What do you Mean? The Role of the Mean Function in Bayesian Optimisation

George De Ath,
Richard Everson and Jonathan Fieldsend

University of Exeter, United Kingdom

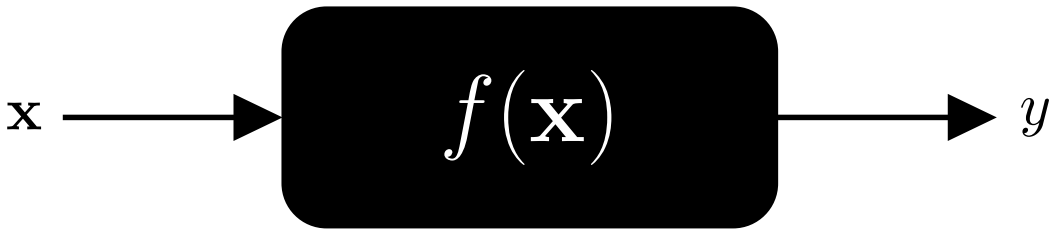
Innovate UK

UNIVERSITY OF EXETER

Global Optimisation of Black-Box Functions

In global optimisation we wish to optimise a function $f : \mathcal{X} \mapsto \mathbb{R}$, defined on a compact domain $\mathcal{X} \subseteq \mathbb{R}^D$:

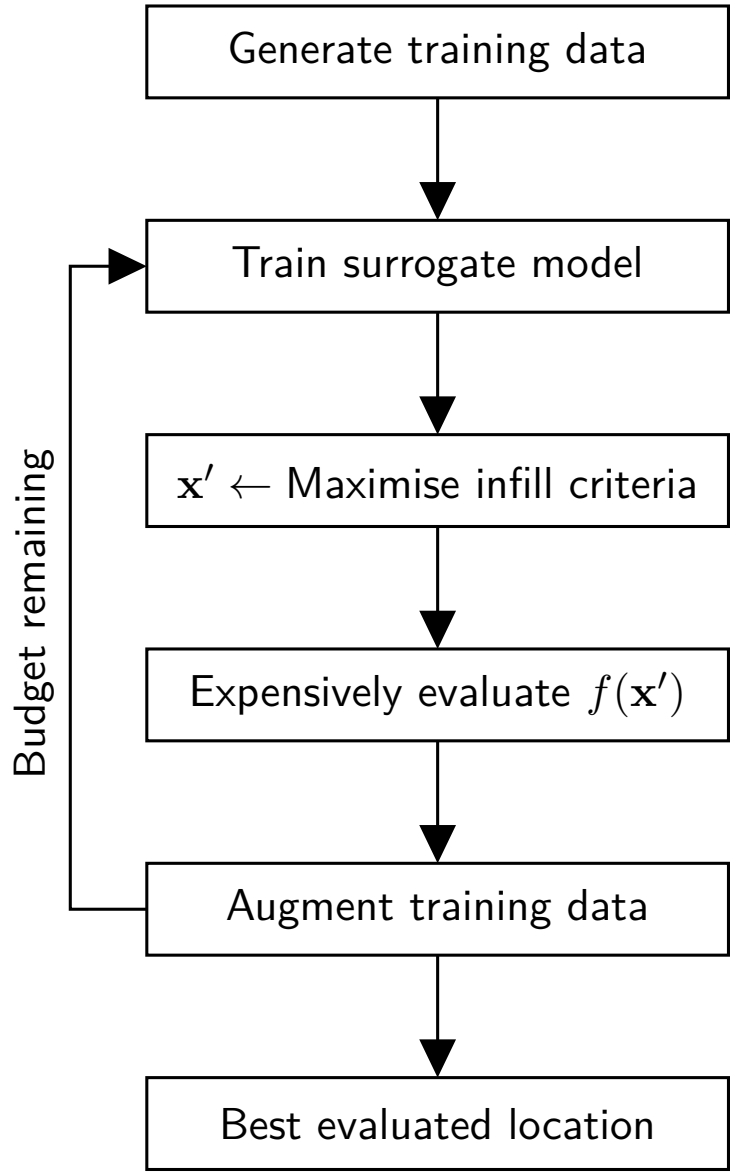
$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmin}}\, f(\mathbf{x})$$



- The functional form of f is *unknown* (i.e. black-box).
- f is assumed to be *non-convex*.
- Evaluations of f are *expensive* in terms of time and/or money.

Bayesian Optimisation

- Widely used in both academia and industry for expensive global black-box optimisation.
- Uses a probabilistic *surrogate model* of the function, created using previously evaluated locations.
- The next location to evaluate is chosen by maximising an *acquisition function* or *infill criterion* that balances exploitation and exploration.



Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

- GP prior over the function: $f \sim \mathcal{GP}(m(\mathbf{x}),\, \kappa(\mathbf{x}, \mathbf{x}'))$.

Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

- GP prior over the function: $f \sim \mathcal{GP}(m(\mathbf{x}),\, \kappa(\mathbf{x}, \mathbf{x}'))$.
- Given some observations $\mathcal{D} = \{(\mathbf{x}_n,\, f(\mathbf{x}_n))\}_{n=1}^{t}$,

Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

- GP prior over the function: $f \sim \mathcal{GP}(m(\mathbf{x}),\, \kappa(\mathbf{x}, \mathbf{x}'))$.
- Given some observations $\mathcal{D} = \{(\mathbf{x}_n,\, f(\mathbf{x}_n))\}_{n=1}^{t}$,
- The posterior distribution of f at any location $\mathbf{x}$ is Gaussian:

$$p(f \,|\, \mathbf{x}, \mathcal{D}) = \mathcal{N}(f \,|\, \mu(\mathbf{x}),\, \sigma^2(\mathbf{x})),$$

with

$$\mu(\mathbf{x} \,|\, \mathcal{D}) = m(\mathbf{x}) + \boldsymbol{\kappa}(\mathbf{x}, X)K^{-1}(\mathbf{f} - \mathbf{m})$$
$$\sigma^2(\mathbf{x} \,|\, \mathcal{D}) = \kappa(\mathbf{x}, \mathbf{x}) - \boldsymbol{\kappa}(\mathbf{x}, X)^\top K^{-1}\kappa(X, \mathbf{x}).$$

Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

- GP prior over the function: $f \sim \mathcal{GP}(m(\mathbf{x}),\, \kappa(\mathbf{x}, \mathbf{x}'))$.
- Given some observations $\mathcal{D} = \{(\mathbf{x}_n,\, f(\mathbf{x}_n))\}_{n=1}^{t}$,
- The posterior distribution of f at any location $\mathbf{x}$ is Gaussian:

$$p(f \,|\, \mathbf{x}, \mathcal{D}) = \mathcal{N}(f \,|\, \mu(\mathbf{x}),\, \sigma^2(\mathbf{x})),$$

with

$$\mu(\mathbf{x} \,|\, \mathcal{D}) = m(\mathbf{x}) + \boldsymbol{\kappa}(\mathbf{x}, X) K^{-1}(\mathbf{f} - \mathbf{m})$$
$$\sigma^2(\mathbf{x} \,|\, \mathcal{D}) = \kappa(\mathbf{x}, \mathbf{x}) - \boldsymbol{\kappa}(\mathbf{x}, X)^\top K^{-1} \kappa(X, \mathbf{x}).$$

Surrogate Model: Gaussian Processes

These are the surrogate model of choice in Bayesian Optimisation, due to their strength in function approximation and uncertainty quantification.

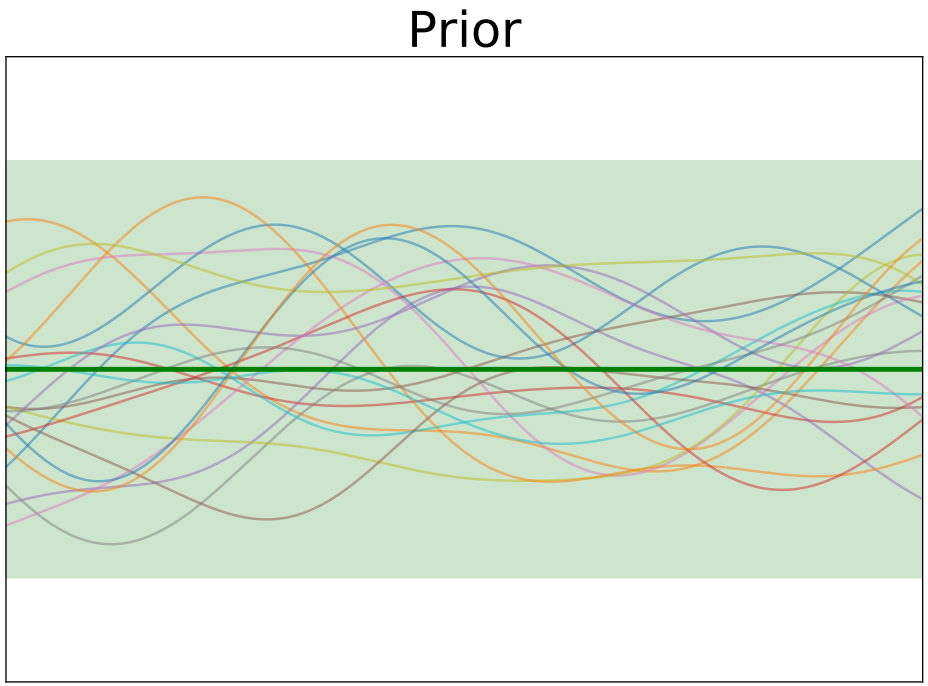
- GP prior over the function: $f \sim \mathcal{GP}(m(\mathbf{x}),\, \kappa(\mathbf{x}, \mathbf{x}'))$.
- Given some observations $\mathcal{D} = \{(\mathbf{x}_n,\, f(\mathbf{x}_n))\}_{n=1}^{t}$,
- The posterior distribution of f at any location $\mathbf{x}$ is Gaussian:

$$p(f \,|\, \mathbf{x}, \mathcal{D}) = \mathcal{N}(f \,|\, \mu(\mathbf{x}),\, \sigma^2(\mathbf{x})),$$

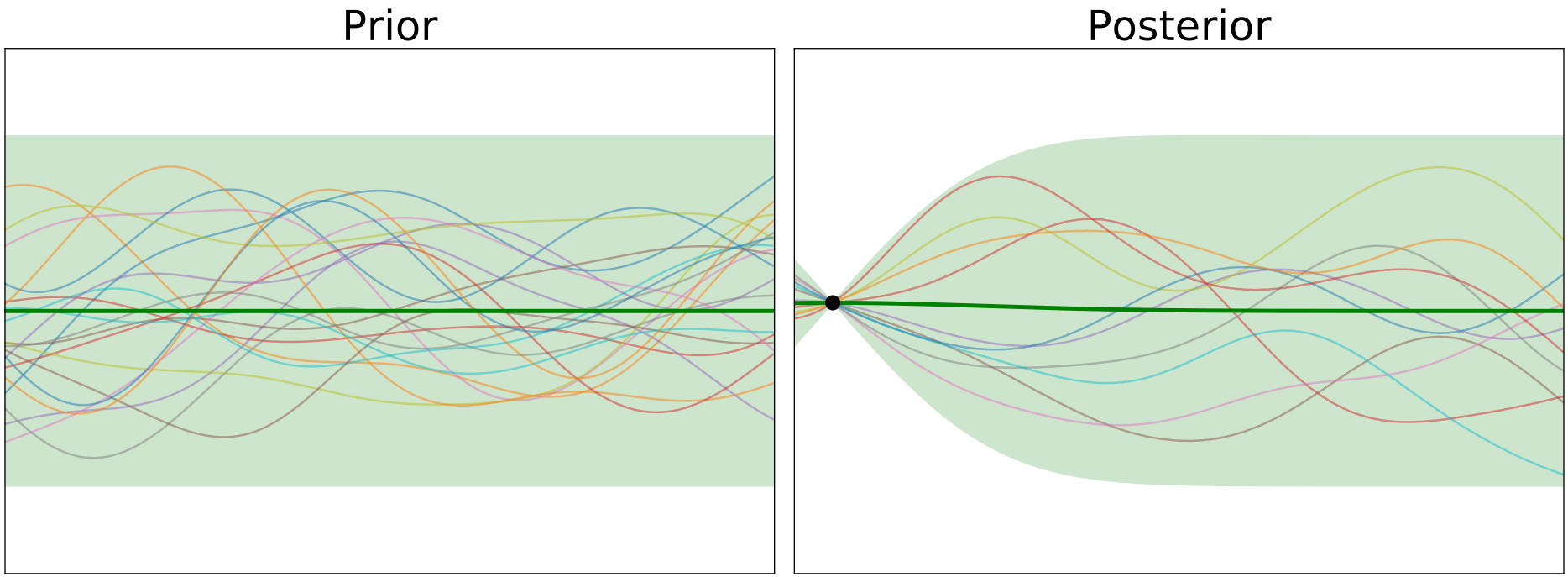
with

$$\mu(\mathbf{x} \,|\, \mathcal{D}) \approx m(\mathbf{x})$$
$$\sigma^2(\mathbf{x} \,|\, \mathcal{D}) \approx \kappa(\mathbf{x}, \mathbf{x}).$$

