ESE 680-004: Learning and Control **Fall 2019**

Lecture 24: Gaussian Processes

Lecturer: Nikolai Matni *Scribes: Alëna Rodionova*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1 Preliminaries

1.1 Gaussian Random Vectors

Definition 1. *Random vector $x \in \mathbb{R}^n$ is **Gaussian** if it has density*

$$p_x(v) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(v-\mu)^\top \Sigma^{-1}(v-\mu)\right) \tag{1}$$

for some $\Sigma = \Sigma^\top \succ 0$, and $\mu \in \mathbb{R}^n$.

We will write a Gaussian random vector as:

$$x \sim \mathcal{N}(\mu, \Sigma).$$

Vector $\mu \in \mathbb{R}^n$ is the *mean* or *expected value* of x and is defined as

$$\mu = \mathbb{E}x = \int v p_x(v)\, \mathrm{d}v \tag{2}$$

and the matrix Σ is the *covariance matrix* of x, given dy

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^\top\right] = \mathbb{E}[xx^\top] - \mathbb{E}x\mu^\top - \mu\mathbb{E}x^\top + \mu\mu^\top = \mathbb{E}[xx^\top] - \mu\mu^\top = \int v p_x(v)\, \mathrm{d}v \tag{3}$$

Variables μ and Σ determine the shape of density. Graphical representation of the probability density function for $x \sim \mathcal{N}(0,1)$ is given in a Figure 1.

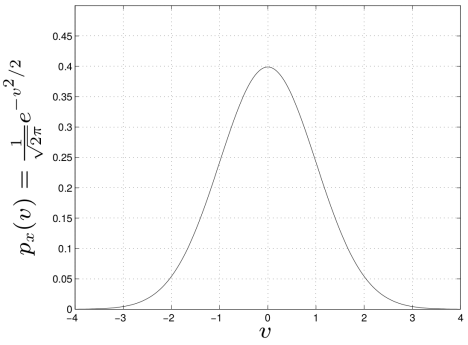

Figure 1: Probability density function for $x \sim \mathcal{N}(0,1)$.

Example 1. *$x \sim \mathcal{N}(0, I)$ means x_i are independent identically distributed (IID) random variables $\mathcal{N}(0,1)$.*

Definition 2. *Mean (norm) square deviation of x from μ is*[1]

$$\mathbb{E}||x-\mu||^2 = \mathbb{E}\,\mathrm{Tr}(x-\mu)(x-\mu)^\top = \mathrm{Tr}\,\Sigma = \sum_{i=1}^n \Sigma_{ii} \tag{4}$$

[1] Using $\mathrm{Tr}\,AB = \mathrm{Tr}\,BA$.

1.2 Confidence Ellipsoids

Using Equation (1) it is easy to show that $p_x(v)$ is constant for $(v-\mu)^\top \Sigma^{-1}(v-\mu) = \alpha$, i.e., on the surface of the ellipsoid

$$\mathcal{E}_\alpha := \{v \mid (v-\mu)^\top \Sigma^{-1}(v-\mu) \leq \alpha\}. \tag{5}$$

Definition 3. *η-confidence set for random variable z is the smallest volume set S such that*

$$\mathbb{P}[z \in S] \geq \eta. \tag{6}$$

Definition 4 (Confidence Ellipsoids)**.** *For Gaussian random variables, the $\mathcal{E}_\alpha$ are the η-confidence sets (using Equation (5)), and are called **confidence ellipsoids**, where α determines confidence level η.*

Mean μ gives the center of the ellipsoid $\mathcal{E}_\alpha$, and semiaxes are defined as $\sqrt{\alpha\lambda_i}u_i$, where u_i are orthonormal eigenvectors of Σ with eigenvalues λ_i.

1.2.1 Confidence Levels

Notice that since x is a Gaussian random variable and $\Sigma \succ 0$, the non-negative random variable $(x-\mu)^\top \Sigma^{-1}(x-\mu)$ has a χ^2_n distribution[2], with a CDF $F_{\chi^2_n}(\alpha)$. Hence, using the above confidence ellipsoid's definition 4, $\mathbb{P}[x \in \mathcal{E}_\alpha] = F_{\chi^2_n}(\alpha)$.

Some good approximations are:

- $\mathcal{E}_n$ gives about 50% of probability mass.
- $\mathcal{E}_{n+2\sqrt{n}}$ gives about 90% of probability mass.

