Introduction to Gaussian Processes

Noah J. Wichrowski

Johns Hopkins University

August 25, 2022

Outline

Prelude: Linear Regression

- Consider data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$.
- We want the weight vector $\boldsymbol{w} \in \mathbb{R}^d$ that yields the best linear fit

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}\,.$$

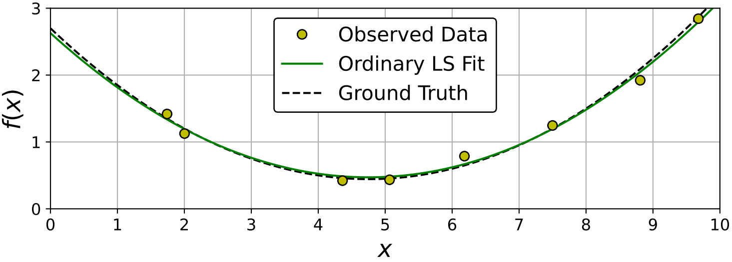
- Let $X \in \mathbb{R}^{n \times d}$ be the design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ the response vector.
- Then $\widehat{\boldsymbol{w}} = X^\dagger \boldsymbol{y} = (X^\top X)^{-1} X^\top \boldsymbol{y}$ minimizes the squared error:

$$X^\dagger \boldsymbol{y} \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2\,.$$

Linear Regression with Basis Functions

- Consider data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \mathbb{R}$.
- For weight vector $\boldsymbol{w} \in \mathbb{R}^N$ and basis functions $\phi_j : \mathbb{R}^d \to \mathbb{R}$,

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{j=1}^{N} w_j \phi_j(\boldsymbol{x})\,.$$



Bayesian Linear Regression

- The standard linear model for "frequentist" regression is given by

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^\top \boldsymbol{x}\,, \\ y_i &= f(\boldsymbol{x}_i) + \varepsilon_i\,, \\ \varepsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)\,. \end{aligned}$$

- In a Bayesian framework, we have $\boldsymbol{y} \,|\, X, \boldsymbol{w} \sim \mathcal{N}(X\boldsymbol{w}, \sigma_\varepsilon^2\, I)$.
- With prior $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_p)$, the resulting posterior is

$$\begin{aligned} \boldsymbol{w} \,|\, X, \boldsymbol{y} &\sim \mathcal{N}(\bar{\boldsymbol{w}}, C)\,, \\ C^{-1} &= \sigma_\varepsilon^{-2} X^\top X + \Sigma_p^{-1}\,, \\ \bar{\boldsymbol{w}} &= \sigma_\varepsilon^{-2} C X^\top \boldsymbol{y}\,. \end{aligned}$$

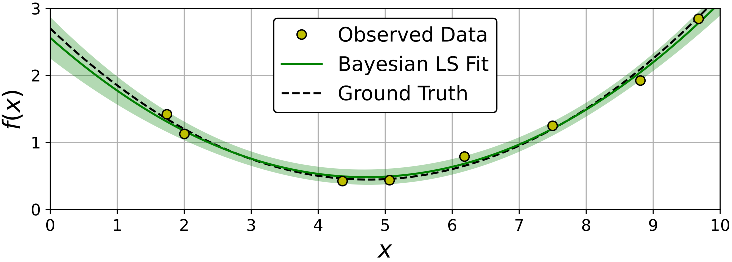
[Rasmussen & Williams, 2006]

Bayesian Linear Regression: Prediction

- We can use the posterior on $\boldsymbol{w}$ to predict new observations.
- For a new input $\boldsymbol{x}_\star \in \mathbb{R}^d$,

$$y_\star \,|\, \boldsymbol{x}_\star, X, \boldsymbol{y} \sim \mathcal{N}(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_\star,\ \boldsymbol{x}_\star^\top C \boldsymbol{x}_\star)\,.$$

[Rasmussen & Williams, 2006]



Gaussian Processes

Definition (Rasmussen and Williams, 2006)

A Gaussian process (GP) is a collection of random variables $\{Y_{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{X}\}$, any finite number of which have a joint Gaussian distribution.

- The index set $\mathcal{X}$ is often an interval $T \subseteq \mathbb{R}$.
- A GP is completely specified by:
  - mean function: $m(\boldsymbol{x}) = \mathbb{E}[Y_{\boldsymbol{x}}]$
  - covariance kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}[Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'}]$

Gaussian Processes

Definition (Rasmussen and Williams, 2006)

A Gaussian process (GP) is a collection of random variables $\{ Y_{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{X} \}$, any finite number of which have a joint Gaussian distribution.

- The index set $\mathcal{X}$ is often an interval $T \subseteq \mathbb{R}$.
- A GP is completely specified by:
  - mean function: $m(\boldsymbol{x}) = \mathbb{E}\left[Y_{\boldsymbol{x}} \right]$
  - covariance kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}[Y_{\boldsymbol{x}}, Y_{\boldsymbol{x}'}]$

Example (Pavliotis, 2014)

Brownian Motion on $\mathbb{R}$:

$$m(t) = 0 \qquad\qquad k(t, t') = \min(t, t')$$

Ornstein-Uhlenbeck Process:

$$m(t) = x_0 e^{-\alpha t} \qquad\qquad k(t, t') = \frac{1}{\alpha\beta}\left(e^{-\alpha|t-t'|} - e^{-\alpha(t+t')} \right)$$

Function-Space View of GPs

- Any GP defines a distribution over functions $f : \mathcal{X} \to \mathbb{R}$:

$$f(\boldsymbol{x}|\omega) = Y_{\boldsymbol{x}}(\omega)$$

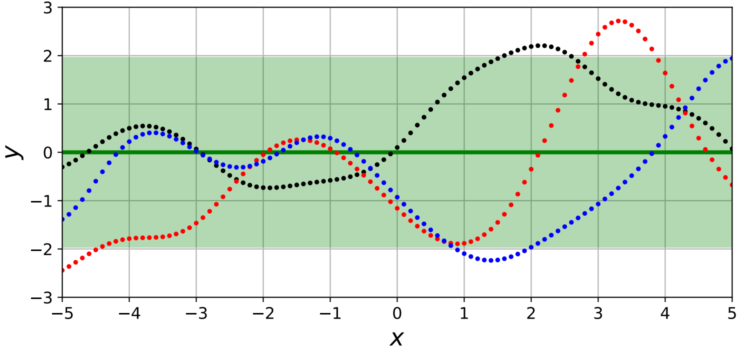
- How do we sample $f \sim \mathcal{GP}\left[m(\cdot), k(\cdot,\cdot)\right]$?
  - Choose $\mathcal{D}_X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.
  - Compute $m_i = m(\boldsymbol{x}_i)$, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
  - Draw $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}, K)$ and take $f(\boldsymbol{x}_i) = y_i$.
- Use conditional distributions to sample $y_\star = f(\boldsymbol{x}_\star)$ for $\boldsymbol{x}_\star \notin \mathcal{D}_X$:

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

[Rasmussen & Williams, 2006]

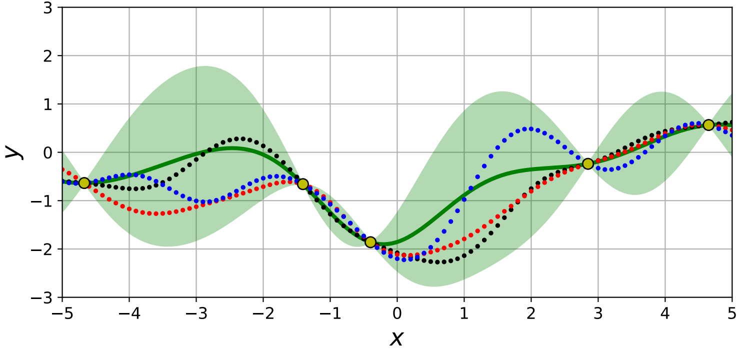
Sampling from a GP

- Prior distribution: $m(x) \equiv 0,\ k(x, x') = \exp\left[-\frac{1}{2}(x - x')^2\right]$
- Can draw (discretized) functions from prior or posterior.
- Plotted points are entries from MVN vectors.



Sampling from a GP

- Prior distribution: $m(x) \equiv 0$, $k(x, x') = \exp\left[-\frac{1}{2}(x - x')^2\right]$
- Can draw (discretized) functions from prior or posterior.
- Plotted points are entries from MVN vectors.



Covariance Functions: Concepts

- A kernel must be symmetric and positive semidefinite.
  - For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$.
  - Given measure μ, for all $f \in L^2(\mathcal{X}, \mu)$,

$$\iint_{\mathcal{X} \times \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}) f(\boldsymbol{x}') \, d\mu(\boldsymbol{x}) \, d\mu(\boldsymbol{x}') \geq 0 \,.$$

- We say k is stationary if $k(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{x} - \boldsymbol{x}')$.
- We say k is isotropic if $k(\boldsymbol{x}, \boldsymbol{x}') = g(\|\boldsymbol{x} - \boldsymbol{x}'\|)$.
- Kernels can be added, multiplied, and scaled:

$$k(\boldsymbol{x}, \boldsymbol{x}') = c_1^2 k_1(\boldsymbol{x}, \boldsymbol{x}') + c_2^2 k_2(\boldsymbol{x}, \boldsymbol{x}') k_3(\boldsymbol{x}, \boldsymbol{x}') \,.$$

[Duvenaud, 2014; Rasmussen & Williams, 2006]

Continuity and Differentiability

Definition (Rasmussen and Williams, 2006)

Let $f \sim \mathcal{GP}\left[m(\cdot), k(\cdot,\cdot)\right]$ be a Gaussian process on $\mathcal{X} \subseteq \mathbb{R}^d$. Then f is continuous in mean square (CMS) at $\boldsymbol{x}_\star \in \mathcal{X}$ if

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_\star} \mathbb{E}\left[\left|f(\boldsymbol{x}) - f(\boldsymbol{x}_\star)\right|^2\right] = 0\,.$$

We say that f is mean-square differentiable (MSD) at $\boldsymbol{x}_\star$ with partial derivatives $\partial f(\boldsymbol{x}_\star)/\partial x_i$ if, for $i \in \{1, \ldots, d\}$,

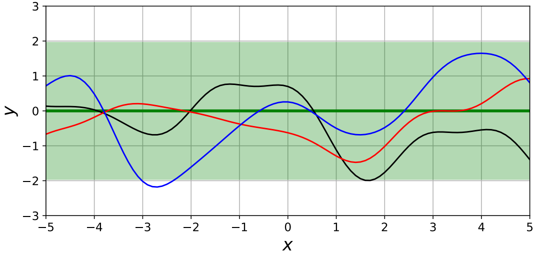
$$\lim_{h \to 0} \mathbb{E}\left[\left(\frac{f(\boldsymbol{x}_\star + h\boldsymbol{e}_i) - f(\boldsymbol{x}_\star)}{h} - \frac{\partial f(\boldsymbol{x}_\star)}{\partial x_i}\right)^2\right] = 0\,.$$

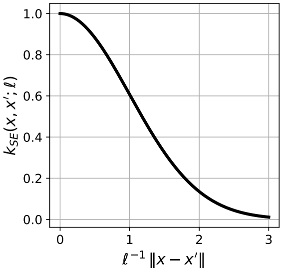
- GP with kernel k is CMS at $\boldsymbol{x}_\star \in \mathcal{X}$ iff k is continuous at $(\boldsymbol{x}_\star, \boldsymbol{x}_\star)$.
- A $2p$-order derivative of $k(\boldsymbol{x}_\star, \boldsymbol{x}_\star)$ ensures f is MSD p times at $\boldsymbol{x}_\star$.