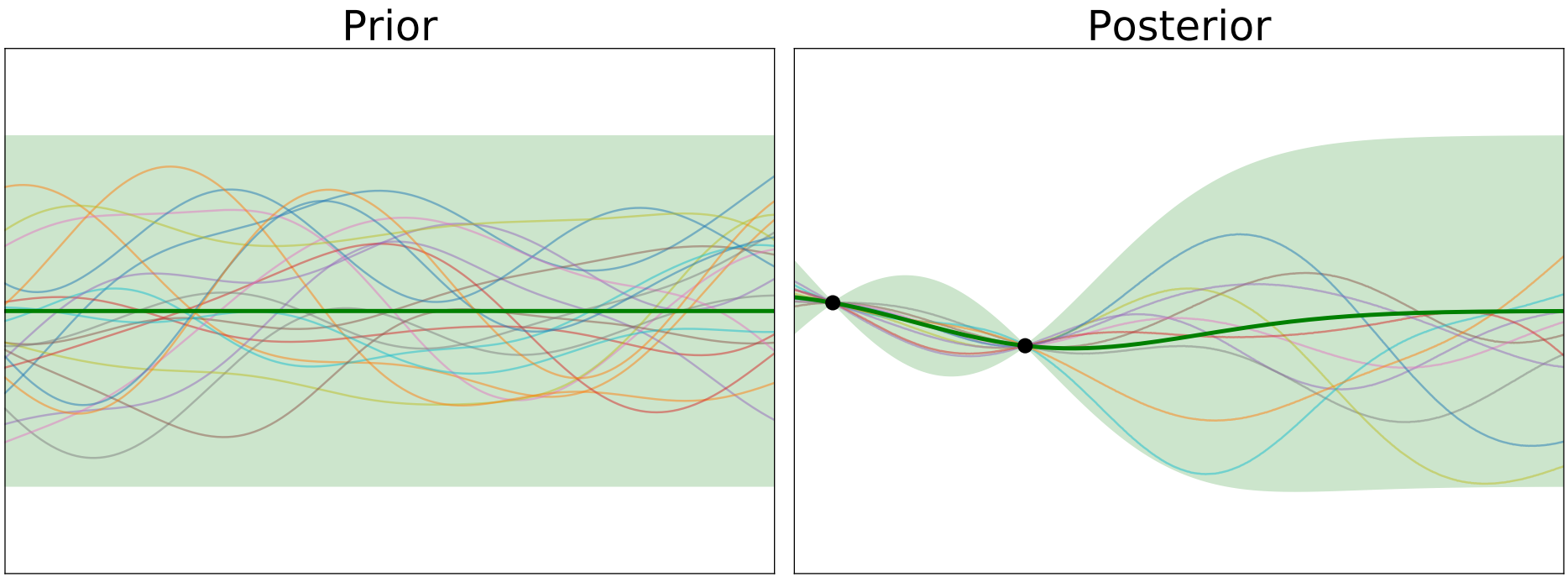
Surrogate Model: Gaussian Processes



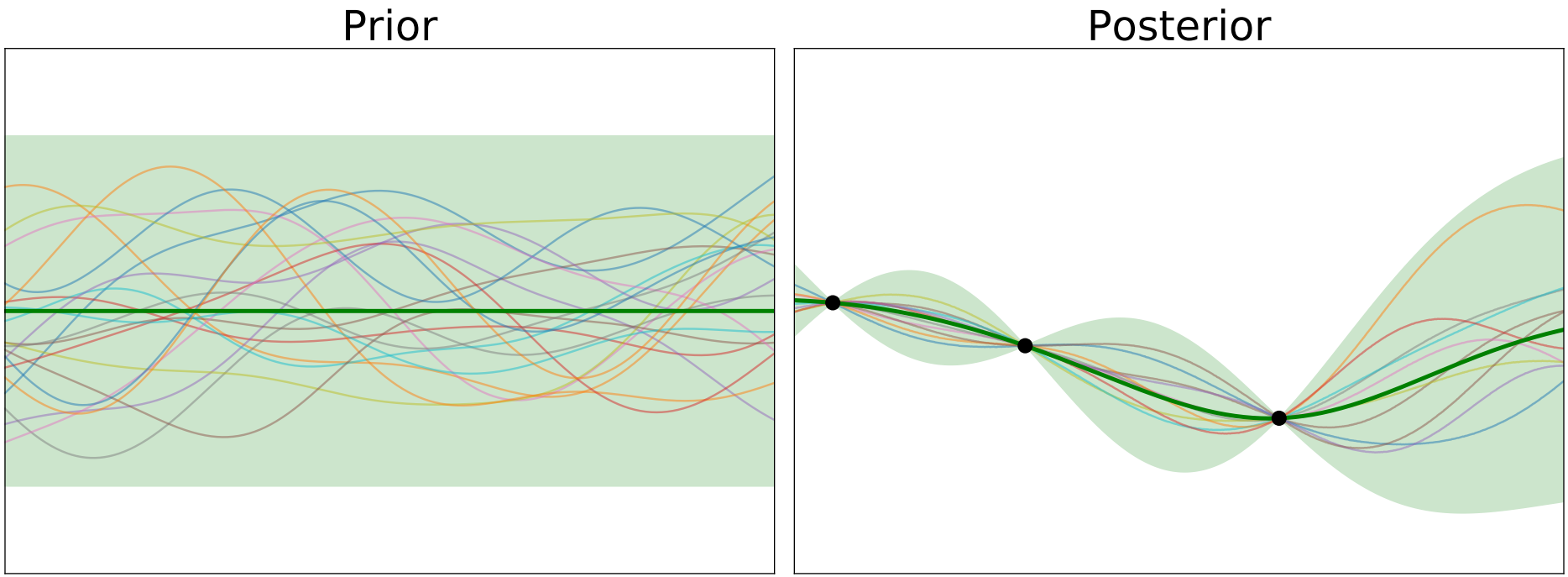
Surrogate Model: Gaussian Processes



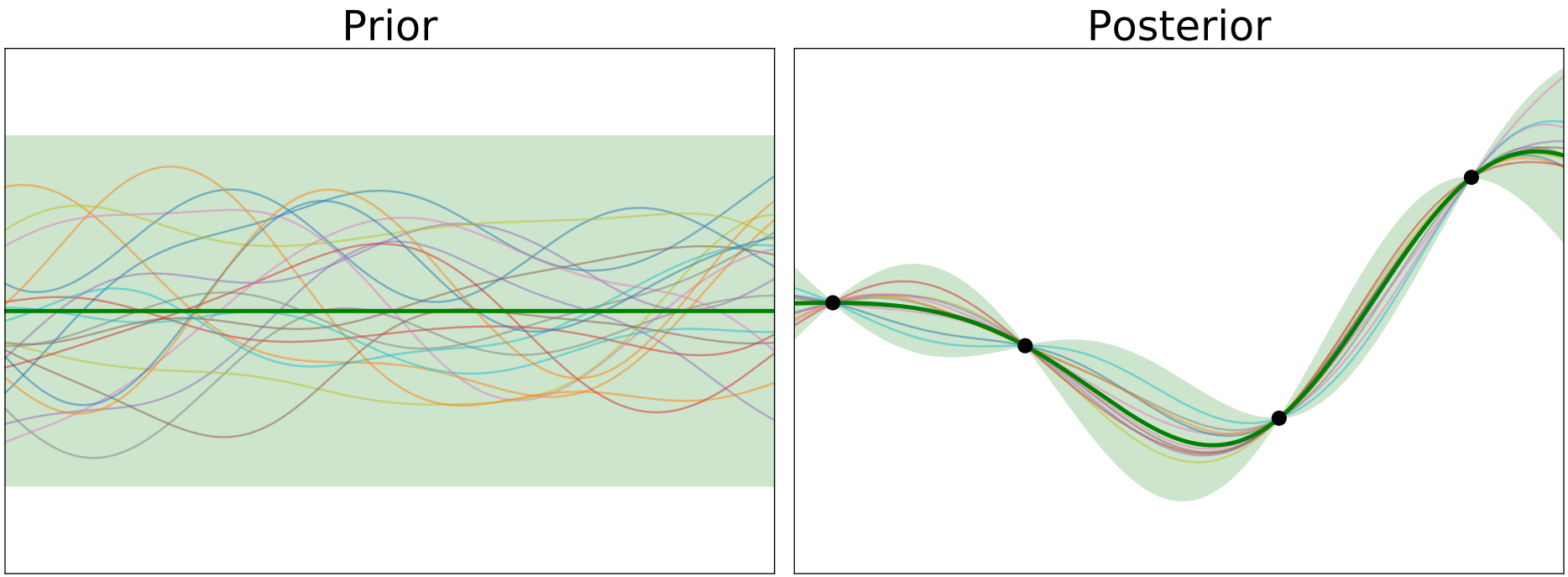
Surrogate Model: Gaussian Processes



Surrogate Model: Gaussian Processes

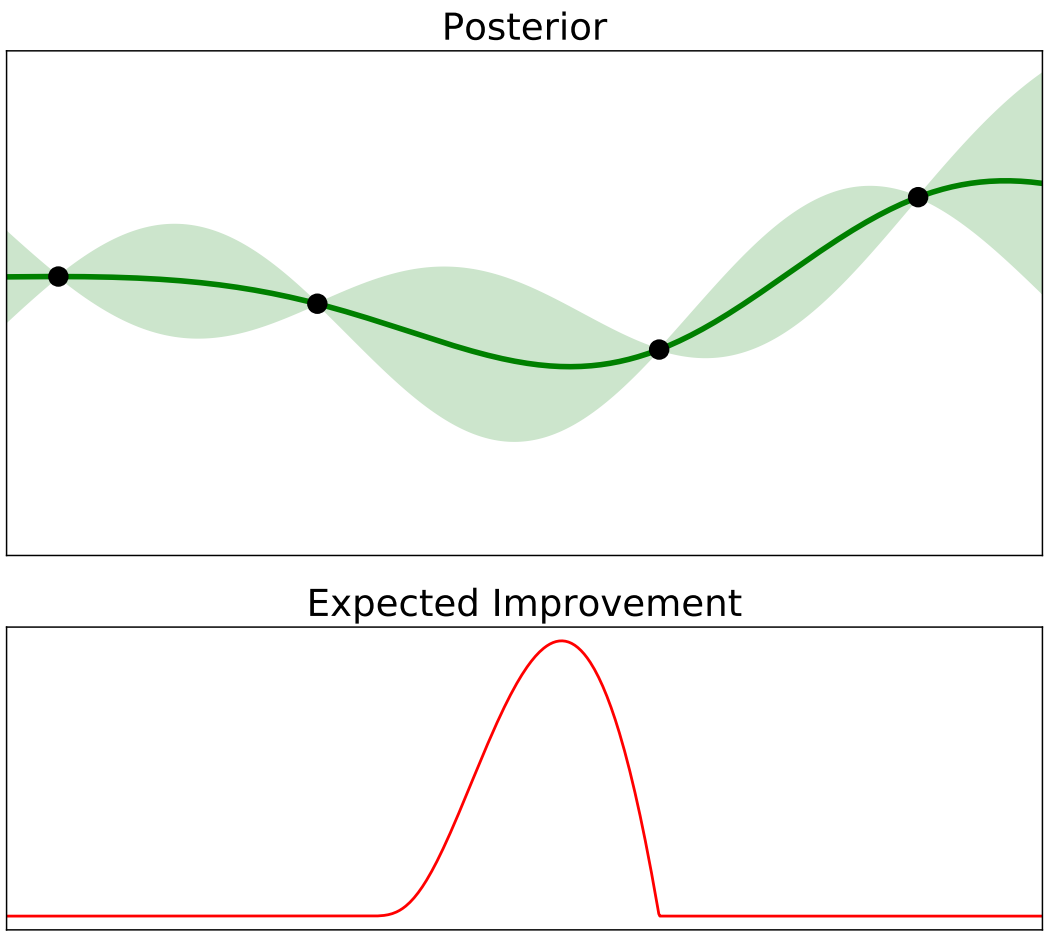


Surrogate Model: Gaussian Processes



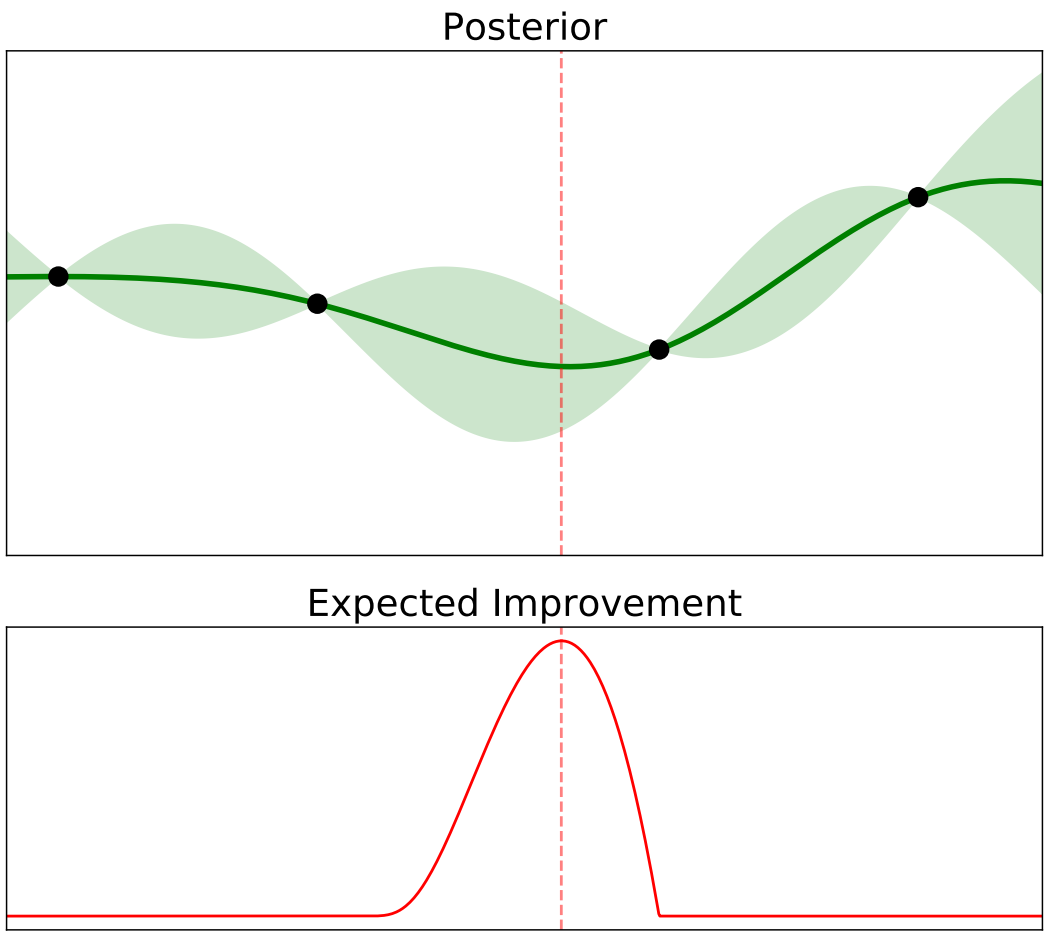
Acquisition Functions

These combine the surrogate model's predictions and their uncertainty to strike a balance between exploitation and exploration.