Example 2. *Let $x \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \begin{bmatrix}2\\1\end{bmatrix}$ and $\Sigma = \begin{bmatrix}2 & 1\\1 & 1\end{bmatrix}$. Then 90% confidence ellipsoid corresponds to $\alpha = 4.6$, see Figure 2. In this experiment, 91 out of 100 samples fall in $\mathcal{E}_{4.6}$.*

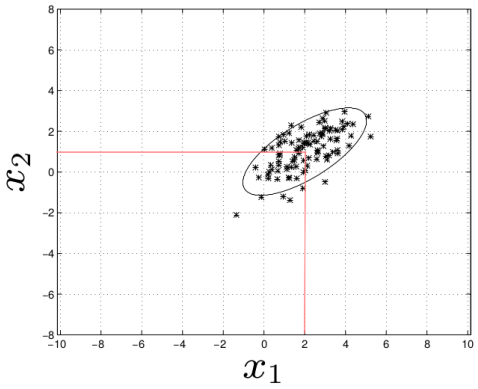

Figure 2: Confidence ellipsoid $\mathcal{E}_\alpha$ for $\alpha = 4.6$.

1.3 Affine Transformations

Suppose that $x \sim \mathcal{N}(\mu, \Sigma_x)$. Consider an affine transformation of x: $z = Ax + b$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Then z is Gaussian with mean

$$\bar{z} = \mathbb{E}z = \mathbb{E}(Ax+b) = A\mathbb{E}x + b = A\mu + b, \tag{7}$$

[2]Chi-squared distribution χ^2_n is the distribution of a sum of the squares of n independent standard normal random variables, `https://en.wikipedia.org/wiki/Chi-squared_distribution`

and covariance

$$\begin{aligned}\Sigma_z &= \mathbb{E}\left[(z-\bar{z})(z-\bar{z})^\top\right] = \\ &= \mathbb{E}\left[(Ax+b-A\mu-b)(Ax+b-A\mu-b)^\top\right] = \\ &= A\mathbb{E}\left[(x-\mu)(x-\mu)\top\right]A^\top = \\ &= A\Sigma_x A^\top\end{aligned} \tag{8}$$

Example 3. *For $w \sim \mathcal{N}(0, I)$ and $x = \Sigma^{1/2}w + \mu$, we have $x \sim \mathcal{N}(\mu, \Sigma)$. Useful for simulating vectors with given mean and covariance*

Example 4. *Conversely, for $x \sim \mathcal{N}(\mu, \Sigma)$ and $z = \Sigma^{-1/2}(x-\mu)$, we have $x \sim \mathcal{N}(0, I)$. So it normalizes and decorrelates. Called whitening or normalizing.*

Example 5. *For $x \sim \mathcal{N}(\mu, \Sigma)$ and $c \in \mathbb{R}^n$ scalar $c^\top x \sim \mathcal{N}(c^\top\mu,\, c^\top\Sigma c)$. This means that we can identify unit length direction of minimum (maximum) variability for x by selecting the orthonormal eigenvector u corresponding to the minimum (maximum) eigenvalue $\lambda_{\min}$ ($\lambda_{\max}$)of Σ:*

$$\Sigma u = \lambda u, \quad ||u|| = 1. \tag{9}$$

Standard deviation of $u_n^\top x$ is $\sqrt{\lambda_{\min}}$.

1.4 Linear Measurements

Suppose that we obtain liear measurements with noise $y = Ax+v$, where $x \in \mathbb{R}^n$ is what we want to estimate, $y \in \mathbb{R}^m$ is a measurement, $A \in \mathbb{R}^{m\times n}$ characterizes sensors or measurements and v is a sensor noise. We also assume that $x \sim \mathcal{N}(\bar{x},\, \Sigma_x)$, so the *prior distribution* of x that describes initial uncertainty about x is given. Another assumption is $v \sim \mathcal{N}(\bar{v},\, \Sigma_v)$, where $\bar{v}$ is noise *bias* or *offset* and Σ_v is noise covariance. x and v are assumed to be independent. Then we have

$$\begin{bmatrix} x \\ v \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{x} \\ \bar{v} \end{bmatrix},\, \begin{bmatrix} \Sigma_x & \\ & \Sigma_v \end{bmatrix}\right). \tag{10}$$

We can write

$$\mathbb{E}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \bar{x} \\ A\bar{x}+\bar{v} \end{bmatrix} \tag{11}$$

because

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}\begin{bmatrix} x \\ v \end{bmatrix} \tag{12}$$

Therefore,

$$\mathbb{E}\left[\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}^\top\right] = \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}\begin{bmatrix} \Sigma_x & \\ & \Sigma_v \end{bmatrix}\begin{bmatrix} I & 0 \\ A & I \end{bmatrix}^\top = \begin{bmatrix} \Sigma_x & \Sigma_x A^\top \\ A\Sigma_x & A\Sigma_x A^\top + \Sigma_v \end{bmatrix} \tag{13}$$

We showed that the covariance of measurement y is $A\Sigma_x A^\top + \Sigma_v$, where first term $A\Sigma_x A^\top$ is called *signal covariance* and second term Σ_v, as mentioned before, *noise covariance*.

