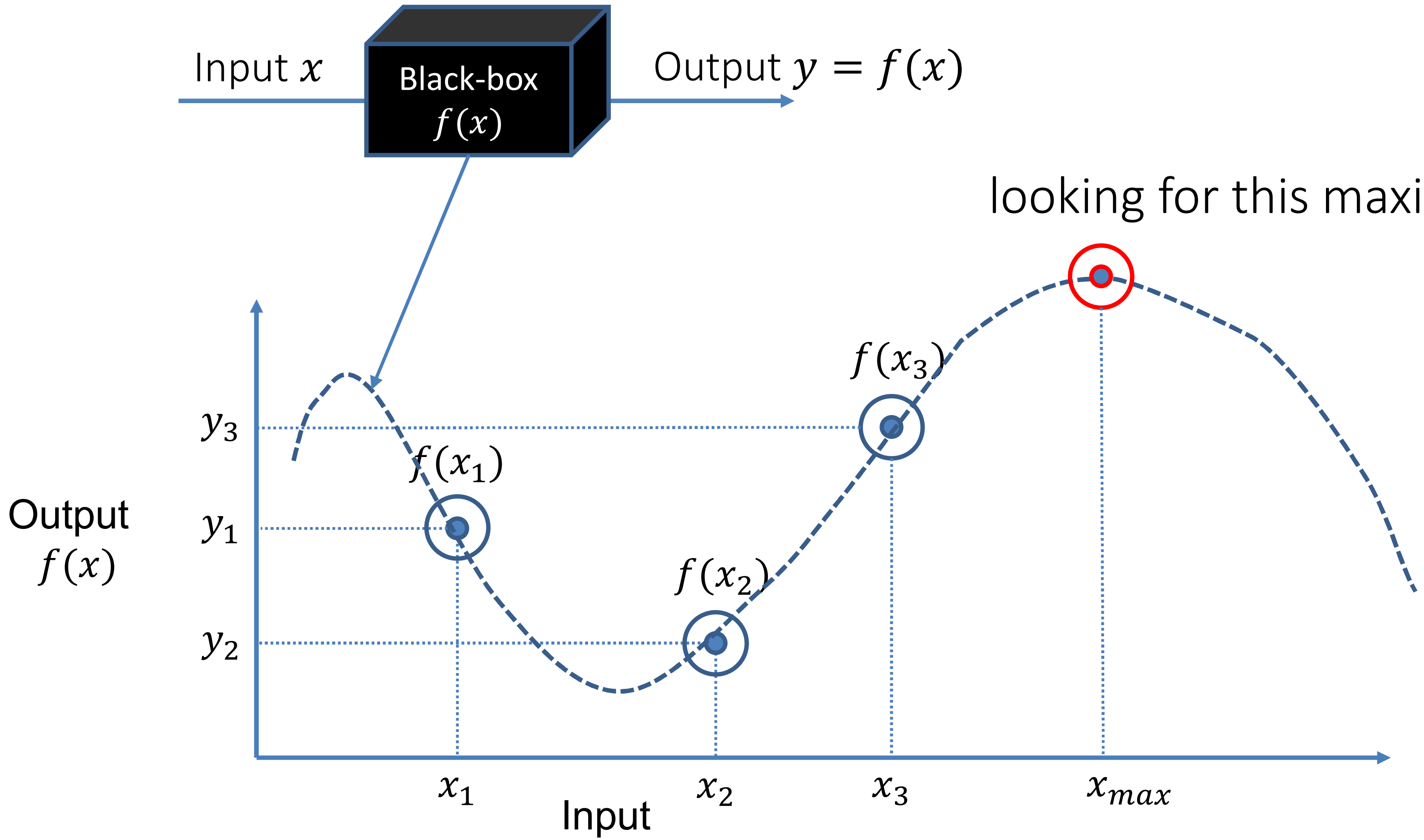
Knowing The What, But Not The Where in Bayesian Optimization

Vu Nguyen & Michael A. Osborne

University of Oxford

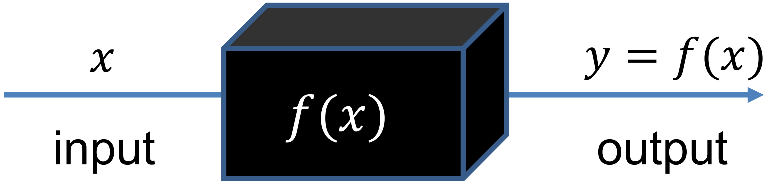
Black-box Optimization

The relationship from x to y is through the black-box.



Properties of Black-box Function

$$f: X \in \mathcal{R}^d \rightarrow Y \in \mathcal{R}^1$$



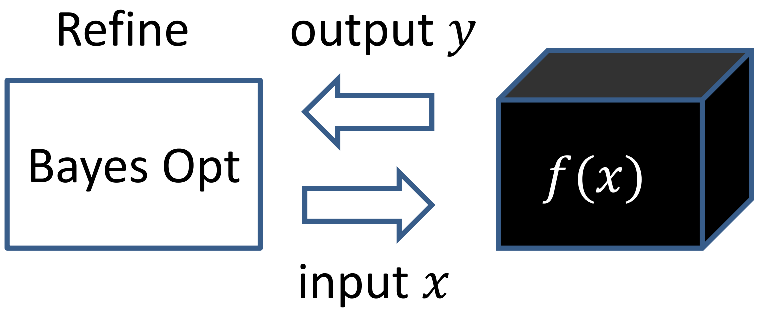
Function form is not known $\cancel{y = ax + b}$

No derivative form $\cancel{\frac{\partial f}{\partial x} = \cdots}$

Expensive to evaluate (in time and cost)

