Bandit Optimisation with Approximations

Kirthevasan Kandasamy

Carnegie Mellon University

École Polytechnique, Paris

April 27, 2017

Slides: www.cs.cmu.edu/∼kkandasa/misc/ecole-slides.pdf

Slides are up on my website: www.cs.cmu.edu/∼kkandasa

www.cs.cmu.edu/~kkandasa/

home publications software misc.

kirthevasan kandasamy

PhD Student, Carnegie Mellon University

[CV] [Google Scholar] [GitHub] [Contact]

I am a fourth year Machine Learning PhD student in the School of Computer Science at Carnegie Mellon University. I am co-advised by Jeff Schneider and Barnabas Poczos. I also work with Larry Wasserman. I am a member of the Auton Lab and the StatML Group. Prior to CMU, I completed my B.Sc in Electronics & Telecommunications Engineering at the University of Moratuwa, Sri Lanka.

My research interests lie in the intersection of statistical and algorithmic Machine Learning. My current research spans bandit problems, Bayesian optimisation, Gaussian processes, nonparametric statistics and graphical models. As of late, I have also hopped on the deep learning bandwagon.

I am generously supported by a Facebook PhD fellowship (2017) and a CMU Presidential fellowship (2015).

Preprints

Slides

Multi-fidelity Bayesian Optimisation with Continuous Approximations
Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, Barnabas Poczos
[arxiv] [EP slides]

Multi-fidelity Gaussian Process Bandit Optimisation
Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, Barnabas Poczos
[arxiv] [code] [UCL slides]

Influence Functions for Machine Learning: Nonparametric Estimators for Entropies, Divergences and Mutual Informations
Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, James Robins

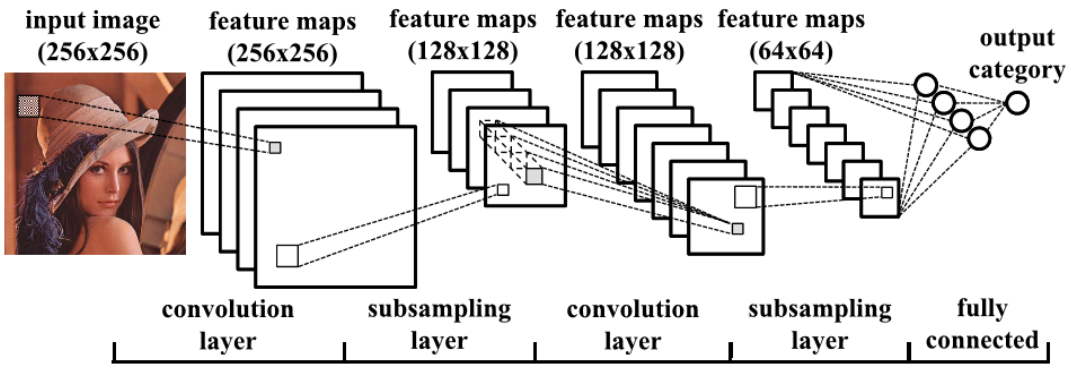
Bandit Optimisation

x → **Neural Network** → $f(x)$

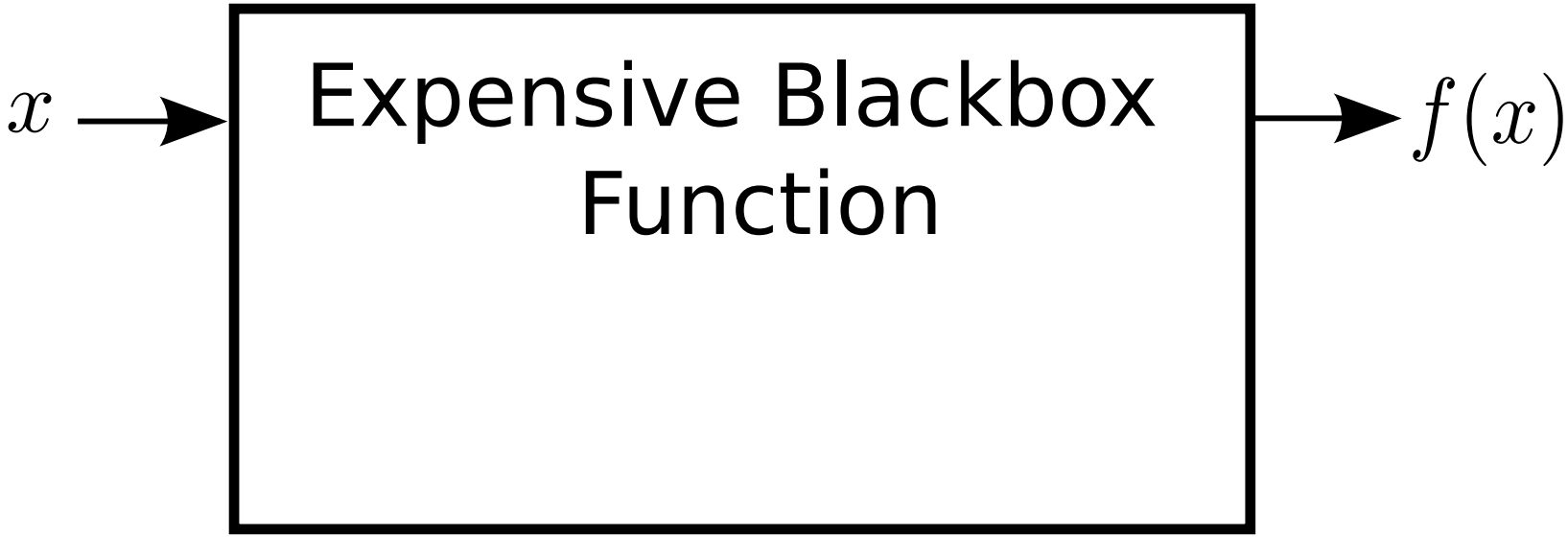
hyper-parameters

- Train NN using given hyper-parameters
- Compute accuracy on validation set

cross validation accuracy



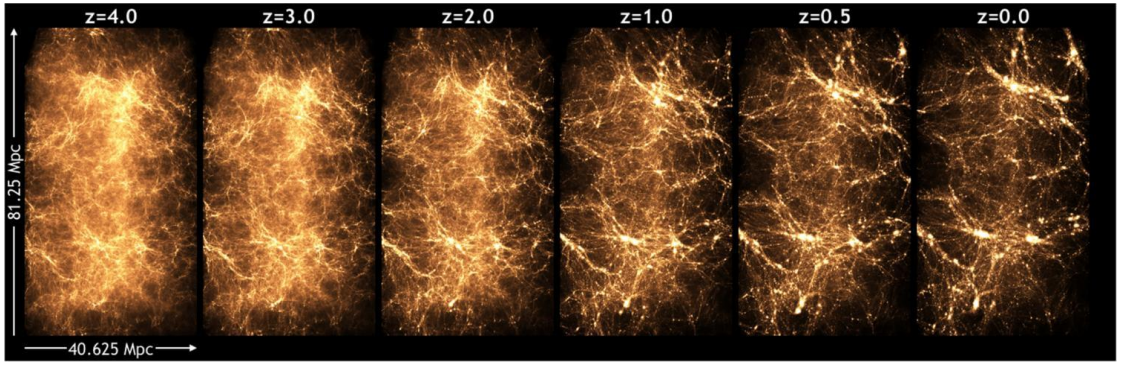
Bandit Optimisation



Bandit Optimisation

$x \longrightarrow$ **Expensive Blackbox Function** $\longrightarrow f(x)$

Other Examples:
- ML estimation in Astrophysics
- Optimal policy in Autonomous Driving
- Synthetic gene design



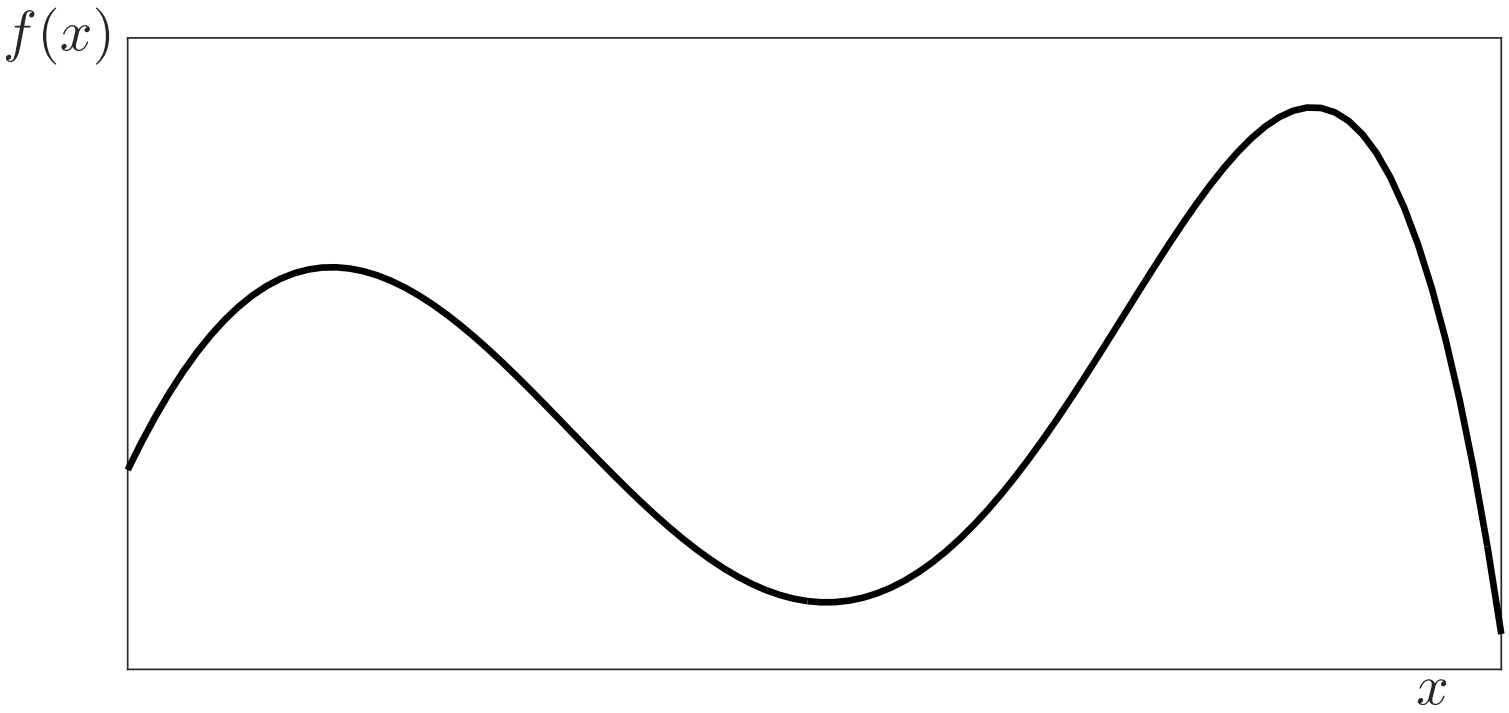
Bandit Optimisation

$x \longrightarrow$ **Expensive Blackbox Function** $\longrightarrow f(x)$

Other Examples:
- ML estimation in Astrophysics
- Optimal policy in Autonomous Driving
- Synthetic gene design

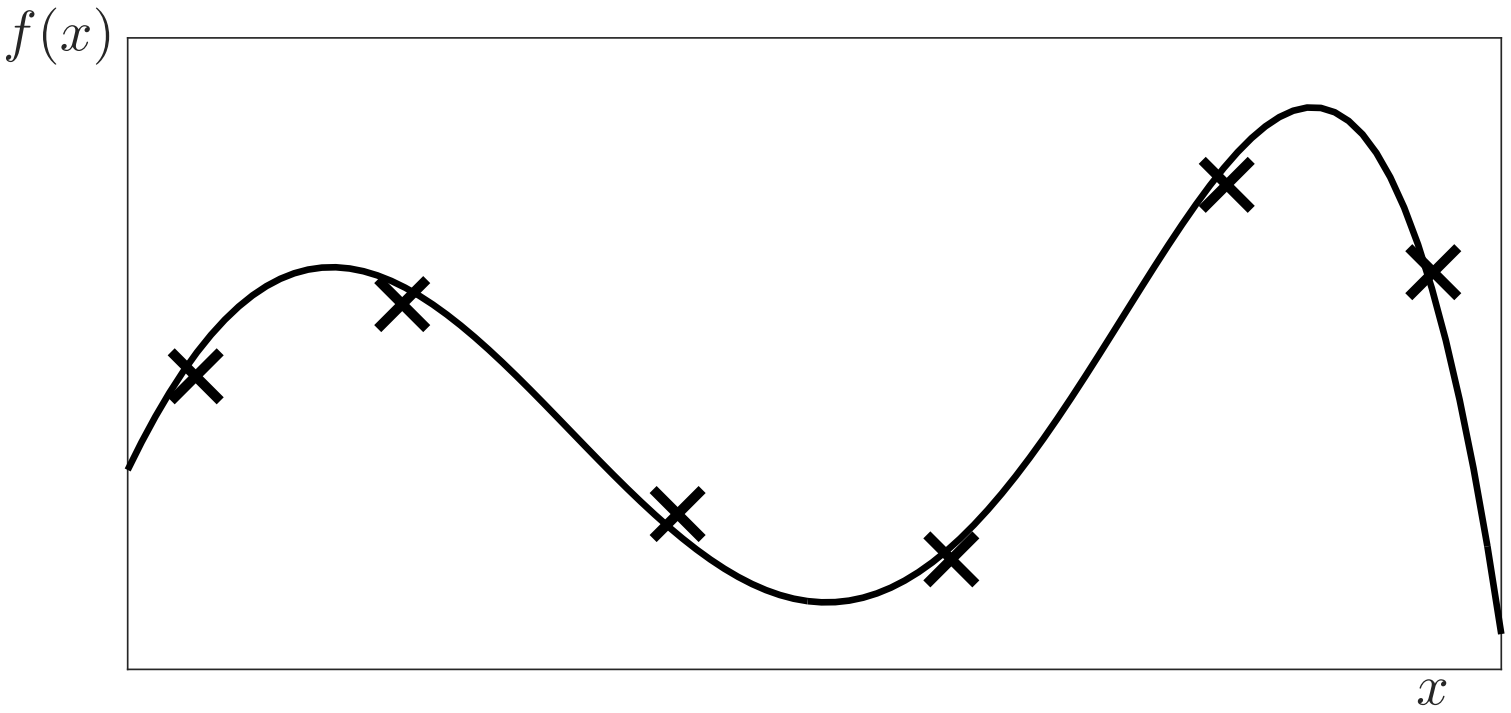
Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.



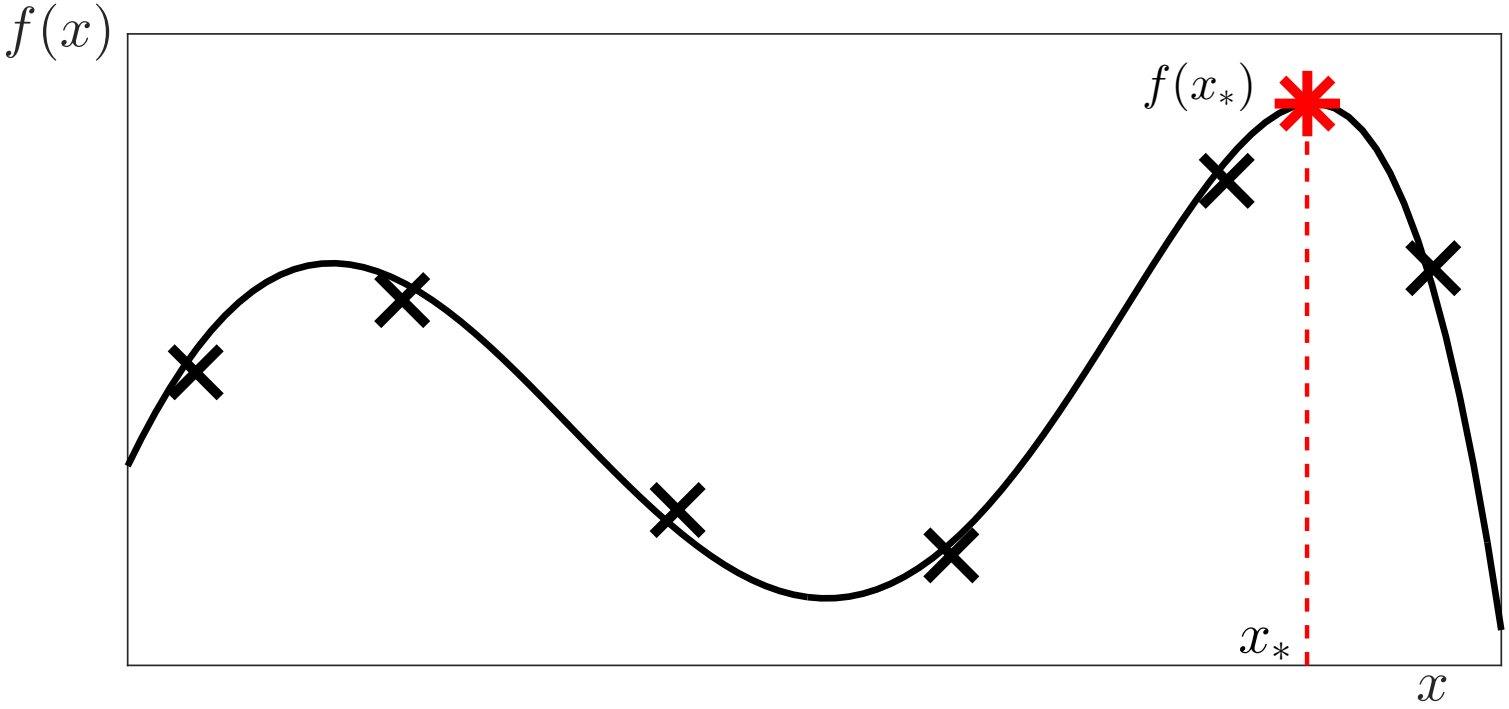
Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.



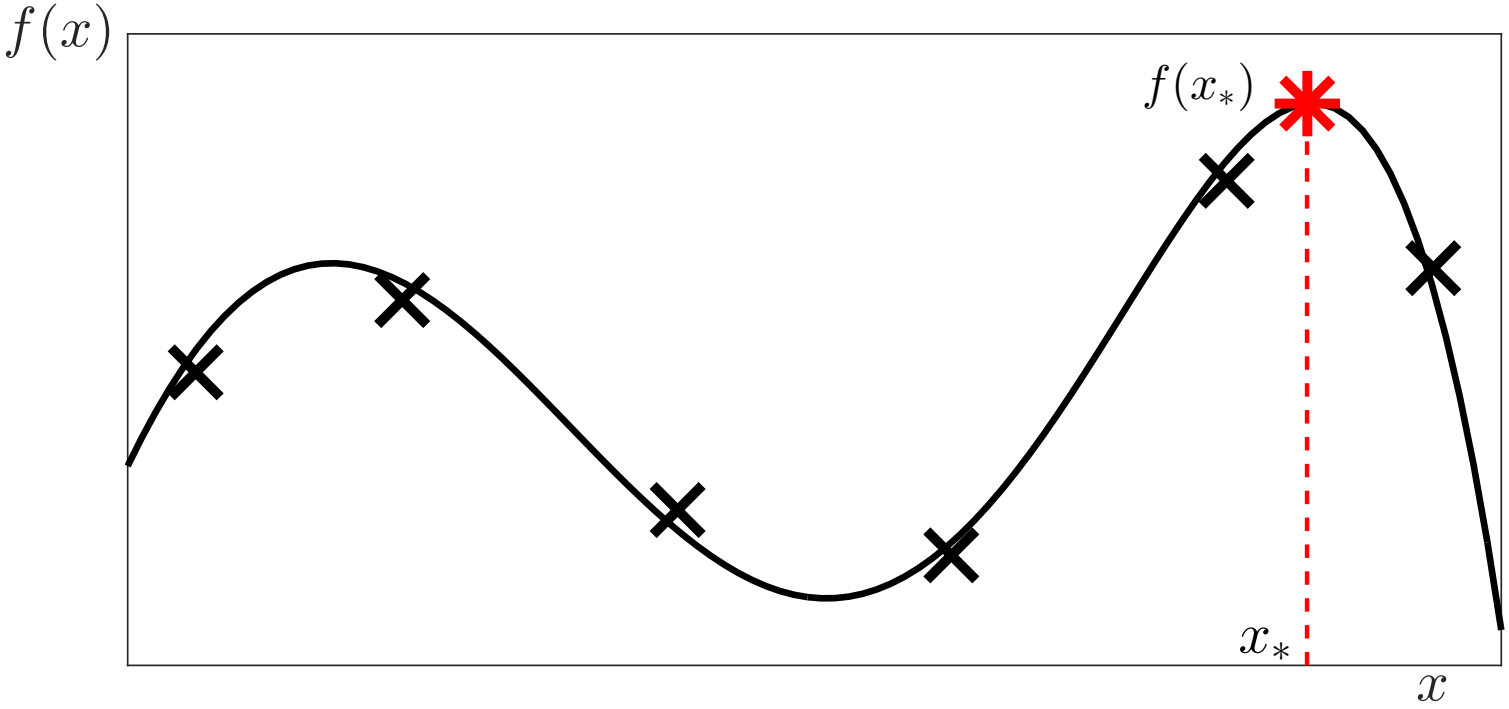
Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x f(x)$.



Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.

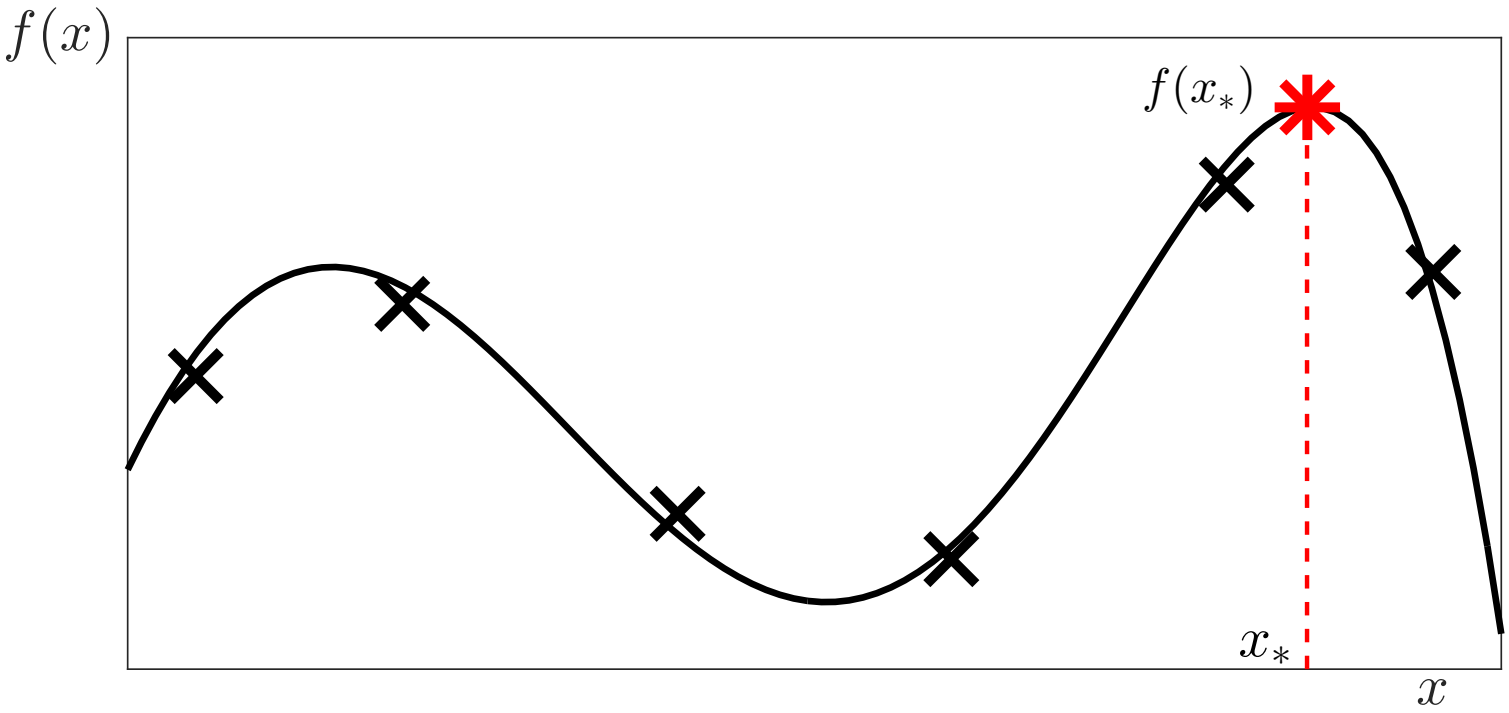


Simple Regret after n evaluations

$$S_n = f(x_\star) \; - \max_{t=1,\ldots,n} f(x_t).$$

Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.

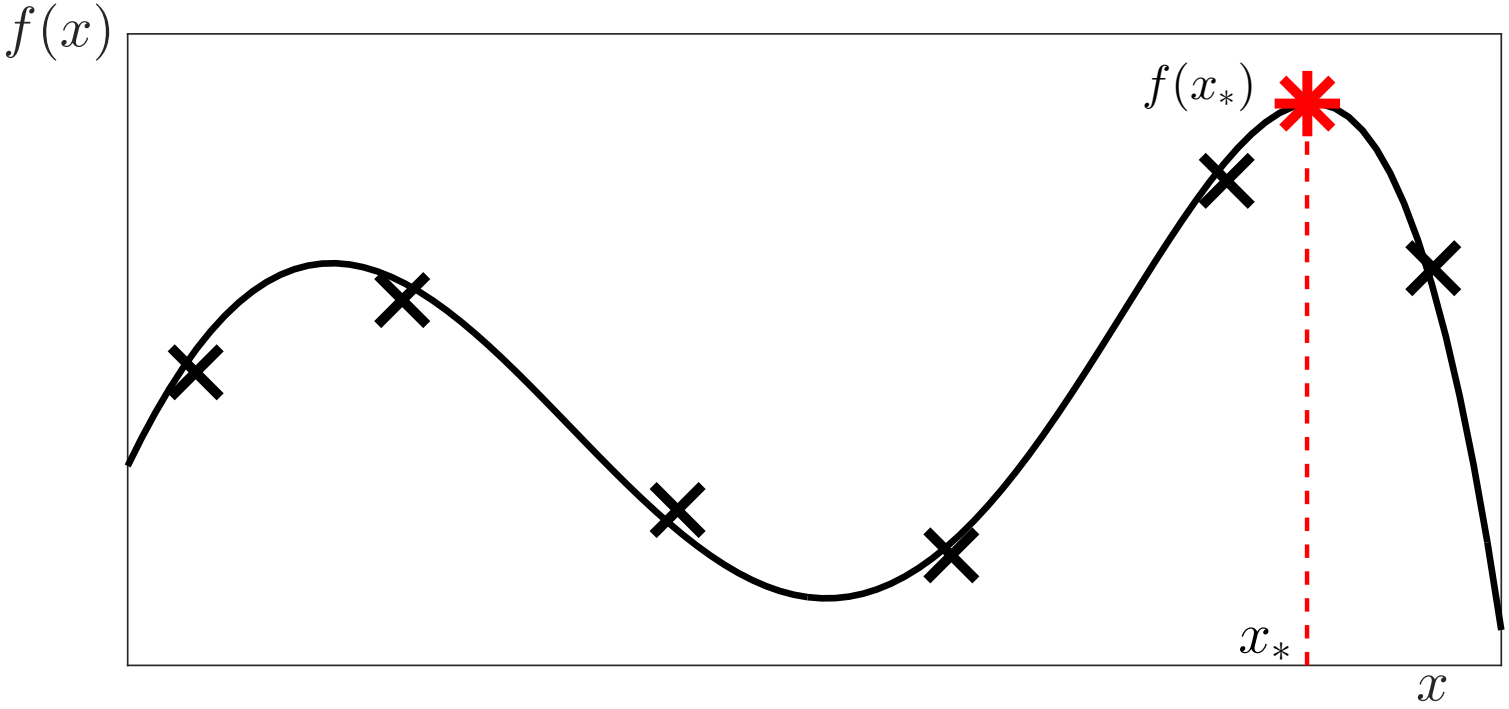


Cumulative Regret after n evaluations

$$R_n = \sum_{t=1}^{n} f(x_\star) \; - f(x_t).$$

Bandit Optimisation

$f : \mathcal{X} \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.



