Multi-fidelity Bandit Optimisation

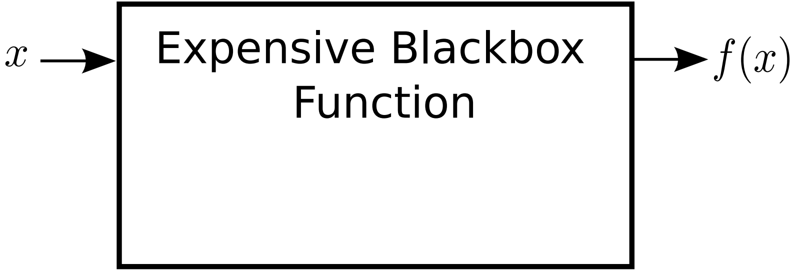
Kirthevasan Kandasamy

Carnegie Mellon University

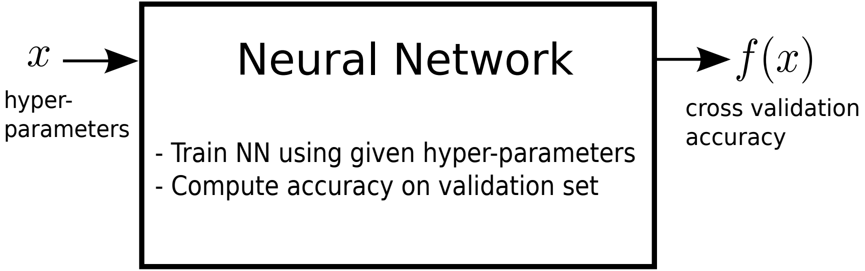
July 12, 2016

University College London

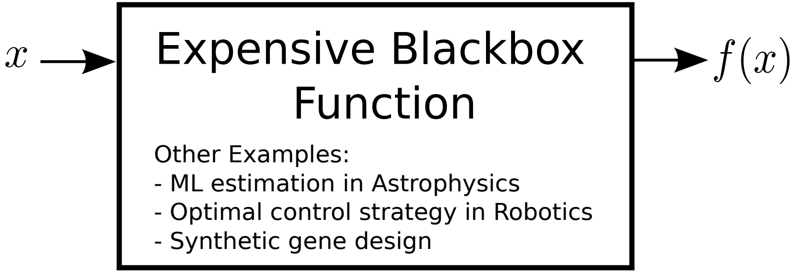
Bandit Optimisation



Bandit Optimisation

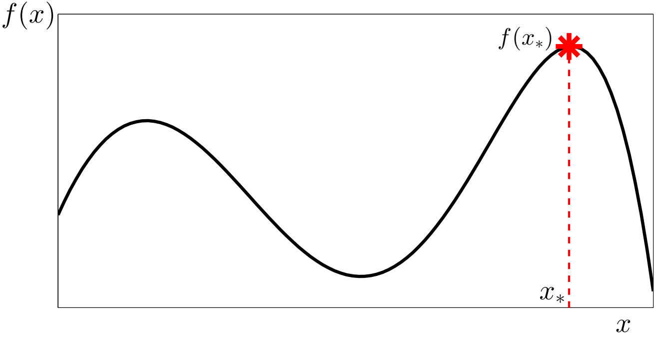


Bandit Optimisation



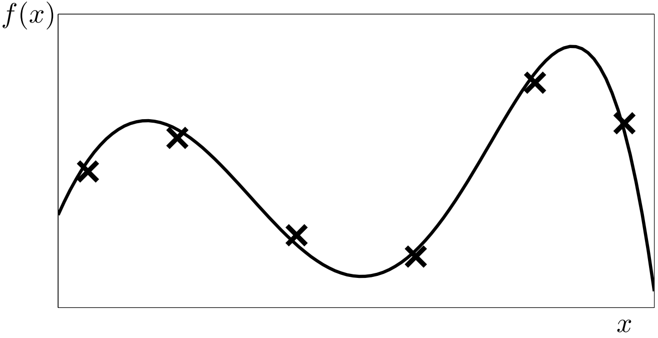
Bandit Optimisation

$f : \mathcal{X} \equiv [0,1]^d \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \operatorname{argmax}_x f(x)$.



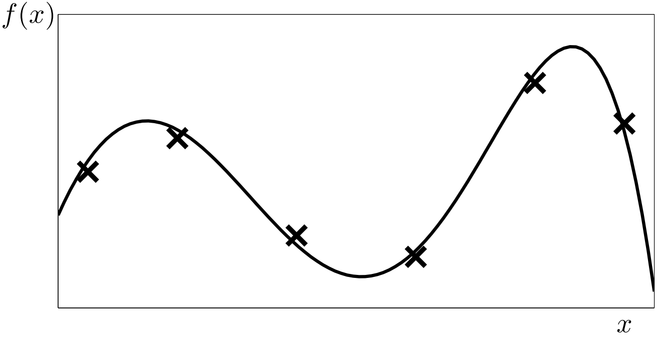
Bandit Optimisation

$f : \mathcal{X} \equiv [0,1]^d \rightarrow \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.



Bandit Optimisation

$f : \mathcal{X} \equiv [0,1]^d \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.

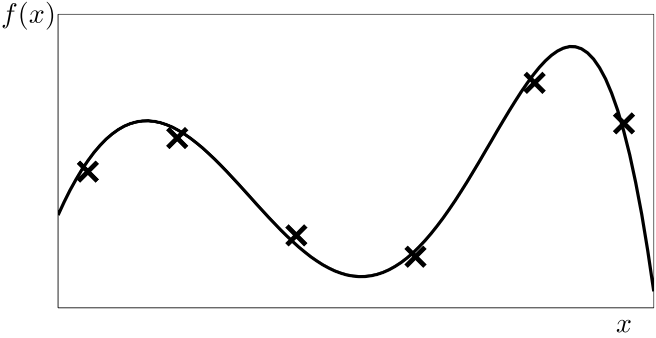


Optimisation $\cong$ Minimise *Simple Regret*.

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

Bandit Optimisation

$f : \mathcal{X} \equiv [0,1]^d \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x\, f(x)$.

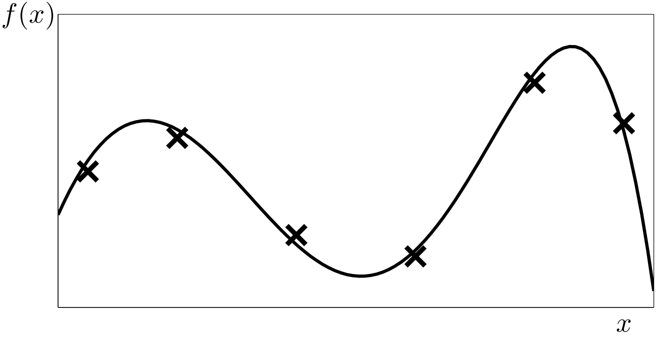


Bandits $\cong$ Minimise *Cumulative Regret*.

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

Bandit Optimisation

$f : \mathcal{X} \equiv [0,1]^d \to \mathbb{R}$ is an expensive, black-box, noisy function.
Let $x_\star = \mathrm{argmax}_x f(x)$.



Both problems are related.

$$S_n \leq \frac{1}{n} R_n$$

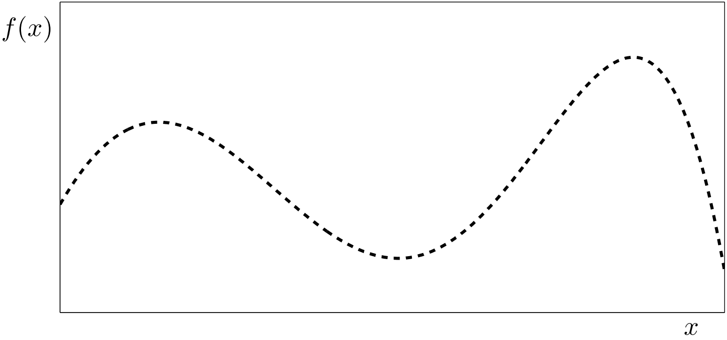
Gaussian Processes ($\mathcal{GP}$)

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

Gaussian Processes ($\mathcal{GP}$)

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

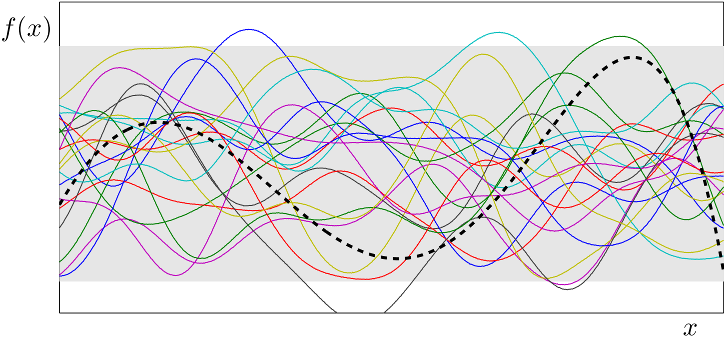
Functions with no observations



Gaussian Processes ($\mathcal{GP}$)

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

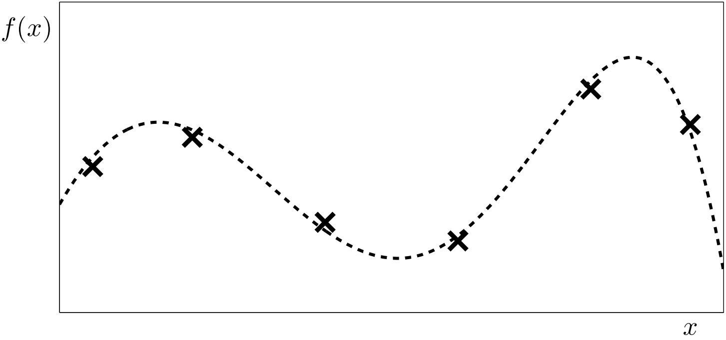
Prior $\mathcal{GP}$



Gaussian Processes ($\mathcal{GP}$)

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

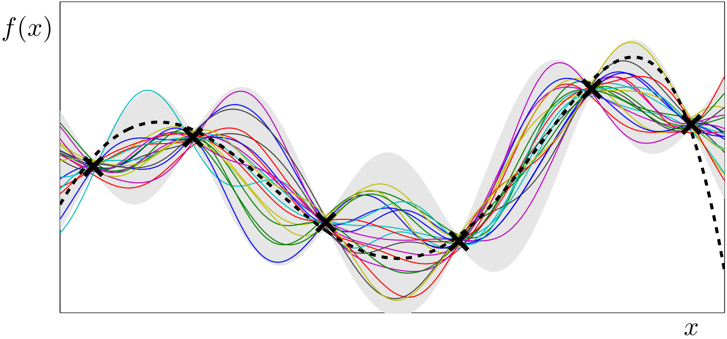
Observations



Gaussian Processes ($\mathcal{GP}$)

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

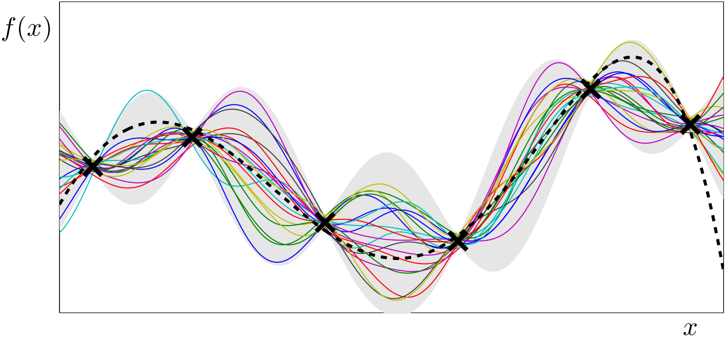
Posterior $\mathcal{GP}$ given Observations



Gaussian Processes ($\mathcal{GP}$)

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

Posterior $\mathcal{GP}$ given Observations

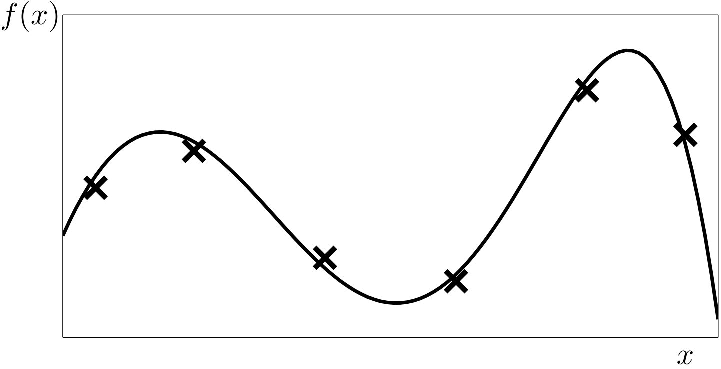


After t observations, $\quad f(x) \sim \mathcal{N}(\,\mu_t(x),\, \sigma_t^2(x)\,)$.

