Applications of Gaussian Processes

EN.540.782: Statistical Uncertainty Quantification

N. Wichrowski

Johns Hopkins University

April 8, 2021

Outline

Review: Gaussian Processes

Definition

A Gaussian Process (GP) is a collection of random variables $\{X_\alpha \mid \alpha \in \mathcal{X}\}$ any finite number of which have a joint Gaussian distribution.

- A GP is completely specified by:
  - mean function: $m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$
  - covariance kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}[f(\boldsymbol{x}), f(\boldsymbol{x}')]$
- Use conditional distributions to predict values at new inputs:
  - Observations: inputs $X \in \mathbb{R}^{n \times d}$, outputs $\boldsymbol{y} \in \mathbb{R}^n$.
  - For new inputs $X_\star \in \mathbb{R}^{n_\star \times d}$,

$$\mathbb{E}[\boldsymbol{y}_\star | X_\star, X, \boldsymbol{y}] = K(X_\star, X) K(X, X)^{-1} \boldsymbol{y}$$
$$\mathrm{Cov}[\boldsymbol{y}_\star | X_\star, X, \boldsymbol{y}] = K(X_\star, X_\star) - K(X_\star, X) K(X, X)^{-1} K(X, X_\star)$$

Problem Setting

- Objective $f : \mathcal{X} \to \mathbb{R}$ on domain $\mathcal{X} \subseteq \mathbb{R}^d$.
- We seek a global minimizer

$$\boldsymbol{x}^\star \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \,.$$

- Bayesian optimization (BO) is most appropriate when:
  - Evaluating f is expensive (and possibly noisy).
  - We have no derivative information.

Basic BO Algorithm

Algorithm 1: Bayesian Optimization

Data: objective function $f : \mathcal{X} \to \mathbb{R}$; initial sample $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

for $i \in \{n+1, \ldots, N\}$ **do**

  Fit a GPR model to data $\mathcal{D}_{i-1}$;

  Choose a new point $\boldsymbol{x}_i \in \mathcal{X}$;

  Evaluate $y_i \leftarrow f(\boldsymbol{x}_i)$;

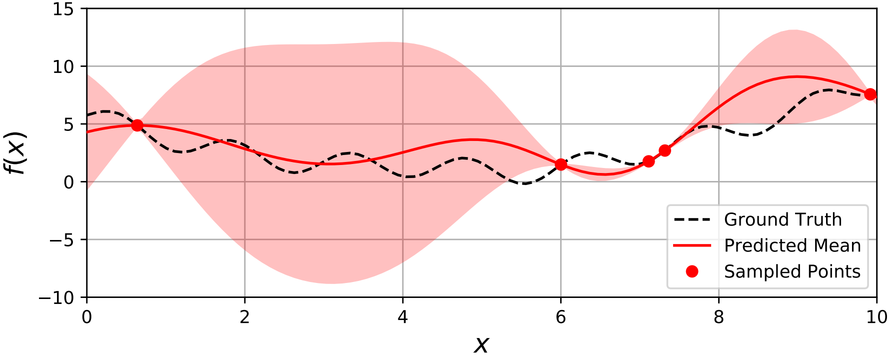
  Update sample $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(\boldsymbol{x}_i, y_i)\}$;

end

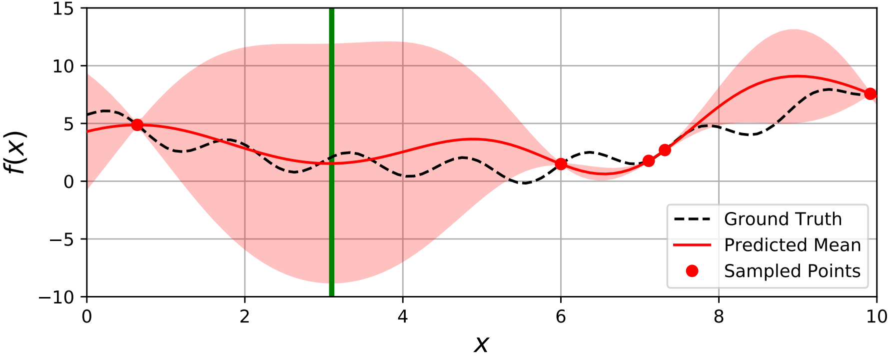
$j \leftarrow \operatorname*{argmin}_{1 \leq i \leq N} y_i$;

