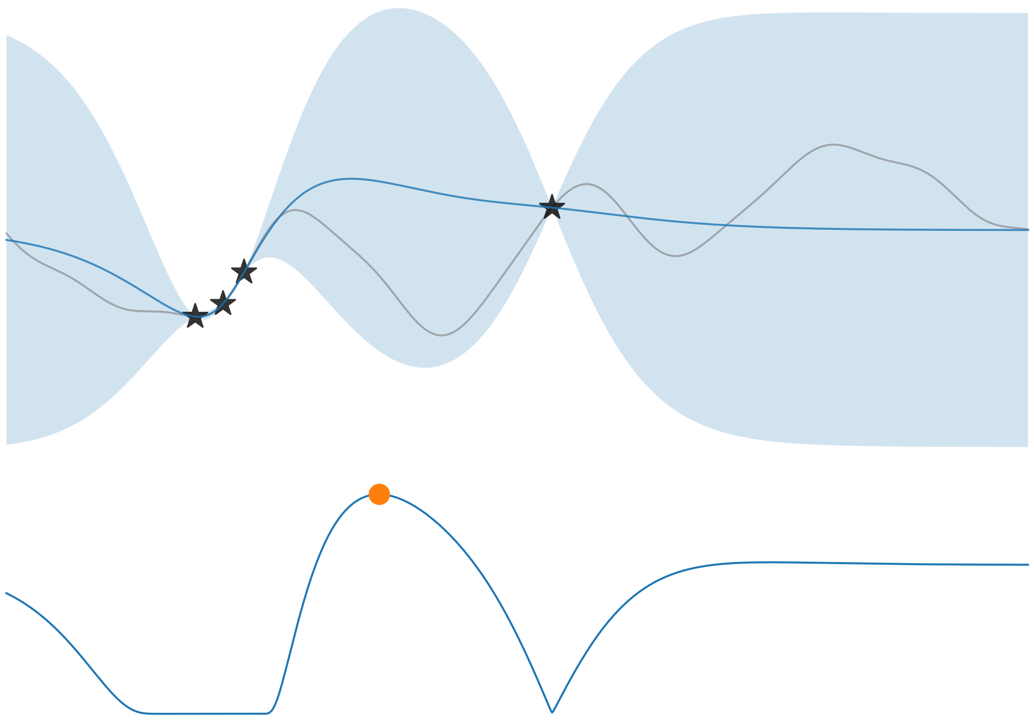
# Cost-aware Bayesian Optimization

| | Uniform costs | Varying costs |
|---|---|---|
| One-step | Expected improvement | Expected improvement per cost |
| Multi-step | Multi-step Lookahead EI | Budgeted Multi-step Lookahead EI |
| | Upper Confidence Bound | ? |
| | Thompson Sampling | ? |
| | ⋮ | ⋮ |

New design principle: Gittins Index

naturally cost-aware

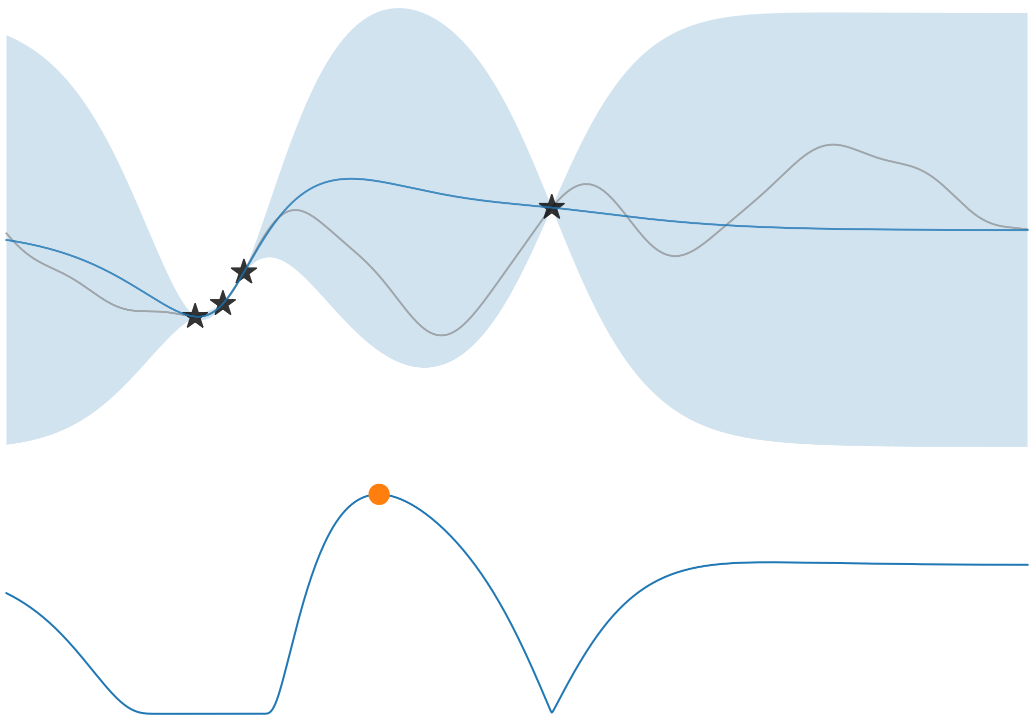
# Expected Improvement



$$\mathrm{EI}(x) = \mathbb{E}[\max(f(x) - y_{\text{best}}, 0) \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

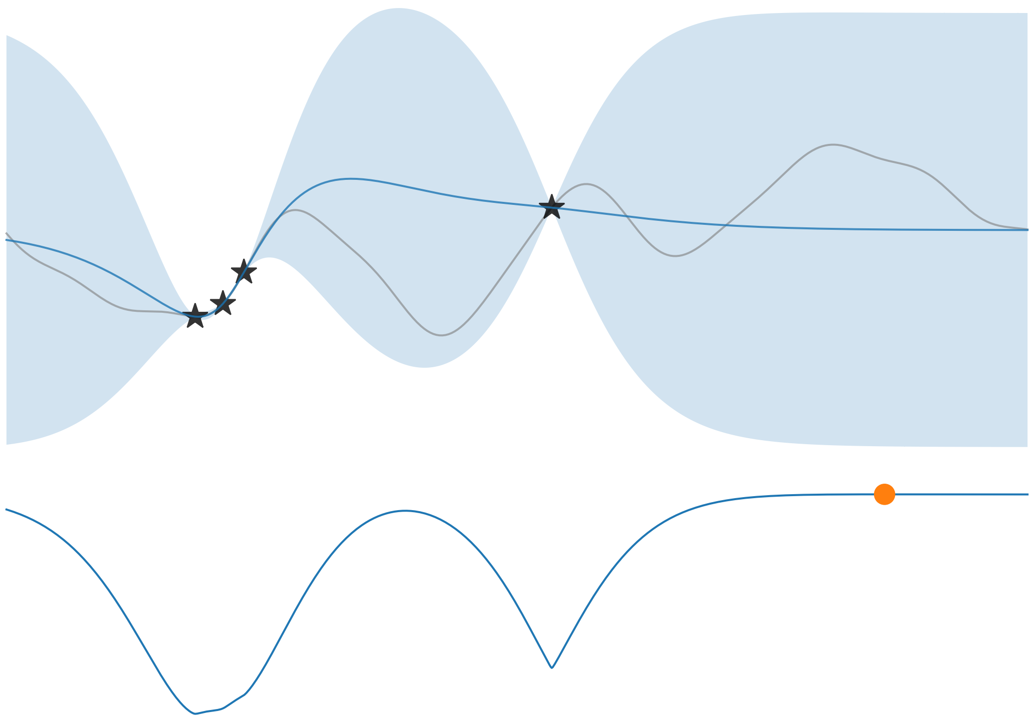
## Expected Improvement



$$\mathrm{EI}(x) = \mathbb{E}[\max(f(x) - y_{\text{best}}, 0) \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