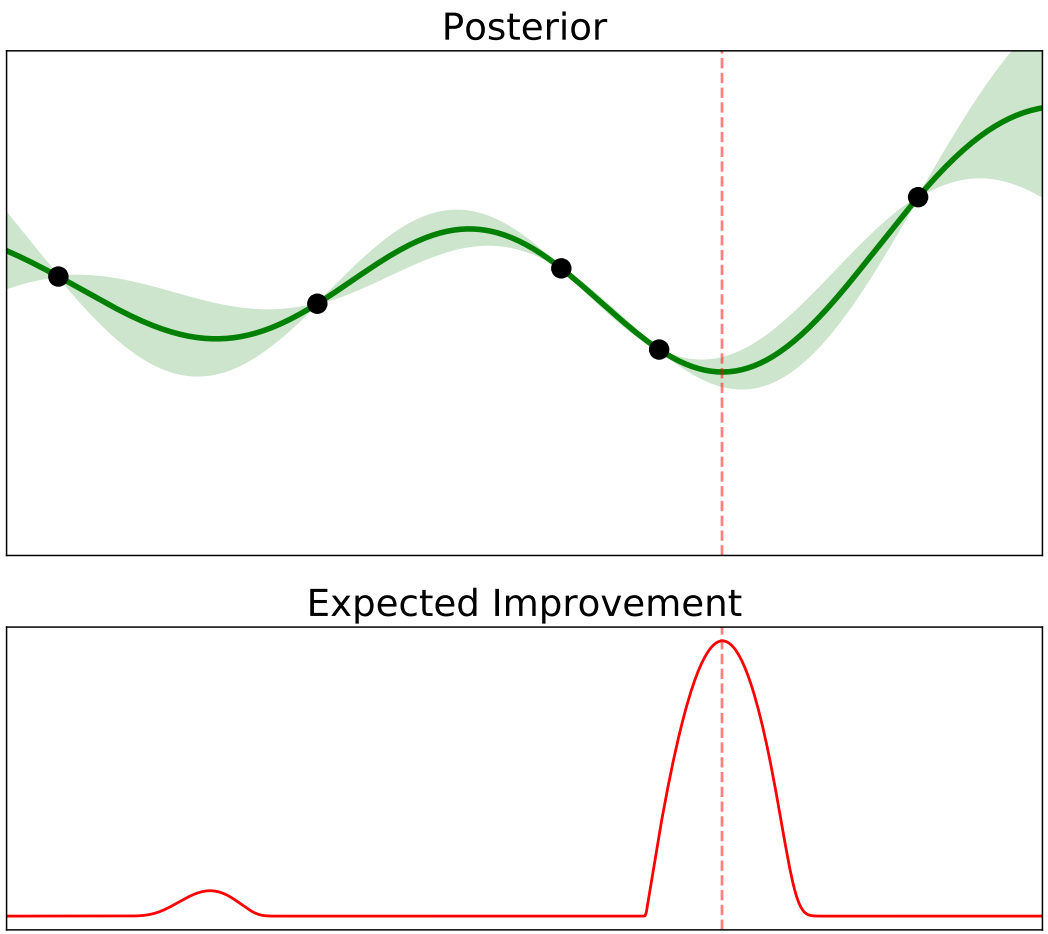
Acquisition Functions

These combine the surrogate model's predictions and their uncertainty to strike a balance between exploitation and exploration.



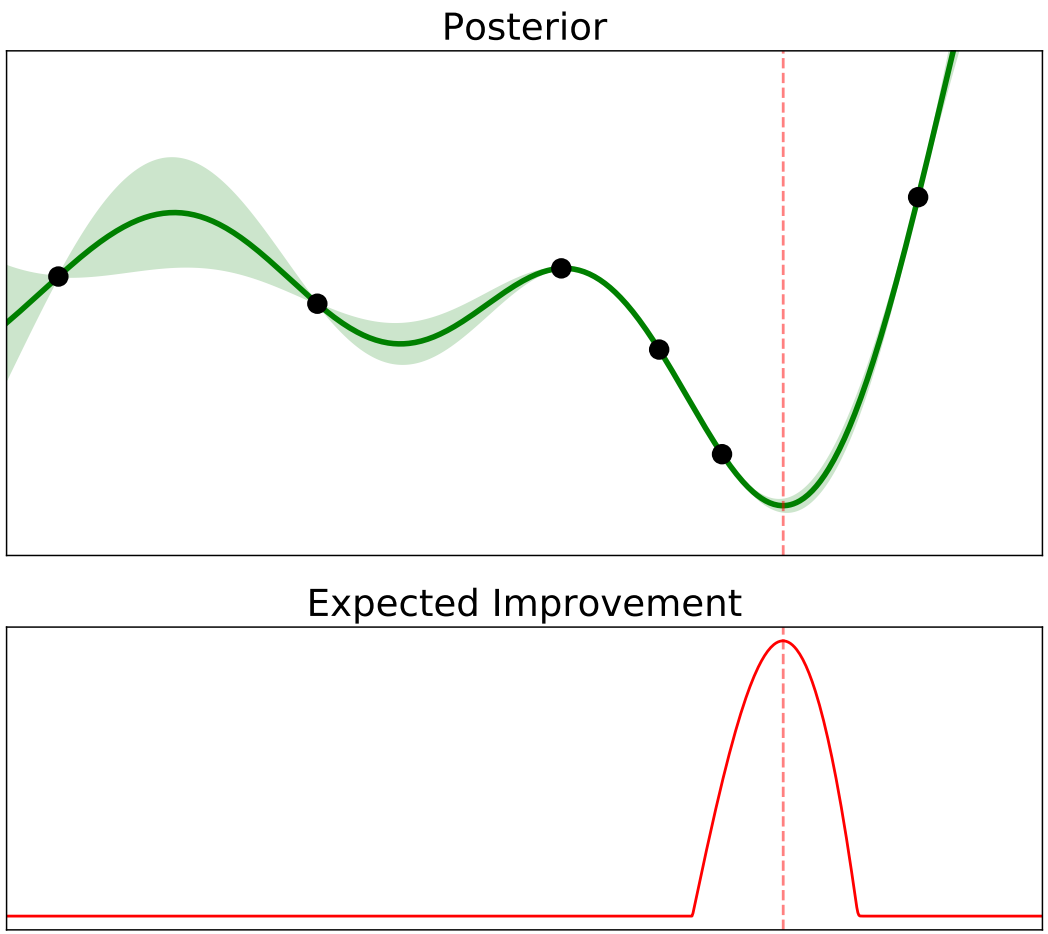
Acquisition Functions

These combine the surrogate model's predictions and their uncertainty to strike a balance between exploitation and exploration.



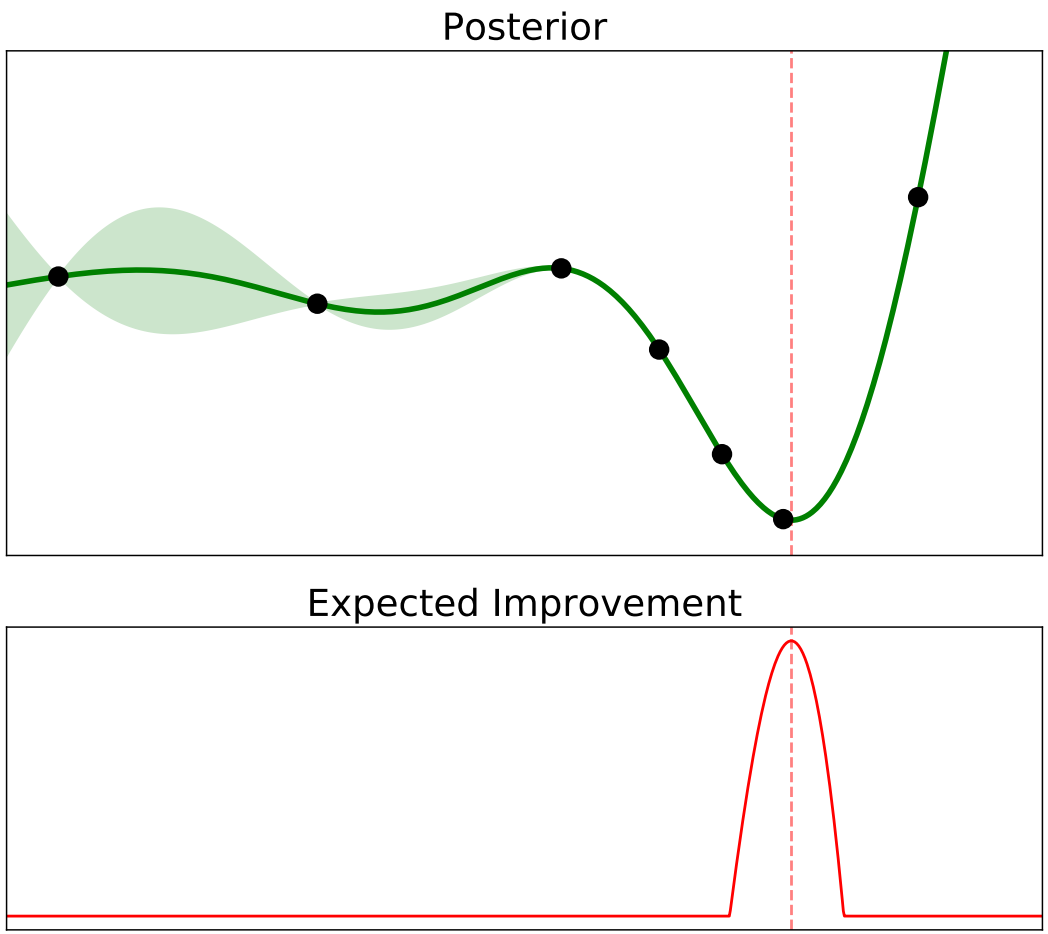
Acquisition Functions

These combine the surrogate model's predictions and their uncertainty to strike a balance between exploitation and exploration.



Acquisition Functions

These combine the surrogate model's predictions and their uncertainty to strike a balance between exploitation and exploration.



Previous Work

Acquisition Functions

Many acquisition functions have been proposed and compared:

- EI, PI, UCB, ES, PES, MES, FITBO, ϵ-greedy, ...

Previous Work

Acquisition Functions

Many acquisition functions have been proposed and compared:

- EI, PI, UCB, ES, PES, MES, FITBO, ϵ-greedy, ...

Kernel Functions

The role of kernels have also been investigated in BO:

- Linear, Squared Exponential (RBF), Matérn, Periodic, ...

Previous Work

Acquisition Functions

Many acquisition functions have been proposed and compared:

- EI, PI, UCB, ES, PES, MES, FITBO, ϵ-greedy, ...

Kernel Functions

The role of kernels have also been investigated in BO:

- Linear, Squared Exponential (RBF), Matérn, Periodic, ...

Mean Functions

Little attention has been paid to the role of the mean function in BO.

- Observations are usually standardised and a mean of 0 is used.

