These notes draw heavily on Rasmussen & Williams, the classic text on Gaussian Processes.

1. What is a Gaussian Process?

A Gaussian Process is "a collection of random variables, any finite number of which have a joint Gaussian distribution." That is, it is a just a Gaussian random distribution, though it is of potentially infinite dimension. Do not be overly distracted by the "process" in the name.

Typically, we use a set of continuous independent variables $\vec{x}$, sometimes called "input variables." The examples I go through will be all one dimensional so I will just refer to x, but increasing the dimensionality of the independent variable space doesn't matter much.

The Gaussian Process outputs will be referred to as $f(x)$, and the distribution of $f(x)$ is fully defined by the mean and covariance:

$$\begin{aligned} m(x) &= \langle f(x) \rangle \\ k(x,x') &= \big\langle (f(x) - m(x))(f(x') - m(x')) \big\rangle \end{aligned} \tag{1}$$

where you will hear $k()$ also referred to as the "kernel." So $k(x,x)$ is just the variance of the output $f(x)$, and for $x \neq x'$ the kernel $k(x,x')$ is an off-diagonal covariance.

Why would we define such a thing as a Gaussian Process? A good example is the case of an irregularly sampled function at points x, with potentially large gaps—say elevation measurements across the surface of the Earth (which would have been a familiar data set to Gauss himself). We want to have some estimate of the elevation in between the measurements, at other points x^*, and we want those estimates to obey some expectations we have about relative smoothness. Another way of expressing smoothness is to say that the variations from point to point are correlated, or covariant. Points close together are well correlated, and points further away are less correlated.

This line of reasoning would motivate making predictions about the new points x^* using a Gaussian Process. Typically we set the kernel, and then use measured $f(x)$ to make predictions for $m(x^*)$. Along with the predictions $m(x^*)$ we also have predictions for their variance and covariance. If the kernel is "good," these variances are extraordinarily useful—they give you some idea of how certain to be of the values you would measure at points x^*, and for a surveyor might allow them to find useful places to take new measurements.

A common example of a kernel is the so-called "squared exponential":

$$k(x,x') = \exp\left(\frac{1}{2}\left|x - x'\right|^2\right) \tag{2}$$

which is definitely not a squared exponential, but that is what is called. Since the Gaussian Process people definitely know what a Gaussian looks like, I assume they have their own reasons for this

nomenclature, possibly to avoid confusing what they call the kernel (which need not be Gaussian) with the very necessary Gaussian distribution of $f(x)$.

2. A two point example

Let us imagine that we have two distinct values of an independent variable, x_1 and x_2, and we are interested in a function $f(x)$ at those points. Let's say we know $f(x_1)$, what can we say about x_2?

A Gaussian Process approach to this question is to assume that $f(x_1)$ and $f(x_2)$ are drawn from a joint Gaussian distribution with zero mean and with some covariance matrix:

$$\mathbf{C} = \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix} \tag{3}$$

and in fact these elements are determined by a kernel function $k(x, x')$, like the squared exponential, so the closer together the x values, the more correlated the variables.

If this is the case, our question becomes, "what is the distribution of $f(x_2)$ conditional on $f(x_1)$?" For any multivariate Gaussian, a conditional distribution on one dimension is also Gaussian. Specifically, it will be defined by its "predictive mean" and "predictive variance":

$$\begin{aligned} \langle f(x_2) \rangle &= \frac{k_{12}}{k_{11}} f(x_1) = \sqrt{\frac{k_{22}}{k_{11}}} r f(x_1) \\ \sigma_{f2}^2 = \left\langle [f(x_2) - \langle f(x_2) \rangle]^2 \right\rangle &= k_{22} - \frac{k_{12}^2}{k_{11}} = k_{22}\left(1 - r^2\right) \end{aligned} \tag{4}$$

where in the last equality, which is shown just because it might be familiar, r is the correlation coefficient $k_{12}/\sqrt{k_{11}k_{22}}$.