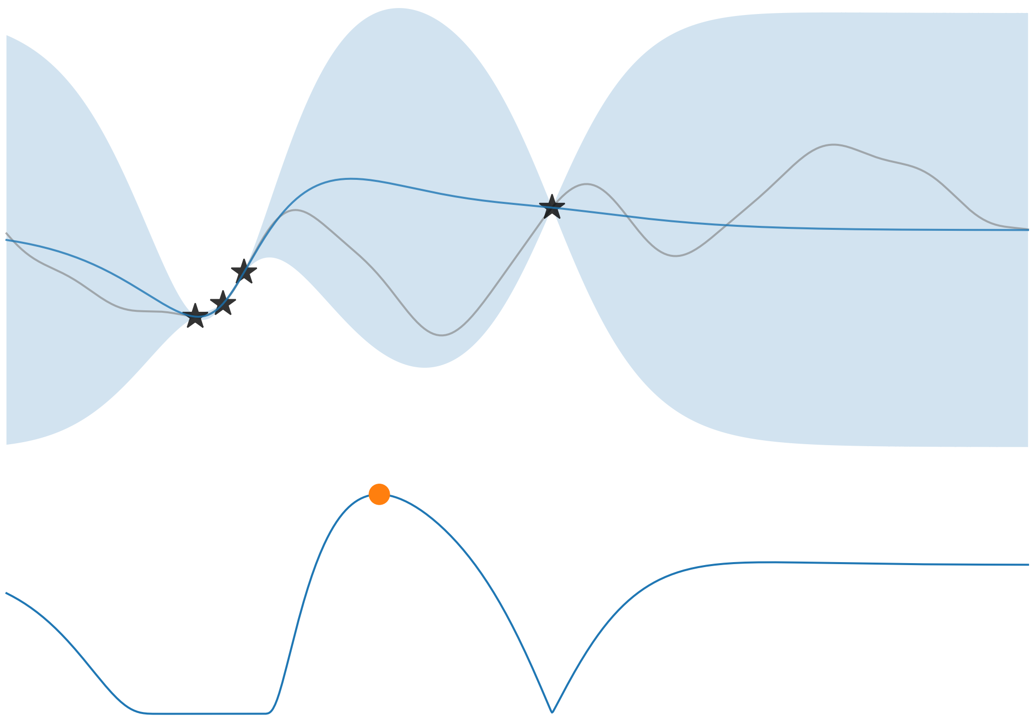
## Gittins Index



$$\mathrm{GI}_{f|D}(x) = g \text{ s.t. } \mathrm{EI}_{f|D}(x; g) = c(x)$$

$$\max_x \mathrm{GI}_{f|D}(x)$$

## Expected Improvement

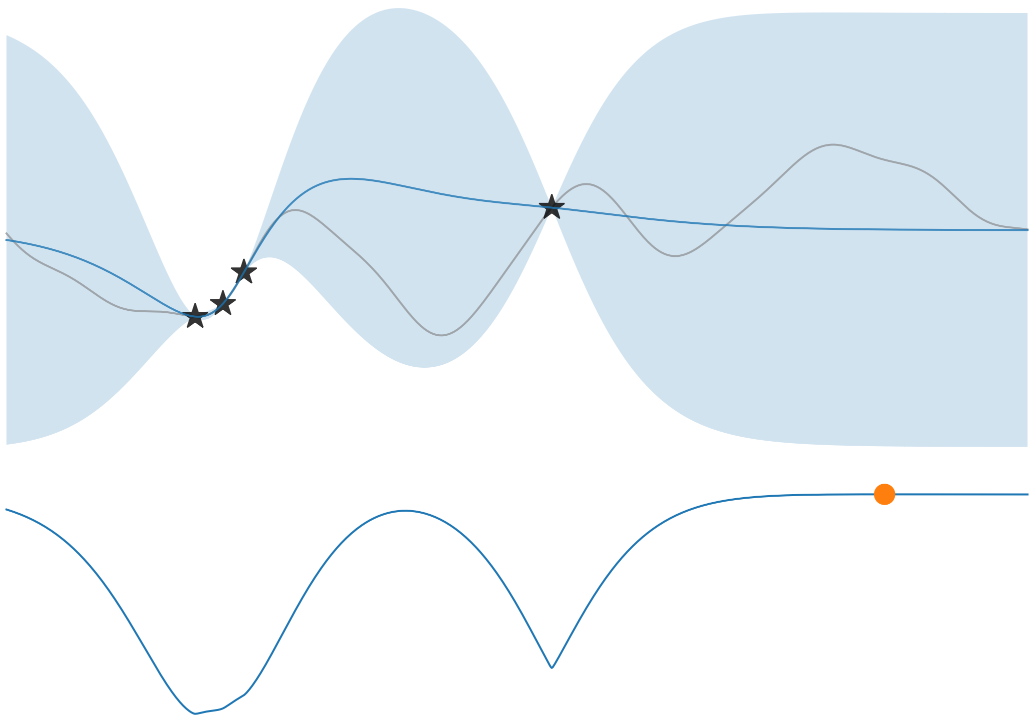


$$\mathrm{EI}(x) = \mathbb{E}[\max(f(x) - y_{\text{best}}, 0) \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

