Train In Data

Sequential Model-Based Optimization

Bayes' Rule

$$\overbrace{P(A|B)}^{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{likelihood}} \times \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

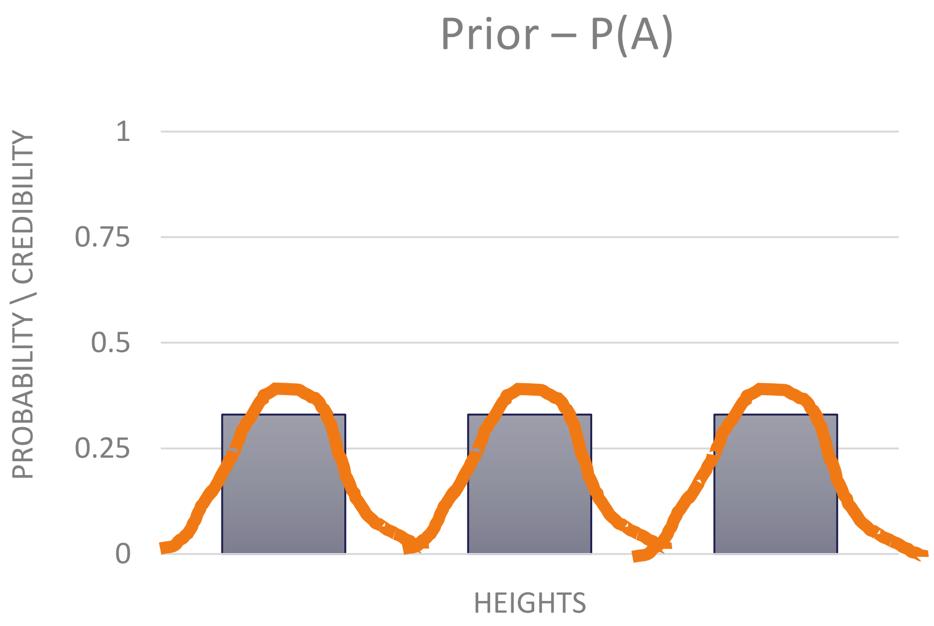
With **Bayes' Rule** we infer a posterior P(A | B) from a Prior (A)

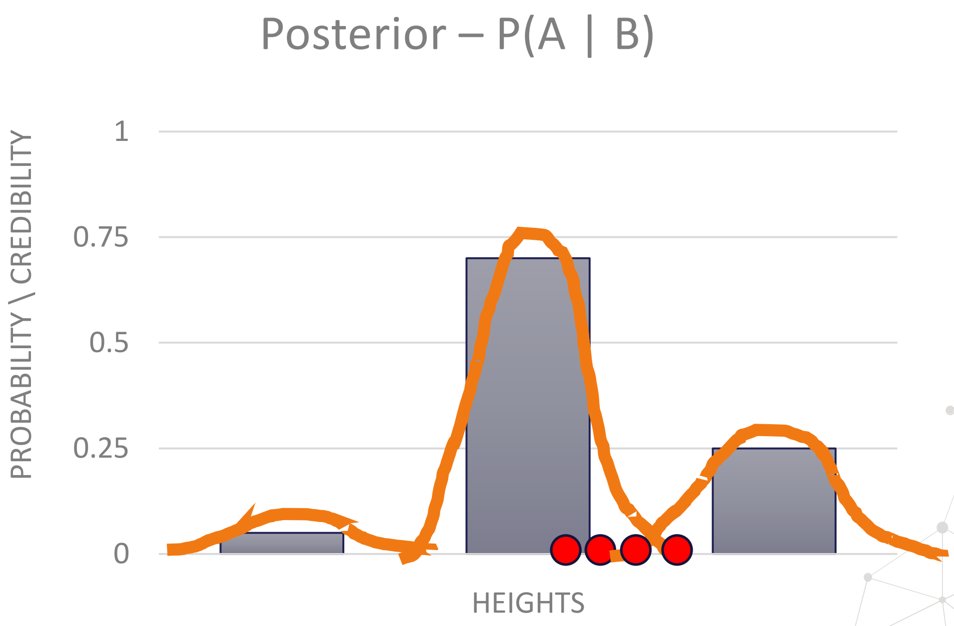
Bayes' Rule

$$\underbrace{P(A\,|B)}_{\text{posterior}} \;\alpha\; \underbrace{P(B\,|A)}_{\text{likelihood}} \times \underbrace{P(A)}_{\text{prior}}$$

With **Bayes' Rule** we infer a posterior P(A | B) from a Prior (A)

Probability reallocation





We **hypothesize** a range of possible distributions, models or generators (priors), and from data we determine their credibility (posterior).

Hyperparameter Optimization

Mathematically, we want to find the global maximizer (or minimizer) of an unknown (black-box) objective function *f*:

$$\mathbf{x}^{\star} = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

where x are the hyperparameters.