1.5 Minimum Mean Square Estimation

Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be random vectors (not necessarily Gaussians). We seek to estimate x given y, thus we seek a function $\phi : \mathbb{R}^m \to \mathbb{R}^n$ such that $\hat{x} = \phi(y)$ is 'near' x. A common criterion is to seek to minimize the mean square estimation error be finding a map ϕ as follows:

$$\phi_{mmse}(y) = \arg\min \mathbb{E}||\phi(y) - x||_2^2. \tag{14}$$

A general solution to this problem is given by the conditional expectation of x given y:

$$\phi_{mmse}(y) = \mathbb{E}[x|y]. \tag{15}$$

Such ϕ_{mmse} is called *minimum mean-square estimator* (MMSE estimator).

If (x, y) are jointly Gaussian, i.e., if

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{bmatrix} \right), \tag{16}$$

then the conditional density is

$$p_{x|y}(v|y) = \frac{1}{\sqrt{(2\pi)^n \det(\Lambda)}} \exp\left(-\frac{1}{2}(v-w)^\top \Lambda^{-1}(v-w) \right), \tag{17}$$

where $\Lambda = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\top$ and $W = \bar{x} + \Sigma_{xy}\Sigma_y^{-1}(y - \bar{y})$. Hence MMSE estimator (i.e., conditional expectation) is

$$\hat{x} = \phi_{mmse}(y) = \mathbb{E}(x|y) = \bar{x} + \Sigma_{xy}\Sigma_y^{-1}(y - \bar{y}) \tag{18}$$

so ϕ_{mmse} is an affine function. MMSE estimation error, $\hat{x} - x$, is a Gaussian random vector

$$\hat{x} - x \sim \mathcal{N}\left(0,\, \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\top\right). \tag{19}$$

Note that $\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\top \leq \Sigma_x$, i.e. covariance of estimation error is always less than prior covariance of x, measurements decrease the variance of our estimation error.

Example 6. *If measurements are linear $y = Ax + v$, $x \sim \mathcal{N}(\bar{x},\, \Sigma_x)$ and $v \sim \mathcal{N}(\bar{v},\, \Sigma_v)$, with x and v independent, we obtain the MMSE as an affine function*

$$\hat{x} = \bar{x} + B(y - \hat{y}) = \bar{x} + \Sigma_x A^\top (A\Sigma_x A^\top + \Sigma_v)^{-1}(y - A\bar{x} - \bar{v}). \tag{20}$$

Before measurement, $\hat{x}$ is the best prior guess of x. The difference $y - \hat{y}$ is the discrepancy between what we actually measure (y) and the expected value of what we measure ($\hat{y}$). Estimator modifies prior guess by B times this discrepancy, estimator blends prior information with measurement. B gives gain from observed discrepancy to estimate. Also note, that B is small if noise term Σ_v in 'denominator' is large.

2 Gaussian Processes

Gaussian processes[3] (GPs) generalize the concept of a Gaussian distribution over discrete random variables to the idea of a Gaussian distribution over continuous functions and inference taking place directly in the space of functions. GPs seen a lot of use in safe learning and control applications because of their ability to track the evolution of both the mean and covariance of the distribution. Hence, they allow for uncertainty quantification.

Definition 5. *A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Same as its finite dimensional counterpart, a GP is completely specified by its *mean function* and *covariance function*.

Definition 6. *For a real process $f(x)$ a mean function $m(x)$ and a covariance function $k(x, x')$ are defined as:*

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] \end{aligned}$$

and we will write a GP as

$$f(x) \sim \mathcal{GP}(m(x),\ k(x, x')) \tag{21}$$

[3] This section was adapted from [1], Chapter 2.