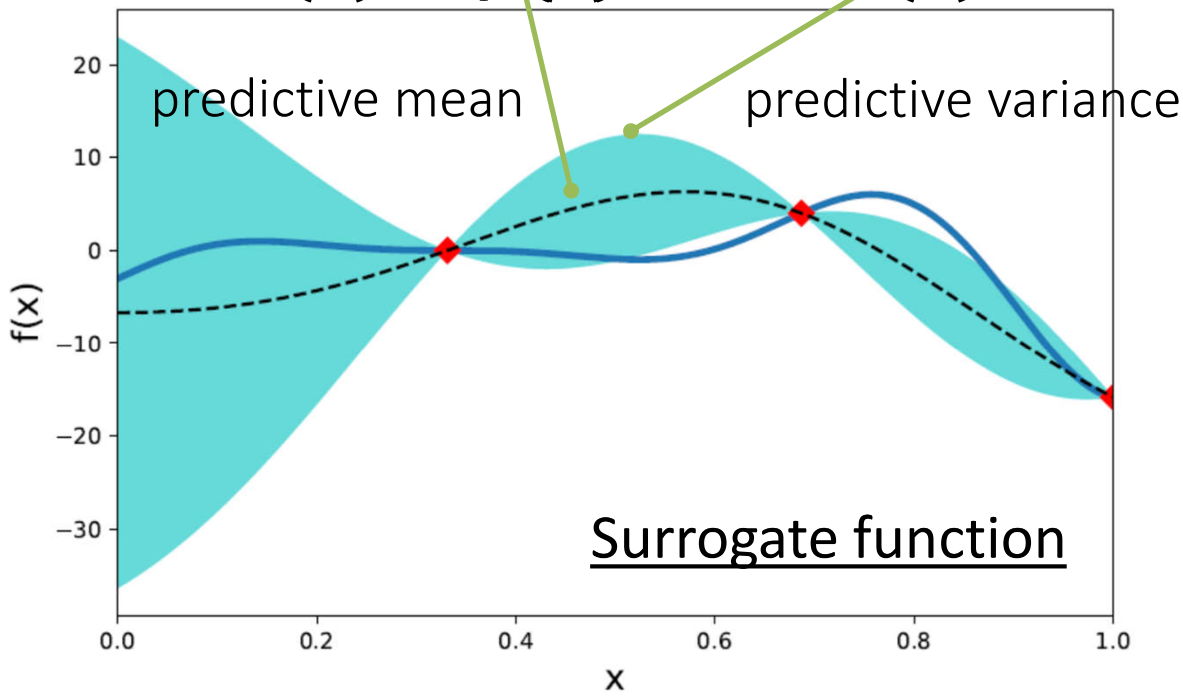
Nothing is known about the function, except a few evaluations $y = f(x)$

Bayesian Optimization Overview



- Make a series of evaluations $x_1, x_2, \ldots x_T$

Acquisition function $\alpha(x) = \overset{\text{exploit}}{\boldsymbol{\mu}(\boldsymbol{x})} + \kappa \times \overset{\text{explore}}{\boldsymbol{\sigma}(x)}$

