Scientific Machine Learning Workshop

Lecture 4: Gaussian Process Regression

Ulisses Braga Neto

Department of Electrical and Computer Engineering
Scientific Machine Learning Lab (SciML Lab)
Texas A&M Institute of Data Science (TAMIDS)
Texas A&M University

Cenpes
August 2025

Gaussian Processes

- Gaussian process regression is a *nonparametric* regression approach that performs Bayesian inference directly on a space of functions using a Gaussian stochastic process prior.
- A *stochastic process* can be seen as an ensemble of random real-valued random functions $\{f(\boldsymbol{x}, \xi); \boldsymbol{x} \in R^d, \xi \in S\}$, where S is a sample space in an appropriate probability space.
- For each $\xi \in S$, $f(\boldsymbol{x}, \xi)$ is an ordinary function of $\boldsymbol{x}$ (a *realization* or *sample function*); for each $\boldsymbol{x} \in R^d$, $f(\boldsymbol{x}, \xi)$ is a random variable.
- Under certain conditions, a stochastic process is uniquely characterized by the distributions of the random vectors $\mathbf{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_k)]$ for all finite sets $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in R^d$, $k \geq 1$.
- If all such random vectors have multivariate Gaussian distributions, then the stochastic process is a *Gaussian process*.

Stationary Gaussian Processes

- Due to the nature of the multivariate Gaussian distribution, a Gaussian stochastic process depends only on the *mean function*

$$m(\boldsymbol{x}) = E[f(\boldsymbol{x})], \quad \boldsymbol{x} \in R^d,$$

and the *covariance function of Gaussian process* or *kernel*

$$k(\boldsymbol{x}, \boldsymbol{x}') = E[f(\boldsymbol{x})f(\boldsymbol{x}')] - m(\boldsymbol{x})m(\boldsymbol{x}'), \quad \boldsymbol{x}, \boldsymbol{x}' \in R^d.$$

We may thus denote a GP by $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$.

- A stochastic process is called *stationary* if the distribution of $\mathbf{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_k)]$ is the same as that of $\mathbf{f_u} = [f(\boldsymbol{x}_1 + \mathbf{u}), \ldots, f(\boldsymbol{x}_k + \mathbf{u})]$, for all $\mathbf{u} \in R^d$ and finite sets of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in R^d$, $k \geq 1$.

- It can be shown that the covariance function of a stationary process can only be a function of $\boldsymbol{x}' - \boldsymbol{x}'$.

Stationary Covariance Functions

- By analogy, a a covariance function is called *stationary* if it is a function only of $\boldsymbol{x} - \boldsymbol{x}'$. (Notice that having a stationary covariance function does not make a stochastic process stationary — it is only a necessary condition).
- The *variance function* $v(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x})$ is the variance of each random variable $f(\boldsymbol{x})$, for $\boldsymbol{x} \in R^d$. If the covariance function is stationary, then the variance function is constant: $v(\boldsymbol{x}) = \sigma_k^2 = k(\boldsymbol{x}, \boldsymbol{x})$, for any $\boldsymbol{x} \in R^d$.
- Finally, a stationary covariance function is *isotropic* if it is a function only of $||\boldsymbol{x} - \boldsymbol{x}'||$. By using the general fact that $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$ we can see that, in the univariate case, a stationary covariance function is automatically isotropic.

Gaussian and Absolute Exponential Kernels

- Two important examples of isotropic stationary covariance functions are the squared exponential ("Gaussian") kernel,

$$k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2\ell^2}\right),$$

  and the absolute exponential kernel,

$$k_{\mathrm{AE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||}{\ell}\right),$$

  where, in both cases, ℓ is the process *length-scale*.

- In the univariate case, the absolute exponential is the double-exponential covariance function $k(\tau) = \sigma_k^2 \exp(-|\tau|/\ell)$, with $\tau = x - x'$, hence the name "absolute exponential."

Matérn Kernel

- The Gaussian and absolute exponential kernels are extremes in a family of isotropic stationary *Matérn* covariance functions:

$$k^{\nu}_{\mathrm{MAT}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell} \right),$$

  where $\nu > 0$ is the order of the kernel, and K_ν is the incomplete Bessel function of the second kind.