Hyperparameter Optimization

Mathematically, we want to find the global maximizer (or minimizer) of an unknown (black-box) objective function *f*:

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$$

Response surface

- **A**: Algorithm
  - A_λ: Hyperparameters
- **X**: Dataset
- **L**: Function to optimise
  - Performance metric

Black Box

where x are the hyperparameters.

Hyperparameter Optimization

- *Let **w** be f(x) and **D** be the available data*
- **w** is unknown, so we treat *it* as a random function and place a **prior** over it ➔ P(w)
- *P(w) captures our beliefs about the possible values of w*
- *Given **D** and the likelihood model **P(D | w)**, we can infer the posterior **P(w | D)** using Bayes' rule*

Hyperparameter Optimization

$$P(w \mid D) = \frac{P(D \mid w) \times P(w)}{P(D)}$$

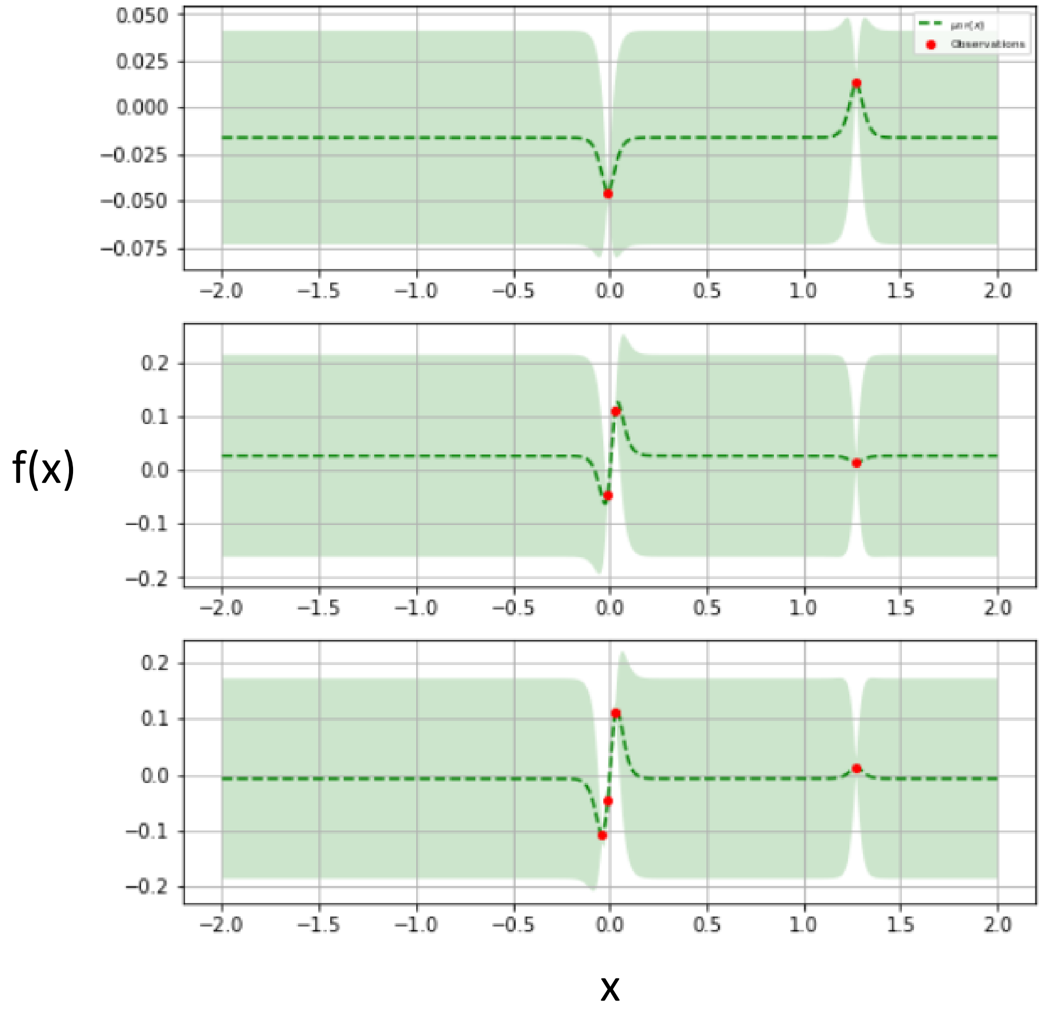
- *Let **w** be f(x) and **D** be the available data*
- **w** is unknown, so we treat *it* as a random function and place a **prior** over it ➔ P(w)
- *P(w) captures our beliefs about the possible values of w*
- *Given **D** and the likelihood model **P(D | w)**, we can infer the posterior **P(w | D)** using Bayes' rule*

Hyperparameter Optimization

$$P(w \mid D) = \frac{P(D \mid w) \times P(w)}{P(D)}$$

- *The posterior* ***P(w | D)*** *represents our updated belief of w after contemplating D.*
- Given the posterior distribution **P(w | D)**, we construct an acquisition function to determine the next query point to sample w.