Covariance Kernels: Squared Exponential

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left[-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right] \qquad \ell > 0$$

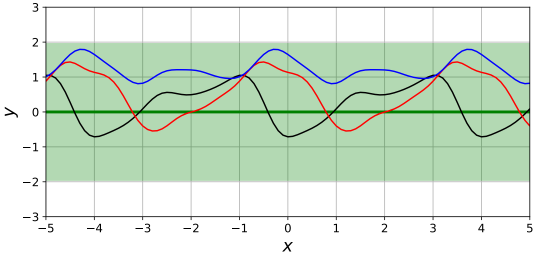
- Infinitely mean-square differentiable
- Often a limiting case of other kernel families

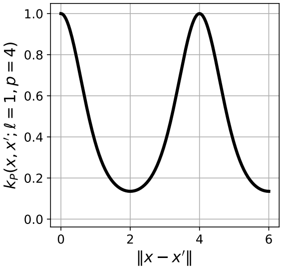




Covariance Kernels: Periodic

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left[-\frac{2\sin^2\left(\pi\|\boldsymbol{x} - \boldsymbol{x}'\|/p\right)}{\ell^2}\right] \qquad \ell, p > 0$$

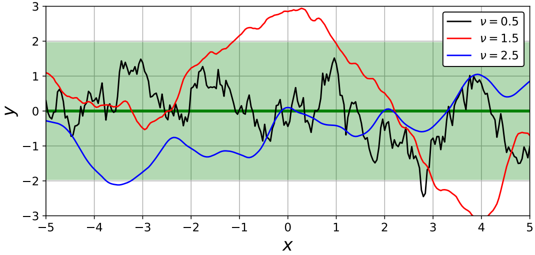


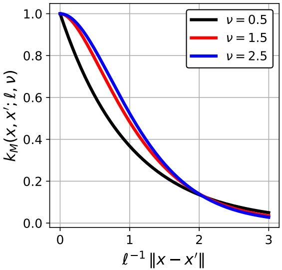


Covariance Kernels: Matérn

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r\sqrt{2\nu}}{\ell} \right)^{\nu} K_{\nu} \left(\frac{r\sqrt{2\nu}}{\ell} \right) \qquad \ell, \nu > 0$$

$K_{\nu}(\cdot)$ is a modified Bessel function





Matérn Kernel Smoothness Parameter

- Mean-square differentiable $\lceil \nu \rceil - 1$ times.
- Reduces to squared exponential as $\nu \to \infty$.
- Simpler forms for $\nu + \frac{1}{2} \in \mathbb{N}$:

$$k_{1/2}(r) = \exp\left[-\frac{r}{\ell}\right] \qquad k_{3/2}(r) = \left(1 + \frac{r\sqrt{3}}{\ell}\right)\exp\left[-\frac{r\sqrt{3}}{\ell}\right]$$

$$k_{5/2}(r) = \left(1 + \frac{r\sqrt{5}}{\ell} + \frac{5r^2}{3\ell^2}\right)\exp\left[-\frac{r\sqrt{5}}{\ell}\right]$$

[Rasmussen & Williams, 2006]

Generalizing Isotropic Kernels

- Most common kernels are defined as $k(\boldsymbol{x}, \boldsymbol{x}') = g(\|\boldsymbol{x} - \boldsymbol{x}'\|)$.
- What if some directions are more important than others?

$$k(\boldsymbol{x}, \boldsymbol{x}') = g\left(\sum_{i=1}^{d} \frac{|x_i - x_i'|}{\ell_i}\right)$$

- For interactions among directions, use a Mahalanobis metric, *e.g.*,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left[-(\boldsymbol{x} - \boldsymbol{x}')^\top M^{-1}(\boldsymbol{x} - \boldsymbol{x}')\right],$$

with M symmetric, positive definite.

Fitting GPs to Data

- Consider data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$.
- Recall the prediction for $y_\star = f(\boldsymbol{x}_\star)$ at $\boldsymbol{x}_\star \notin \mathcal{D}_X$:

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

- Must account for measurement noise when choosing a model.
- Given $\mathcal{D}$, the GPR model is specified by our choice of kernel.

[Goldberg *et al.*, 1997; Rasmussen & Williams, 2006]

Model Selection

- GP Regression requires:
  - Selecting a kernel function (model).
  - Tuning hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^p$.
- The log-marginal likelihood is

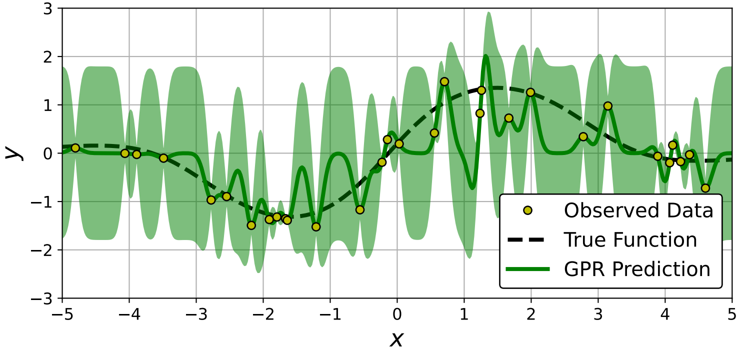
$$\log p(\boldsymbol{y} \mid X, \boldsymbol{\theta}) = -\frac{1}{2}\, \boldsymbol{y}^\top K_{\boldsymbol{\theta}}^{-1} \boldsymbol{y} - \frac{1}{2} \log |K_{\boldsymbol{\theta}}| - \frac{n}{2} \log(2\pi)\,,$$
$$K_{\boldsymbol{\theta}} = K(X, X; \boldsymbol{\theta}) + N(\boldsymbol{\theta})\,,$$

  where $N(\boldsymbol{\theta})$ models noise in outputs, $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.
- Cross-validation is also possible.
  - Block matrix inversion speeds up prediction.
  - Loss minimization requires expensive derivatives.

[Rasmussen & Williams, 2006]

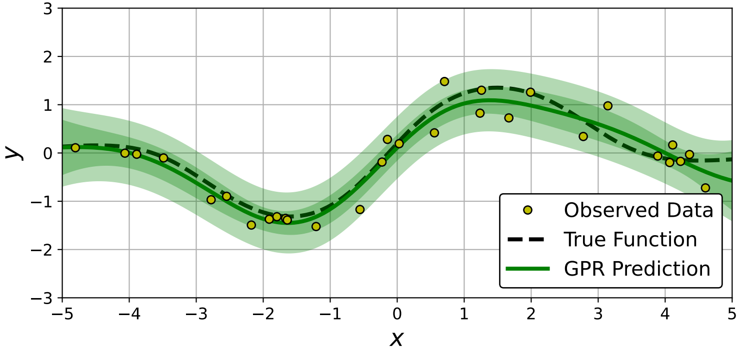
Model Selection: An Illustration

- Observed outputs are corrupted by iid $\mathcal{N}(0, 0.3^2)$ noise.
- Assuming outputs are noiseless: length scale $\ell = 0.0972$



Model Selection: An Illustration

- Observed outputs are corrupted by iid $\mathcal{N}(0, 0.3^2)$ noise.
- Including noise-level hyperparmater: length scale $\ell = 1.42$