Result: approximate global minimizer $\boldsymbol{x}_j$

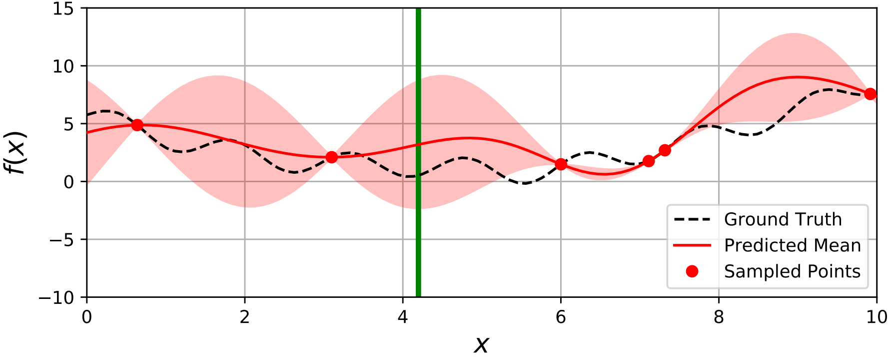
An Illustrated Example



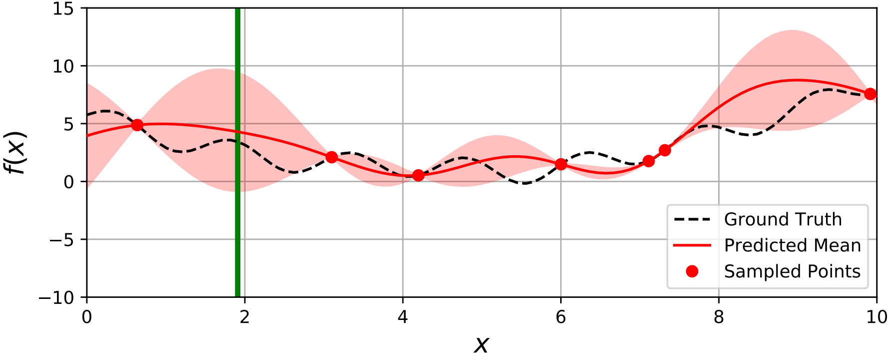
An Illustrated Example



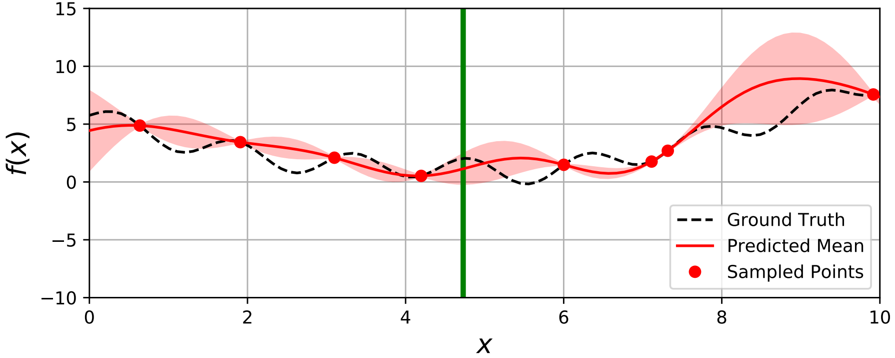
An Illustrated Example



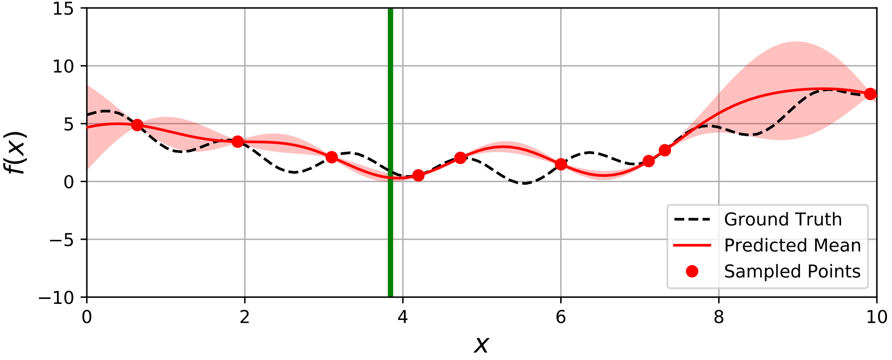
An Illustrated Example



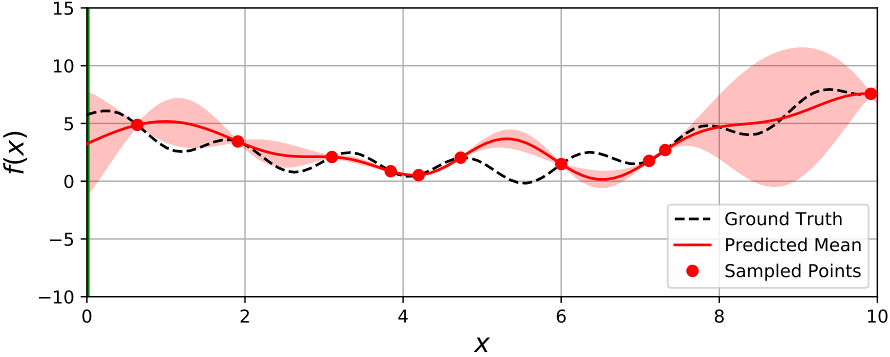
An Illustrated Example