- It is possible to show that $\nu = 1/2$ leads to the absolute exponential kernel, while $\nu \gg 1$ approximates closely the Gaussian kernel (in fact, it converges to the Gaussian kernel as $\nu \to \infty$).

Matérn Kernel

- The cases of most interest in Gaussian process regression are the case $\nu = 3/2$,

$$k_{\mathrm{MAT}}^{\nu=3/2}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \left(1 + \frac{\sqrt{3}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell}\right)$$

  and the case $\nu = 5/2$,

$$k_{\mathrm{MAT}}^{\nu=5/2}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \left(1 + \frac{\sqrt{5}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell} + \frac{5\,||\boldsymbol{x} - \boldsymbol{x}'||^2}{3\ell^2}\right) \times \exp\left(-\frac{\sqrt{5}\,||\boldsymbol{x} - \boldsymbol{x}'||}{\ell}\right).$$

- These covariance functions are plotted on the next slide for the univariate case.

Kernel Comparison

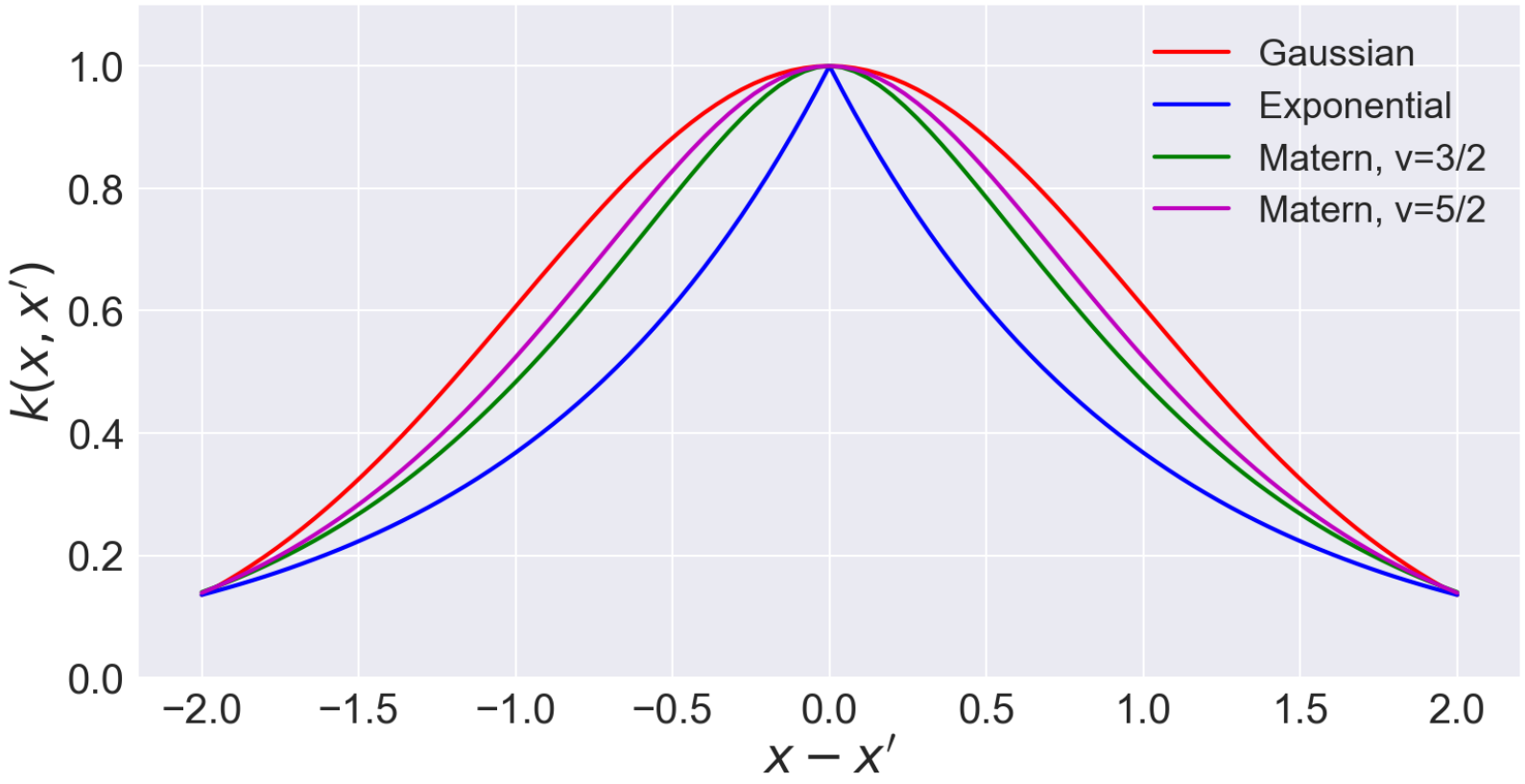

Figure: Univariate kernels used in Gaussian Process regression (plots generated by `c11_GPkern.py`).