Gaussian Process Bandit (Bayesian) Optimisation

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

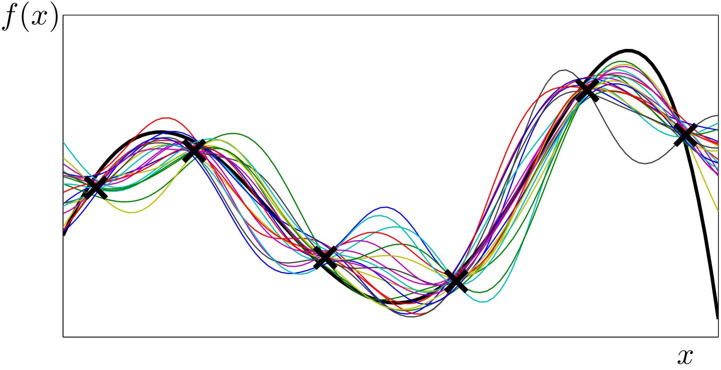
GP-UCB (Srinivas et al. 2010).



Gaussian Process Bandit (Bayesian) Optimisation

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

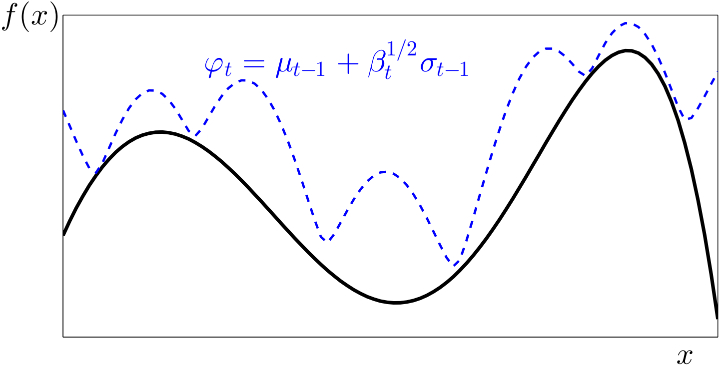
GP-UCB (Srinivas et al. 2010).



Gaussian Process Bandit (Bayesian) Optimisation

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

GP-UCB (Srinivas et al. 2010).

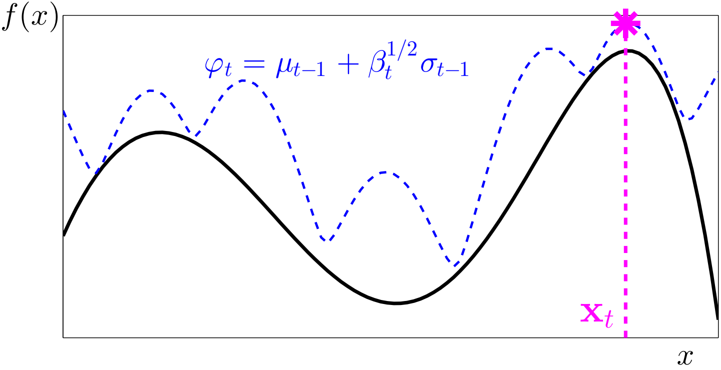


Construct Upper Conf. Bound: $\varphi_t(x) = \mu_{t-1}(x) + \beta_t^{1/2}\sigma_{t-1}(x)$.

Gaussian Process Bandit (Bayesian) Optimisation

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

GP-UCB (Srinivas et al. 2010).



Maximise Upper Confidence Bound.

GP-UCB

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

GP-UCB

$$\mathbf{x}_t = \underset{x}{\operatorname{argmax}} \;\; \mu_{t-1}(x) + \beta_t^{1/2}\sigma_{t-1}(x)$$

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration

GP-UCB

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

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration

- β_t controls the tradeoff. $\beta_t \asymp \log t$.

GP-UCB

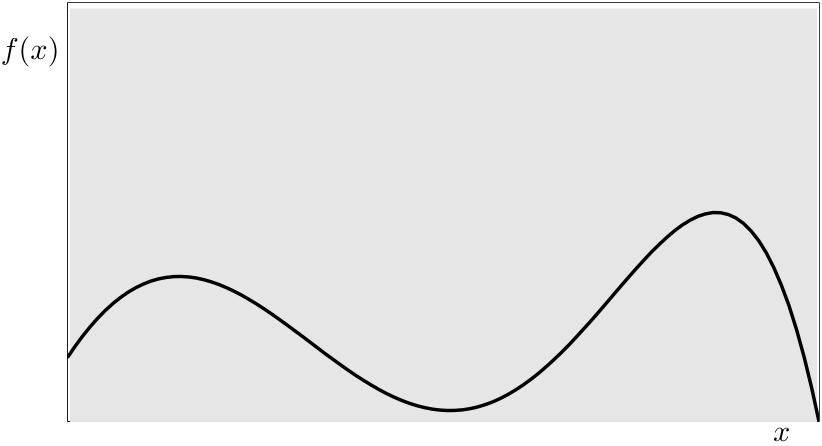
$$\mathbf{x}_t = \underset{x}{\operatorname{argmax}} \;\; \mu_{t-1}(x) + \beta_t^{1/2} \sigma_{t-1}(x)$$

- μ_{t-1}: Exploitation
- σ_{t-1}: Exploration

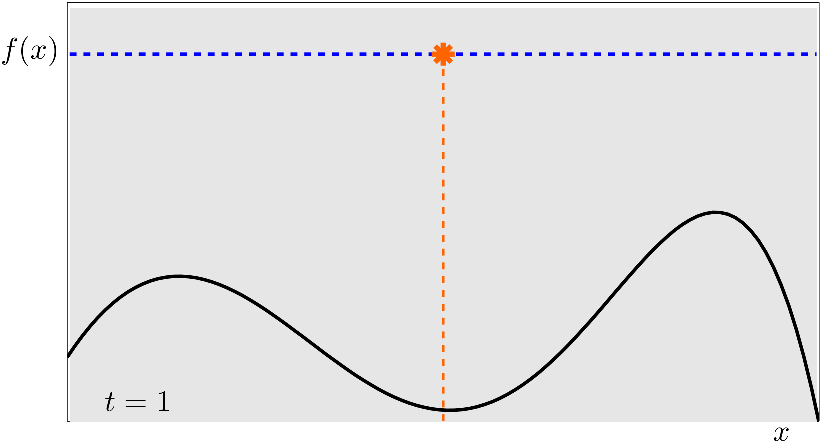
- β_t controls the tradeoff. $\quad \beta_t \asymp \log t$.

- The upper bound $\mu_{t-1} + \beta_t^{1/2} \sigma_{t-1}$ becomes tighter around the optimum $x_\star$.