Simple Regret after n evaluations

$$S_n = f(x_\star) \; - \max_{t=1,\ldots,n} f(x_t).$$

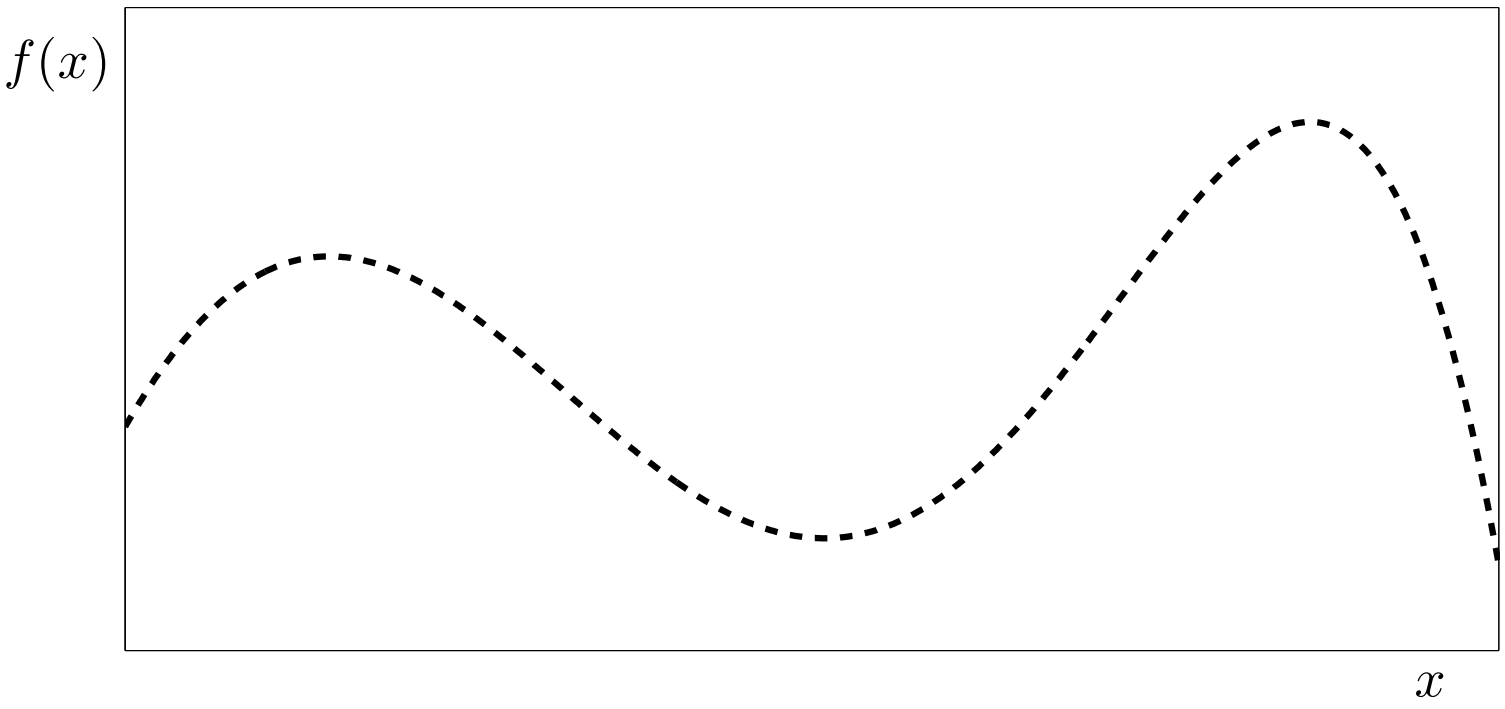
Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

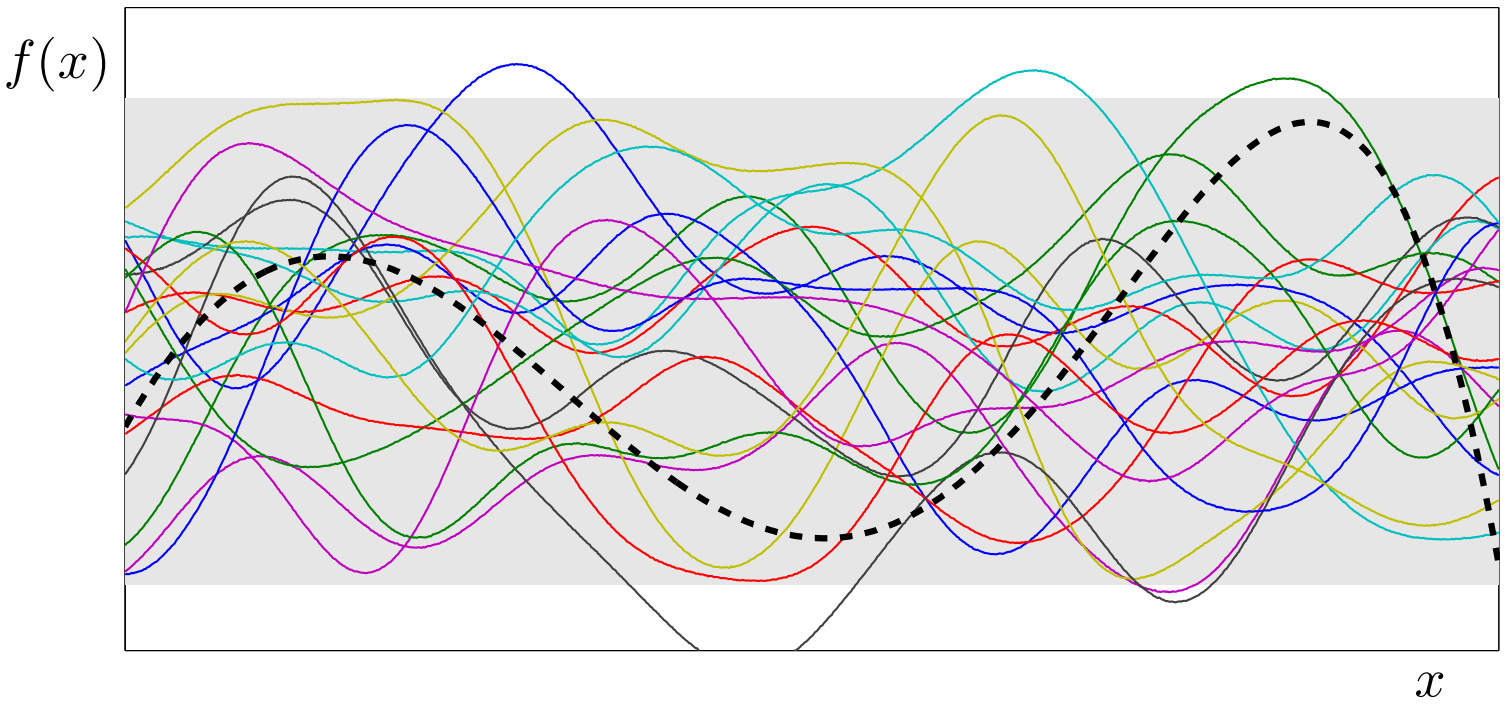
Functions with no observations



Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

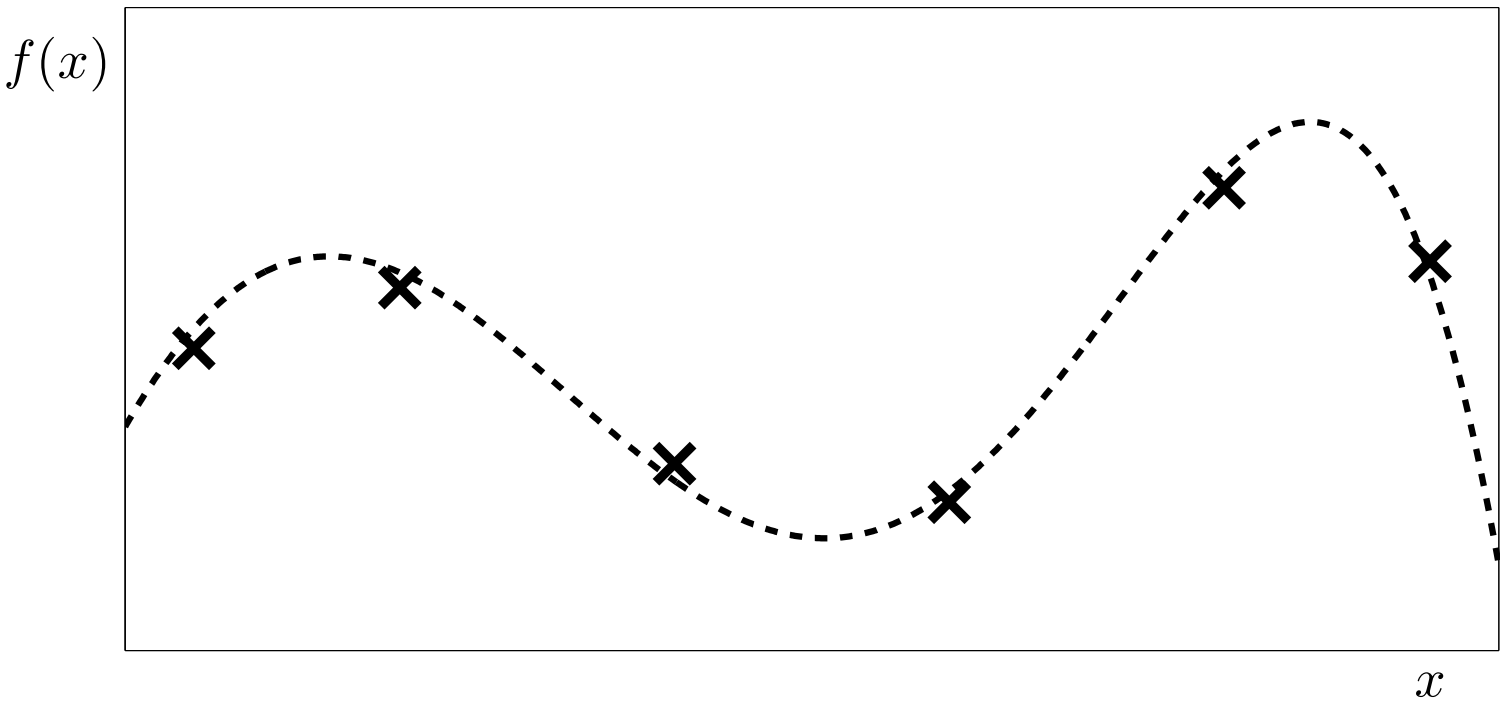
Prior $\mathcal{GP}$



Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

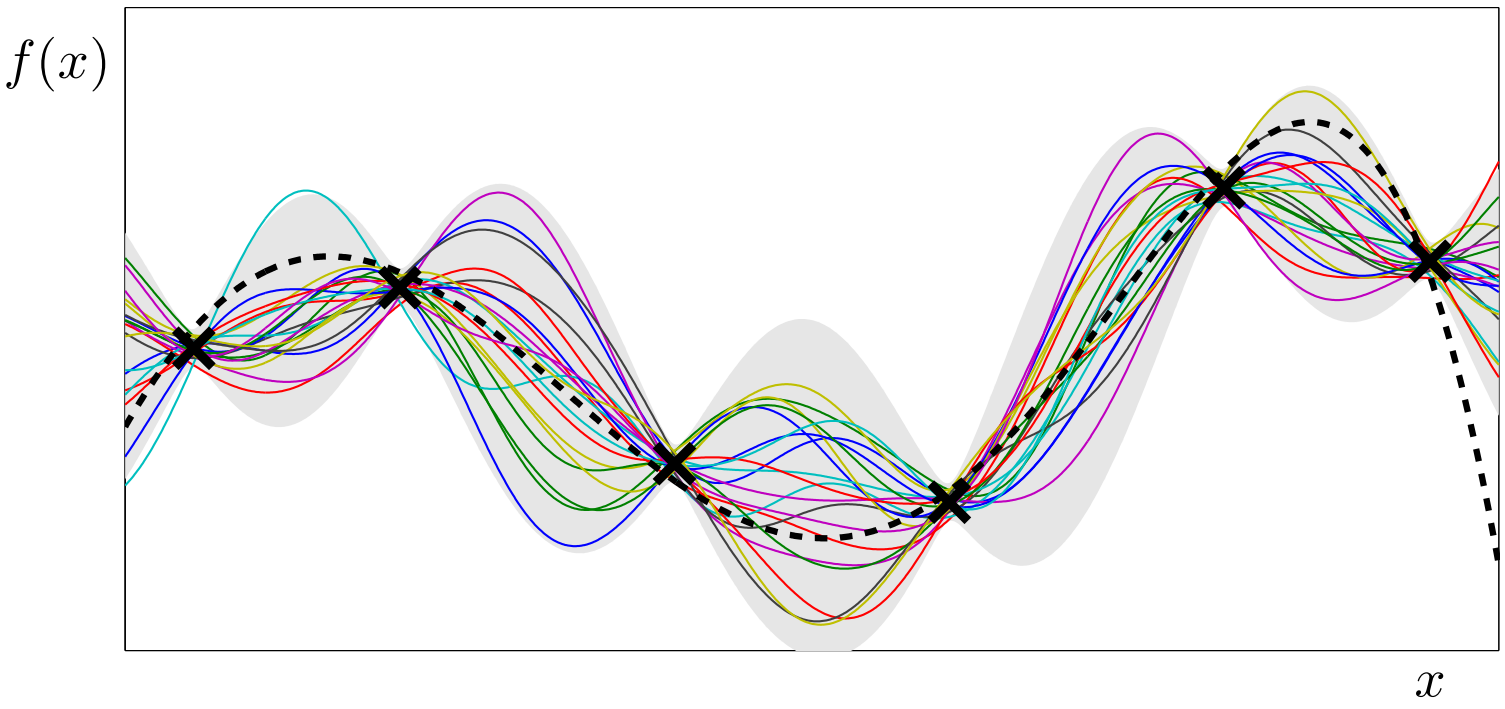
Observations



Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

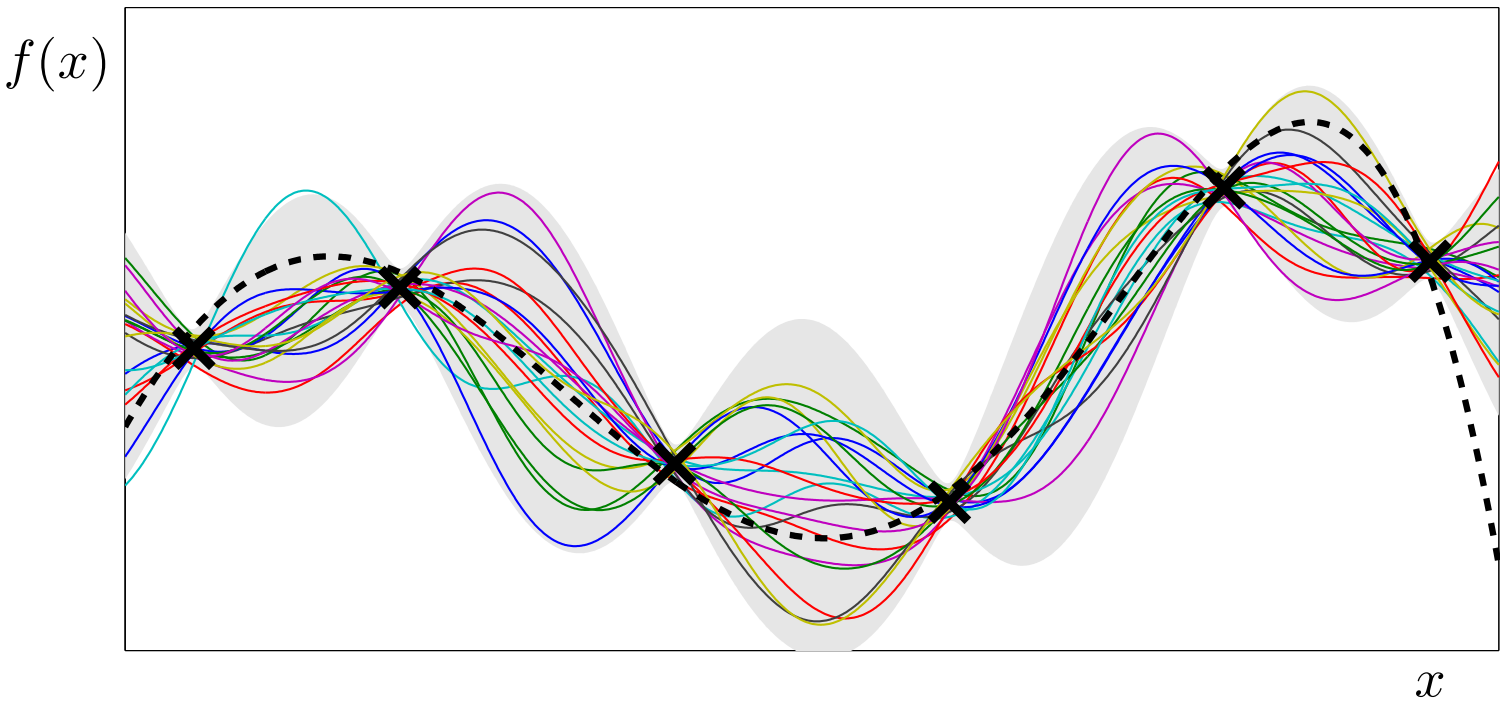
Posterior $\mathcal{GP}$ given observations



Gaussian Processes ($\mathcal{GP}$)

$\mathcal{GP}(\mu, \kappa)$: A distribution over functions from $\mathcal{X}$ to $\mathbb{R}$.
Mean $\mu : \mathcal{X} \to \mathbb{R}$, Covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$.

Posterior $\mathcal{GP}$ given observations



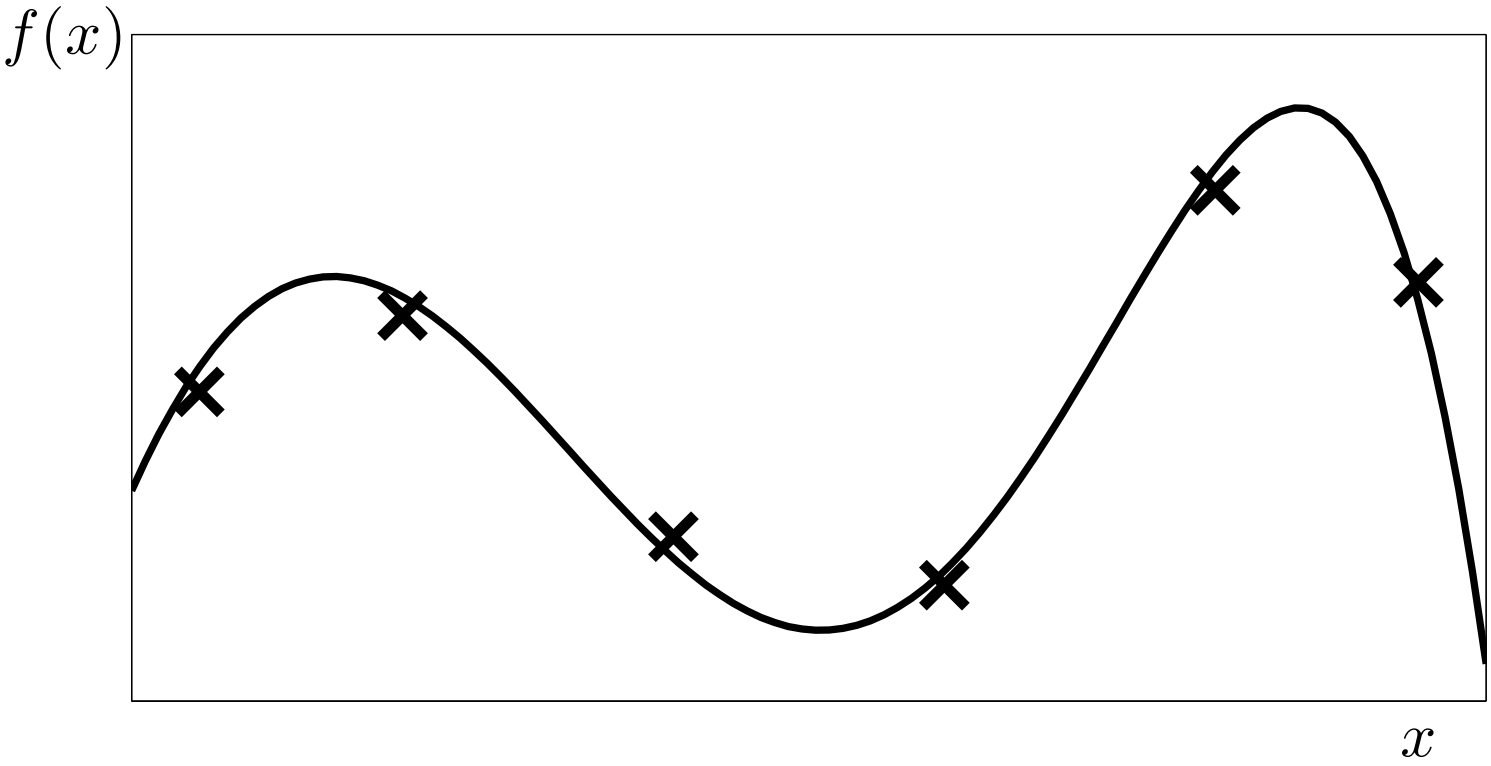
After t observations, $f(x) \sim \mathcal{N}(\,\mu_t(x), \sigma_t^2(x)\,)$.

Gaussian Process Bandit (Bayesian) Optimisation

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$.

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010).

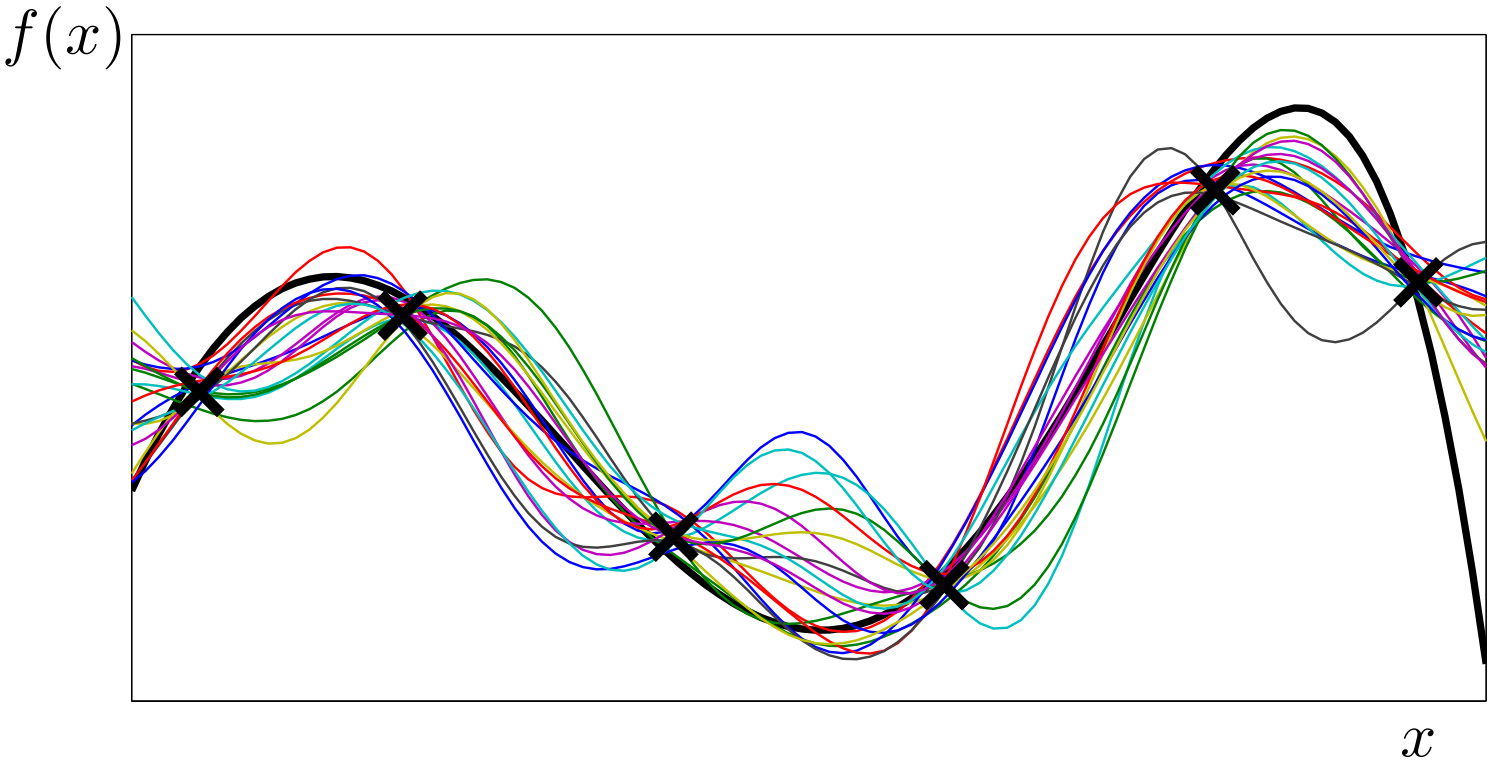


Gaussian Process Bandit (Bayesian) Optimisation

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$.

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010).

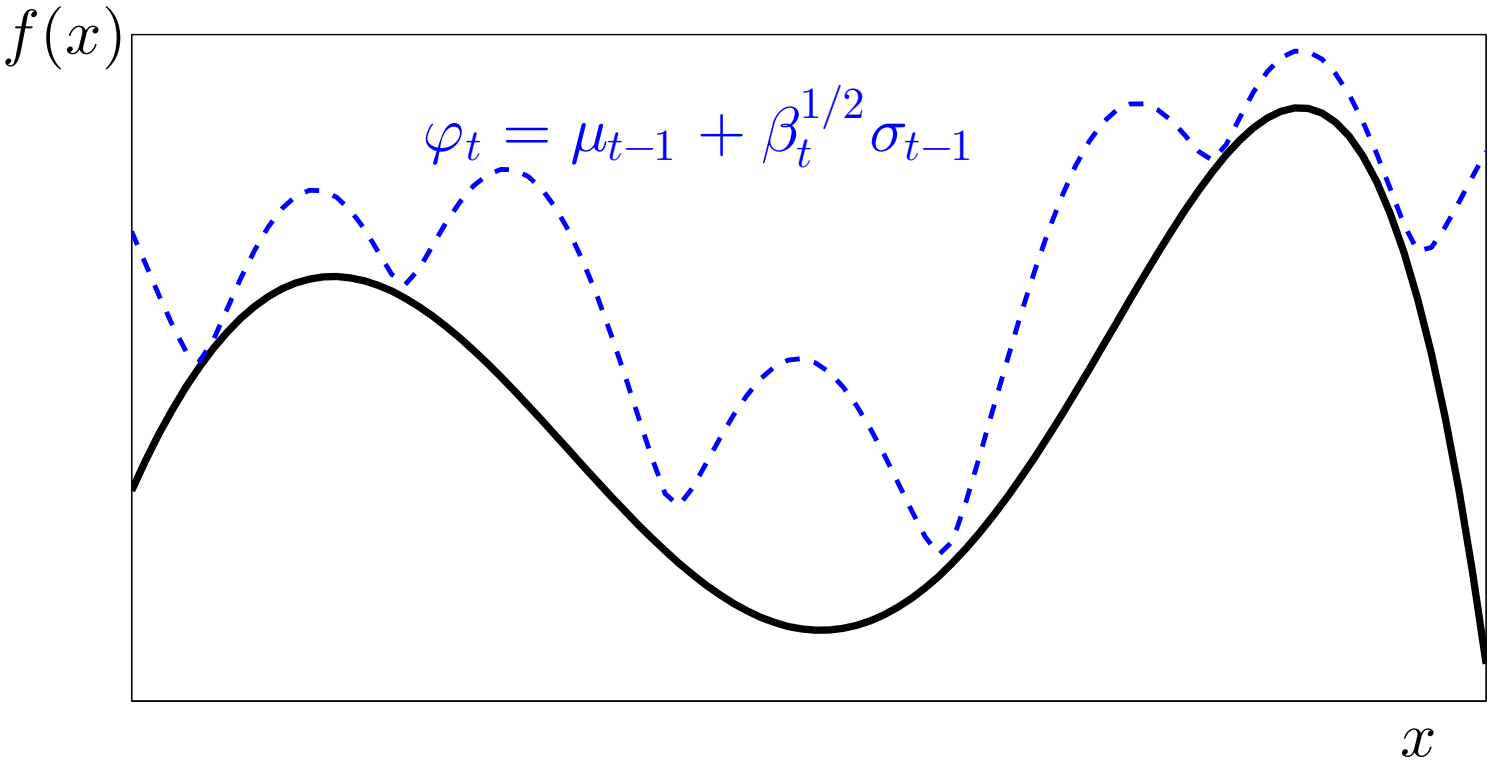


Gaussian Process Bandit (Bayesian) Optimisation

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$.

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010).



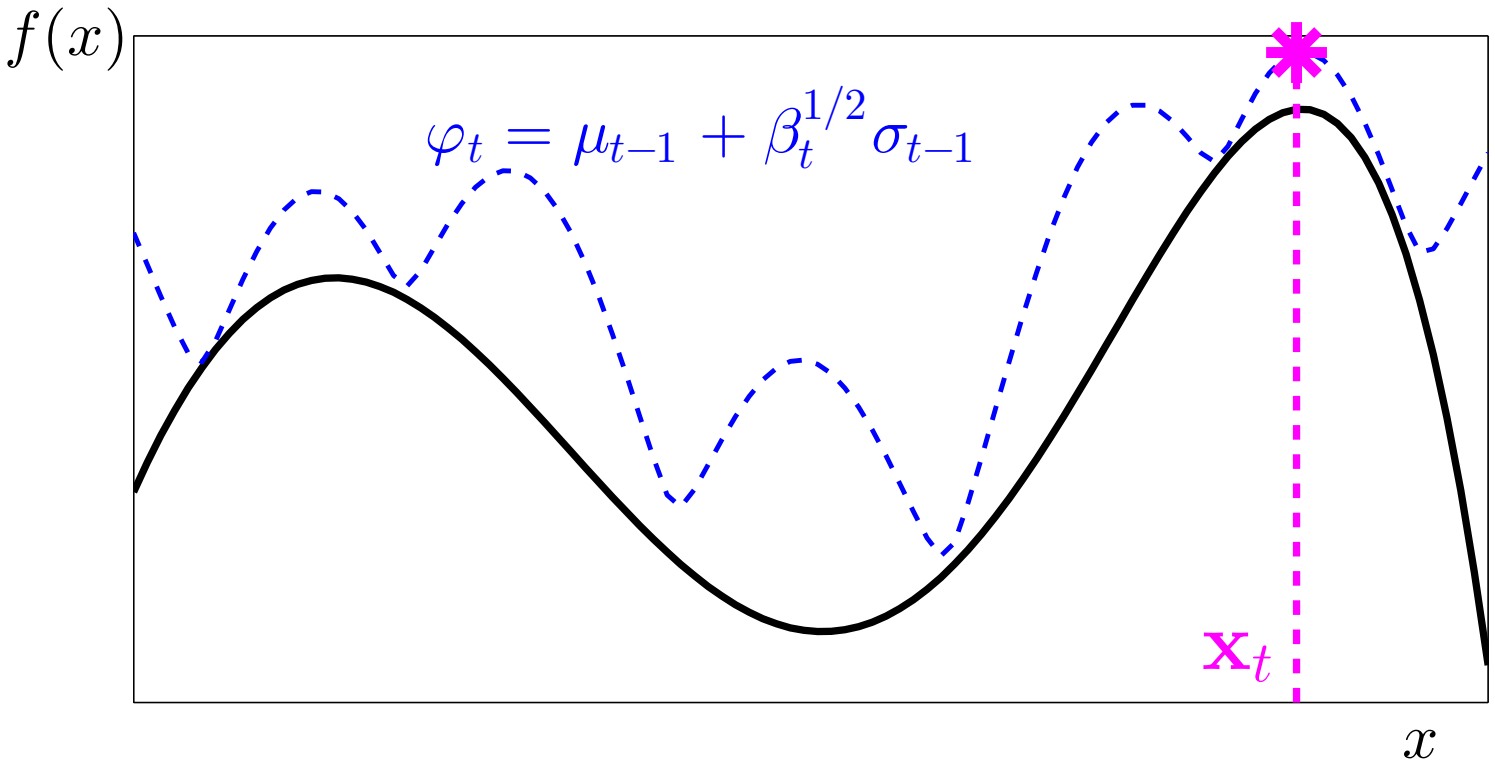
Construct upper conf. bound: $\varphi_t(x) = \mu_{t-1}(x) + \beta_t^{1/2}\sigma_{t-1}(x)$.

Gaussian Process Bandit (Bayesian) Optimisation

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$.

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010).



Maximise upper confidence bound.

GP-UCB

$$x_t = \operatorname*{argmax}_x \ \ \mu_{t-1}(x) + \beta_t^{1/2} \sigma_{t-1}(x)$$

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration
- β_t controls the tradeoff. $\quad \beta_t \asymp \log t$.

GP-UCB

$$x_t = \operatorname*{argmax}_x \ \ \mu_{t-1}(x) + \beta_t^{1/2}\sigma_{t-1}(x)$$

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration
- β_t controls the tradeoff. $\quad \beta_t \asymp \log t$.

GP-UCB (Srinivas et al. 2010)

w.h.p $\quad S_n \ = \ f(x_\star) - \max_{t=1,\dots,n} f(x_t) \ \lesssim \ \sqrt{\dfrac{\Psi_n(\mathcal{X})}{n}}$

$\Psi_n(\mathcal{X}) \leftarrow$ Maximum Information Gain.

GP-UCB

$$x_t = \operatorname*{argmax}_x \ \mu_{t-1}(x) + \beta_t^{1/2}\sigma_{t-1}(x)$$

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration
- β_t controls the tradeoff. $\beta_t \asymp \log t$.