Gaussian Process Example

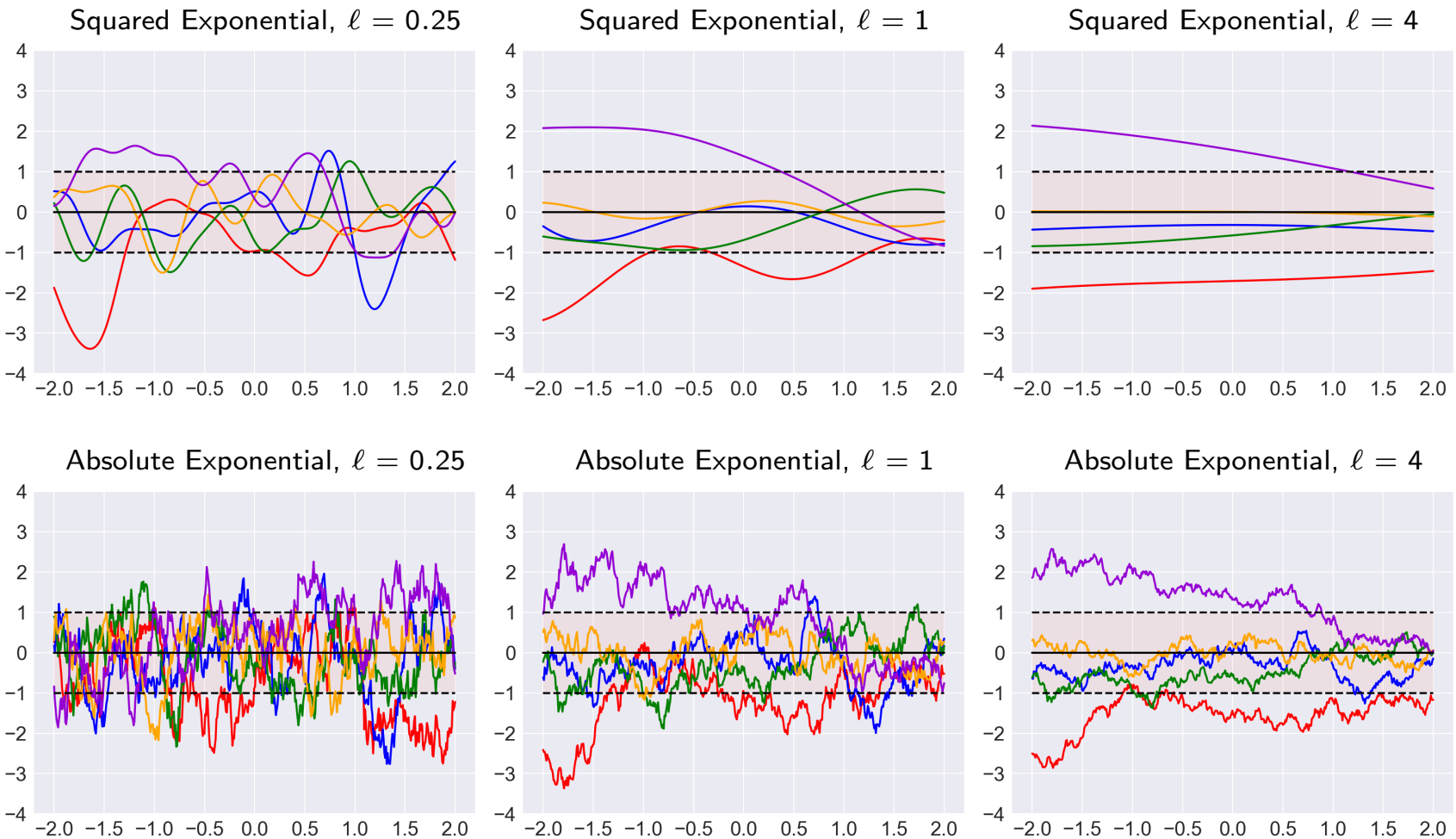

Figure: Sample functions from unit-variance, zero-mean Gaussian processes. The constant mean and variance functions are also displayed (plots generated by `c11_GPsamp.py`).

Gaussian Process Example

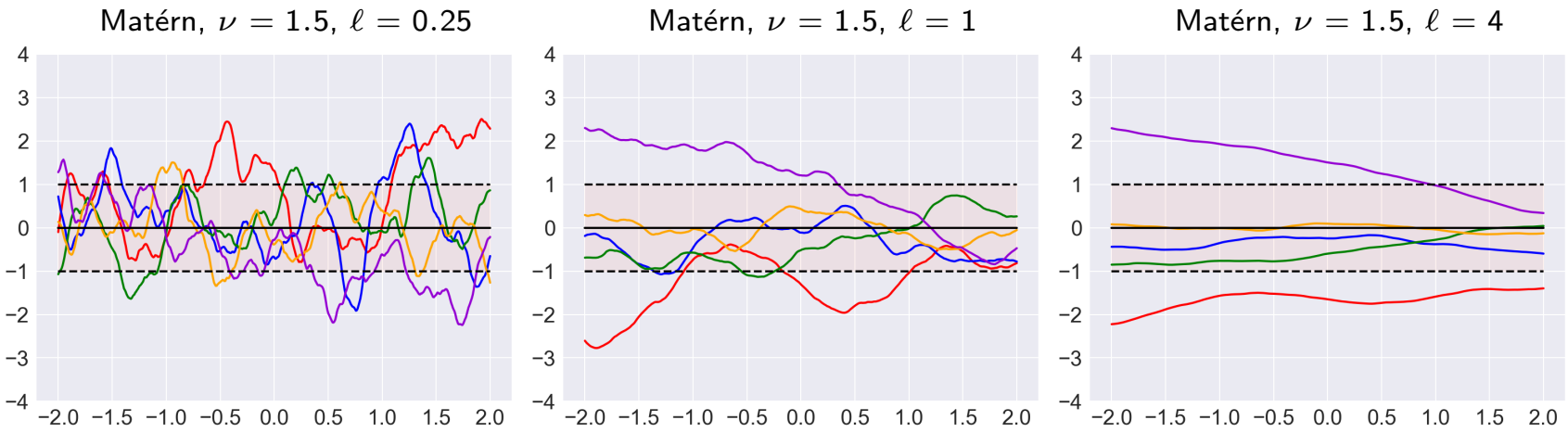

Figure: Sample functions from unit-variance, zero-mean Gaussian processes (continued).

Gaussian Process Example

- In the plot of covariance functions, we see that the Gaussian kernel produces the largest correlations at all distances $x - x'$. This implies that the sample functions are more likely to be smooth for the Gaussian kernel.
- The Gaussian kernel is infinitely differentiable at the origin, and hence the associated process is infinitely differentiable everywhere in the mean-square sense, leading to smooth sample functions.
- The absolute exponential kernel is not differentiable at all at the origin, so that the associated process is not mean-square differentiable anywhere, leading to very rough sample functions.
- A stochastic process with the Matérn covariance function of order ν is mean-square differentiable $\lceil \nu \rceil - 1$ times, producing intermediate behavior; e.g. processes with $\nu = 3/2$ and $\nu = 5/2$ are once and twice mean-square differentiable, respectively.