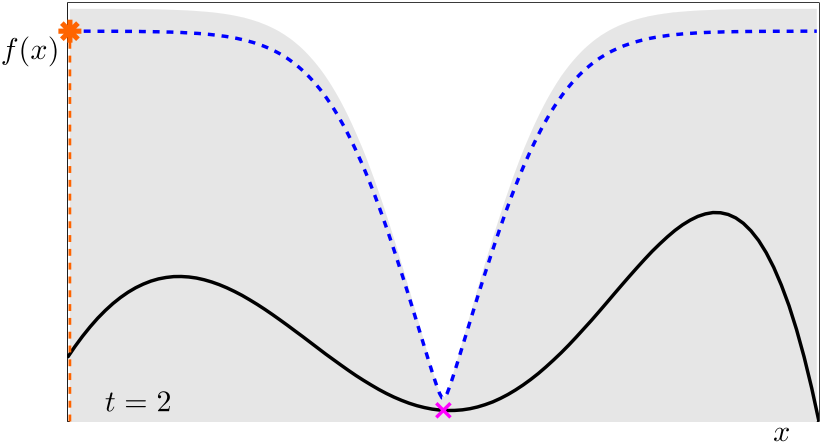
GP-UCB



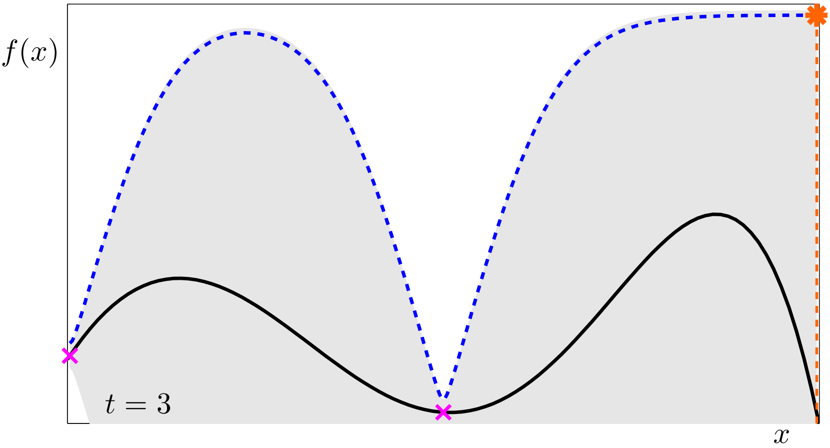
GP-UCB



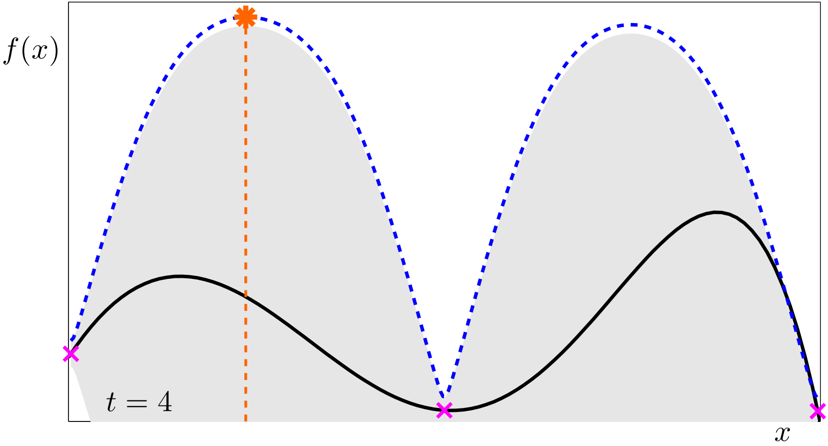
GP-UCB



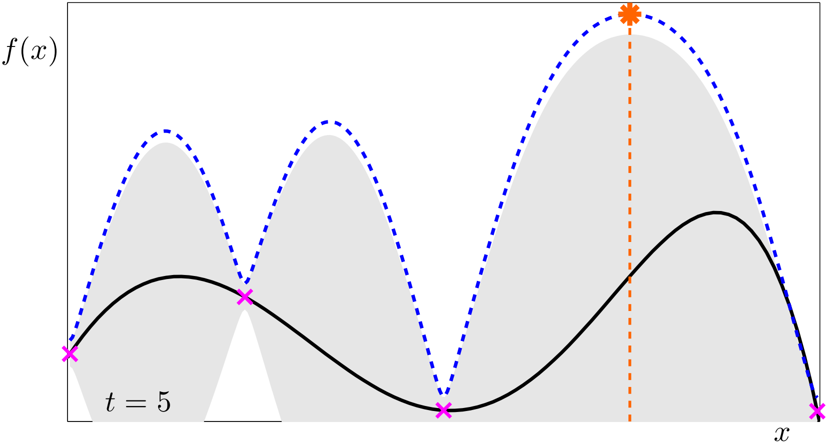
GP-UCB



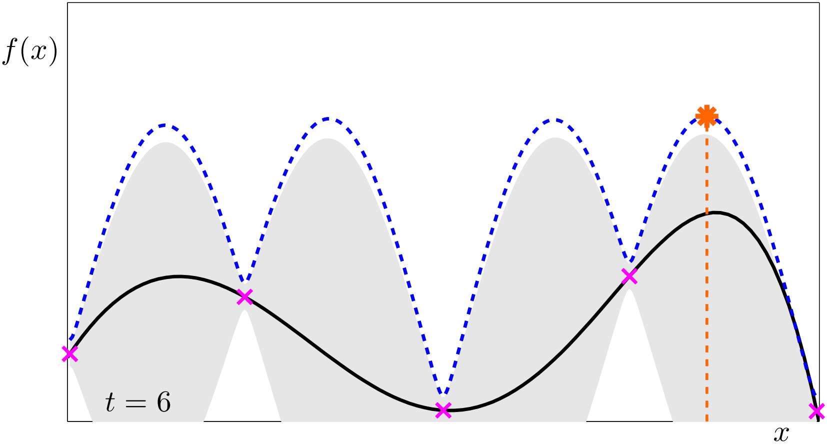
GP-UCB



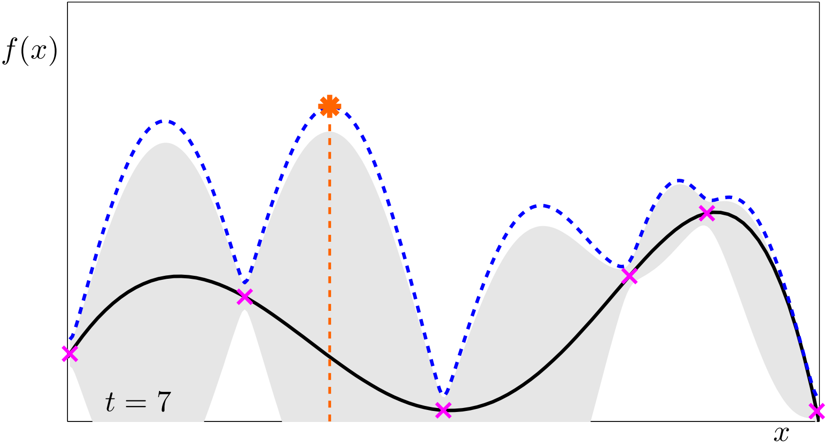
GP-UCB



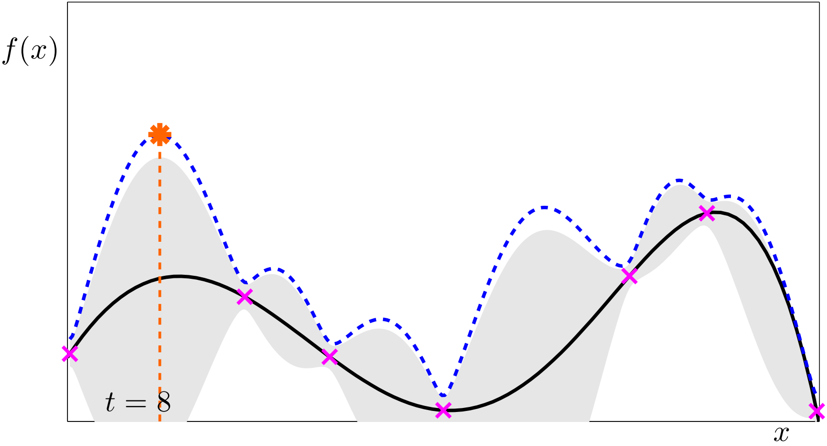
GP-UCB



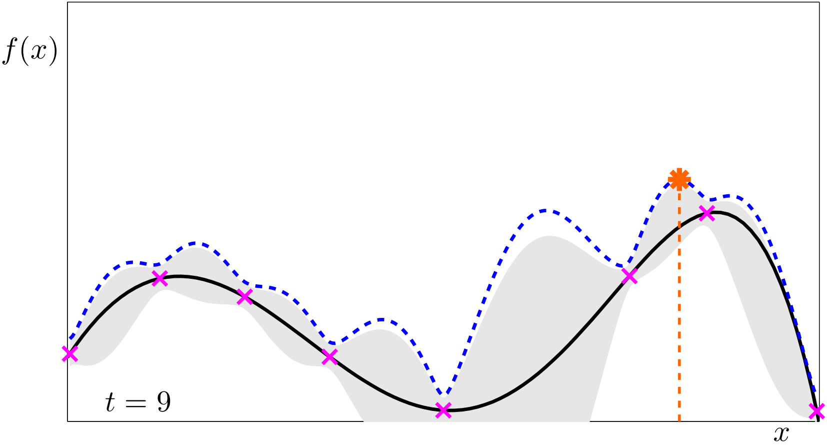
GP-UCB



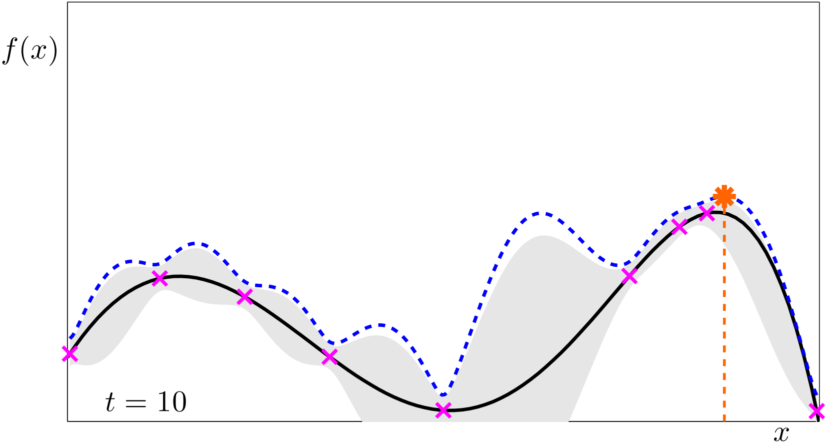
GP-UCB



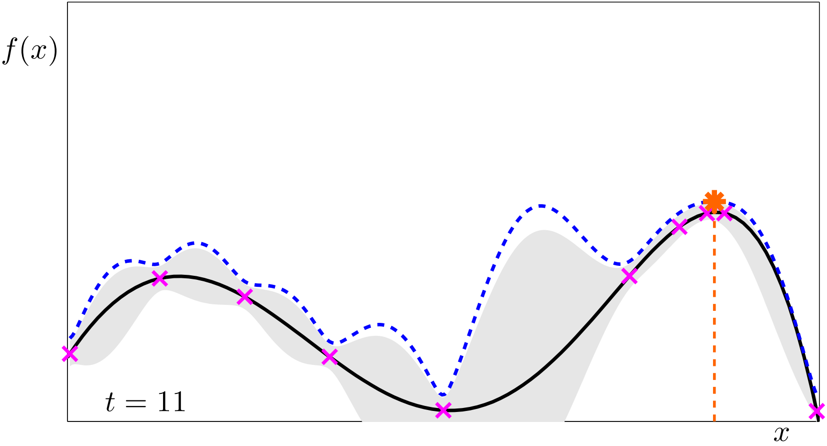
GP-UCB



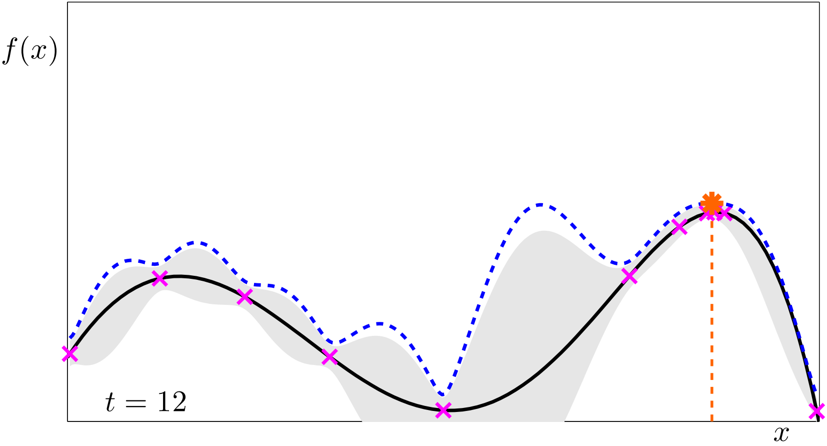
GP-UCB



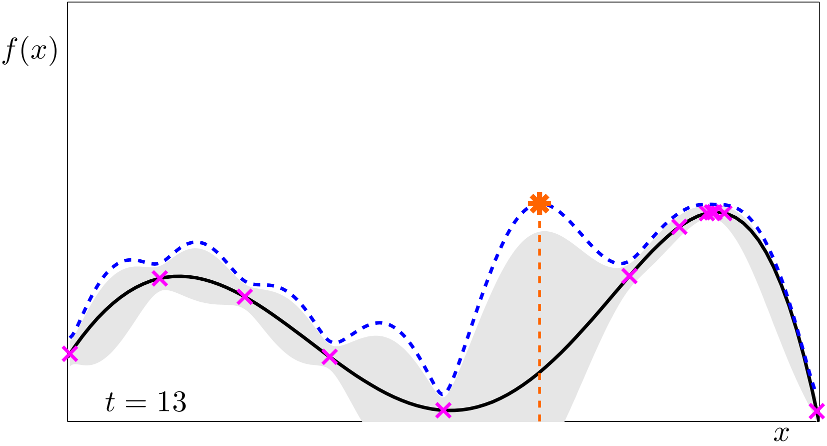
GP-UCB



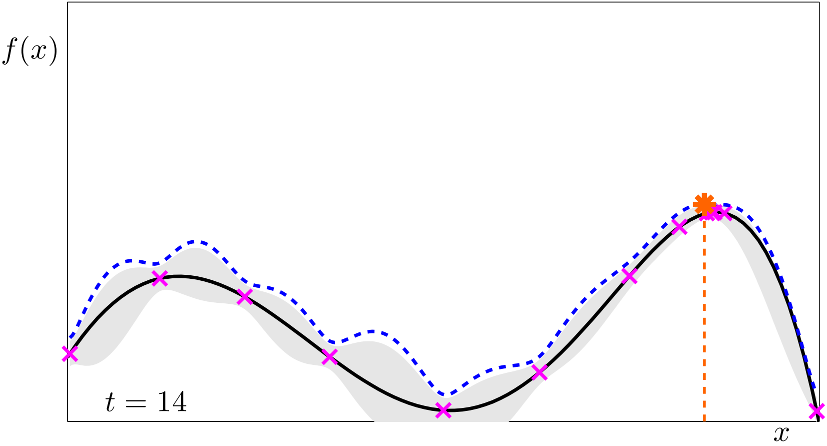
GP-UCB



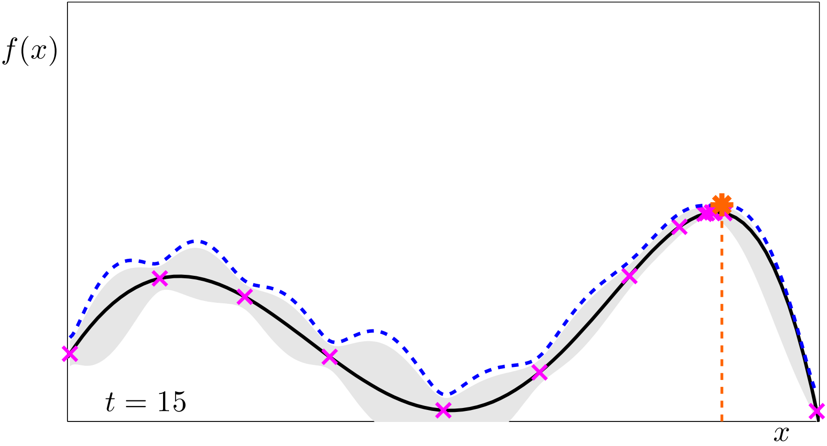
GP-UCB



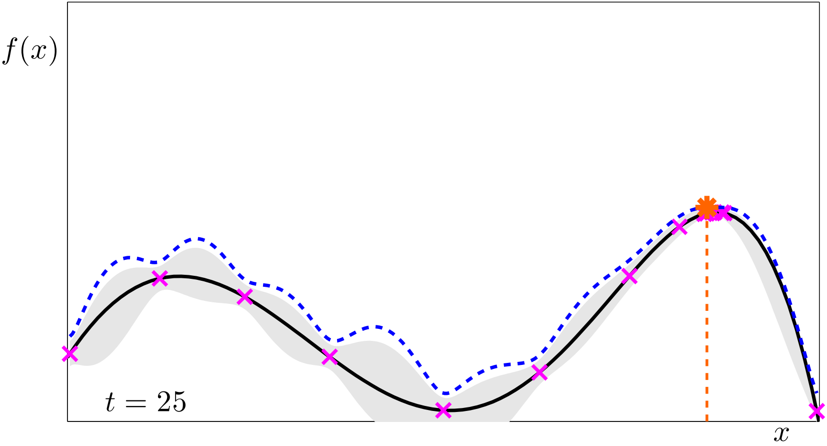
GP-UCB



GP-UCB



GP-UCB



What if we have cheap approximations to f ?