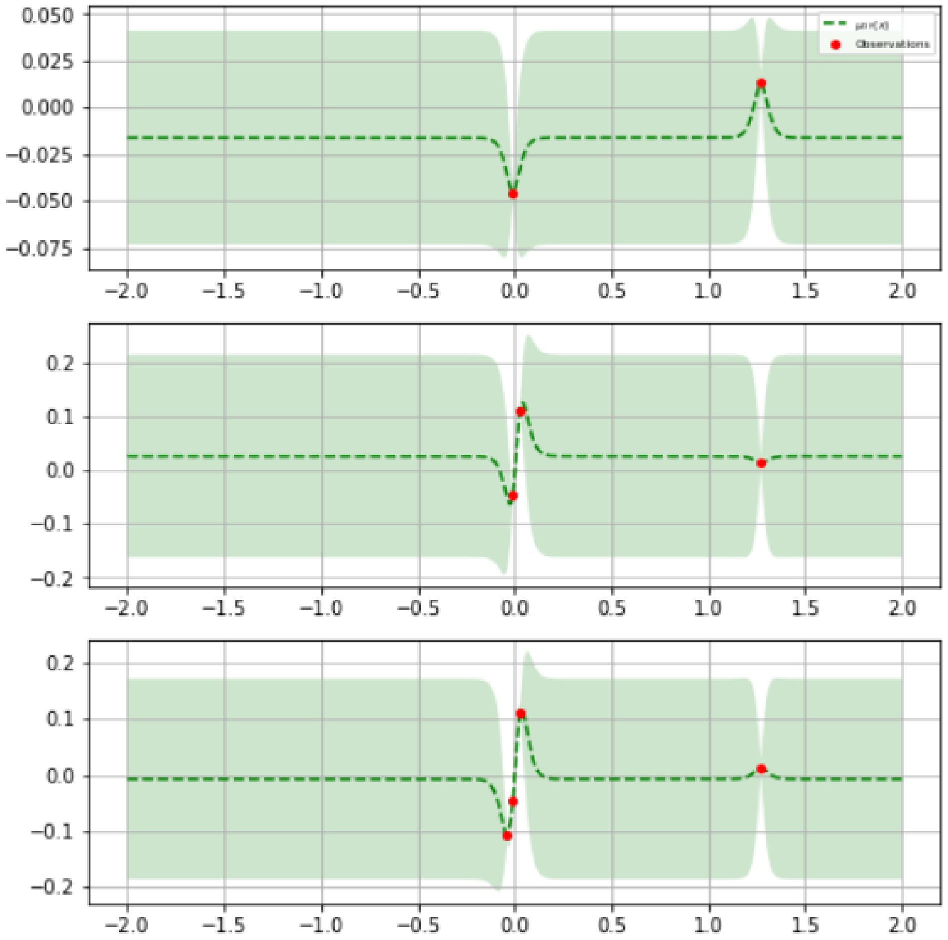
Hyperparameter Optimization



- *The green area is our Prior, $P(w)$*
- *D is the red dots, where we evaluate $f(x)$*
- *We adjust the posterior $P(w \mid D)$ based on D.*
- The uncertainty decreases around the evidence.