Example 7 (Bayesian linear model)**.** *Consider a linear regression model* $f(x) = \phi(x)^\top w$ *for* $x \in \mathbb{R}^n$, $\phi(x) : \mathbb{R}^n \to \mathbb{R}^p$ *and prior* $w \in \mathbb{R}^p$, $w \sim \mathcal{N}(0, \Sigma_p)$. *Then* $f(x)$ *defines a GP with*

$$\mathbb{E}[f(x)] = \phi(x)\mathbb{E}[w] = 0$$
$$\mathbb{E}[f(x)f(x')] = \phi(x)^\top \mathbb{E}[ww^\top]\phi(x') = \phi(x)^\top \Sigma_p \phi(x')$$

Thus, $f(x)$ *and* $f(x')$ *are jointly Gaussian with zero mean and covariance given by* $\phi(x)^\top \Sigma_p \phi(x')$.

In fact, the function values $f(x_1), \ldots, f(x_n)$, *for any number* $n > 0$, *are jointly Gaussian. However, if* $p < n$ *then this Gaussian is singular.*

Our running example of a covariance function will be the *squared exponential* (SE) or a *Radial Basis Function* (RBF):

$$\text{cov}(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\left(-\frac{1}{2}||x_p - x_q||_2^2\right) \tag{22}$$

Note, that the covariance between the *outputs* is written as a function of the *inputs*. This covariance function (or kernel) corresponds to a linear model with an infinite number of basis functions (see Section 4.3.1 in [1]), this avoiding the singularity issues raised in the above example.

2.1 Predictions. Sampling from a GP

For simplicity, we will assume that $m(x) = 0$ (not necessary). Thus, the GP is fully defined by the specified covariance function. If one specifies a set of input points $X_* = (X_*^1, \ldots, X_*^n)$, and constructs the kernel matrix $K(X_*, X_*)$, it is possible to generate a sample function at these inputs by simply drawing $f_* \sim \mathcal{N}(0, K(X_*, X_*))$. You can see such example in a Figure 3a. Here we plot 3 different function realizations at 50 points sampled from a GP with SE kernel from eq. (22).

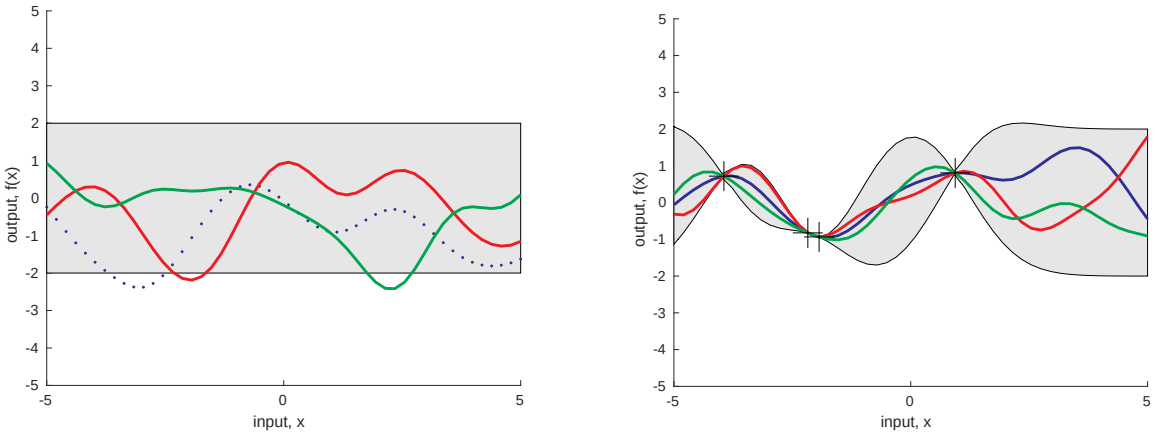

Figure 3: Panel (a) shows three functions drawn at random from a GP prior; the dots indicate values of y actually generated by GP; the two other functions (in red and green) have been drawn as lines by joining a large number of evaluated points. Panel (b) shows three random functions drawn from the posterior, i.e. the prior conditioned on the four noise free observations marked with cross symbols. In both plots the shaded area represents the pointwise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region), for the prior and posterior respectively.

2.1.1 Predictions from noise free observations

We are usually not primarily interested in drawing random functions from the prior, but want to incorporate the knowledge that the training data provides about the function. Initially, we will consider the simple special case where the observations are *noise free*, that is assume we are provided with a sample set:

$$S = \{(x_i, f(x_i)) \mid i = 1, \ldots, n\}.$$

According to the GP prior, the joint distribution of the training outputs given by $f = (f_i) = (f(x_i))$, and the test outputs $f = f(x_*)$ is defined by

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{23}$$

If there are n training points and n_* test points then $K(X, X_*)$ denotes the $n \times n_*$ matrix of the covariances evaluated at all pairs of training and test points, and similarly for the other entries $K(X,X) \in \mathbb{R}^{n\times n}$, $K(X_*,X_*) \in \mathbb{R}^{n_*\times n_*}$ and $K(X_*,X) \in \mathbb{R}^{n_*\times n}$.