Dealing with Noisy Outputs

- Often, we can only observe $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.
- Standard approach: assume independent $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$
- Predictive distribution is given by

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{y}_\star | X_\star, X, \boldsymbol{y}\right] &= K(X_\star, X) K_{\boldsymbol{y}}^{-1} \boldsymbol{y} \\
\mathrm{Cov}\left[\boldsymbol{y}_\star | X_\star, X, \boldsymbol{y}\right] &= K_\star - K(X_\star, X) K_{\boldsymbol{y}}^{-1} K(X, X_\star) \\
K_{\boldsymbol{y}} &= K(X, X) + \sigma_\varepsilon^2 I \\
K_\star &= K(X_\star, X_\star) + \sigma_\varepsilon^2 I
\end{aligned}
$$

Uncertainty vs. Variability

- Often wrongly used as synonyms.
  - Uncertainty: Lack of knowledge of a deterministic quantity
  - Variability: Differences in nominally interchangeable objects
- Distinct, but intertwined, concepts.
  - Uncertainty is quantified in probabilistic terms.
  - Variability is expressed in the language of statistics.
- Two main classes of uncertainty:
  - Aleatoric: difference in outcomes from repeated experiments
  - Epistemic: imprecise results from incomplete information

[Begg *et al.*, 2014; Der Kiureghian & Ditlevsen, 2009]

An Important Distinction

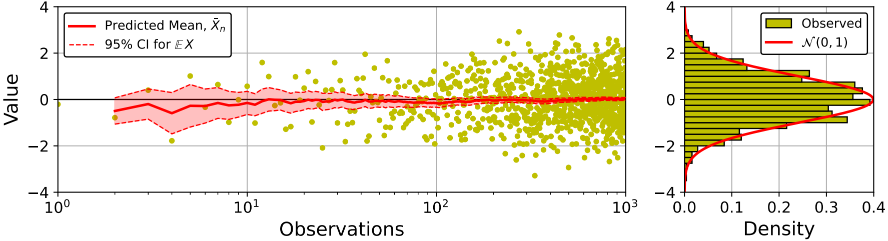
Consider independent variables $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

<u>Estimate the Mean</u>

$$\mathbb{V}\left[\bar{X}_n\right] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \to 0$$

<u>Predict the Next Value</u>

$$\mathbb{V}\left[X_{n+1}\right] = 1 \;\; \forall\, n \in \mathbb{N}$$



Uncertainty Quantification in GPR

- Predictive variance as a measure of confidence in model prediction
- For noiseless observations, no distinction: $\mathbb{V}[y] = \mathbb{V}[f(\boldsymbol{x})]$
- When noise is independent of input $\boldsymbol{x}$,

$$\mathbb{V}[y] = \mathbb{V}[f(\boldsymbol{x}) + \varepsilon] = \mathbb{V}[f(\boldsymbol{x})] + \sigma_\varepsilon^2$$

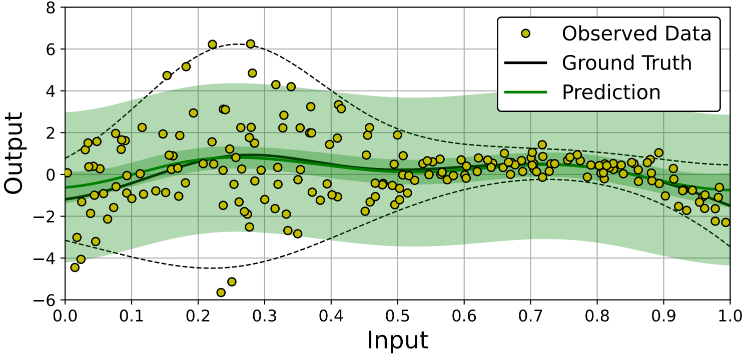
- We can quantify uncertainty "for the mean" or "for outputs."

What if the Noise Level Varies?

- Single hyperparameter σ_ε^2 implies homoscedasticity.
- Estimate $v_i \approx \sigma_\varepsilon^2(\boldsymbol{x}_i)$ and use $K_{\boldsymbol{y}} = K(X, X) + \text{diag}(\boldsymbol{v})$.
- For complete results, also need $v_\star \approx \sigma_\varepsilon^2(\boldsymbol{x}_\star)$ at test points.

GPR with Heteroscedastic Noise

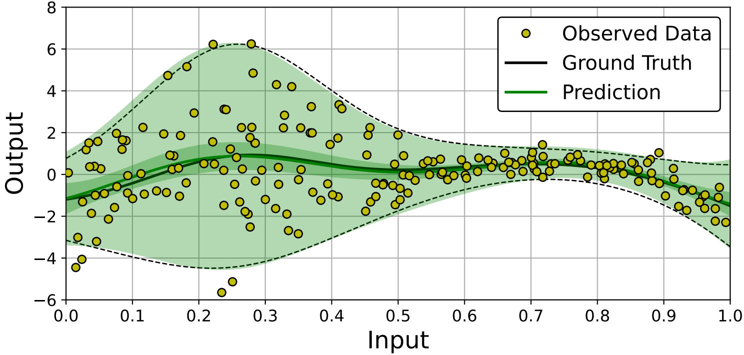
- Use a secondary GPR to model the varying noise level.
- Incorporate noise estimates into likelihood of primary model.



[Kersting *et al.*, 2007; Zhang *et al.*, 2020]

GPR with Heteroscedastic Noise

- Use a secondary GPR to model the varying noise level.
- Incorporate noise estimates into likelihood of primary model.



[Kersting *et al.*, 2007; Zhang *et al.*, 2020]

GPR with Qualitative Inputs

- Model a system as $f : \mathbb{R}^d \times S \to \mathbb{R}$ with $S = \{1, 2, \dots, n_s\}$.
- 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.