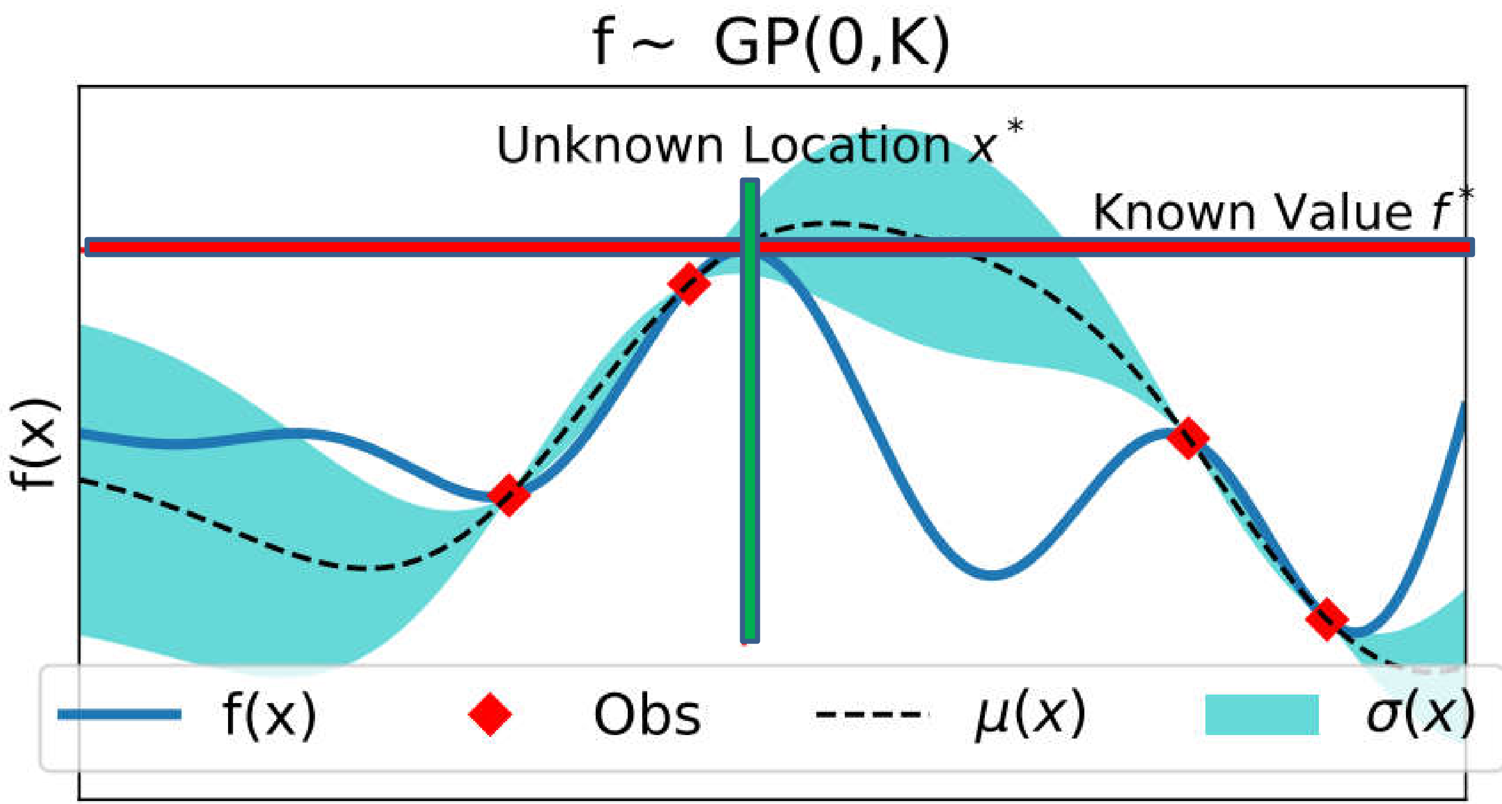


Outline

- Bayesian Optimization
- Bayes Opt with Known Optimum Value

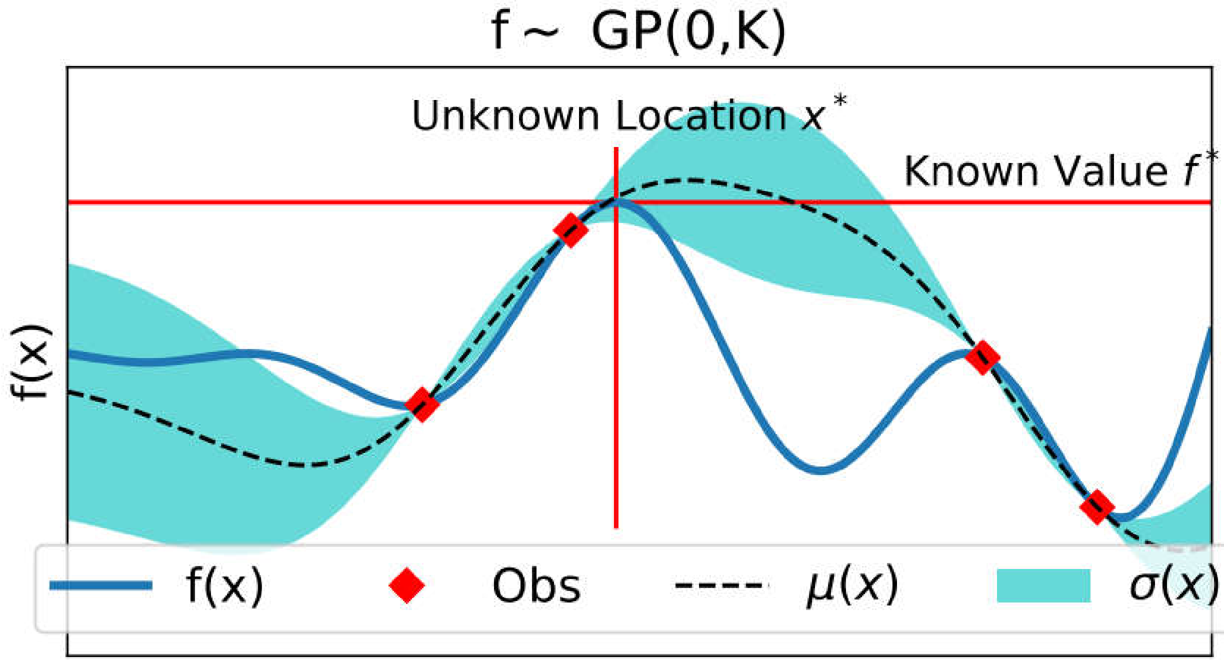
Knowing Optimum Value of The Black-Box

- We consider situations where the **optimum value** is known.
- $f^* = \max f(x)$ and the goal is to find $x^* = \arg\max f(x)$.



Examples of Knowing Optimal Value of The Black-Box

- Deep reinforcement learning:
  - CartPole: 200
  - Pong: 18
  - Frozen Lake: 0.79 ± 0.05
  - InvertedPendulum: 950



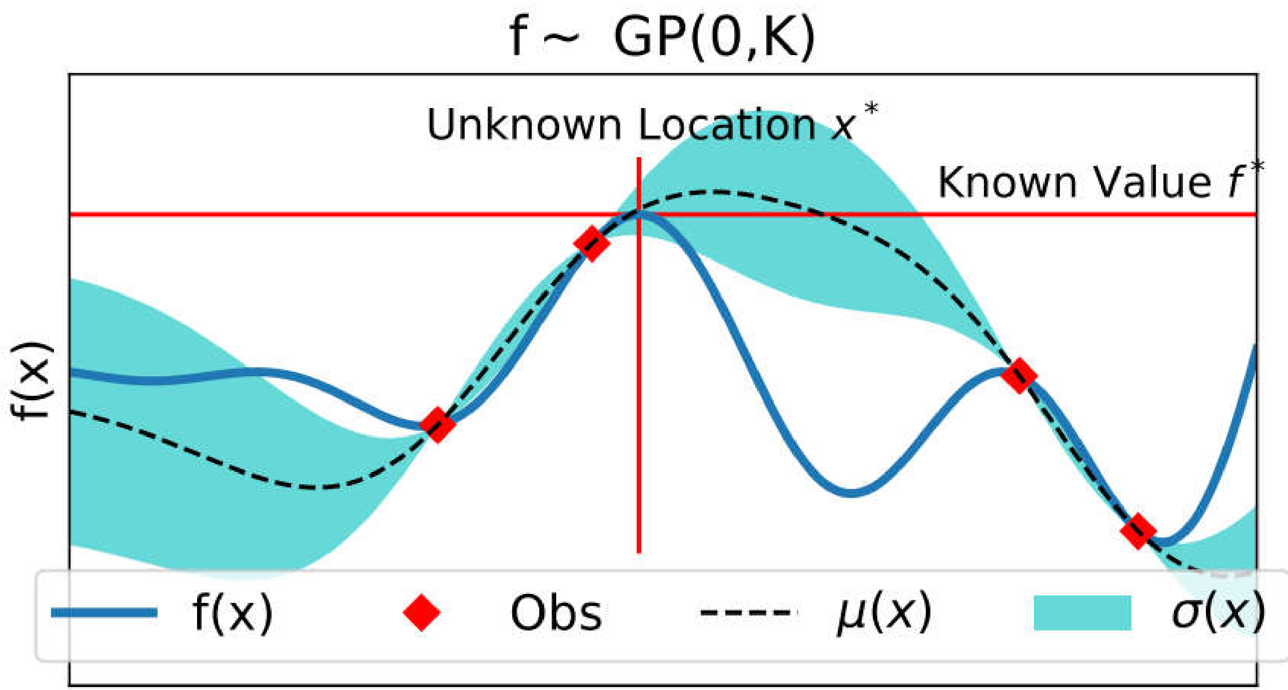
- Classification:
  - Skin dataset: Accuracy 100

- Inverse optimization:
  - Given a database and a target property t, identifying a corresponding data point x^*.

What can f^* tell us about f ?

1. f^* tells us about the upper bound: $f^* \geq f(x), \forall x$

2. f^* tells us that the function is reaching f^* at some points.



Transformed Gaussian process

$$f(x) = f^* - \frac{1}{2}\underbrace{g^2(x)}_{\geq 0}$$

$$g(x) \sim GP(\sqrt{2f^*}, K)$$

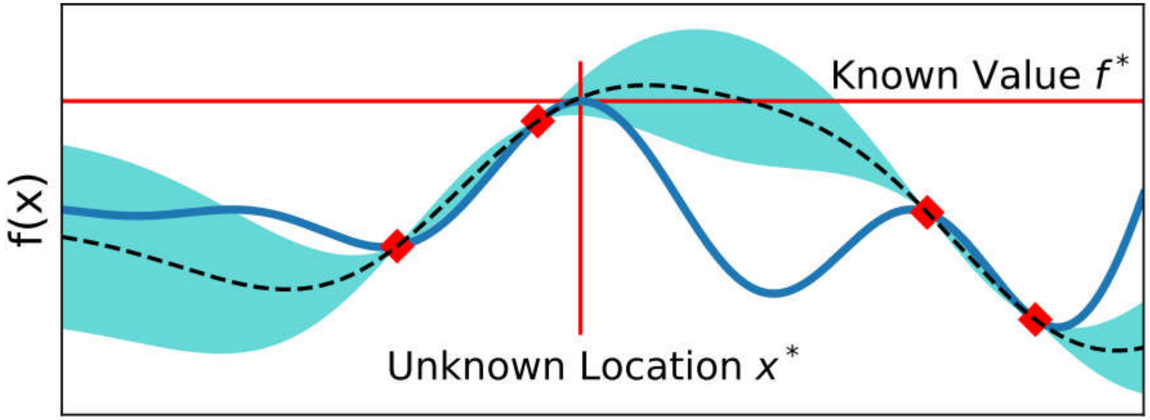
This condition ensures that $f^* \geq f(x), \forall x$ 1