Simulation of a Gaussian Process

- In practice, Gaussian processes are simulated by:
  - Picking a uniformly-spaced finite set of *testing points* $X^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*)$,
  - Generating a random vector

$$\mathbf{f}^* \sim \mathcal{N}(m(X^*), K(X^*, X^*)), \tag{1}$$

where

$$m(X^*) = (m(\boldsymbol{x}_1^*), \ldots, m(\boldsymbol{x}_m^*))^T, \tag{2}$$

and

$$K(X^*, X^*) = \begin{bmatrix} k(\boldsymbol{x}_1^*, \boldsymbol{x}_1^*) & k(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*) & \cdots & k(\boldsymbol{x}_1^*, \boldsymbol{x}_m^*) \\ k(\boldsymbol{x}_2^*, \boldsymbol{x}_1^*) & k(\boldsymbol{x}_2^*, \boldsymbol{x}_2^*) & \cdots & k(\boldsymbol{x}_2^*, \boldsymbol{x}_m^*) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_m^*, \boldsymbol{x}_1^*) & k(\boldsymbol{x}_m^*, \boldsymbol{x}_2^*) & \cdots & k(\boldsymbol{x}_m^*, \boldsymbol{x}_m^*) \end{bmatrix}. \tag{3}$$

  - Applying linear interpolation (connecting the dots with lines).

- Hence, in practice we deal only with finite numbers of multivariate Gaussian random vectors.

Gaussian Process Regression

- The goal of Gaussian process regression is to use training observations to predict the value of the unknown function f at a given arbitrary set of test points.
- Consider a set of training points $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and corresponding set of noisy observations $\mathbf{Y} = \mathbf{f} + \boldsymbol{\varepsilon}$, where $\mathbf{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ — we consider only the homoskedastic zero-mean Gaussian noise case, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$ and $\mathcal{N}$ is independent of X.
- Given testing points $X^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*)$, we would like to *predict* the value of $\mathbf{f}(X^*) = (f(\boldsymbol{x}_1^*), \ldots, f(\boldsymbol{x}_m^*))$ in a way that is *consistent* with the training data.
- The Bayesian paradigm is to determine the *posterior distribution*, i.e., the conditional distribution of the vector $\mathbf{f}^*$ given $\mathbf{Y}$, and then make inferences on that — where the *prior distribution* of $\mathbf{f}^*$ is described by (1)–(3) on the previous slide.

Gaussian Process Conditional Distribution

- For Gaussian processes, it is possible to obtain this conditional distribution in closed form.
- If $\boldsymbol{X}$ and $\boldsymbol{X}'$ are jointly distributed Gaussian vectors, with multivariate distribution

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{X}'} \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right),$$

  then it can be shown that $\boldsymbol{X} \mid \boldsymbol{X}'$ is again a multivariate Gaussian distribution,

$$\boldsymbol{X} \mid \boldsymbol{X}' \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{X}} + CB^{-1}(\boldsymbol{X}' - \boldsymbol{\mu}_{\boldsymbol{X}'}), A - CB^{-1}C^T \right). \quad (4)$$

Gaussian Process Conditional Distribution

- Now, the joint distribution of the vectors $\mathbf{Y}$ and $\mathbf{f}^*$ is multivariate Gaussian:

$$\begin{bmatrix} \mathbf{f}^* \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X^*, X^*) & K(X^*, X) \\ K(X^*, X)^T & K(X, X) + \sigma^2 I \end{bmatrix}\right) \tag{5}$$

where $K(X^*, X^*)$ is given by (3), while

$$K(X^*, X) = \begin{bmatrix} k(\boldsymbol{x}_1^*, \boldsymbol{x}_1) & k(\boldsymbol{x}_1^*, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1^*, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2^*, \boldsymbol{x}_1) & k(\boldsymbol{x}_2^*, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_2^*, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_m^*, \boldsymbol{x}_1) & k(\boldsymbol{x}_m^*, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_m^*, \boldsymbol{x}_n) \end{bmatrix},$$

and

$$K(X, X) = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & k(\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}.$$