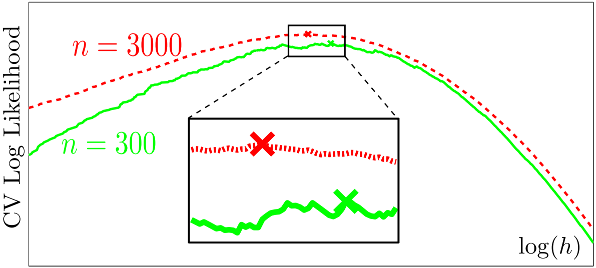
What if we have cheap approximations to f ?

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

What if we have cheap approximations to f ?

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

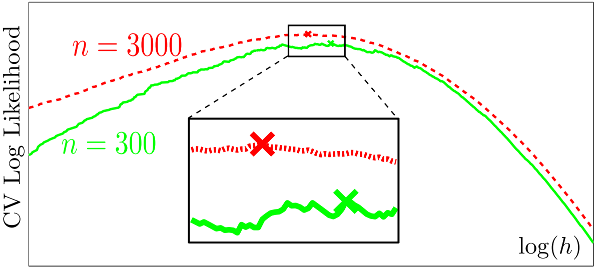
   E.g. Bandwidth (h) selection in kernel density estimation.



What if we have cheap approximations to f ?

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

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

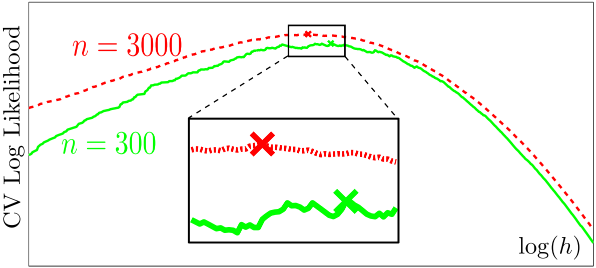


2. Robotics: Simulation vs Real world experiment.

What if we have cheap approximations to f ?

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

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



2. Robotics: Simulation vs Real world experiment.

3. Compuatational Astrophysics: Cosmological simulations with less granularity.

Outline

1. Multi-fidelity Bandit Optimisation
   - Formalism & Challenges

2. **MF-GP-UCB**: Multi-fidelity optimisation using GPs
   - Single Approximation/ 2 fidelity setting
   - Theoretical Results & Proof Sketches

3. MF-GP-UCB with multiple fidelities.

4. Experiments

Multi-fidelity Bandit Optimisation

Goal:

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

- **But ..**

Multi-fidelity Bandit Optimisation

Goal:

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

- **But ..** we have $M-1$ cheap approximations $f^{(1)}, f^{(2)}, \ldots, f^{(M-1)}$ to the function of interest $f = f^{(M)}$.

Multi-fidelity Bandit Optimisation

Goal:

- Optimise f. $\quad x_\star = \mathrm{argmax}_x f(x)$.

- **But ..** we have $M-1$ cheap approximations $f^{(1)}, f^{(2)}, \ldots, f^{(M-1)}$ to the function of interest $f = f^{(M)}$.

- $f^{(m)}$ costs $\lambda^{(m)}$. $\quad \lambda^{(1)} < \lambda^{(2)} < \ldots \lambda^{(M-1)} < \lambda^{(M)}$.
  "cost": could be computation time, money etc.

Multi-fidelity Bandit Optimisation

Goal:

- Optimise f. $\quad x_\star = \mathrm{argmax}_x\, f(x)$.
- **But ..** we have $M-1$ cheap approximations $f^{(1)}, f^{(2)}, \ldots, f^{(M-1)}$ to the function of interest $f = f^{(M)}$.
- $f^{(m)}$ costs $\lambda^{(m)}$. $\quad \lambda^{(1)} < \lambda^{(2)} < \ldots \lambda^{(M-1)} < \lambda^{(M)}$.
  "cost": could be computation time, money etc.
- Assumptions
  - $f^{(m)} \sim \mathcal{GP}(0, \kappa)$ for all $m = 1, \ldots, M$.
  - $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.
    $\zeta^{(m)}$'s are decreasing with m and are *known*.

Outline for a Sequential Strategy

At each step:

- Determine the point $\mathbf{x}_t \in \mathcal{X}$ and fidelity $\mathbf{m}_t$ at which you want to query.

Outline for a Sequential Strategy

At each step:

- Determine the point $\mathbf{x}_t \in \mathcal{X}$ and fidelity $\mathbf{m}_t$ at which you want to query.

- At time t, we have queried previously at any one of M fidelities. Use all these information to determine next query.

Outline for a Sequential Strategy

At each step:

- Determine the point $\mathbf{x}_t \in \mathcal{X}$ and fidelity $\mathbf{m}_t$ at which you want to query.
- At time t, we have queried previously at any one of M fidelities. Use all these information to determine next query.
- **End Goal:** Maximise $f^{(M)}$. We don't really care much about the value of the query at the lower fidelities.

Outline for a Sequential Strategy

At each step:

- Determine the point $\mathbf{x}_t \in \mathcal{X}$ and fidelity $\mathbf{m}_t$ at which you want to query.
- At time t, we have queried previously at any one of M fidelities. Use all these information to determine next query.
- **End Goal:** Maximise $f^{(M)}$. We don't really care much about the value of the query at the lower fidelities.
- But use $f^{(1)}, \ldots, f^{(M-1)}$ to guide search for $x_\star$ at $f^{(M)}$.

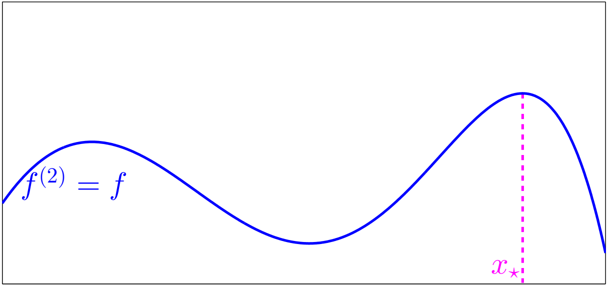
Outline for a Sequential Strategy

At each step:

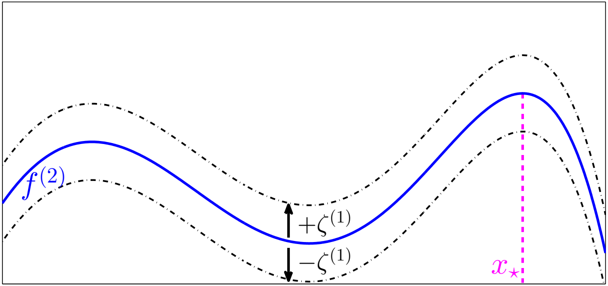
- Determine the point $\mathbf{x}_t \in \mathcal{X}$ and fidelity $\mathbf{m}_t$ at which you want to query.
- At time t, we have queried previously at any one of M fidelities. Use all these information to determine next query.
- **End Goal:** Maximise $f^{(M)}$. We don't really care much about the value of the query at the lower fidelities.
- But use $f^{(1)}, \ldots, f^{(M-1)}$ to guide search for $x_\star$ at $f^{(M)}$.

MF-GP-UCB: Multi-fidelity Gaussian Process Upper Confidence Bound

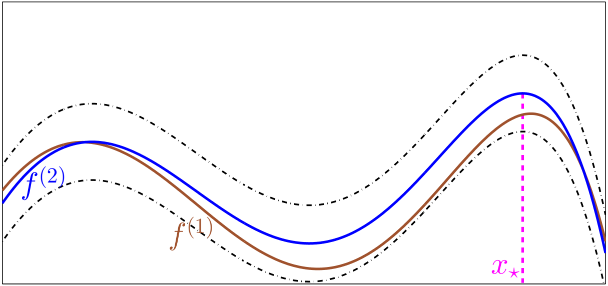
Challenges (in 2 fidelities)



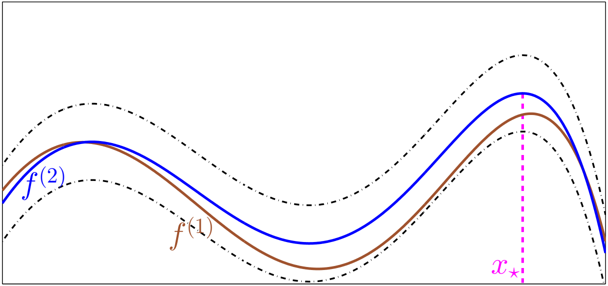
Challenges (in 2 fidelities)



Challenges (in 2 fidelities)

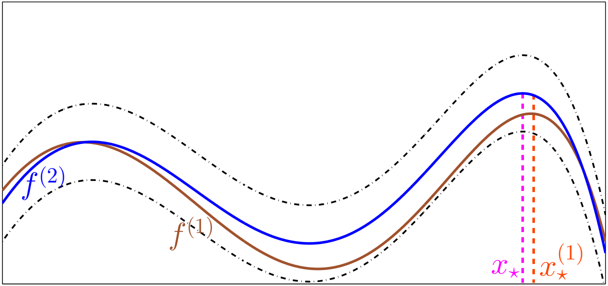


Challenges (in 2 fidelities)



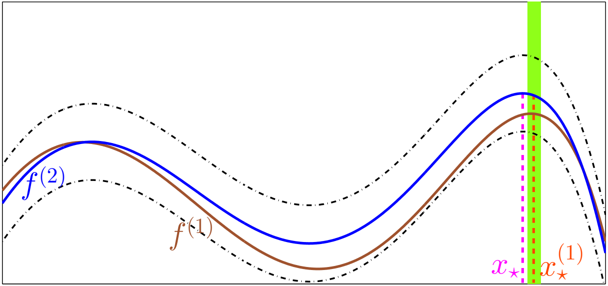
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.

Challenges (in 2 fidelities)



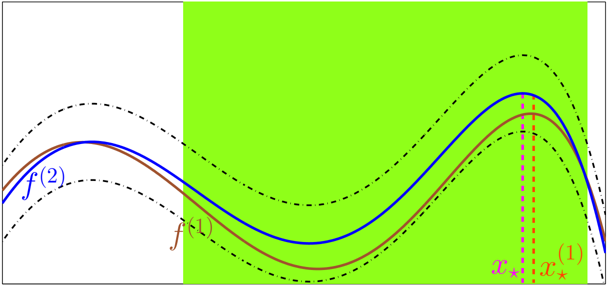
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges (in 2 fidelities)



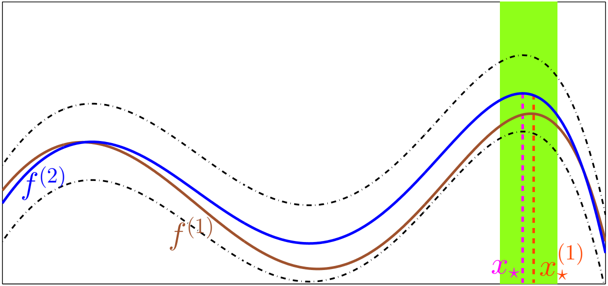
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges (in 2 fidelities)



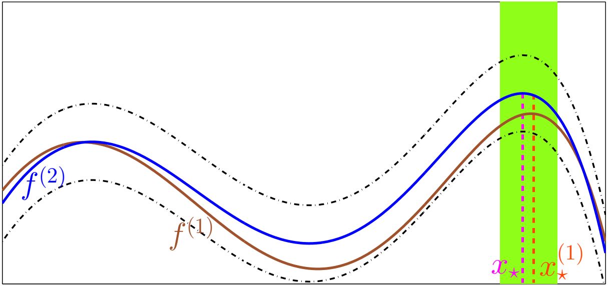
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges (in 2 fidelities)



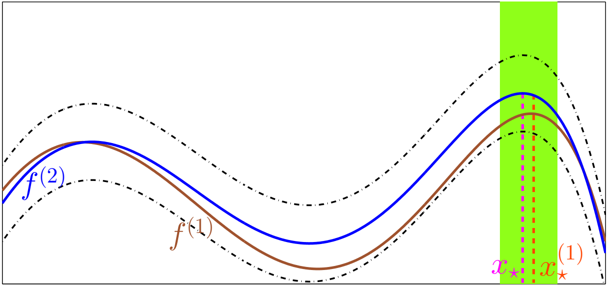
- $f^{(1)}$ is not just a noisy version of $f^{(2)}$.
- Cannot just maximise $f^{(1)}$. $x_\star^{(1)}$ is suboptimal for $f^{(2)}$.