Mean Functions

Mean function can be represented as a set of basis functions $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})]$ with corresponding weights $\mathbf{w} = [w_1, \ldots, w_k]$:

$$m(\mathbf{x}) = \sum_{i=1}^{k} w_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top \mathbf{w}.$$

Mean Functions

Mean function can be represented as a set of basis functions $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})]$ with corresponding weights $\mathbf{w} = [w_1, \ldots, w_k]$:

$$m(\mathbf{x}) = \sum_{i=1}^{k} w_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top \mathbf{w}.$$

A constant mean function with value c can be written as

$$\mathbf{h}(\mathbf{x}) = c\mathbf{1} \quad \text{with} \quad \mathbf{w} = \mathbf{1}.$$

Mean Functions

Mean function can be represented as a set of basis functions $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})]$ with corresponding weights $\mathbf{w} = [w_1, \ldots, w_k]$:

$$m(\mathbf{x}) = \sum_{i=1}^{k} w_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top \mathbf{w}.$$

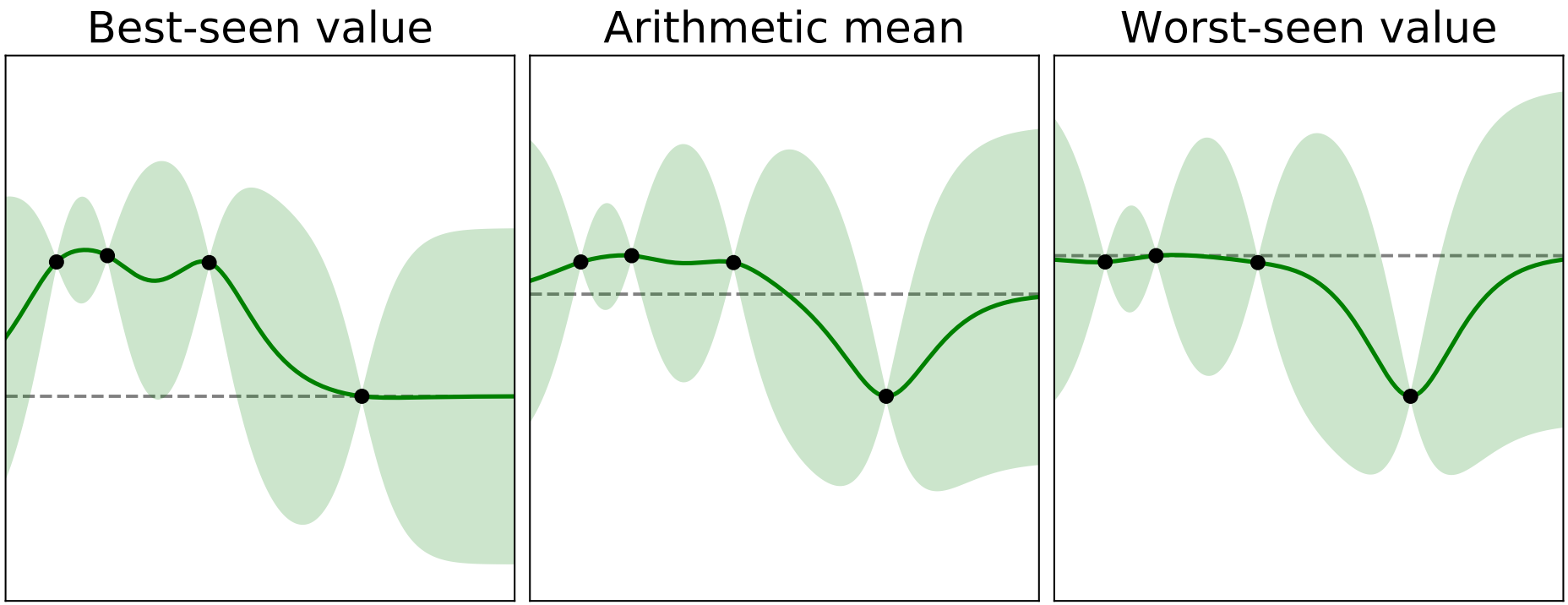
A constant mean function with value c can be written as

$$\mathbf{h}(\mathbf{x}) = c\mathbf{1} \quad \text{with} \quad \mathbf{w} = \mathbf{1}.$$

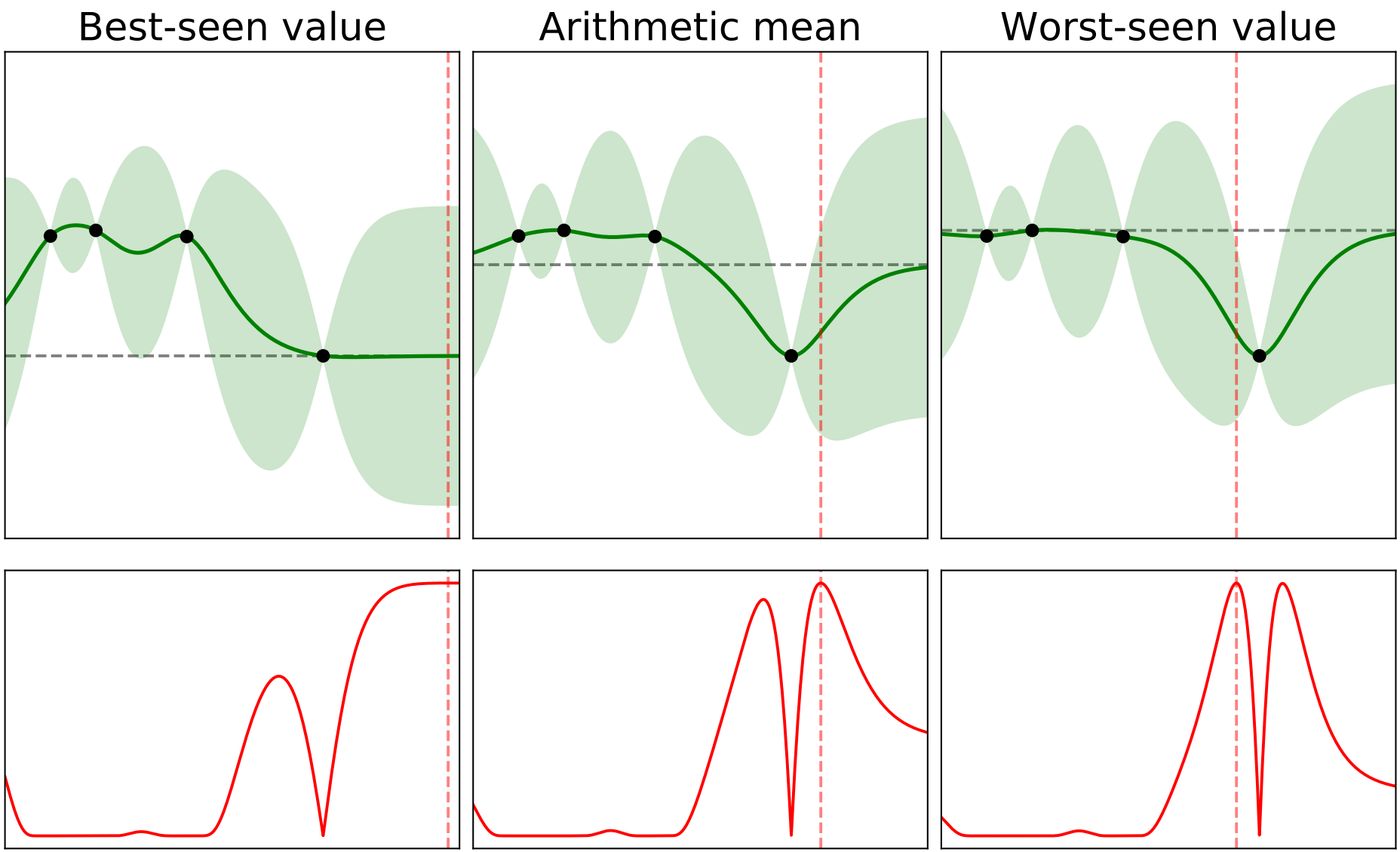
Given the observations $\mathcal{D} = \{(\mathbf{x}_n,\ f_n \triangleq f(\mathbf{x}_n))\}_{n=1}^{t}$, we consider four constant mean functions:

- Arithmetic mean: $c = t^{-1} \sum_{n=1}^{t} f_n$.
- Best-seen value: $c = \min\{f_1, \ldots, f_t\}$.
- Worst-seen value: $c = \max\{f_1, \ldots, f_t\}$.
- Median: $c = \text{median}\{f_1, \ldots, f_t\}$.

Mean Functions: Constant

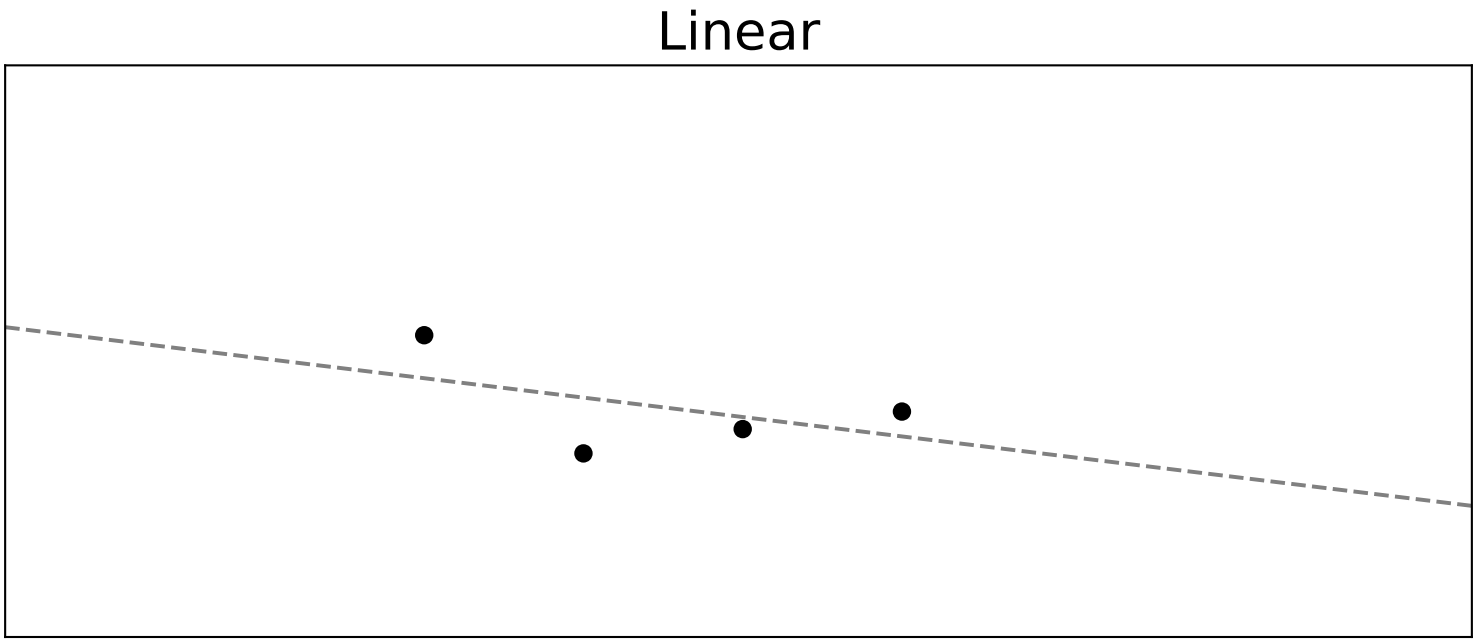


Mean Functions: Constant



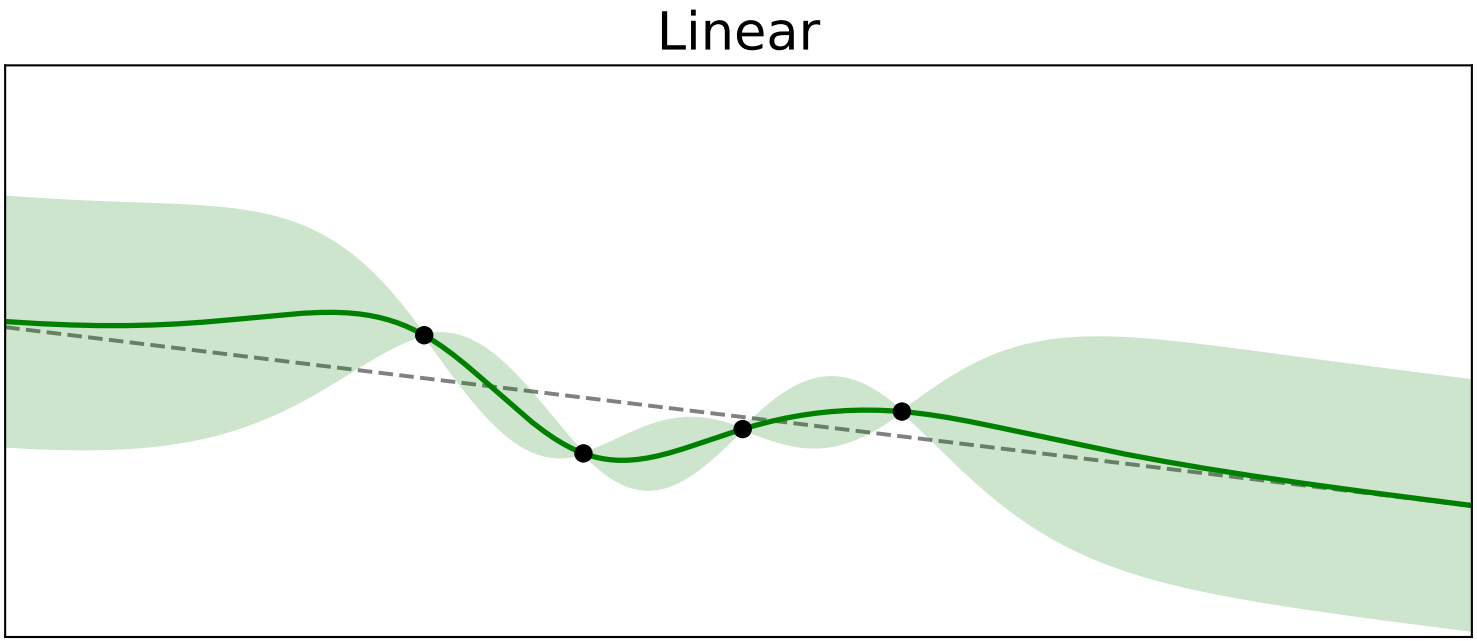
Mean Functions: Linear

$$\mathbf{h}(\mathbf{x}) = [1, x_1, x_2, \ldots, x_d] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^{d+1}.$$