Gaussian Process Conditional Distribution

- Using the conditional Gaussian result (4), we gather that the posterior distribution is

$$\mathbf{f}^* \mid \mathbf{Y} \sim \mathcal{N}(\bar{\mathbf{f}}^*, \mathrm{Var}(\mathbf{f}^*)),$$

with posterior mean vector and posterior covariance matrix

$$\begin{aligned} \bar{\mathbf{f}}^* &= K(X^*, X)[K(X, X)^{-1} + \sigma^2 I_n]^{-1}\mathbf{Y} \\ \mathrm{Var}(\mathbf{f}^*) &= K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma^2 I]^{-1}K(X, X^*). \end{aligned} \tag{6}$$

- Clearly, even if the prior Gaussian process is zero mean and has a stationary covariance function, this is no longer the case, in general, for the posterior Gaussian process.

Gaussian Process Regression

- In Gaussian process regression, we estimate the value of $\mathbf{f}^*$ at the test points by the conditional mean $\bar{\mathbf{f}}^*$, and the conditional regression error at each test point by the corresponding element in the diagonal of $\mathrm{Var}(\mathbf{f}^*)$.
- Notice that, in practice, σ^2 is rarely known, so the value used in (6) becomes a parameter σ_p^2 to be selected.
- If desired, an estimate of the entire function f can be obtained by interpolating the conditional mean values (and the conditional variance values) over a dense set of test points.
- In the univariate case, this could be done by simply joining the estimated values with lines. In a multivariate setting, more advanced interpolation methods are required.