We want to control the surrogate using f^*

1 Push down: the surrogate must not go above f^*

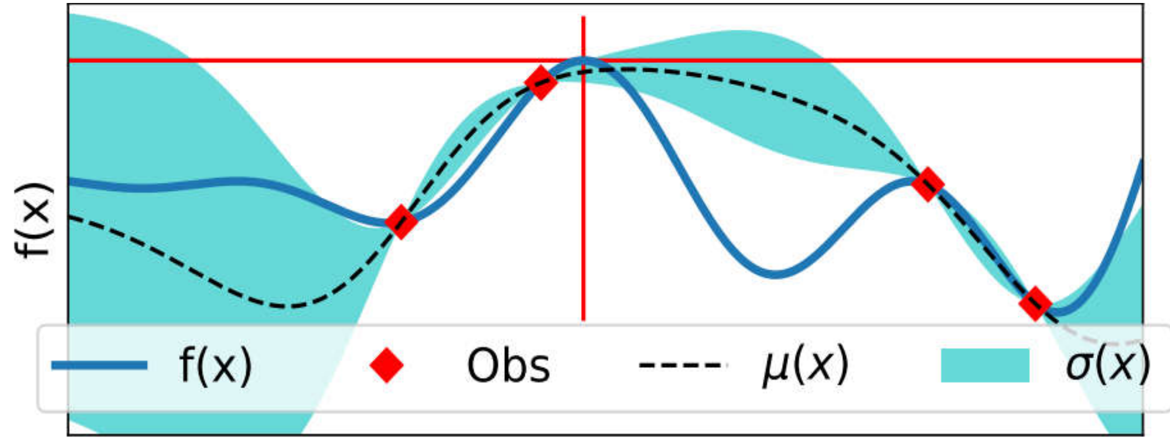
standard GP

$f(x)$ is above f^*



transformed GP

below f^*



Transformed Gaussian process

- $f(x) = f^* - \frac{1}{2}\underbrace{g^2(x)}_{\geq 0}$ $\qquad g(x) \sim GP(0, K)$

Zero mean prior !

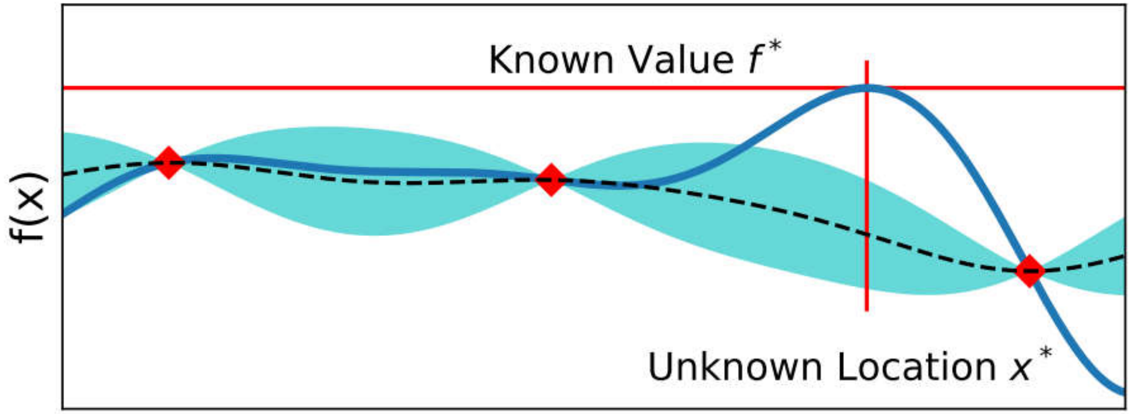
- This condition encourages that there is a point where $g(x) = 0$ and thus $f^* = f(x)$ 2

We want to control the surrogate using f^*

2 Lift up: the surrogate should reach f^*

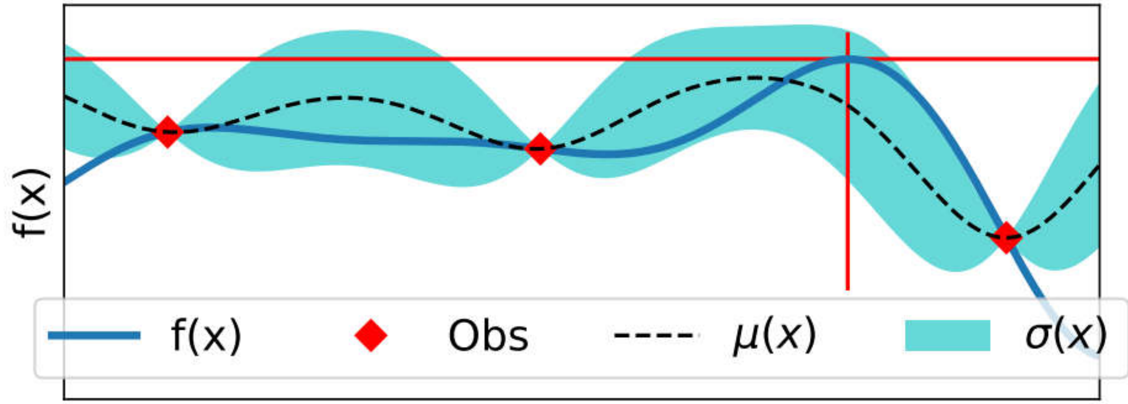
standard GP

$f(x)$ does not reach f^*



transformed GP

reach f^*



Transformed Gaussian process

- Linearization using Taylor expansion

$$f(x) \approx f^* - \frac{1}{2}\mu_g^2(x) - \mu_g(x)\left[g(x) - \mu_g(x)\right]$$
$$= f^* + \frac{1}{2}\mu_g^2(x) - \mu_g(x)g(x)$$

- Linear transformation of a GP remains Gaussian

$$\mu(x) = f^* - \frac{1}{2}\mu_g^2(x)$$
$$\sigma(x) = \mu_g(x)\sigma_g(x)\mu_g(x)$$