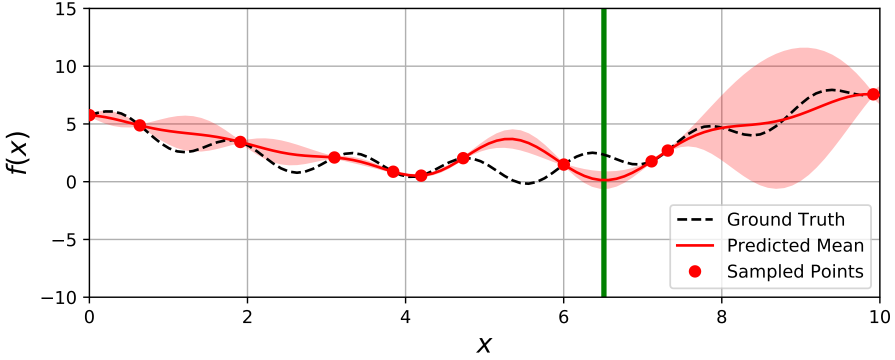
An Illustrated Example



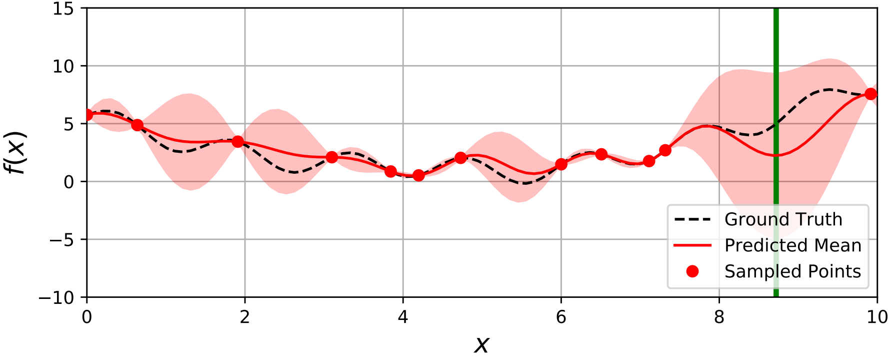
An Illustrated Example



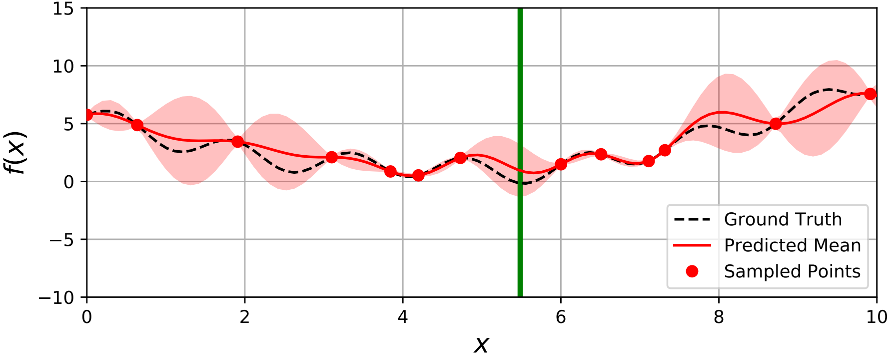
An Illustrated Example



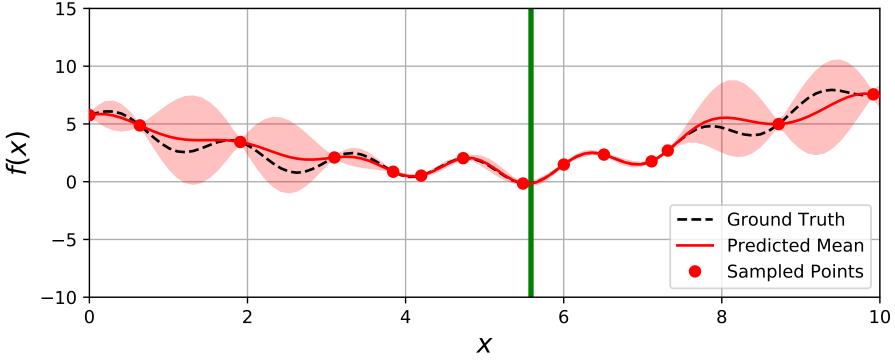
An Illustrated Example



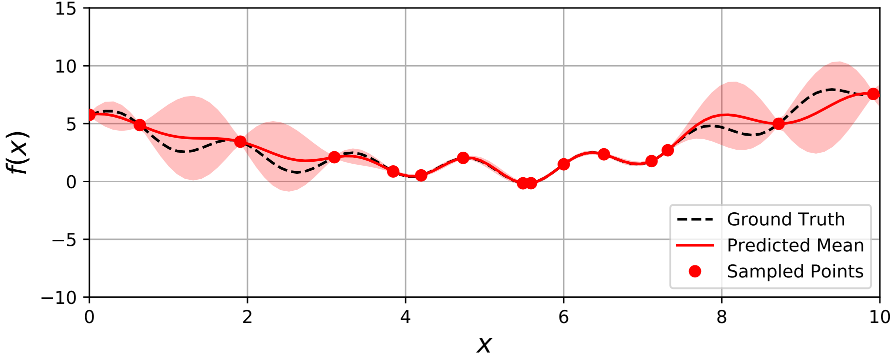
An Illustrated Example



An Illustrated Example



An Illustrated Example



Selecting the Next Sample Point

- Use current knowledge to inform our choice.
- Exploration vs. Exploitation
- An acquisition function measures utility of sampling at $\boldsymbol{x} \in \mathcal{X}$:

$$\boldsymbol{x}_{i+1} = \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}} \alpha(\boldsymbol{x}; \mathcal{D}_i)$$

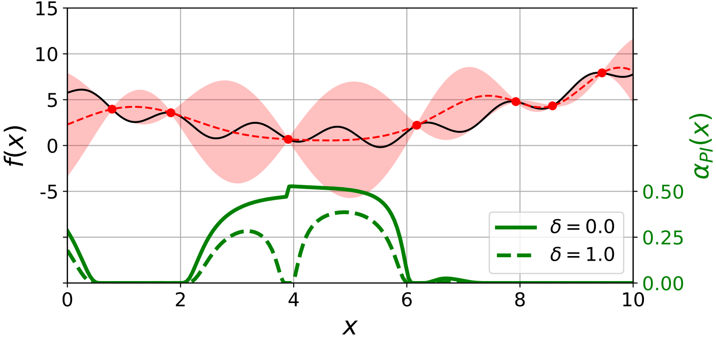
- How is this better? Optimizing α does not require evaluating f.