Challenges (in 2 fidelities)



- $f^{(1)}$ 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.

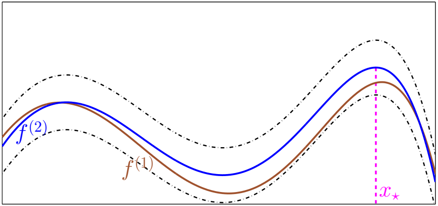
Challenges (in 2 fidelities)



- $f^{(1)}$ 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: MF-GP-UCB will explore $\mathcal{X}$ using $f^{(1)}$ and use $f^{(2)}$ *mostly* in a "good" set $\mathcal{X}_g$, determined via $f^{(1)}$.

MF-GP-UCB with 2 fidelities

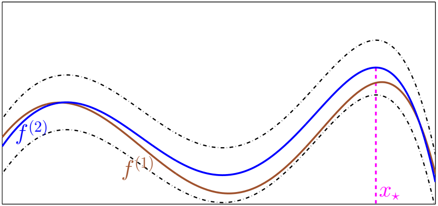


Upper Confidence Bound: Maintain 2 upper bounds for $f^{(2)}$.

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

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

MF-GP-UCB with 2 fidelities



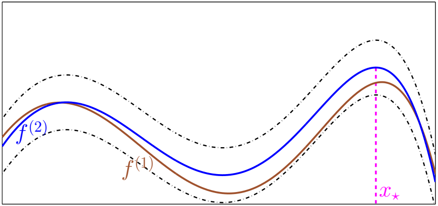
Upper Confidence Bound: Maintain 2 upper bounds for $f^{(2)}$.

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

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

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

MF-GP-UCB with 2 fidelities



Upper Confidence Bound: Maintain 2 upper bounds for $f^{(2)}$.

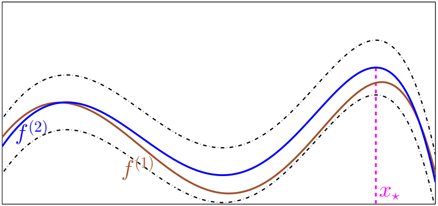
$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}$$

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

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

- Choose $\mathbf{x}_t = \mathrm{argmax}_{x\in\mathcal{X}}\, \varphi_t(x)$.

MF-GP-UCB with 2 fidelities



Upper Confidence Bound: Maintain 2 upper bounds for $f^{(2)}$.

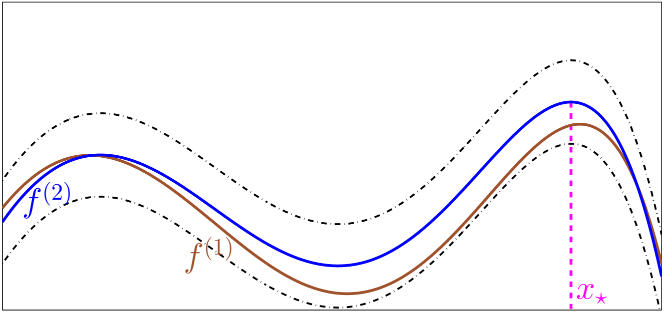
$$\varphi_t^{(1)}(x) = \mu_{t-1}^{(1)}(x) + \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) + \zeta^{(1)}$$

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

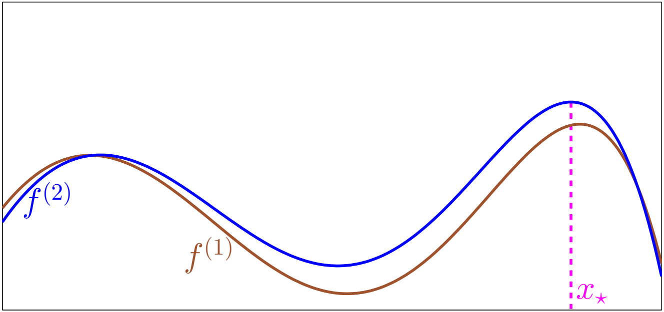
$$\varphi_t(x) = \min\{\varphi_t^{(1)}(x),\ \varphi_t^{(2)}(x)\}$$

- Choose $\mathbf{x}_t = \mathrm{argmax}_{x\in\mathcal{X}}\,\varphi_t(x)$.
- $\mathbf{m}_t = \begin{cases} 1 & \text{if } \beta_t^{1/2}\sigma_{t-1}^{(1)}(x) > \gamma^{(1)} \\ 2 & \text{otherwise.} \end{cases}$

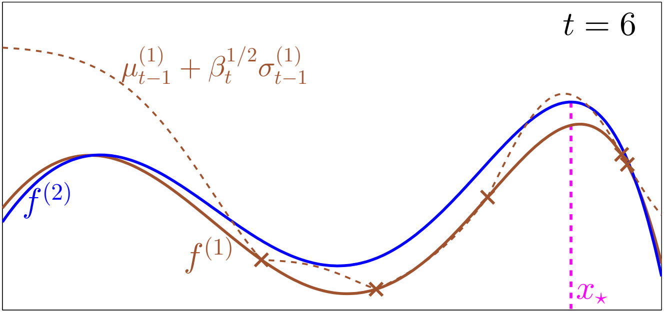
MF-GP-UCB



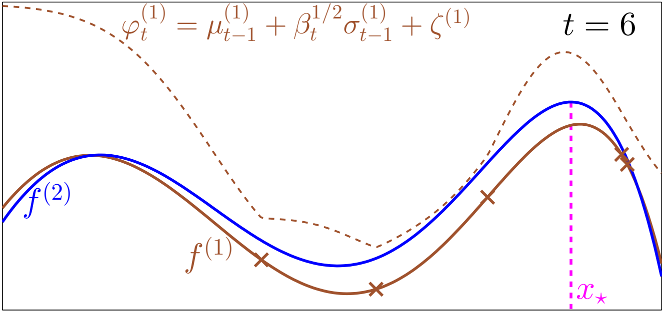
MF-GP-UCB



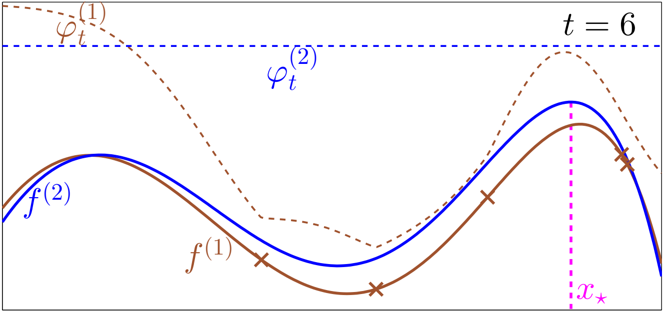
MF-GP-UCB



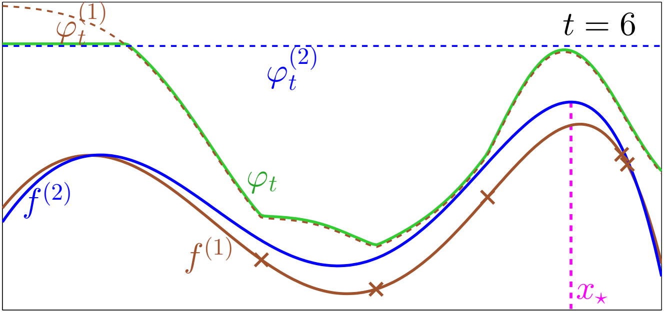
MF-GP-UCB



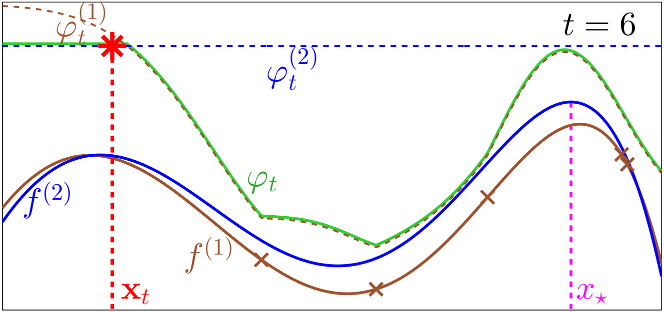
MF-GP-UCB



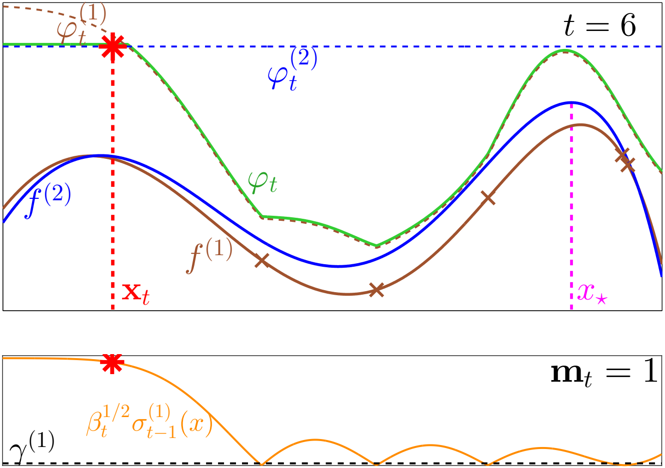
MF-GP-UCB



MF-GP-UCB

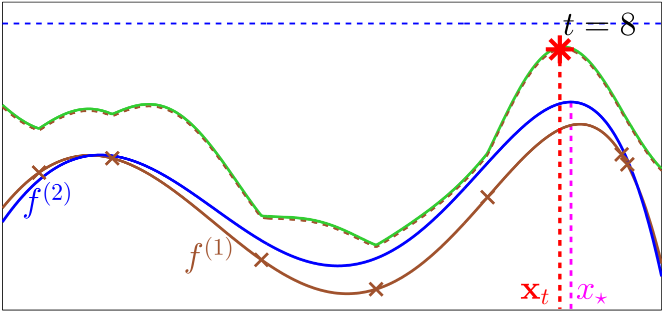


MF-GP-UCB

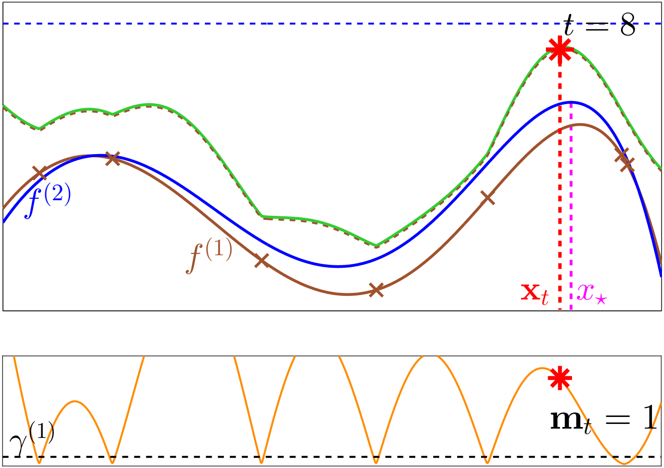


$t = 6$
$\varphi_t^{(1)}$
$\varphi_t^{(2)}$
φ_t
$f^{(2)}$
$f^{(1)}$
$\mathbf{x}_t$
$x_\star$
$\mathbf{m}_t = 1$
$\beta_t^{1/2}\sigma_{t-1}^{(1)}(x)$
$\gamma^{(1)}$