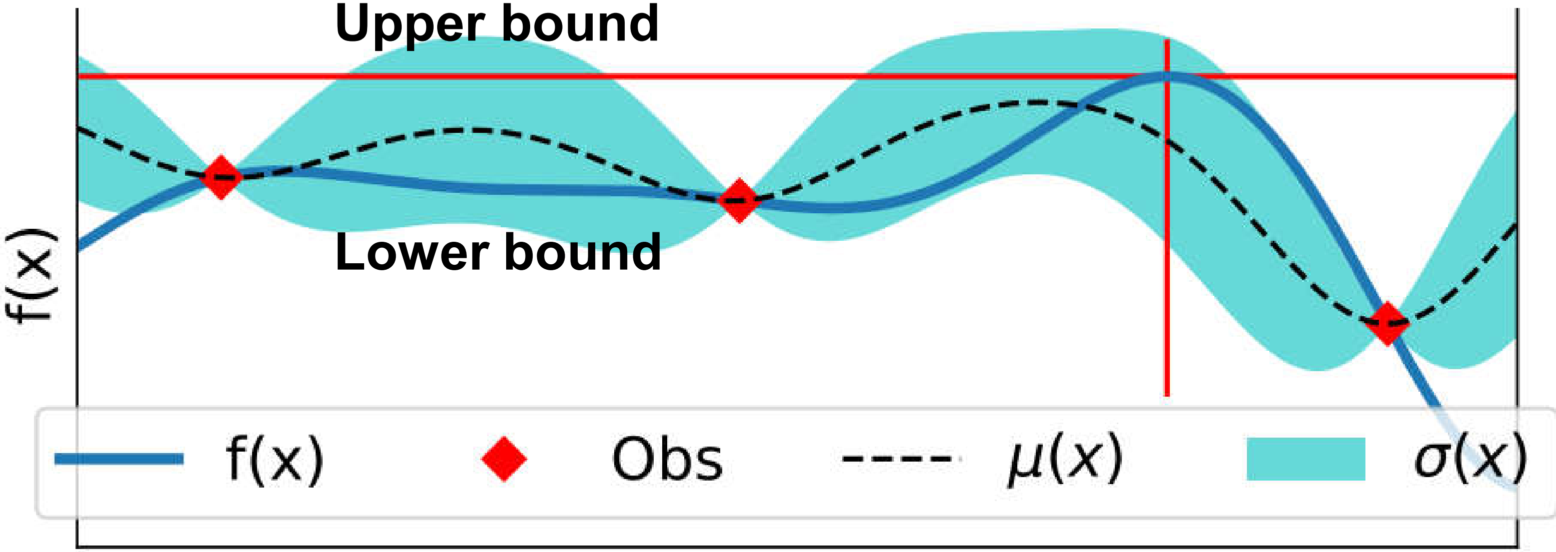
- The predictive distribution $p(x) \sim \mathcal{N}(\mu(x), \sigma(x))$
- Taylor expansion is very accurate at the mode which is $\mu_g^2(\boldsymbol{x})$

Outline

- Bayesian Optimization
- Bayes Opt with Known Optimum Value f^*
  - Problem definition
  - Exploiting f^*
    - Building better surrogate model
    - Making informed decision

Confidence Bound Minimization

- Under GP surrogate model, we have this condition w.h.p.



where β_t is defined following [Srinivas et al 2010]. This means

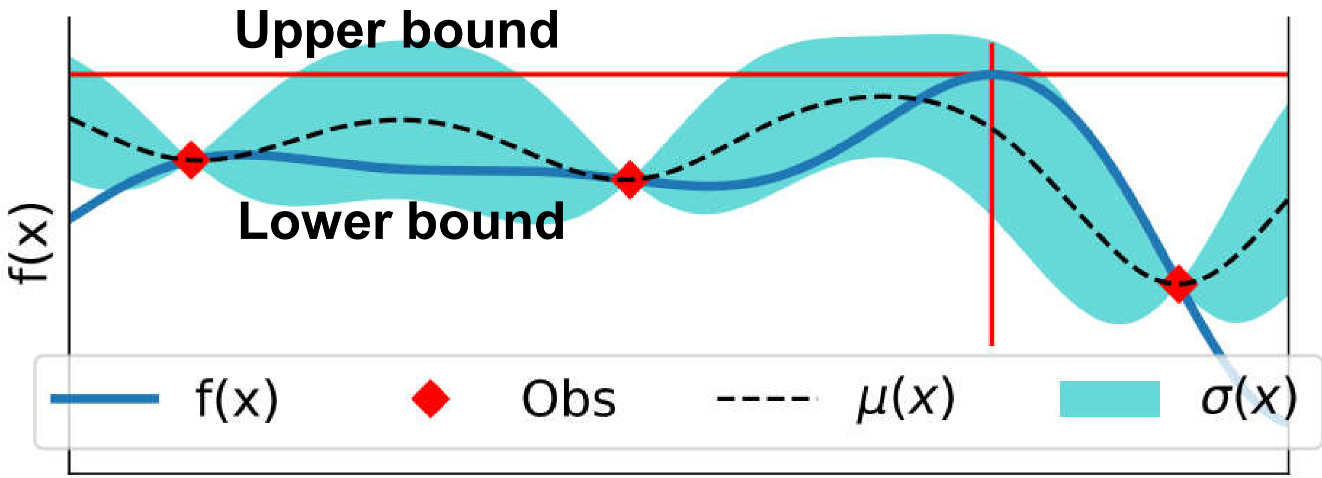
$$\underbrace{\mu(x^*) - \sqrt{\beta_t}\sigma(x^*)}_{\text{Lower bound}} \le \underset{\text{unknown}}{f(x^*)} = \underset{\text{known}}{f^*} \le \underbrace{\mu(x^*) + \sqrt{\beta_t}\sigma(x^*)}_{\text{Upper bound}}$$

can be estimated $\forall x$

Confidence Bound Minimization

- The best candidate for x^* is where the bound is tight

$$x_t = \arg\min |\mu(x) - f^*| + \sqrt{\beta_t}\sigma(x)$$



- The inequality becomes equality at the true x^* location where

$$\underbrace{\mu(x^*) - \sqrt{\beta_t}\sigma(x^*)}_{\text{Lower bound}} = \underset{\text{known}}{f^*} = \underbrace{\mu(x^*) + \sqrt{\beta_t}\sigma(x^*)}_{\text{Upper bound}}$$

when $\mu(x^*) = f^*$ and $\sigma(x^*) = 0$

Expected Regret Minimization

- Regret $r = f^* - f(x_t)$ where $f^* = \max f(x), \forall x$

- Finding the optimum location x^* = minimizing the regret.

- We can select the next point by minimizing the expected regret.