Gaussian Process Regression Example

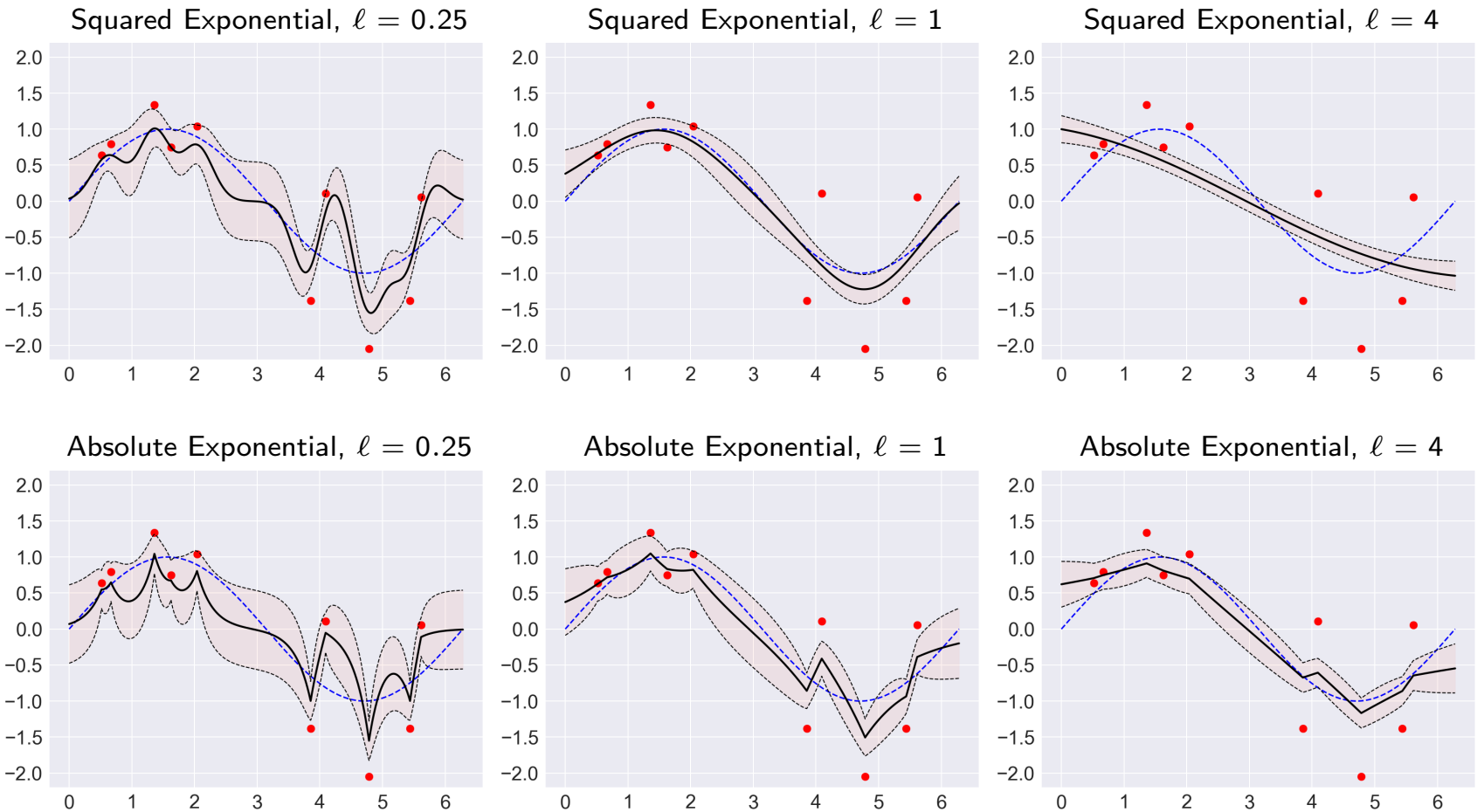

Figure: Gaussian process regression example. Red circles represent the training data, while the black solid and dashed blue curves represent the estimated and optimal regression, respectively. A one standard-deviation confidence band is displayed as well. (plots generated by `c11_GPfit.py`).

Gaussian Process Regression Example

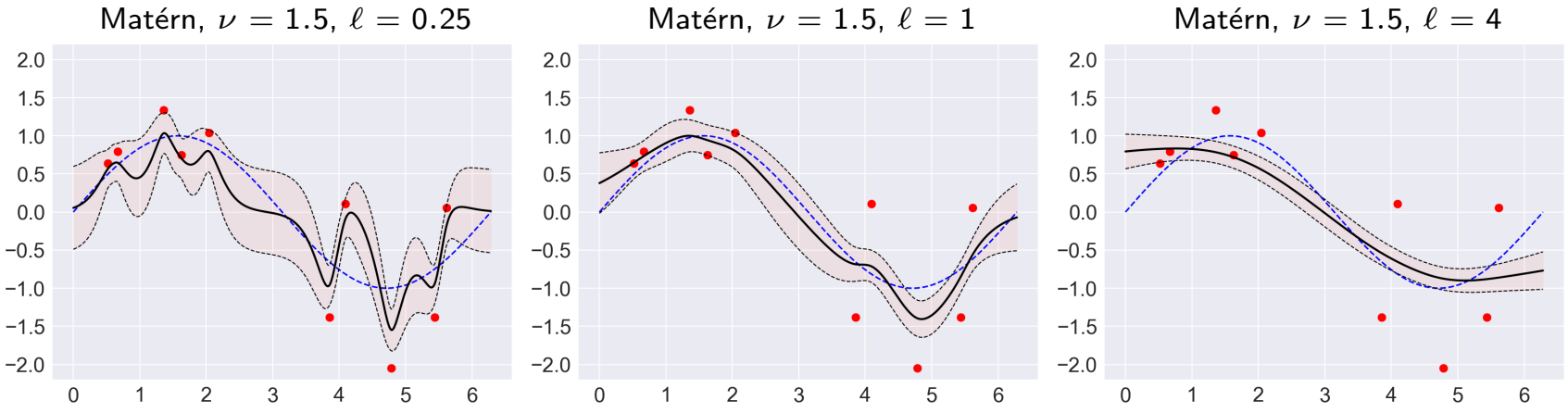

Figure: Gaussian process regression example. (Continued).