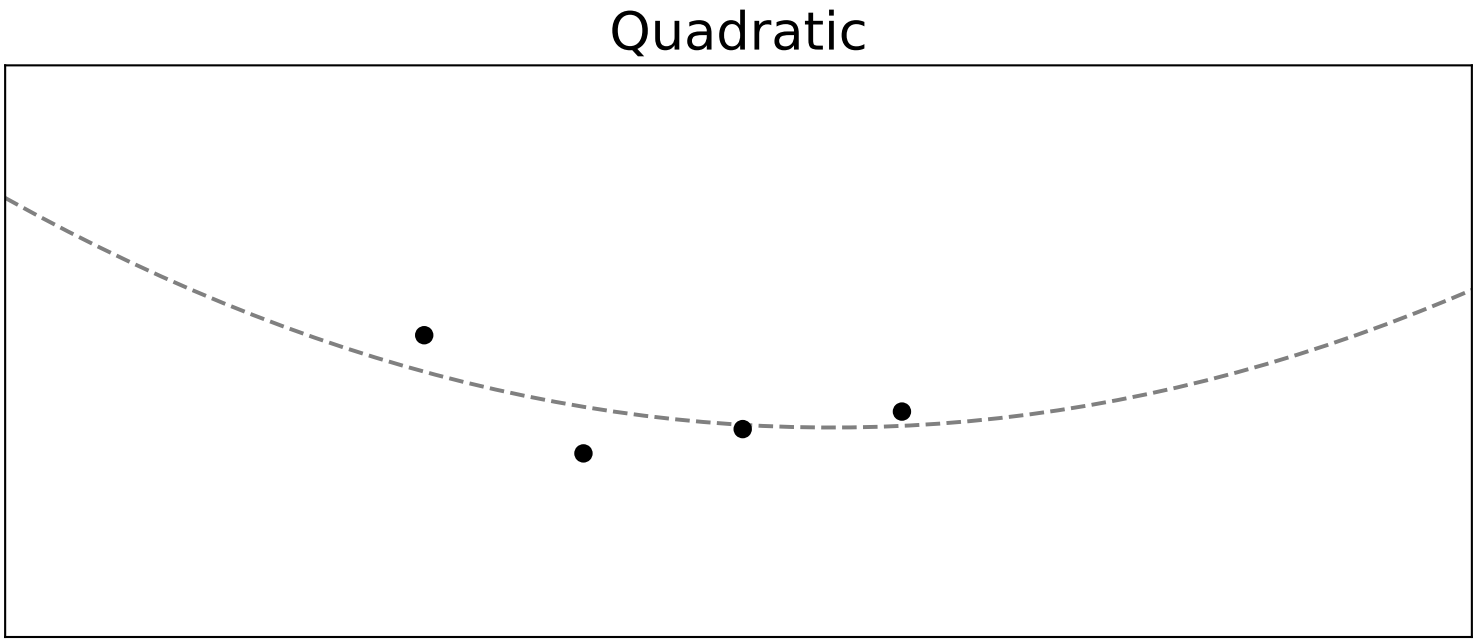
Mean Functions: Linear

$$\mathbf{h}(\mathbf{x}) = [1, x_1, x_2, \ldots, x_d] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^{d+1}.$$



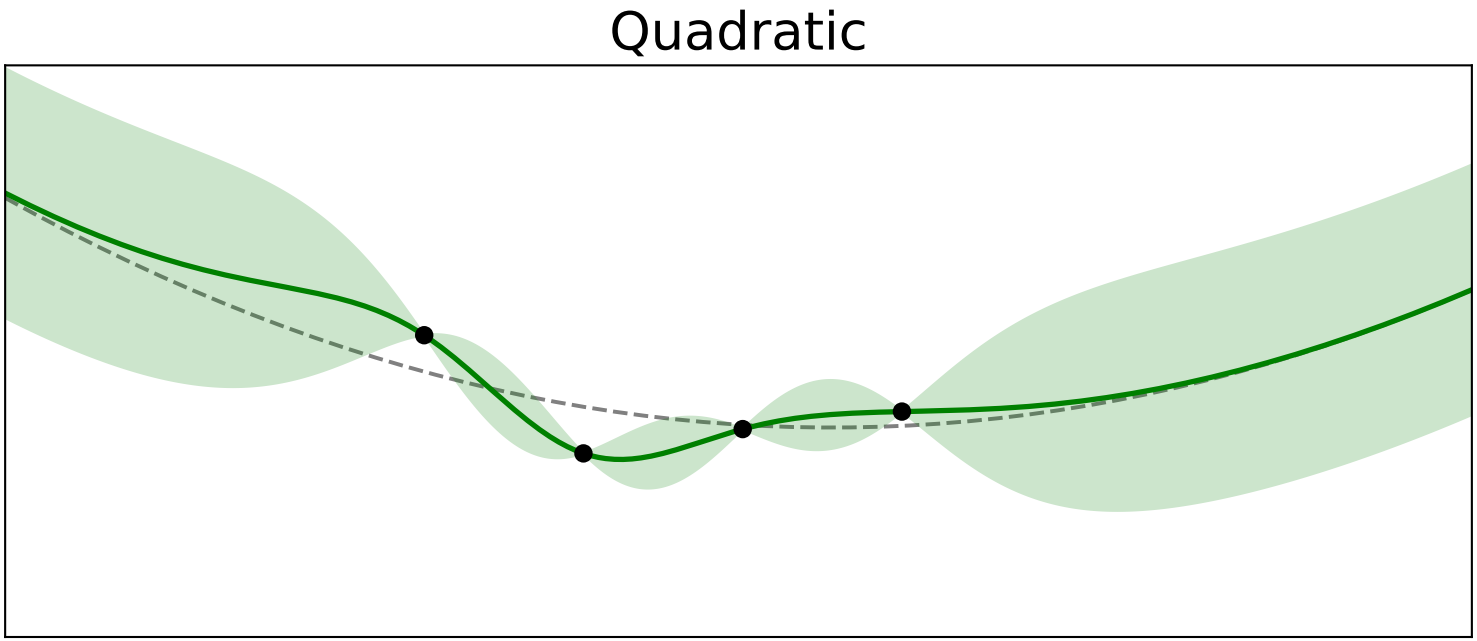
Mean Functions: Quadratic

$$\mathbf{h}(\mathbf{x}) = [1, x_1, x_1^2, x_1 x_2, x_2, \cdots, x_{d-1} x_d] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^{\binom{d+2}{d}}.$$



Mean Functions: Quadratic

$$\mathbf{h}(\mathbf{x}) = [1, x_1, x_1^2, x_1 x_2, x_2, \cdots, x_{d-1} x_d] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^{\binom{d+2}{d}}.$$

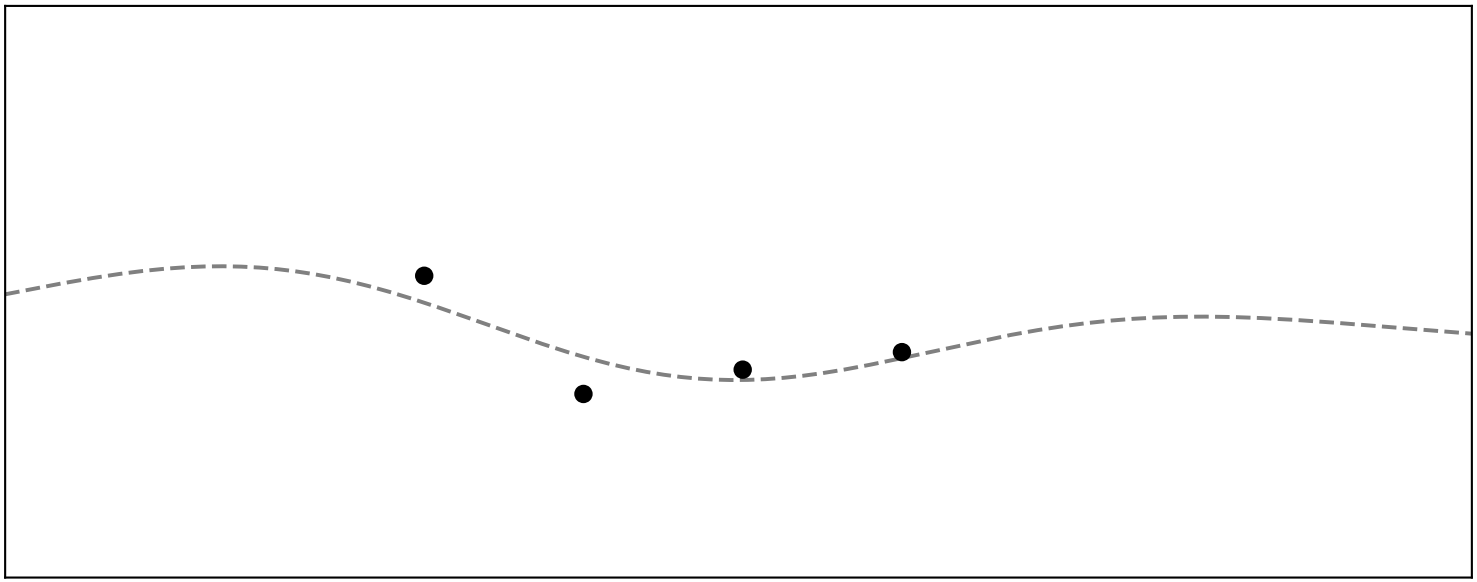


Mean Functions: RBF Network

- Radial basis function networks consist of a set of basis functions that depend only on the distance from a fixed centre.
- Gaussian RBFs $\phi_i(\mathbf{z}) = \exp(-\gamma\|\mathbf{z} - \mathbf{x}_i\|^2)$ are placed at each of the t previously-evaluated locations.

$$\mathbf{h}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_t(\mathbf{x})] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^t.$$

RBF

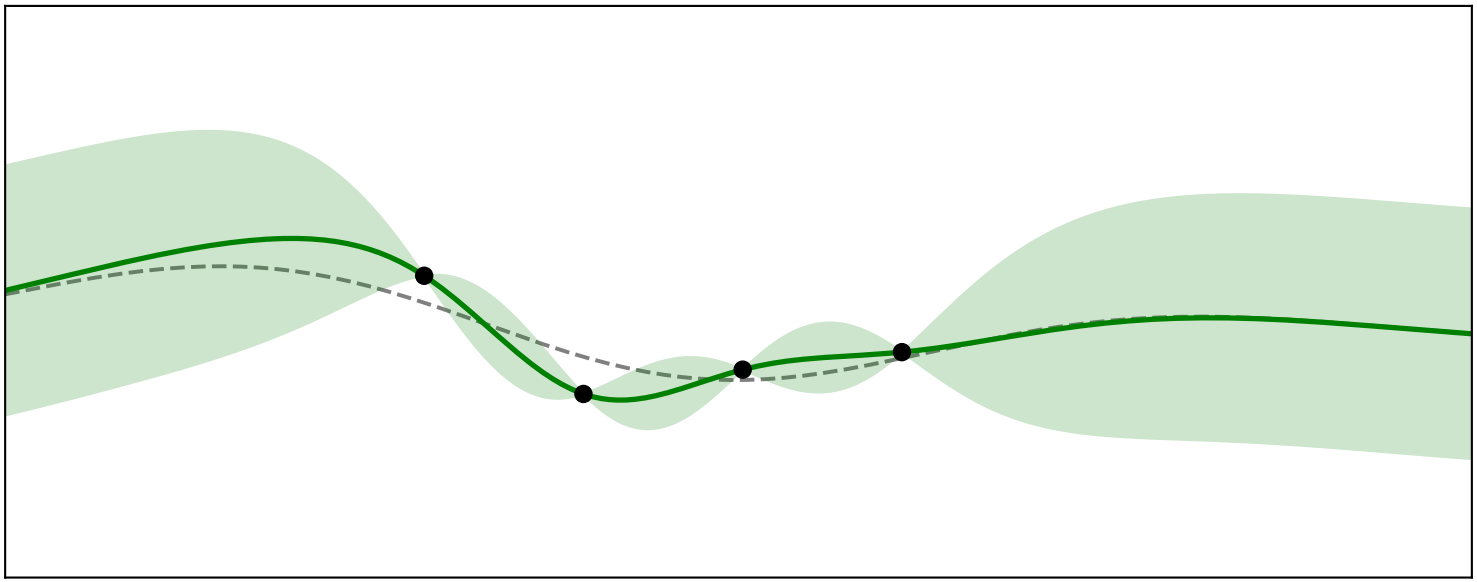


Mean Functions: RBF Network

- Radial basis function networks consist of a set of basis functions that depend only on the distance from a fixed centre.
- Gaussian RBFs $\phi_i(\mathbf{z}) = \exp(-\gamma\|\mathbf{z} - \mathbf{x}_i\|^2)$ are placed at each of the t previously-evaluated locations.

$$\mathbf{h}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_t(\mathbf{x})] \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^t.$$

RBF



Mean Functions: Fitting

- To avoid overfitting to the data, the weights are learnt using a regularised least-squares approximation. The optimal weights $\mathbf{w}^*$ are calculated by solving

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{f} - H\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2,$$

  where $H = [\mathbf{h}(\mathbf{x}_1), \mathbf{h}(\mathbf{x}_2), \ldots, \mathbf{h}(\mathbf{x}_t)]$ and $\lambda \geq 0$.