So what is the behavior of this? At some given distance $x_1 - x_2$, the expected mean value of $f(x_2)$ is some fraction of $f(x_1)$. The closer the points, the closer the ratio is to unity, and so the closer the values. As the points get very far apart, the expectation for $f(x_2)$ reverts to the mean of zero. Meanwhile, the expected variance of the points starts out relatively small at close distances, but grows up to the auto-covariance at large distances. If you ask, as r differs from unity, how does the mean and standard deviation of $f(x_2)$ change, you find that the standard deviation changes rapidly and the mean more slowly, so that the standard deviation always exceeds it—i.e. although the expectation value of $f(x_2)$ drifts towards zero, the associated uncertainties always are consistent with $f(x_2) = f(x_1)$.

3. Gaussian Process Prediction

The usual case is that we have measurements of $f(x)$ at a few points $\vec{x}$ and we want to make predictions at some other points $\vec{x}^*$. This is just the same as the two point case, but we condition

on all the $\vec{x}$, and look at the resulting (still Gaussian!) mean and variance of $\vec{x}^*$.

There are some important identities of multivariate Gaussians that make this simple. We will use the math-y notation that "$\sim$" means "is distributed as" and $\mathcal{N}(\vec{\mu}, \mathbf{C})$ indicates a Gaussian distribution with mean $\vec{\mu}$ and covariance matrix $\mathbf{C}$. The full Gaussian distribution of all the values at xs of interest is:

$$\begin{bmatrix} f(\vec{x}) \\ f(\vec{x}^*) \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \mathbf{K}(\vec{x}, \vec{x}) & \mathbf{K}(\vec{x}, \vec{x}^*) \\ \mathbf{K}(\vec{x}^*, \vec{x}) & \mathbf{K}(\vec{x}^*, \vec{x}^*) \end{bmatrix}\right) \tag{5}$$

where the vector and matrix are written in block form, and $\mathbf{K}(\vec{x}, \vec{x})$ just is a block with $k(x_i, x_j)$ as each element K_{ij} and similar in the other blocks.

It turns out that if I condition this matrix on $f(\vec{x})$, I can write:

$$f(\vec{x}^*)|f(\vec{x}) \sim \mathcal{N}\left(\mathbf{K}(\vec{x}^*, \vec{x}) \cdot \mathbf{K}(\vec{x}, \vec{x})^{-1} \cdot f(\vec{x}), \mathbf{K}(\vec{x}^*, \vec{x}^*) - \mathbf{K}(\vec{x}^*, \vec{x}) \cdot \mathbf{K}(\vec{x}, \vec{x})^{-1} \cdot \mathbf{K}(\vec{x}, \vec{x}^*)\right) \tag{6}$$

If you look carefully, this is the same equation we used before for the two point case, now in matrix form.

In implementation, instead of the inverse of $K(\vec{x}, \vec{x})$, often its Cholesky decomposition is used:

$$\mathbf{L} \cdot \mathbf{L}^T = \mathbf{K} \tag{7}$$

Then this calculation:

$$\mathbf{K}^{-1} \cdot f(\vec{x}) = \vec{q} \tag{8}$$

can be rewritten as:

$$\begin{aligned} \mathbf{K} \cdot f(\vec{x}) &= \vec{q} \\ \mathbf{L} \cdot \left(\mathbf{L}^T \cdot \vec{q}\right) &= f(\vec{x}) \end{aligned} \tag{9}$$

So if you solve the equation:

$$\mathbf{L} \cdot \vec{r} = f(\vec{x}) \tag{10}$$

and then

$$\mathbf{L}^T \cdot \vec{q} = f(\vec{x}) \tag{11}$$

You then can write the predictive mean as:

$$\mathbf{K}(\vec{x}^*, \vec{x}) \cdot \vec{q}. \tag{12}$$

Similarly the predictive variance can be written using:

$$\begin{aligned} \mathbf{V} &= \mathbf{L}^{-1} \cdot \mathbf{K}(\vec{x}, \vec{x}^*) \\ \mathbf{L} \cdot \mathbf{V} &= \mathbf{K}(\vec{x}, \vec{x}^*) \end{aligned} \tag{13}$$

and then:

$$\mathbf{K}(\vec{x}^*, \vec{x}^*) - \mathbf{V}^T \cdot \mathbf{V} \tag{14}$$

This procedure is generally more stable.

The very most stable thing to do is to solve these systems using the pseudo-inverse (i.e. use the SVD decomposition of $\mathbf{L}$, which can be achieved with `numpy.linalg.lstsq`.