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathscr{X}} \alpha^{\text{ERM+}f^*}(\mathbf{x}) = \arg\min_{\mathbf{x} \in \mathscr{X}} \mathbb{E}\left[r(\mathbf{x})\right]$$

Expected Regret Minimization

- Using analytical derivation, we derive the closed-form computation for ERM.

$$\alpha^{ERM+f^*}(x) = \sigma(x) \times \phi(z) + [f^* - \mu(x)] \times \Phi(x)$$

$$z = \frac{[f^* - \mu(x)]}{\sigma(x)}$$

GP variance

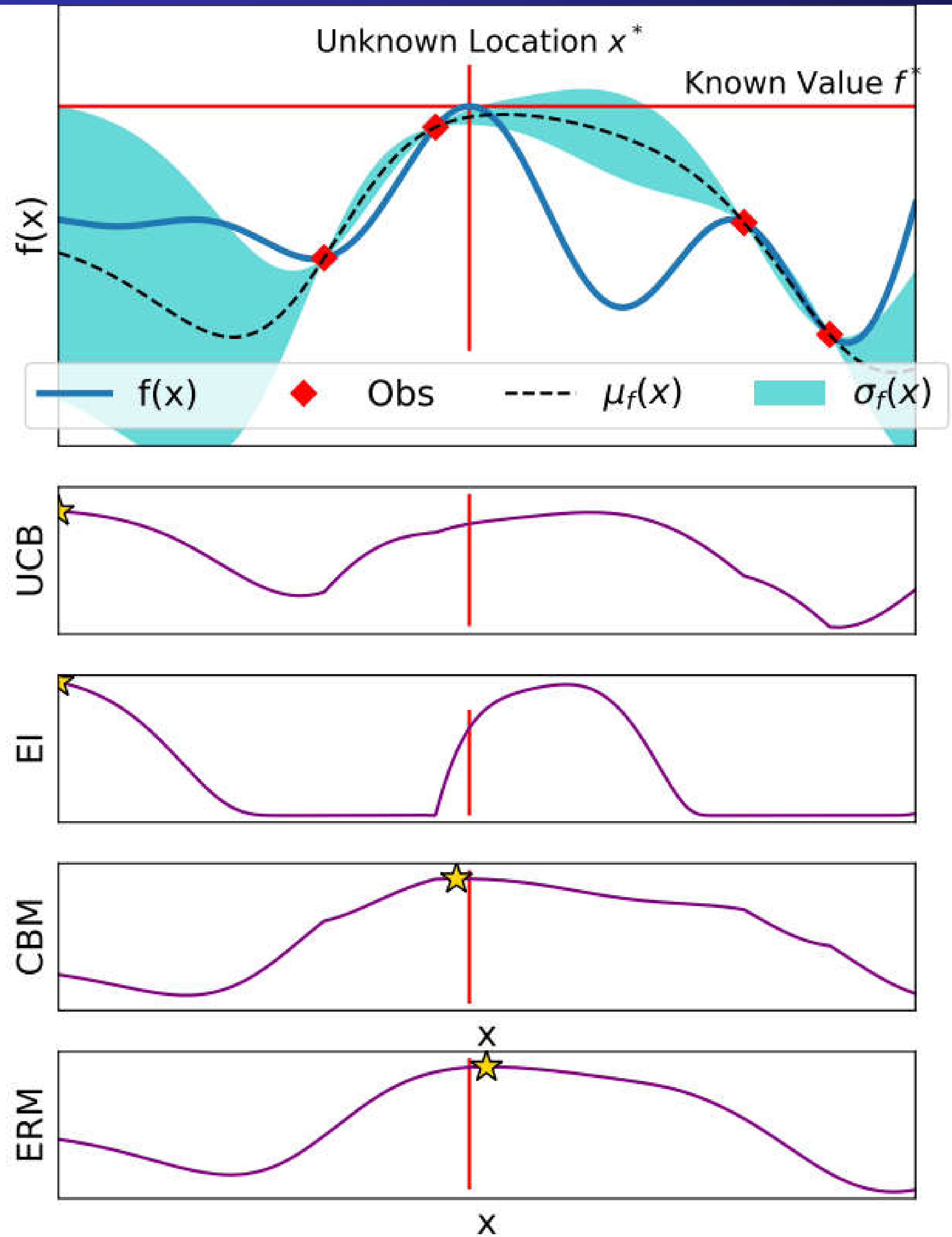
Gaussian PDF

GP mean

Gaussian CDF

- See the paper for details!