To get the posterior distribution over functions we need to restrict this joint prior distribution to contain only those functions which agree with the observed data points, i.e. we need to take conditional expectation. The multivariate Gaussian distribution has the property that any conditional distribution is also Gaussian. Therefore, the distribution $f_*|X_*, X, f$ can be fully described with a mean and covariance matrix. We can describe that mean and covariance using the standard multivariate Gaussian conditional formula:

Lemma 1. *Function values f_* corresponding to test inputs X_* can be sampled from the joint posterior distribution by evaluating the mean and covariance matrix as following:*

$$f_*|X_*, X, f \sim \mathcal{N}\left(K(X_*,X)K(X,X)^{-1}f, \quad K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)\right). \tag{24}$$

Proof. See Appendix A.1. □

Figure 3b shows the results of these computations given the four data points marked with cross symbols.

Note, that the formula (24) is the same formula (29) with $\Sigma_x = K(X_*,X_*)$, $\Sigma_{xy} = K(X_*,X)$, $\Sigma_y = K(X,X)$.

2.1.2 Predictions from noisy observations

It is typical for more realistic modeling situations that we do not have access to function values $f(x)$ themselves, but only noisy versions of the form $y = f(x) + \epsilon$. Assuming additive independent identically distributed Gaussian noise, $\epsilon \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_n^2)$, then the prior on the noisy observations becomes

$$\mathrm{cov}(y_p, y_q) = k(x_p, x_q)) + \sigma_n^2 \delta_{pq} \tag{25}$$

where δ_{pq} is a Kronecker delta which is one iff $p = q$ and zero otherwise. Equivalently, using the matrix notation the above equation can be written as

$$\mathrm{cov}(y) = k(X, X) + \sigma_n^2 I \tag{26}$$

It follows from the independence assumption about the noise, that a diagonal matrix is added, in comparison to the noise free case, eq. (22). Introducing the noise term in eq. (23) we can write the joint distribution of the observed target values and the function values at the test locations under the prior as

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{27}$$

Deriving the conditional distribution corresponding to eq. (24) we arrive at the key predictive equations for Gaussian process regression

$$
\begin{aligned}
f_*|X_*, X, y \sim \mathcal{N}\left(\bar{f}_*, \mathrm{cov}(f_*)\right), \quad \text{where} \\
\bar{f}_* = \mu_{f_*} = K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} y \\
\mathrm{cov}(f_*) = K(X_*, X_*) - K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} K(X, X_*)
\end{aligned}
\tag{28}
$$

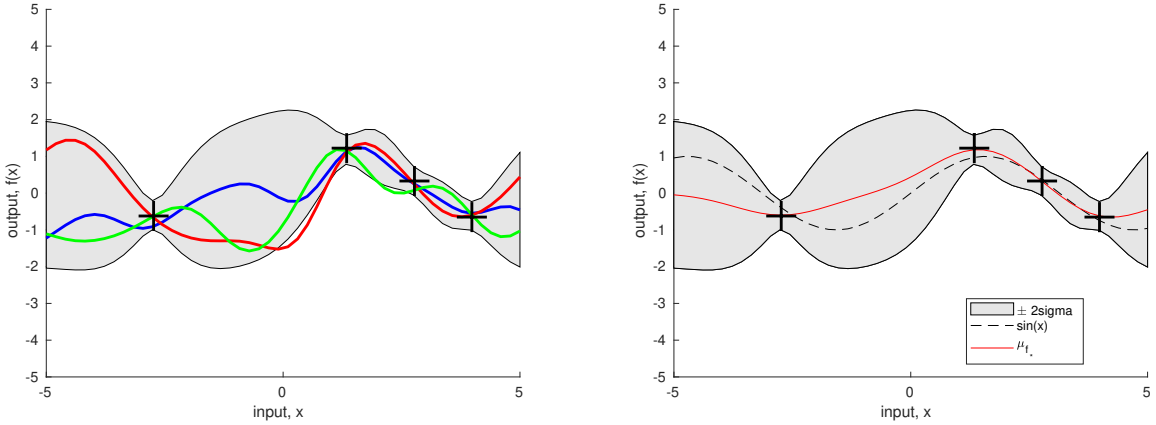

Figure 4: Observations with noise, $\sigma_n = 0.2$: Panel (a) shows three random functions drawn from the noisy posterior, i.e. the prior conditioned on the four observations with noise marked with cross symbols. Panel (b) shows the underlying signal $\sin(x)$ and the predicted mean $\bar{f}_*$ signal.