GP-UCB (Srinivas et al. 2010)

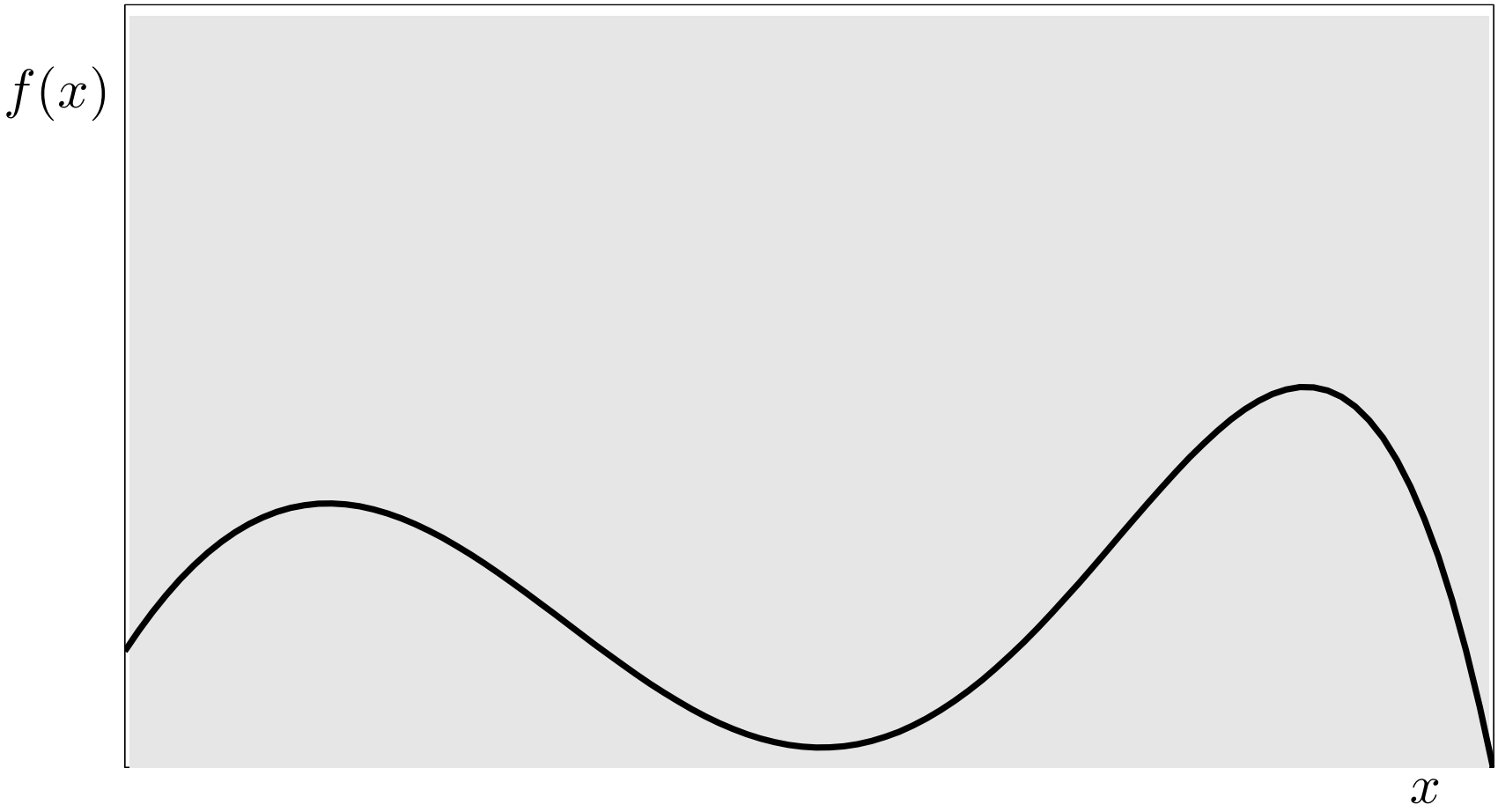
w.h.p $$S_n = f(x_\star) - \max_{t=1,\dots,n} f(x_t) \lesssim \sqrt{\frac{\Psi_n(\mathcal{X})}{n}}$$

$\Psi_n(\mathcal{X}) \leftarrow$ Maximum Information Gain.

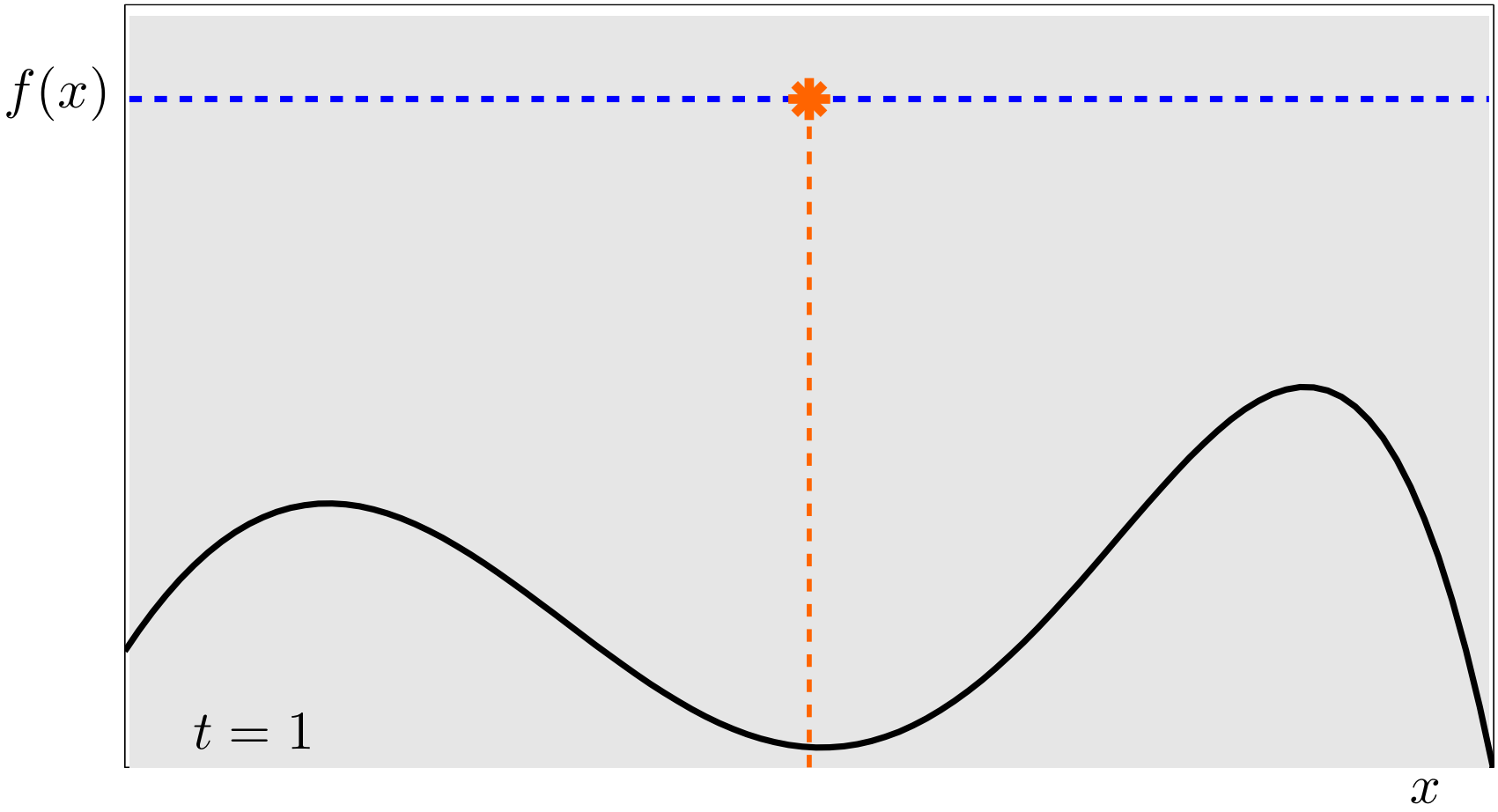
When $\mathcal{X} \subset \mathbb{R}^d$, SE kernel: $\Psi_n(\mathcal{X}) \asymp d^d \log(n)^d \cdot \mathrm{vol}(\mathcal{X})$.

Matérn kernel: $\Psi_n(\mathcal{X}) \asymp n^{1-\frac{1}{d^2}} \cdot \mathrm{vol}(\mathcal{X})$.

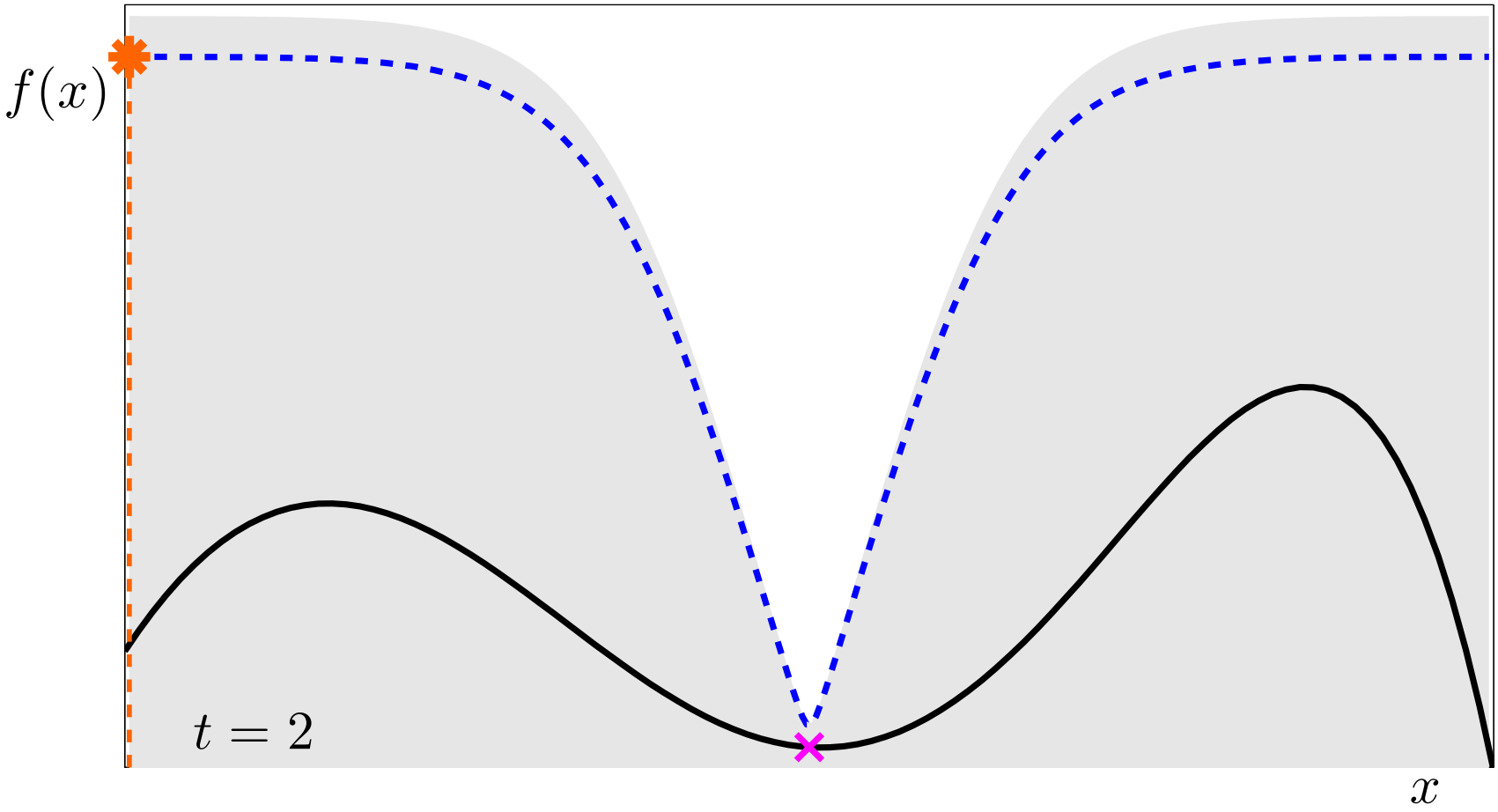
GP-UCB (Srinivas et al. 2010)



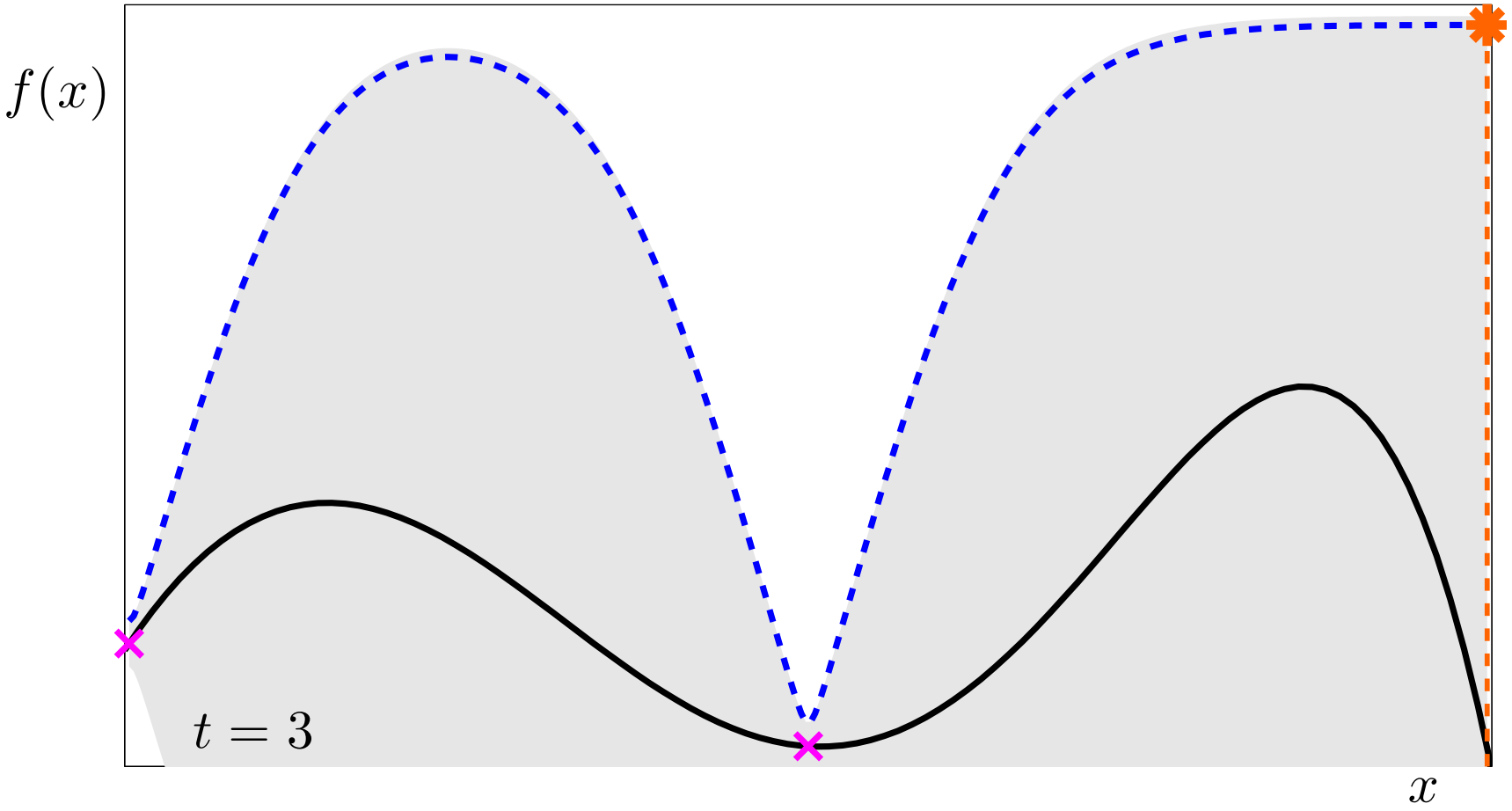
GP-UCB (Srinivas et al. 2010)



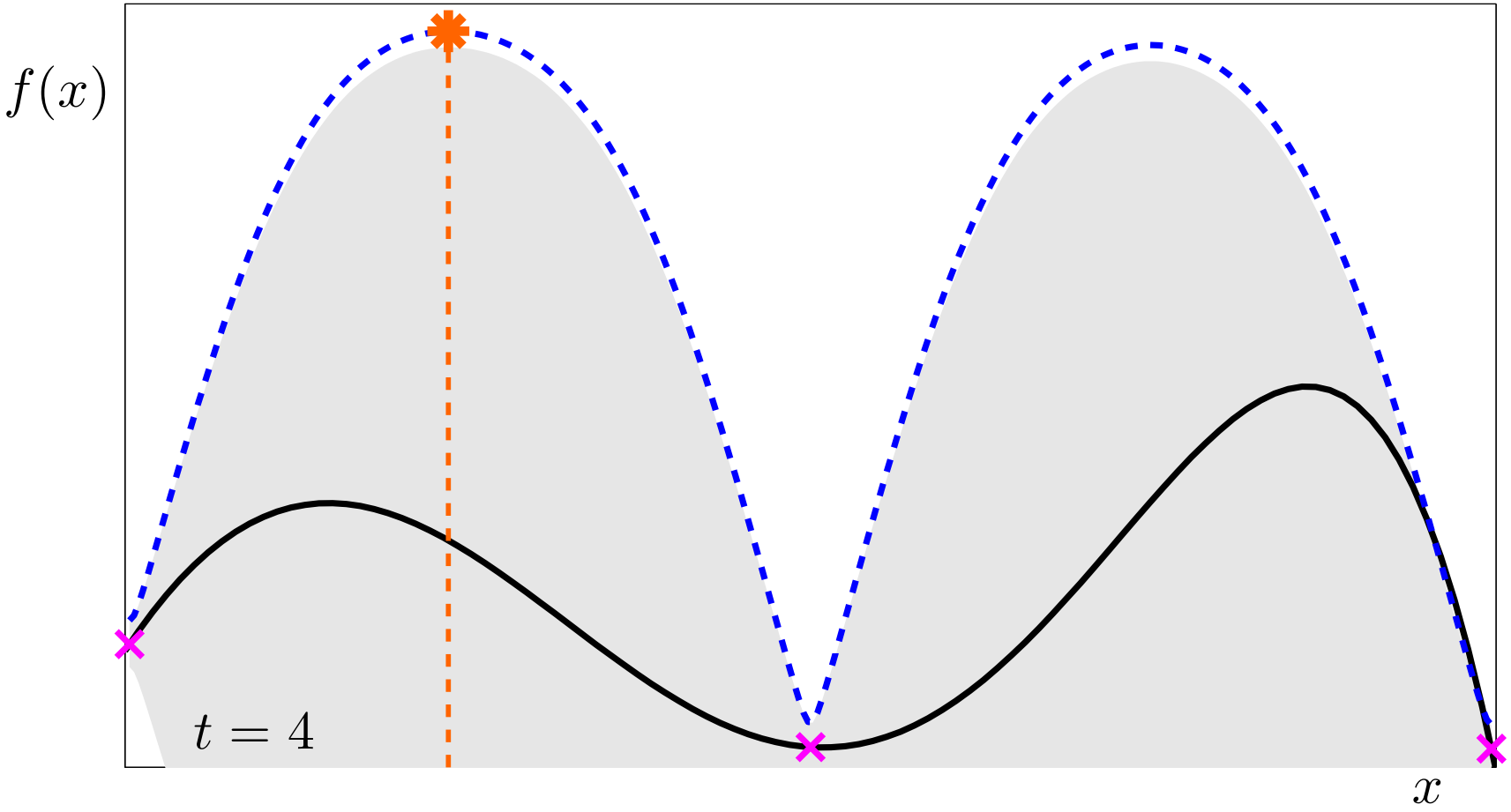
GP-UCB (Srinivas et al. 2010)



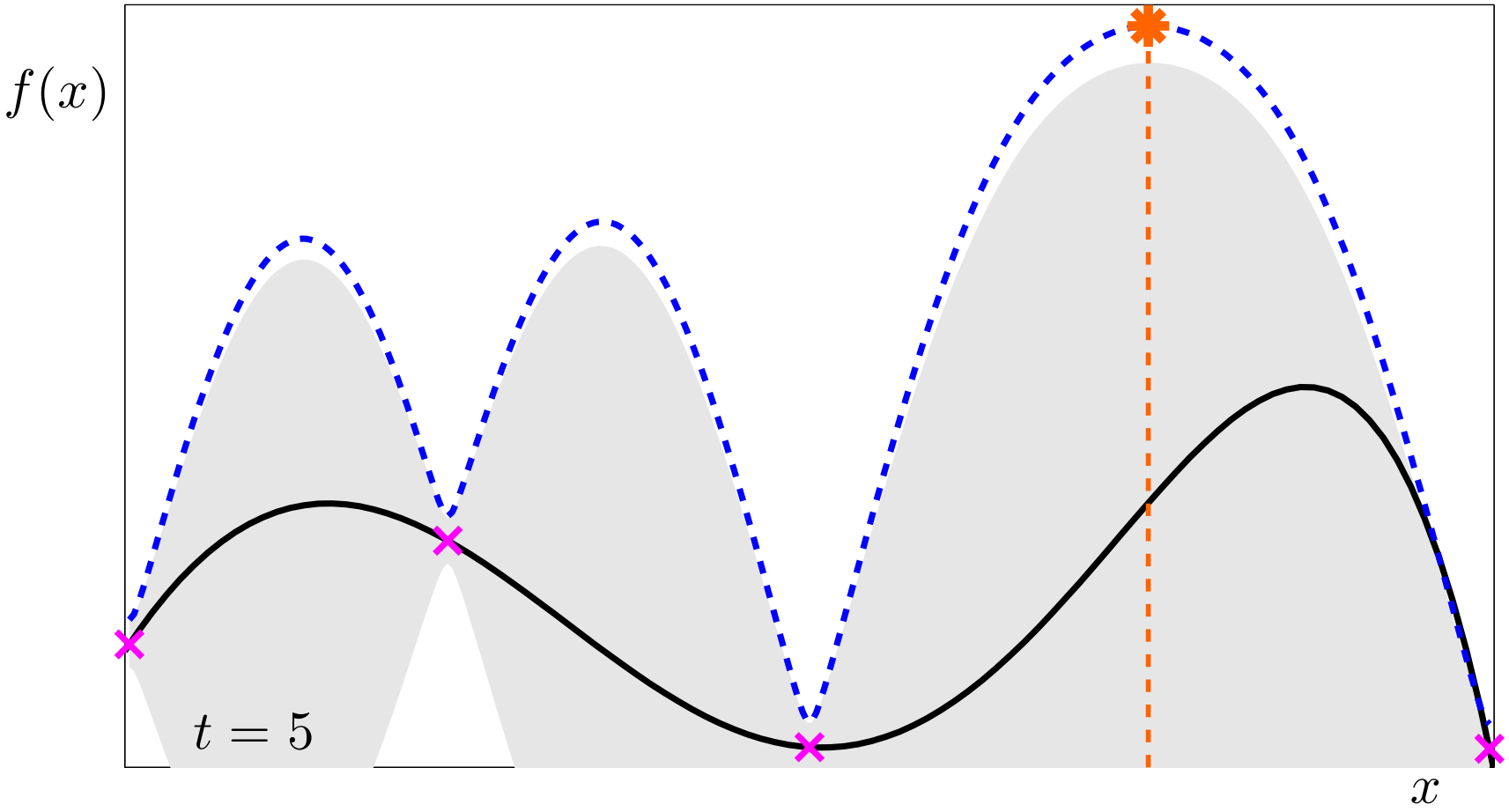
GP-UCB (Srinivas et al. 2010)



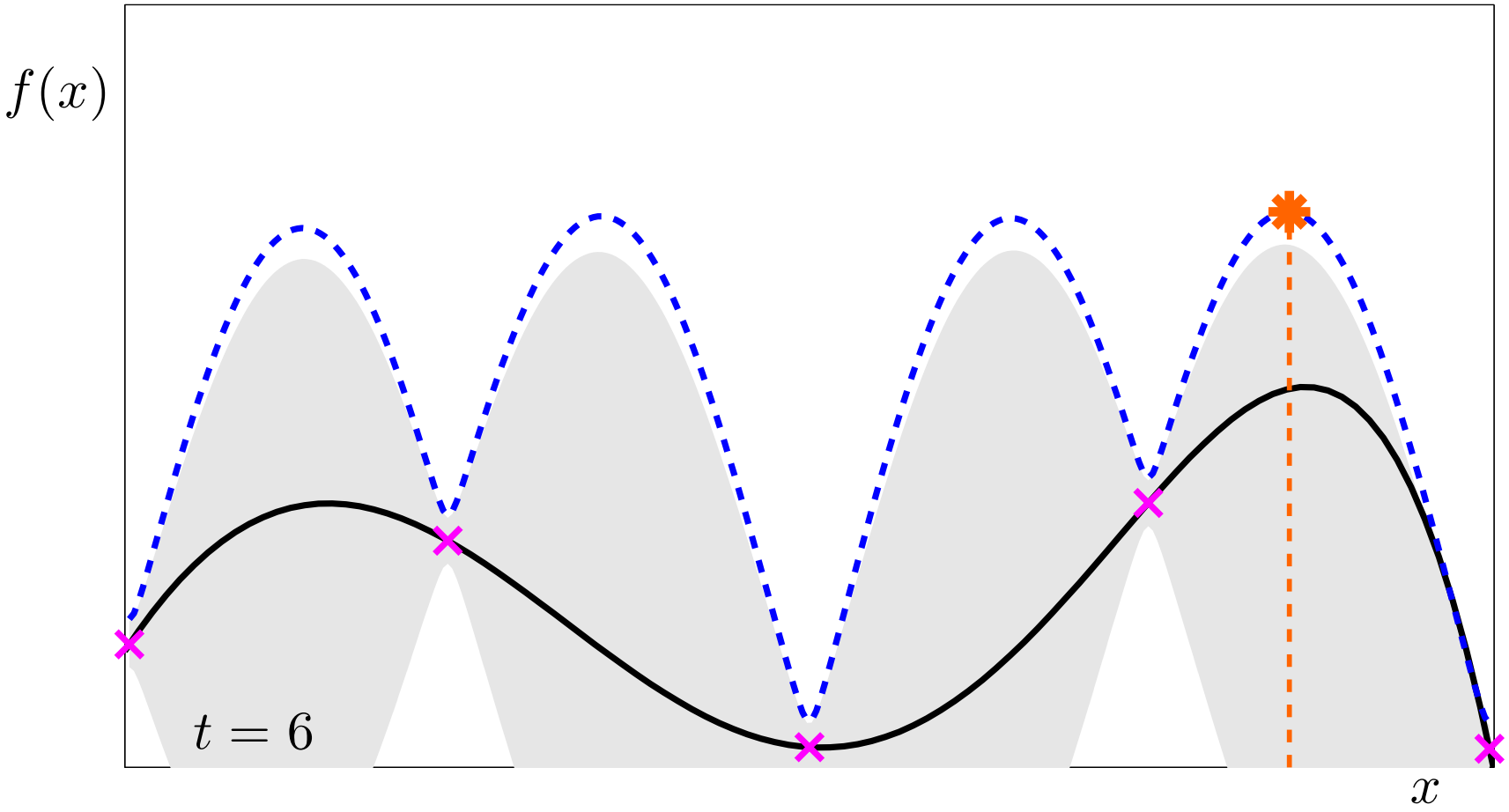
GP-UCB (Srinivas et al. 2010)



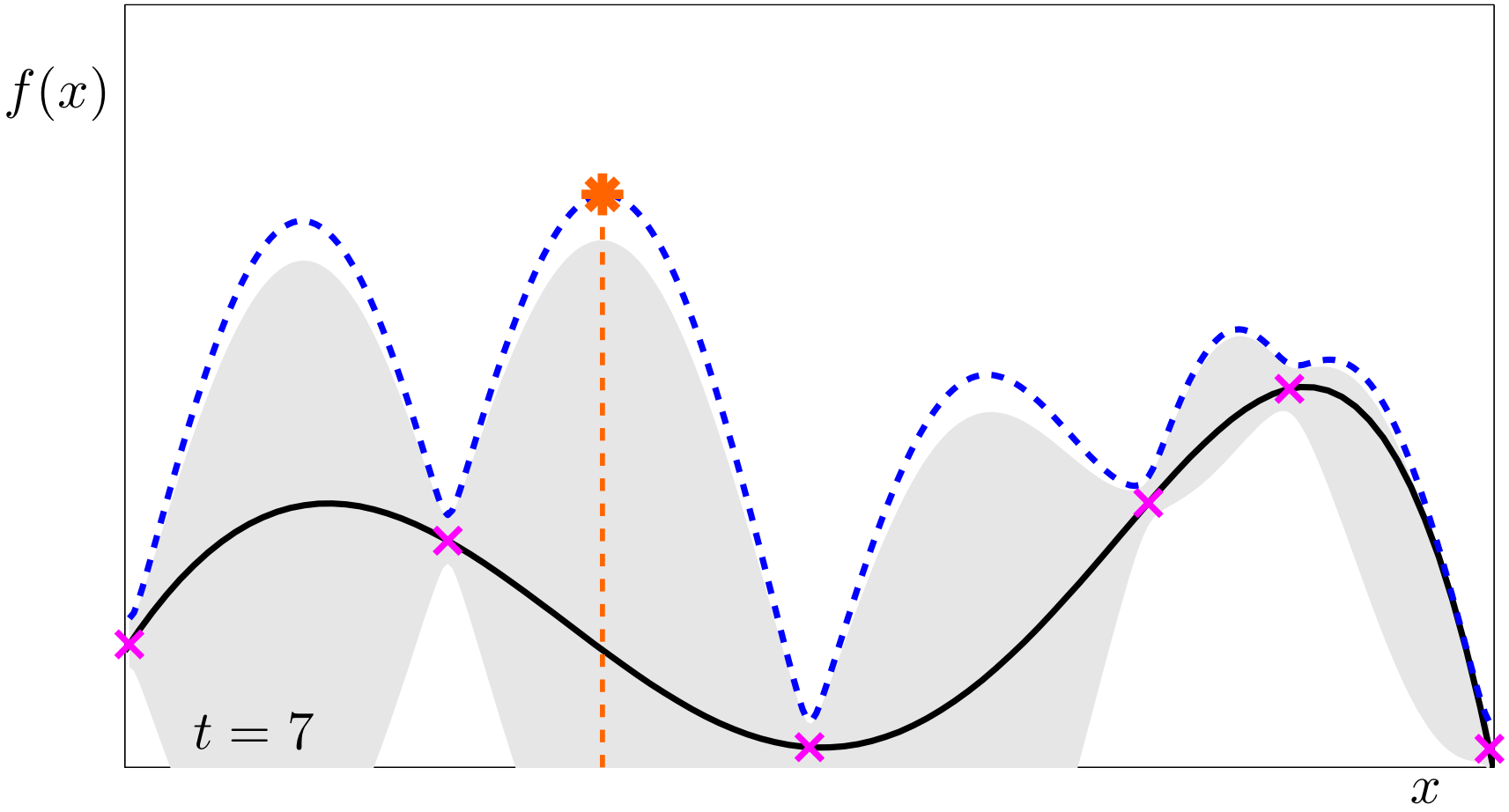
GP-UCB (Srinivas et al. 2010)



