Basic Theory of Gaussian Processes

EN.540.782: Statistical Uncertainty Quantification

N. Wichrowski

Johns Hopkins University

March 16, 2021

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

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

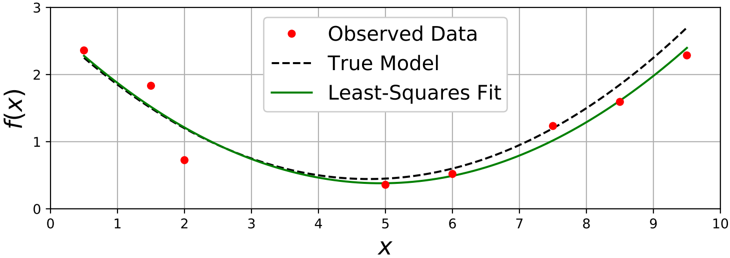
- Let $X \in \mathbb{R}^{n \times d}$ be the design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ the response vector.
- Then $\hat{\boldsymbol{w}} = X^\dagger \boldsymbol{y} = (X^\top X)^{-1} X^\top \boldsymbol{y}$ minimizes the squared error:

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

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

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

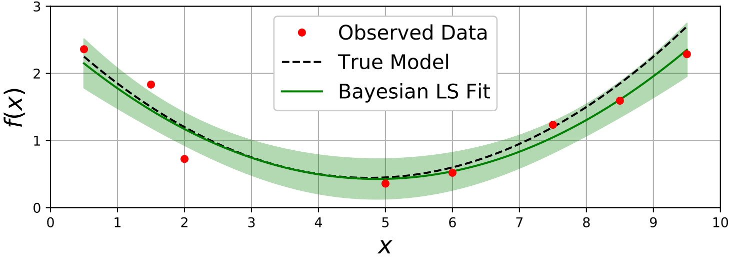
- In a Bayesian framework, we have $\boldsymbol{y} \,|\, X, \boldsymbol{w} \sim \mathcal{N}(X\boldsymbol{w}, \sigma_n^2 I)$.
- With prior $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_p)$, the resulting posterior is

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

Bayesian Linear Regression: Prediction

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

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



Gaussian Processes

Definition

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

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

Gaussian Processes

Definition

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

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

Example

Brownian motion on $\mathbb{R}$:

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

Ornstein-Uhlenbeck Process:

$$m(t) = 0 \qquad\qquad k(t, t') = e^{-|t-t'|}$$

Function-Space View of GPs

- The mean and covariance functions of a GP define a distribution over real-valued functions defined on the index set $\mathcal{X}$.
- How do we sample $f \sim \mathcal{GP}[m(\cdot), k(\cdot, \cdot)]$?
  - 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 $\boldsymbol{x}_\star \notin \mathcal{D}_X$:

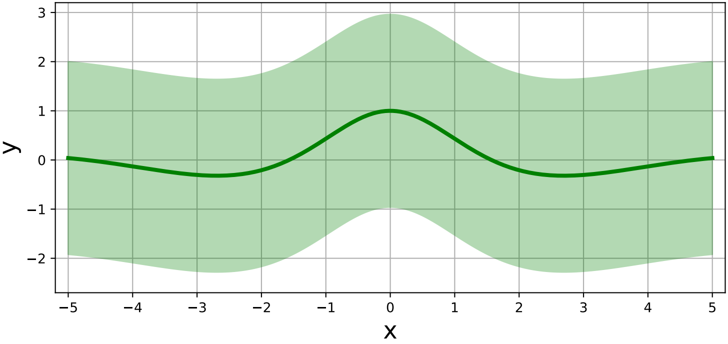
$$\mathbb{E}[\boldsymbol{y}_\star | X_\star, X, \boldsymbol{y}] = K(X_\star, X) K(X, X)^{-1} \boldsymbol{y}$$

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

Sampling from a GP

Consider the following prior distribution on univariate functions:

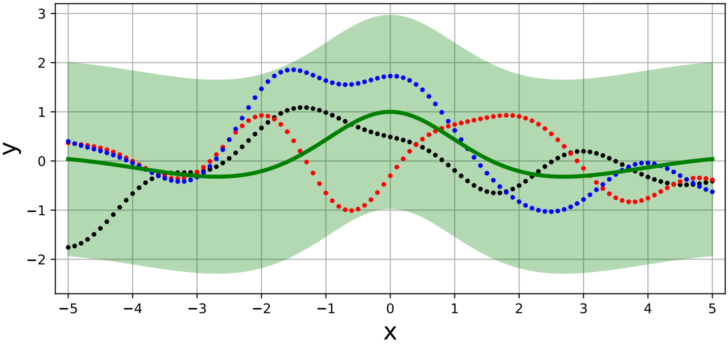
$$m(x) = \frac{\cos(x)}{1 + 0.25x^2}$$
$$k(x, x') = \exp\left[-\frac{1}{2}\left(x - x'\right)^2\right]$$



Sampling from a GP

Consider the following prior distribution on univariate functions:

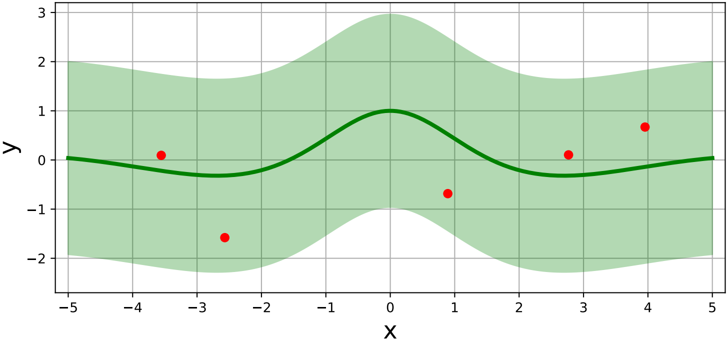
$$m(x) = \frac{\cos(x)}{1 + 0.25x^2}$$
$$k(x, x') = \exp\left[-\frac{1}{2}\left(x - x'\right)^2\right]$$



Sampling from a GP

Consider the following prior distribution on univariate functions:

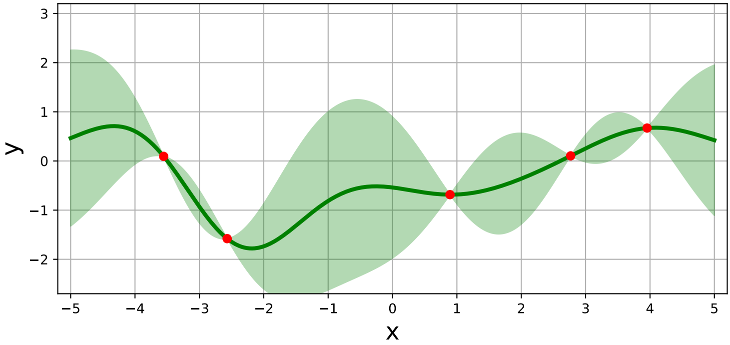
$$m(x) = \frac{\cos(x)}{1 + 0.25x^2}$$
$$k(x, x') = \exp\left[-\frac{1}{2}\left(x - x'\right)^2\right]$$



Sampling from a GP

Consider the following prior distribution on univariate functions:

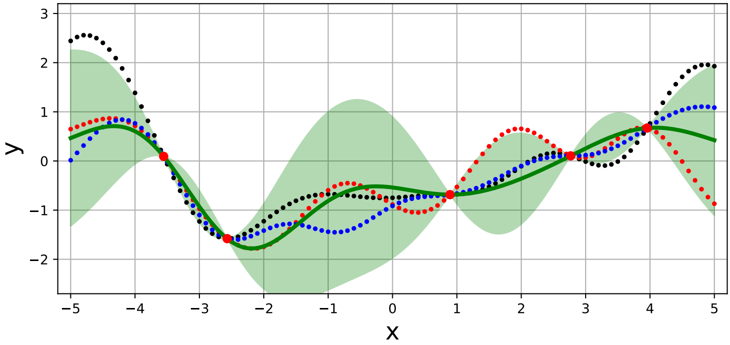
$$m(x) = \frac{\cos(x)}{1 + 0.25x^2}$$
$$k(x, x') = \exp\left[-\frac{1}{2}\left(x - x'\right)^2\right]$$