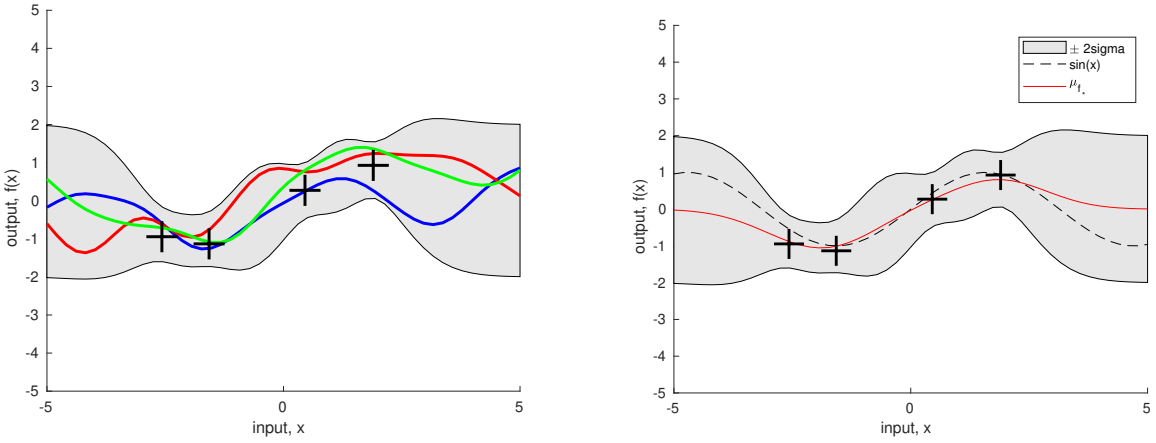

Figure 5: Observations with noise, $\sigma_n = 0.4$: Panel (a) shows three random functions drawn from the noisy posterior, i.e. the prior conditioned on the four observations with noise marked with cross symbols. Panel (b) shows the underlying signal $\sin(x)$ and the predicted mean $\bar{f}_*$ signal.

Consider the data given by $y = f(x) + \epsilon$ for $\sigma_n = 0.2$ in a Figure 4 and $\sigma_n = 0.4$ in a Figure 5, correspondingly. Subplot (a) in both figures present three realizations from posterior distribution, and subplot (b) draws the underlying signal $f(x) = sin(x)$ and predicted mean signal μ_{f_*}. The figures also show

the 2 standard-deviation error bars. Notice how the error bars get larger in sampled points as the variance σ_n gets higher ($\sigma_n = 0.2$ in Figure 4 in comparison with $\sigma_n = 0.4$ in Figure 5).

Note that the variance in eq. (28) does not depend on the observed targets y, but only on the inputs X and X_*. This is a property of the Gaussian distribution. The variance is the difference between two terms: the first term $K(X_*, X_*)$ is simply the prior covariance. From that is subtracted a (positive) term, representing the information the observations gives us about the function. We can very simply compute the predictive distribution of test targets y_* by adding $\sigma_n^2 I$ to the variance in the expression for $\text{cov}(f_*)$.

Let us now examine f_* evaluated at a single test point x_*. Then $K(x_*, X)$ is a vector $[k(x_*, x_i)]_{i=1,\ldots,n}$ and

$$\bar{f}_*(x_*) = \sum_{i=1}^{n} \alpha_i k(x_*, x_i), \tag{29}$$

for $\alpha = [K(X, X) + \sigma_n^2 I]^{-1} y$. So it as a linear combination of n kernel functions, each one centered on a training point. The fact that the mean prediction for $f(x_*)$ can be written as eq. (29) despite the fact that the GP can be represented in terms of a possibly infinite number of basis functions is one manifestation of the *representer theorem*, see Section 6.2 in [1].

2.2 Gaussian Process Regression as Linear Smoother

GP regression aims to reconstruct the underlying signal f by removing the contaminating noise ϵ[4]. To do this it computes a weighted average of the noisy observations y as

$$\bar{f}_*(x_*) = k(x_*)^\top (K + \sigma_n^2 I)^{-1} y \tag{30}$$

as $\bar{f}_*(x_*)$ is a linear combination of the y values, GP regression is a *linear smoother* [2].