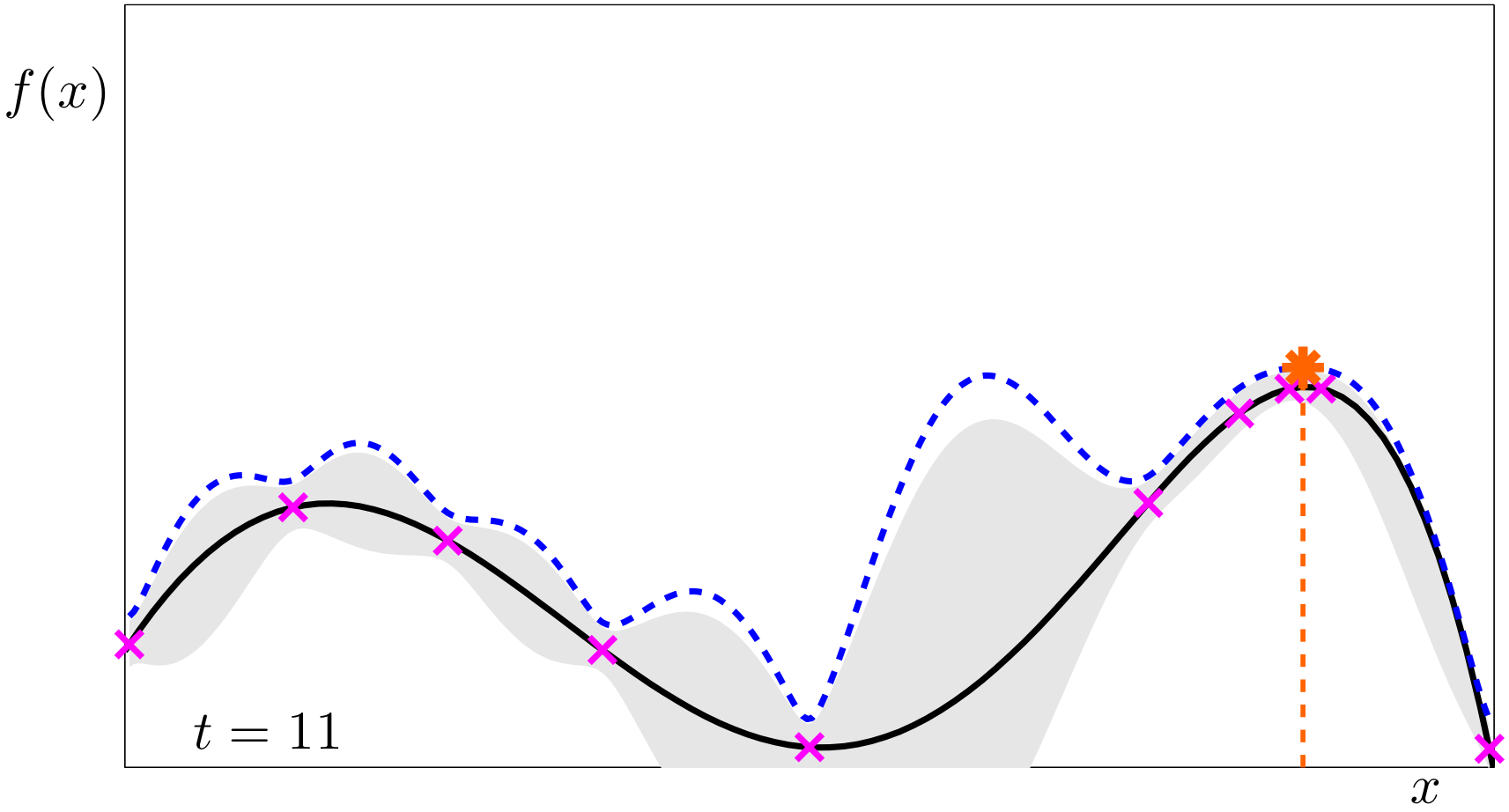
GP-UCB (Srinivas et al. 2010)



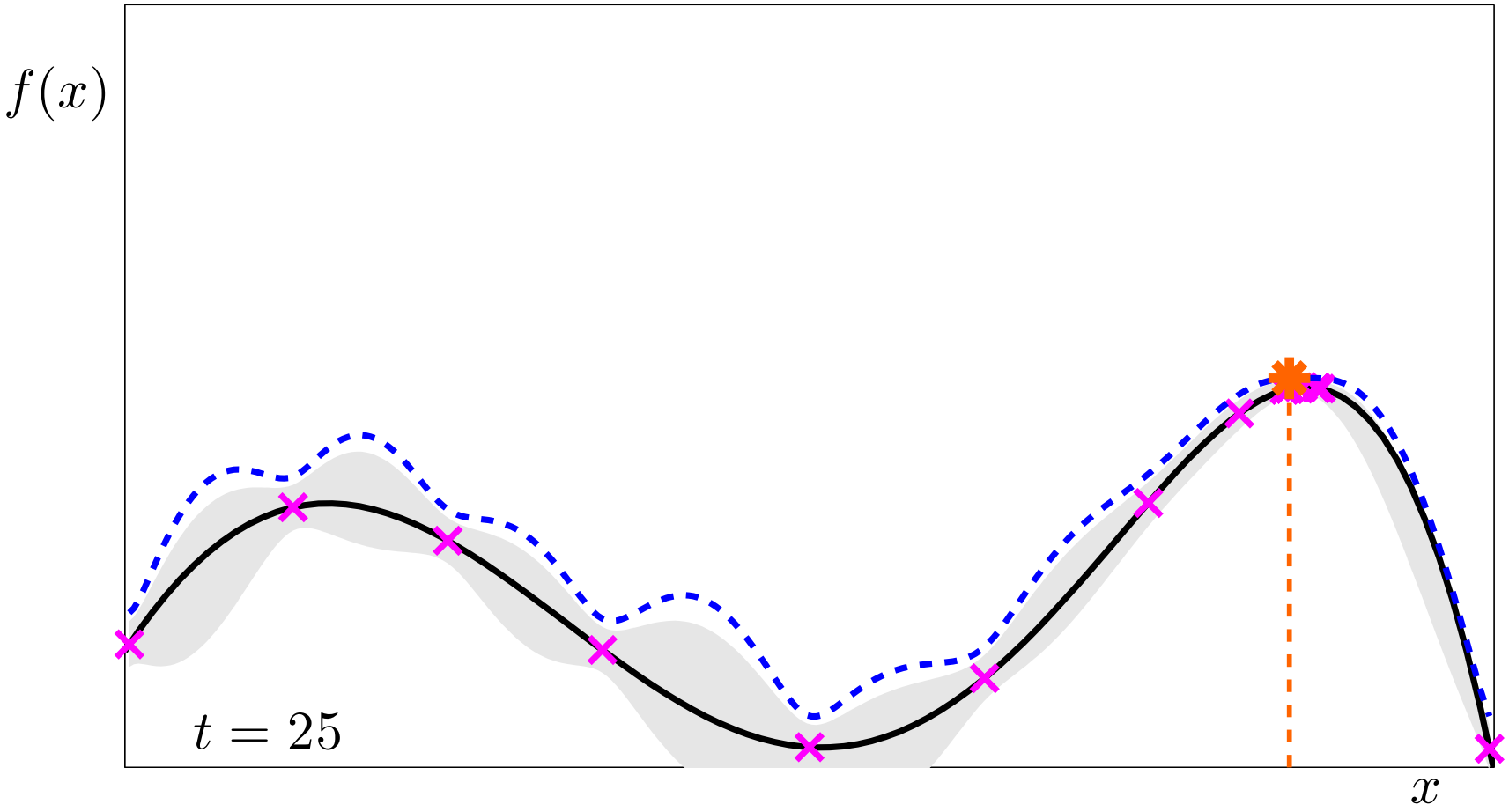
GP-UCB (Srinivas et al. 2010)



GP-UCB (Srinivas et al. 2010)



GP-UCB (Srinivas et al. 2010)

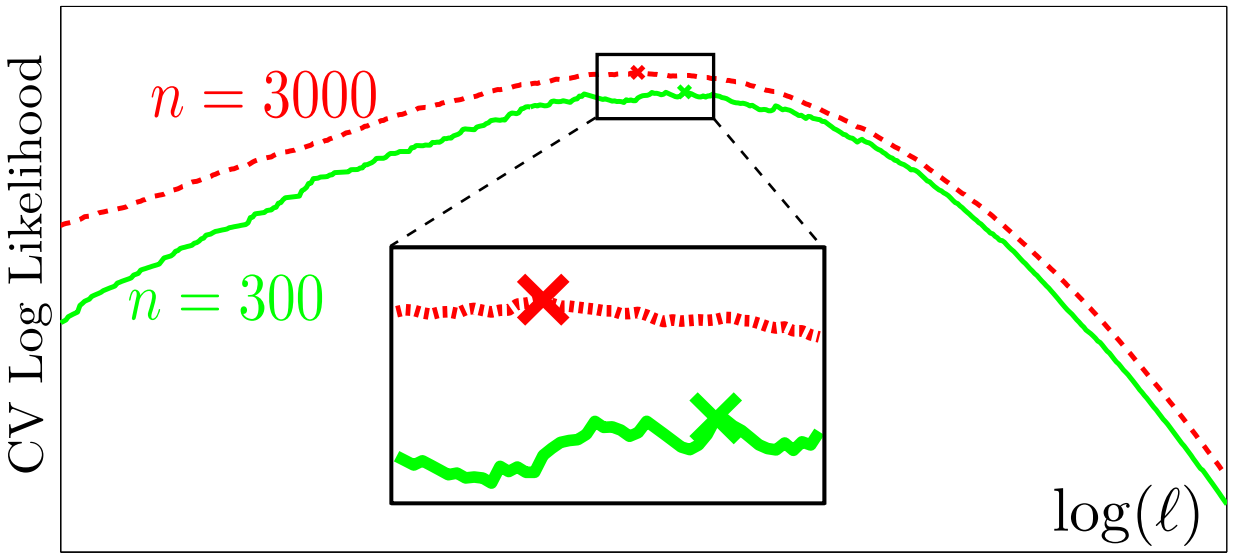


What if we have cheap approximations to f ?

What if we have cheap approximations to f ?

1. Hyper-parameter tuning: Train & validate with a subset of the data, and/or early stopping before convergence.

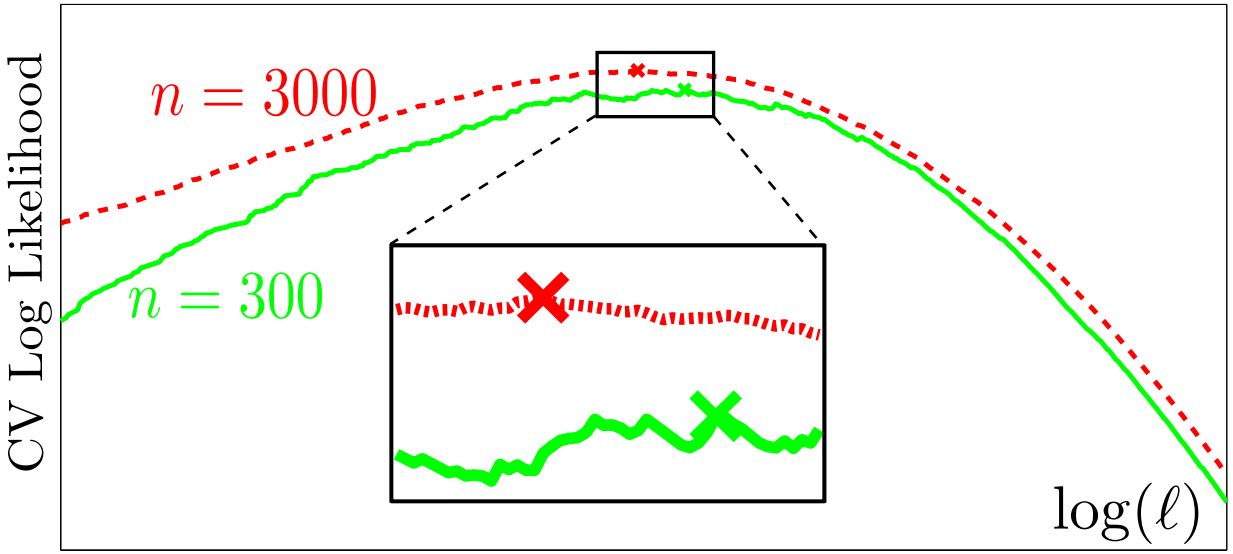
   E.g. Bandwidth (ℓ) selection in kernel density estimation.



What if we have cheap approximations to f ?

1. Hyper-parameter tuning: Train & validate with a subset of the data, and/or early stopping before convergence.

   E.g. Bandwidth (ℓ) selection in kernel density estimation.



2. Autonomous driving: simulation vs real world experiment.

3. Computational astrophysics: cosmological simulations and numerical computations with less granularity.

Prior work in **Multi-fidelity** Methods

For specific applications,

- Industrial design (Forrester et al. 2007)
- Hyper-parameter tuning (Agarwal et al. 2011, Klein et al. 2015, Li et al. 2016)
- Active learning (Zhang & Chaudhuri 2015)
- Robotics (Cutler et al. 2014)

Multi-fidelity optimisation

(Huang et al. 2006, Forrester et al. 2007, March & Wilcox 2012, Poloczek et al. 2016)

Multi-fidelity GP Bandit Optimisation

1. A finite number of approximations

   (Kandasamy et al. NIPS 2016b)

   - Formalism and challenges
   - Algorithm
   - Theoretical results & proof sketches
   - Experiments

2. A continuous spectrum of approximations

   (Kandasamy et al. Arxiv 2017)

   - Formalism
   - Algorithm
   - Theoretical results
   - Experiments

Multi-fidelity GP Bandit Optimisation

1. A finite number of approximations

   (Kandasamy et al. NIPS 2016b)

   - Formalism and challenges
   - Algorithm
   - Theoretical results & proof sketches
   - Experiments

2. A continuous spectrum of approximations

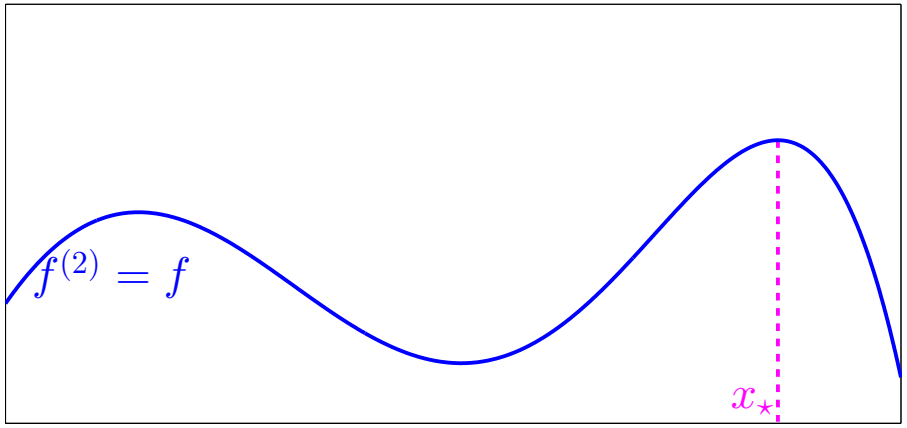
   (Kandasamy et al. Arxiv 2017)

   - Formalism
   - Algorithm
   - Theoretical results
   - Experiments

Extends beyond GPs.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

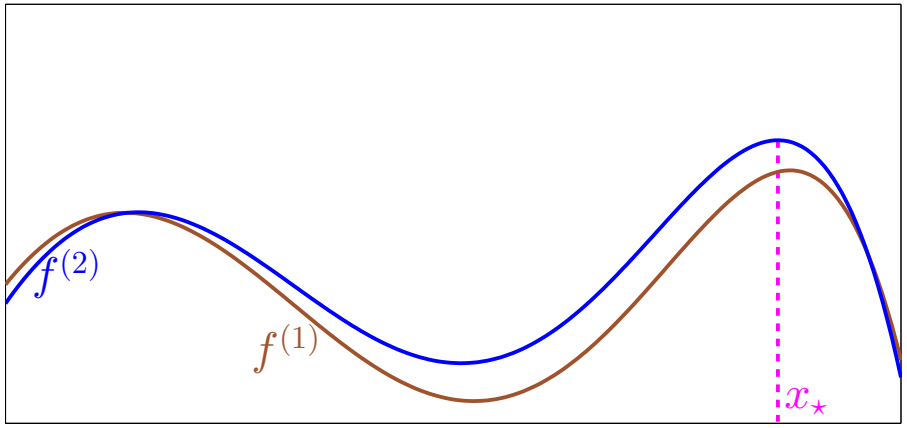
(Kandasamy et al. NIPS 2016b)



- Optimise $f = f^{(2)}$. $\quad x_\star = \mathrm{argmax}_x \, f^{(2)}(x)$.
- **But ..**

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

(Kandasamy et al. NIPS 2016b)

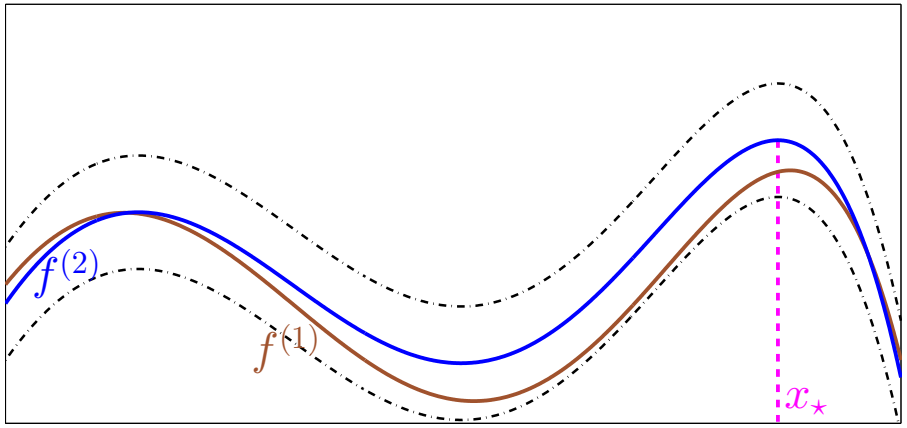


- Optimise $f = f^{(2)}$. $\quad x_\star = \mathrm{argmax}_x \, f^{(2)}(x)$.
- **But ..** we have an approximation $f^{(1)}$ to $f^{(2)}$.
- $f^{(1)}$ costs $\lambda^{(1)}$, $\quad f^{(2)}$ costs $\lambda^{(2)}$. $\quad \lambda^{(1)} < \lambda^{(2)}$.

  "cost": could be computation time, money etc.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

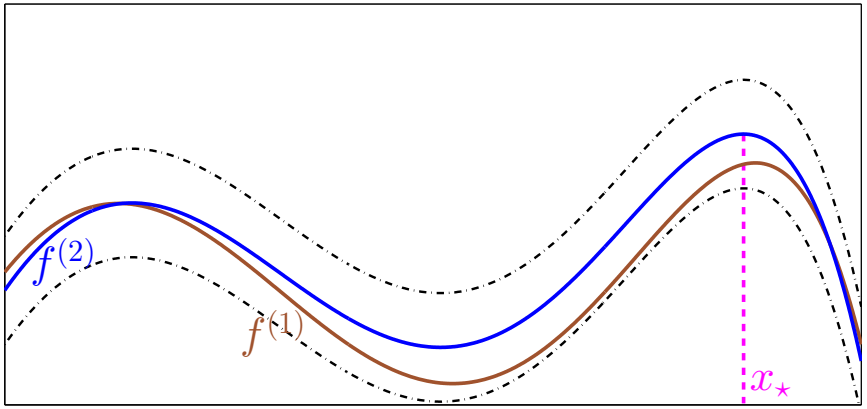
(Kandasamy et al. NIPS 2016b)



- Optimise $f = f^{(2)}$. $\quad x_\star = \mathrm{argmax}_x f^{(2)}(x)$.
- **But ..** we have an approximation $f^{(1)}$ to $f^{(2)}$.
- $f^{(1)}$ costs $\lambda^{(1)}$, $\quad f^{(2)}$ costs $\lambda^{(2)}$. $\quad \lambda^{(1)} < \lambda^{(2)}$.
  "cost": could be computation time, money etc.
- $f^{(1)}, f^{(2)} \sim \mathcal{GP}(0, \kappa)$.
- $\|f^{(2)} - f^{(1)}\|_\infty \leq \zeta^{(1)}$. $\quad \zeta^{(1)}$ is known.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

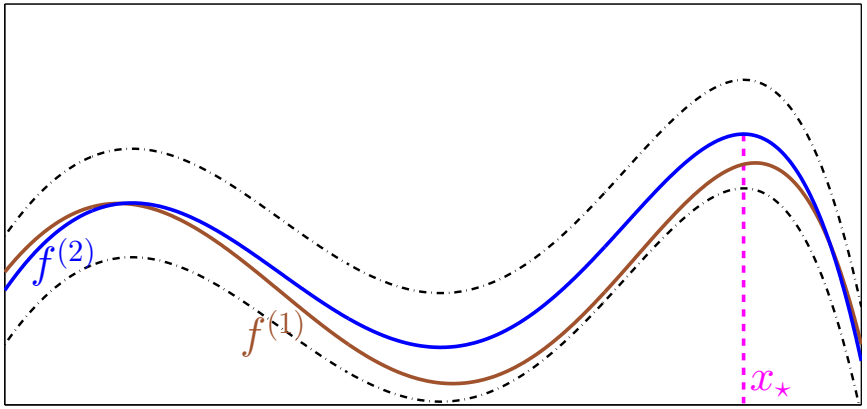
(Kandasamy et al. NIPS 2016b)



At time t: Determine the point $x_t \in \mathcal{X}$ and fidelity $m_t \in \{1, 2\}$ for querying.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

(Kandasamy et al. NIPS 2016b)

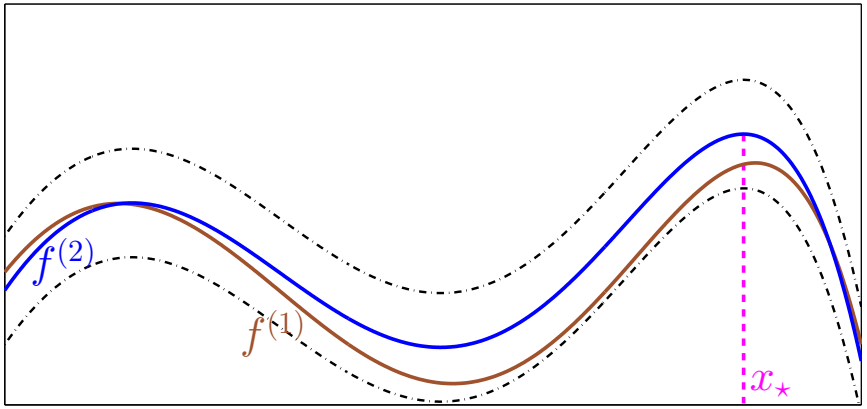


At time t: Determine the point $x_t \in \mathcal{X}$ and fidelity $m_t \in \{1, 2\}$ for querying.

End Goal: Maximise $f^{(2)}$. Don't care for maximum of $f^{(1)}$.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

(Kandasamy et al. NIPS 2016b)



At time t: Determine the point $x_t \in \mathcal{X}$ and fidelity $m_t \in \{1, 2\}$ for querying.

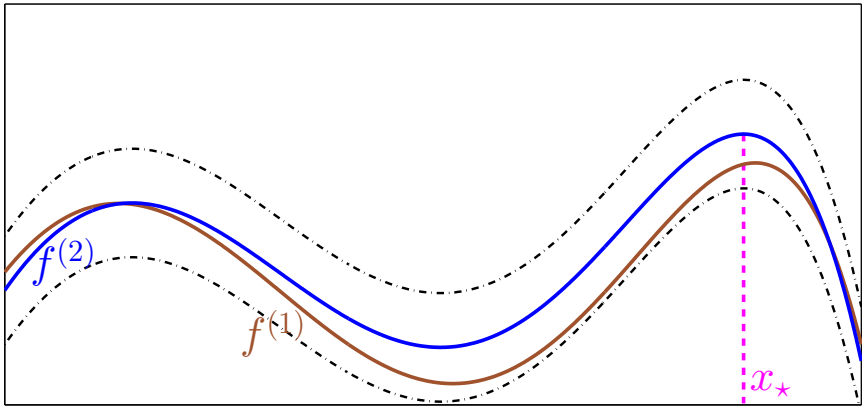
End Goal: Maximise $f^{(2)}$. Don't care for maximum of $f^{(1)}$.

Simple Regret: $$S(\Lambda) = f^{(2)}(x_\star) - \max_{t\,:\,m_t=2} f^{(2)}(x_t)$$

$S(\Lambda) = +\infty$ if we haven't queried $f^{(2)}$ yet.

Multi-fidelity Bandit Optimisation in 2 Fidelities (1 Approximation)

(Kandasamy et al. NIPS 2016b)



At time t: Determine the point $x_t \in \mathcal{X}$ and fidelity $m_t \in \{1, 2\}$ for querying.

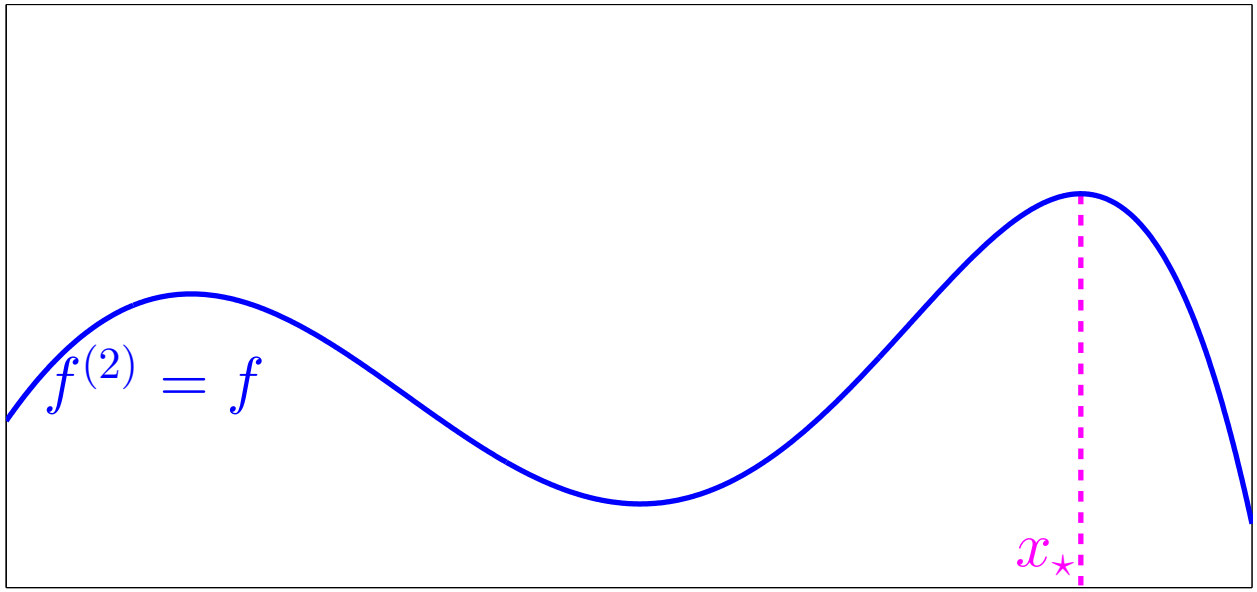
End Goal: Maximise $f^{(2)}$. Don't care for maximum of $f^{(1)}$.

Simple Regret: $$S(\Lambda) = f^{(2)}(x_\star) - \max_{t\,:\,m_t=2} f^{(2)}(x_t)$$

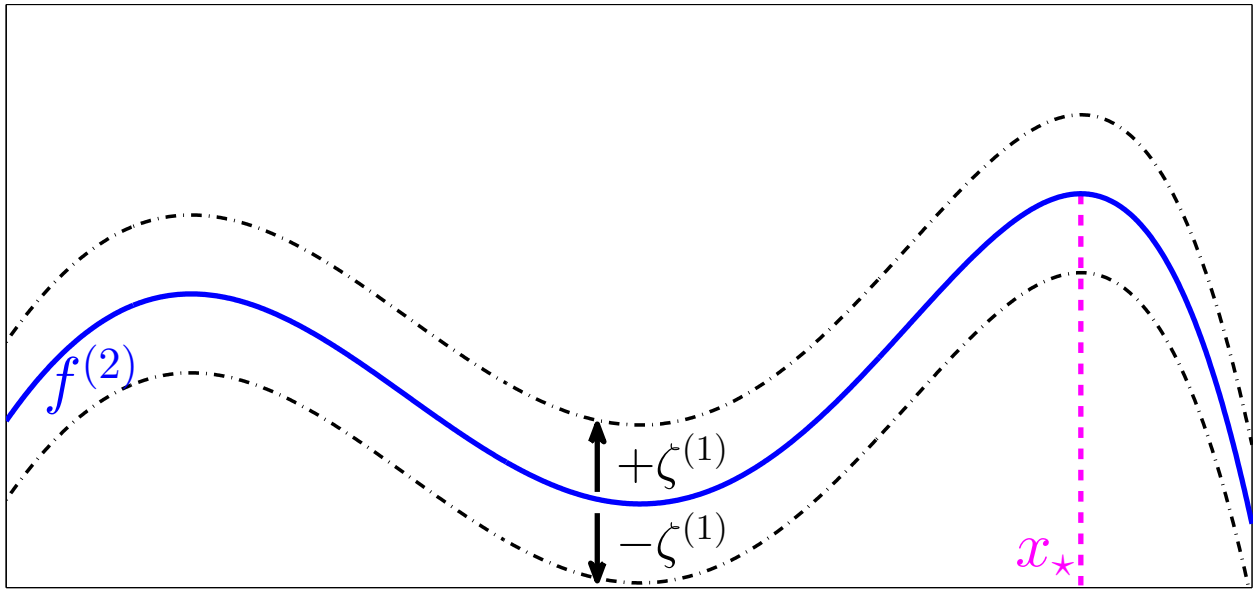
$S(\Lambda) = +\infty$ if we haven't queried $f^{(2)}$ yet.

$\rightarrow$ But use $f^{(1)}$ to guide search for $x_\star$ at $f^{(2)}$.