Sampling from a GP

Consider the following prior distribution on univariate functions:

$$m(x) = \frac{\cos(x)}{1 + 0.25x^2}$$
$$k(x, x') = \exp\left[-\frac{1}{2}\left(x - x'\right)^2\right]$$



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:

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

Continuity and Differentiability

Definition

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

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

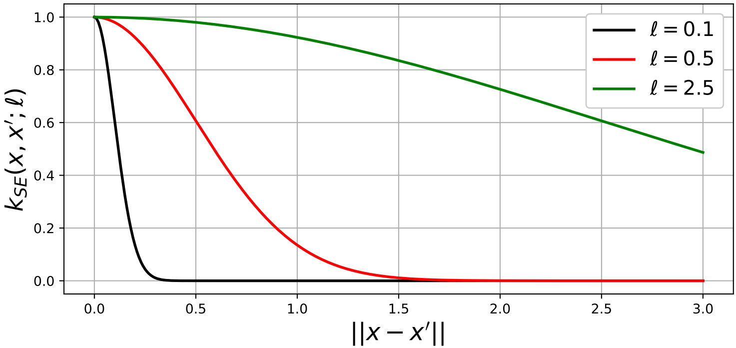
We say that f is mean-square differentiable (MSD) at $\boldsymbol{x}_\star$ with partial derivatives $\partial f(\boldsymbol{x}_\star)/\partial x_i$ if, for $i \in \{1, \ldots, d\}$,

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

- 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 $\boldsymbol{x}_\star$.

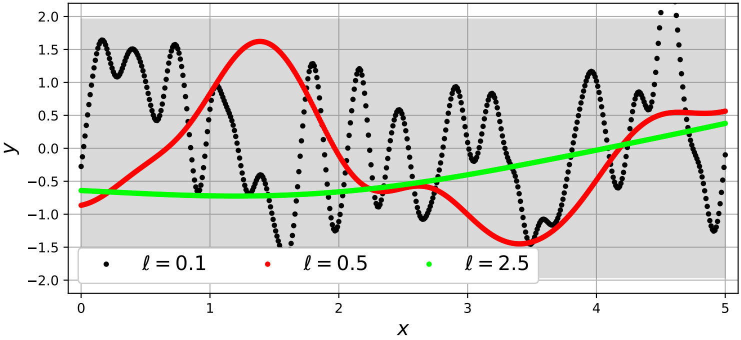
Covariance Functions: Squared Exponential

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



Covariance Functions: Squared Exponential

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

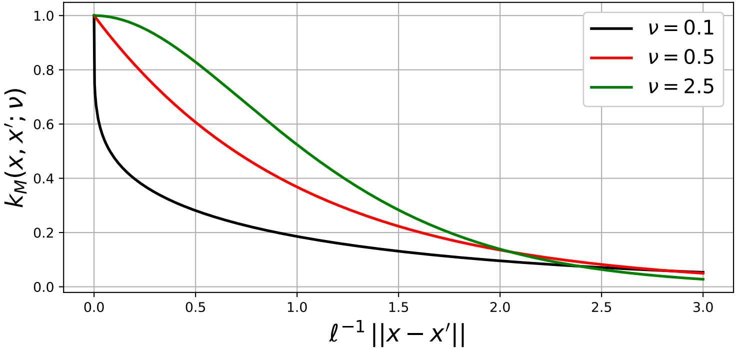


Covariance Functions: Matérn

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

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

Can be expressed more simply if $\nu + \frac{1}{2} \in \mathbb{N}$ (see Appendix).