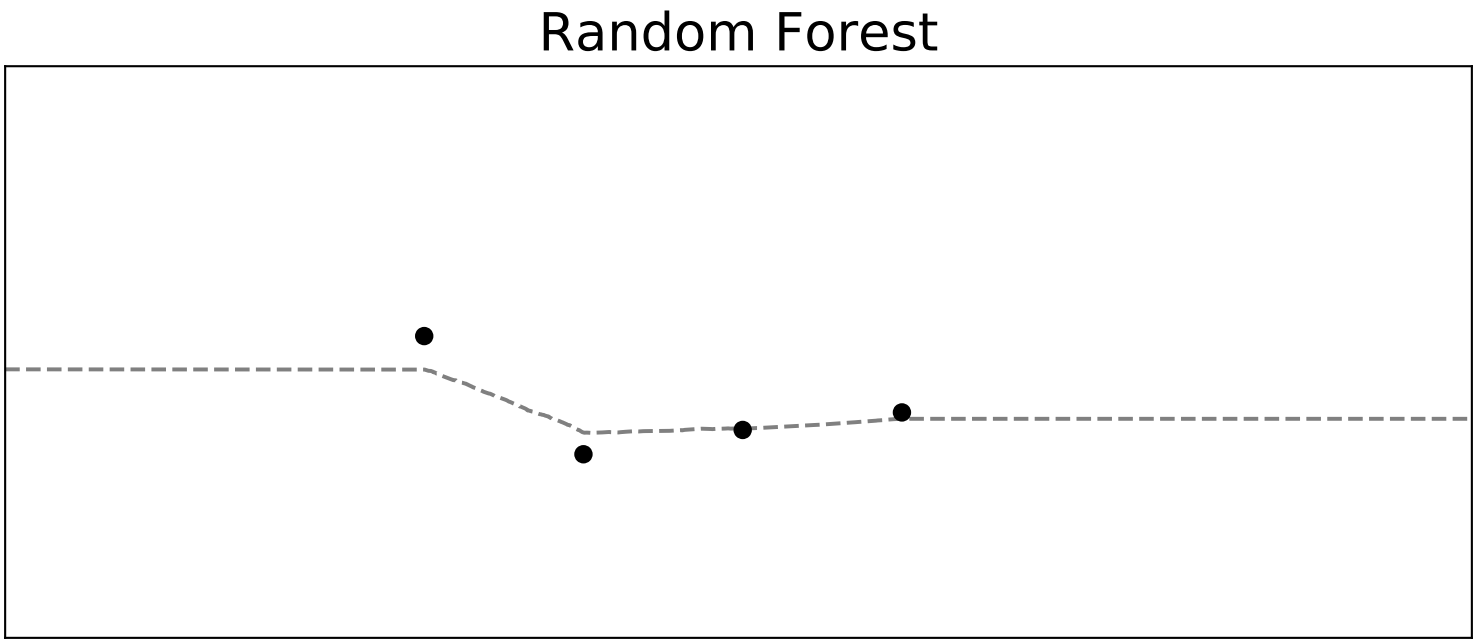
- The ordinary least-squares estimator for this is

$$\mathbf{w}^* = \left(H^\top H + \lambda \mathbf{I}\right)^{-1} H^\top \mathbf{f}.$$

- λ (and γ) chosen via five-fold cross-validation.

Mean Functions: Random Forests (Extra-Trees)

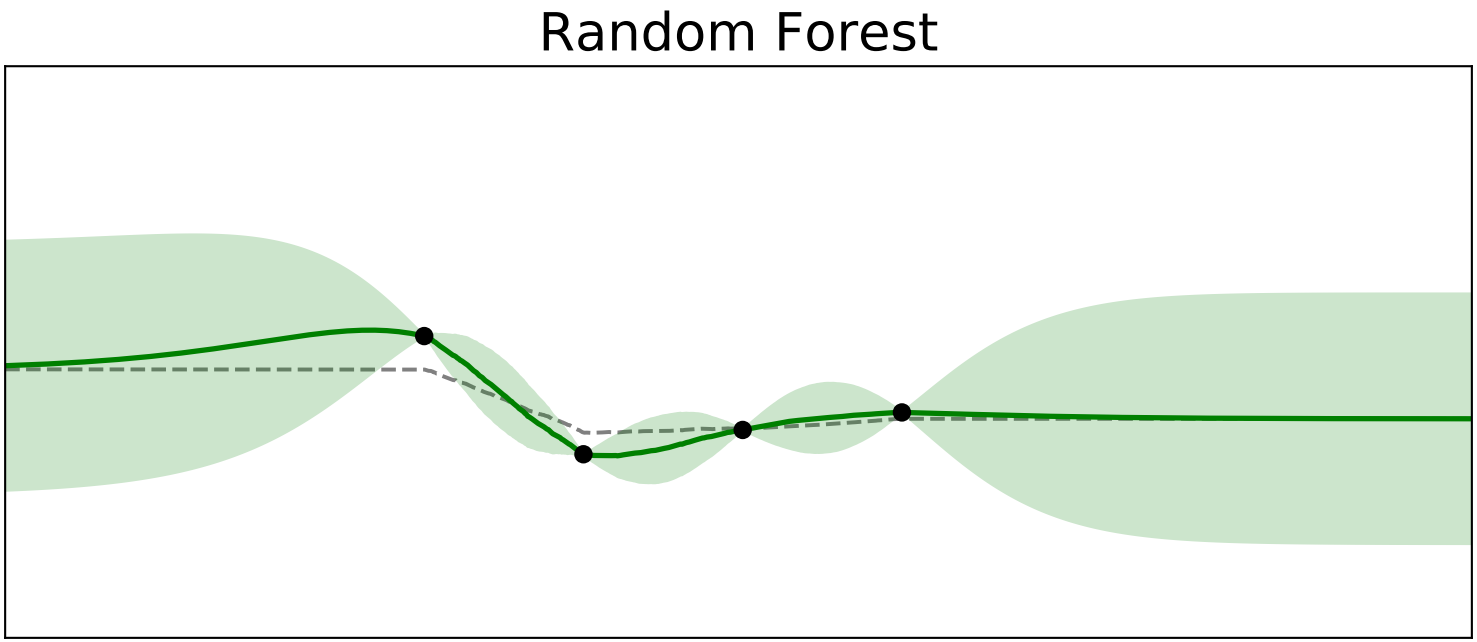
Random Forests are ensembles of regression trees that are each trained on randomly chosen subsets of the data and select each tree split optimally. Extremely randomised trees (Geurts et al. 2006) select each split as the best from a small, randomly-chosen set of splits, resulting in a much smoother regression.



P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. Machine Learning 63, 1 (2006), 3–42.

Mean Functions: Random Forests (Extra-Trees)

Random Forests are ensembles of regression trees that are each trained on randomly chosen subsets of the data and select each tree split optimally. Extremely randomised trees (Geurts et al. 2006) select each split as the best from a small, randomly-chosen set of splits, resulting in a much smoother regression.

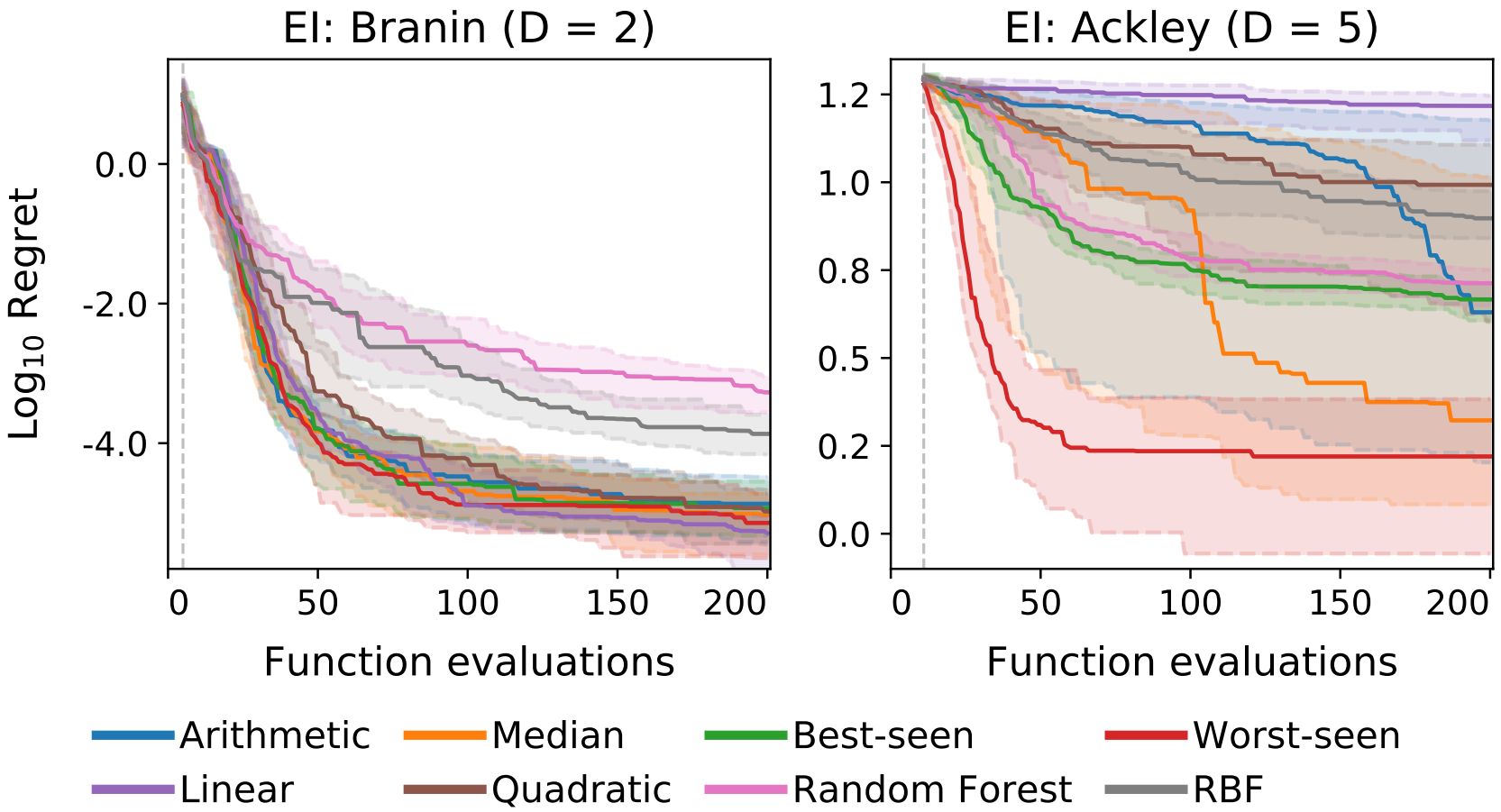


P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. Machine Learning 63, 1 (2006), 3–42.

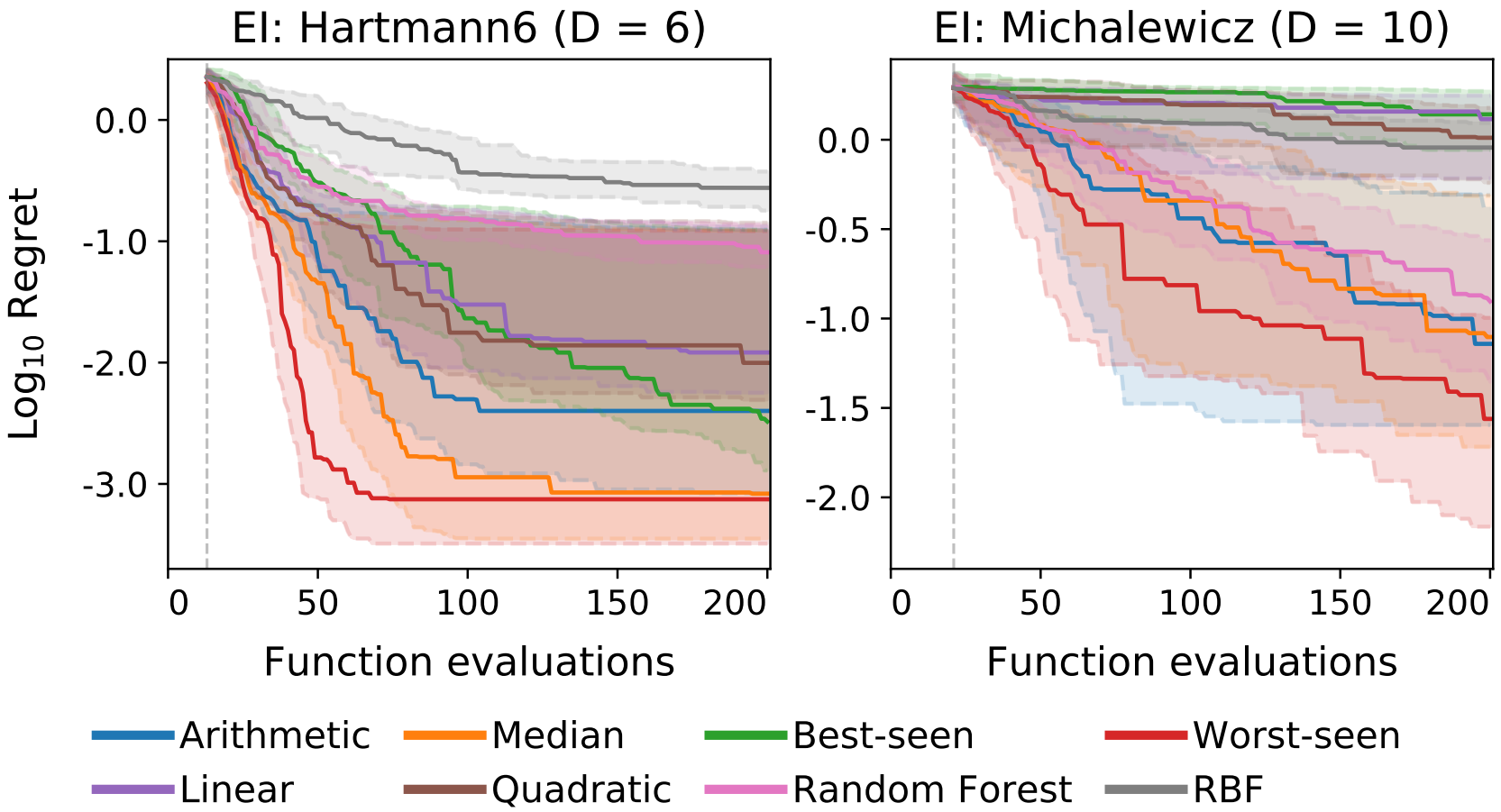
Experimental Evaluation

- Eight mean functions:
  - *Arithmetic*, *Median*, *Best-seen* and *Worst-seen*.
  - *Linear* and *Quadratic*.
  - *Random Forest* (Extra-Trees).
  - *RBF* (RBF network).
- Three sets of experiments:
  - 10 synthetic functions.
  - 2 robot pushing (active learning) problems.
  - Computational fluid dynamics problem.
- Experimental set-up:
  - Acquisition function: EI (UCB in paper).
  - 200 function evaluation budget (including initial samples).
  - 51 optimisation runs per mean function.

