Gaussian Processes

We previously considered the **kernel trick,** which allowed us to define the **kernel function**

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \mathbf{\Sigma} \phi(\mathbf{x}').$$

Then, when doing Bayesian linear regression with prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{\Sigma}),$$

and data $\mathcal{D}$, we derived the predictive distribution

$$p(\mathbf{y}^* \mid \mathbf{X}^*, \mathcal{D}, \sigma^2) = \mathcal{N}(\mathbf{y}^*; \boldsymbol{\mu}_{\mathbf{y}^*|\mathcal{D}}, \mathbf{K}_{\mathbf{y}^*|\mathcal{D}}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{y}^*|\mathcal{D}} &= \mathbf{K}^{*T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\
\mathbf{K}_{\mathbf{y}^*|\mathcal{D}} &= \mathbf{K}^{**} - \mathbf{K}^{*T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}^* + \sigma^2\mathbf{I}^*,
\end{aligned}$$

and we have defined

$$\mathbf{K} = K(\mathbf{X}, \mathbf{X}) \qquad \mathbf{K}^* = K(\mathbf{X}, \mathbf{X}^*) \qquad \mathbf{K}^{**} = K(\mathbf{X}^*, \mathbf{X}^*).$$

Examining the form of this predictive distribution, we barely need to perform inference over the weight vector $\mathbf{w}$ at all; its covariance simply appears in the definition of K. If we treat K as a black-box, maybe we can skip the inference over $\mathbf{w}$ entirely and simply work with K directly. This line of thinking motivates **Gaussian processes.**

Instead of reasoning about $\mathbf{w}$ and ϕ, the Gaussian process approach to regression directly reasons about the *function* f: we construct an explicit prior distribution for f, that we reason about via the Bayesian method. This seems like it might be difficult, because f might in general be infinite-dimensional. A Gaussian process provides a natural extension of the familiar multivariate Gaussian distribution to potentially infinite-dimensional function spaces. The key definition is the following.

Definition 1. *A* Gaussian process *(GP) is a (potentially infinite) collection of random variables such that the joint distribution of any finite number of them is multivariate Gaussian.*

These infinitely many random variables can be thought of as the values of the function f at every location $x \in \mathcal{X}$. The way we extend the finite-dimensional Gaussian distribution to such a collection of variables is by allowing arbitrary inputs but only ever reasoning about finitely many of them at a time.

A GP prior distribution on f is written

$$p(f) = \mathcal{GP}(f; \mu, K),$$

and just like the multivariate Gaussian distribution, is parameterized by its first two moments (now functions):

- $\mathbb{E}[f] = \mu \colon \mathcal{X} \to \mathbb{R}$, the **mean function,** and
- $\mathbb{E}\big[\big(f(x) - \mu(x)\big)\big(f(x') - \mu(x')\big)\big] = K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a positive semidefinite **covariance function** or **kernel.**[1]

[1] A function is **positive semidefinite** if, for every finite set of points $\mathbf{X}$, the **Gram matrix** $\mathbf{K} = K(\mathbf{X}, \mathbf{X})$ is positive semidefinite. Note that a matrix is a valid covariance matrix if and only if it is positive semidefinite.

The function K should look familiar! Rather than explicitly construct K, we simply select any positive semidefinite function, any one of which will define a GP.

The mean function encodes the central tendency of the function, and is often assumed to be a constant (usually zero). The covariance function encodes information about the shape and structure we expect the function to have. A simple and very common example is the **squared exponential** covariance:

$$K(\mathbf{x}, \mathbf{x}'; \lambda, \ell) = \lambda^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right),$$

which encodes the notation that "nearby points should have similar function values."

Suppose we have selected a GP prior $\mathcal{GP}(f; \mu, K)$ for the function f. Consider a finite set of points $\mathbf{X} \subseteq \mathcal{X}$. The GP prior on f, by definition, implies the following joint distribution on the associated function values $\mathbf{f} = f(\mathbf{X})$:

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})).$$

That is, we simply evaluate the mean and covariance functions at $\mathbf{X}$ and compute the associated multivariate Gaussian distribution.

Connection to Bayesian linear regression

As it turns out, we have been specifying a Gaussian process on a latent function $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ all along! Let us again write down the abstract problem of regression. We have an unknown function $f \colon \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is an arbitrary input space (for example $\mathcal{X} = \mathbb{R}^d$).