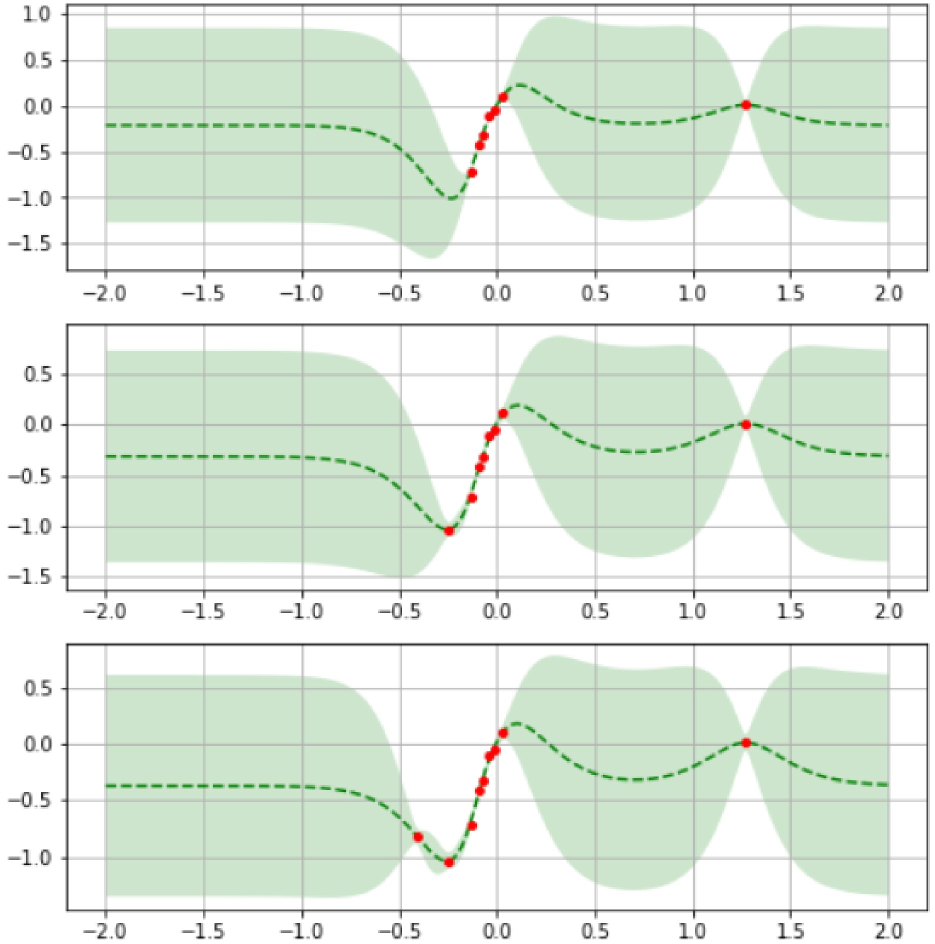
Hyperparameter Optimization

f(x)



x

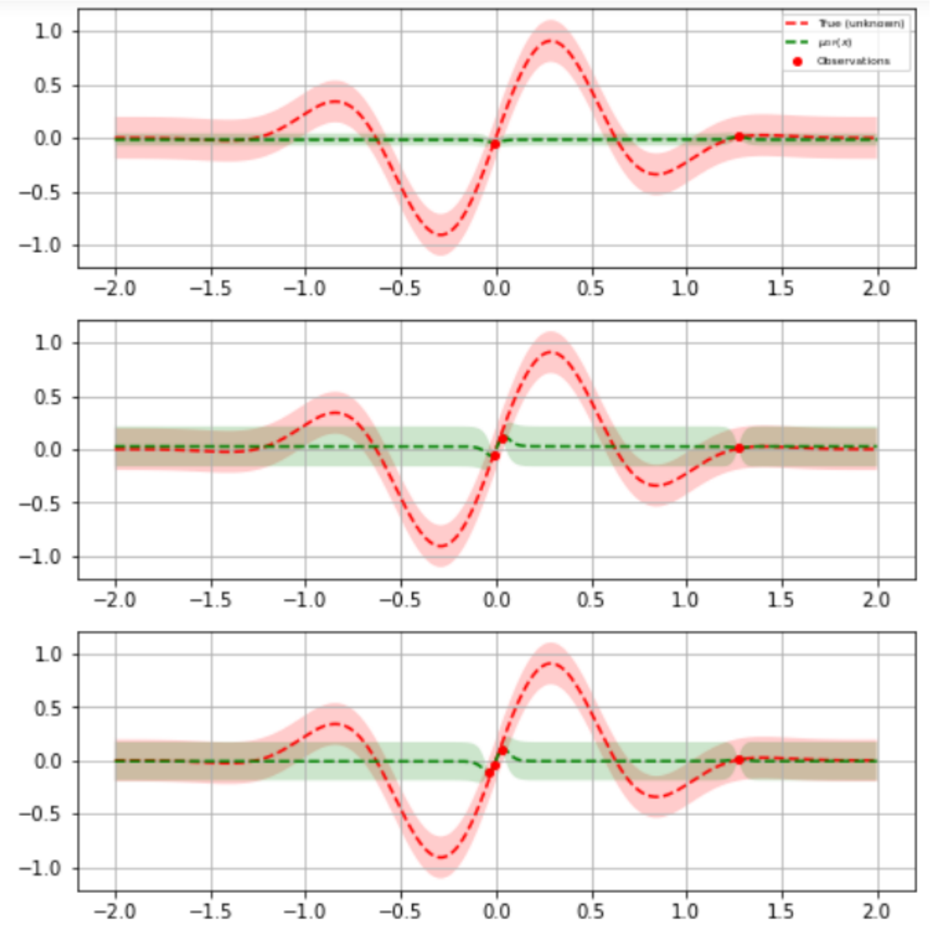
f(x)



x

Hyperparameter Optimization

f(x)