[Santner *et al.*, 2018]

Modeling Cross-Correlation

- Describe correlation between response surfaces $f(\cdot, s)$.
- Must have $C_{s,s} = 1$ and $|C_{s,s'}| \leq 1$ for all $s, s' \in S$.
- Options include:
  - Exchangable model:

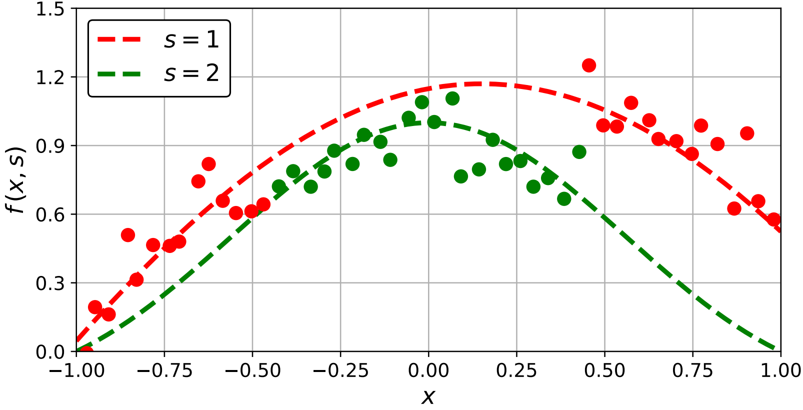
$$C_{s,s'} = \begin{cases} 1 & : s = s' \\ \rho & : s \neq s' \end{cases}$$

  - Toeplitz model for ordinal categories: $C_{s,s'} = e^{-\gamma|s-s'|}$
  - Specify all $\frac{1}{2}n_s(n_s - 1)$ superdiagonal entries.

[Qian *et al.*, 2008; Santner *et al.*, 2018]

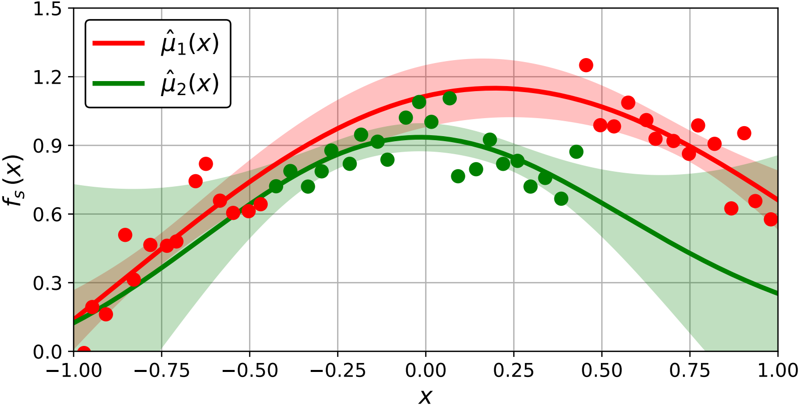
Mixed-Input Example: Data

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



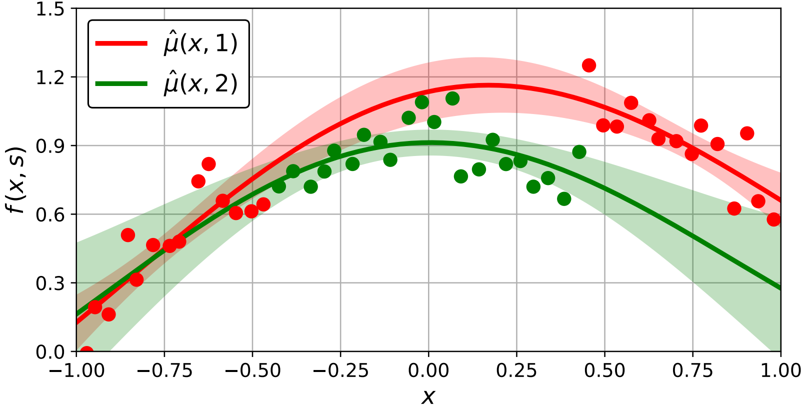
Mixed-Input Example: Disjoint Fits

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



Mixed-Input Example: Combined Fit

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

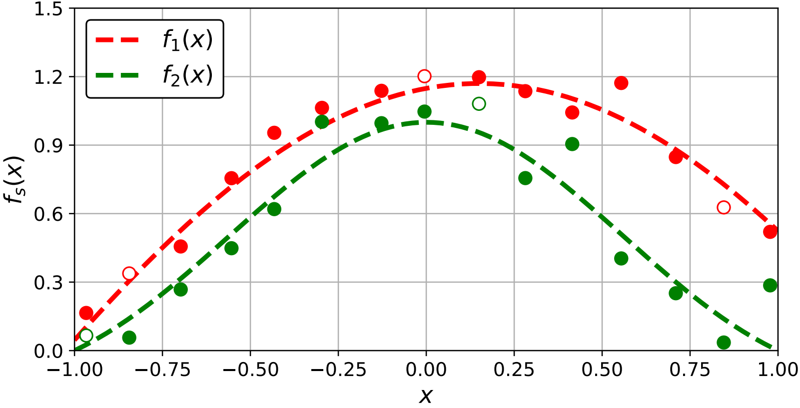


Multiple Regression

- 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.



[Santner *et al.*, 2018]

Bayesian Optimization

- 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.

[Frazier, 2018]

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

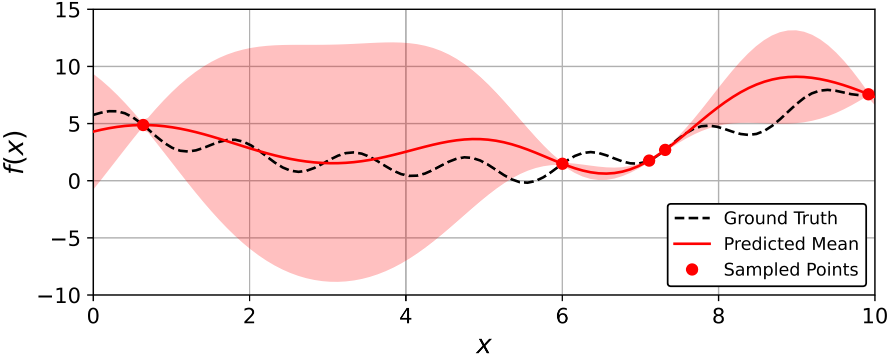
  Update sample $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(\boldsymbol{x}_i, y_i)\}$;

end

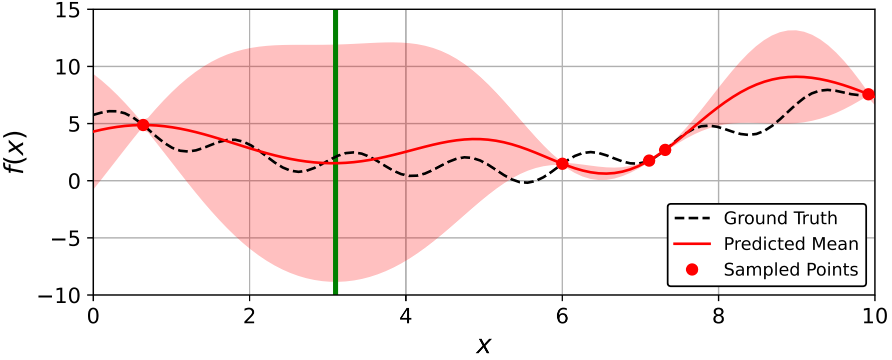
$j \leftarrow \operatorname*{argmin}_{1 \le i \le N} y_i$;