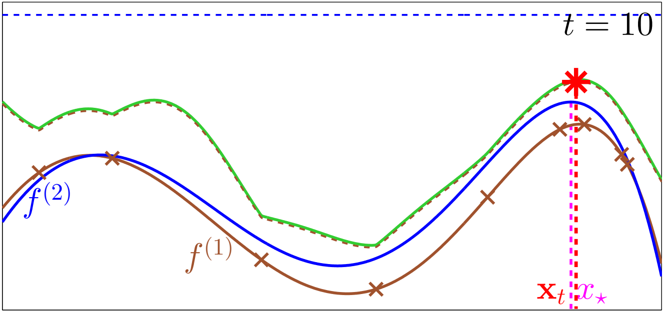
MF-GP-UCB



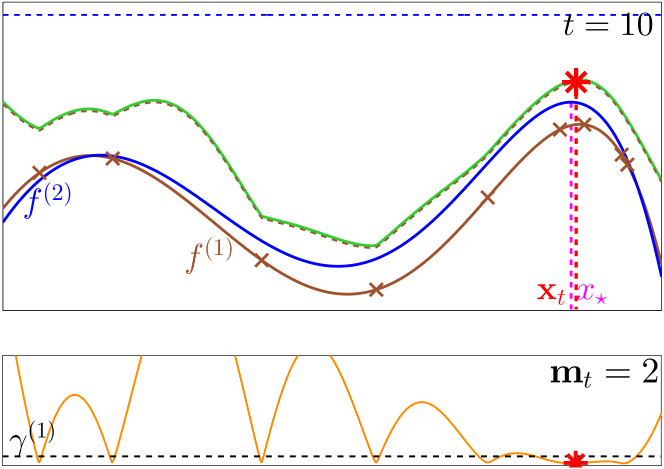
MF-GP-UCB



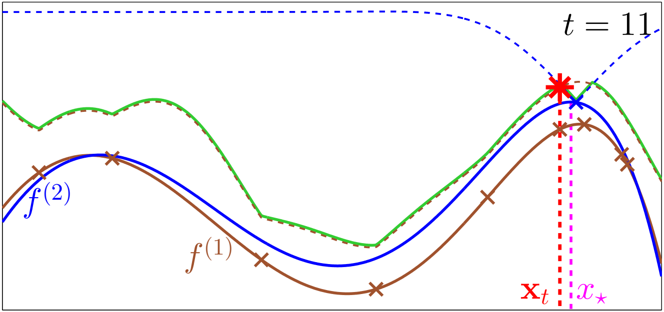
MF-GP-UCB



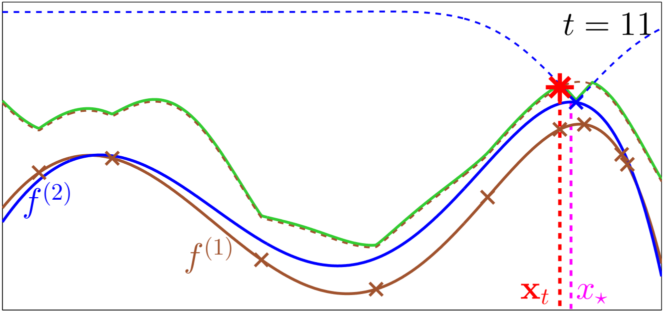
MF-GP-UCB

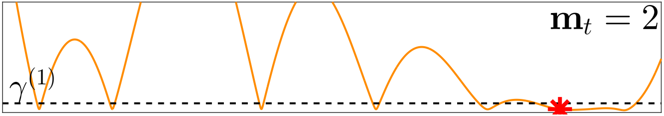


MF-GP-UCB

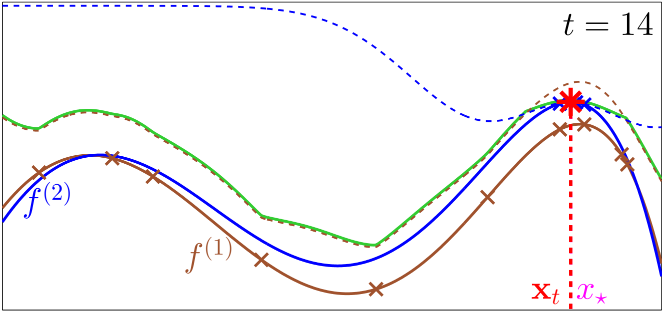


MF-GP-UCB

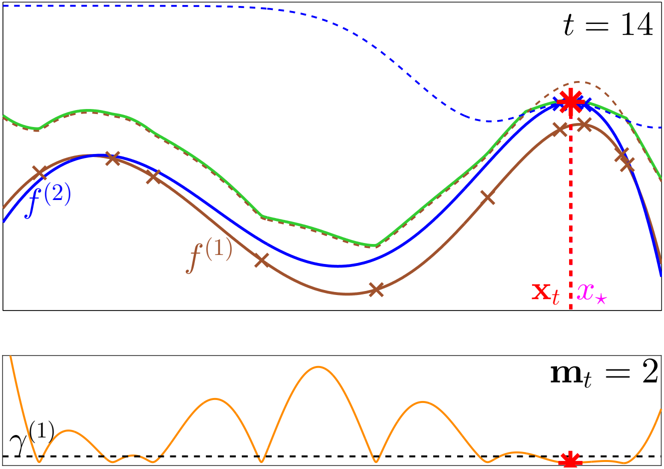




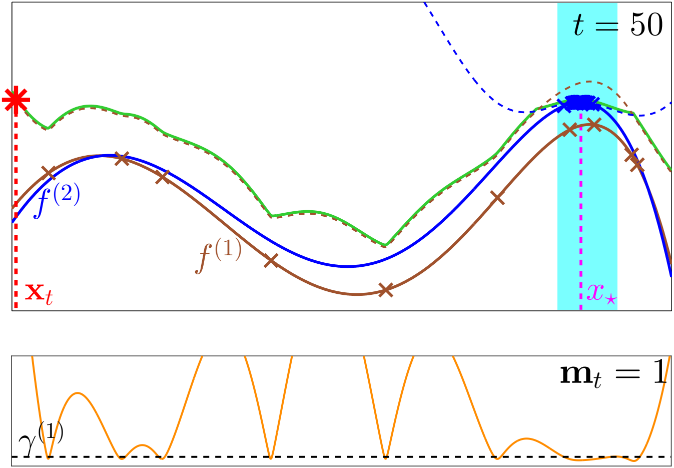
MF-GP-UCB



MF-GP-UCB



MF-GP-UCB



Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$.

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$. $\rightarrow$ $\Psi_n(A) \propto \mathrm{vol}(A)$.

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$. $\rightarrow$ $\Psi_n(A) \propto \mathrm{vol}(A)$.

GP-UCB (Srinivas et. al. 2010)

$$S(\Lambda) \lesssim \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$. $\rightarrow$ $\Psi_n(A) \propto \mathrm{vol}(A)$.

GP-UCB (Srinivas et. al. 2010)

$$\lambda^{(2)} S(\Lambda) \lesssim \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$. $\rightarrow$ $\Psi_n(A) \propto \mathrm{vol}(A)$.

GP-UCB (Srinivas et. al. 2010)

$$\lambda^{(2)} S(\Lambda) \lesssim \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

Can we achieve?

$$\lambda^{(2)} S(\Lambda) \lesssim \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g)}{n_\Lambda}} + \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g^c)}{n_\Lambda}}$$

Theoretical Results

Simple regret after *capital* Λ,

$$S(\Lambda) = f^{(2)}(x_\star) - \max_{t:\mathbf{m}_t=2} f^{(2)}(\mathbf{x}_t).$$

$n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$ is number of queries by GP-UCB within capital Λ.

$\Psi_n(A)$: Maximum Information Gain of $A \subset \mathcal{X}$. $\rightarrow$ $\Psi_n(A) \propto \mathrm{vol}(A)$.

GP-UCB (Srinivas et. al. 2010)

$$\lambda^{(2)} S(\Lambda) \;\lesssim\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X})}{n_\Lambda}}$$

Can we achieve?

$$\lambda^{(2)} S(\Lambda) \;\lesssim\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g)}{n_\Lambda}} \;+\; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g^c)}{n_\Lambda}}$$

Ideal Scenario: $\lambda^{(1)} \ll \lambda^{(2)}$ and

$$\mathrm{vol}(\mathcal{X}_g) \ll \mathrm{vol}(\mathcal{X}_g^c) \implies \Psi_{n_\Lambda}(\mathcal{X}_g) \ll \Psi_{n_\Lambda}(\mathcal{X}_g).$$

The "Good" Set $\mathcal{X}_g$

$\mathcal{X}_g$ is completely determined by $f_\star$ and $f^{(1)}$.

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

The “Good” Set $\mathcal{X}_g$

$\mathcal{X}_g$ is completely determined by $f_\star$ and $f^{(1)}$.

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

- Contains $x_\star$.
- Need not be contiguous.

The "Good" Set $\mathcal{X}_g$

$\mathcal{X}_g$ is completely determined by $f_\star$ and $f^{(1)}$.

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \le \zeta^{(1)}\}.$$

- Contains $x_\star$.
- Need not be contiguous.
- Is "fundamental" to the problem: any strategy must explore $f^{(2)}$ well within this region.
  - Lower bounds in the K-armed multi-fidelity bandit.

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g)}{n_\Lambda}} \; + \; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}_g^c)}{n_\Lambda}}$$

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \le \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \le \zeta^{(1)} + \rho\gamma\} \quad \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda}} \; + \; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}}$$

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)} + \rho\gamma\} \quad \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda}} \; + \; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}}$$

$$+ \; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda^\alpha}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda^{2-\alpha}}}$$

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)} + \rho\gamma\} \quad \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \quad \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda}} \;+\; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}}$$

$$+ \;\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda^\alpha}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda^{2-\alpha}}}$$

- Statement true for all $\alpha > 0$ for $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$.

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)} + \rho\gamma\} \;\; \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \;\; \lesssim \;\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda}} \; + \; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}}$$

$$+ \;\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda^\alpha}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda^{2-\alpha}}} \; + \; \lambda^{(1)} \frac{\mathrm{vol}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda} \frac{1}{{\gamma^{(1)}}^d}$$

- Statement true for all $\alpha > 0$ for $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$.

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \le \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \le \zeta^{(1)} + \rho\gamma\} \quad \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \quad \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho,n})}{n_\Lambda}} \;+\; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}} \;+\; \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda^\alpha}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda^{2-\alpha}}} \;+\; \lambda^{(1)} \frac{\mathrm{vol}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda} \frac{1}{{\gamma^{(1)}}^d}$$

- Statement true for all $\alpha > 0$ for $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$.
- $\widetilde{\mathcal{X}}_{g,\rho,n} \to \widetilde{\mathcal{X}}_{g,\rho}$ as $n \to \infty$.

Theoretical Results

$$\mathcal{X}_g = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)}\}.$$

We will consider a slightly inflated set.

$$\widetilde{\mathcal{X}}_{g,\rho} = \{x \in \mathcal{X} : f_\star - f^{(1)}(x) \leq \zeta^{(1)} + \rho\gamma\} \quad \supset \mathcal{X}_g.$$

Theorem (Simple Regret for MF-GP-UCB):

$$\lambda^{(2)} S(\Lambda) \quad \lesssim \quad \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda}} \; + \; \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda}}$$

$$+ \; \lambda^{(2)} \sqrt{\frac{\Psi_{n^\alpha_\Lambda}(\widetilde{\mathcal{X}}^c_{g,\rho})}{n_\Lambda^{2-\alpha}}} \; + \; \lambda^{(1)} \frac{\mathrm{vol}(\widetilde{\mathcal{X}}_{g,\rho})}{n_\Lambda} \frac{1}{{\gamma^{(1)}}^d}$$

- Statement true for all $\alpha > 0$ for $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$.
- $\widetilde{\mathcal{X}}_{g,\rho,n} \to \widetilde{\mathcal{X}}_{g,\rho}$ as $n \to \infty$.

Proof Sketch

$N \leftarrow$ Number of plays by MF-GP-UCB within capital Λ.

Proof Sketch

$N \leftarrow$ Number of plays by MF-GP-UCB within capital Λ.
Since $\lambda^{(1)} < \lambda^{(2)}$, N could be much larger than $n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

Proof Sketch

$N \leftarrow$ Number of plays by MF-GP-UCB within capital Λ.
Since $\lambda^{(1)} < \lambda^{(2)}$, N could be much larger than $n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

But .. we show $N \leq 2n_\Lambda$ with high probability.

Proof Sketch

$N \leftarrow$ Number of plays by MF-GP-UCB within capital Λ.
Since $\lambda^{(1)} < \lambda^{(2)}$, N could be much larger than $n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

But .. we show $N \leq 2n_\Lambda$ with high probability.

We need to bound the following 4 quantities.