One-step approximation to MDP

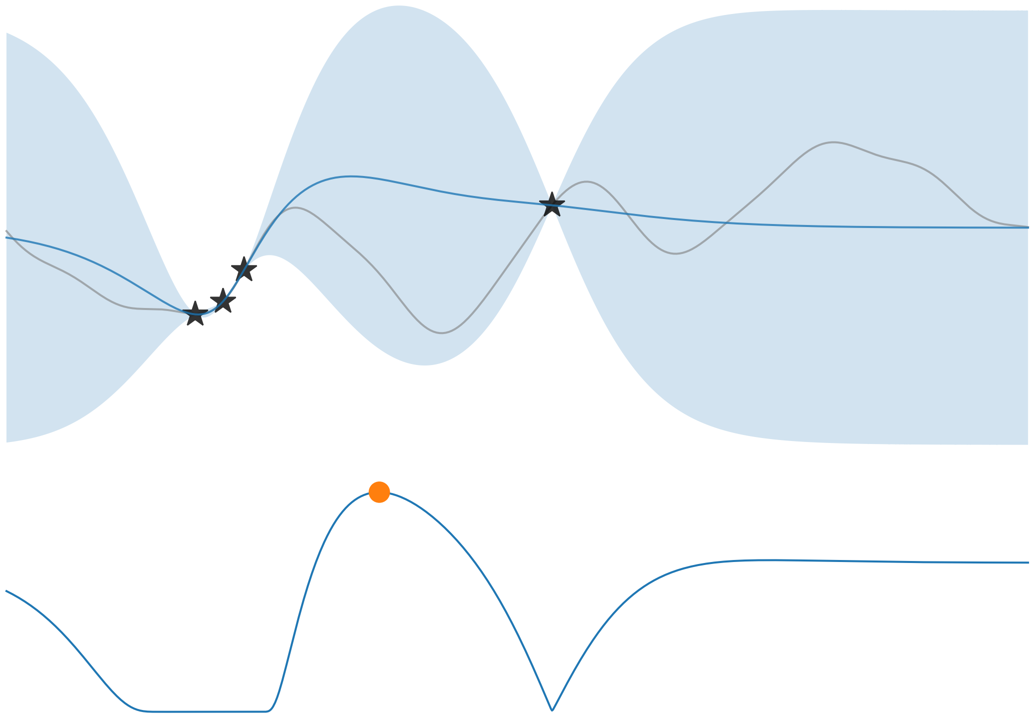
## Gittins Index



$$\mathrm{GI}_{f|D}(x) = g \text{ s.t. } \mathrm{EI}_{f|D}(x; g) = c(x)$$

$$\max_x \mathrm{GI}_{f|D}(x)$$

## Expected Improvement

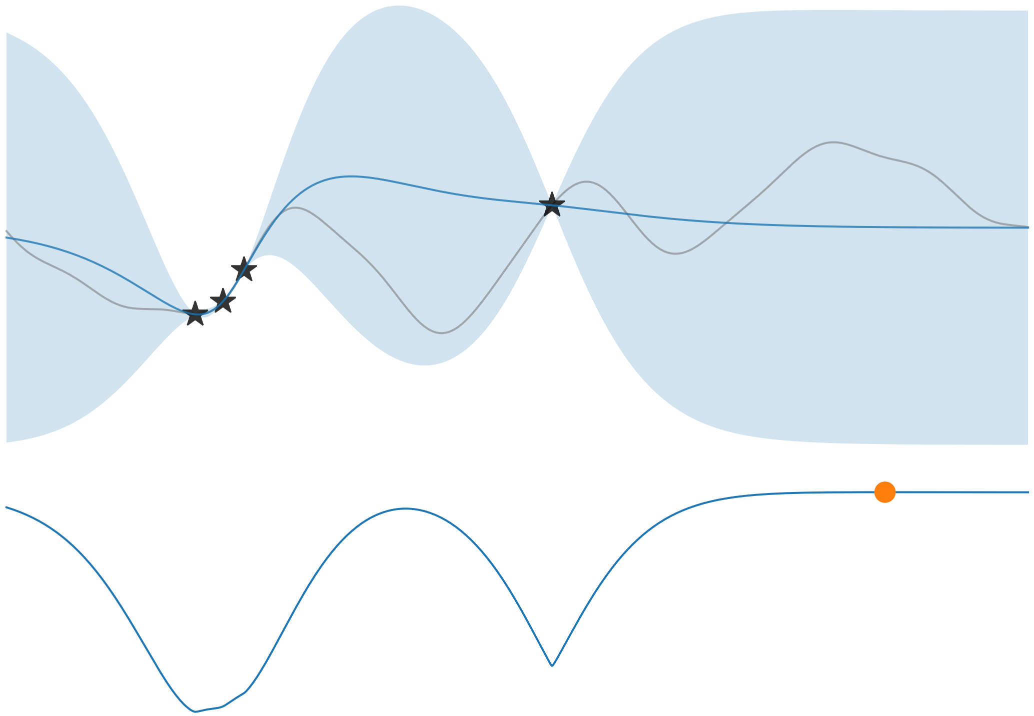


$$\mathrm{EI}(x) = \mathbb{E}[\max(f(x) - y_{\mathrm{best}}, 0) \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\mathrm{best}})$$