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

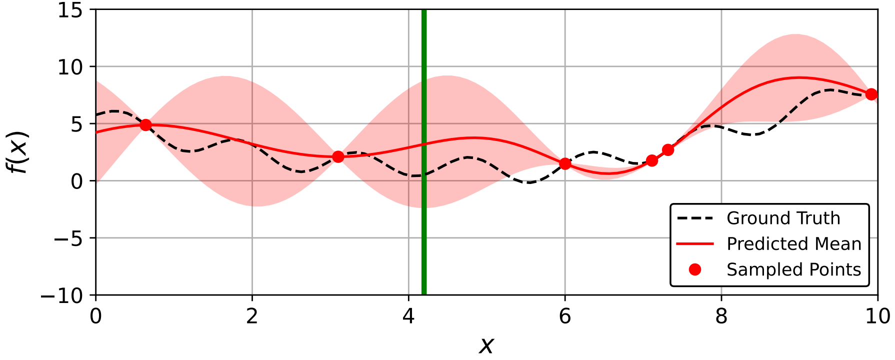
An Illustrated Example



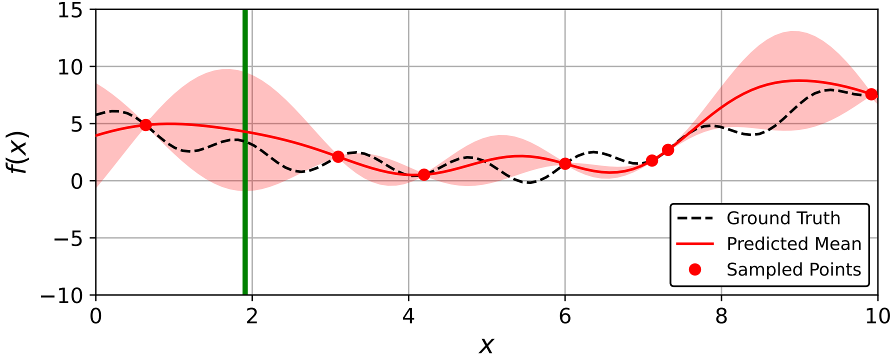
An Illustrated Example



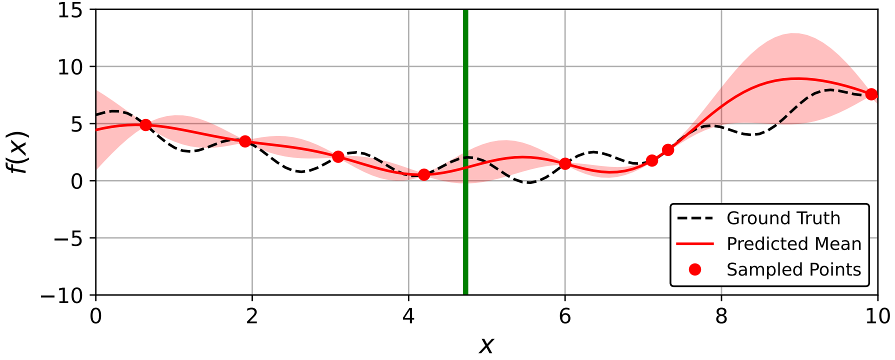
An Illustrated Example



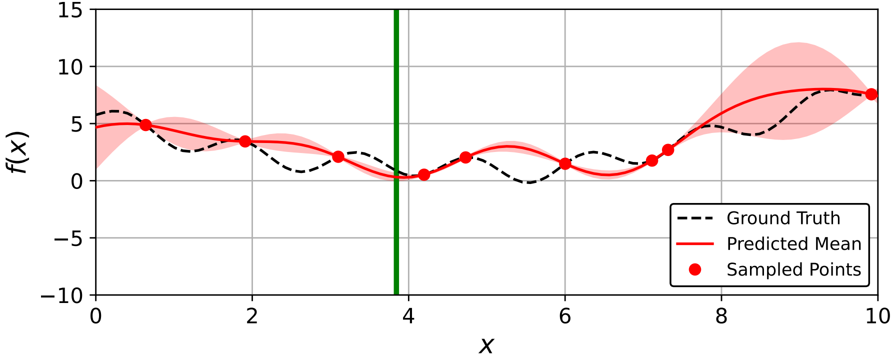
An Illustrated Example



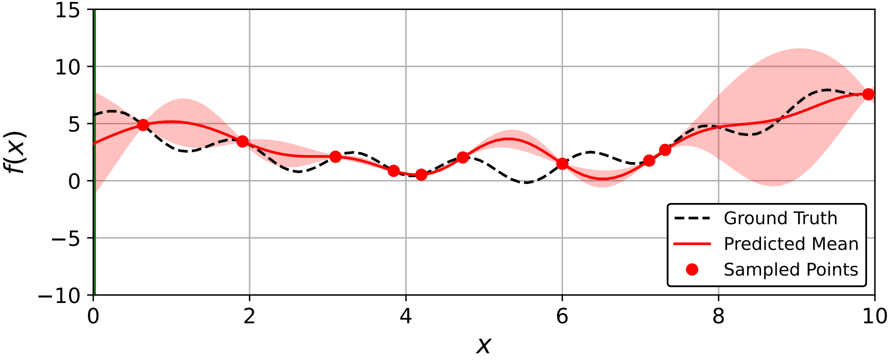
An Illustrated Example



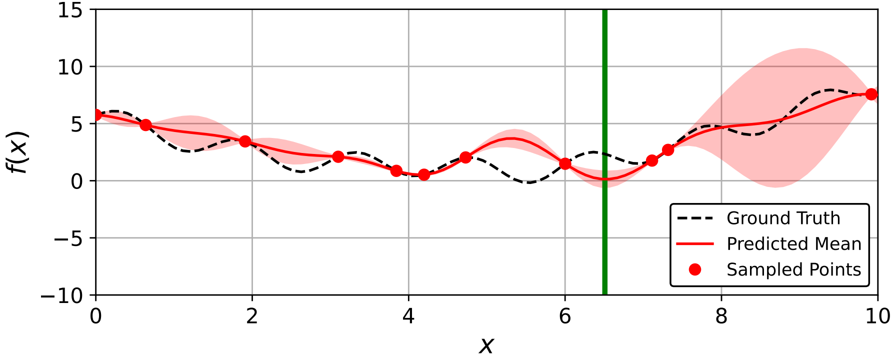