In linear regression, we made the assumption that f was linear: $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ where ϕ is some arbitrary feature expansion. Given a Gaussian prior on $\mathbf{w}$, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and input locations $\mathbf{X}$, the prior distribution for the vector of latent values $\mathbf{f} = f(\mathbf{X}) = \boldsymbol{\Phi}\mathbf{w}$ is a linear transformation of the Gaussian-distributed vector $\mathbf{w}$:

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\Phi}\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top).$$

Therefore we have shown that the joint distribution of any finite combination of our random variables $\{f(\mathbf{x})\}_{\mathbf{x} \in \mathcal{X}}$ is multivariate Gaussian, and thus according to the above definition we have a Gaussian process! We may identify the mean and covariance functions of the Bayesian linear regression Gaussian process:

$$\mu(\mathbf{x}) = \phi(\mathbf{x})^\top \boldsymbol{\mu}$$
$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \boldsymbol{\Sigma} \phi(\mathbf{x}).$$

By starting over "from the beginning," we may free ourselves from the constraint of needing to explicitly specify the feature expansion function ϕ. Instead, we may choose arbitrary mean and covariance functions μ and K and reason about f in an essentially identical manner.

Posterior GPs

Given a prior GP belief on f, $\mathcal{GP}(f; \mu, K)$, and after observing some training data $\mathcal{D} = (\mathbf{X}, \mathbf{f})$, suppose we wish to make predictions about the value of f at some test points $\mathbf{X}^*$.

We begin by writing the joint distribution between the training function values $f(\mathbf{X}) = \mathbf{f}$ and the test function values $f(\mathbf{X}^*) = \mathbf{f}^*$:

$$p(\mathbf{f}, \mathbf{f}^*) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f} \\ \mathbf{f}^*\end{bmatrix}; \begin{bmatrix}\mu(\mathbf{X}) \\ \mu(\mathbf{X}^*)\end{bmatrix}, \begin{bmatrix}K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*)\end{bmatrix}\right).$$

We can then condition this multivariate Gaussian on the known training values $\mathbf{f}$, something we already know how to do!

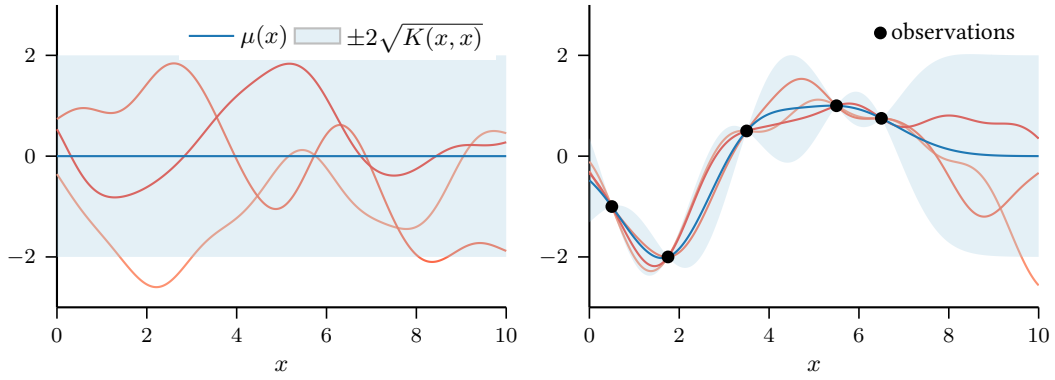
$$p(\mathbf{f}^* \mid \mathbf{X}^*, \mathcal{D}) = \mathcal{N}\big(\mathbf{f}^*; \mu_{f|\mathcal{D}}(\mathbf{X}^*), K_{f|\mathcal{D}}(\mathbf{X}^*, \mathbf{X}^*)\big),$$

where

$$\mu_{f|\mathcal{D}}(\mathbf{x}) = \mu(\mathbf{x}) + K(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}\big(\mathbf{f} - \mu(\mathbf{X})\big)$$
$$K_{f|\mathcal{D}}(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}K(\mathbf{X}, \mathbf{x}').$$

Notice that the functions $\mu_{f|\mathcal{D}}$ and $K_{f|\mathcal{D}}$ are valid mean and covariance functions, respectively. This means that the posterior distribution over f is itself a Gaussian process! This can also be inferred from the fact that the posterior distribution over any finite set of test points $\mathbf{X}^*$ is a multivariate Gaussian.

Figure 1 shows an example of going from a GP prior to a GP posterior; the prior is a zero-mean GP with a squared exponential covariance function where $\lambda = \ell = 1$.



(a) GP prior with samples

(b) GP posterior with samples

Figure 1: GP beliefs before and after observing data. Each figure also shows samples drawn from the depicted GP belief to give a sense for the kind of functions modeled by the belief: note how the squared exponential kernel corresponds to smooth sample functions.