- $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho})$: # of second fidelity queries in $\widetilde{\mathcal{X}}_{g,\rho}$.
- $T_N^{(2)}(\widetilde{\mathcal{X}}^c_{g,\rho})$: # of second fidelity queries in $\widetilde{\mathcal{X}}^c_{g,\rho}$.
- $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$, $T_N^{(1)}(\widetilde{\mathcal{X}}^c_{g,\rho})$.

Proof Sketch

$N \leftarrow$ Number of plays by MF-GP-UCB within capital Λ.
Since $\lambda^{(1)} < \lambda^{(2)}$, N could be much larger than $n_\Lambda = \lfloor \Lambda/\lambda^{(2)} \rfloor$.

But .. we show $N \leq 2n_\Lambda$ with high probability.

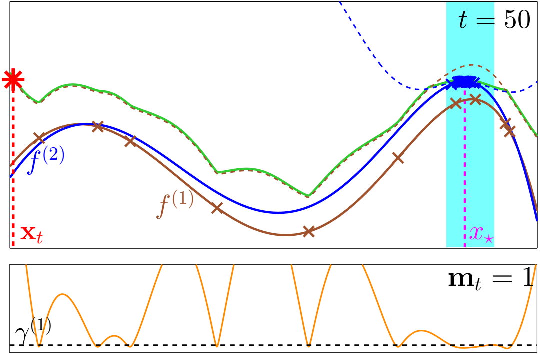
We need to bound the following 4 quantities.

- $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho})$: # of second fidelity queries in $\widetilde{\mathcal{X}}_{g,\rho}$.
- $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$: # of second fidelity queries in $\widetilde{\mathcal{X}}_{g,\rho}^c$.
- $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$, $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$.

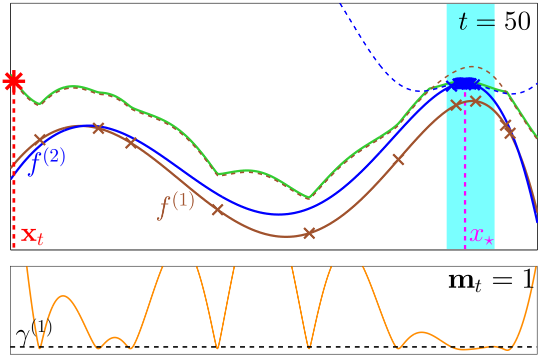
We will use, $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$, $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}) \leq N$. Gives us

$$\lambda^{(2)}\sqrt{\frac{\Psi_N(\widetilde{\mathcal{X}}_{g,\rho})}{N}} \; + \; \lambda^{(1)}\sqrt{\frac{\Psi_N(\widetilde{\mathcal{X}}_{g,\rho}^c)}{N}}$$

Proof Sketch: Bounding $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$

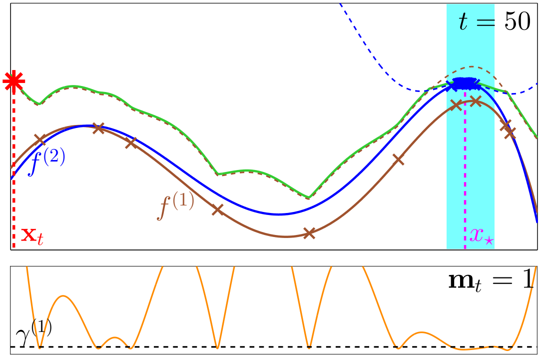


Proof Sketch: Bounding $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$



$$\mathbb{P}\left(T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c) > N^\alpha \right) < \text{something small}$$

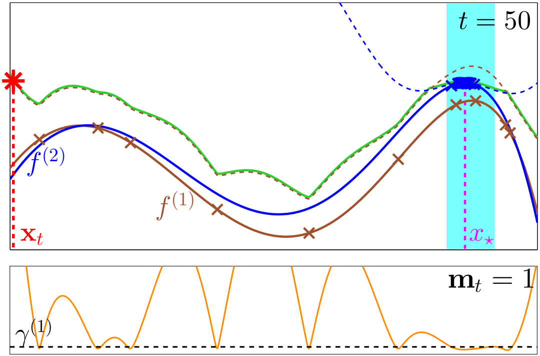
Proof Sketch: Bounding $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$



$$\mathbb{P}\left(T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c) \; > \; N^\alpha \right) \; < \; \text{something small}$$

Holds for all $\alpha > 0$ if $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$. This result is *strong*.

Proof Sketch: Bounding $T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c)$

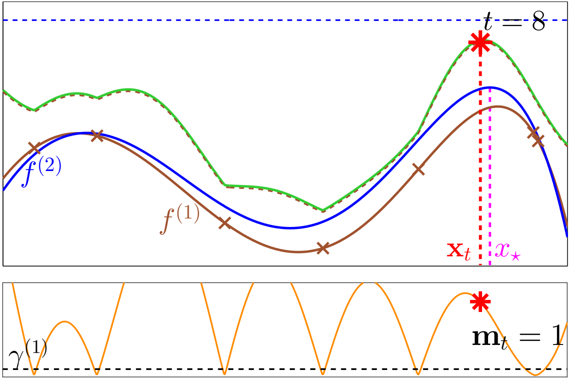


$$\mathbb{P}\left(T_N^{(2)}(\widetilde{\mathcal{X}}_{g,\rho}^c) \; > \; N^\alpha \right) \;\; < \;\; \text{something small}$$

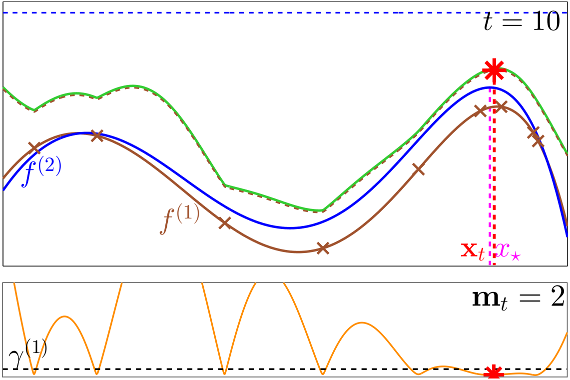
Holds for all $\alpha > 0$ if $\rho \asymp 1 + \frac{1}{\sqrt{\alpha}}$. This result is *strong*.

This gives us the third term $\lambda^{(2)} \sqrt{\frac{\Psi_{N^\alpha}(\widetilde{\mathcal{X}}_{g,\rho}^c)}{N^{2-\alpha}}}$.

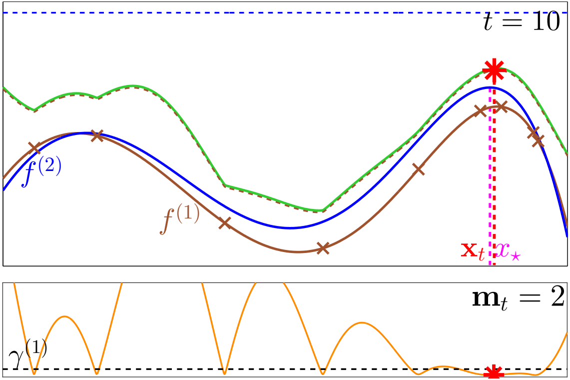
Proof Sketch: Bounding $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$



Proof Sketch: Bounding $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$

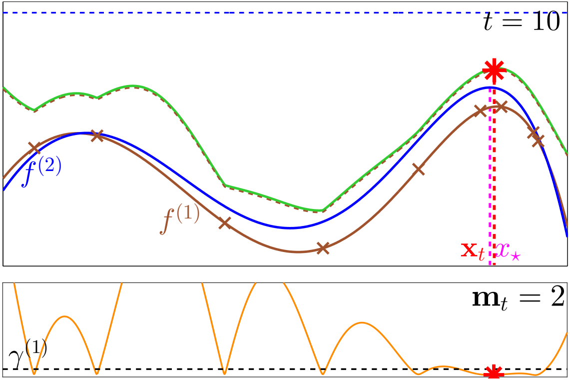


Proof Sketch: Bounding $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$



$T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$ cannot be large due to the switching criterion. Proof uses a covering argument and bounds on the GP posterior variance.

Proof Sketch: Bounding $T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$



$T_N^{(1)}(\widetilde{\mathcal{X}}_{g,\rho})$ cannot be large due to the switching criterion. Proof uses a covering argument and bounds on the GP posterior variance.

This gives us the last term $\lambda^{(1)} \frac{\mathrm{vol}(\widetilde{\mathcal{X}}_{g,\rho})}{N} \frac{1}{{\gamma^{(1)}}^d}$

MF-GP-UCB with M fidelities

Setting: $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.

MF-GP-UCB with M fidelities

Setting: $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.

MF-GP-UCB:

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

MF-GP-UCB with M fidelities

Setting: $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.

MF-GP-UCB:

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

$$\varphi_t(x) = \min_{m=1,\ldots,M} \varphi_t^{(m)}(x)$$

MF-GP-UCB with M fidelities

Setting: $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.

MF-GP-UCB:

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

$$\varphi_t(x) = \min_{m=1,\ldots,M} \varphi_t^{(m)}(x)$$

- Choose $\mathbf{x}_t = \mathrm{argmax}_{x\in\mathcal{X}}\, \varphi_t(x)$.

MF-GP-UCB with M fidelities

Setting: $\|f^{(M)} - f^{(m)}\|_\infty \leq \zeta^{(m)}$ for all $m = 1, \ldots, M-1$.

MF-GP-UCB:

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

$$\varphi_t(x) = \min_{m=1,\ldots,M} \varphi_t^{(m)}(x)$$

- Choose $\mathbf{x}_t = \operatorname{argmax}_{x\in\mathcal{X}} \varphi_t(x)$.
- Choosing $\mathbf{m}_t$:

  for $m = 1, \ldots, M$:

  if $\beta_t^{1/2}\sigma_{t-1}^{(m)}(\mathbf{x}_t) > \gamma^{(m)}$, **break**;

  $\mathbf{m}_t = m$.

Regret Bound: MF-GP-UCB with M fidelities

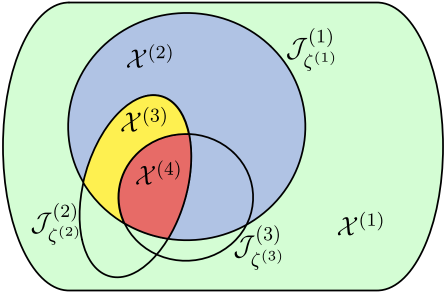
"Ideal" Bound:

$$\lambda^{(M)} S(\Lambda) \;\lesssim\; \lambda^{(M)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(M)})}{n_\Lambda}} + \ldots + \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(2)})}{n_\Lambda}} + \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(1)})}{n_\Lambda}}$$

Regret Bound: MF-GP-UCB with M fidelities

"Ideal" Bound:

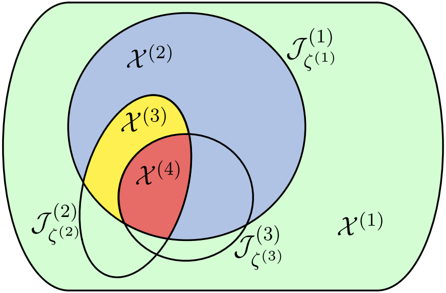
$$\lambda^{(M)} S(\Lambda) \lesssim \lambda^{(M)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(M)})}{n_\Lambda}} + \ldots + \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(2)})}{n_\Lambda}} + \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(1)})}{n_\Lambda}}$$



Regret Bound: MF-GP-UCB with M fidelities

"Ideal" Bound:

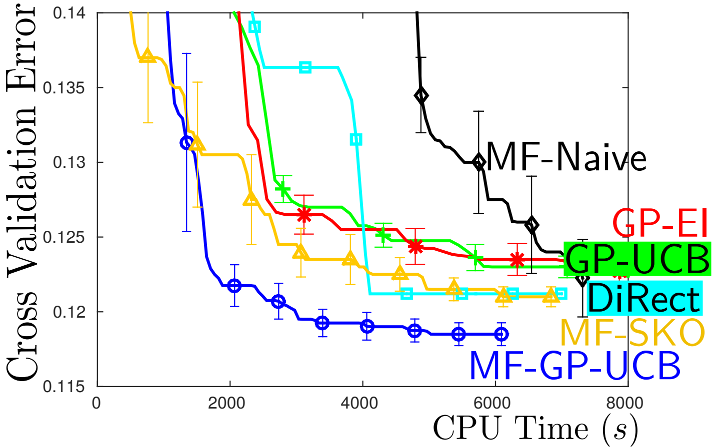
$$\lambda^{(M)} S(\Lambda) \;\lesssim\; \lambda^{(M)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(M)})}{n_\Lambda}} + \ldots + \lambda^{(2)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(2)})}{n_\Lambda}} + \lambda^{(1)} \sqrt{\frac{\Psi_{n_\Lambda}(\mathcal{X}^{(1)})}{n_\Lambda}}$$



Theorem: Similar to above but contains $\gamma^{(m)}$ dependent inflations and other subdominant terms as in the two fidelity setting.