Temporal simplification to MDP

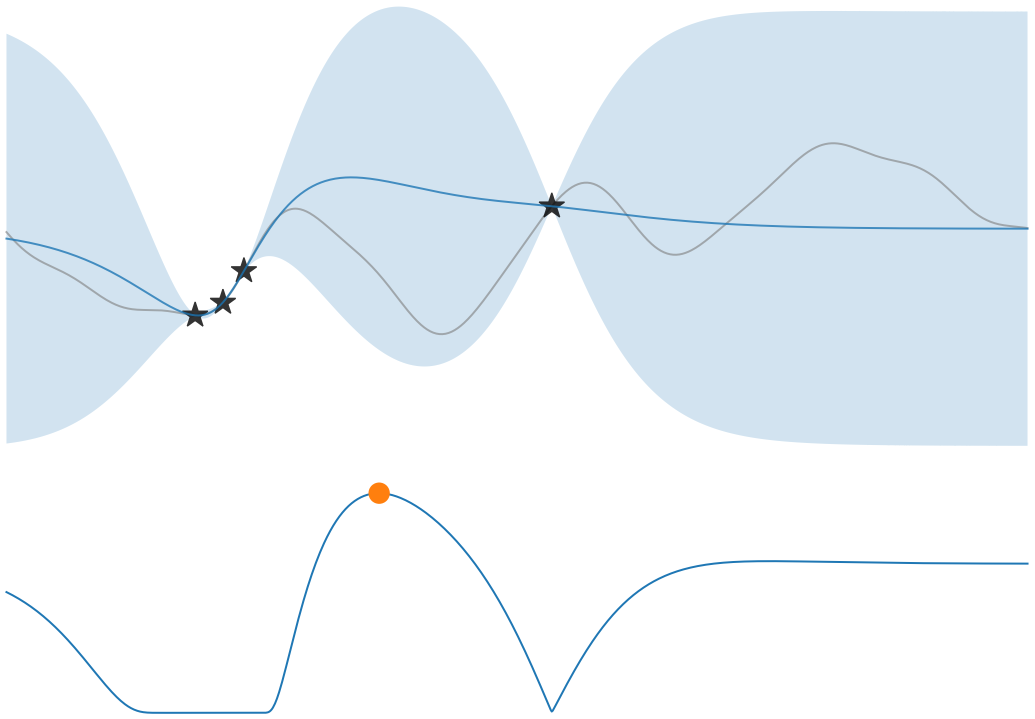
## Gittins Index



$$\mathrm{GI}_{f|D}(x) = g \text{ s.t. } \mathrm{EI}_{f|D}(x; g) = c(x)$$

$$\max_x \mathrm{GI}_{f|D}(x)$$

## Expected Improvement

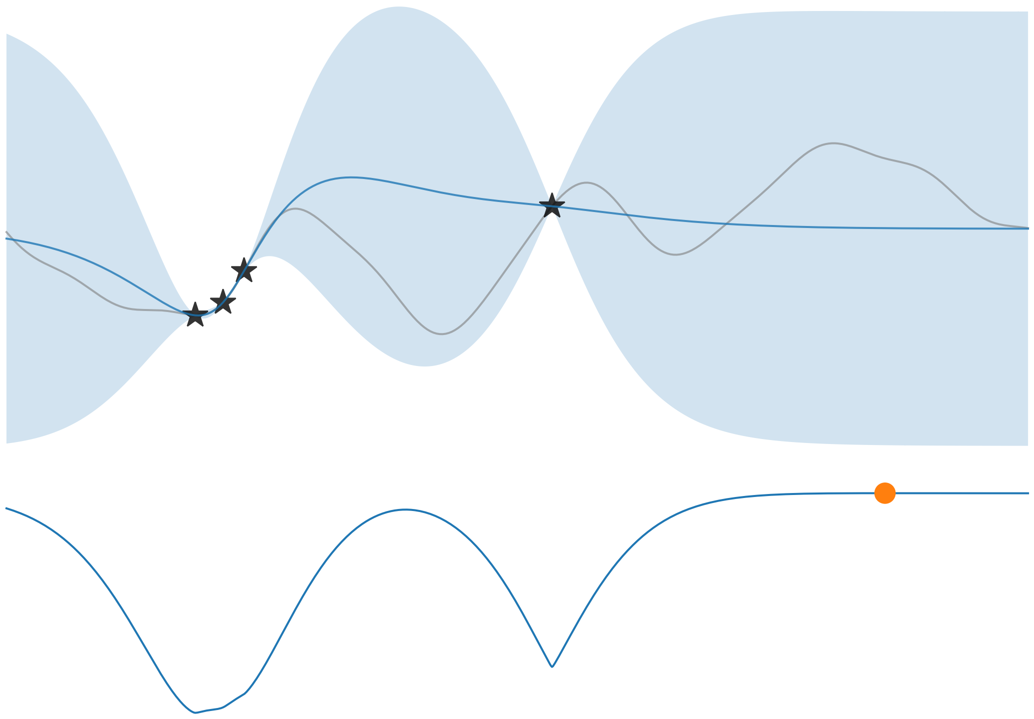


$$\mathrm{EI}(x) = \mathbb{E}[\max(f(x) - y_{\text{best}}, 0) \mid D]$$

$$\max_x \mathrm{EI}_{f|D}(x; y_{\text{best}})$$

Temporal simplification to MDP

## Gittins Index

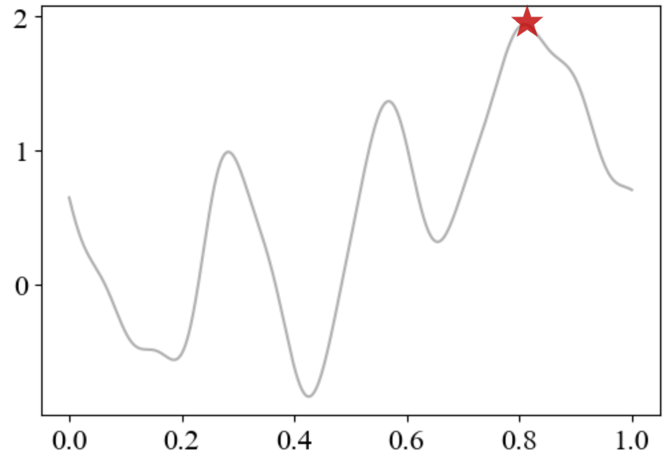


$$\mathrm{GI}_{f|D}(x) = g \text{ s.t. } \mathrm{EI}_{f|D}(x; g) = c(x)$$

$$\max_x \mathrm{GI}_{f|D}(x)$$

Spatial simplification to MDP

# Our Approach: Spatial Simplification

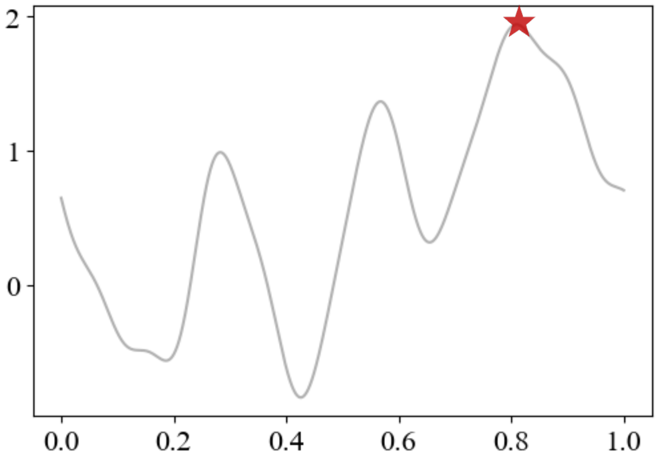



Bayesian Optimization

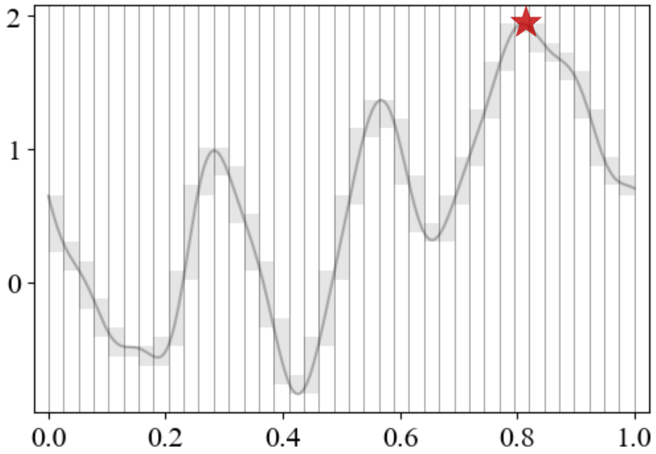
Continuous

Correlated

# Our Approach: Spatial Simplification




⟹

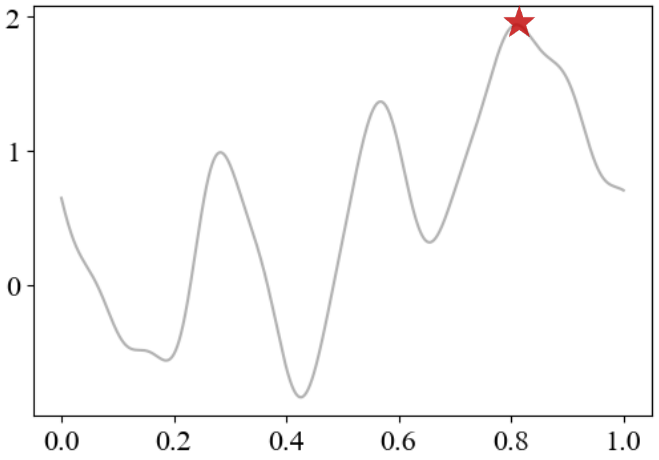



Bayesian Optimization

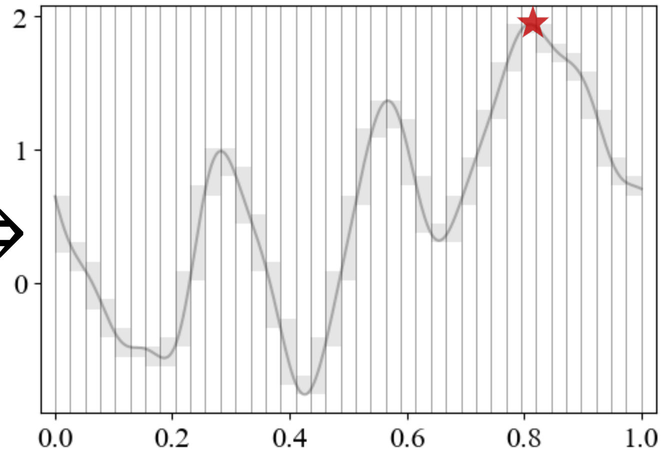
Continuous ⟹ Discrete

Correlated

# Our Approach: Spatial Simplification




$\Longrightarrow$

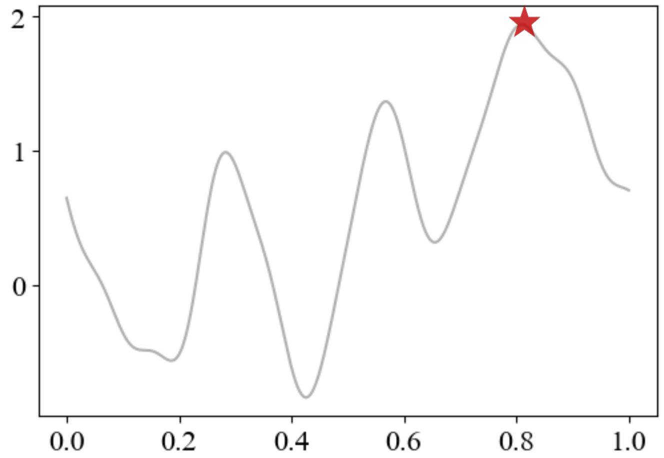



Bayesian Optimization

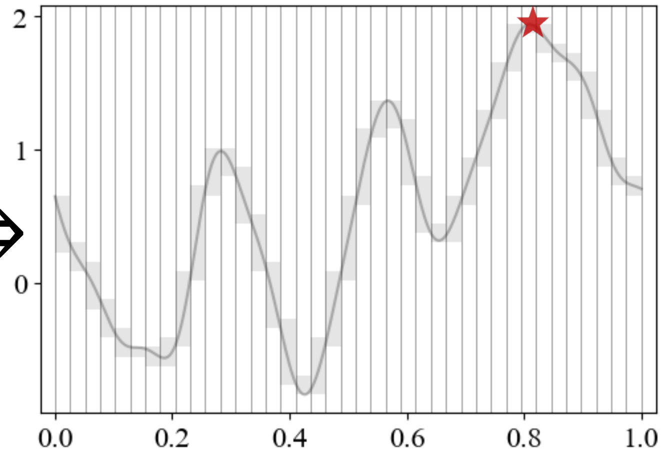