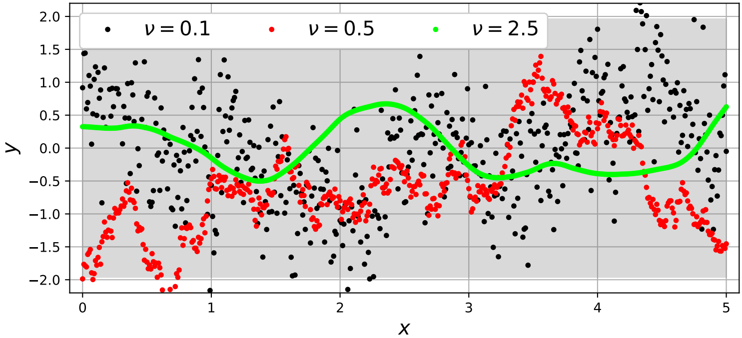


Covariance Functions: Matérn

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

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

Can be expressed more simply if $\nu + \frac{1}{2} \in \mathbb{N}$ (see Appendix).



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 $\boldsymbol{x}_\star \notin \mathcal{D}_X$:

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

- If noisy measurements, then replace $K(X, X)$ by $K(X, X) + \sigma_n^2 I$.
- Given $\mathcal{D}$, the GPR model is specified by our choice of kernel.

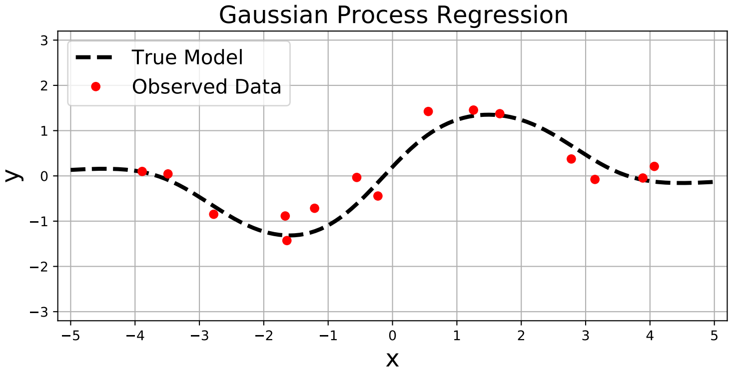
Model Selection

- GP Regression requires:
  - Selecting a kernel function (model).
  - Tuning the hyperparameters.
- "Minimize the generalization error."
- Log marginal likelihood:

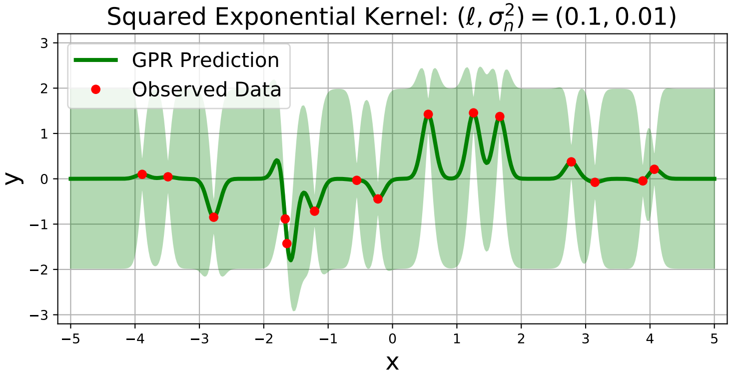
$$\log p(\boldsymbol{y} \mid X, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^\top K_y \boldsymbol{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log(2\pi)$$
$$K_y = K(X, X; \boldsymbol{\theta}) + \sigma_n^2 I$$