The predicted mean values $\bar{\mathbf{f}}$ at the training points are given by

$$\bar{\mathbf{f}} = K(K + \sigma_n^2 I)^{-1} y, \quad \text{for } K = K(X, X) \tag{31}$$

Let K have the eigendecomposition

$$K = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \tag{32}$$

where λ_i is the i-th eigenvalue and $\mathbf{u}_i$ is the corresponding eigenvalue. As K is real and symmetric positive semidefinite, its eigenvalues are real and non-negative, and its eigenvectors are mutually orthogonal. Let $\mathbf{y} = \sum_{i=1}^{n} \gamma_i \mathbf{u}_i$ for some coefficients $\gamma_i = \mathbf{u}_i^\top \mathbf{y}$. Then

$$\bar{\mathbf{f}} = \sum_{i=1}^{n} \frac{\gamma_i \lambda_i}{\lambda_i + \sigma_n^2} \mathbf{u}_i. \tag{33}$$

Note that

- if $\lambda_i \ll \sigma_n^2$ then summand term ≈ 0, so the component in $\mathbf{y}$ along $\mathbf{u}_i$ is effectively eliminated.
- if $\lambda_i \gg \sigma_n^2$ then summand term ≈ 1, so the component is **not** eliminated.

Therefore, for most covariance functions that are used in practice the eigenvalues are larger for more slowly varying eigenvectors (e.g. fewer zero-crossings). This means that high-frequency components in $\mathbf{y}$ are smoothed out. It acts as a filter, letting through only those "terms above the noise".

3 Non-zero mean GPs: Incorporating Explicit Basis Functions

It is common but not necessary to consider GPs with a zero mean function[5]. Note that this is not necessarily a drastic limitation, since the mean of the posterior process is not confined to be zero. Yet there are several

[4]This section was adapted from Section 2.6 in [1].
[5]This section was adapted from Section 2.7 in [1].

reasons why one might wish to explicitly model a mean function. The use of explicit basis functions is a way to specify a *non-zero mean* over functions.

Using a *fixed* (deterministic) mean function $m(x)$ is trivial: Simply apply the usual zero mean GP to the *difference* between the observations and the fixed mean function. With

$$f(x) \sim \mathcal{GP}(m(x),\ k(x, x')) \tag{34}$$

the predictive mean becomes

$$\bar{f}_* = m(X_*) + K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1}(y - m(X)) \tag{35}$$

and the predictive variance $\mathrm{cov}(f_*)$ remains unchanged from eq. (28).

However, in practice it can often be difficult to specify a fixed mean function. In many cases it may be more convenient to specify a few fixed basis functions, whose coefficients, β, are to be inferred from the data. Consider

$$g(x) = f(x) = h(x)^\top \beta, \quad \text{where} \quad f(x) \sim \mathcal{GP}(0,\ k(x, x')), \tag{36}$$

here $f(x)$ is a zero mean GP, $h(x)$ are a set of fixed basis functions (for example, polynomial $h(x) = (1, x, x^2, \ldots)$), and β are additional parameters. This formulation expresses that the data is close to a global linear model with the residuals being modelled by a GP. When fitting the model, one could optimize over the parameters β jointly with the hyperparameters of the covariance function. Alternatively, if we take the prior on β to be Gaussian, $\beta \sim \mathcal{N}(b, B)$, we can also integrate out these parameters. Following [3], the obtained GP is

$$g(x) \sim \mathcal{GP}(h(x)^\top b,\ k(x, x') + h(x)^\top B h(x')). \tag{37}$$

The contribution in the covariance function caused by the uncertainty in the parameters of the mean. If we plug in the mean and covariance functions of $g(x)$ into eq. (36) and (28), we obtain

$$\bar{g}(X_*) = H_*^\top \bar{\beta} + K_*^\top K_y^{-1}(y - H^\top \bar{\beta}) = \bar{f}(X_*) + R^\top \bar{\beta}, \tag{38}$$

$$\mathrm{cov}(g_*) = \mathrm{cov}(f_*) + R^\top (B^{-1} + H K_y^{-1} H^\top)^{-1} R \tag{39}$$

where the H matrix collects the $h(x)$ vectors for all training cases, H_* collects all test points, and $\bar{\beta}$ and R are defined as

$$\begin{aligned} \bar{\beta} &= (B^{-1} + H K_y^{-1} H^\top)^{-1}(H K_y^{-1} y + B^{-1} b) \\ R &= H_* - H K_y^{-1} K_*. \end{aligned} \tag{40}$$

The interpretation of the mean expression $\bar{\beta}$, eq. (38): is the mean of the global linear model parameters, being a compromise between the data term and prior, and the predictive mean is simply the mean linear output plus what the GP model predicts from the residuals. The covariance $\mathrm{cov}(g_*)$, eq. (39), is the sum of the usual covariance term $\mathrm{cov}(f_*)$ and a new non-negative contribution.