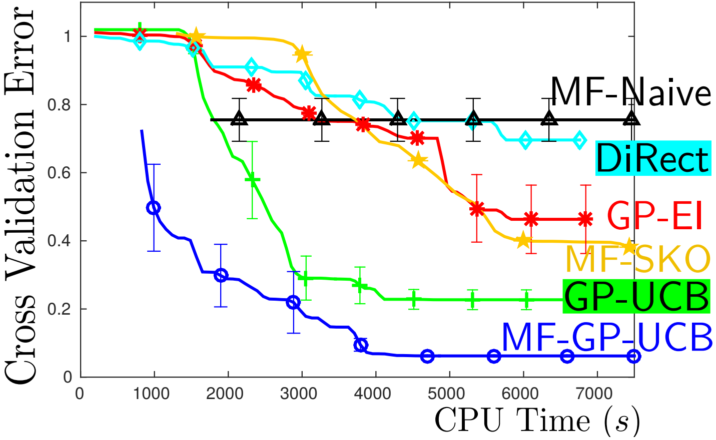
Experiment: Support Vector Classification

2 hyper-parameters, 2 fidelities ($n_{tr} = \{500, 2000\}$)



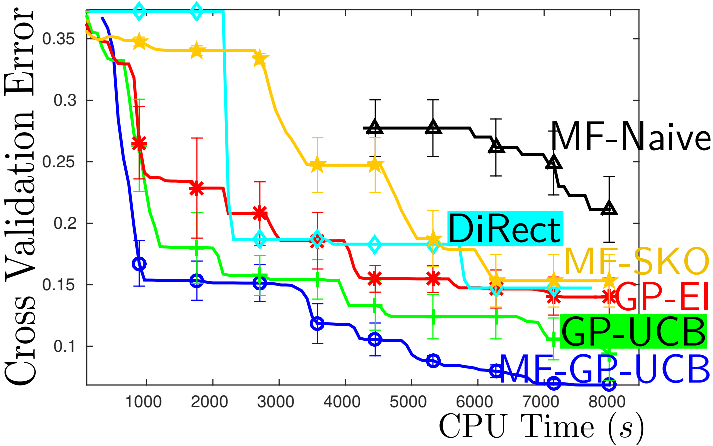
Experiment: SALSA

6 hyper-parameters, 3 fidelities ($n_{tr} = \{2000, 4000, 8000\}$)



Experiment: Viola & Jones Face Detection

22 hyper-parameters, 2 fidelities ($n_{tr} = \{300, 3000\}$)

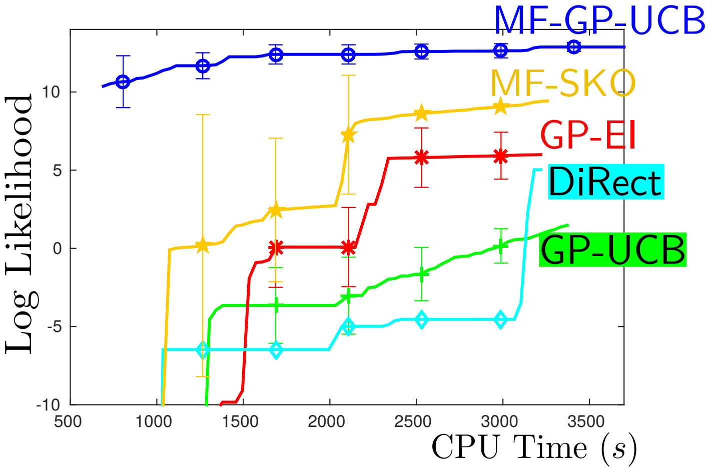


Experiment: Cosmological Maximum Likelihood Inference

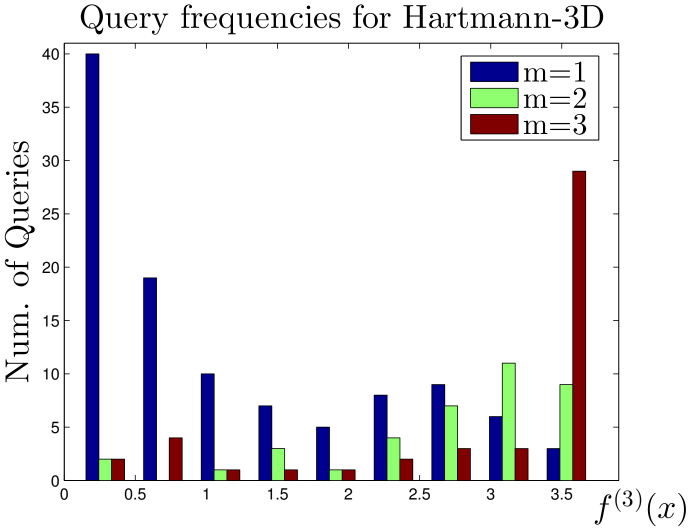
- Type Ia Supernovae Data
- Maximum likelihood inference for 3 cosmological parameters:
  - Hubble Constant H_0
  - Dark Energy Fraction Ω_Λ
  - Dark Matter Fraction Ω_M
- Likelihood: Robertson Walker metric
  Requires numerical integration for each point in the dataset.

Experiment: Cosmological Maximum Likelihood Inference

3 cosmological parameters, 3 fidelities ($\mathrm{grid} = \{10^2, 10^4, 10^6\}$)



Synthetic Experiment: Hartmann-3D



Summary

- A novel framework and algorithm for Multi-fidelity Bandit Optimisation.
- MF-GP-UCB: intuitive algorithm using UCB principles.

Summary

- A novel framework and algorithm for Multi-fidelity Bandit Optimisation.
- MF-GP-UCB: intuitive algorithm using UCB principles.
- Theoretical Results
  - Lower fidelities are used to eliminate bad regions.
  - Higher fidelities are used in successively smaller regions.

Summary

- A novel framework and algorithm for Multi-fidelity Bandit Optimisation.
- MF-GP-UCB: intuitive algorithm using UCB principles.
- Theoretical Results
  - Lower fidelities are used to eliminate bad regions.
  - Higher fidelities are used in successively smaller regions.
- Outperforms naive strategies and other multi-fidelity methods in practice.

Collaborators

Gautam Dasarathy

Junier Oliva

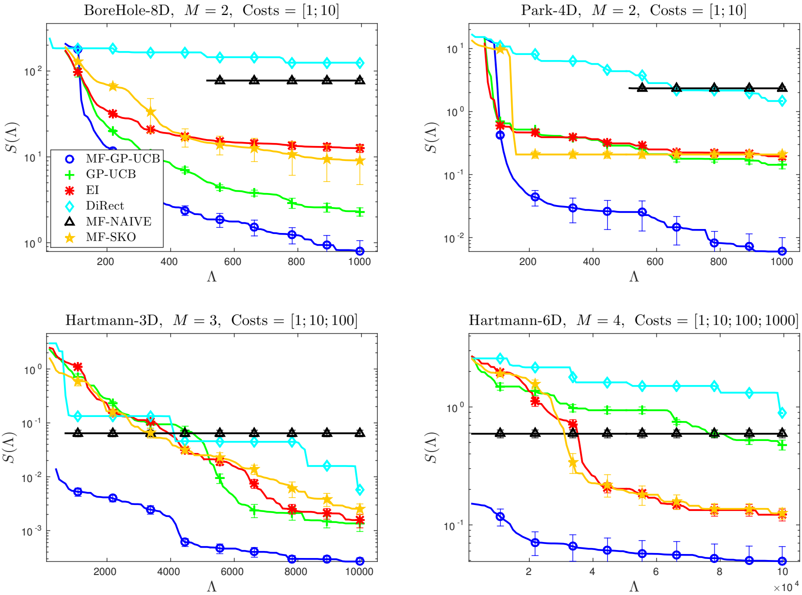
Jeff Schneider

Barnabas Poczos

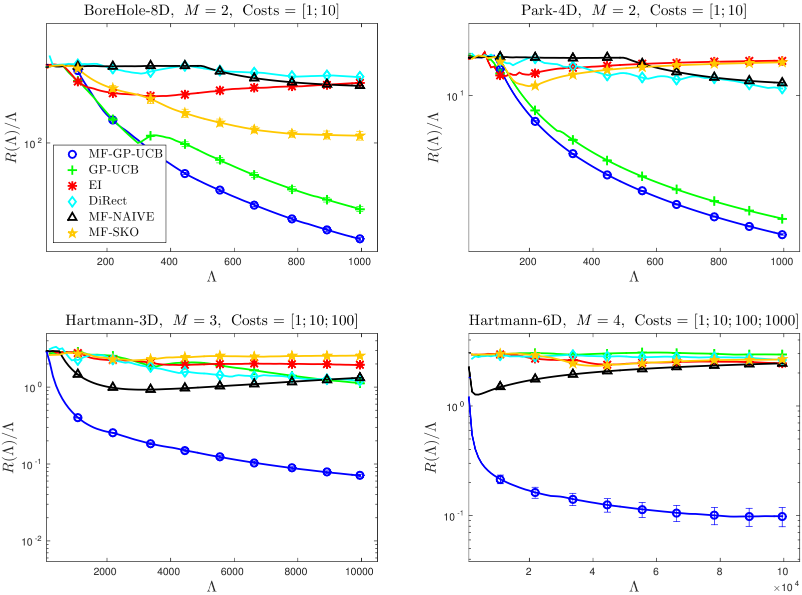
Thank you.

Paper and slides are up on my website.
Code will be up online soon.

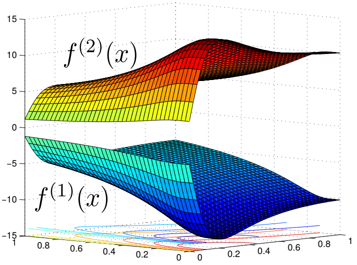
Appendix: Simple Regret

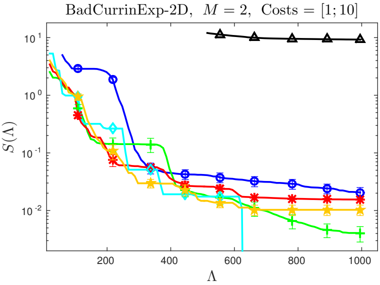


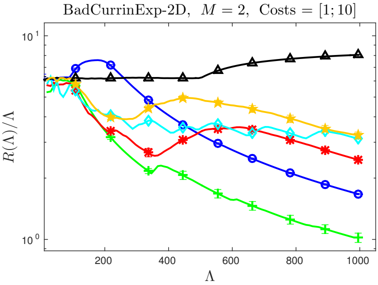
Appendix: Cumulative Regret



Appendix: Bad Approximations







Appendix: Cumulative Regret Definition

$$\text{Instantaneous Reward} \quad q_t = \begin{cases} -B & \text{if } \mathbf{m}_t \neq M \\ f^{(M)}(\mathbf{x}_t) & \text{if } \mathbf{m}_t = M \end{cases}$$

$$\text{Instantaneous Regret} \quad r_t = f_\star - q_t \ = \begin{cases} f_\star - B & \text{if } \mathbf{m}_t \neq M \\ f_\star - f^{(M)}(\mathbf{x}_t) & \text{if } \mathbf{m}_t = M \end{cases}$$

$$\begin{aligned} R(\Lambda) \ &= \ \Lambda f_\star - \left[\sum_{t=1}^{N} \lambda^{(m_t)} q_t + \left(\Lambda - \sum_{t=1}^{N} \lambda^{(m_t)} \right) (-B) \right] \\ &\leq \ 2B \underbrace{\left(\Lambda - \sum_{t=1}^{N} \lambda^{(m_t)} \right)}_{\Lambda_{res}} + \sum_{t=1}^{N} \lambda^{(m_t)} r_t \end{aligned}$$