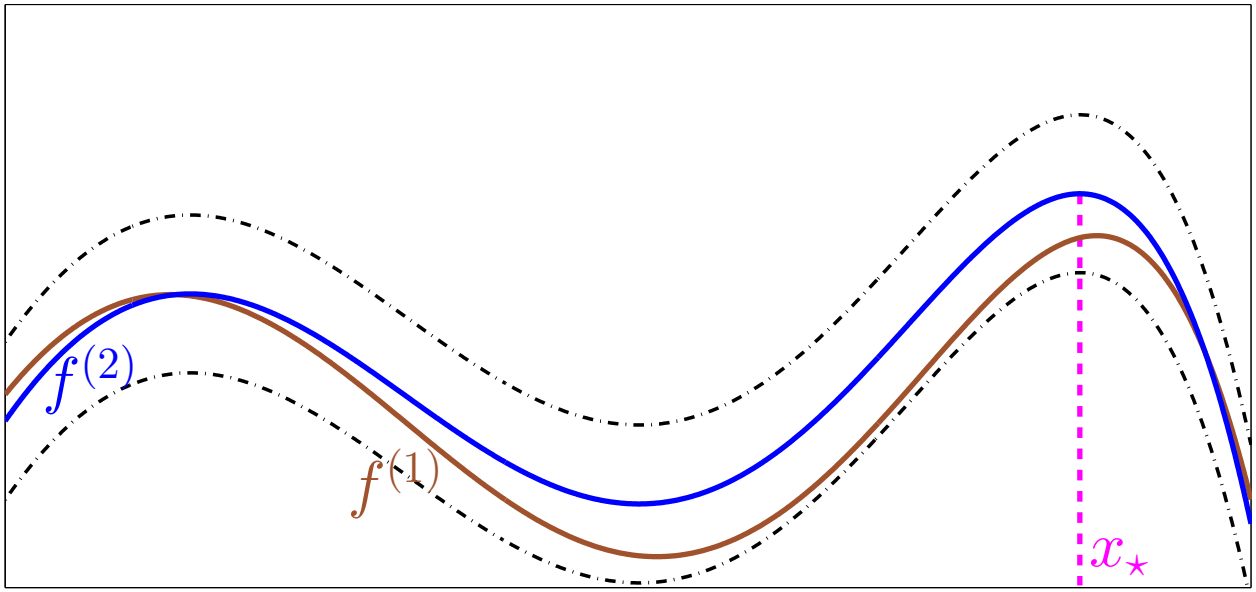
Challenges



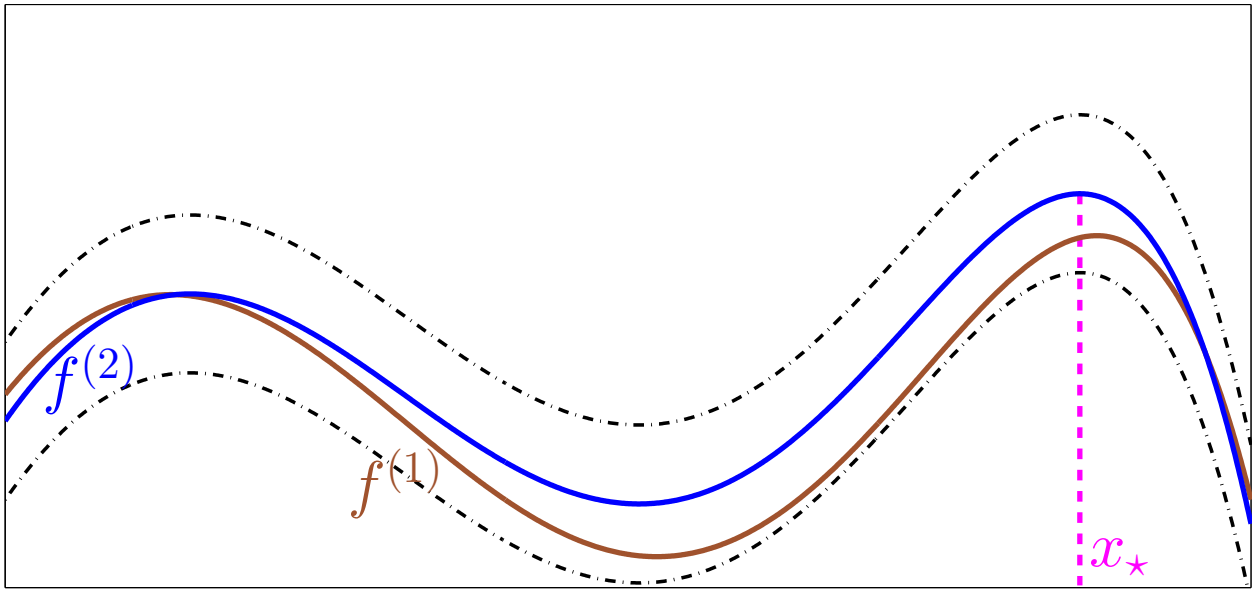
Challenges



Challenges

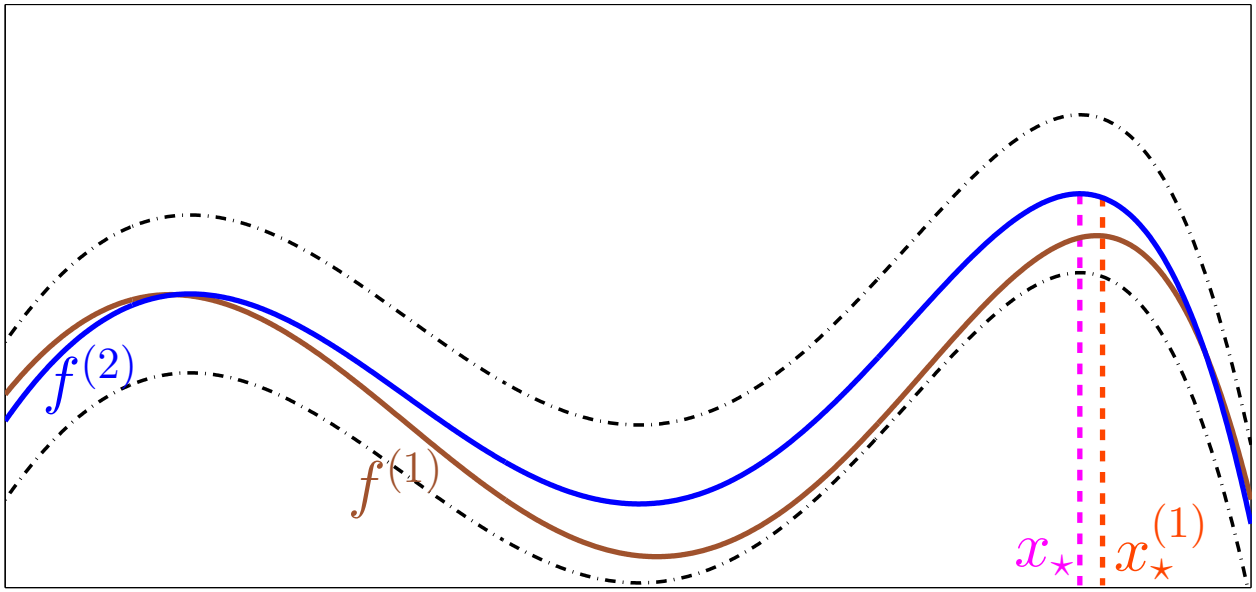


Challenges



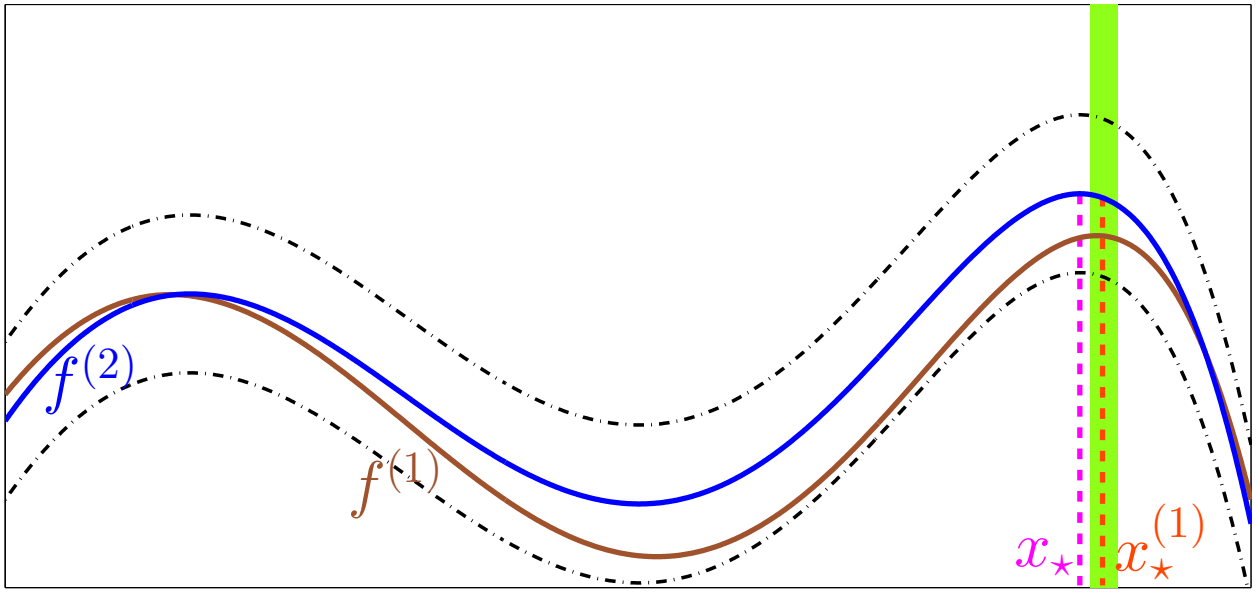
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.

Challenges



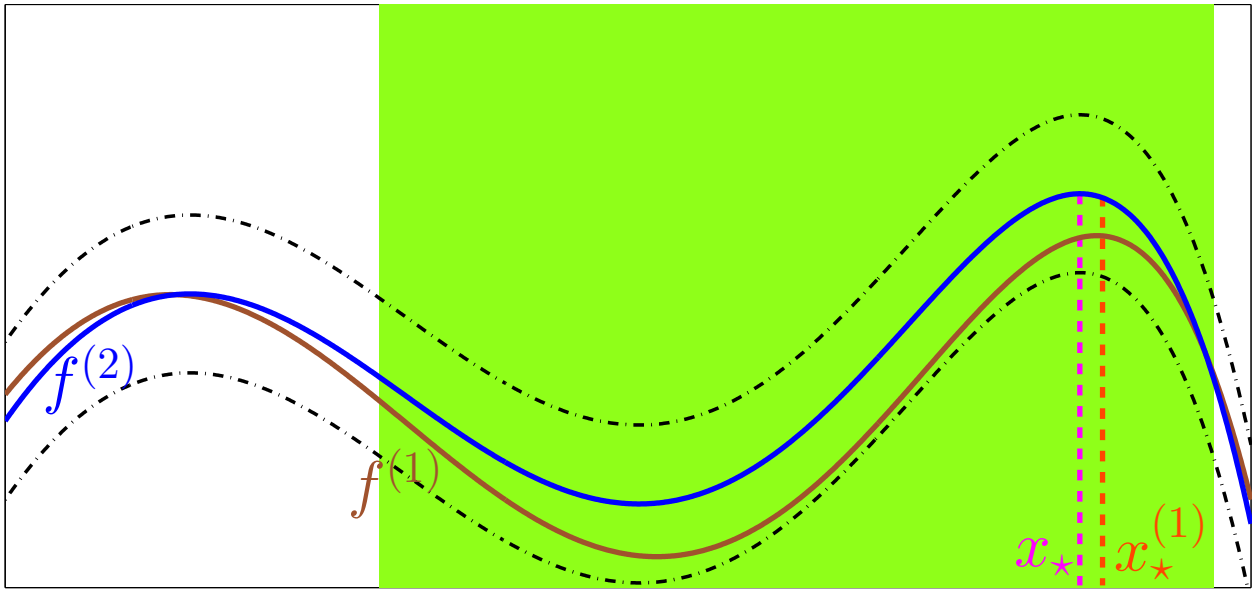
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges



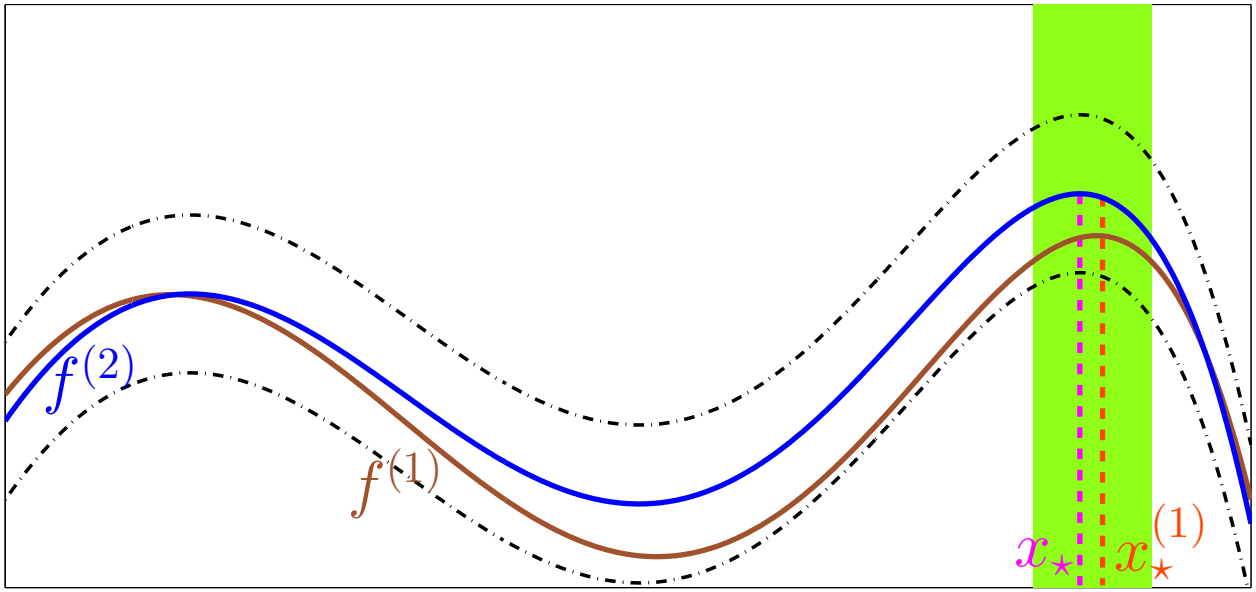
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges



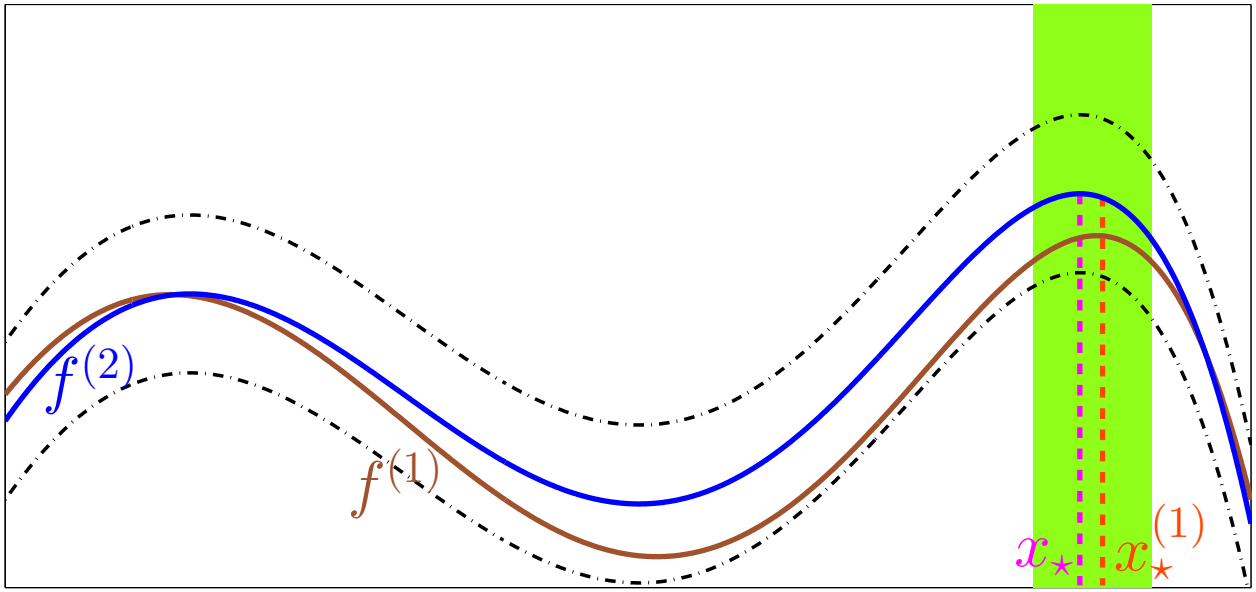
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges



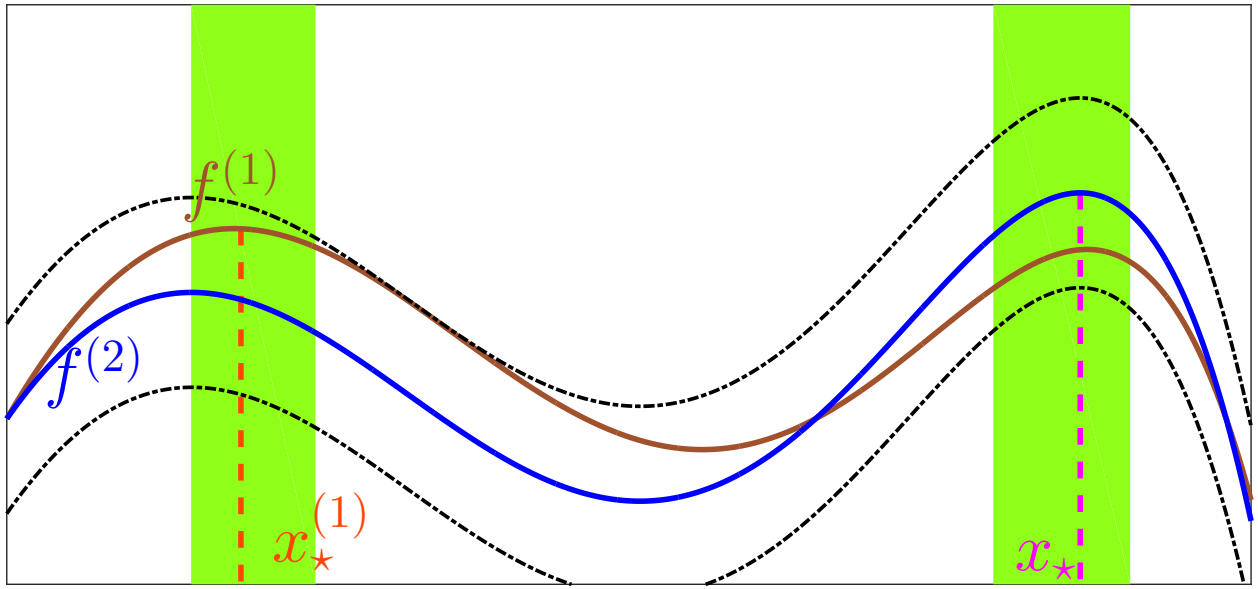
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges



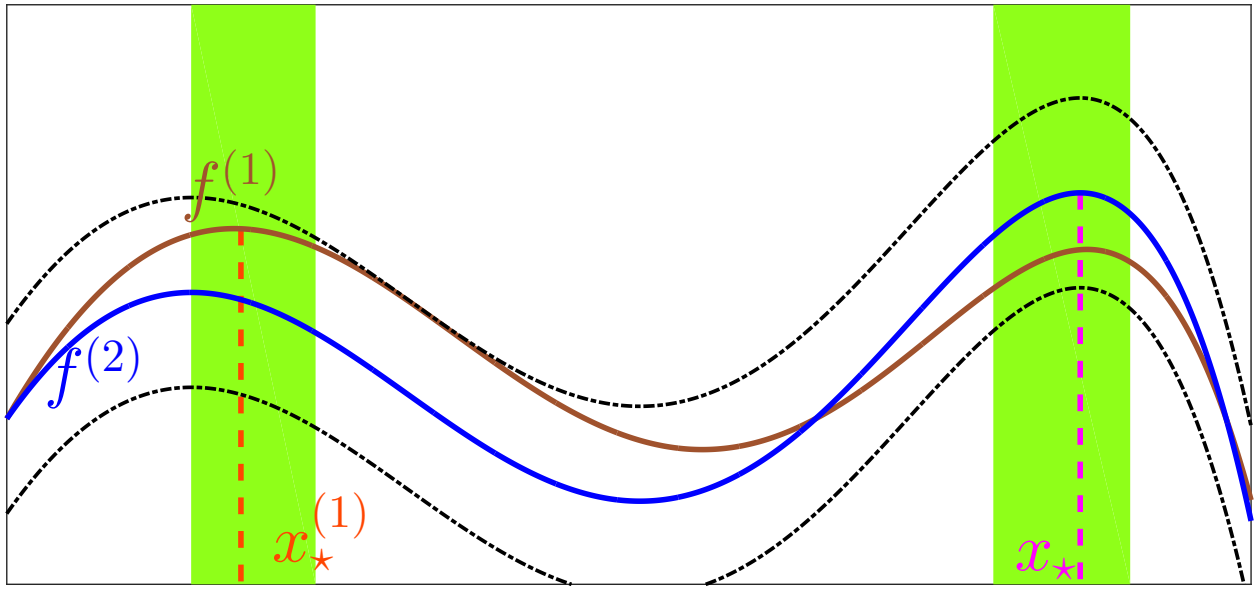
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.
- Need to explore $f^{(2)}$ sufficiently well around the *high valued regions* of $f^{(1)}$ – but at a not too large region.

Challenges



- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.
- Need to explore $f^{(2)}$ sufficiently well around the *high valued regions* of $f^{(1)}$ – but at a not too large region.

Challenges

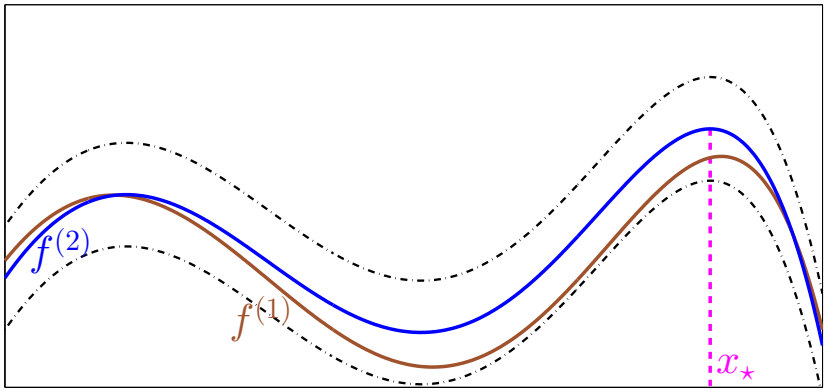


- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.
- Need to explore $f^{(2)}$ sufficiently well around the *high valued regions* of $f^{(1)}$ – but at a not too large region.

Key Message: We will explore $\mathcal{X}$ using $f^{(1)}$ and use $f^{(2)}$ *mostly* in a promising region $\mathcal{X}_\alpha$.

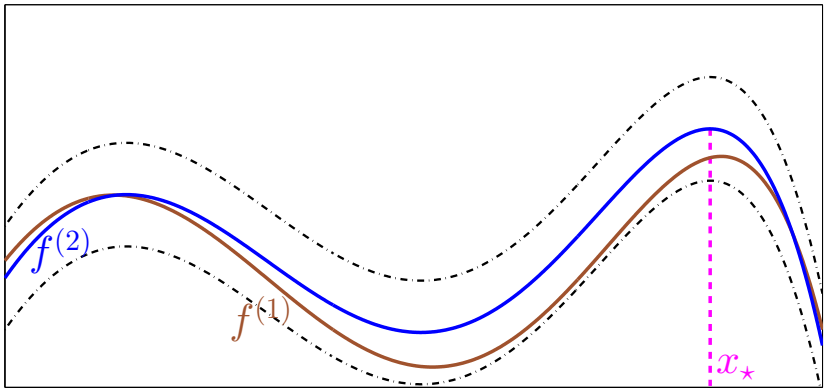
MF-GP-UCB (Kandasamy et al. NIPS 2016b)

Multi-fidelity Gaussian Process Upper Confidence Bound



MF-GP-UCB (Kandasamy et al. NIPS 2016b)

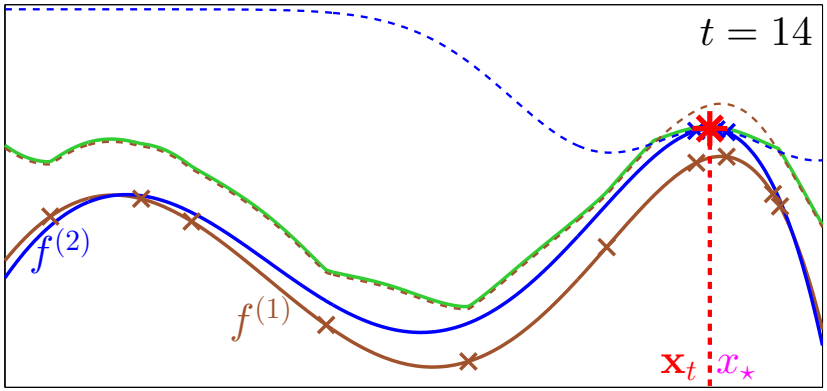
Multi-fidelity Gaussian Process Upper Confidence Bound



- Construct Upper Confidence Bound φ_t for $f^{(2)}$.
  Choose point $x_t = \mathrm{argmax}_{x \in \mathcal{X}} \varphi_t(x)$.

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

Multi-fidelity Gaussian Process Upper Confidence Bound



- Construct Upper Confidence Bound φ_t for $f^{(2)}$.

  Choose point $x_t = \mathrm{argmax}_{x \in \mathcal{X}} \varphi_t(x)$.

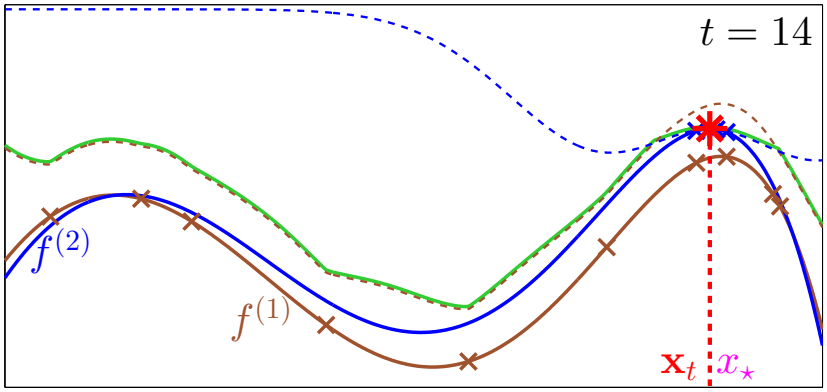
$$\varphi_t^{(1)}(x) = \boxed{\mu_{t-1}^{(1)}(x) + \beta_t^{1/2} \sigma_{t-1}^{(1)}(x)} + \zeta^{(1)}$$

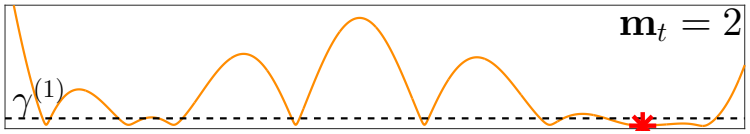
$$\varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2} \sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\, \varphi_t^{(1)}(x),\ \varphi_t^{(2)}(x) \,\}$$

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

Multi-fidelity Gaussian Process Upper Confidence Bound





- Construct Upper Confidence Bound φ_t for $f^{(2)}$.

  Choose point $x_t = \mathrm{argmax}_{x \in \mathcal{X}} \varphi_t(x)$.

$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2} \sigma_{t-1}^{(1)}(x) + \zeta^{(1)}$$

$$\varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2} \sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{ \varphi_t^{(1)}(x),\ \varphi_t^{(2)}(x) \}$$

- Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2} \sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{otherwise.} \end{cases}$

Theoretical Results for MF-GP-UCB

GP-UCB (Srinivas et al. 2010)

w.h.p $$S(\Lambda) = f^{(2)}(x_\star) - \max_{t\,:\,m_t=2} f^{(2)}(x_t) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

$\Psi_{n_\Lambda}(A)$ = Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

Theoretical Results for MF-GP-UCB

GP-UCB (Srinivas et al. 2010)

w.h.p $$S(\Lambda) = f^{(2)}(x_\star) - \max_{t\,:\,m_t=2} f^{(2)}(x_t) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

$\Psi_{n_\Lambda}(A)$ = Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

Theoretical Results for MF-GP-UCB

GP-UCB (Srinivas et al. 2010)

$$\text{w.h.p} \quad S(\Lambda) = f^{(2)}(x_\star) - \max_{t\,:\,m_t=2} f^{(2)}(x_t) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

$\Psi_{n_\Lambda}(A)$ = Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

$$\text{w.h.p} \quad \forall \alpha > 0, \quad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \leq C_\alpha \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Proof Sketches

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

w.h.p $$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \lesssim \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Proof Sketches

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

w.h.p $$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \lesssim \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Number of (random) queries after capital Λ $\leftarrow N$,

$$n_\Lambda = \frac{\Lambda}{\lambda^{(2)}} \leq N \leq \frac{\Lambda}{\lambda^{(1)}}.$$

Proof Sketches

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

w.h.p $$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \lesssim \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Number of (random) queries after capital Λ $\leftarrow N$,

$$n_\Lambda = \frac{\Lambda}{\lambda^{(2)}} \leq N \leq \frac{\Lambda}{\lambda^{(1)}}.$$

But we show $N \in \mathcal{O}(n_\Lambda)$.

Proof Sketches

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

w.h.p $$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \lesssim \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Number of (random) queries after capital Λ $\leftarrow N$,

$$n_\Lambda = \frac{\Lambda}{\lambda^{(2)}} \leq N \leq \frac{\Lambda}{\lambda^{(1)}}.$$

But we show $N \in \mathcal{O}(n_\Lambda)$.

$$N = T_N^{(1)}(\mathcal{X}_\alpha) + T_N^{(1)}(\mathcal{X}_\alpha^c) + T_N^{(2)}(\mathcal{X}_\alpha) + T_N^{(2)}(\mathcal{X}_\alpha^c)$$

Proof Sketches

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

MF-GP-UCB (Kandasamy et al. NIPS 2016b)

w.h.p $$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha^c)}{n_\Lambda^{2-\alpha}}}$$

$\mathcal{X}_\alpha = \{x : f^{(2)}(x_\star) - f^{(1)}(x) \lesssim \zeta^{(1)}\}$.

Good approximation $\implies \mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X}) \implies \Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Number of (random) queries after capital Λ $\leftarrow N$,

$$n_\Lambda = \frac{\Lambda}{\lambda^{(2)}} \leq N \leq \frac{\Lambda}{\lambda^{(1)}}.$$

But we show $N \in \mathcal{O}(n_\Lambda)$.

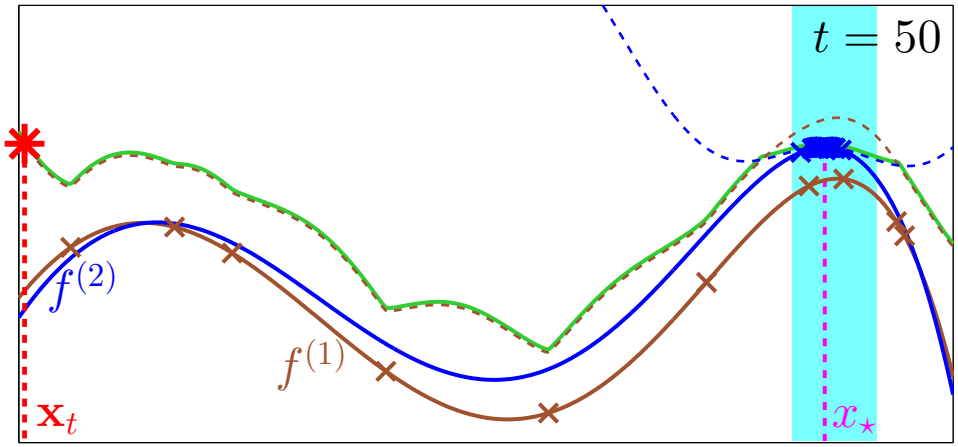
$$N = \underbrace{T_N^{(1)}(\mathcal{X}_\alpha)}_{\mathrm{polylog}(N)} + \underbrace{T_N^{(1)}(\mathcal{X}_\alpha^c)}_{\mathrm{sublinear}(N)} + T_N^{(2)}(\mathcal{X}_\alpha) + \underbrace{T_N^{(2)}(\mathcal{X}_\alpha^c)}_{N^\alpha}$$

$T_N^{(2)}(\mathcal{X}_\alpha^c) \leq N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \leq C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.

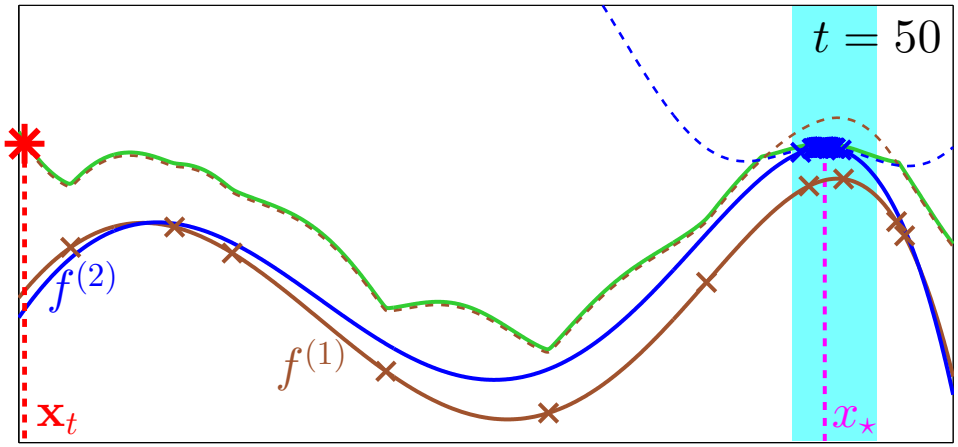


$T_N^{(2)}(\mathcal{X}_\alpha^c) \leq N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \leq C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.