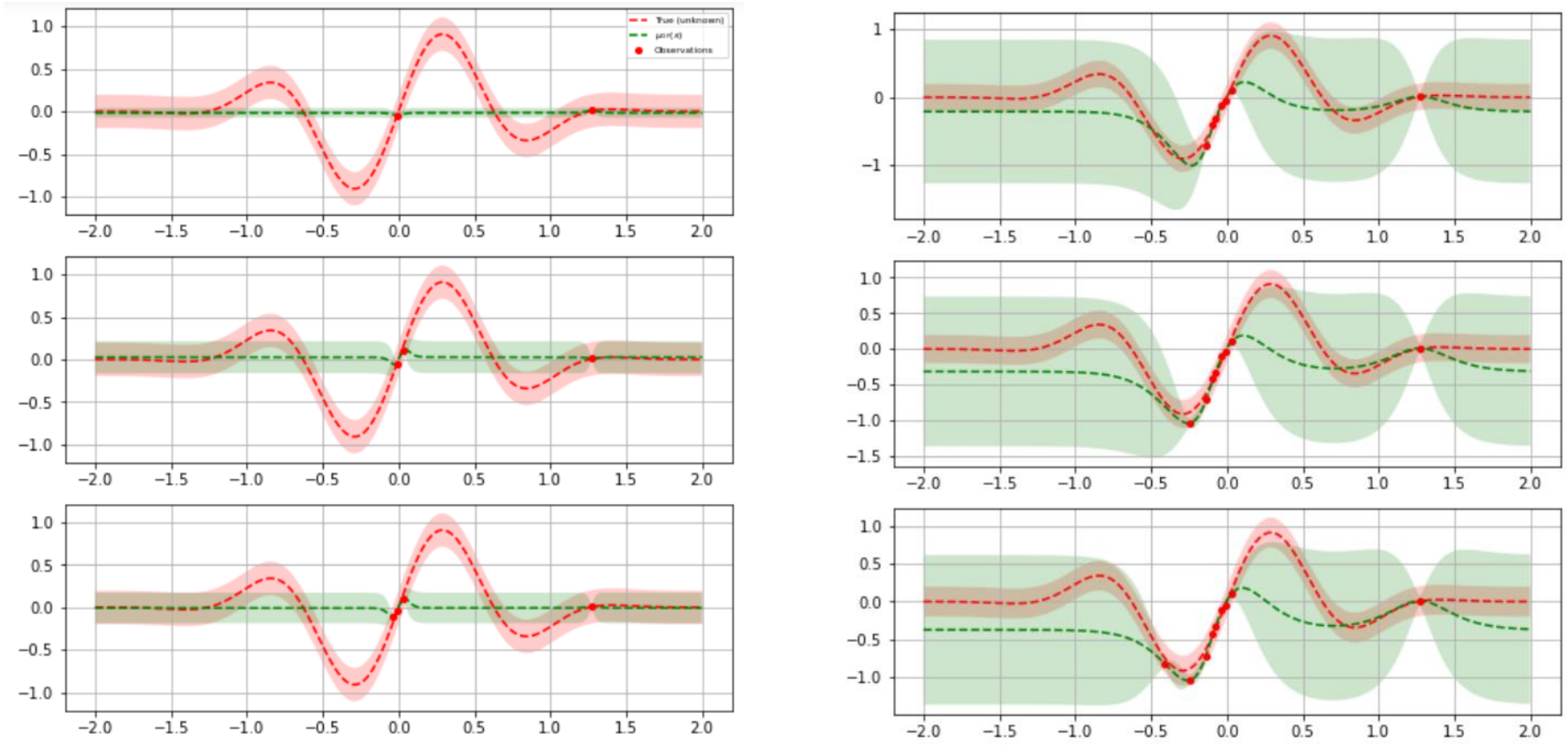


x

- *Red is f(x)*
- *The green area is our Prior, P(w)*
- *D is the red dots, where we evaluate f(x)*
- *We adjust the posterior P(w | D) based on D.*
- Note how with Bayes Optimization we find the minimum of f(x).

Hyperparameter Optimization

f(x)