Continuous $\Longrightarrow$ Discrete

Correlated

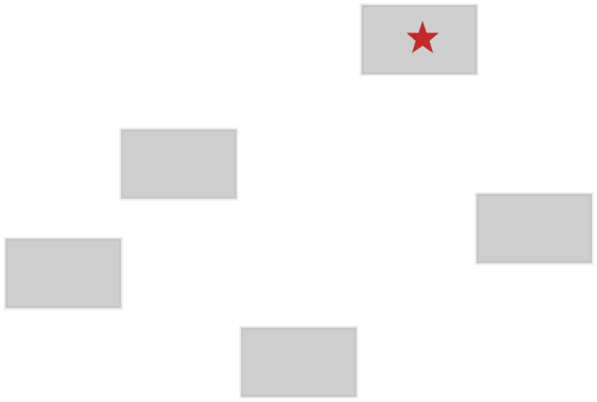
# Our Approach: Spatial Simplification




$\Longrightarrow$



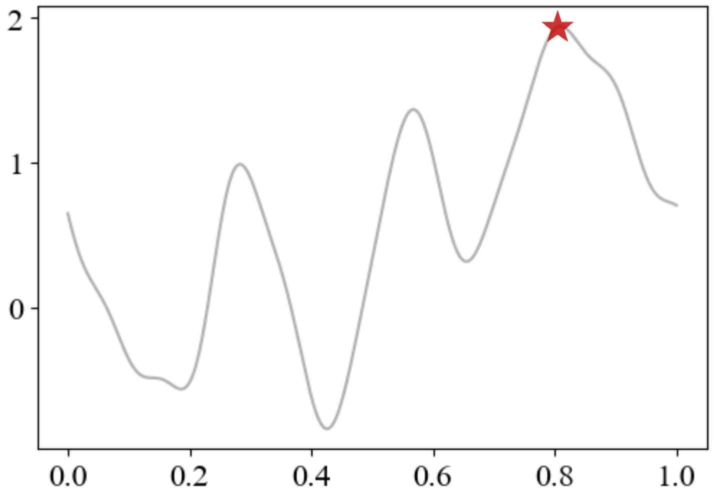

$\Longrightarrow$



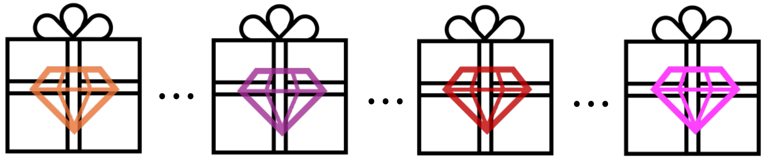
Bayesian Optimization

| | | |
|---|---|---|
| Continuous | $\Longrightarrow$ | Discrete |
| Correlated | $\Longrightarrow$ | Independent |

# Our Approach: Spatial Simplification




$\Longrightarrow$



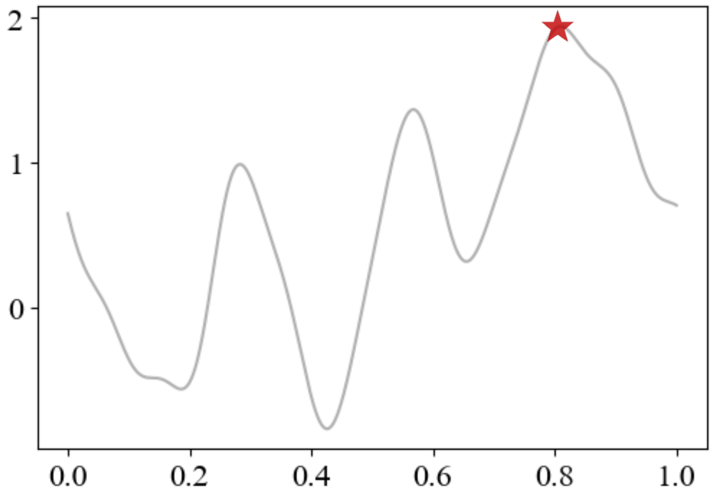
Bayesian Optimization

Pandora's Box [Weitzman'79]

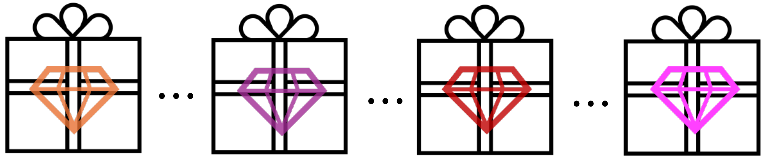
Continuous $\Longrightarrow$ Discrete

Correlated $\Longrightarrow$ Independent

# Our Approach: Spatial Simplification




$\Longrightarrow$



Bayesian Optimization

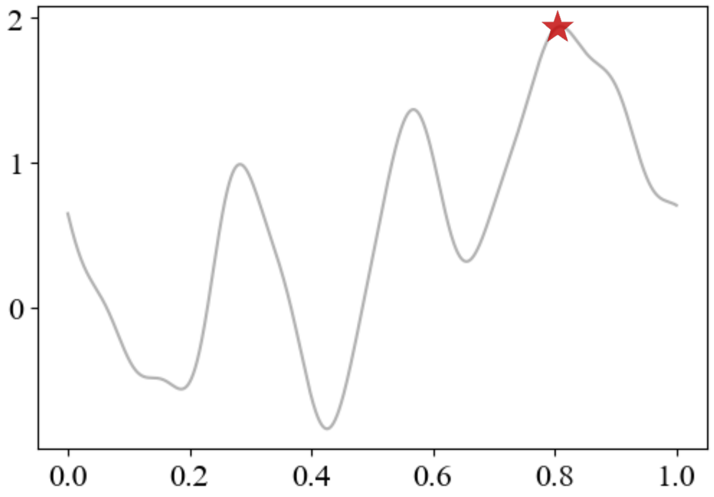
Pandora's Box [Weitzman'79]

Continuous $\Longrightarrow$ Discrete

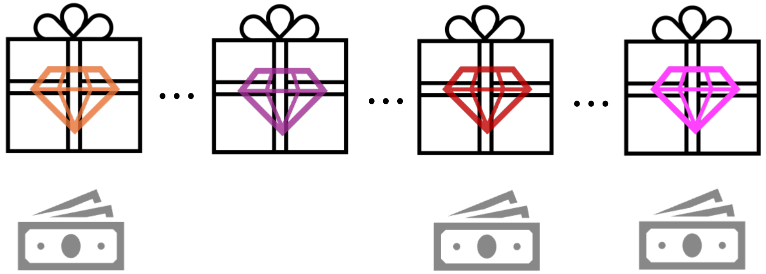
Correlated $\Longrightarrow$ Independent

Optimal policy: Gittins index

# Our Approach: Spatial Simplification




$\Longrightarrow$



Bayesian Optimization

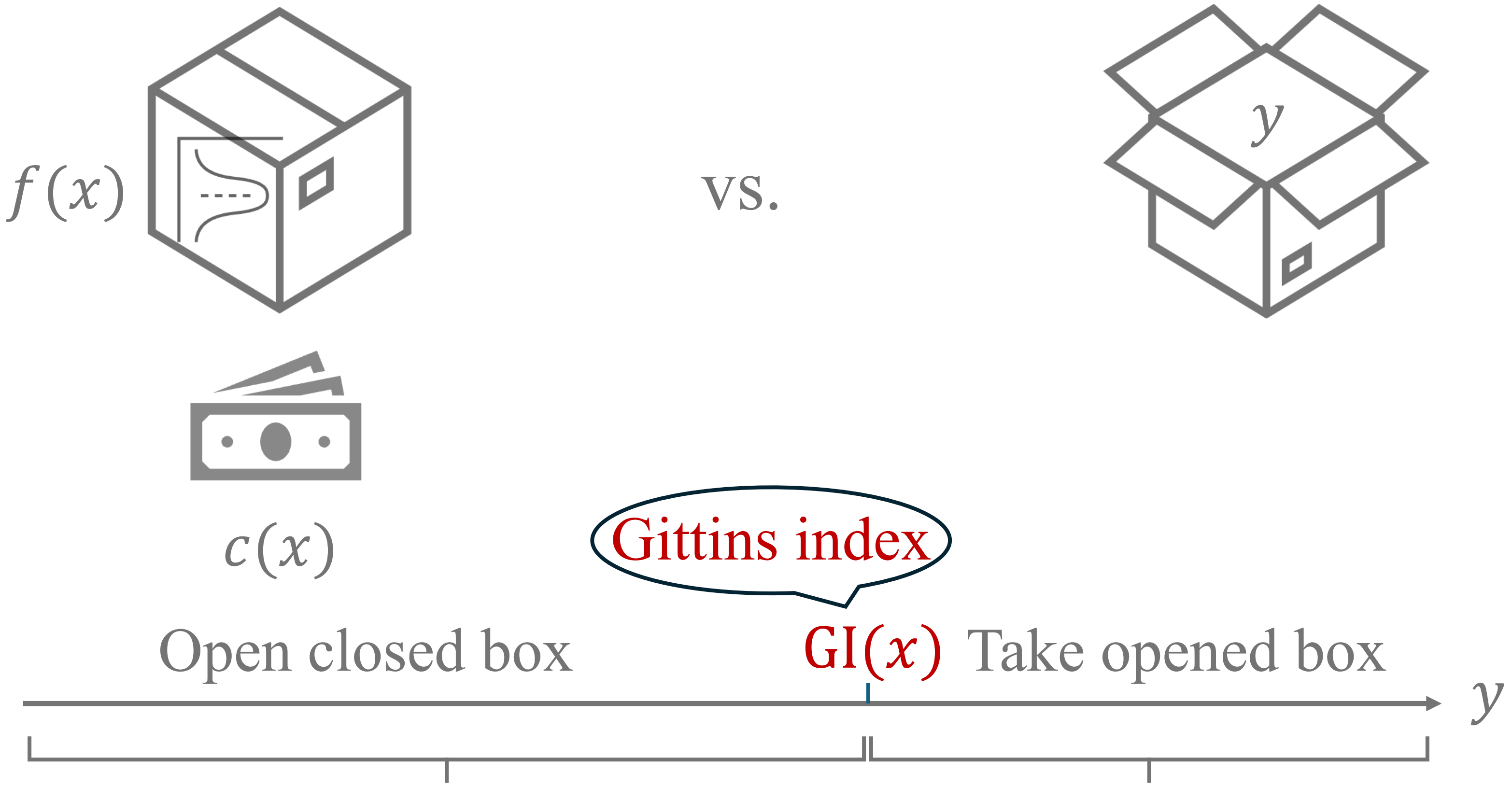
Pandora's Box [Weitzman'79]