$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}, \qquad \varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x), \varphi_t^{(2)}(x)\}, \qquad x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \varphi_t(x) \quad \rightarrow [1].$$

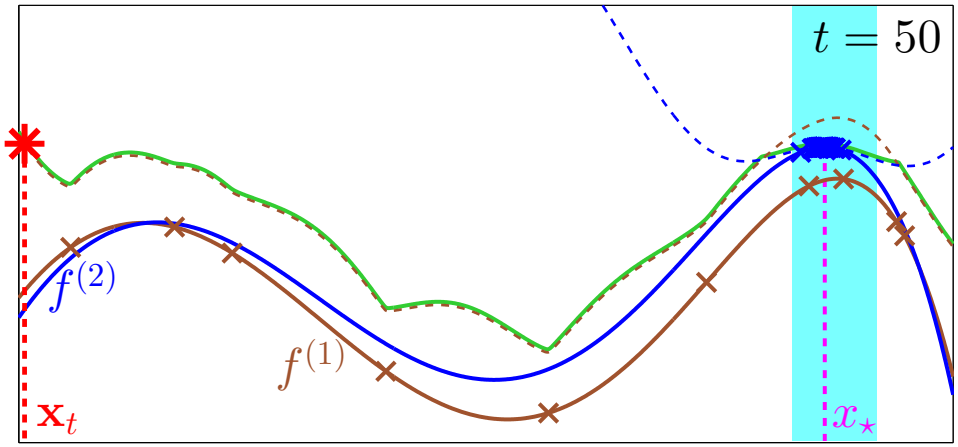
Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \leq \gamma^{(1)} \end{cases}$ $\rightarrow [2]$.

$T_N^{(2)}(\mathcal{X}_\alpha^c) \leq N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \leq C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.



$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}, \qquad \varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x), \varphi_t^{(2)}(x)\}, \qquad x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \varphi_t(x) \quad \rightarrow [1].$$

Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \leq \gamma^{(1)} \end{cases}$ $\rightarrow [2]$.

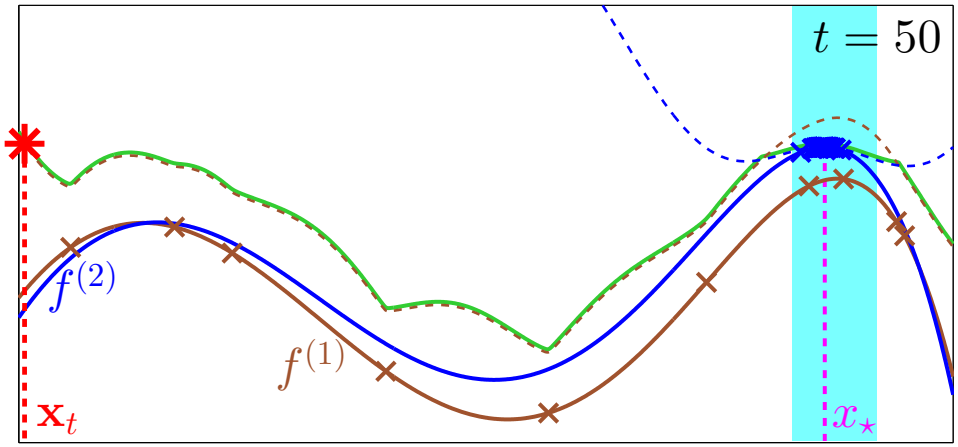
Argument: If $x_t \in \mathcal{X}_\alpha^c$ in [1], then $m_t = 2$ is unlikely in [2].

$T_N^{(2)}(\mathcal{X}_\alpha^c) \le N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \le C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.



$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}, \qquad \varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x), \varphi_t^{(2)}(x)\}, \qquad x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \varphi_t(x) \quad \to [1].$$

Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \le \gamma^{(1)} \end{cases} \quad \to [2].$

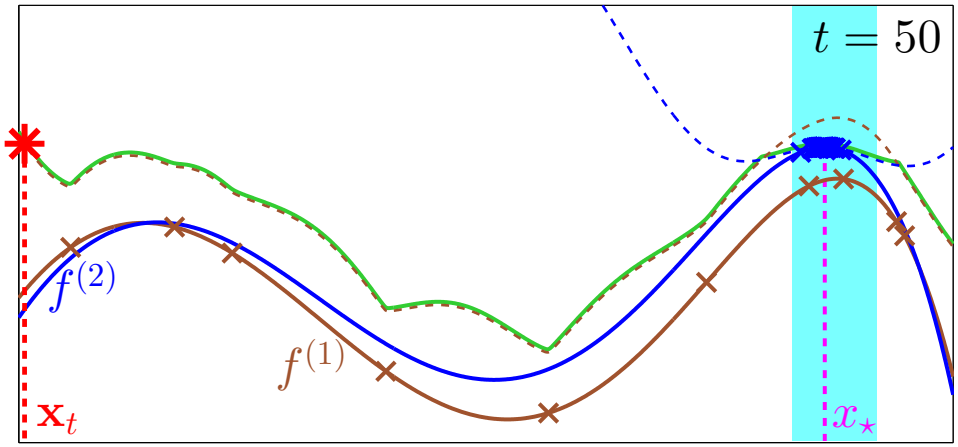
Argument: If $x_t \in \mathcal{X}_\alpha^c$ in [1], then $m_t = 2$ is unlikely in [2].
$m_t = 2 \implies \sigma_{t-1}^{(1)}(x_t)$ is small $\implies$ Several $f^{(1)}$ queries near x_t

$T_N^{(2)}(\mathcal{X}_\alpha^c) \leq N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \leq C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.



$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}, \qquad \varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x), \varphi_t^{(2)}(x)\}, \qquad x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \varphi_t(x) \quad \to [1].$$

Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \leq \gamma^{(1)} \end{cases} \quad \to [2].$

Argument: If $x_t \in \mathcal{X}_\alpha^c$ in [1], then $m_t = 2$ is unlikely in [2].
$m_t = 2 \implies \sigma_{t-1}^{(1)}(x_t)$ is small $\implies$ Several $f^{(1)}$ queries near $x_t \implies$

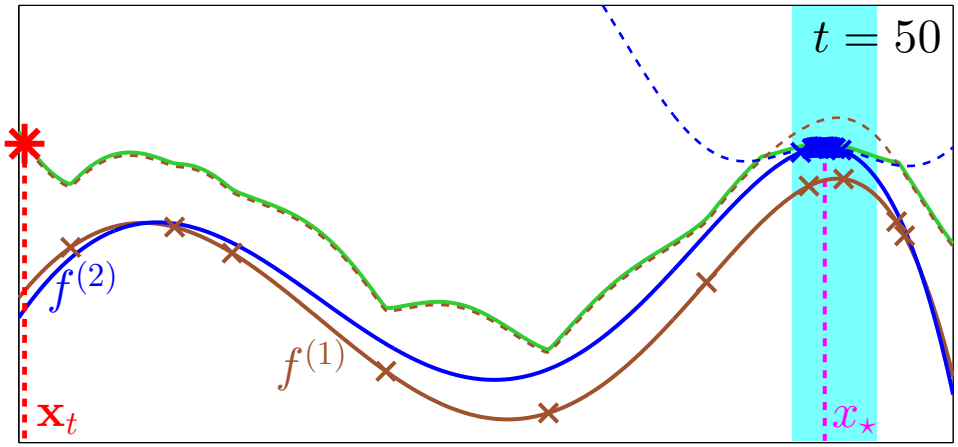
$\mu_{t-1}^{(1)}(x_t) \approx f^{(1)}(x_t) \implies \varphi_t^{(1)}(x_t)$ is small $\implies$

$T_N^{(2)}(\mathcal{X}_\alpha^c) \le N^\alpha$ for all $\alpha > 0$

$\lambda^{(2)}$ expensive $>$ $\lambda^{(1)}$ cheap

For $x \in \mathcal{X}_\alpha$,
$f^{(2)}(x_\star) - f^{(1)}(x) \le C_\alpha \zeta^{(1)}$.

$f^{(1)}$ is small in $\mathcal{X}_\alpha^c$.



$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}, \qquad \varphi_t^{(2)}(x) = \mu_{t-1}^{(2)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(2)}(x)$$

$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x), \varphi_t^{(2)}(x)\}, \qquad x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \varphi_t(x) \quad \to [1].$$

Choose fidelity $m_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) > \gamma^{(1)} \\ 2 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x_t) \le \gamma^{(1)} \end{cases} \quad \to [2].$

Argument: If $x_t \in \mathcal{X}_\alpha^c$ in [1], then $m_t = 2$ is unlikely in [2].
$m_t = 2 \implies \sigma_{t-1}^{(1)}(x_t)$ is small $\implies$ Several $f^{(1)}$ queries near $x_t \implies$

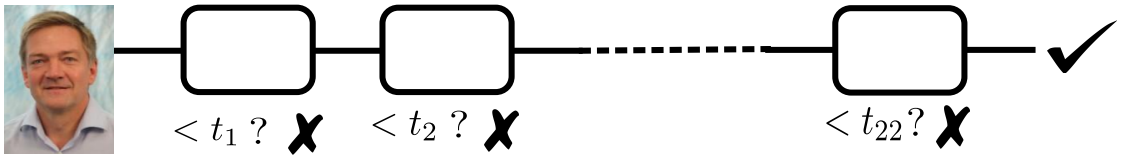
$\mu_{t-1}^{(1)}(x_t) \approx f^{(1)}(x_t) \implies \varphi_t^{(1)}(x_t)$ is small $\implies x_t$ won't be arg-max.

MF-GP-UCB with multiple approximations

MF-GP-UCB with multiple approximations

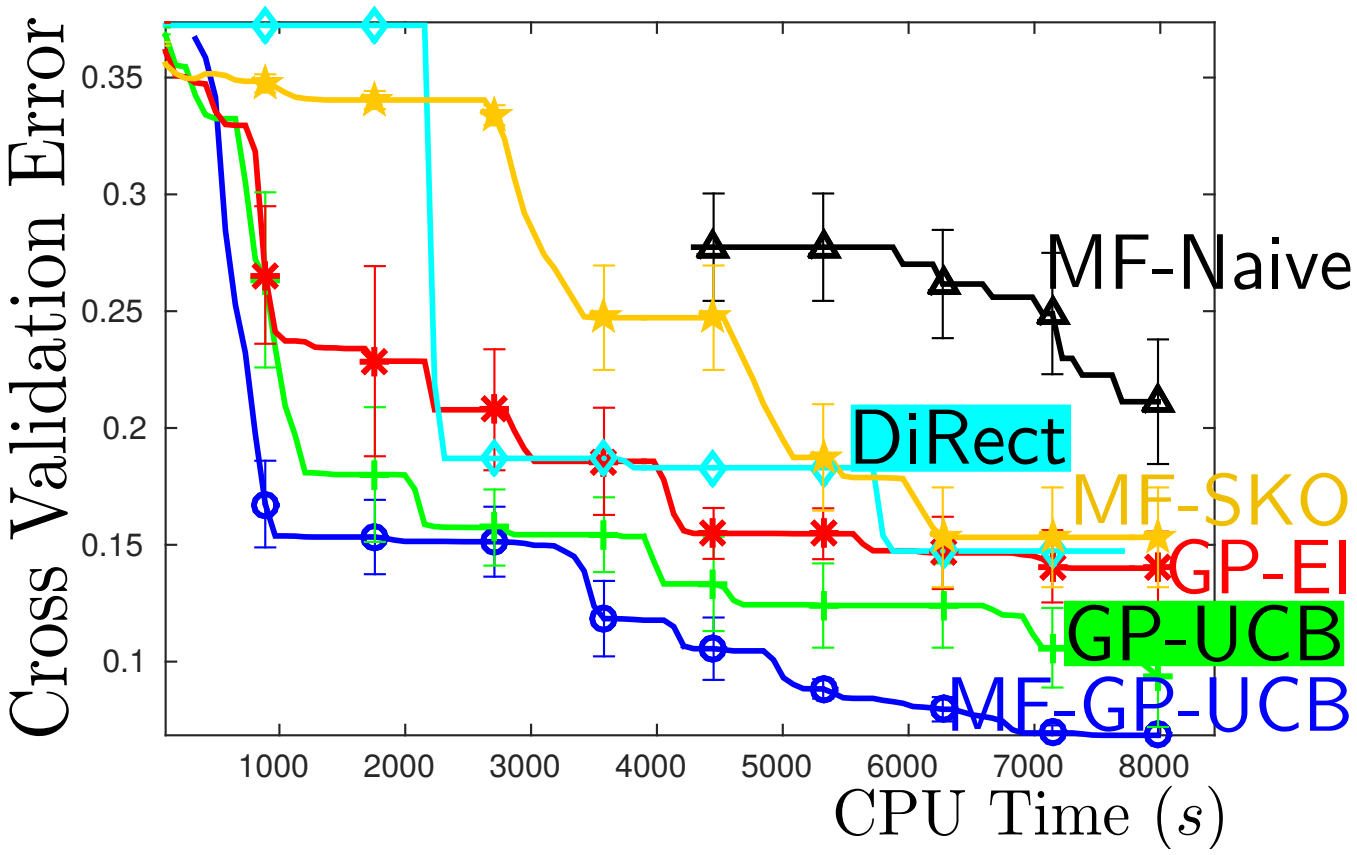
Things work out.

Experiment: Viola & Jones Face Detection



22 Threshold values for each cascade. $(d = 22)$

Fidelities with dataset sizes $(300, 3000)$. $(M = 2)$



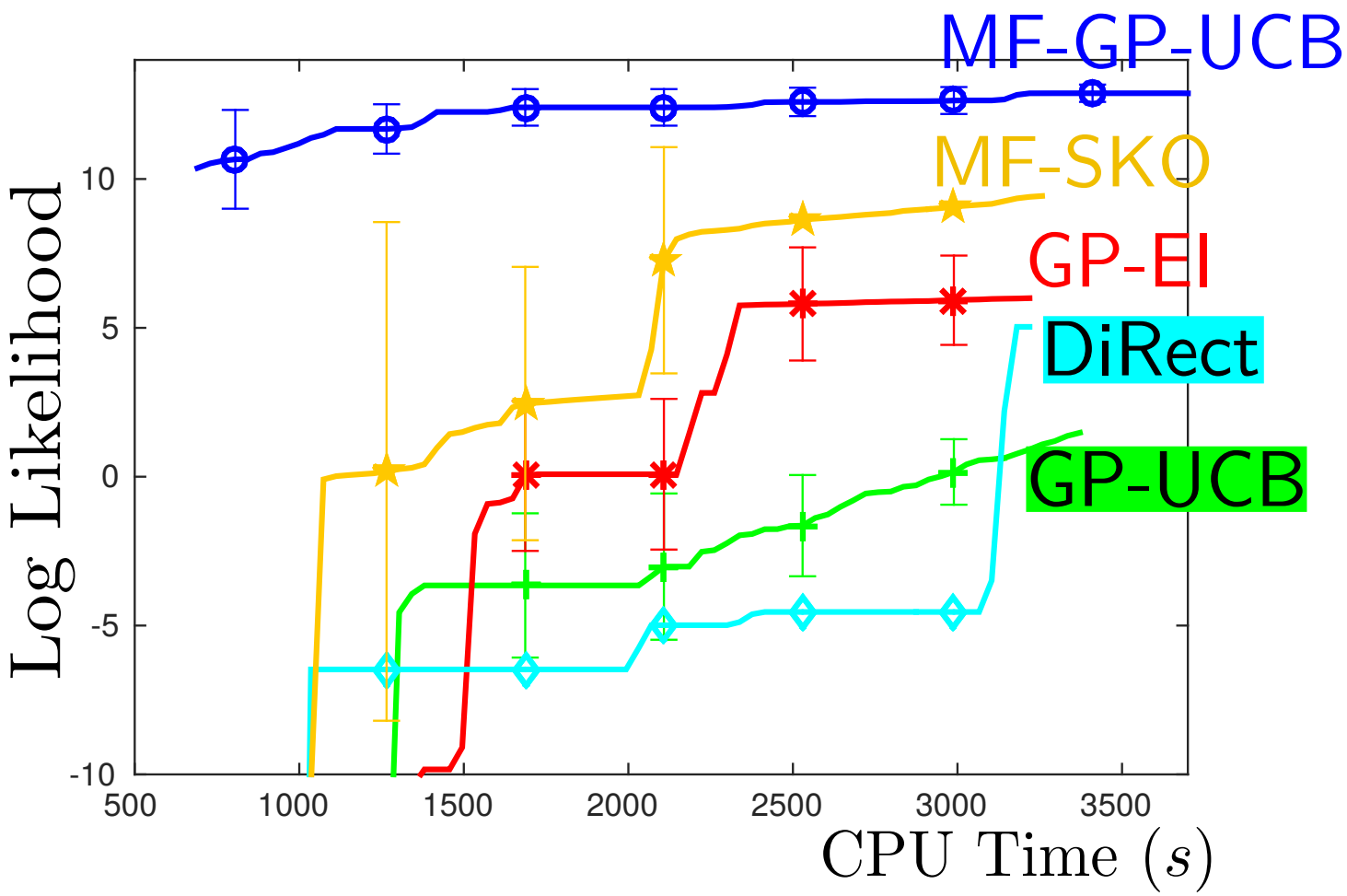
Experiment: Cosmological Maximum Likelihood Inference

- Type Ia Supernovae Data
- Maximum likelihood inference for 3 cosmological parameters:
  - Hubble Constant H_0
  - Dark Energy Fraction Ω_Λ
  - Dark Matter Fraction Ω_M
- Likelihood: Robertson Walker metric (Robertson 1936)
  Requires numerical integration for each point in the dataset.

Experiment: Cosmological Maximum Likelihood Inference

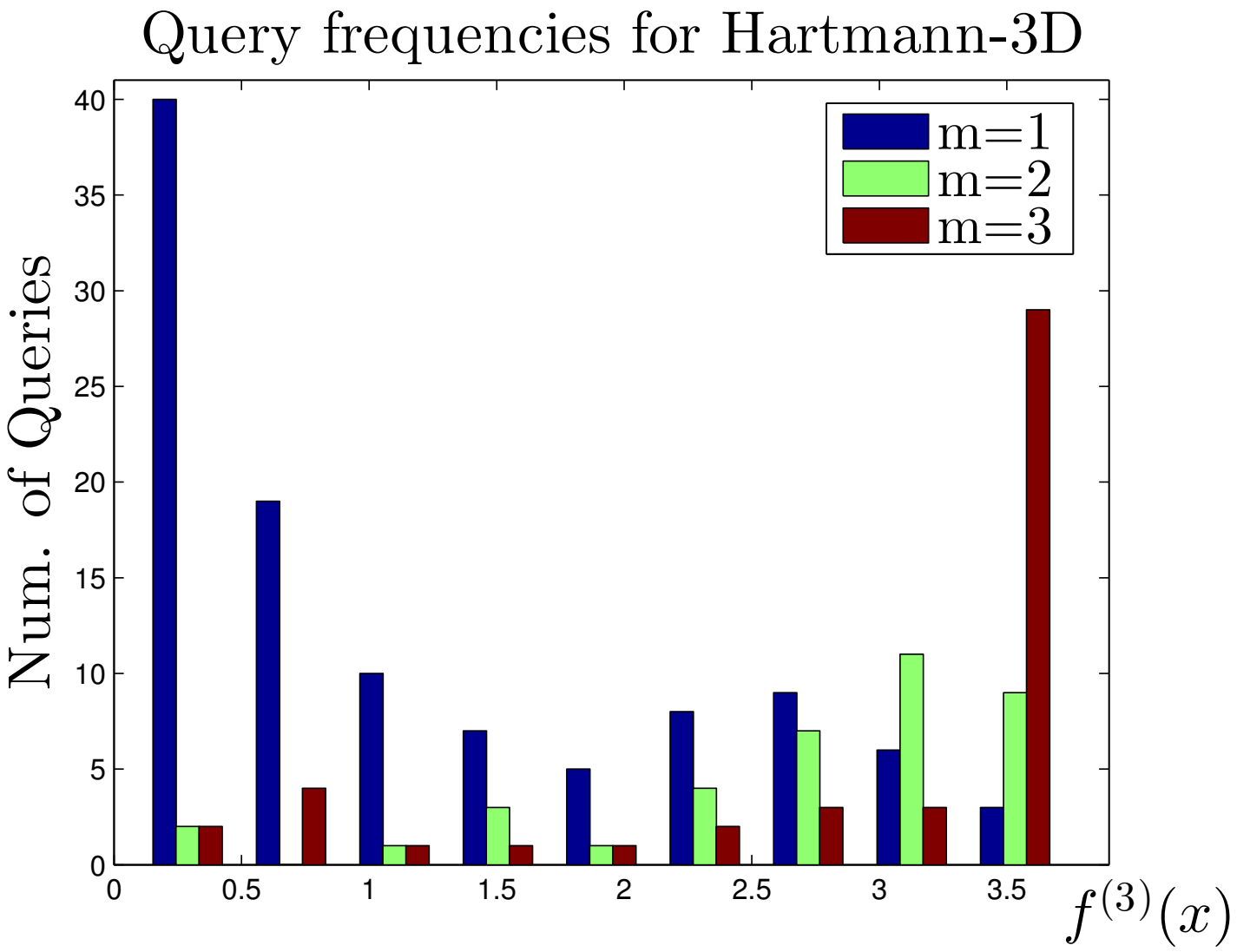
3 cosmological parameters. $(d = 3)$

Fidelities: integration on grids of size $(10^2, 10^4, 10^6)$. $(M = 3)$

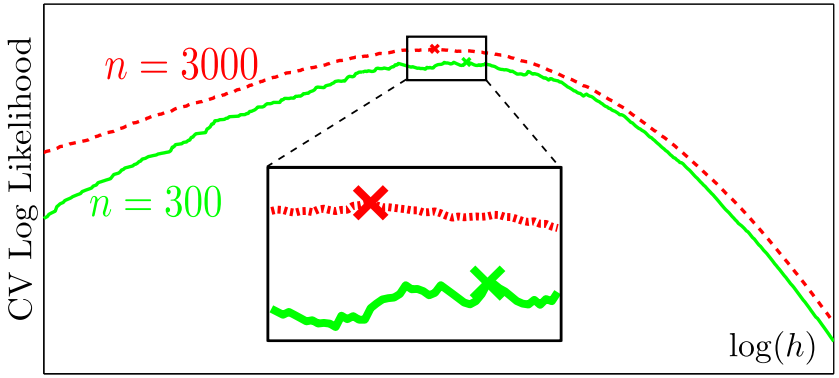


MF-GP-UCB Synthetic Experiment: Hartmann-3D

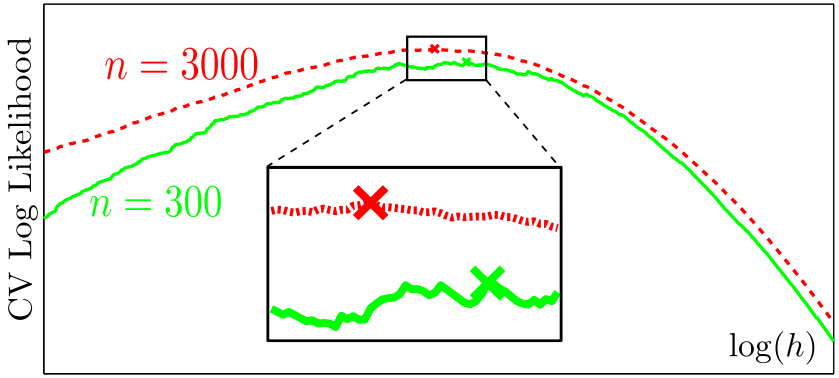
$d = 3, \quad M = 3$



Multi-fidelity Optimisation with Continuous Approximations

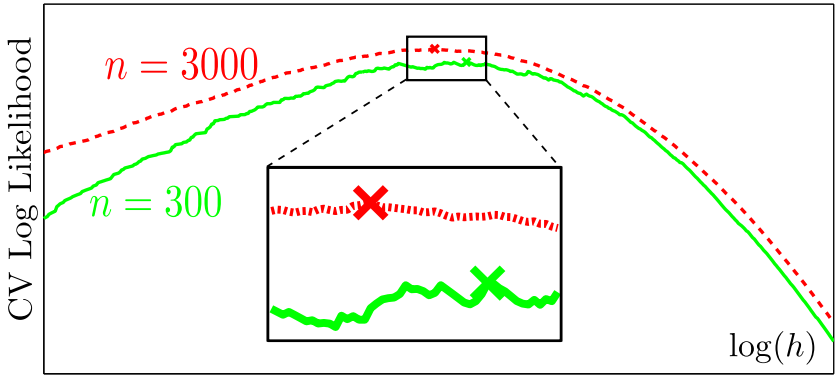


Multi-fidelity Optimisation with Continuous Approximations



- Use an arbitrary amount of data?
- Iterative algorithms: use arbitrary number of iterations?

Multi-fidelity Optimisation with Continuous Approximations

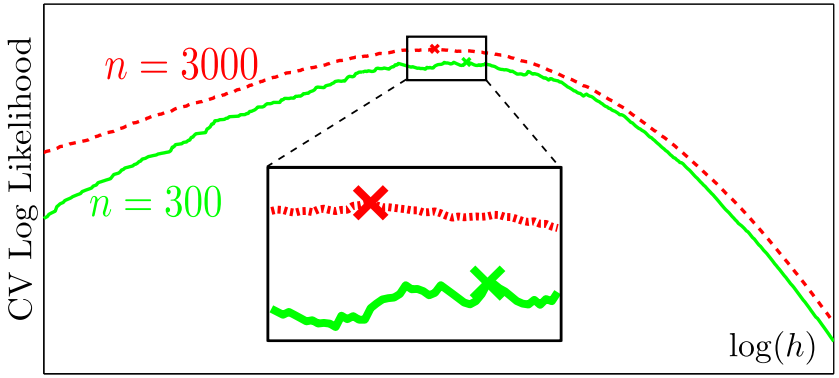


- Use an arbitrary amount of data?

- Iterative algorithms: use arbitrary number of iterations?

E.g. Train an ML model with $N_\bullet$ data and $T_\bullet$ iterations.

Multi-fidelity Optimisation with Continuous Approximations



- Use an arbitrary amount of data?

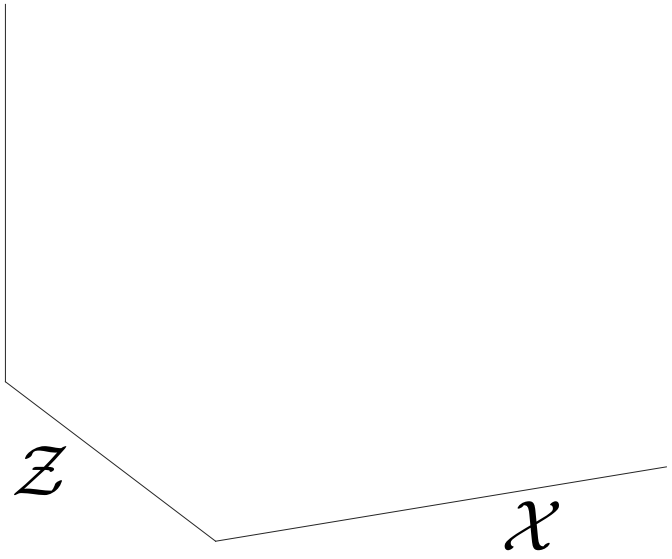
- Iterative algorithms: use arbitrary number of iterations?

E.g. Train an ML model with $N_\bullet$ data and $T_\bullet$ iterations.
But use $N < N_\bullet$ data and $T < T_\bullet$ iterations to approximate cross validation performance.

Approximations from a *continuous* 2D "fidelity space" (N, T).

Multi-fidelity Optimisation with Continuous Approximations

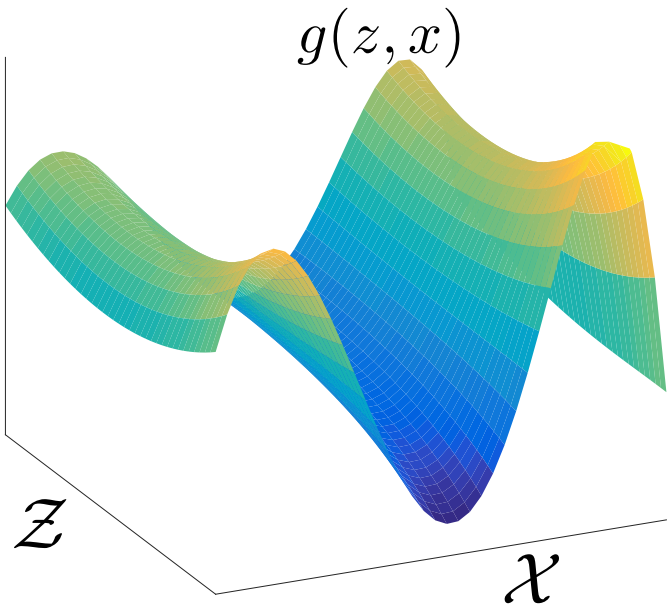
(Kandasamy et al. Arxiv 2017)



A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$.

Multi-fidelity Optimisation with Continuous Approximations

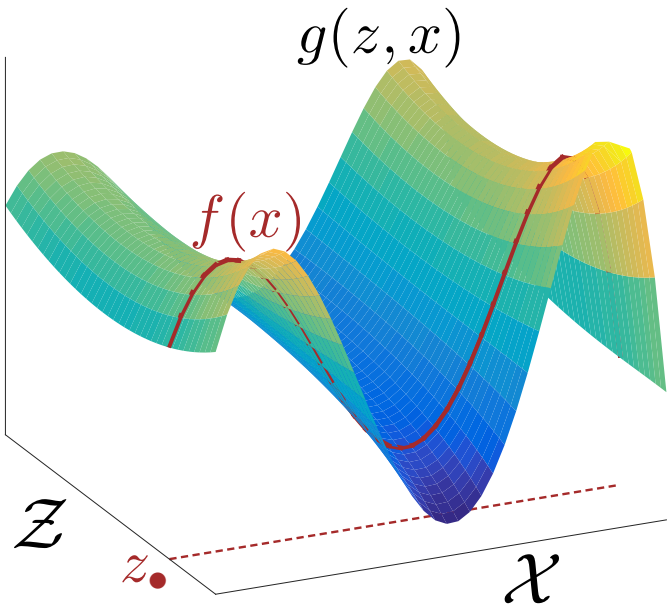
(Kandasamy et al. Arxiv 2017)



A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

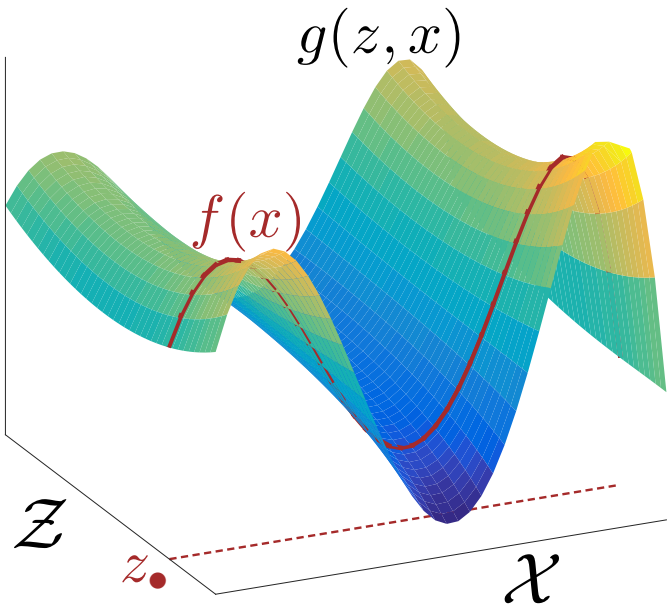


A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

We wish to optimise $f(x) = g(z_\bullet, x)$ where $z_\bullet \in \mathcal{Z}$.