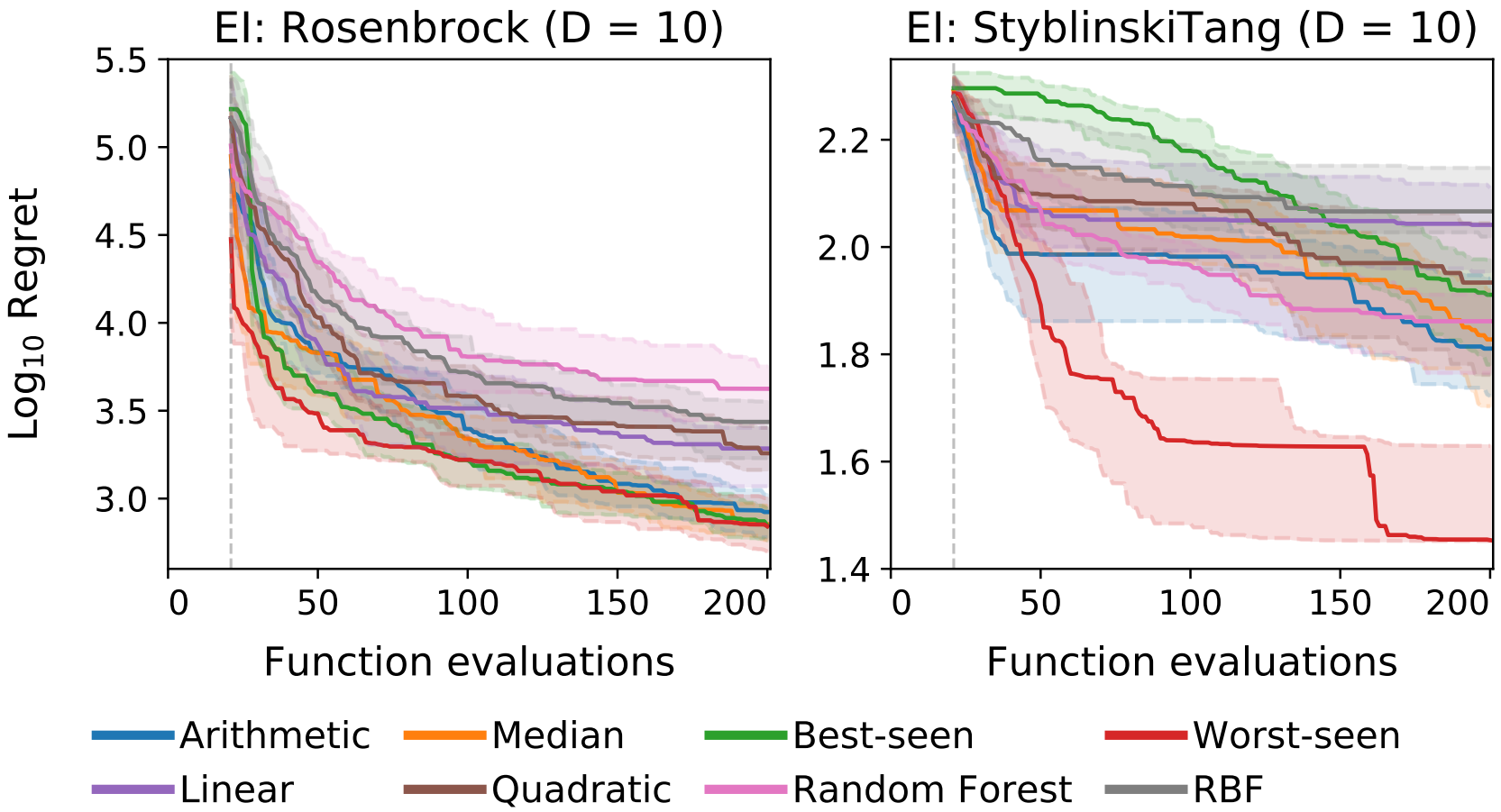
Synthetic Functions: Convergence



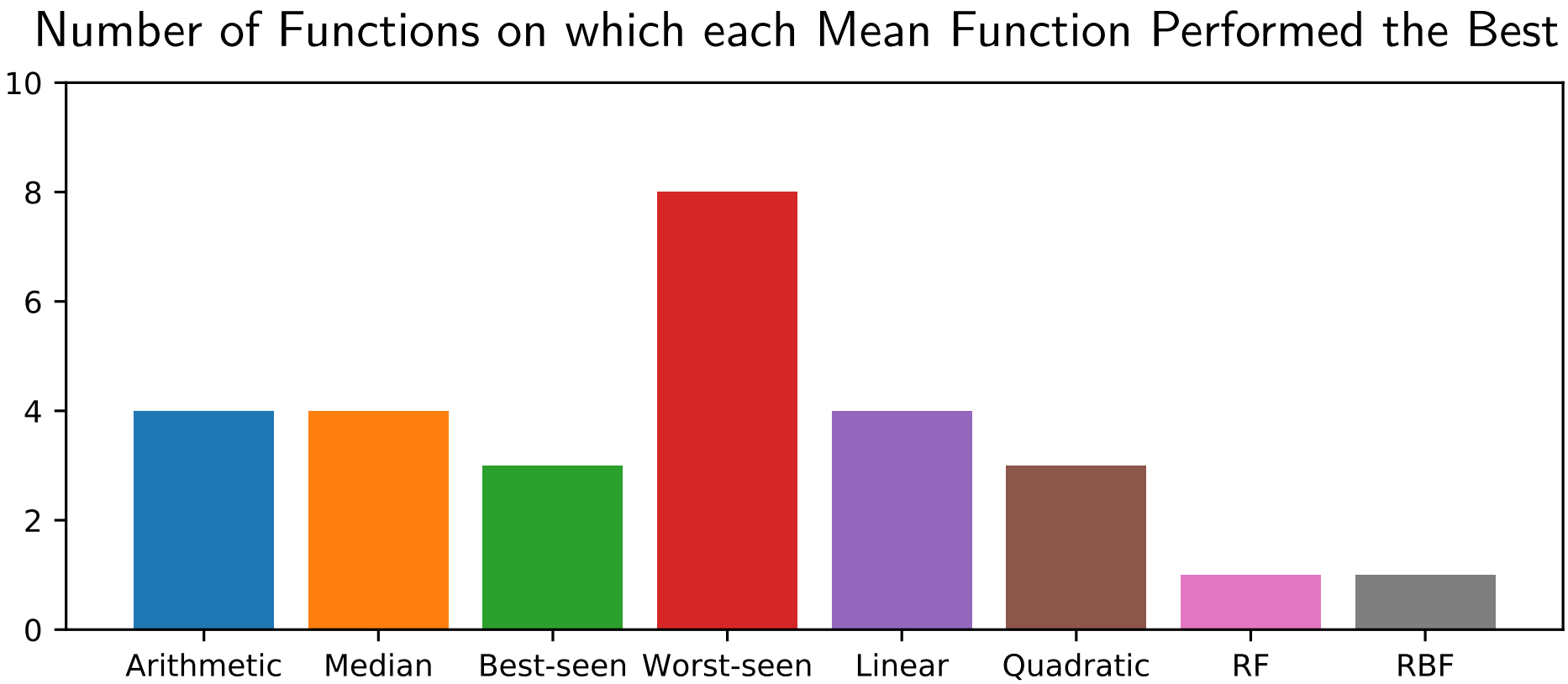
Synthetic Functions: Convergence



Synthetic Functions: Convergence



Synthetic Functions: Best Performing



Why is Using the Worst-seen Value Good?

It makes the acquisition function more *greedy*.

- Exploitation leads to faster convergence in higher dimensions.

Recent works support the *greed is good* strategy:

Why is Using the Worst-seen Value Good?

It makes the acquisition function more *greedy*.

- Exploitation leads to faster convergence in higher dimensions.

Recent works support the *greed is good* strategy:

- Greed is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation
  G. De Ath, R. Everson, A. Rahat, and J. Fieldsend. 2020.
  In ACM Transactions on Evolutionary Learning and Optimization (to appear).

Why is Using the Worst-seen Value Good?

It makes the acquisition function more *greedy*.

- Exploitation leads to faster convergence in higher dimensions.

Recent works support the *greed is good* strategy:

- Greed is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation
  G. De Ath, R. Everson, A. Rahat, and J. Fieldsend. 2020.
  In ACM Transactions on Evolutionary Learning and Optimization (to appear).
- Expected Improvement versus Predicted Value in Surrogate-Based Optimization.
  F. Rehbach, M. Zaefferer, B. Naujoks, and T. Bartz-Beielstein. 2020.
  In Genetic and Evolutionary Computation Conference (GECCO '20).

Why is Using the Worst-seen Value Good?

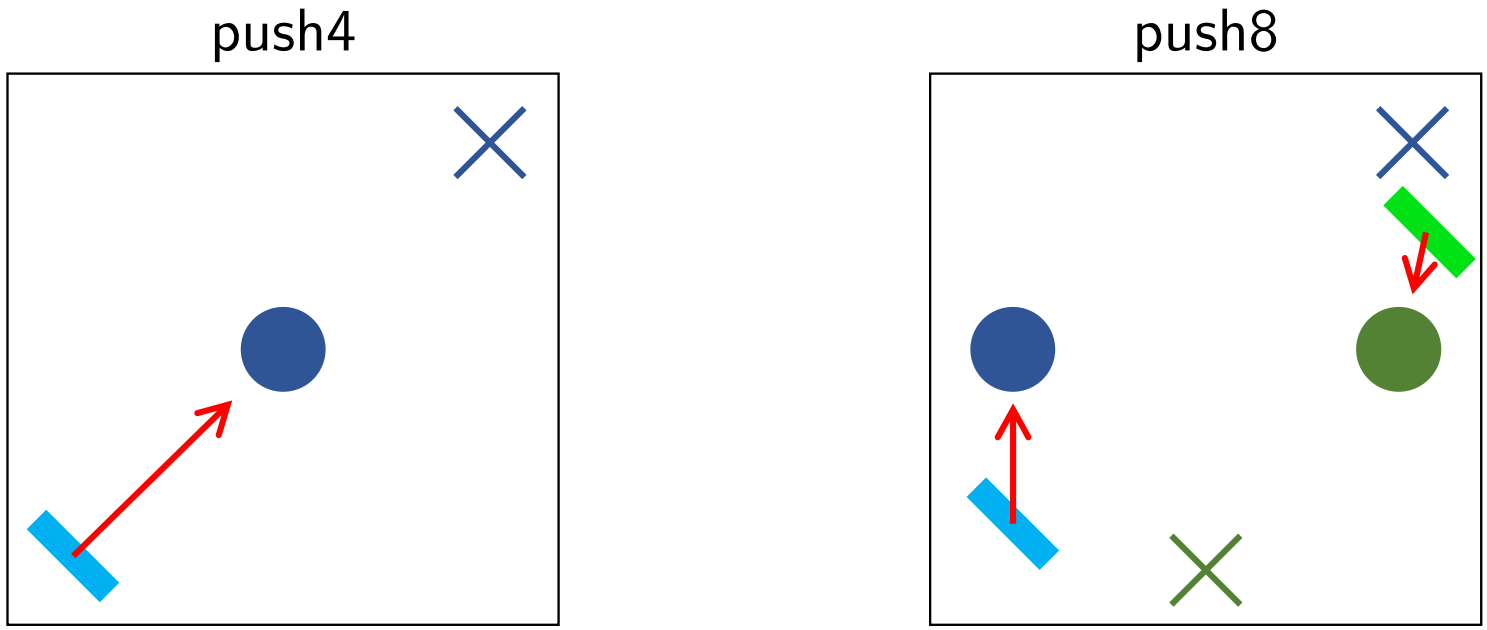
It makes the acquisition function more *greedy*.

- Exploitation leads to faster convergence in higher dimensions.

Recent works support the *greed is good* strategy:

- Greed is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation
  G. De Ath, R. Everson, A. Rahat, and J. Fieldsend. 2020.
  In ACM Transactions on Evolutionary Learning and Optimization (to appear).
- Expected Improvement versus Predicted Value in Surrogate-Based Optimization.
  F. Rehbach, M. Zaefferer, B. Naujoks, and T. Bartz-Beielstein. 2020.
  In Genetic and Evolutionary Computation Conference (GECCO '20).
- ϵ-shotgun: ϵ–greedy Batch Bayesian Optimisation.
  G. De Ath, R. Everson, J. Fieldsend, and A. Rahat. 2020.
  In Genetic and Evolutionary Computation Conference (GECCO '20).