Acquisition Function: Probability of Improvement

- Which inputs are likely to improve the current best observation?
- Gaussian distribution provides an analytical expression:

$$\alpha_{PI}(\boldsymbol{x}; \mathcal{D}) = \mathbb{P}\left[f(\boldsymbol{x}) < \tau \mid \mathcal{D}\right] = \Phi\left(\frac{\tau - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)$$

- Threshold τ usually written as $\tau = y_{\min} - \delta$ for $\delta \geq 0$.

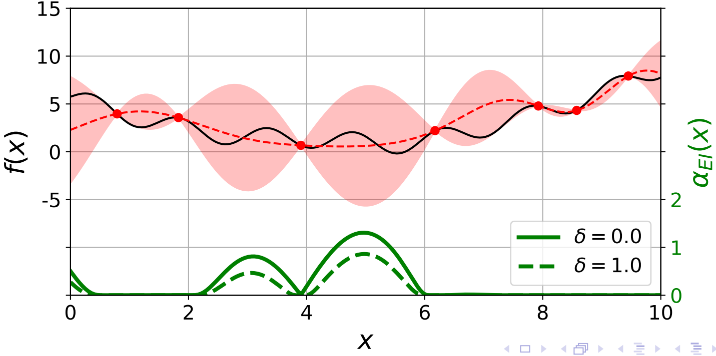


Acquisition Function: Expected Improvement

- Not all improvements are equally helpful.
- Expected improvement compared to threshold τ is

$$\alpha_{EI}(\boldsymbol{x};\mathcal{D}) = \mathbb{E}\left[\max(0,\, \tau - \mathcal{N}\left[\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\right])\right]$$
$$= (\tau - \mu(\boldsymbol{x}))\,\Phi\left(\frac{\tau - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) + \sigma(\boldsymbol{x})\,\phi\left(\frac{\tau - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right).$$

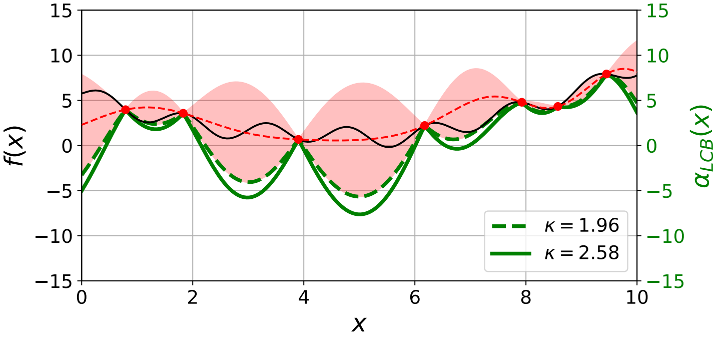
- Less prone than PI to become stuck in a local minimum.



Acquisition Function: Lower Confidence Bound

- Which inputs have superior best-case outputs?
- Usually expressed as a loss to be minimized:

$$\alpha_{LCB}(\boldsymbol{x}; \mathcal{D}) = \mu(\boldsymbol{x}) - \kappa\sigma(\boldsymbol{x})\,.$$



Computational Cost

- Inverting $K(X, X) \in \mathbb{R}^{n \times n}$ requires $O(n^3)$ operations.
- Cholesky decomposition must be updated each iteration.
- Approximation techniques exchange accuracy for speed.
- Using a sparse kernel for the GPR may help.