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)



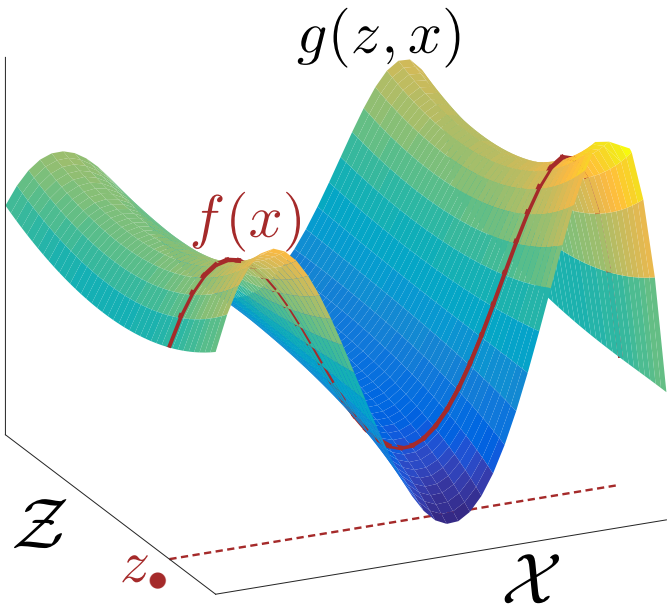
A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

We wish to optimise $f(x) = g(z_\bullet, x)$ where $z_\bullet \in \mathcal{Z}$.

previous e.g.: $\quad \mathcal{Z} =$ all (N, T) values, $\quad z_\bullet = [N_\bullet, T_\bullet]$.

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)



A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

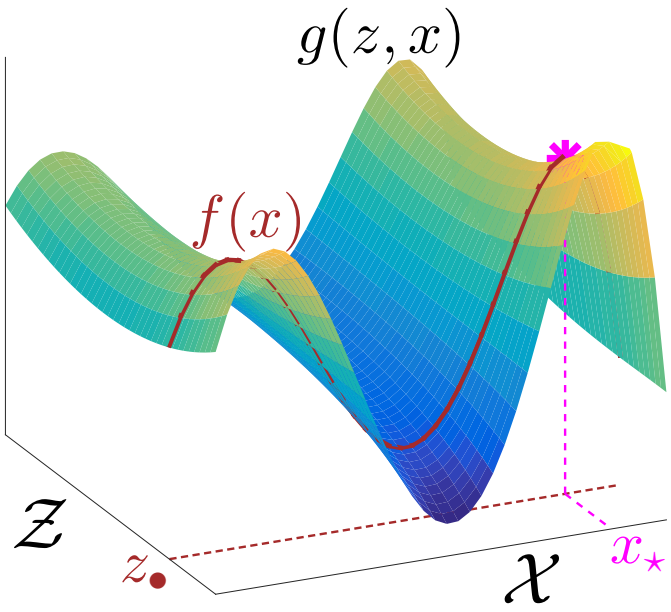
We wish to optimise $f(x) = g(z_\bullet, x)$ where $z_\bullet \in \mathcal{Z}$.

previous e.g.: $\quad \mathcal{Z} =$ all (N, T) values, $\quad z_\bullet = [N_\bullet, T_\bullet]$.

A cost function, $\lambda : \mathcal{Z} \to \mathbb{R}_+$. $\qquad$ e.g.: $\lambda(z) = \lambda(N, T) = \mathcal{O}(N^2 T)$

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)



A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

We wish to optimise $f(x) = g(z_\bullet, x)$ where $z_\bullet \in \mathcal{Z}$.

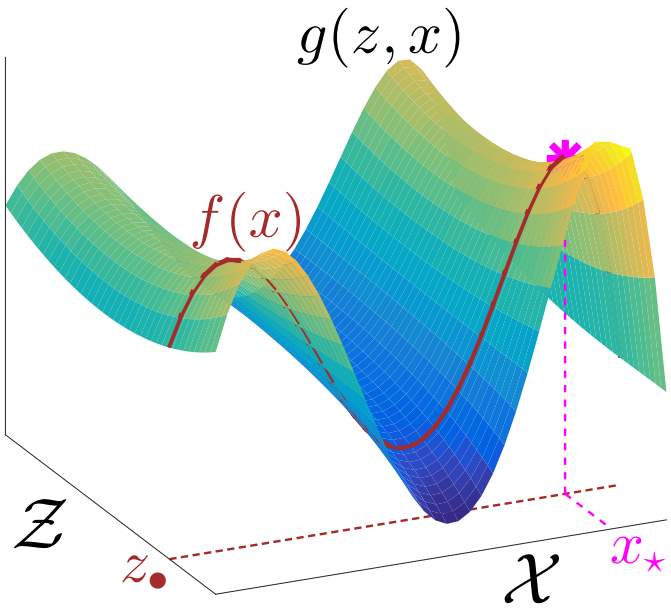
previous e.g.: $\quad \mathcal{Z} =$ all (N, T) values, $\quad z_\bullet = [N_\bullet, T_\bullet]$.

A cost function, $\lambda : \mathcal{Z} \to \mathbb{R}_+$. $\qquad$ e.g.: $\lambda(z) = \lambda(N, T) = \mathcal{O}(N^2 T)$

$x_\star = \mathrm{argmax}_x \, f(x)$.

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)



A fidelity space $\mathcal{Z} \subset \mathbb{R}^p$ and domain $\mathcal{X} \subset \mathbb{R}^d$. $\quad g : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$.

We wish to optimise $f(x) = g(z_\bullet, x)$ where $z_\bullet \in \mathcal{Z}$.

previous e.g.: $\quad \mathcal{Z} =$ all (N, T) values, $\quad z_\bullet = [N_\bullet, T_\bullet]$.

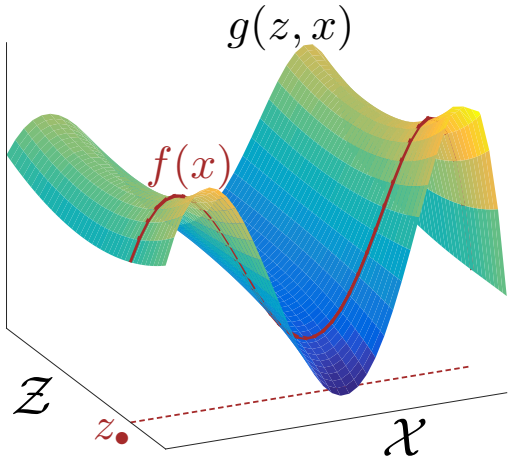
A cost function, $\lambda : \mathcal{Z} \to \mathbb{R}_+$. $\qquad$ e.g.: $\lambda(z) = \lambda(N, T) = \mathcal{O}(N^2 T)$

$x_\star = \mathrm{argmax}_x f(x)$. $\quad$ Simple Regret: $\quad S(\Lambda) = f(x_\star) - \max_{t:\, z_t = z_\bullet} f(x_t)$.

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

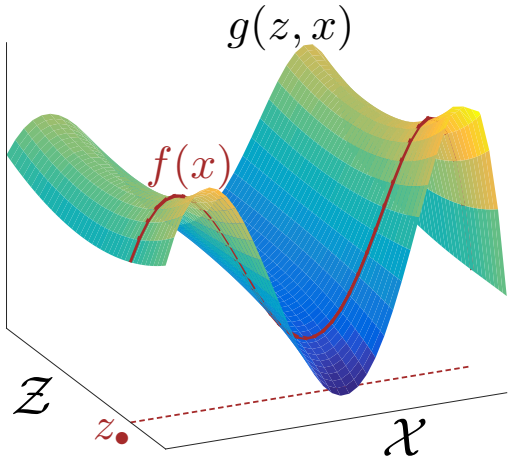
$$g \sim \mathcal{GP}(\mathbf{0}, \kappa),$$



Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

$$g \sim \mathcal{GP}(\mathbf{0}, \kappa), \quad \kappa : (\mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}.$$

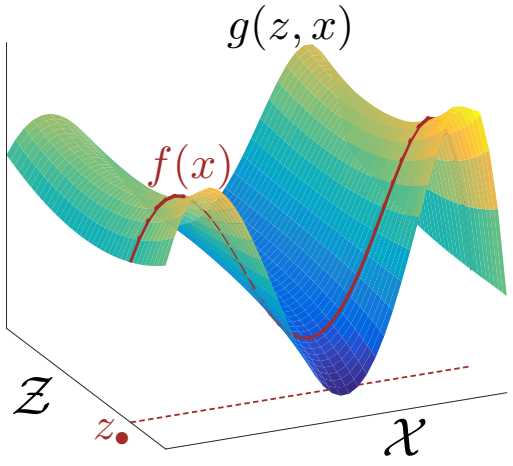


Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

$$g \sim \mathcal{GP}(\mathbf{0}, \kappa), \quad \kappa : (\mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}.$$

$$\kappa([z, x], [z', x']) = \kappa_{\mathcal{X}}(x, x') \cdot \kappa_{\mathcal{Z}}(z, z')$$

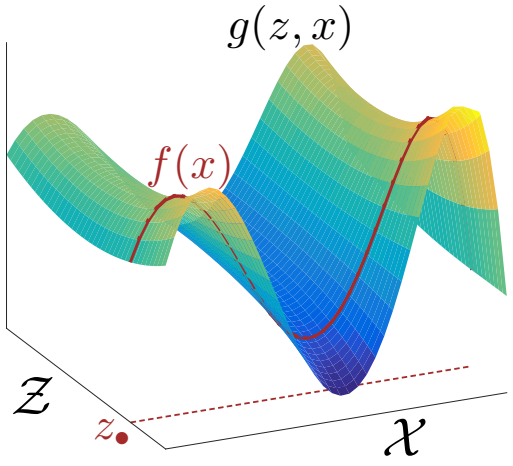


Multi-fidelity Optimisation with Continuous Approximations

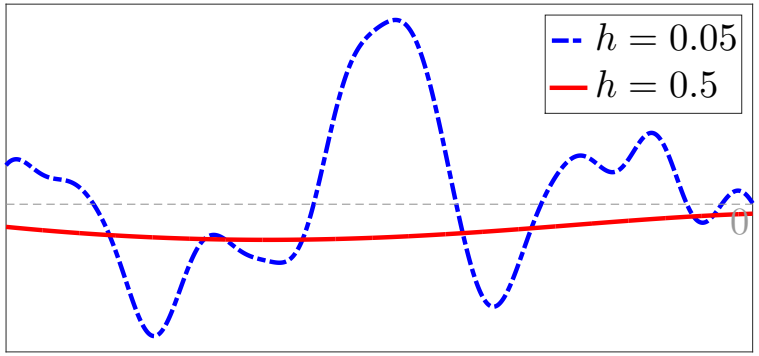
(Kandasamy et al. Arxiv 2017)

$$g \sim \mathcal{GP}(\mathbf{0}, \kappa), \quad \kappa : (\mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}.$$

$$\kappa([z, x], [z', x']) = \kappa_{\mathcal{X}}(x, x') \cdot \kappa_{\mathcal{Z}}(z, z')$$



SE kernel:

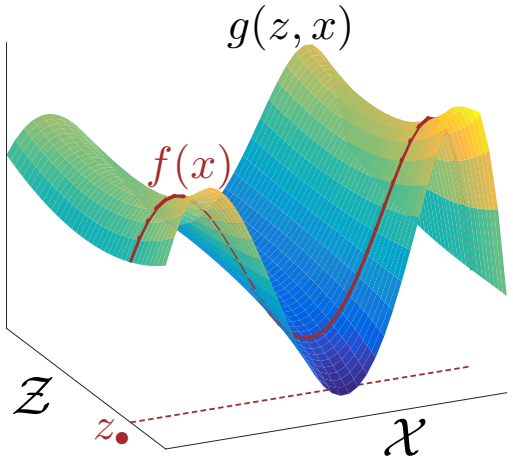


Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

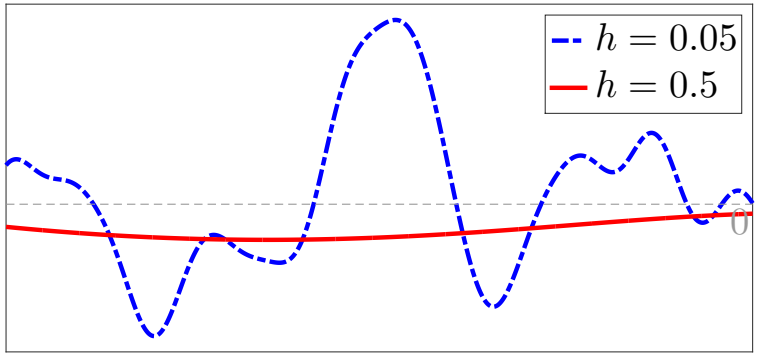
$$g \sim \mathcal{GP}(\mathbf{0}, \kappa), \quad \kappa : (\mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}.$$

$$\kappa([z, x], [z', x']) = \kappa_{\mathcal{X}}(x, x') \cdot \kappa_{\mathcal{Z}}(z, z')$$



Information Gap $\xi : \mathcal{Z} \to \mathbb{R}$

- measures the price (in information) for querying at $z \neq z_\bullet$.



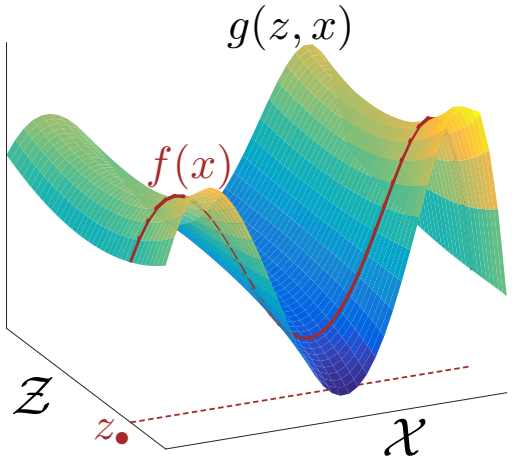
SE kernel:

Multi-fidelity Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

$$g \sim \mathcal{GP}(\mathbf{0}, \kappa), \quad \kappa : (\mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}.$$

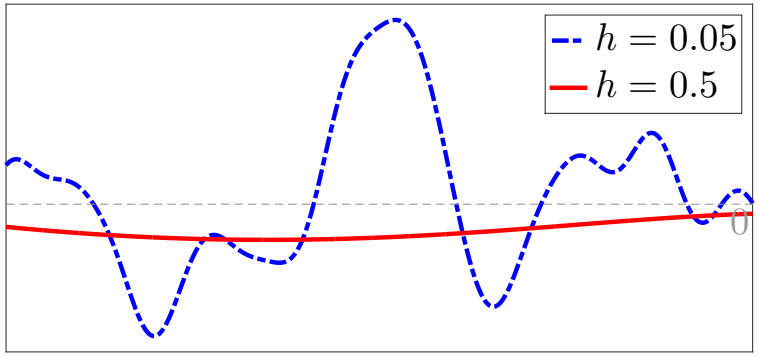
$$\kappa([z, x], [z', x']) = \kappa_{\mathcal{X}}(x, x') \cdot \kappa_{\mathcal{Z}}(z, z')$$



Information Gap $\xi : \mathcal{Z} \to \mathbb{R}$

- measures the price (in information) for querying at $z \neq z_\bullet$.

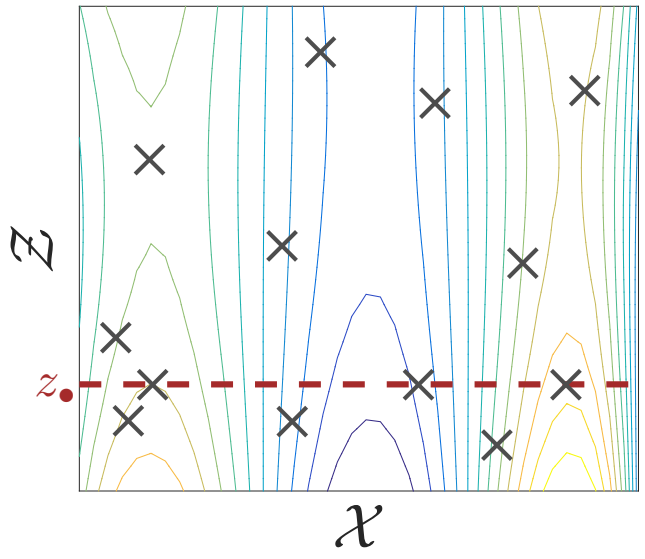
SE kernel: $\xi(z) \lesssim \frac{\|z - z_\bullet\|}{h}$.



BOCA: Bayesian Optimisation with Continuous Approximations

(Kandasamy et al. Arxiv 2017)

At time t we have $t-1$ previous evaluations. $\{(z_i, x_i, y_i)\}_{i=1}^{t-1}$.



BOCA: Bayesian Optimisation with Continuous Approximations

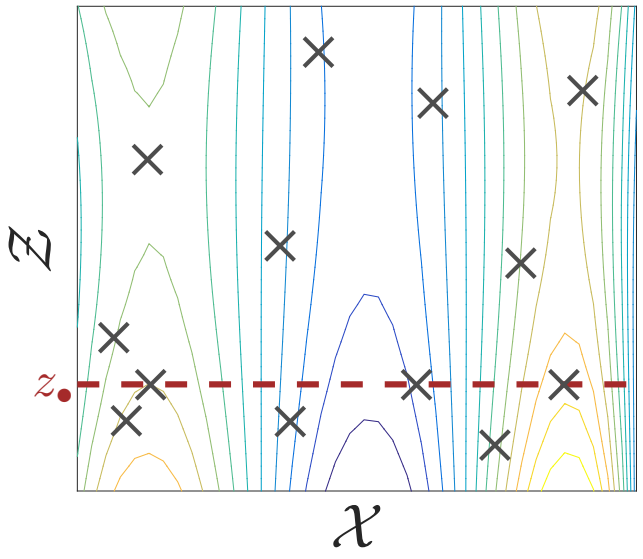
(Kandasamy et al. Arxiv 2017)

At time t we have $t-1$ previous evaluations. $\{(z_i, x_i, y_i)\}_{i=1}^{t-1}$.

Construct posterior $\mathcal{GP}$ for g:

mean $\mu_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$

std-dev $\sigma_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}_+$



BOCA: Bayesian Optimisation with Continuous Approximations

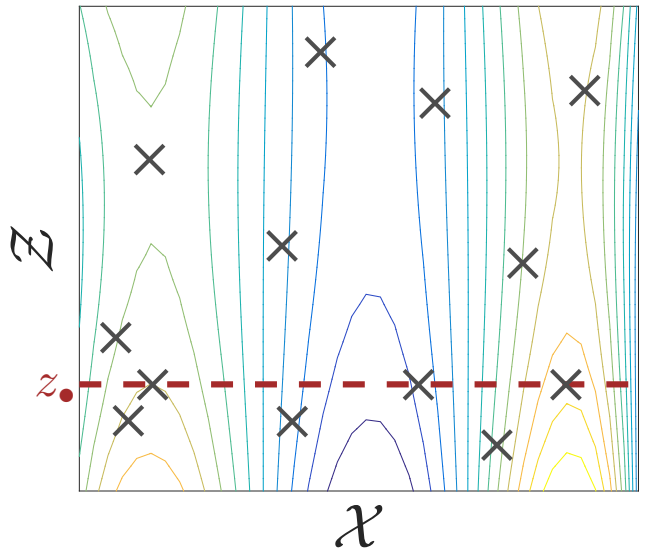
(Kandasamy et al. Arxiv 2017)

At time t we have $t-1$ previous evaluations. $\{(z_i, x_i, y_i)\}_{i=1}^{t-1}$.

Construct posterior $\mathcal{GP}$ for g:

mean $\mu_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$

std-dev $\sigma_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}_+$



$x_t \leftarrow$ maximise upper confidence bound for f.

$$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; \mu_{t-1}(z_\bullet, x) + \beta_t^{1/2} \sigma_{t-1}(z_\bullet, x)$$

BOCA: Bayesian Optimisation with Continuous Approximations

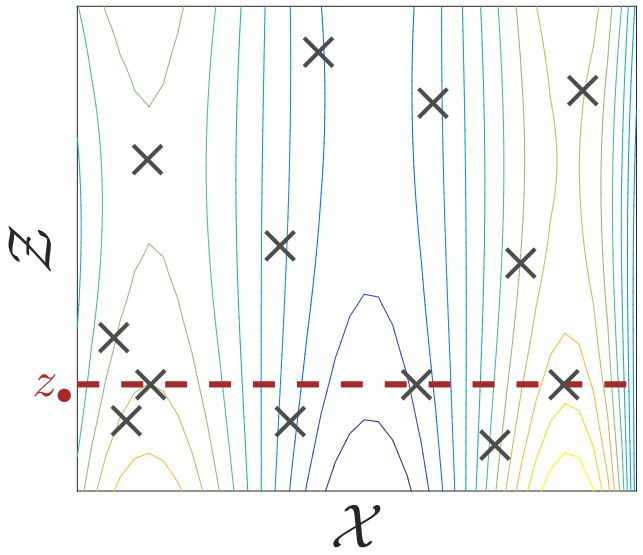
(Kandasamy et al. Arxiv 2017)

At time t we have $t-1$ previous evaluations. $\{(z_i, x_i, y_i)\}_{i=1}^{t-1}$.

Construct posterior $\mathcal{GP}$ for g:

mean $\quad \mu_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$

std-dev $\quad \sigma_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}_+$



$x_t \leftarrow$ maximise upper confidence bound for f.

$$x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \; \mu_{t-1}(z_\bullet, x) + \beta_t^{1/2} \sigma_{t-1}(z_\bullet, x)$$

$$\mathcal{Z}_t \approx \{z_\bullet\} \cup \left\{ z : \sigma_{t-1}(z, x_t) \geq \gamma(z) \right\}$$

$$z_t = \operatorname*{argmin}_{z \in \mathcal{Z}_t} \lambda(z) \qquad \text{(cheapest } z \text{ in } \mathcal{Z}_t\text{)}$$

BOCA: Bayesian Optimisation with Continuous Approximations

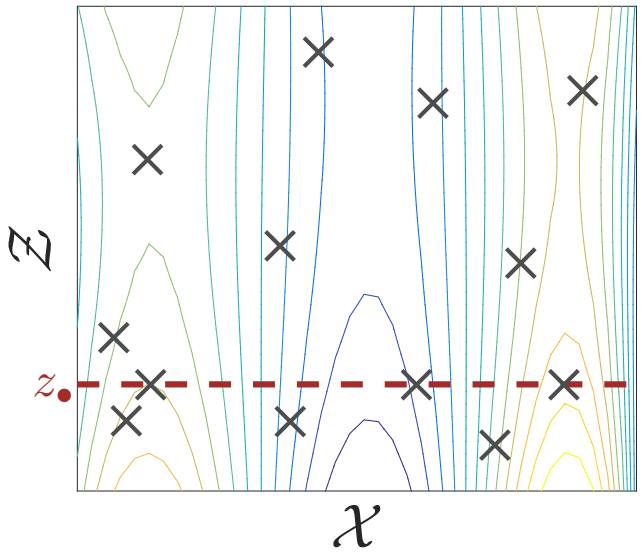
(Kandasamy et al. Arxiv 2017)

At time t we have $t-1$ previous evaluations. $\{(z_i, x_i, y_i)\}_{i=1}^{t-1}$.

Construct posterior $\mathcal{GP}$ for g:

mean $\quad \mu_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$

std-dev $\quad \sigma_{t-1} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}_+$



$x_t \leftarrow$ maximise upper confidence bound for f.

$$x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \; \mu_{t-1}(z_\bullet, x) + \beta_t^{1/2} \sigma_{t-1}(z_\bullet, x)$$

$$\mathcal{Z}_t \approx \{z_\bullet\} \cup \left\{ z : \sigma_{t-1}(z, x_t) \geq \gamma(z) = \left(\frac{\lambda(z)}{\lambda(z_\bullet)} \right)^q \xi(z) \right\}$$

$$z_t = \operatorname*{argmin}_{z \in \mathcal{Z}_t} \lambda(z) \qquad \text{(cheapest } z \text{ in } \mathcal{Z}_t\text{)}$$

Theoretical Results for BOCA

GP-UCB (Srinivas et al. 2010)

$$\text{w.h.p} \qquad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda / \lambda(z_\bullet) \rfloor$.

$\Psi_{n_\Lambda}(A) =$ Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

Theoretical Results for BOCA

GP-UCB (Srinivas et al. 2010)

$$\text{w.h.p} \qquad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda / \lambda(z_\bullet) \rfloor$.

$\Psi_{n_\Lambda}(A) =$ Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

BOCA (Kandasamy et al. Arxiv 2017)

$$\text{w.h.p} \quad \forall \alpha > 0, \quad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda^{2-\alpha}}}$$

$$\mathcal{X}_\alpha = \left\{ x; \;\; f(x_\star) - f(x) \lesssim C_\alpha \frac{1}{h} \right\}$$

Theoretical Results for BOCA

GP-UCB (Srinivas et al. 2010)

$$\text{w.h.p} \qquad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

$n_\Lambda = \lfloor \Lambda / \lambda(z_\bullet) \rfloor$.

$\Psi_{n_\Lambda}(A) =$ Maximum Information Gain $\rightarrow$ Scales with $\mathrm{vol}(A)$.

BOCA (Kandasamy et al. Arxiv 2017)

$$\text{w.h.p} \quad \forall \alpha > 0, \quad S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_\alpha)}{n_\Lambda}} + \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda^{2-\alpha}}}$$

$$\mathcal{X}_\alpha = \left\{ x; \;\; f(x_\star) - f(x) \lesssim C_\alpha \frac{1}{h} \right\}$$

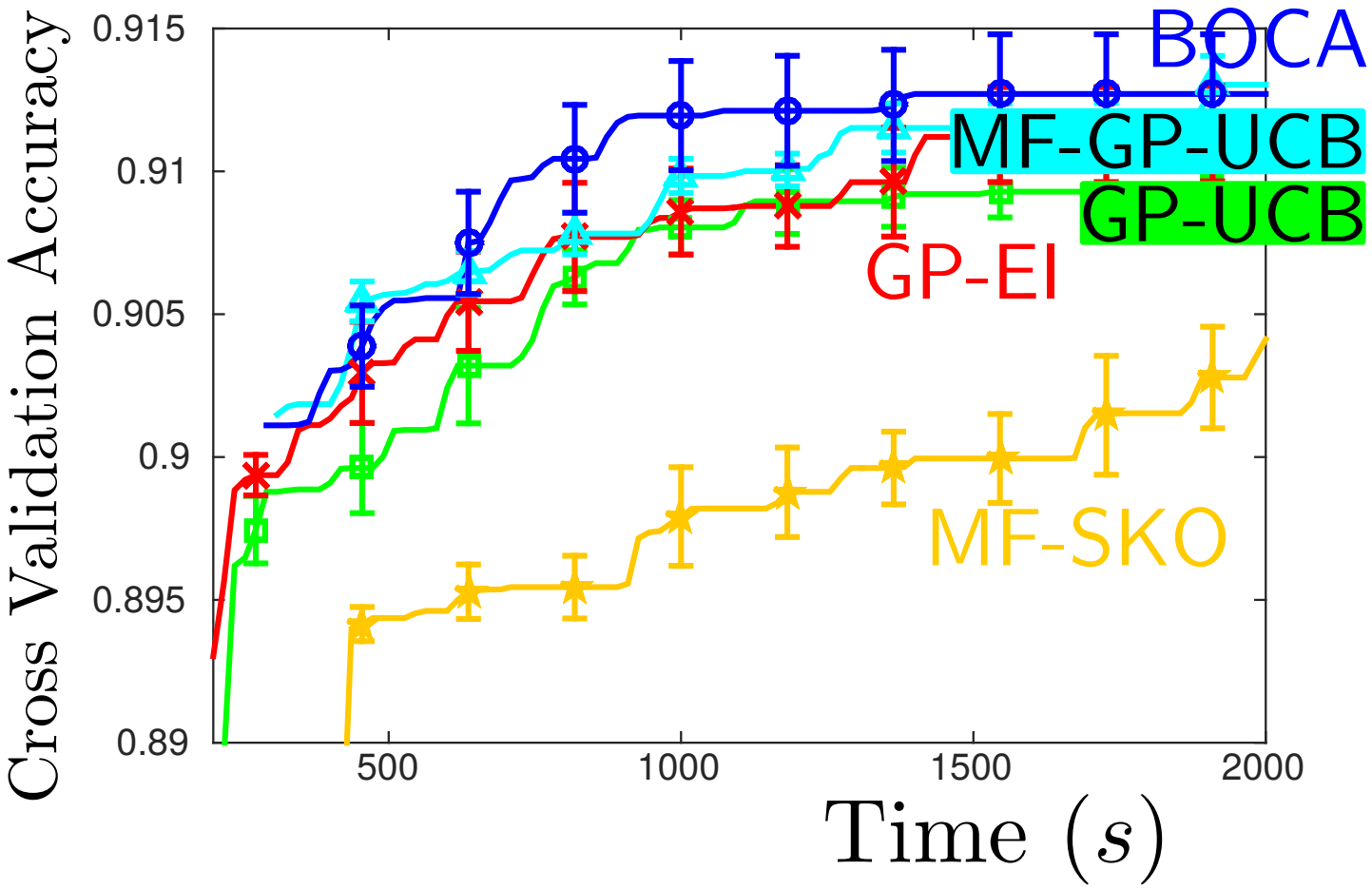
If h is large, $\mathrm{vol}(\mathcal{X}_\alpha) \ll \mathrm{vol}(\mathcal{X})$, $\Psi_{n_\Lambda}(\mathcal{X}_\alpha) \ll \Psi_{n_\Lambda}(\mathcal{X})$.

Experiment: SVM with 20 News Groups

Tune two hyper-parameters for the SVM. $(d = 2)$
Dataset has $N_\bullet = 15K$ data and use $T_\bullet = 100$ iterations.
But can choose $N \in [5K, 15K]$ or $T \in [20, 100]$. $(p = 2)$

Experiment: SVM with 20 News Groups

Tune two hyper-parameters for the SVM. $(d = 2)$
Dataset has $N_\bullet = 15K$ data and use $T_\bullet = 100$ iterations.
But can choose $N \in [5K, 15K]$ or $T \in [20, 100]$. $(p = 2)$



Summary

Multi-fidelity K-armed bandits (Kandasamy et al. NIPS 2016a)

- An algorithm MF-UCB and an upper bound on the regret.
- An almost matching lower bound.

Summary

Multi-fidelity K-armed bandits (Kandasamy et al. NIPS 2016a)

- An algorithm MF-UCB and an upper bound on the regret.
- An almost matching lower bound.

Key takeaways (Kandasamy et al. NIPS 2016a, Kandasamy et al. NIPS 2016b, Kandasamy et al. Arxiv 2017)

- Upper confidence bound strategy
- Choose higher fidelity only after controlling uncertainty/variance at lower fidelities.
- Explore the entire space using cheap low fidelities and reserve expensive higher fidelities for promising candidates.