- Cross-validation

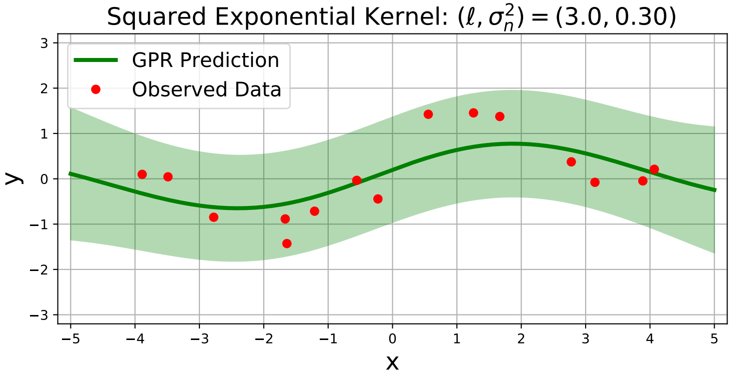
Hyperparameter Tuning: An Illustration



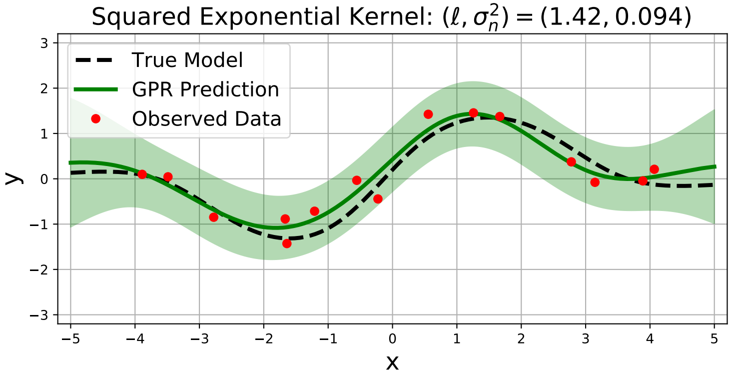
Hyperparameter Tuning: An Illustration



Hyperparameter Tuning: An Illustration



Hyperparameter Tuning: An Illustration



Final Thoughts

- GPs offer benefits for both regression and fitting.
- Precise functional forms are not needed.
- Bayesian framework provides a natural measure of uncertainty.

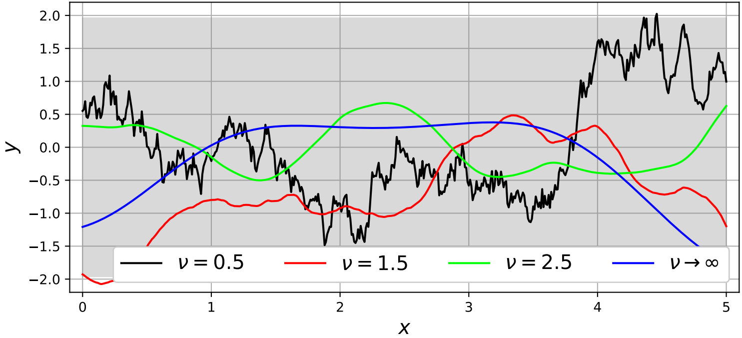
References

- Duvenaud, D.K. *Automatic Model Construction with Gaussian Processes* [Ph.D. dissertation], Pembroke College: Cambridge, England (2014).
- 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).
- Jun, M. *Spatial Statistics* [Lecture Slides], Texas A&M University: College Station, TX (2009).
- Lifshits, M. *Lectures on Gaussian Processes*, Springer: New York, NY (2012).
- Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*, MIT Press: Cambridge, MA (2006).

Matérn Kernel with $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$

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

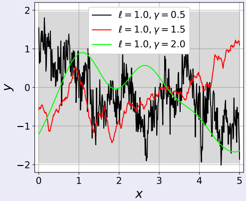
$$k_{5/2}(r) = \left(1 + \frac{r\sqrt{5}}{\ell} + \frac{5r^2}{3\ell^2}\right)\exp\left[-\frac{r\sqrt{5}}{\ell}\right]$$



Additional (Isotropic) Families of Covariance Kernels

Gamma-Exponential

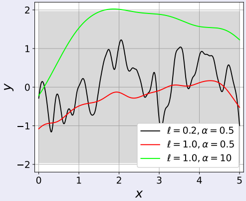
$$k(r) = \exp\left[-\left(\frac{r}{\ell}\right)^{\gamma}\right] \qquad 0 < \gamma \le 2, \ell > 0$$



- Reduces to Squared Exponential when $\gamma = 2$.
- Not mean-square differentiable for any $\gamma < 2$.

Rational Quadratic

$$k(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha} \qquad \alpha, \ell > 0$$



- Reduces to Squared Exponential as $\alpha \to \infty$.
- Mean-square differentiable for all $\alpha > 0$.