Pre-processing Data

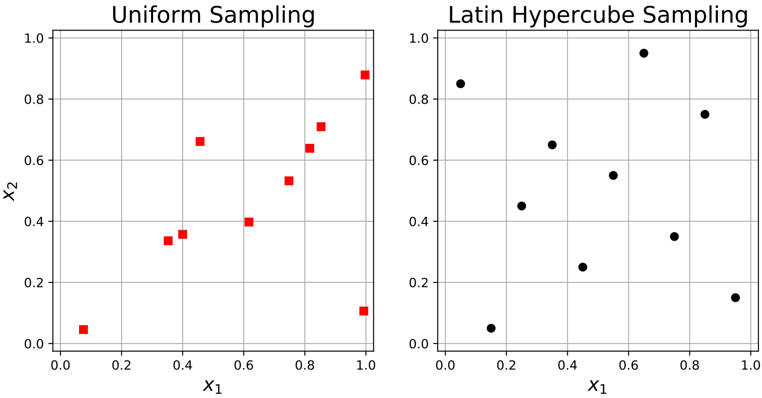
- Transforming sample data can facilitate GPR performance.
  - Normalize observations:

$$\tilde{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{s_x} \qquad \tilde{y}_i = \frac{y_i - \bar{y}}{s_y}$$

  - If outputs are necessarily positive: $\tilde{y}_i = \log(y_i)$.
  - Dimensionality reduction: focus on important directions.
- Apply knowledge of problem when deciding how to process data.
- Must reverse transformations to recover interpretable quantities.

Generating a Useful Initial Sample

- Samples drawn uniformly may not fill the domain.
- Latin hypercube sampling offers better samples for regression.



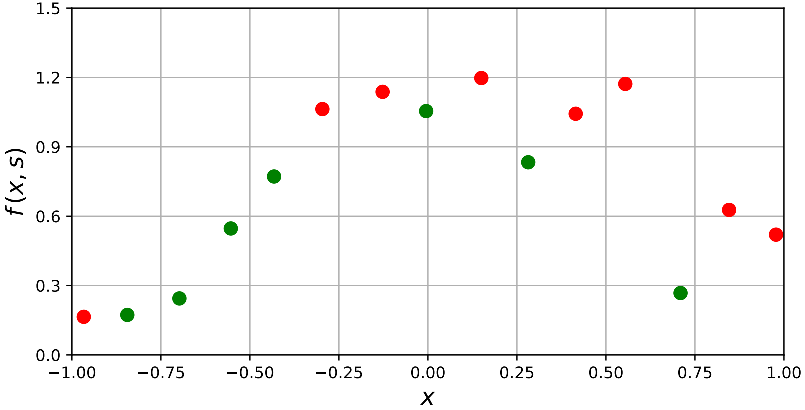
GPR with Qualitative Inputs

- Model a system as $f : \mathbb{R}^d \times S \to \mathbb{R}$ with $S = \{1, \ldots, n_s\}$.
- Can we do better than n_s separate fits?
- If the n_s surfaces are correlated, we can use expanded kernel

$$k\left((\boldsymbol{x}, s), (\boldsymbol{x}', s')\right) = C_{s,s'}\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')\,,$$

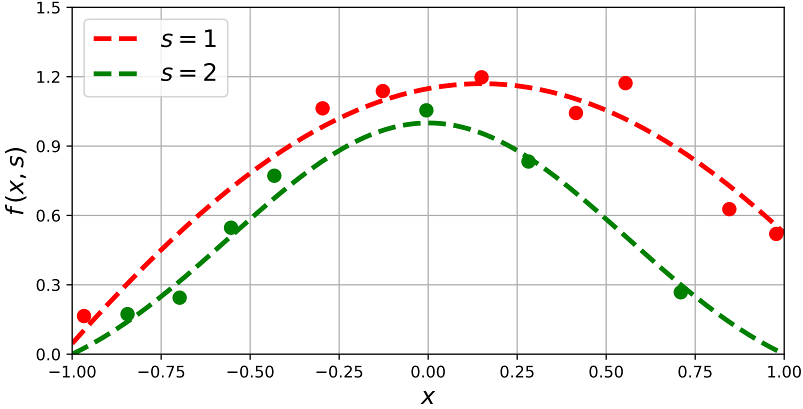
where $C \in \mathbb{R}^{n_s \times n_s}$ is a cross-correlation matrix.

