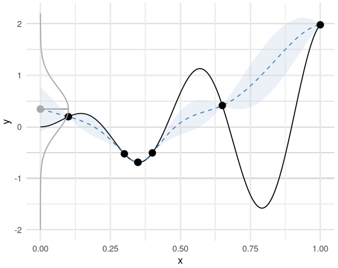
Optimization in Machine Learning

Bayesian Optimization
Posterior Uncertainty and Acquisition Functions I

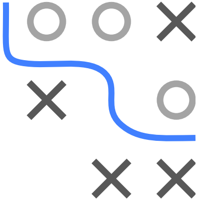


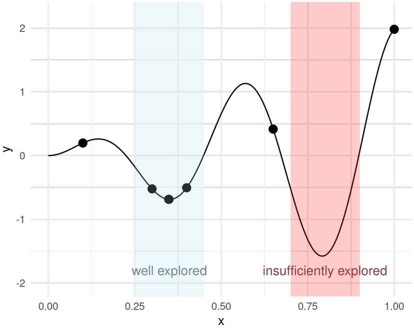
Learning goals

- Bayesian surrogate modeling
- Acquisition functions
- Lower confidence bound

BAYESIAN SURROGATE MODELING

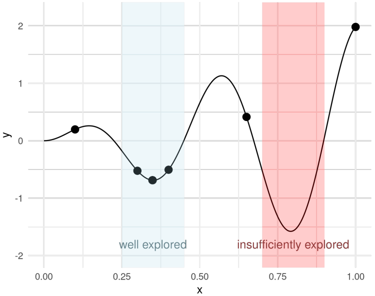
Goal:
Find trade-off between **exploration** (areas we have not visited yet) and **exploitation** (search around good design points)

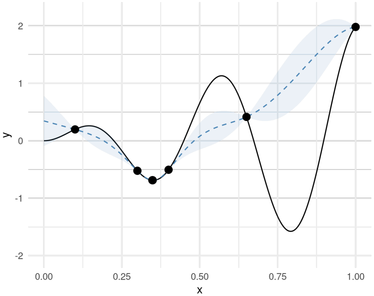




BAYESIAN SURROGATE MODELING / 2

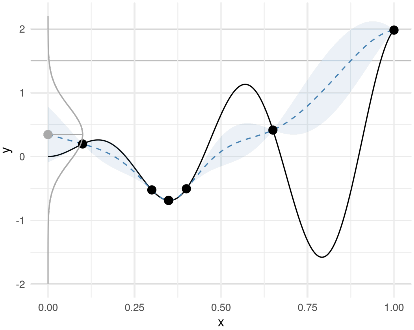
- **Idea**: Use a **Bayesian approach** to build SM that yields estimates for the posterior mean $\hat{f}(\mathbf{x})$ and the posterior variance $\hat{s}^2(\mathbf{x})$
- $\hat{s}^2(\mathbf{x})$ expresses “confidence”/“certainty” in prediction





BAYESIAN SURROGATE MODELING / 3

- Denote by $Y \mid \mathbf{x}, \mathcal{D}^{[t]}$ the (conditional) RV associated with the posterior predictive distribution of a new point $\mathbf{x}$ under a SM; will abbreviate it as $Y(\mathbf{x})$
- Most prominent choice for a SM is a **Gaussian process**, here $Y(\mathbf{x}) \sim \mathcal{N}\left(\hat{f}(\mathbf{x}), \hat{s}^2(\mathbf{x})\right)$

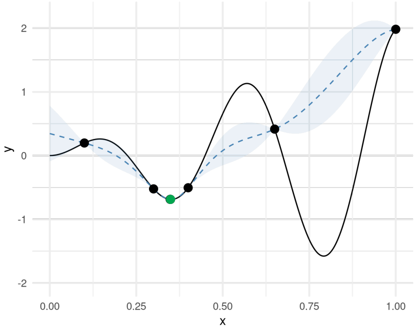


For now we assume an interpolating SM; $\hat{f}(\mathbf{x}) = f(\mathbf{x})$ and $\hat{s}(\mathbf{x}) = 0$ for training points

ACQUISITION FUNCTIONS

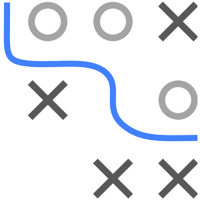
To sequentially propose new points based on the SM, we make use of so-called acquisition functions $a : \mathcal{S} \to \mathbb{R}$

Let $f_{\min} := \min \left\{ f(\mathbf{x}^{[1]}), \ldots, f(\mathbf{x}^{[t]}) \right\}$ denote the best observed value so far (visualized in green - we will need this later!)



In the examples before we simply used the posterior mean $a(\mathbf{x}) = \hat{f}(\mathbf{x})$ as acquisition function - ignoring uncertainty

LOWER CONFIDENCE BOUND



Goal: Find $\mathbf{x}^{[t+1]}$ that minimizes the **Lower Confidence Bound** (LCB):

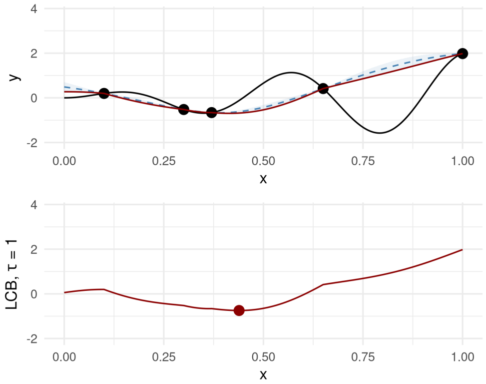
$$a_{\text{LCB}}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \tau \hat{s}(\mathbf{x})$$

where $\tau > 0$ is a constant that controls the “mean vs. uncertainty” trade-off

The LCB is conceptually very simple and does **not** rely on distributional assumptions of the posterior predictive distribution under a SM

LOWER CONFIDENCE BOUND / 2

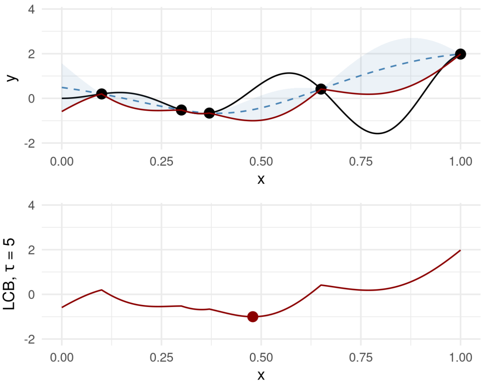
$\tau = 1$



Top: Design points and SM showing $\hat{f}(\mathbf{x})$ (blue) and $\hat{f}(\mathbf{x}) - \tau\hat{s}(\mathbf{x})$ (red)
Bottom: the red point depicts $\arg\min_{\mathbf{x}\in\mathcal{S}} a_{\text{LCB}}(\mathbf{x})$

LOWER CONFIDENCE BOUND / 3

$\tau = 5$



LOWER CONFIDENCE BOUND / 4

$\tau = 10$