Dealing with noise

GPs deal with observation noise in effectively the same way as Bayesian linear regression. Here we will continue to assume that the observed values $\mathbf{y}$ are generated by adding zero-mean, independent Gaussian noise with variance σ^2 to the true function values $\mathbf{f} = f(\mathbf{X})$. Probabilistically, we write

$$p(\mathbf{y} \mid \mathbf{f}, \sigma^2) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2\mathbf{I}).$$

Starting with a GP prior on f, $\mathcal{GP}(f; \mu, K)$, but this time given noisy observations $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, we again write the joint distribution between the training function values $\mathbf{y}$ and the test function values $\mathbf{f}^*$:

$$p(\mathbf{y}, \mathbf{f}^*) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix}; \begin{bmatrix}\mu(\mathbf{X})\\ \mu(\mathbf{X}^*)\end{bmatrix}, \begin{bmatrix}K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I} & K(\mathbf{X}, \mathbf{X}^*)\\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*)\end{bmatrix}\right).$$

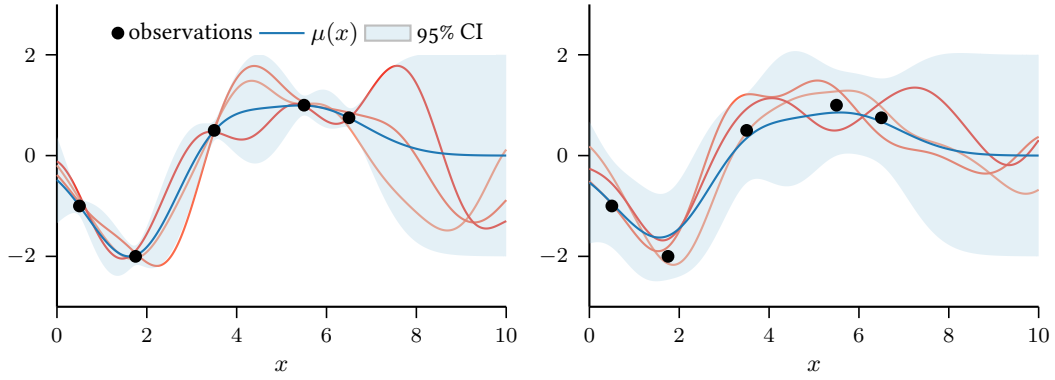
Conditioning as before, we end up with:

$$p(\mathbf{f}^* \mid \mathbf{X}^*, \mathcal{D}) = \mathcal{N}\big(\mathbf{f}^*; \mu_{f|\mathcal{D}}(\mathbf{X}^*), K_{f|\mathcal{D}}(\mathbf{X}^*, \mathbf{X}^*)\big),$$

where

$$\mu_{f|\mathcal{D}}(\mathbf{x}) = \mu(\mathbf{x}) + K(\mathbf{x}, \mathbf{X})\big(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}\big(\mathbf{y} - \mu(\mathbf{X})\big)$$
$$K_{f|\mathcal{D}}(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{X})\big(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}K(\mathbf{X}, \mathbf{x}').$$

Figure 2 shows the effect of different noise levels σ, using the same GP prior and data set depicted in Figure 1.



(a) $\sigma = 0.1$ (b) $\sigma = 0.5$

Figure 2: Noisy GP posteriors with samples: note how in the setting with higher noise, the mean and samples no longer pass directly through the observations as more of the observations' behavior is ascribed to noise.

Hyperparameters

So far, we have assumed that the Gaussian process prior distribution on f has been specified *a priori*. But this prior distribution itself will generally have parameters that we need to specify: for example, the kernel length scale ℓ, the kernel output scale λ, and the noise variance σ^2. As parameters of a prior distribution, we call these **hyperparameters.**

For notational convenience, we will write θ to denote the vector of all hyperparameters of the model (including of μ and K). Assume we have chosen a prior

$$p(f \mid \theta) = \mathcal{GP}\big(f; \mu(\mathbf{x}; \theta), K(\mathbf{x}, \mathbf{x}'; \theta)\big)$$

where the dependence of μ and K on the parameters θ has been made explicit. We will measure the quality of the fit to our training data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ with the **marginal likelihood,** the probability of observing $\mathcal{D}$ under our prior:

$$p(\mathbf{y} \mid \mathbf{X}, \theta) = \int p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, \theta)\, \mathrm{d}\mathbf{f},$$

where we have marginalized the unknown function values $\mathbf{f}$ (hence, marginal likelihood).

Thankfully, this is an integral we can do analytically under the additive Gaussian noise assumption!