| | | | |
|---|---|---|---|
| Continuous | $\Longrightarrow$ | Discrete | allow varying costs |
| Correlated | $\Longrightarrow$ | Independent | |

Optimal policy: Gittins index

# Intuition Behind Pandora's Box Gittins Index

## 1.5-box problem:

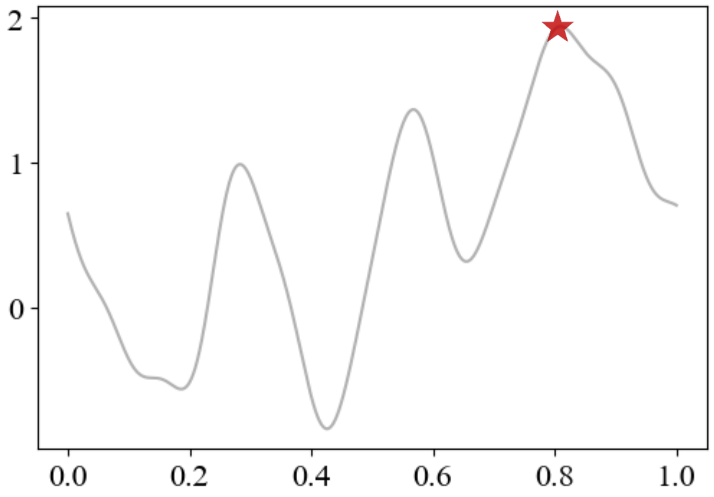



Profit: $\mathbb{E}[\max(f(x), y)] - c(x)$

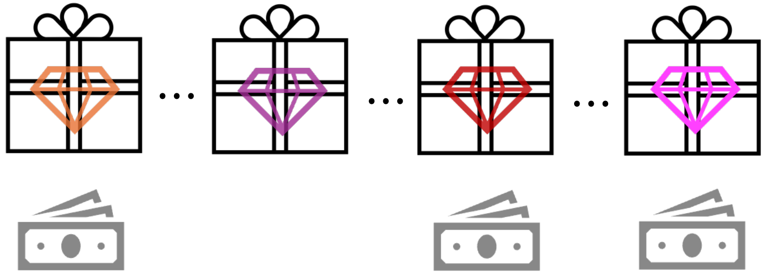
Profit: $y$

$\mathrm{GI}(x)$: solution $y$ to $\mathbb{E}[\max(f(x), y)] - c(x) = y$

# How to translate Gittins index?




$\Longrightarrow$



Bayesian Optimization

Pandora's Box [Weitzman'79]

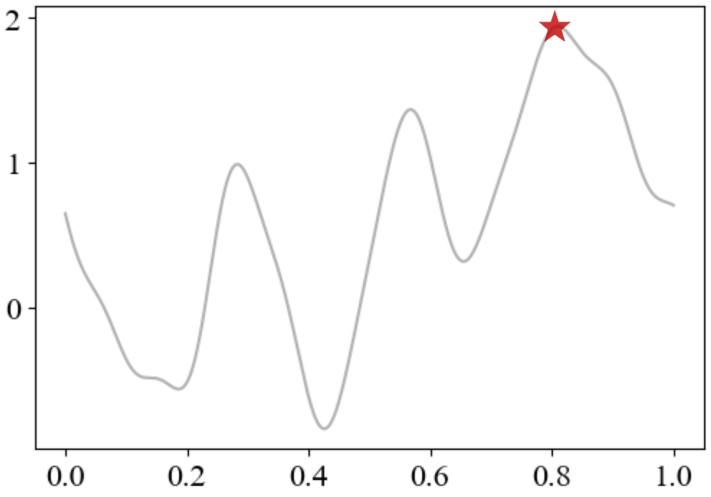
Continuous $\Longrightarrow$ Discrete

Correlated $\Longrightarrow$ Independent

How to translate?

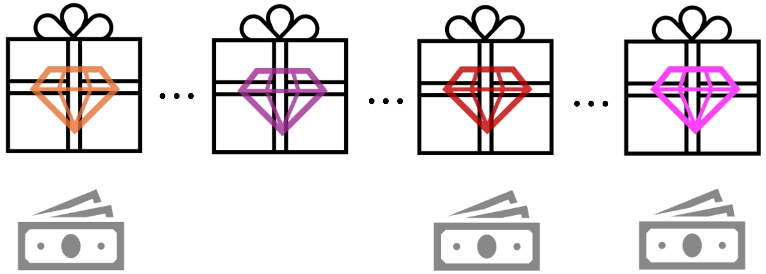
$\Longleftarrow$ Optimal policy: Gittins index

# How to translate Gittins index?




Bayesian Optimization

$\Longrightarrow$




Pandora's Box [Weitzman'79]

Continuous $\Longrightarrow$ Discrete

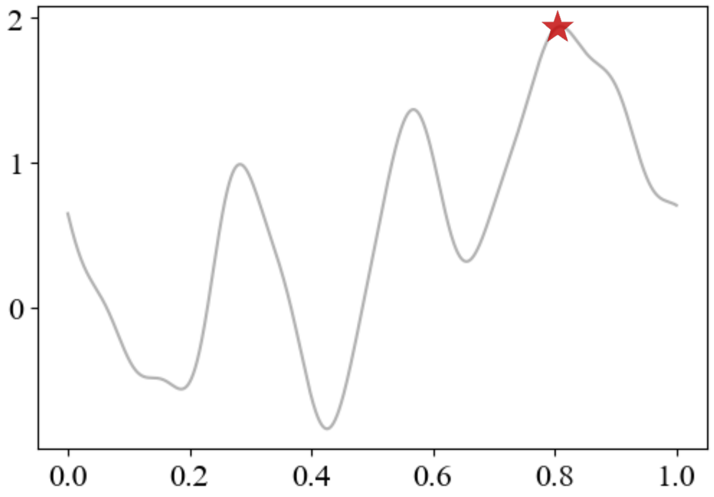
Correlated $\Longrightarrow$ Independent

incorporate posterior

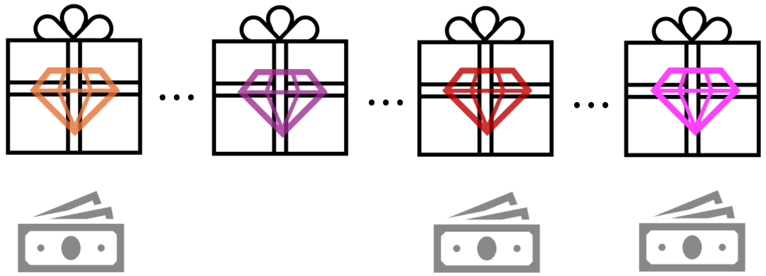
Our policy $\Longleftarrow$ Optimal policy: Gittins index