Illustration



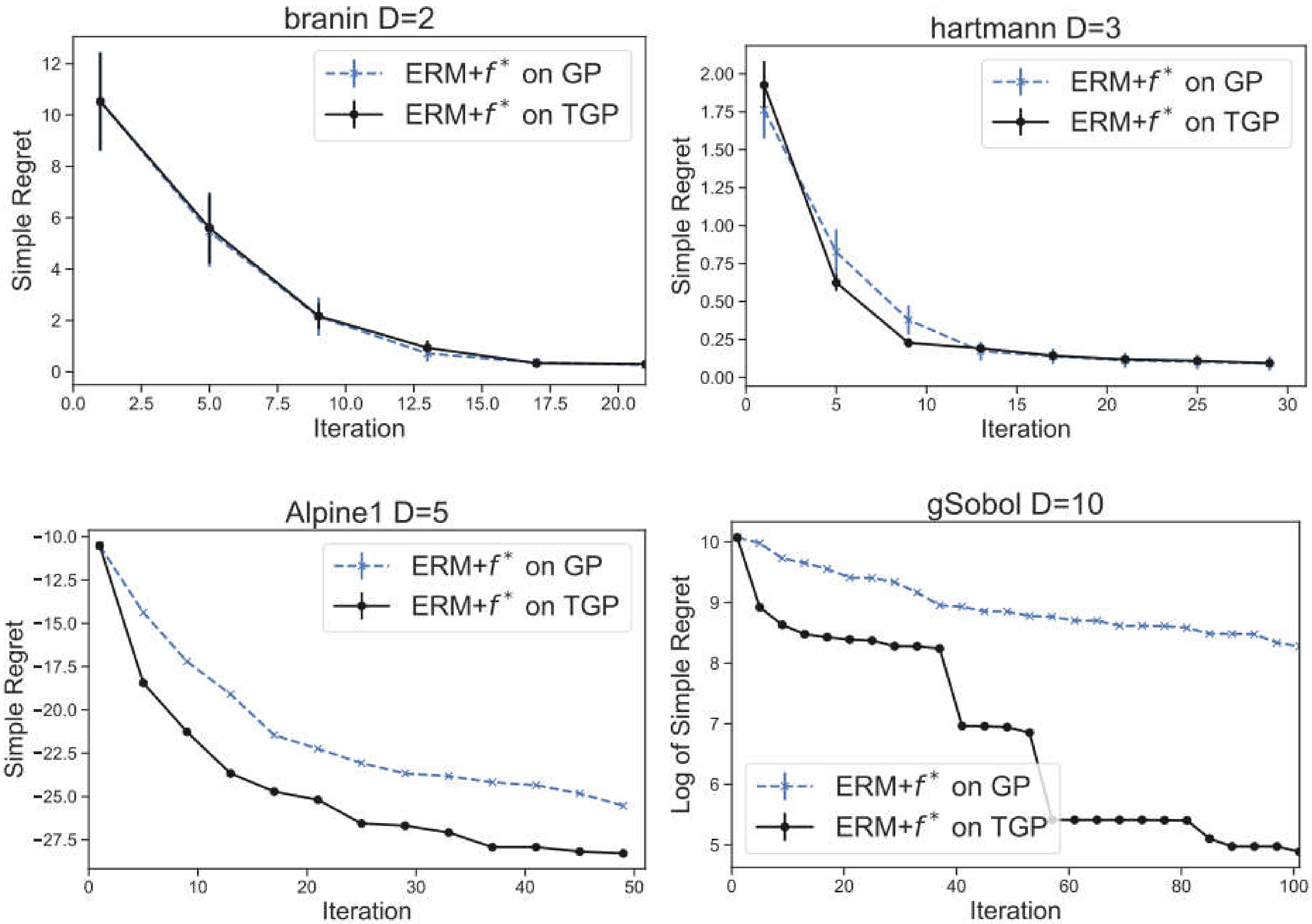
Existing Baselines

The Proposed

Tend to explore elsewhere

Correctly identify the true (unknown) location

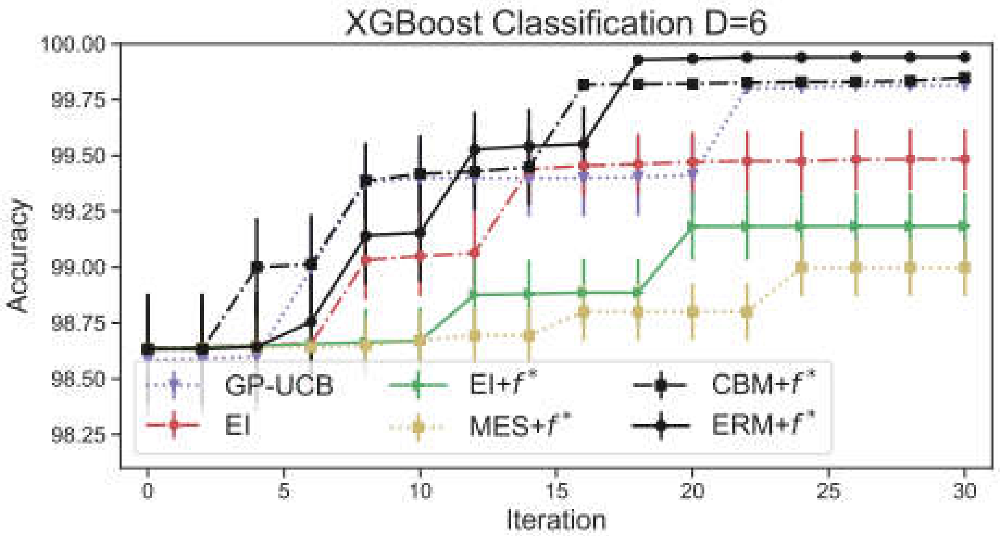
The GP transformation is helpful in high dimension



XGBoost Classification and DRL

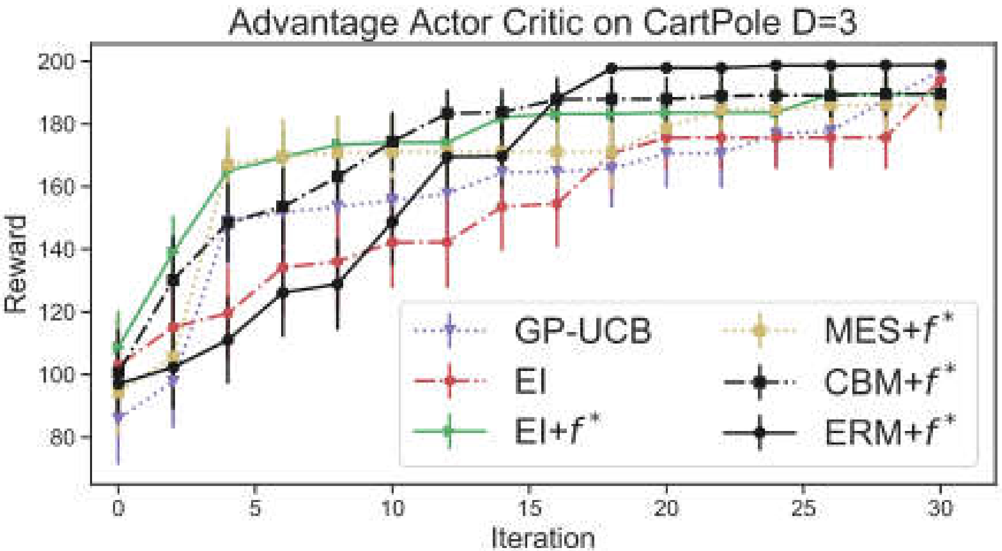
- Skin dataset UCI $f^* = 100$

Variables	Min	Max	Found $\mathbf{x}^*$
min child weight	1	20	4.66
colsample bytree	0.1	1	0.99
max depth	5	15	9.71
subsample	0.5	1	0.77
alpha	0	10	0.82
gamma	0	10	0.51