An Illustrated Example



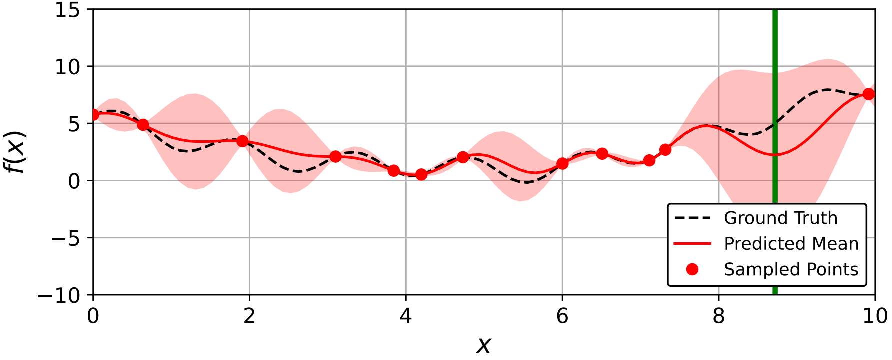
An Illustrated Example



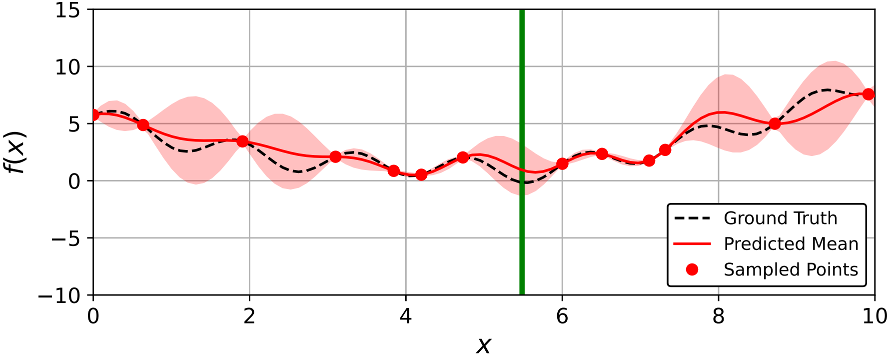
An Illustrated Example



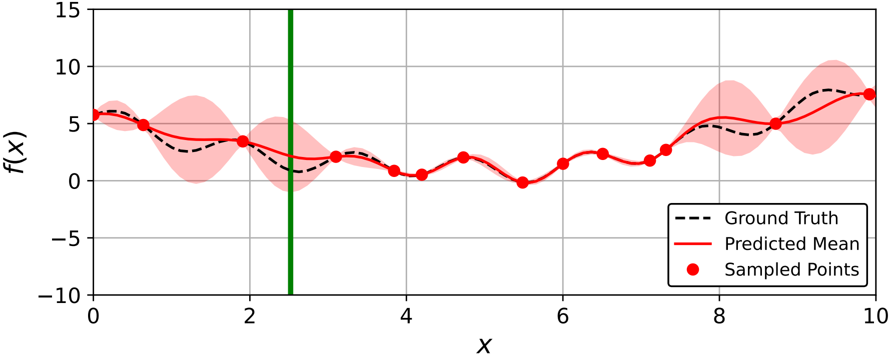
An Illustrated Example



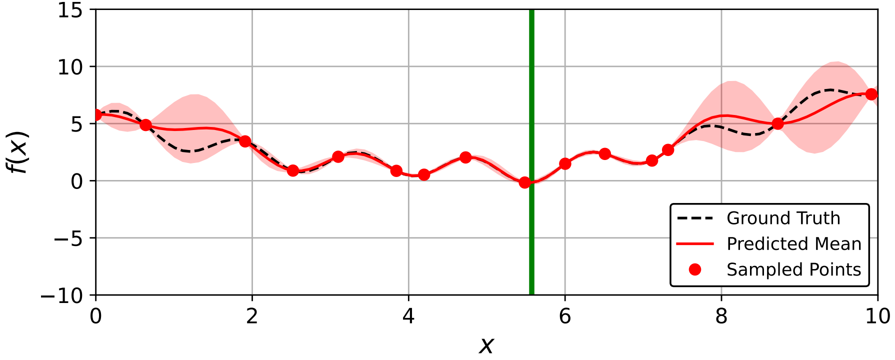
An Illustrated Example



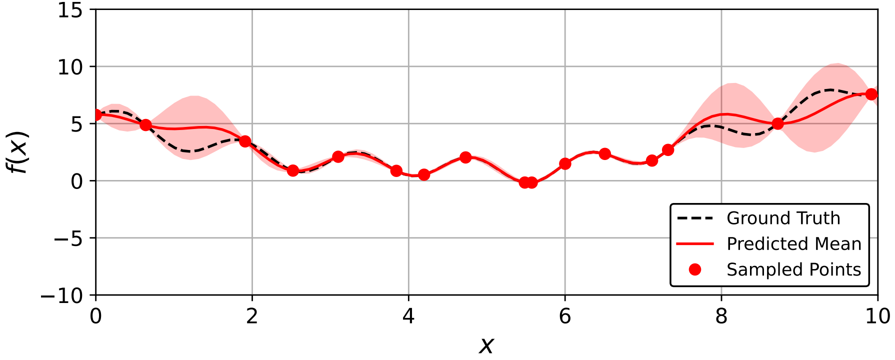
An Illustrated Example



An Illustrated Example



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.