A Mixed-Model Example



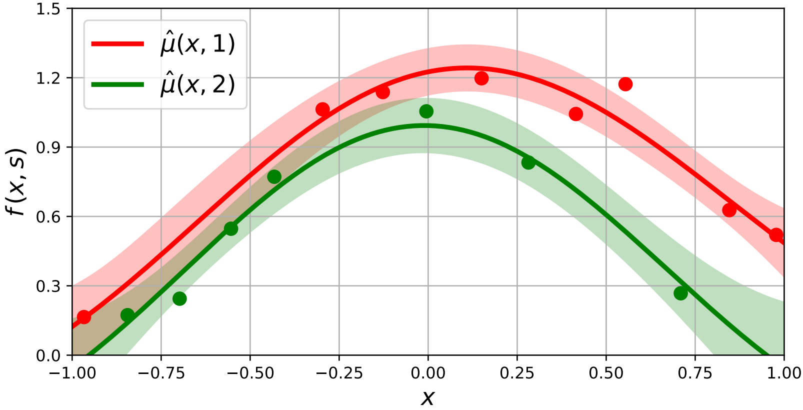
A Mixed-Model Example

$$f : [-1, 1] \times \{1, 2\} : (x, s) \mapsto \begin{cases} 1.9\cos(x - 0.15) - 0.73 & : s = 1 \\ (1 - x^2)e^{-x^2} & : s = 2 \end{cases}$$



A Mixed-Model Example

$$f : [-1, 1] \times \{1, 2\} : (x, s) \mapsto \begin{cases} 1.9\cos(x - 0.15) - 0.73 & : s = 1 \\ (1 - x^2)e^{-x^2} & : s = 2 \end{cases}$$

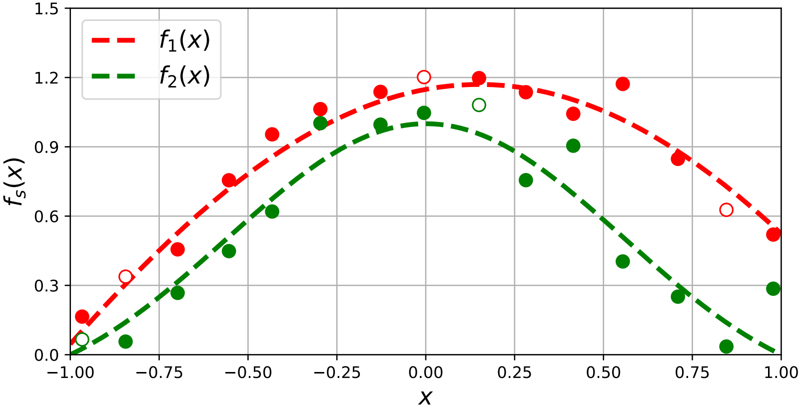


Multiple Regression

- The qualitative input scheme also works for multiple outputs:

$$\boldsymbol{y} = \mathbf{f}(\boldsymbol{x}) \in \mathbb{R}^{d'} \quad \longleftrightarrow \quad y_s = f(\boldsymbol{x}, s)$$

- Such an approach can accommodate missing output data.



References

- Frazier, P.I. "A Tutorial on Bayesian Optimization," *arXiv Preprint*, 22 pp. (2018). arXiv:1807.02811
- Luong, P.; Gupta, S.; Nguyen, D.; Rana, S.; Venkatesh, S. "Bayesian Optimization with Discrete Variables," in *AI 2019: Advances in Artificial Intelligence*, 473–84 (2019).
- Močkus, J. "On Bayesian Methods for Seeking the Extremum," in *Optimization Techniques: IFIP Technical Conference*, 400–04 (1975).
- Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*, MIT Press: Cambridge, MA (2006).
- Santner, T.J.; Williams, B.J.; Notz, W.I. *The Design and Analysis of Computer Experiments*," 2nd ed., Springer: New York, NY (2018).
- Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proc. IEEE*, 104(1), 148–75 (2015).