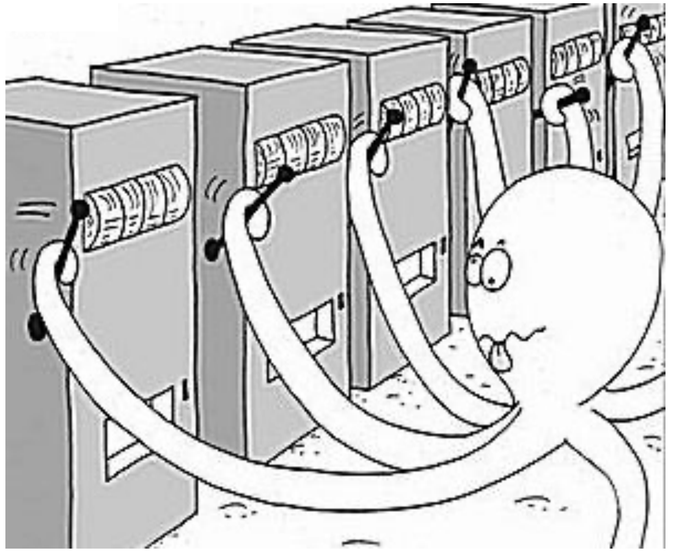
Bandits (the $H = 1$ case)

(And thinking about $H > 1$ / RL!)

Multi-armed bandits

How should we allocate
T tokens to A "arms"
to maximize our return?

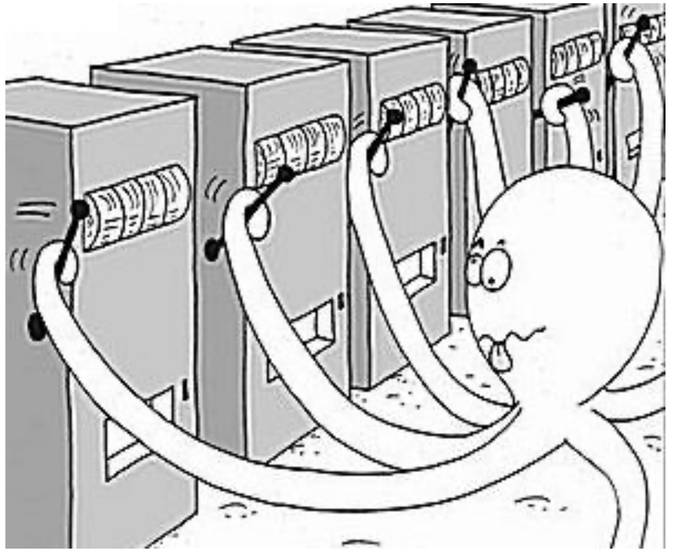
[Robins '52, Gittins'79, Lai & Robbins '85 …]



Multi-armed bandits

How should we allocate
T tokens to A "arms"
to maximize our return?

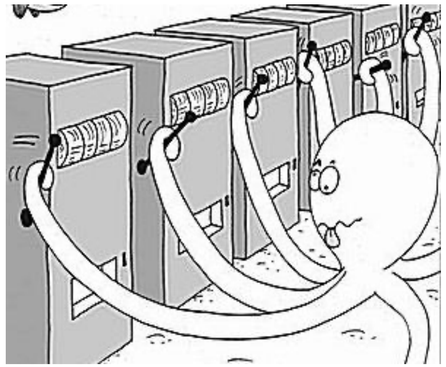
[Robins '52, Gittins'79, Lai & Robbins '85 …]



- Very successful algo when A is small.
- What can we do when the number of arms A is large?

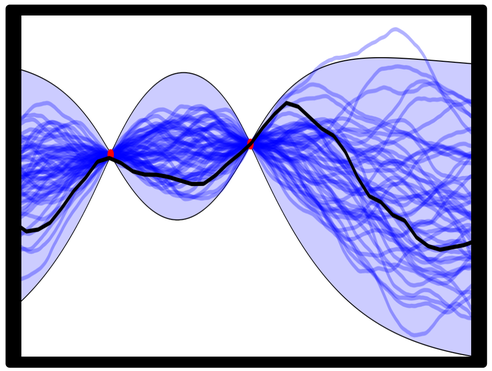
Dealing with the large action case

Bandits



- decision: pull an arm

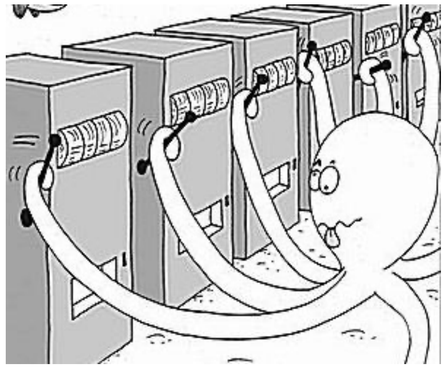
Linear (RKHS) Bandits



- decision: choose some $x \in \mathcal{X}$
- e.g. $x \in R$

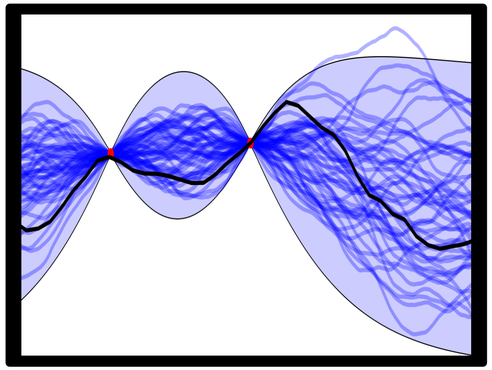
Dealing with the large action case

Bandits



- decision: pull an arm

Linear (RKHS) Bandits

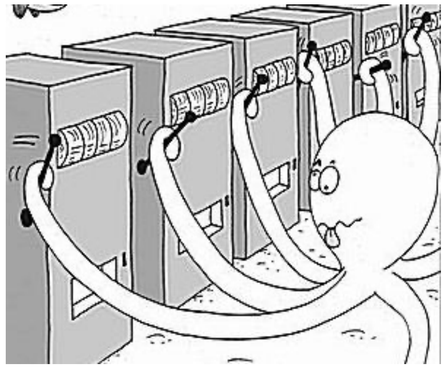


- decision: choose some $x \in \mathcal{X}$
- e.g. $x \in R$

- widely used generalization: The “linear bandit” model [Abe & Long+ ’99]
  successful in many applications: scheduling, ads…