$\max_x \mathrm{GI}_{f|D}(x)$ $\max_x \mathrm{GI}(x)$

# Is Gittins good in Bayesian Optimization?




$\Longrightarrow$



Bayesian Optimization

Pandora's Box [Weitzman'79]

Continuous $\Longrightarrow$ Discrete

Correlated $\Longrightarrow$ Independent

Is Gittins index good? $\xLeftarrow{\text{incorporate posterior}}$ Gittins index is optimal

$\max_x \mathrm{GI}_{f|D}(x)$ $\qquad$ $\max_x \mathrm{GI}(x)$

# Experiment Setup: Objective Functions

Synthetic

Samples from prior

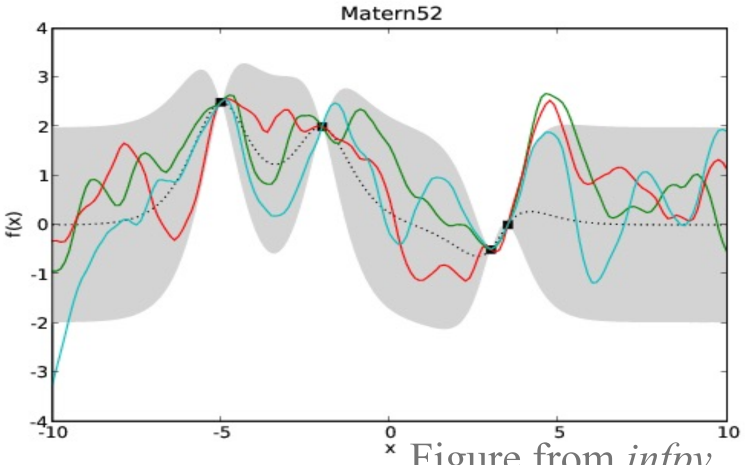



Figure from *infpy*

Ackley function

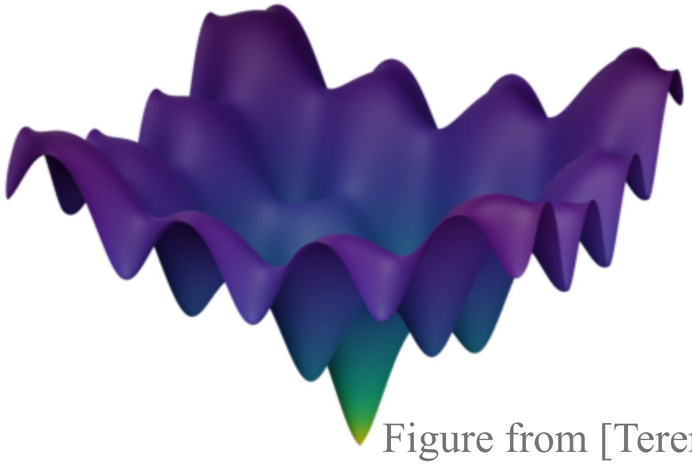


Figure from [Terenin'22]

Empirical

Pest Control

Figure from ChatGPT

Lunar Lander

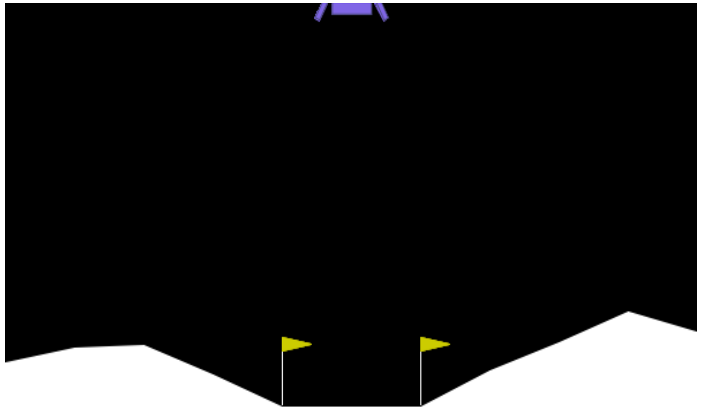
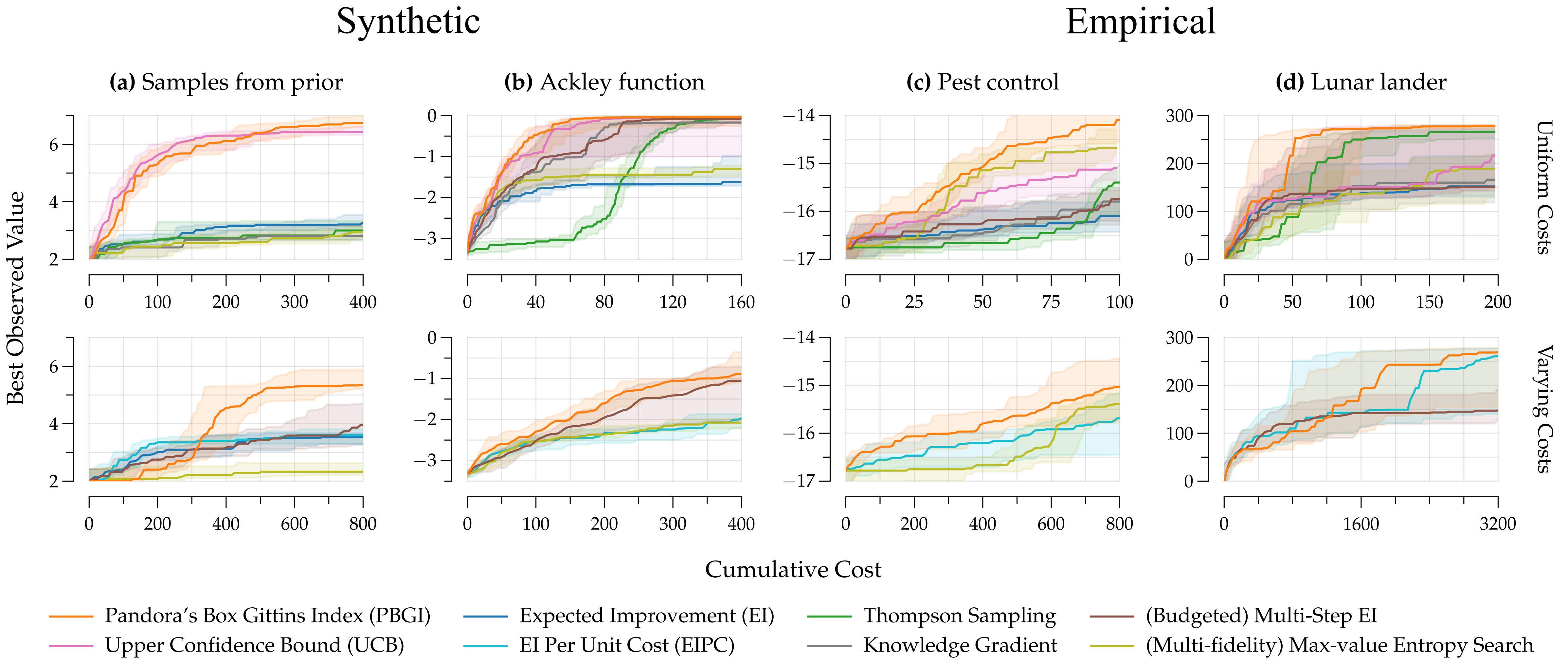


Figure from OpenAI Gym

# Experiment Results



# FAQ

- Easy-to-compute?

# FAQ

- Easy-to-compute?

  Yes, EI + bisection

# FAQ

- Easy-to-compute?

    Yes, EI + bisection
- Any theoretical results?

# FAQ

- Easy-to-compute?