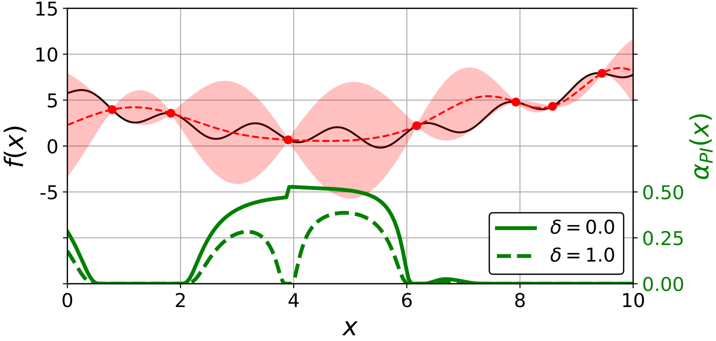
[Shahriari *et al.*, 2015]

Acquisition Function: Probability of Improvement

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

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

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

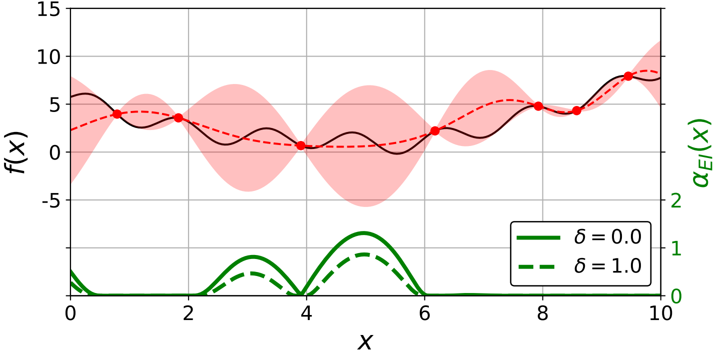


Acquisition Function: Expected Improvement

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

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

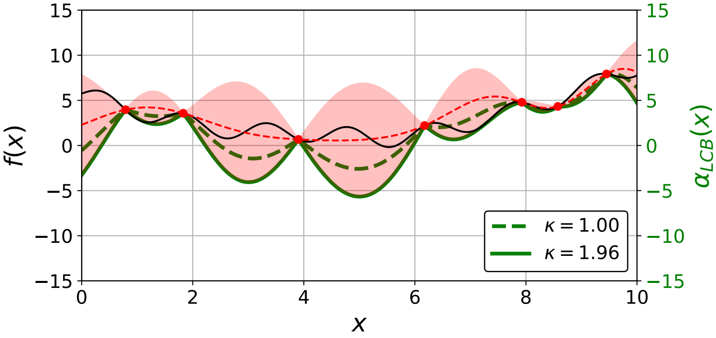
- Less prone than PI to become stuck in a local minimum.



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.

[Duvenaud, 2014; Shahriari *et al.*, 2015]

Pre-processing Data

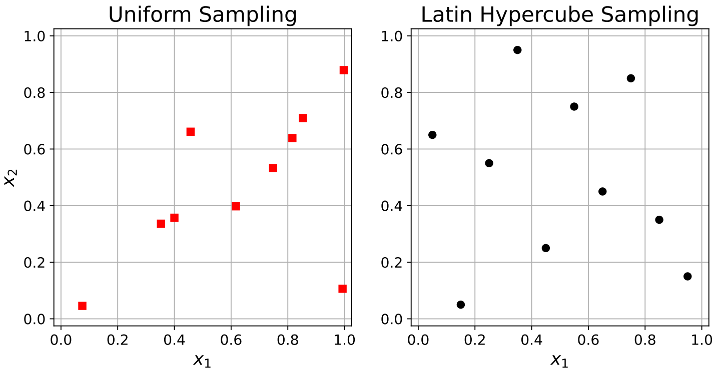
- Transforming sample data can facilitate GPR performance.
  - Normalize observations:

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

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

Generating a Useful Initial Sample

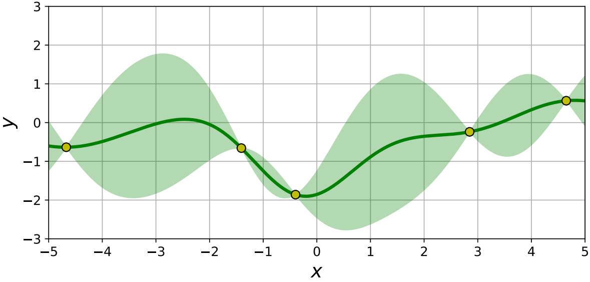
- Samples drawn uniformly may not fill the domain.
- Latin hypercube sampling offers better samples for regression.



[Santner *et al.*, 2018]

Final Thoughts

- GPs offer a rich environment for mathematical modeling.
- Regression is possible without explicit functional forms.
- Bayesian framework provides a natural measure of uncertainty.
- Expensive optimization problems benefit from efficient sampling.



References I

- Begg, S.H.; Bratvold, R.B.; Welsh, M.B. "Uncertainty vs. Variability: What's the Difference and Why is it Important?" in *Society of Petroleum Engineers Hydrocarbon Economics and Evaluation Symposium*, Houston, TX, **2014**.
- Der Kiureghian, A.; Ditlevsen, O. "Aleatory or Epistemic? Does it Matter?" *Struct. Saf.*, 31, 105–112, **2009**.
- Duvenaud, D.K. *Automatic Model Construction with Gaussian Processes* [Ph.D. dissertation], Pembroke College: Cambridge, England, **2014**.
- Frazier, P.I. "A Tutorial on Bayesian Optimization," *arXiv Preprint*, 22 pp., **2018**. arXiv:1807.02811
- Goldberg, P.W.; Williams, C.K.I.; Bishop, C.M. "Regression with Input-Dependent Noise: A Gaussian Process Treatment," *Adv. Neural Inf. Process. Syst.*, 10, 493–499, **1997**.

References II

- Kersting, K.; Plagemann, C.; Pfaff, P.; Burgard, W. “Most Likely Heteroscedastic Gaussian Process Regression,” *Proc. 24th ICML*, 393–400, **2007**.
- Pavliotis, G.A. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, Springer: New York, NY, **2014**.
- Qian, P.Z.G.; Wu, H.; Wu, C.F.J. “Gaussian Process Models for Computer Experiments with Qualitative and Quantitative Factors,” *Technometrics*, 50(3), 383–396, **2008**.
- 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**.

References III

- 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**.
- Zhang, Q.H.; Ni, Y.Q. "Improved Most Likely Heteroscedastic Gaussian Process Regression via Bayesian Residual Moment Estimator," *IEEE Trans. Signal Process.*, 68, 3450–3460, **2020**.