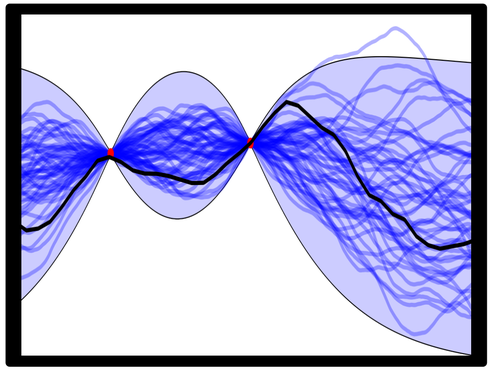
Dealing with the large action case

Bandits



- decision: pull an arm

Linear (RKHS) Bandits



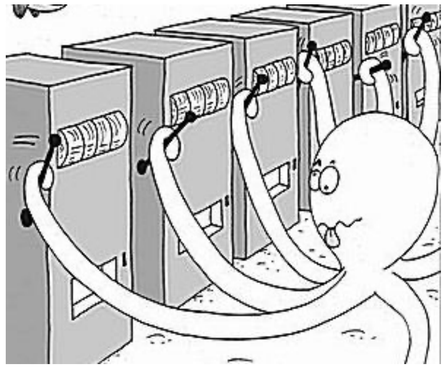
- decision: choose some $x \in \mathcal{X}$
- e.g. $x \in R$

- widely used generalization: The “linear bandit” model [Abe & Long+ ’99]
  successful in many applications: scheduling, ads…
- decision: x_t, reward: r_t, reward model:

$$r_t = f(x_t) + \text{noise}, \qquad f(x) = w^\star \cdot \phi(x)$$

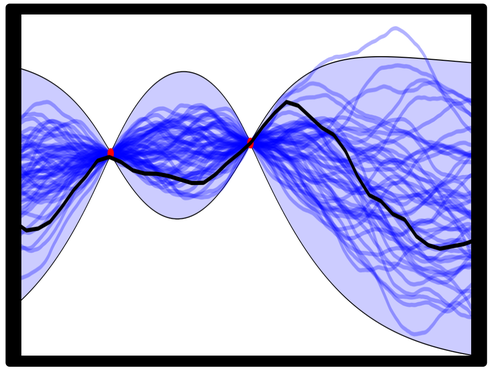
Dealing with the large action case

Bandits



- decision: pull an arm

Linear (RKHS) Bandits



- decision: choose some $x \in \mathcal{X}$
- e.g. $x \in R$

- widely used generalization: The "linear bandit" model [Abe & Long+ '99]
  successful in many applications: scheduling, ads…
- decision: x_t, reward: r_t, reward model:

$$r_t = f(x_t) + \text{noise}, \qquad f(x) = w^\star \cdot \phi(x)$$

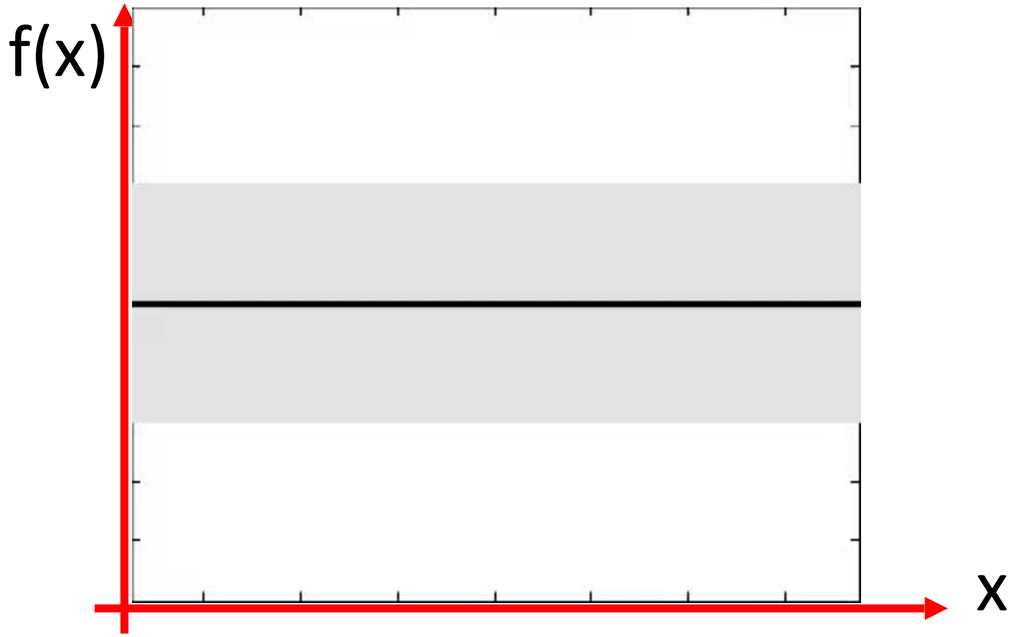
- Hypothesis class $\mathcal{F}$ is set of linear/RKHS functions

Linear-UCB/GP-UCB:

Algorithmic Principle: Optimism in the face of uncertainty

Pick input that maximizes upper confidence bound:

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$

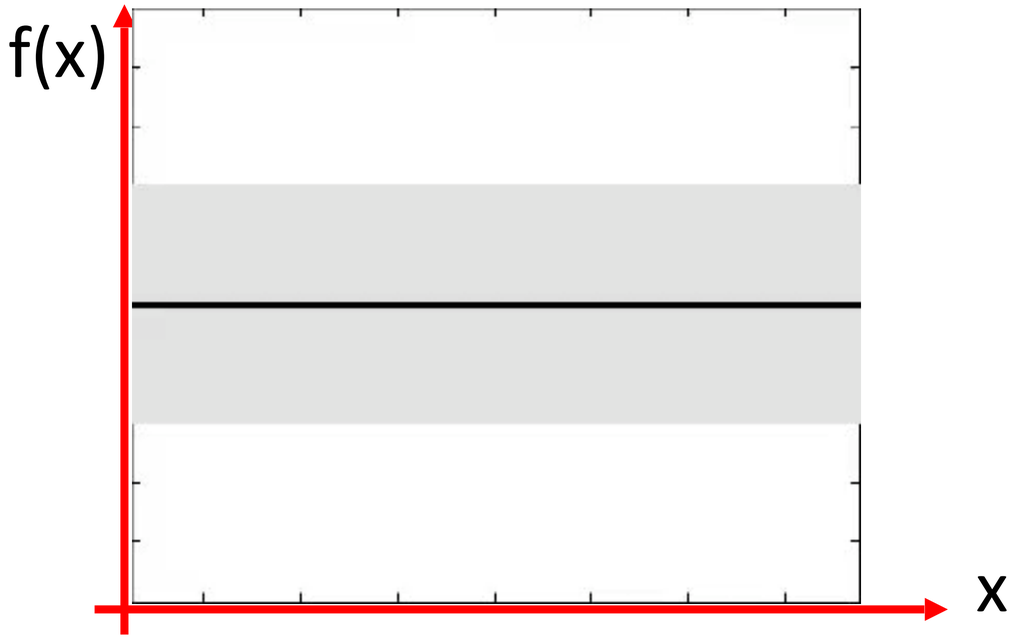


Linear-UCB/GP-UCB:

Algorithmic Principle: Optimism in the face of uncertainty

Pick input that maximizes upper confidence bound:

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$



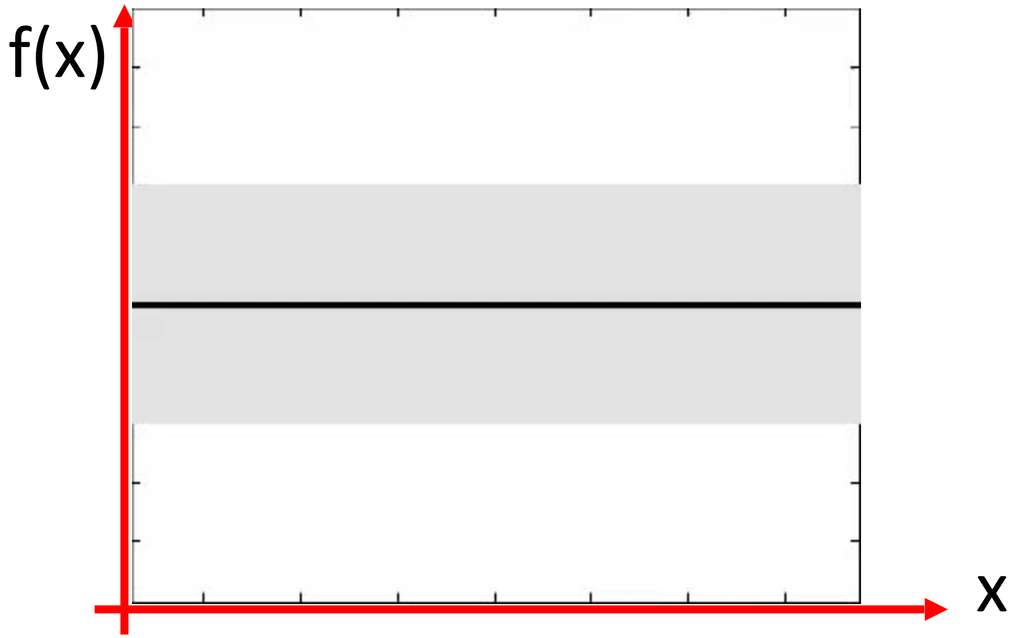
Naturally trades off exploration and exploitation

Linear-UCB/GP-UCB:

Algorithmic Principle: Optimism in the face of uncertainty

Pick input that maximizes upper confidence bound:

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$



Naturally trades off exploration and exploitation

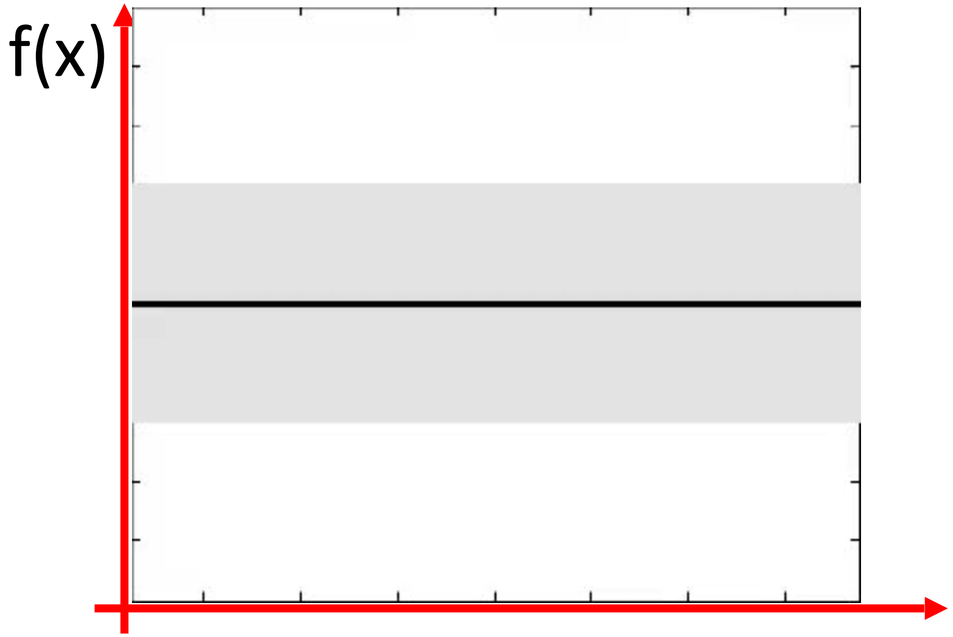
Only picks plausible maximizers

Linear-UCB/GP-UCB:

Algorithmic Principle: Optimism in the face of uncertainty

Pick input that maximizes upper confidence bound:

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$



How should we choose β_t?

Naturally trades off exploration and exploitation

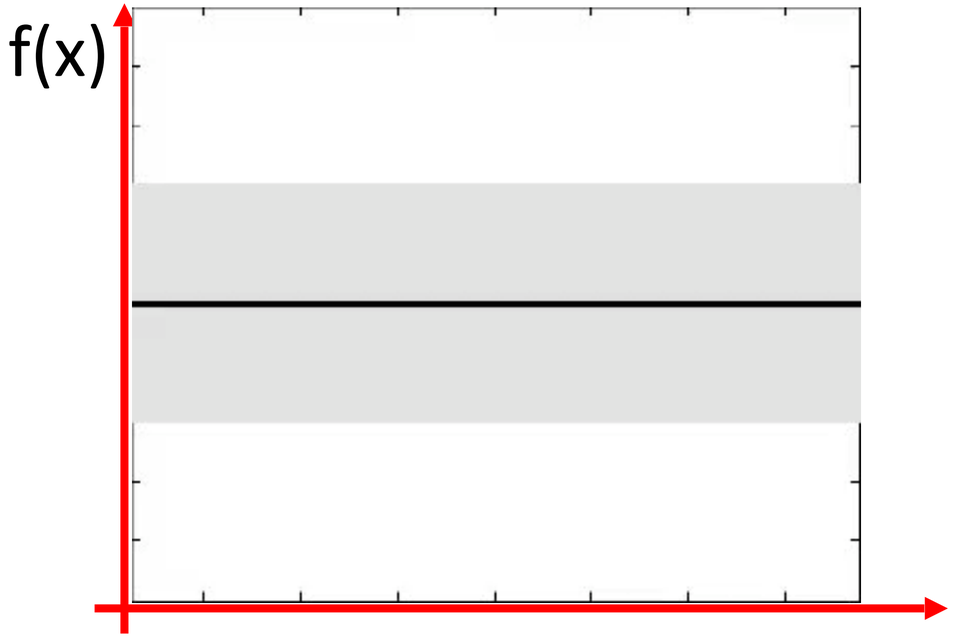
Only picks plausible maximizers

Linear-UCB/GP-UCB:

Algorithmic Principle: Optimism in the face of uncertainty

Pick input that maximizes upper confidence bound:

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$$

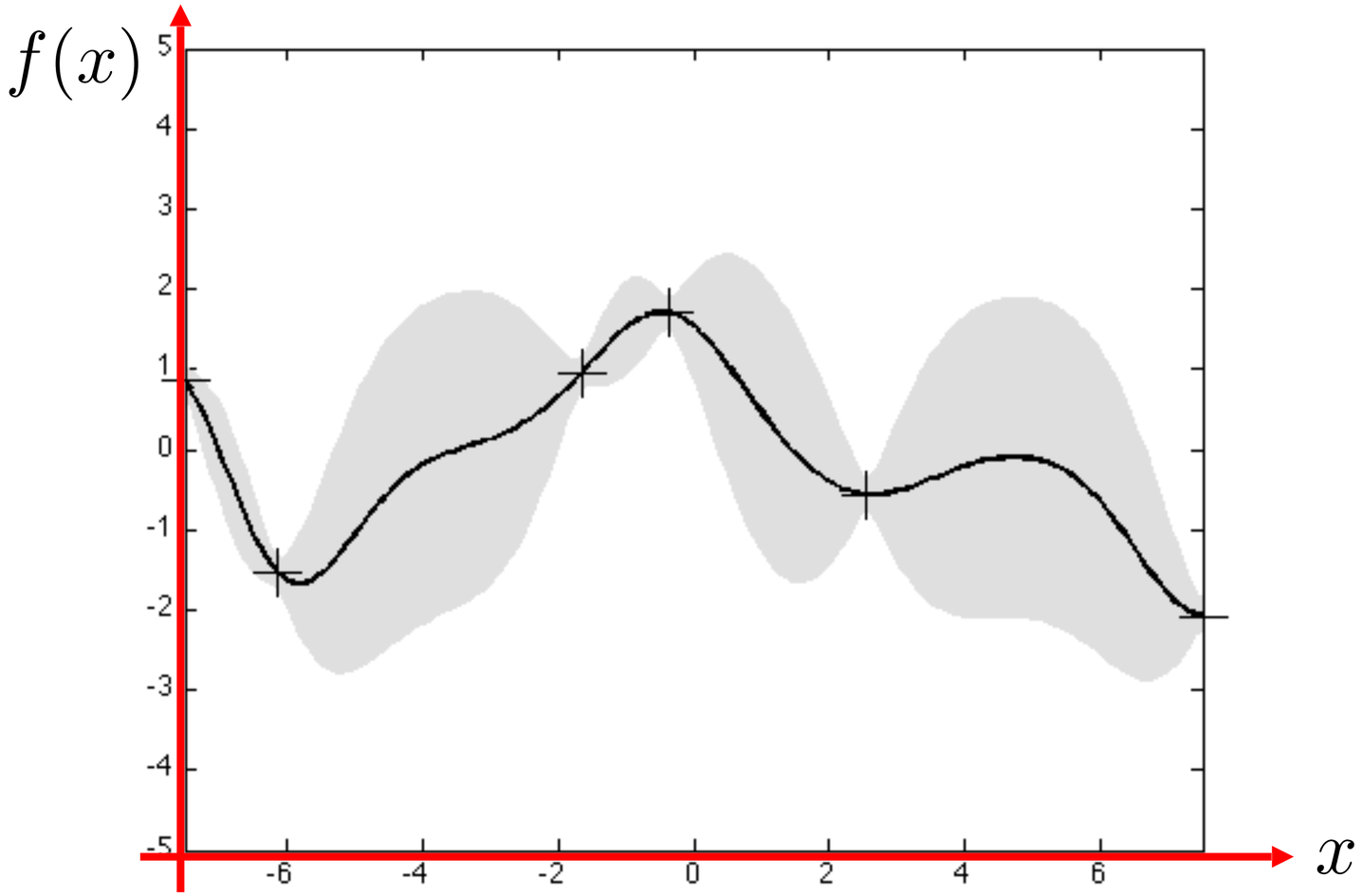


How should we choose β_t?

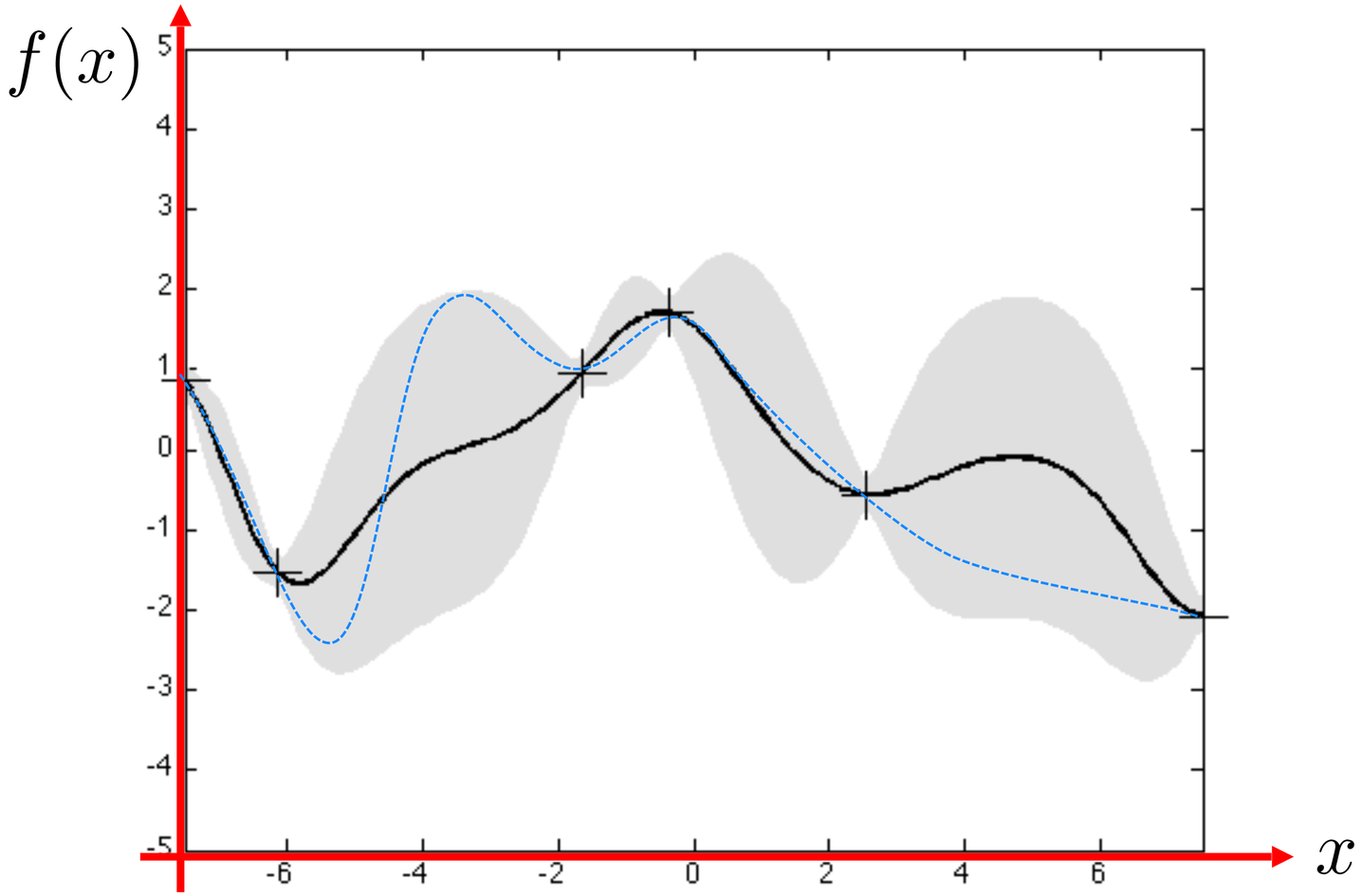
Naturally trades off exploration and exploitation