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{X}, \theta) &= \int p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, \theta)\, \mathrm{d}\mathbf{f} \\ &= \int \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I})\, \mathcal{N}\big(\mathbf{f}; \mu(\mathbf{X}; \theta), K(\mathbf{X}, \mathbf{X}; \theta)\big)\, \mathrm{d}\mathbf{f} \\ &= \mathcal{N}\big(\mathbf{y}; \mu(\mathbf{X}; \theta), K(\mathbf{X}, \mathbf{X}; \theta) + \sigma^2 \mathbf{I}\big) \end{aligned}$$

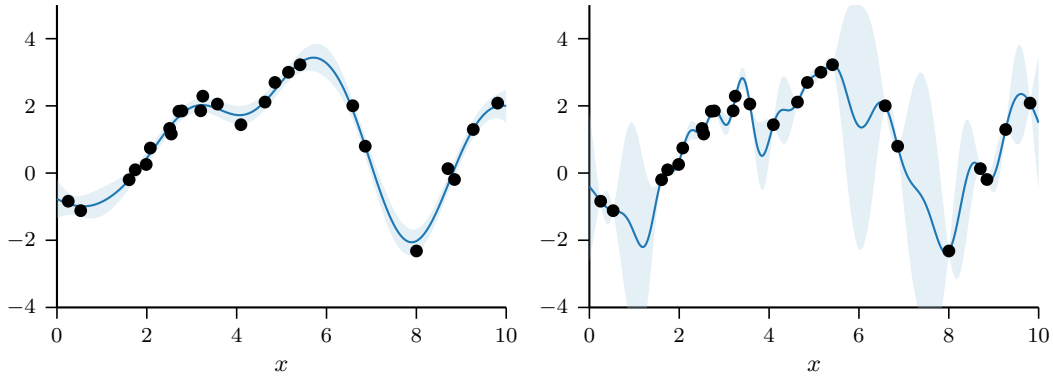
where we have made use of the closure under convolutions property of Gaussians.

The log-likelihood of our data under the chosen prior is then

$$\begin{aligned} \log p(\mathbf{y} \mid \mathbf{X}, \theta) = &- \frac{(\mathbf{y} - \mu(\mathbf{X}; \theta))^\top \big(K(\mathbf{X}, \mathbf{X}; \theta) + \sigma^2 \mathbf{I}\big)^{-1} (\mathbf{y} - \mu(\mathbf{X}; \theta))}{2} \\ &- \frac{\log \det \big(K(\mathbf{X}, \mathbf{X}; \theta) + \sigma^2 \mathbf{I}\big)}{2} - \frac{N \log 2\pi}{2} \end{aligned}$$

There is an inherent trade off between the first two terms in this expression: the first term is large when the data fit the model well and the second term is large when the volume of the prior covariance is small, i.e., when the model is simpler.

Figure 3 shows the effect of different settings of the model hyperparameters, (λ, ℓ, σ), when conditioning a GP prior on a fixed set of observations to arrive at a GP posterior: note how in the setting with higher noise, the mean and samples no longer pass directly through the observations as more of the observations' behavior is ascribed to noise.



(a) $\theta = (\lambda, \ell, \sigma) = (1, 1, 1/5)$ (b) $\theta = (\lambda, \ell, \sigma) = (2, 1/3, 1/20)$

Figure 3: GPs with different hyperparameters fit to the same data set: the log-likelihood of the data under the hyperparameter setting in Figure 3a is -27.6 and under the hyperparameter setting in Figure 3b, it is -46.5.

To be fully Bayesian, we would choose a **hyperprior** over θ, $p(\theta)$, and marginalize the unknown hyperparameters when making predictions:

$$p(f^* \mid \mathbf{x}^*, \mathcal{D}) = \frac{\int p(f^* \mid \mathbf{x}^*, \mathcal{D}, \theta) p(\mathbf{y} \mid \mathbf{X}, \theta) p(\theta)\, \mathrm{d}\theta}{\int p(\mathbf{y} \mid \mathbf{X}, \theta) p(\theta)\, \mathrm{d}\theta}$$

Unfortunately, this integral cannot generally be resolved analytically (of course…).

Instead, if we believe the posterior distribution over θ is well-concentrated (for example, if we have many training examples), we may approximate $p(\theta \mid \mathcal{D})$ with a Dirac delta distribution at the point with maximum marginal likelihood:

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathbf{y} \mid \mathbf{X}, \theta).$$

This is called **maximum likelihood-II** (ML-II) inference and effectively makes the approximation

$$p(f^* \mid \mathbf{x}^*, \mathcal{D}) \approx p(f^* \mid \mathbf{x}^*, \mathcal{D}, \theta_{\text{MLE}}).$$

As long as $\mu(\mathbf{X}; \theta)$ and $K(\mathbf{X}, \mathbf{X}; \theta)$ are differentiable w.r.t. θ, then we can compute the gradient $\partial \log p(\mathbf{y}|\mathbf{X},\theta)/\partial\theta$. This allows us to find θ_{MLE} by minimizing the negative log likelihood (which is equivalent to maximizing the likelihood) using off–the–shelf gradient–based methods.