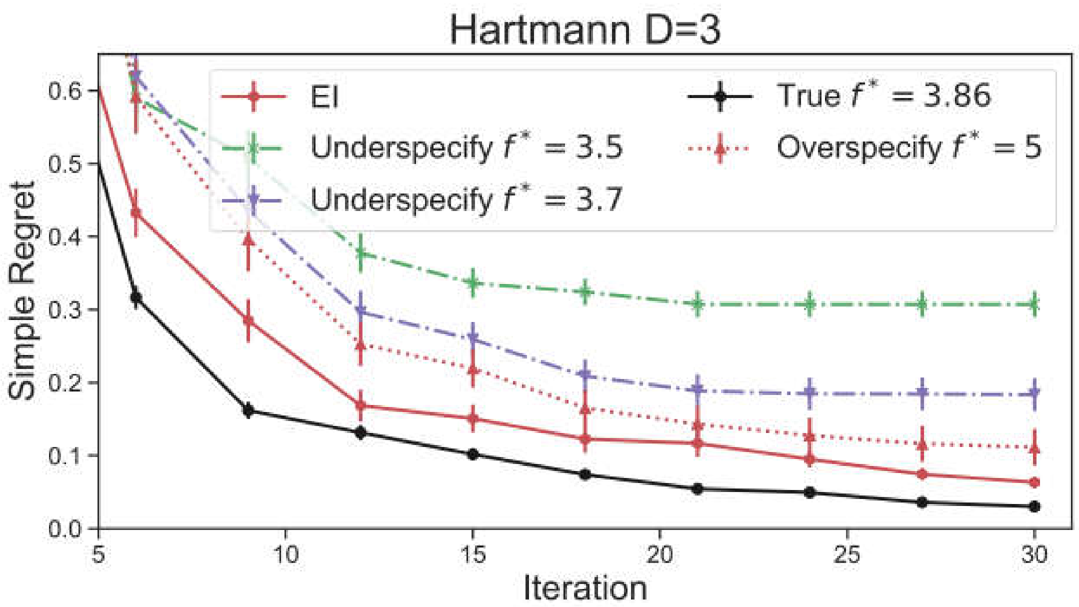
- CartPole DRL $f^* = 200$

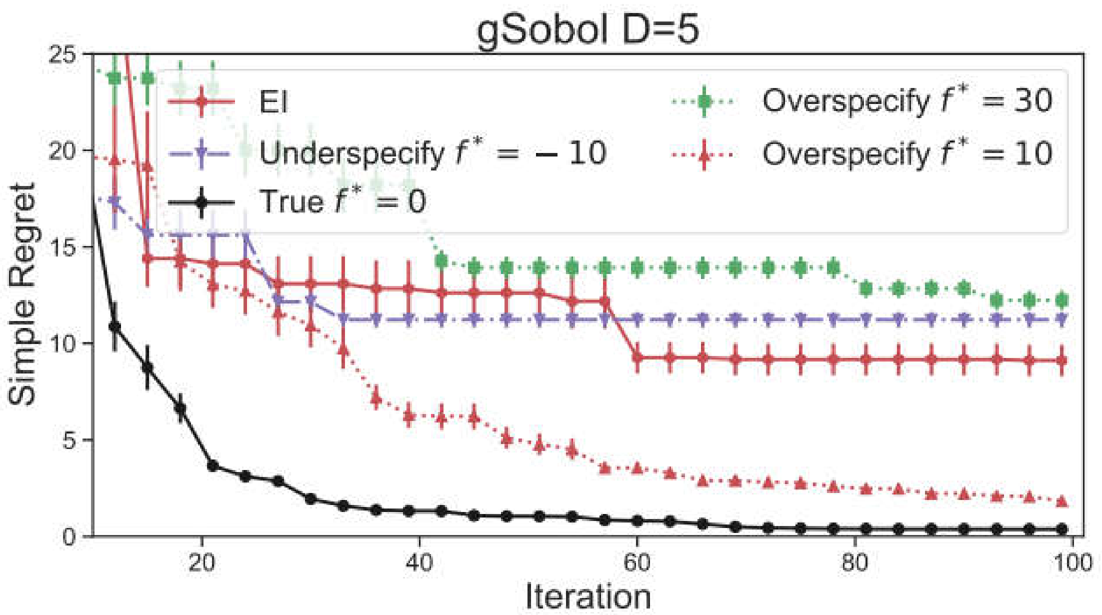
Variables	Min	Max	Best Parameter $\mathbf{x}^*$
γ discount factor	0.9	1	0.95586
learning rate q model	$1e^{-6}$	0.01	0.00589
learning rate v model	$1e^{-6}$	0.01	0.00037



Mis-specified f^* will degrade the performance

- Under-specified f^* smaller than the true f^*
  - More serious, as the algorithm will get stuck.
- Over-specified f^* greater than the true f^*
  - Less serious, but still poor performance.

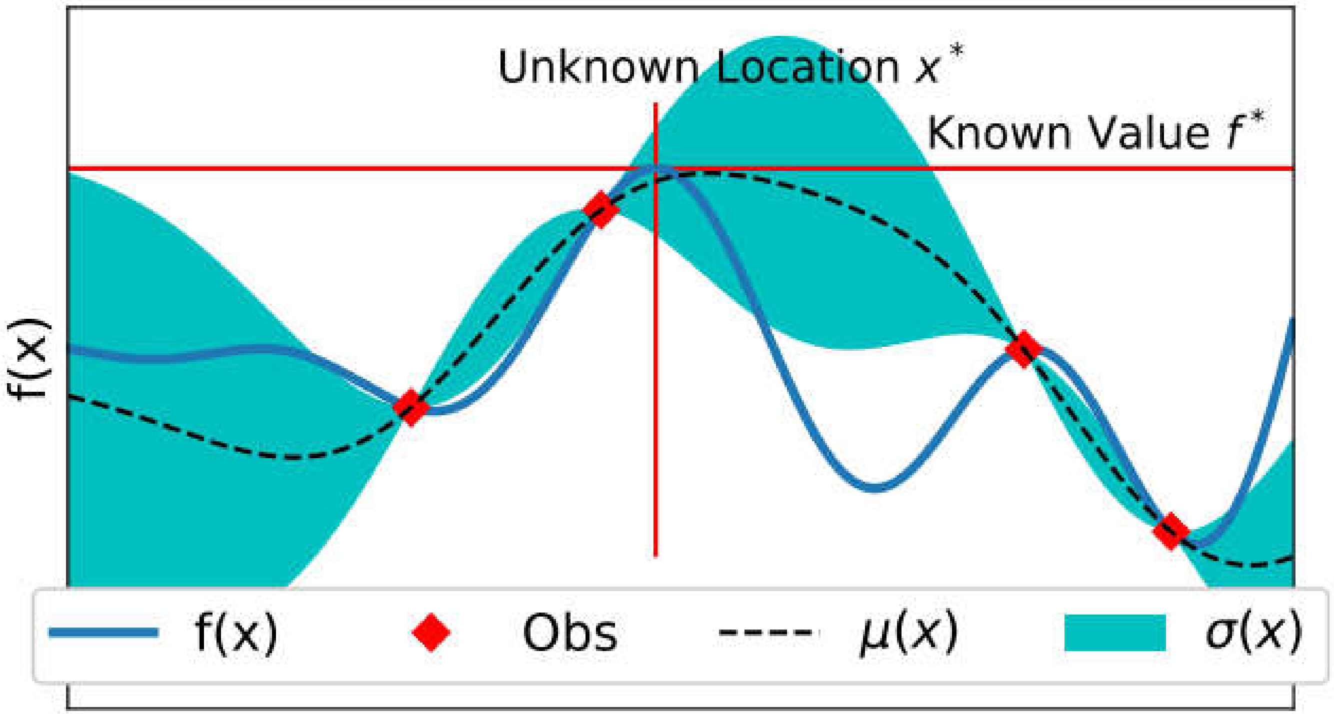




Take Home Messages

- Bayes opt is efficient for optimizing the black-box function

- When the optimum value is known, we can exploit this knowledge for better optimization.



Question and Answer

vu@robots.ox.ac.uk

@nguyentienvu

https://ntienvu.github.io