Only picks plausible maximizers

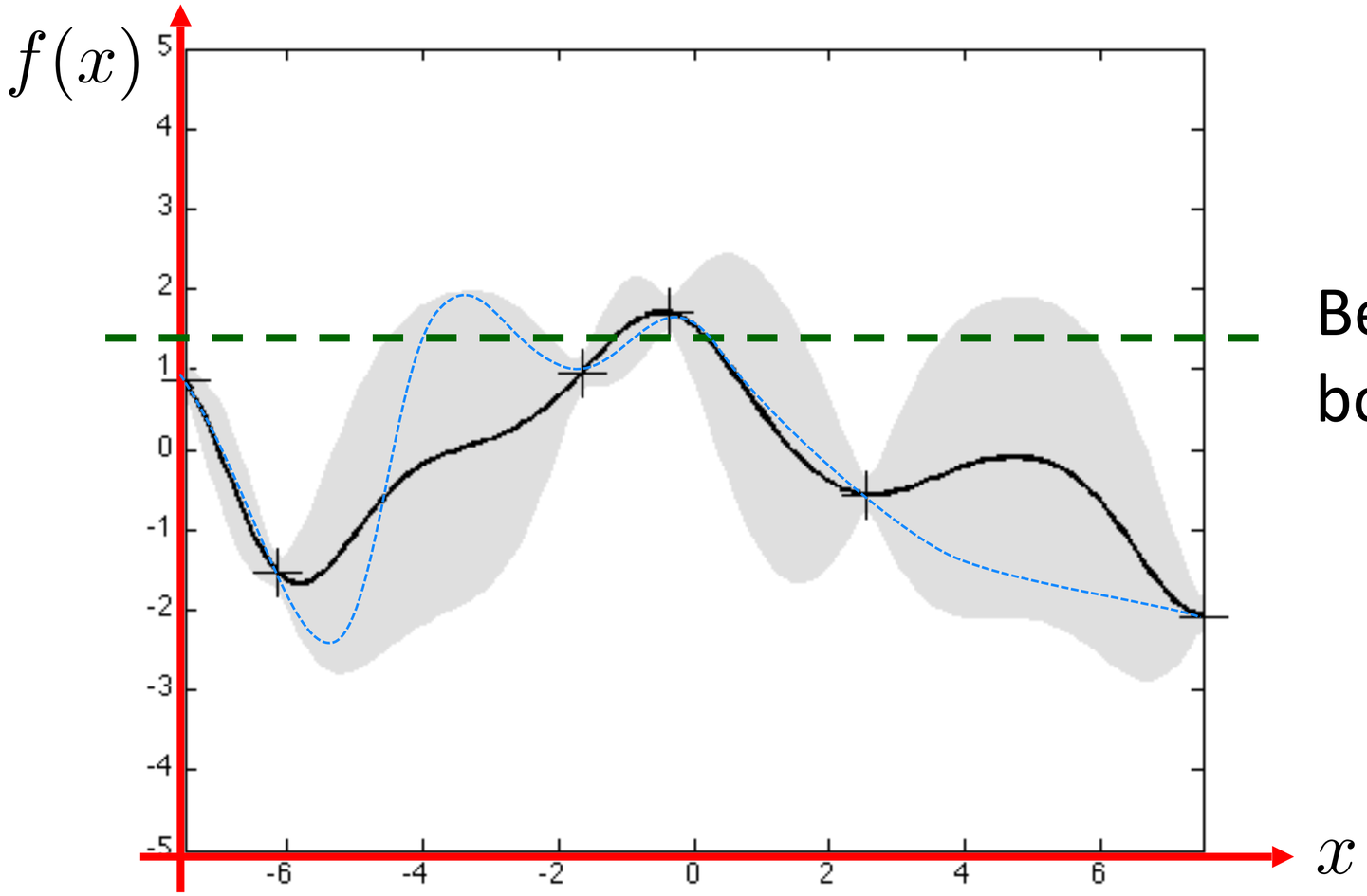
Optimistic Bayesian Optimization with GPs