Exploring the limit of the above expressions as the prior on the β parameter becomes vague, $B^{-1} \to O$ (where O is the matrix of zeros), we obtain a predictive distribution which is independent of b

$$\bar{g}(X_*) = \bar{f}(X_*) + R^\top \bar{\beta}, \tag{41}$$

$$\mathrm{cov}(g_*) = \mathrm{cov}(f_*) + R^\top (H K_y^{-1} H^\top)^{-1} R, \tag{42}$$

where the limiting $\bar{\beta} = (H K_y^{-1} H^\top)^{-1} H K_y^{-1} y$.

References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[2] Trevor Hastie and Robert Tibshirani. Generalized additive models. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.8665`, 1990.

[3] Anthony O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

[4] Thomas B Schön and Fredrik Lindsten. Manipulating the multivariate gaussian density. *Div. Automat. Control, Linköping Univ., Linkping, Sweden, Tech. Rep*, 2011.

A Appendix

A.1 Partitioned Gaussian Densities

Theorem 1. *[4] Let the random vector x be Gaussian $x \sim \mathcal{N}(\mu, \Sigma)$ with mean and variance as following*

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_{bb} \end{bmatrix}, \tag{43}$$

then the conditional density $p(x_a|x_b)$ is given by

$$p(x_a|x_b) = \mathcal{N}(\mu_{a|b},\, \Sigma_{a|b},\,) \tag{44}$$

where

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \tag{45}$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab}^\top \tag{46}$$

Proof. We will make use of the fact that

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)} \tag{47}$$

which is according to the definition of the normal distribution (see eq.(1)) is

$$p(x_a|x_b) = \frac{\sqrt{\det \Sigma_{bb}}}{(2\pi)^{n_a/2}\sqrt{\det \Sigma}} \exp(E) \tag{48}$$

$$E = -\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) - \frac{1}{2}(x_b - \mu_b)^\top \Sigma_{bb}^{-1}(x_b - \mu_b) \tag{49}$$

Since

$$\det \Sigma = \det \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_{bb} \end{pmatrix} = \det \Sigma_{bb} \det(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab}^\top) \tag{50}$$

the constant in front of the exponential in (48) results in the following expression

$$\frac{\sqrt{\det \Sigma_{bb}}}{(2\pi)^{n_a/2}\sqrt{\det \Sigma}} = \frac{1}{(2\pi)^{n_a/2}\sqrt{\det(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab}^\top)}} = \frac{1}{(2\pi)^{n_a/2}\sqrt{\det \Lambda_{aa}^{-1}}} \tag{51}$$

The precision matrix E, eq. (49), is given by

$$\begin{aligned}
E &= -\frac{1}{2}(x-\mu)^\top \Lambda (x-\mu) - \frac{1}{2}(x_b - \mu_b)^\top \Sigma_{bb}^{-1}(x_b - \mu_b) = \\
&= -\frac{1}{2}(x_a - \mu_a)^\top \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^\top \Lambda_{ab}^{-1}(x_b - \mu_b) - \\
&\quad - \frac{1}{2}(x_b - \mu_b)^\top \Lambda_{ba}^{-1}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^\top (\Lambda_{bb} - \Sigma_{bb}^{-1})(x_b - \mu_b) = \\
&= -\frac{1}{2}x_a^\top \Lambda_{aa} x_a + x_a^\top(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)) - \\
&\quad - \frac{1}{2}\mu_a^\top \Lambda_{aa}\mu_a + \frac{1}{2}\mu_a^\top \Lambda_{ab}(x_b - \mu_b) - \frac{1}{2}(x_b - \mu_b)^\top (\Lambda_{bb} - \Sigma_{bb}^{-1})(x_b - \mu_b).
\end{aligned}$$

Using the block matrix inversion result

$$\Sigma_{bb}^{-1}\Lambda_{bb} - \Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab} \tag{52}$$

and completing the squares results in

$$E = -\frac{1}{2}(x_a - (\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)))^\top \Lambda_{aa}(x_a - (\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b))) \tag{53}$$

Finally, combining (51)and (53) results in

$$p(x_a|x_b) = \frac{1}{(2\pi)^{n_a/2}\sqrt{\det \Lambda_{aa}^{-1}}} \exp(E) \tag{54}$$

which concludes the proof. □