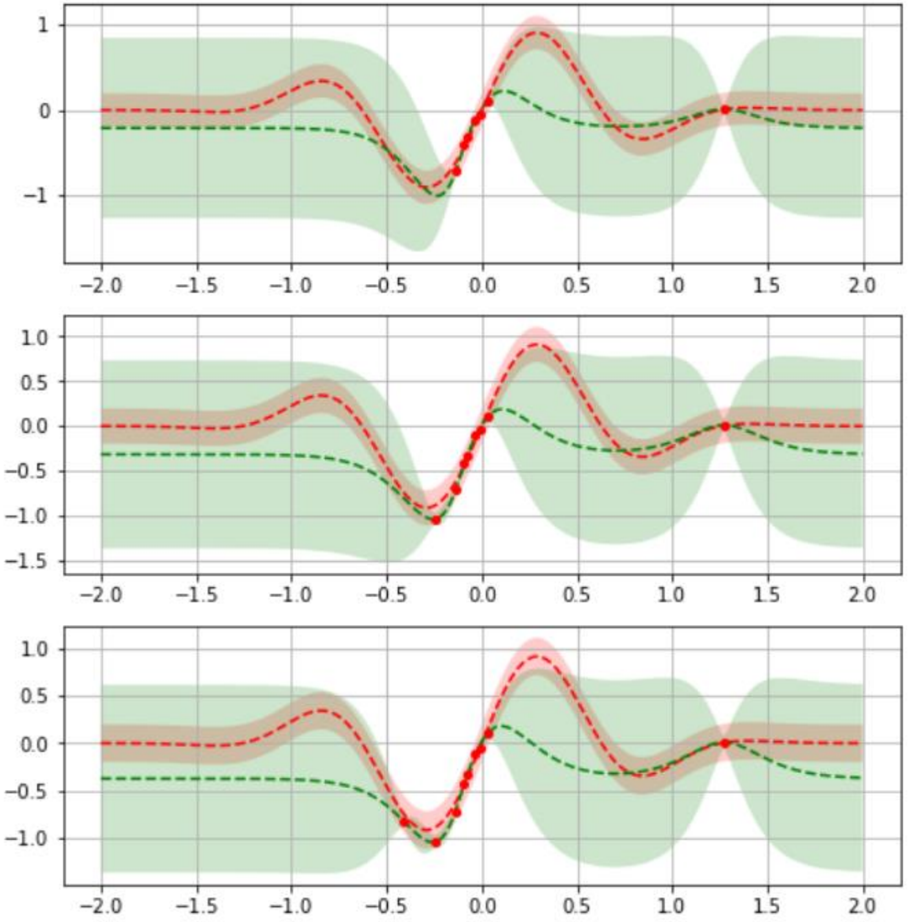


x

Hyperparameter Optimization

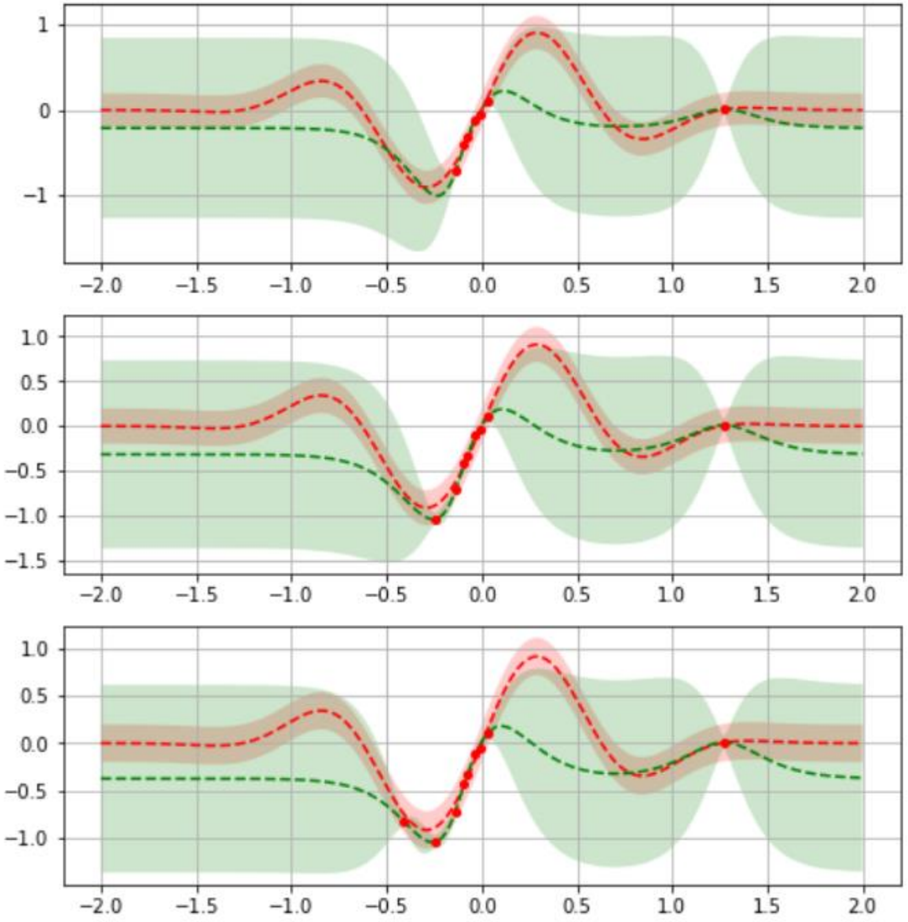
- *Red is f(x)*

- *As we gather more data, the posterior P(w | D) increases its uncertainty in unexplored areas and decreases its uncertainty in explored areas*