Gautam Dasarathy

Junier Oliva

Jeff Schneider

Barnabas Poczos

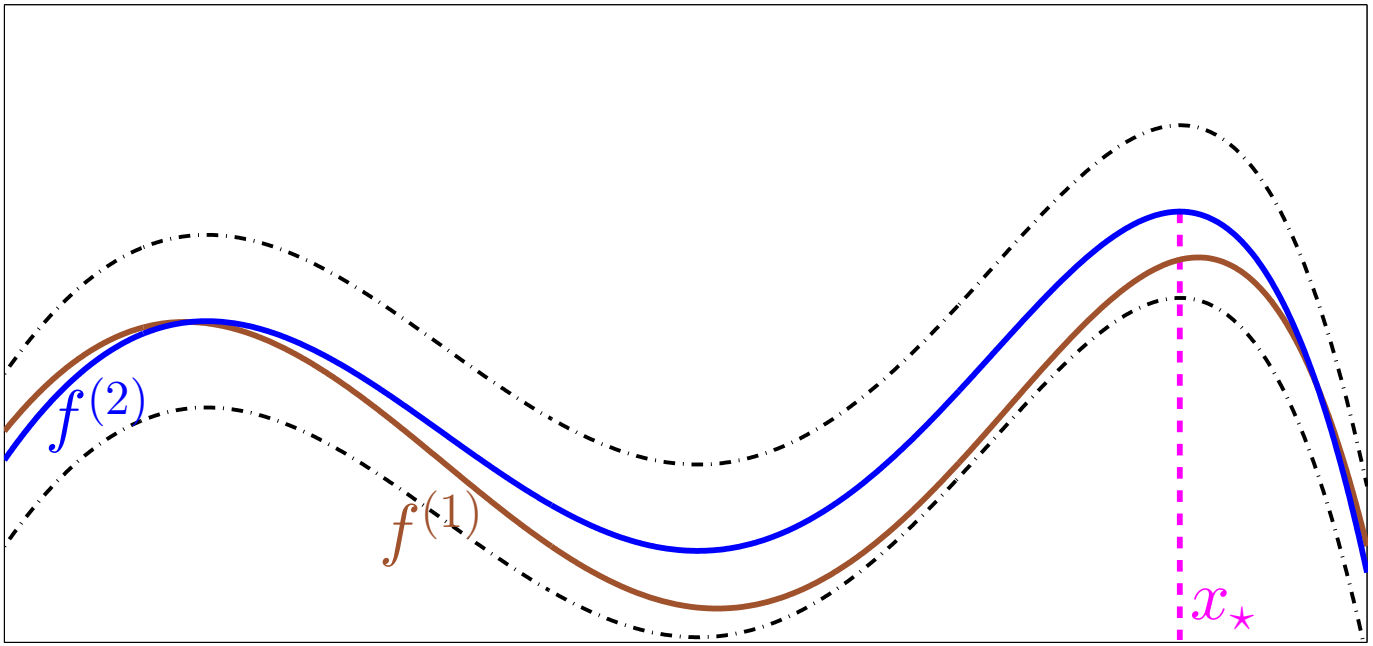
Thank you.

Code for MF-GP-UCB:

`https://github.com/kirthevasank/mf-gp-ucb`

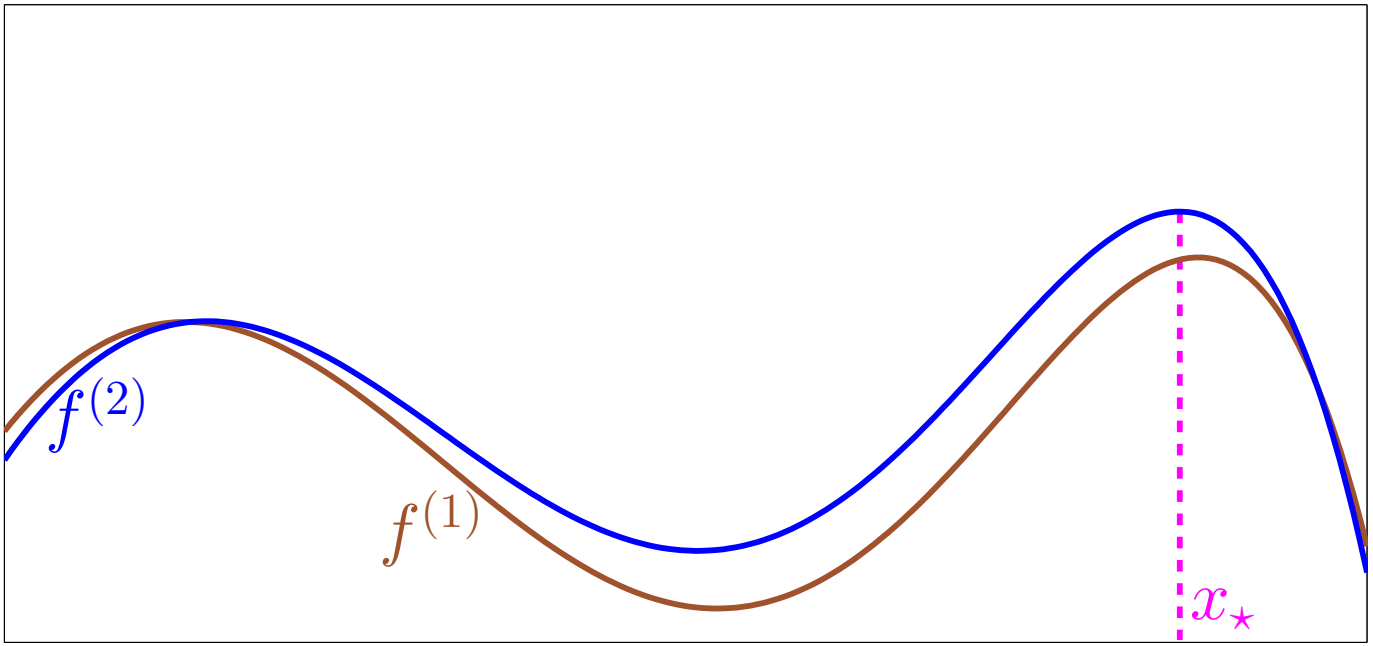
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



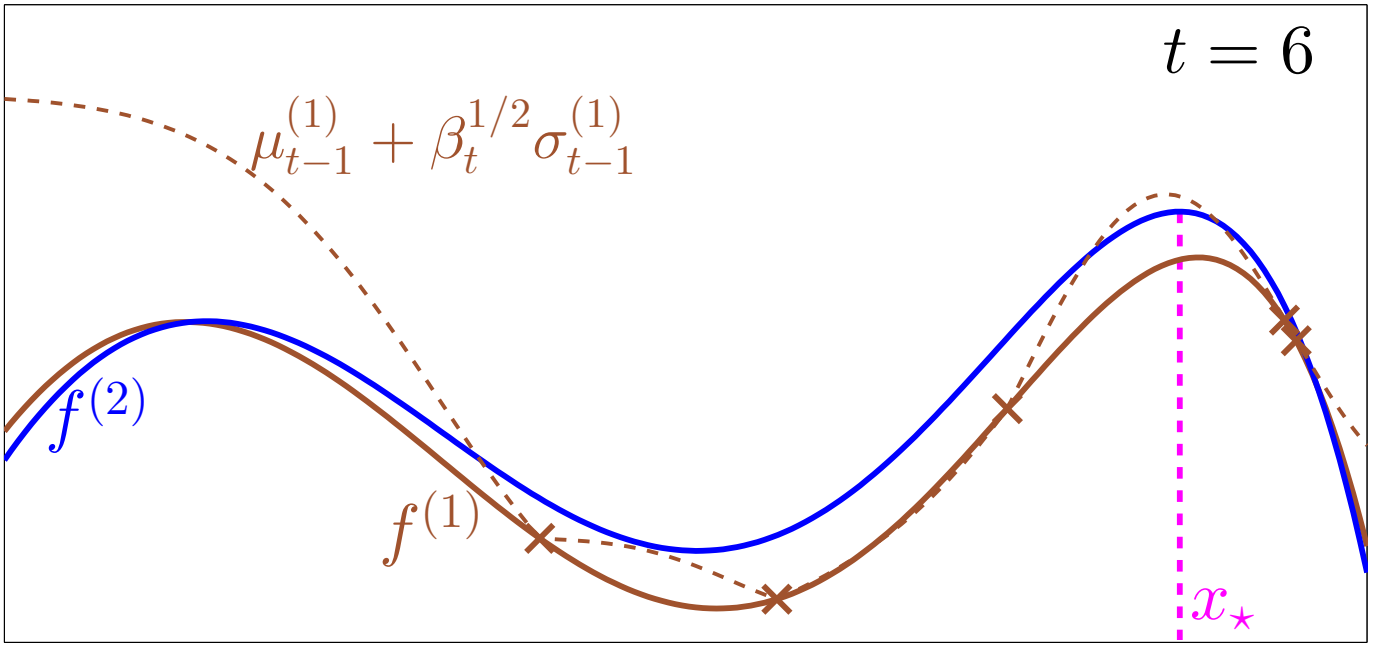
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



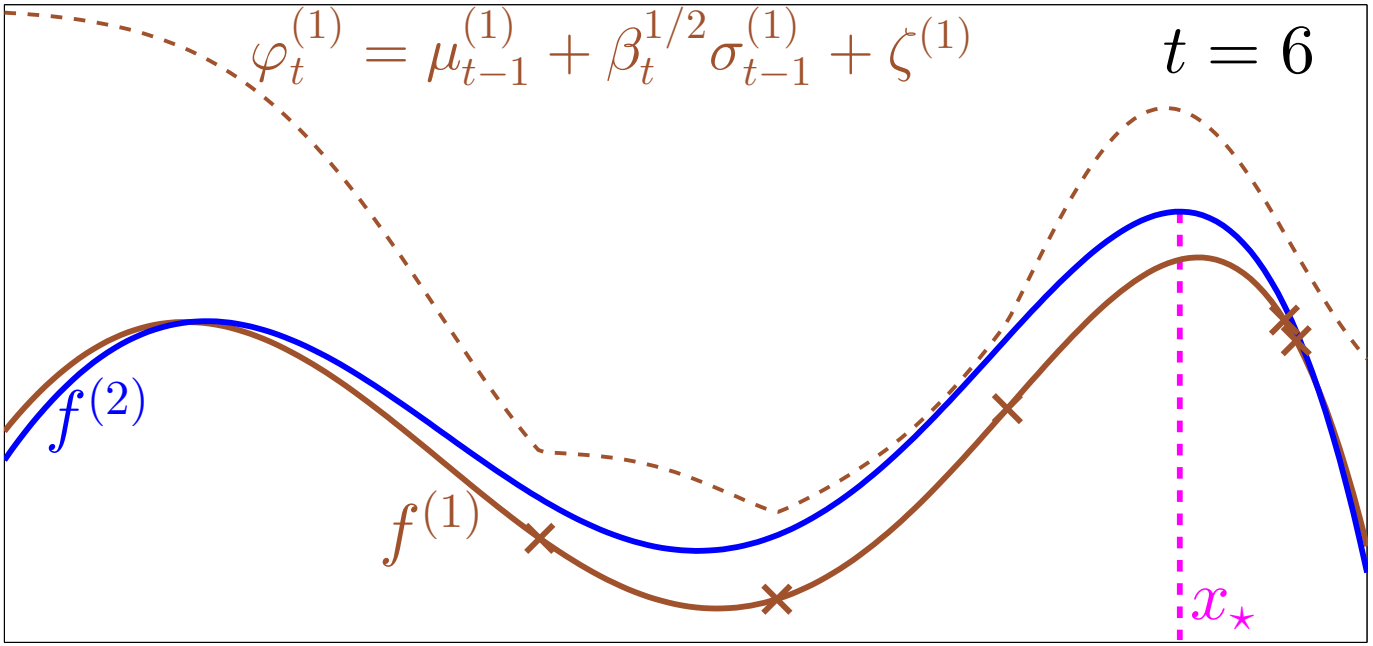
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



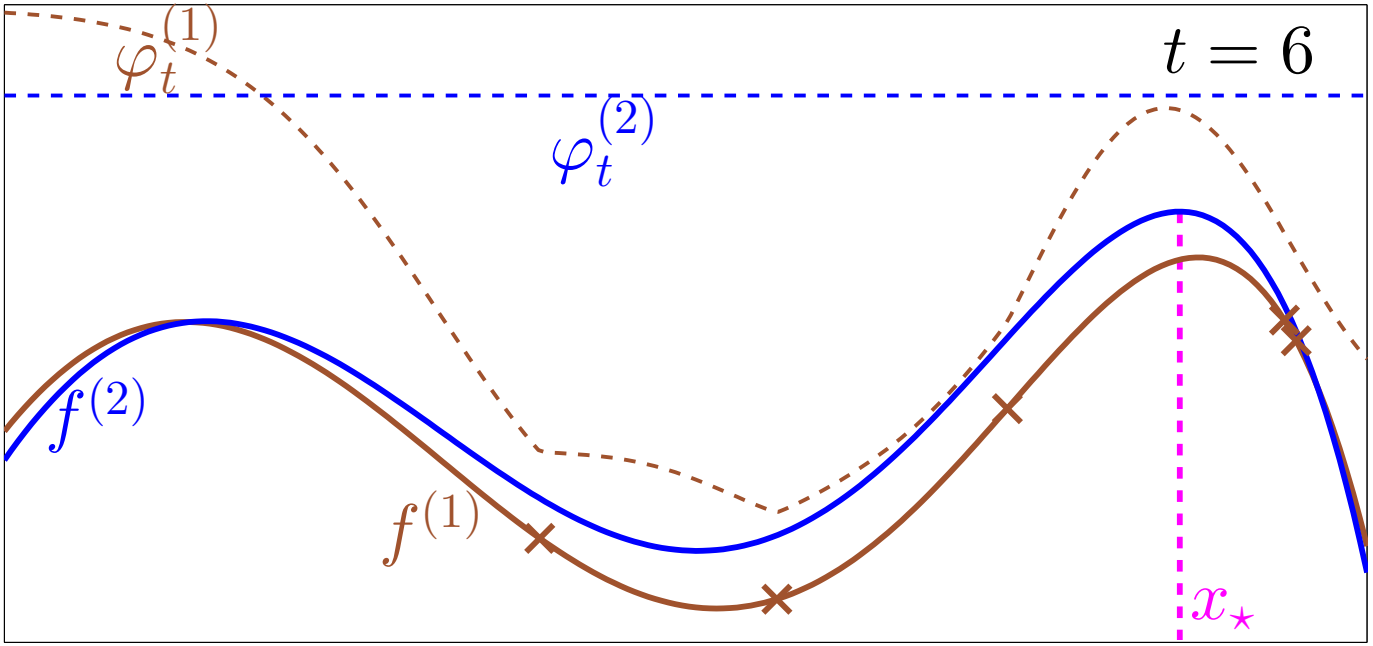
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



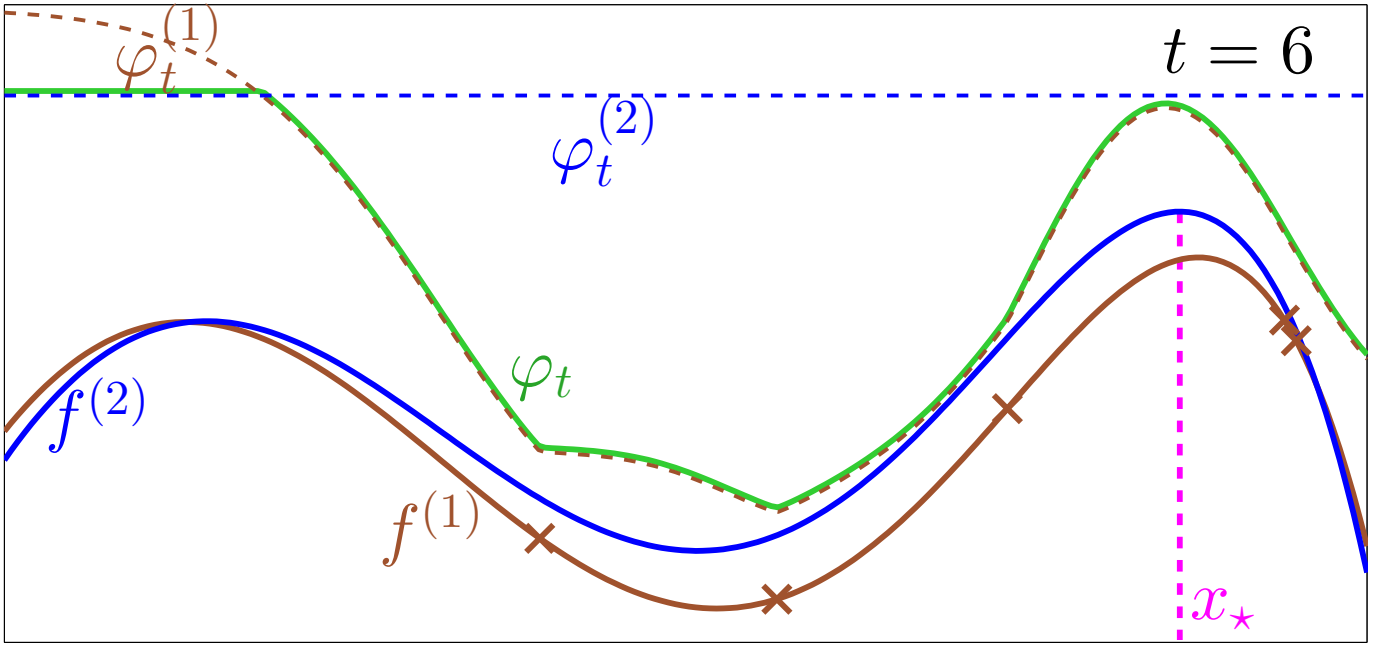
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



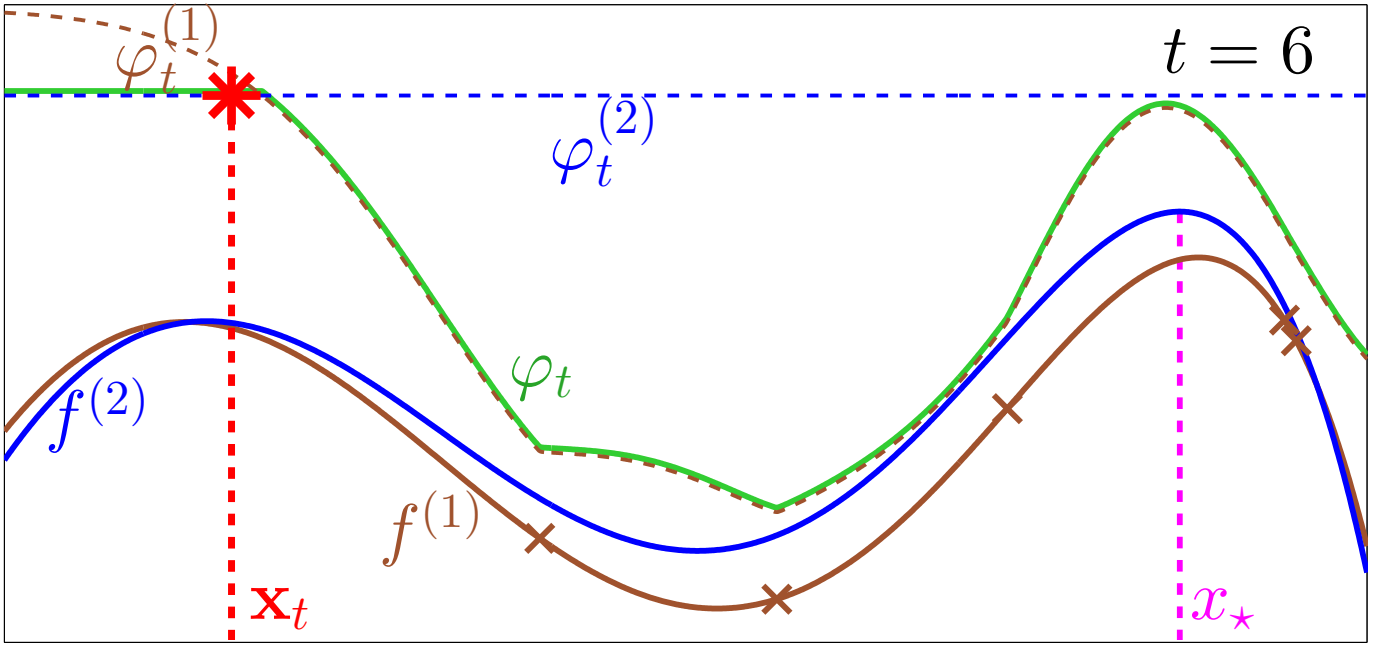
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



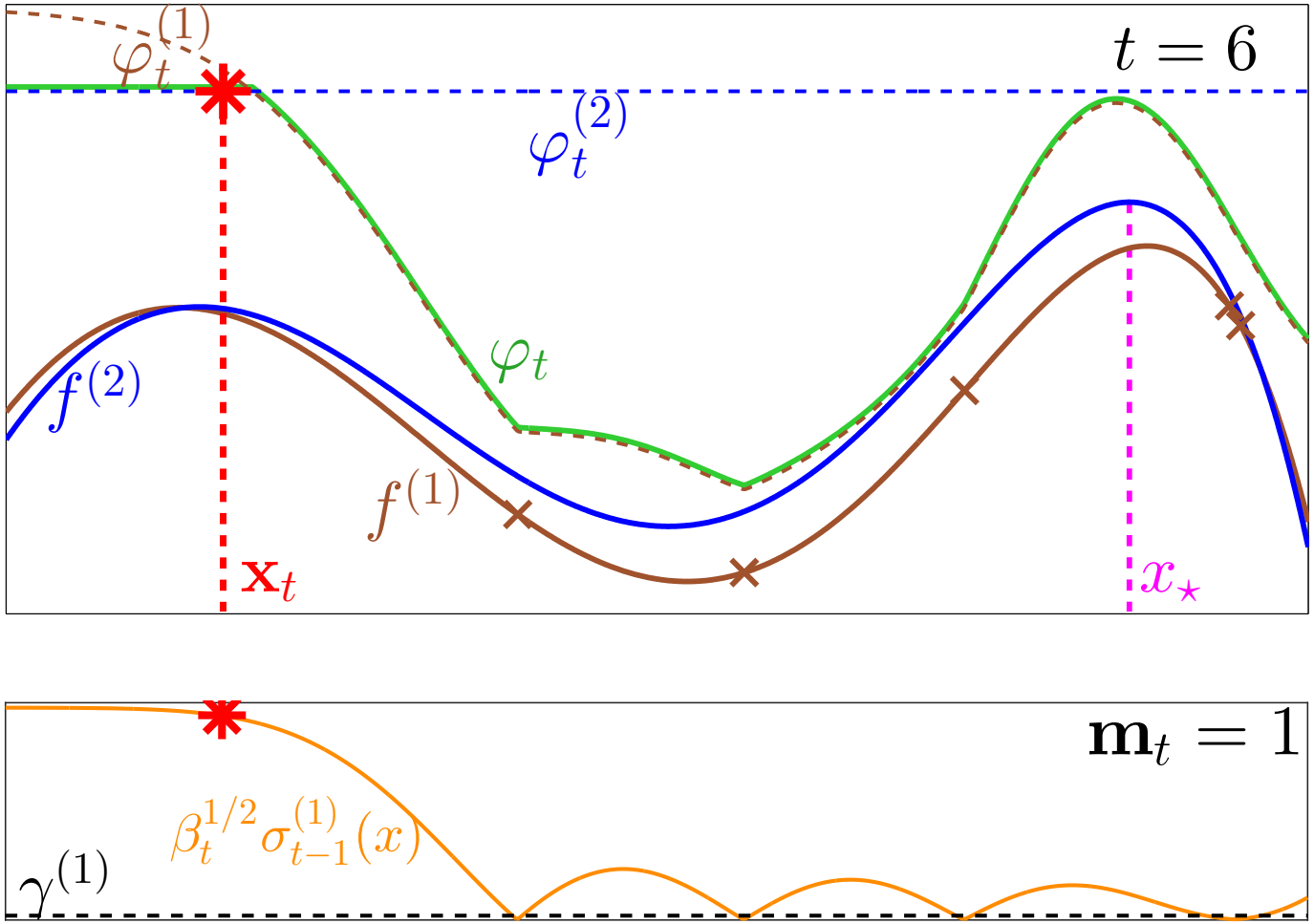
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



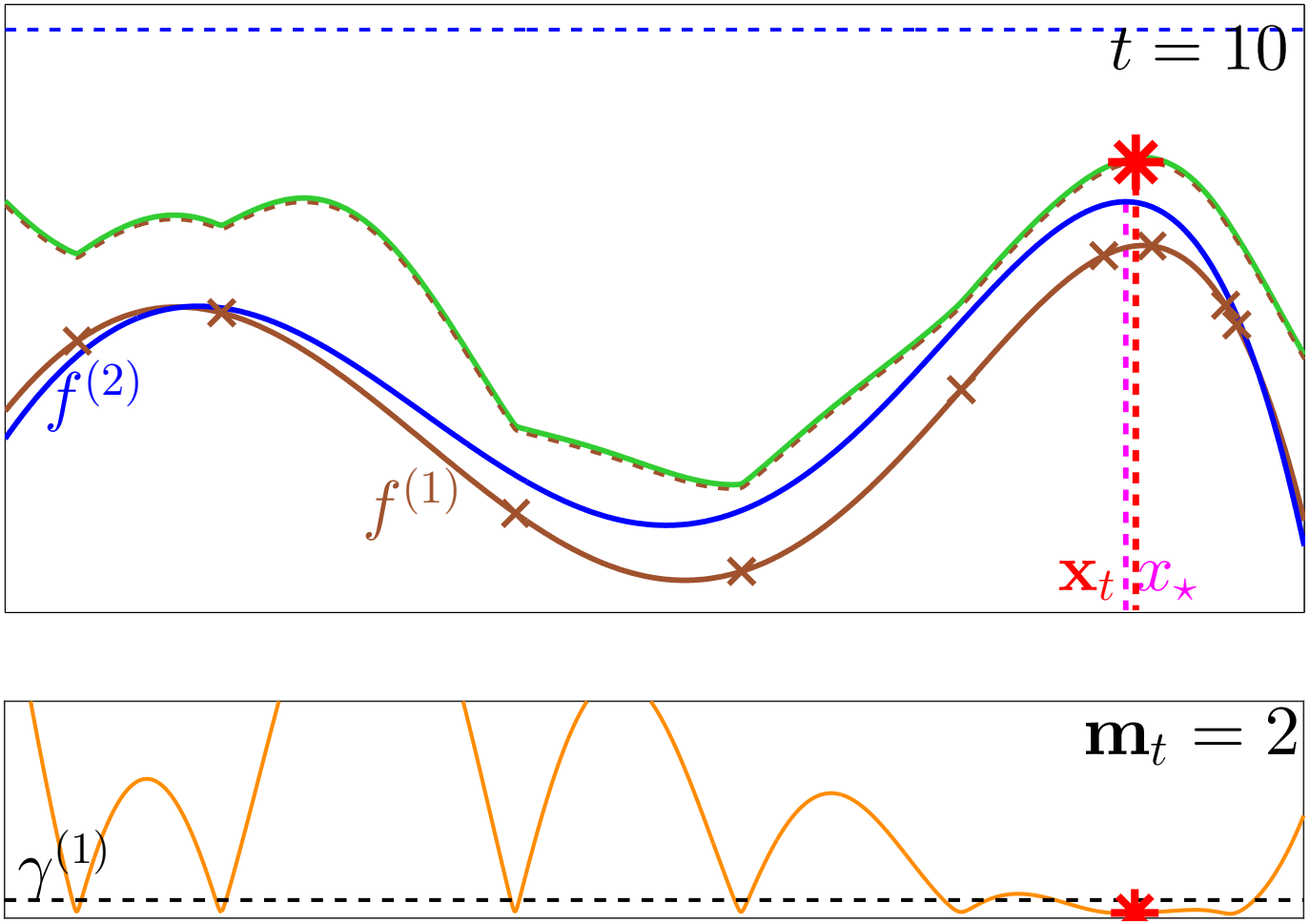
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



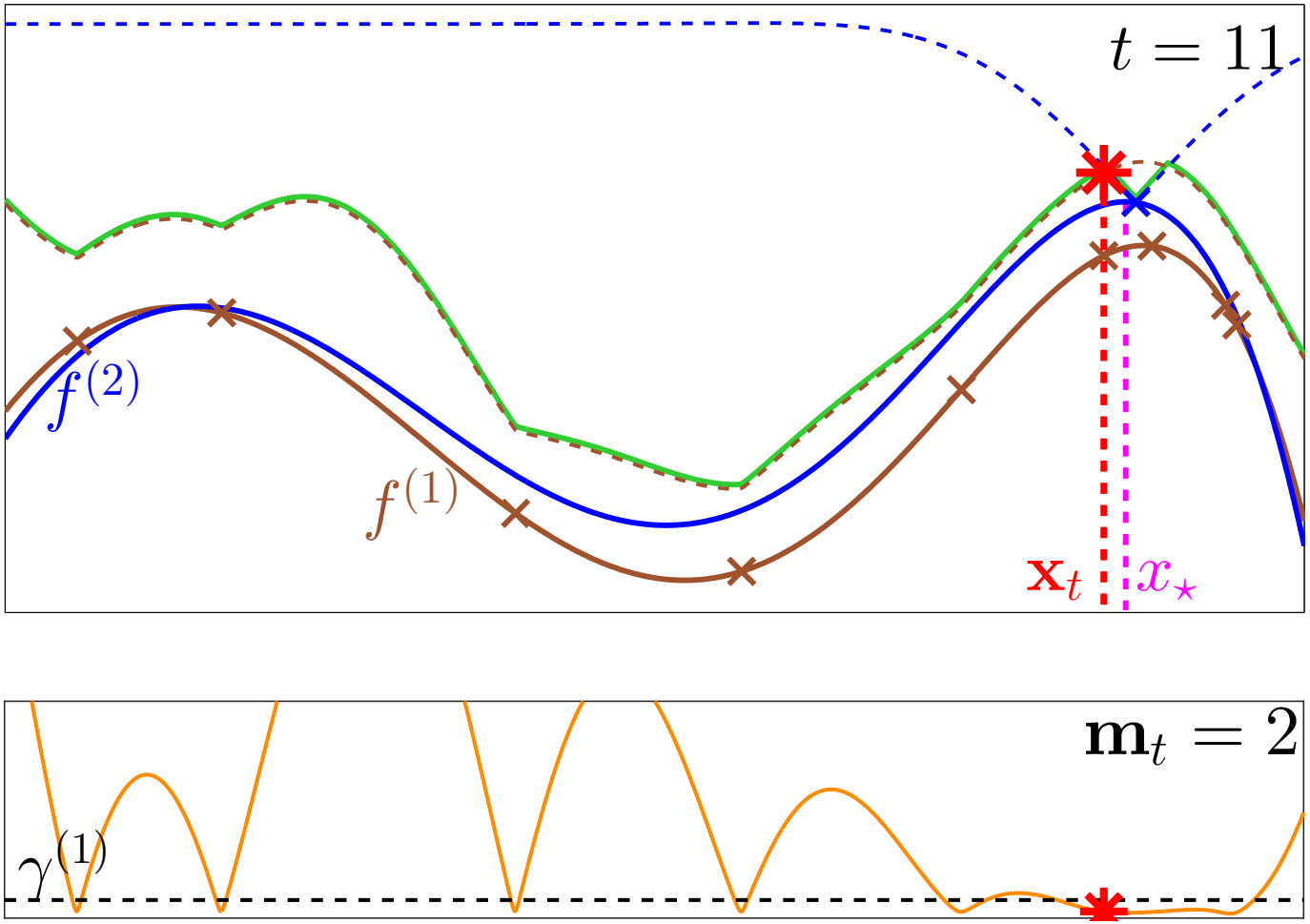
MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



MF-GP-UCB

(Kandasamy et al. NIPS 2016b)



MF-GP-UCB

(Kandasamy et al. NIPS 2016b)