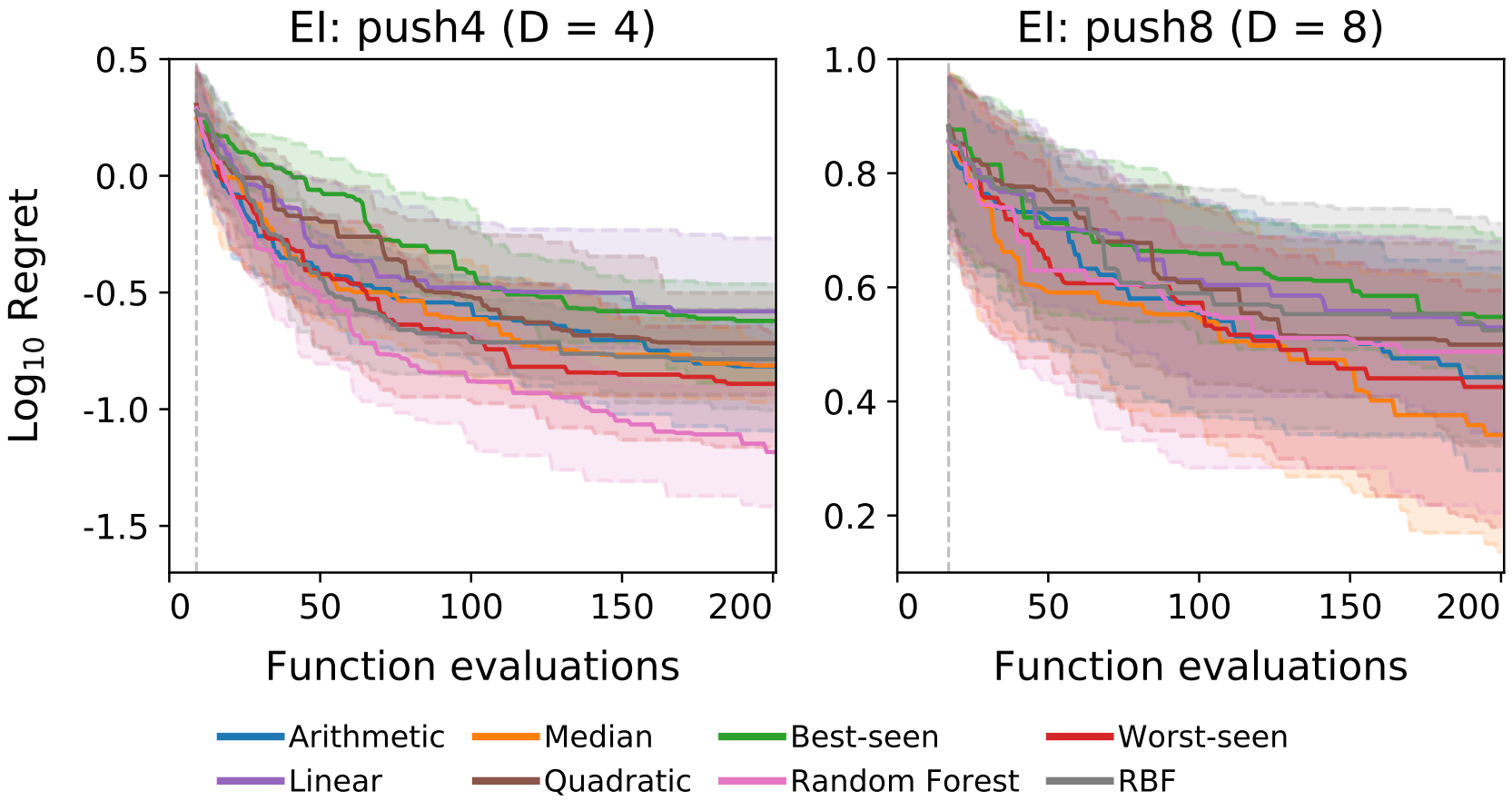
Robot Pushing (Wang and Jegelka, 2017)



- Parameters to be optimised:
  - Robot's initial location, hand orientation and time spent pushing.
- Fitness function: final distance of pushed object to target.
- Optimisation over problem instances.

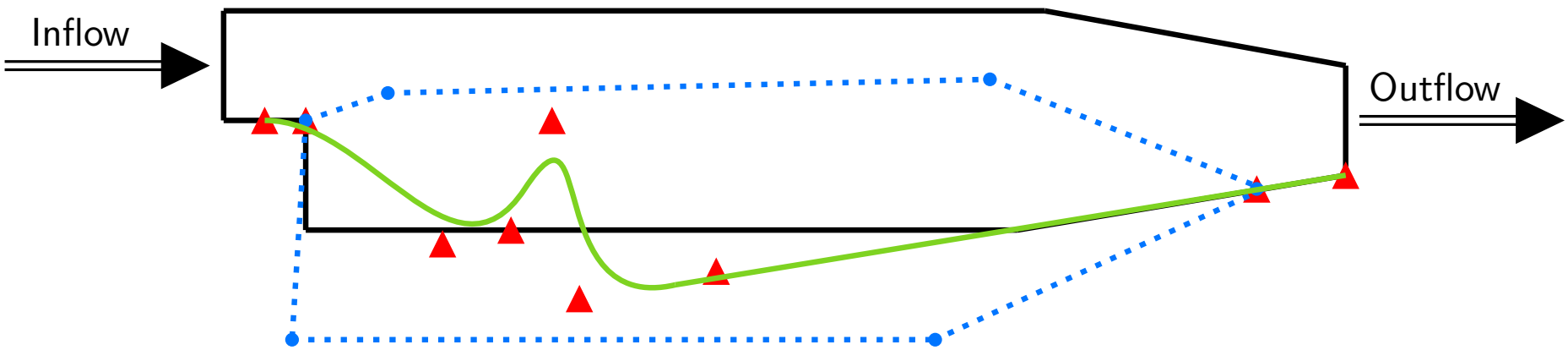
Z. Wang and S. Jegelka. 2017. Max-value entropy search for efficient Bayesian optimization. In Proceedings of the 34th International Conference on Machine Learning. PMLR, 3627–3635.

Robot Pushing: Convergence



Pipe Shape Optimisation

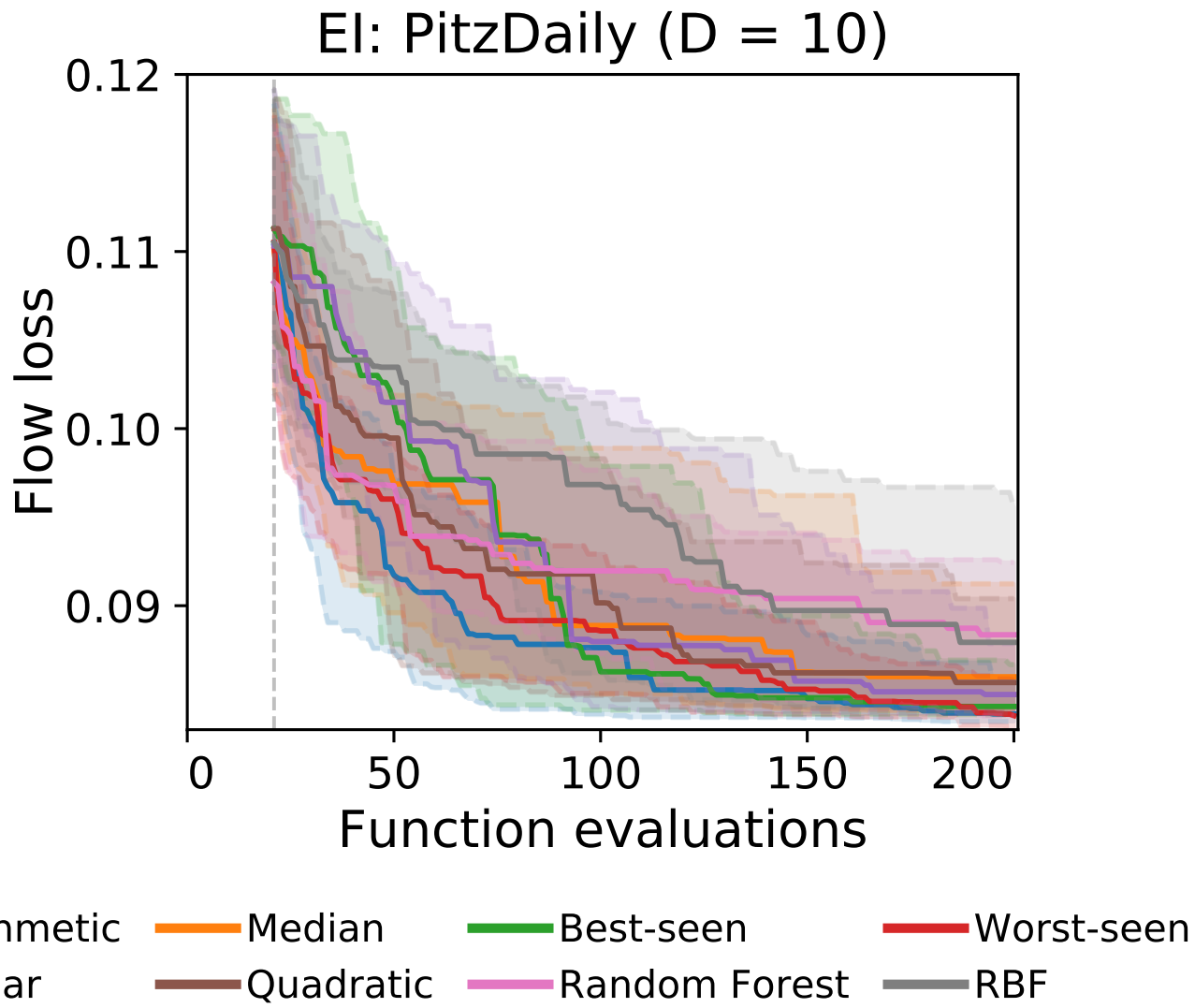
- PitzDaily computational fluid dynamics problem (Daniels et al. 2018).
- Minimise the pressure loss between the inflow and outflow by changing the shape of the lower wall of the pipe.



- Catmull-Clark subdivision curve, defined by 5 control points (▲), and thus resulting in a 10-dimensional problem.
- Constrained to lie within the blue polygon.

S. Daniels, A. Rahat, R. Everson, G. Tabor, and J. Fieldsend. 2018. A Suite of Computationally Expensive Shape Optimisation Problems Using Computational Fluid Dynamics. In Parallel Problem Solving from Nature – PPSN XV. Springer, 296–307.

Pipe Shape Optimisation: Convergence



Concluding Remarks

UNIVERSITY OF EXETER

Greed is good

Using a constant mean equal to the worst-seen value is often better than the arithmetic mean, particularly in higher dimensions.

- Real-world tasks: no optimum choice, highlighting a lack of consistency between synthetic and real-world problems.
- Python code is available at:
  `https://github.com/georgedeath/bomean`
- Future work:
  - Jointly learning the mean function and GP parameters.
  - Fully Bayesian approaches.
  - Adaptation of the BO pipeline to suit the problem structure.

Arithmetic mean vs Worst-seen value

Mean function	**Branin (2)**		**Eggholder (2)**		**GoldsteinPrice (2)**		**SixHumpCamel (2)**		**Shekel (4)**	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Arithmetic	1.35×10^{-5}	1.82×10^{-5}	1.58	1.93	4.18×10^{-2}	5.32×10^{-2}	2.51×10^{-5}	2.58×10^{-5}	8.13×10^{-2}	1.19×10^{-1}
Worst-seen	7.21×10^{-6}	8.81×10^{-6}	2.69	2.48	6.87×10^{-2}	7.40×10^{-2}	1.52×10^{-5}	1.84×10^{-5}	7.16×10^{-2}	1.05×10^{-1}

Mean function	**Ackley (5)**		**Hartmann6 (6)**		**Michalewicz (10)**		**Rosenbrock (10)**		**StyblinskiTang (10)**	
	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Arithmetic	4.27	6.06	4.00×10^{-3}	5.46×10^{-3}	7.22×10^{-2}	1.03×10^{-1}	8.38×10^{2}	3.27×10^{2}	6.47×10^{1}	2.58×10^{1}
Worst-seen	1.66	1.23	7.47×10^{-4}	9.88×10^{-4}	2.75×10^{-2}	3.72×10^{-2}	6.95×10^{2}	3.64×10^{2}	2.84×10^{1}	2.08×10^{1}

GP Modelling Error: NRMSE

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(f_n - \hat{f}_n)^2}}{f_{max} - f_{min}}$$

	Arithmetic	Median	Best-seen	Worst-seen	Linear	Quadratic	RF	RBF
Branin	0.271	0.271	0.271	0.263	0.270	0.263	0.121	0.114
Eggholder	0.217	0.165	0.216	0.177	0.161	0.163	0.198	0.166
GoldsteinPrice	0.136	0.136	0.136	0.136	0.136	0.136	0.069	0.136
SixHumpCamel	0.232	0.231	0.231	0.223	0.215	0.216	0.142	0.228
Shekel	0.265	0.215	0.494	0.058	0.083	0.094	0.142	0.053
Ackley	0.663	0.716	0.316	0.899	0.107	0.378	0.334	0.117
Hartmann6	0.222	0.394	0.184	0.154	0.140	0.185	0.195	0.097
Michalewicz	0.271	0.311	0.417	0.316	0.205	0.193	0.360	0.190
Rosenbrock	0.362	0.368	0.371	0.270	0.257	0.151	0.355	0.157
StyblinskiTang	0.325	0.347	0.360	0.151	0.267	0.199	0.220	0.168
push4	0.202	0.200	0.256	0.254	0.153	0.152	0.222	0.151
push8	0.205	0.202	0.238	0.168	0.136	0.129	0.191	0.125