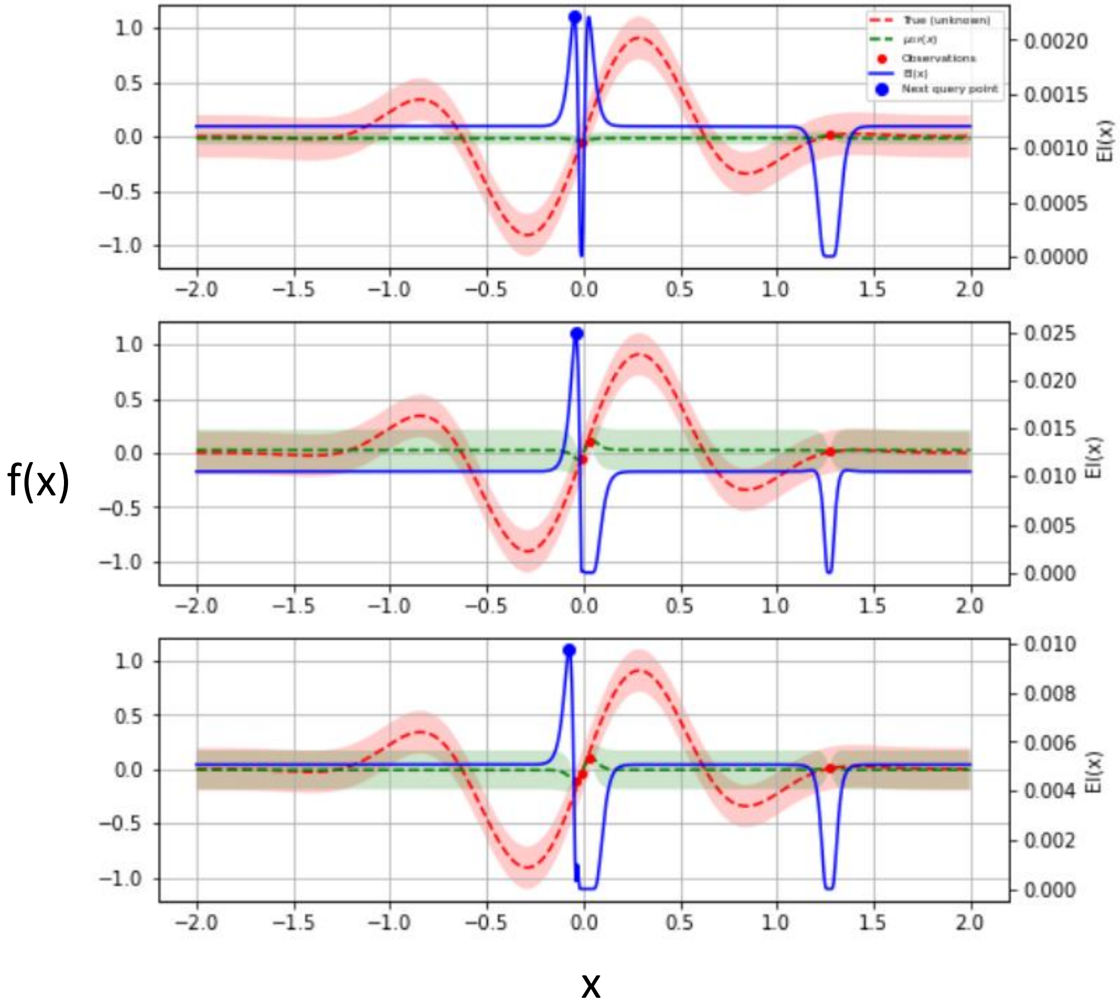


Hyperparameter Optimization

- *f(x) is unknown, we do not know its shape*
- *We need a way to produce a prior, and a posterior ➔ **Surrogate model for f(x)***
- *Gaussian Process Regression*