Optimistic Bayesian Optimization with GPs

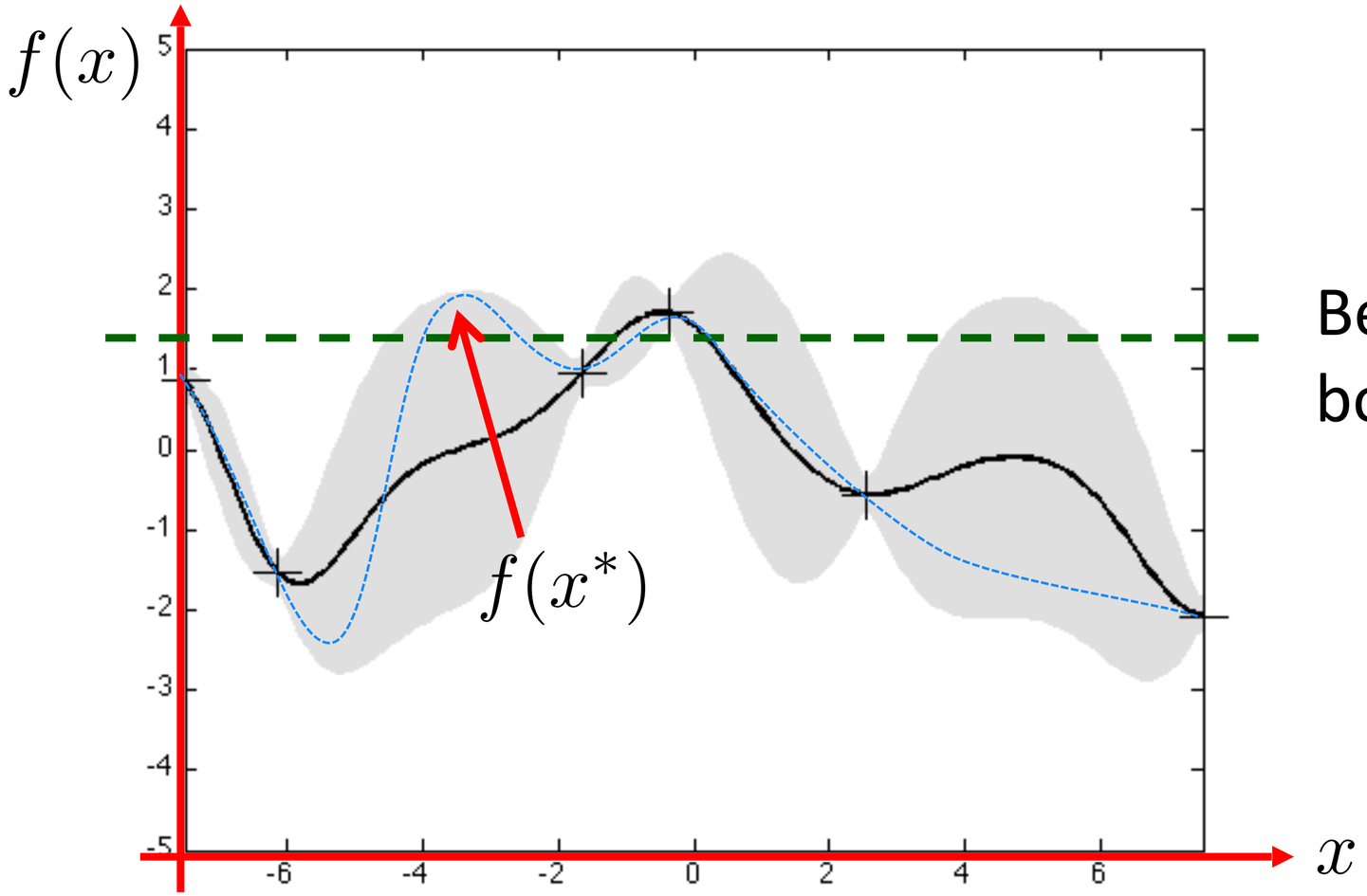


Optimistic Bayesian Optimization with GPs



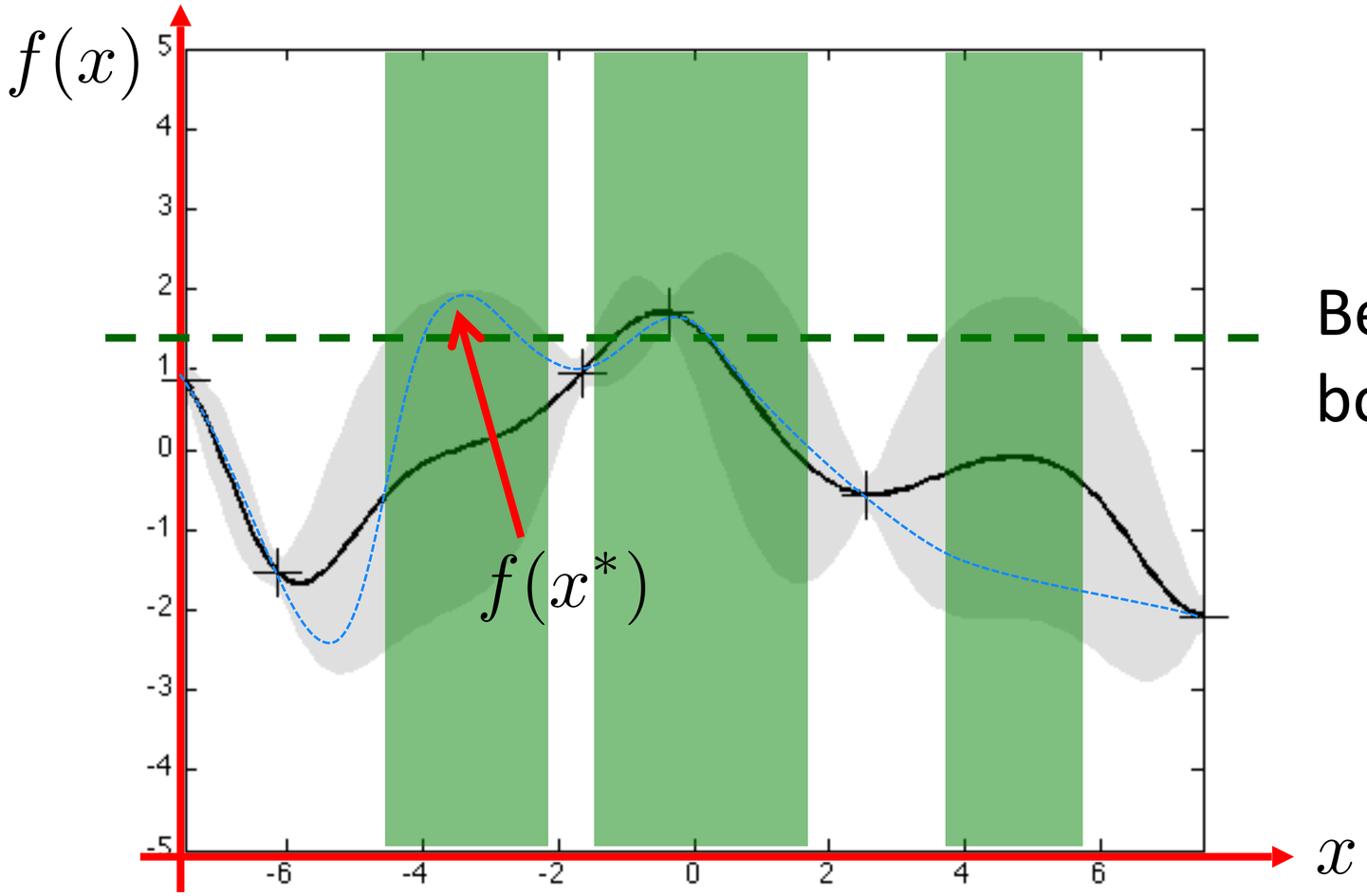
Best lower bound

Optimistic Bayesian Optimization with GPs



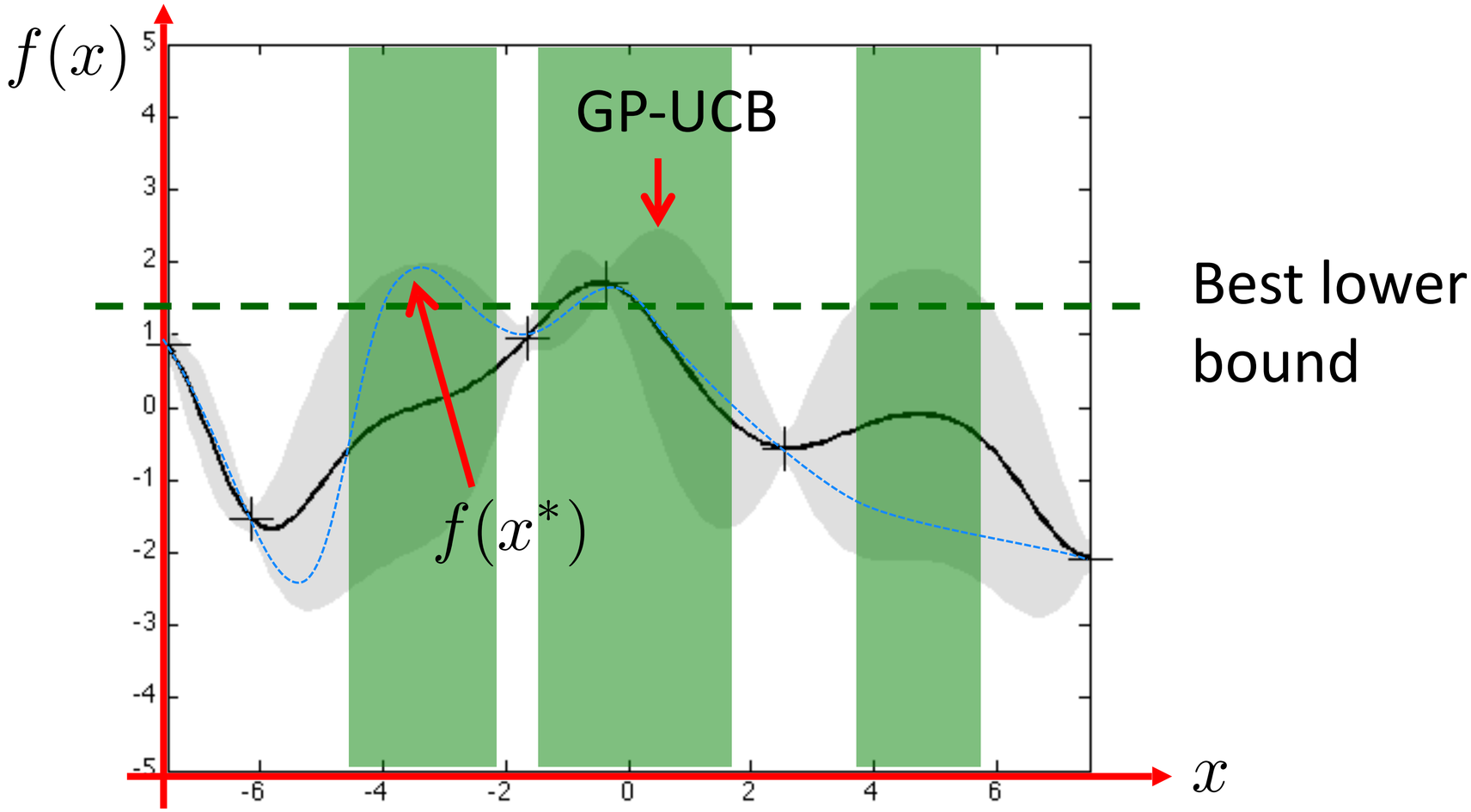
Best lower bound

Optimistic Bayesian Optimization with GPs



Best lower bound

Optimistic Bayesian Optimization with GPs



Key idea: Focus exploration on plausible maximizers
(upper confidence bound ≥ best lower bound)

Regret of Lin-UCB/GP-UCB

(generalization in action space)

Theorem: [Dani, Hayes, & K. ’08], [Srinivas, Krause, K. & Seeger '10]

Assuming $\mathscr{F}$ is an RKHS (with bounded norm), if we choose β_t “correctly”,

$$\frac{1}{T}\sum_{t=1}^{T}[f(x^*) - f(x_t)] = \mathcal{O}^*\left(\sqrt{\frac{\gamma_T}{T}}\right)$$

where $\gamma_T := \max_{x_0 \ldots x_{T-1} \in \mathcal{X}} \log\det\left(I + \sum_{t=0}^{T-1} \phi(x_t)\phi(x_t)^\top\right)$

Regret of Lin-UCB/GP-UCB

(generalization in action space)

Theorem: [Dani, Hayes, & K. ’08], [Srinivas, Krause, K. & Seeger '10]

Assuming $\mathscr{F}$ is an RKHS (with bounded norm), if we choose β_t “correctly”,

$$\frac{1}{T}\sum_{t=1}^{T}[f(x^*) - f(x_t)] = \mathcal{O}^*\left(\sqrt{\frac{\gamma_T}{T}}\right)$$

where $\gamma_T := \max_{x_0 \ldots x_{T-1} \in \mathcal{X}} \log\det\left(I + \sum_{t=0}^{T-1} \phi(x_t)\phi(x_t)^\top\right)$

- Key complexity concept: *“maximum information gain”* γ_T determines the regret
  - $\gamma_T \approx d \log T$ for ϕ in d-dimensions
  - Think of γ_T as the “effective dimension"
- Easy to incorporate context