Gaussian Process Regression Hyperparameter Selection

- In the previous example, σ_k^2, ℓ, and σ_p^2 were set in ad-hoc fashion.
- A more principled way to set these hyperparameters is based on maximizing the *marginal likelihood* $p(\mathbf{Y} \mid X, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\sigma_p^2, \sigma_k^2, \ell)$ is the vector of hyperparameters.
- From (5), we obtain $p(\mathbf{Y} \mid X, \boldsymbol{\theta}) = \mathcal{N}(0, K(X,X) + \sigma^2 I)$. The log of the marginal likelihood is thus given by:

$$\begin{aligned} \ln p(\mathbf{Y} \mid X, \boldsymbol{\theta}) = &-\frac{1}{2}\mathbf{Y}^T (K(X,X) + \sigma^2 I)^{-1} \mathbf{Y} \\ &-\frac{1}{2}\ln |K(X,X) + \sigma^2 I| - \frac{n}{2}\ln 2\pi \,. \end{aligned} \tag{7}$$

- The first term on the right hand side represents the empirical fit to the data, the second one is a complexity penalty term, while the last term is a normalization constant.
- Numerical maximization of (7) by gradient descent produces the required value of the hyperparameters.

Gaussian Process Regression Example

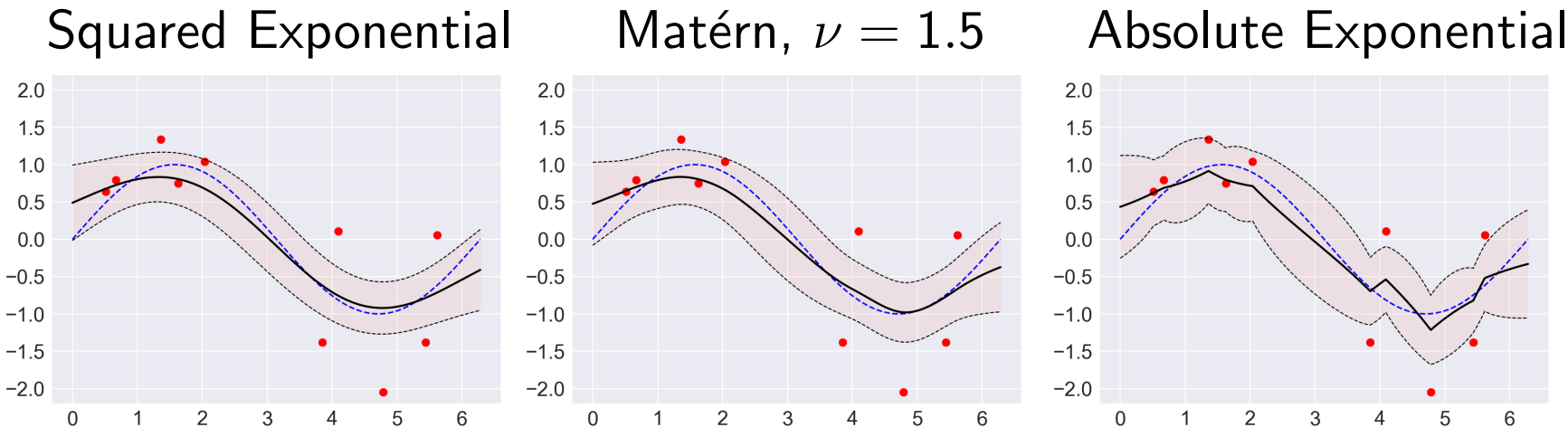

Figure: Gaussian process regression results using hyperparameter values selected by maximization of the marginal likelihood:
$\sigma_k^2 = 0.63, \ell = 1.39, \sigma_p^2 = 0.53$ for the squared exponential kernel,
$\sigma_k^2 = 0.65, \ell = 1.51, \sigma_p^2 = 0.53$ for the Matérm kernel, and
$\sigma_k^2 = 0.74, \ell = 1.45, \sigma_p^2 = 0.45$ for the absolute exponential kernel (plots generated by `c11_GPfitML.py`).