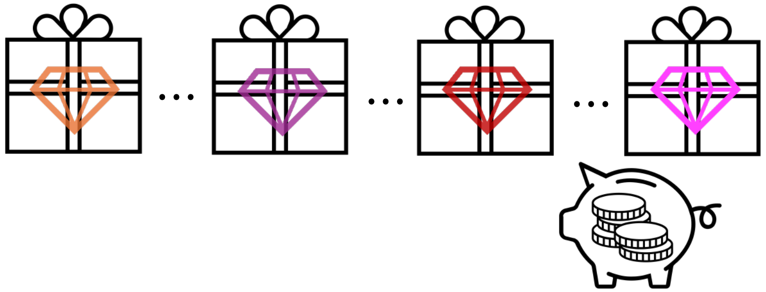
  Yes, EI + bisection

- Any theoretical results?

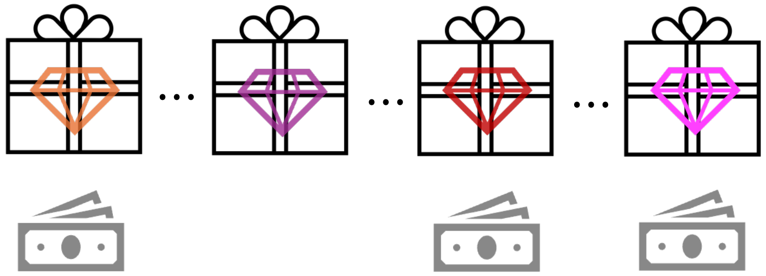
  Yes, expected-budget-constrained $\cong$ cost-per-sample

# Theoretical Result

**max** best observed under budget



**max** (best observed – costs)



Budgeted Pandora's Box

Expected budget constraint

Optimal policy?

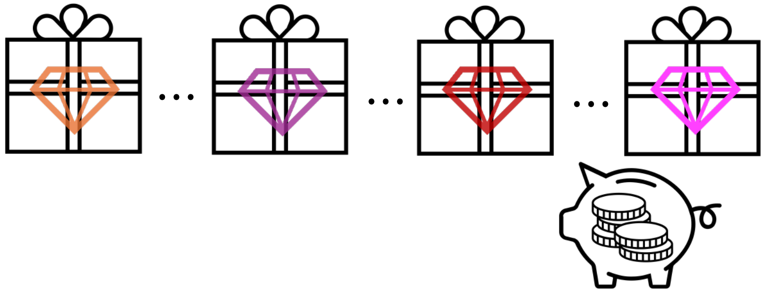
Pandora's Box

Cost per sample

Optimal policy: Gittins index

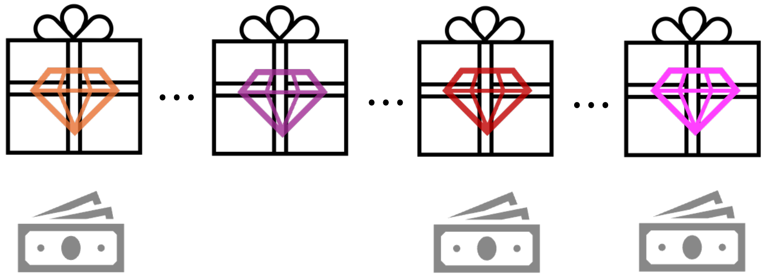
# Theoretical Result

**max** best observed under budget



$\Longleftarrow$

**max** (best observed – scaled costs)



Budgeted Pandora's Box

Expected budget constraint

Optimal policy ✓

extension to [Aminian, Manshadi, Niazadeh'24]

Pandora's Box

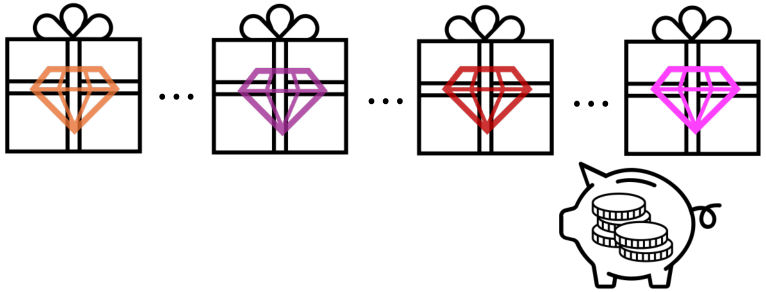
Cost per sample

Optimal policy: Gittins index

# Theoretical Result

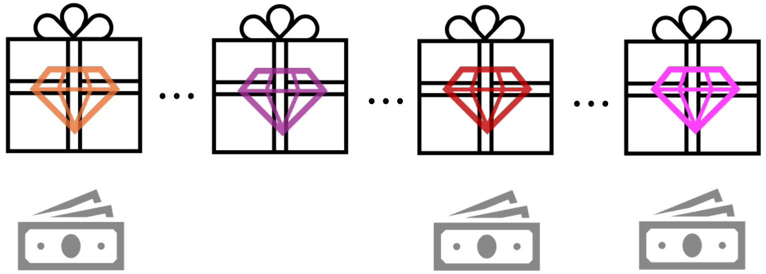
budget-dependent

**max** best observed under budget



$\Longleftarrow$

**max** (best observed – scaled costs)



| Budgeted Pandora’s Box | Pandora’s Box |
|---|---|
| Expected budget constraint | Cost per sample |
| Optimal policy ✔ | Optimal policy: Gittins index |

extension to [Aminian, Manshadi, Niazadeh’24]

# FAQ

- Easy-to-compute?

  Yes, EI + bisection

- Any theoretical results?

  Yes, expected-budget-constrained $\cong$ cost-per-sample

- Tuning parameters?

# FAQ

- Easy-to-compute?

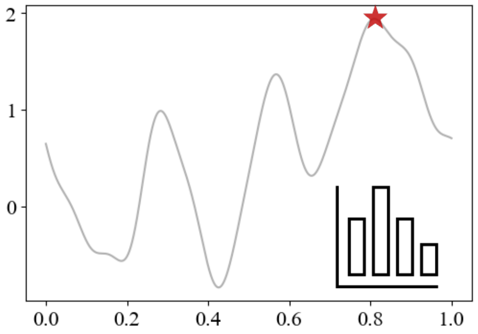
  Yes, EI + bisection
- Any theoretical results?

  Yes, expected-budget-constrained ≅ cost-per-sample
- Tuning parameters?

  Yes, control unit conversion

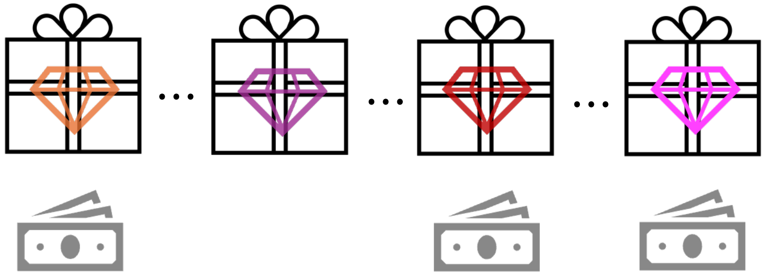
# New Design Principle: Gittins Index

## Problem



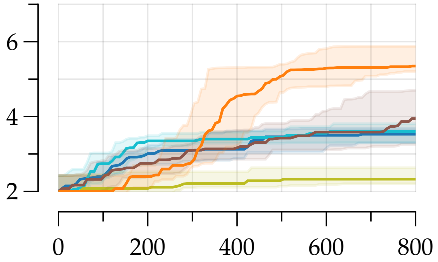

Bayesian optimization
with varying costs

## Key idea



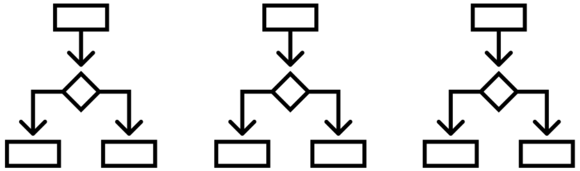
link to Pandora's box
and Gittins index theory

## Impact




competitive performance

## Future potential



black-box processes
with partial feedback

# Check our paper on ArXiv!

"Cost-aware Bayesian Optimization via the Pandora's Box Gittins Index."