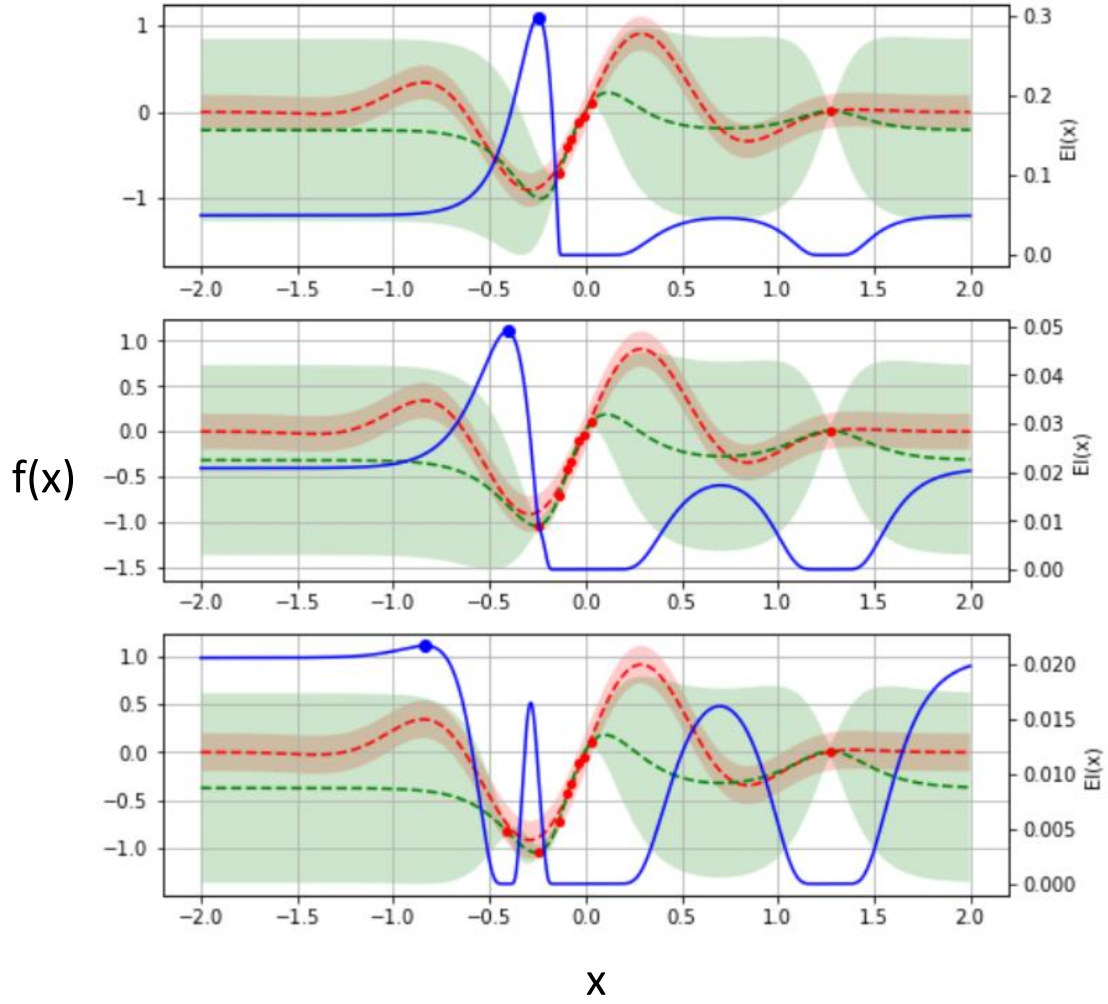


Hyperparameter Optimization



- *We also need a way to estimate where to sample next* ➔ ***acquisition function***

Hyperparameter Optimization



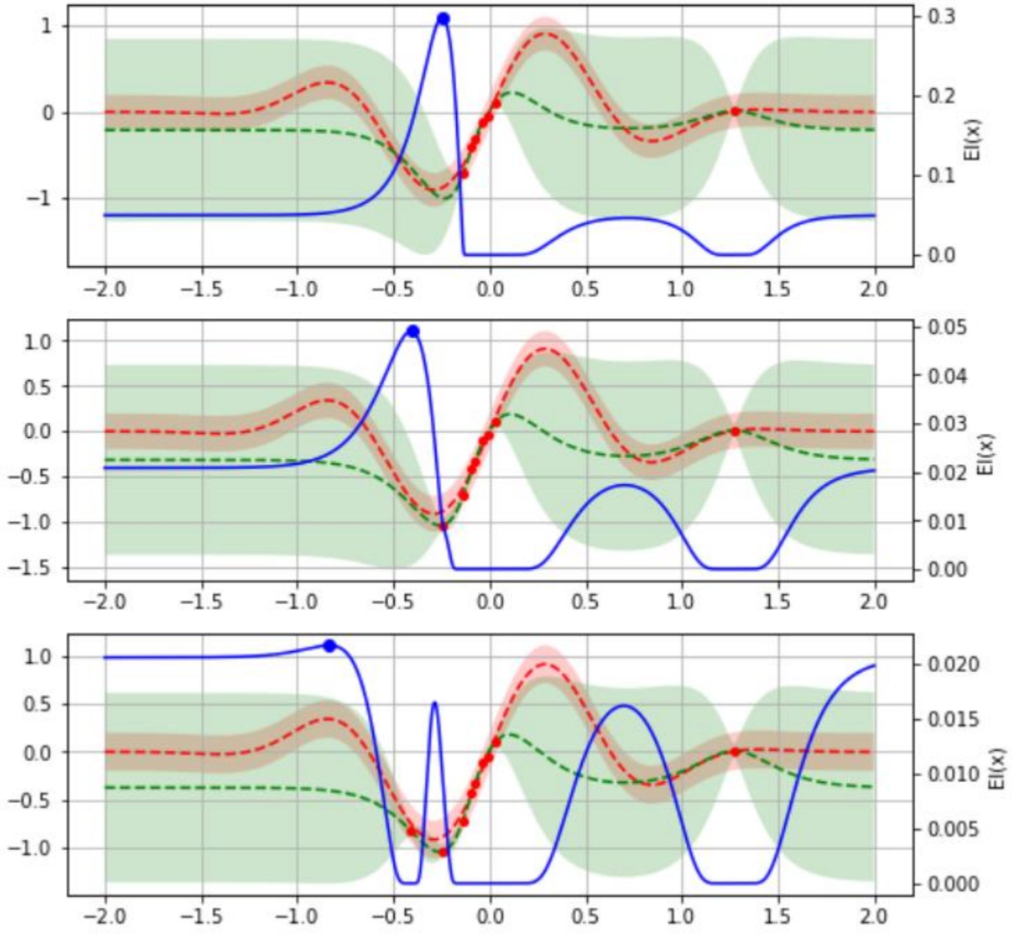
- *We also need a way to estimate where to sample next* ➔ ***acquisition function***

- *Acquisition function values are high where the mean of f(x) is low (or high)* ➔ ***exploitation***.

- *Acquisition function values are high when the variance of f(x) is high* ➔ ***exploration***.

Hyperparameter Optimization

- *f(x) is unknown, we do not know its shape*
- *We need a way to produce a prior, and a posterior ➔ **Surrogate model for f(x)***
  - *Gaussian Process Regression*
- *We need an acquisition function to determine where to sample next*
- ***The surrogate and the acquisition function need to be cheaper to evaluate than f(x)***



THANK YOU

www.trainindata.com