

Math 106 – Notes

Week 3: January 22, 2026

Ethan Levien

Gaussian process (5.4)

Gaussian process as functional generalizations of multivariate normals

Throughout this section we assume all random variables and stochastic processes have mean zero. This assumption simplifies notation and calculations and does not affect any of the conceptual conclusions.

Maximum entropy principle (informal statement). Among all probability distributions on \mathbb{R}^n with a prescribed covariance matrix Σ , the mean-zero multivariate Gaussian distribution with covariance Σ is the unique distribution of maximal entropy. Mathematically, this means the Gaussian density,

$$\rho(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}x^\top \Sigma^{-1}x\right),$$

satisfies the optimization problem

$$\max_{\rho} \left\{ - \int_{\mathbb{R}^n} \rho(x) \log \rho(x) dx : \int_{\mathbb{R}^n} x \rho(x) dx = 0, \int_{\mathbb{R}^n} xx^\top \rho(x) dx = \Sigma \right\}.$$

where the max is taken over all probability densities. In this sense it is the “least informative” choice compatible with the covariance Σ .

This maximum-entropy characterization provides a natural motivation for Gaussian processes. Suppose we wish to model a random function $X = \{X_t\}_{t \in \mathbb{R}}$ while specifying only its second-order statistics, encoded by a covariance function

$$K(s, t) = \text{cov}(X_s, X_t), \quad s, t \in \mathbb{R}. \tag{1}$$

The guiding principle is to choose a stochastic process that is otherwise as unstructured as possible, subject only to this covariance constraint. The reasonable way to construct such a process would be for the finite dimensional distributions to be Gaussian, and the Kolmogorov extension Theorem tells us this process can be extended to the real line or any interval. Thus we have the following definition.

Definition 1 (Gaussian process (D5.9)). *A stochastic process $\{X_t\}_{t \in [0, T]}$ is called a Gaussian process (GP) with covariance function K if for every finite collection of indices $t_1, \dots, t_n \in [0, T]$, the random vector*

$$(X_{t_1}, \dots, X_{t_n}) \tag{2}$$

is multivariate normal with mean zero and covariance matrix

$$\Sigma_{i,j} = K(t_i, t_j). \tag{3}$$

Gaussian process in the context of regression modeling

The most obvious application of a GP is as a prior distribution for an unknown function when performing interpolation, or smoothing. In particular, in application we often have measurements Y_1, Y_2, \dots, Y_n which are assumed to be noisy measurements of some true function $f(t)$:

$$Y_i = f(t_i) + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma_\epsilon^2). \quad (4)$$

Often we proceed by picking some parametric class of functions for $f(t)$ and fitting the parameters. To achieve this using least squares the dependence on the parameters $\{\beta_k\}_k$ must be linear and therefore $f(t)$ is expressed as an expansion in basis functions $\{\phi_k\}_k$; that is,

$$f(t) = \sum_{k=1}^m \beta_k \phi_k(t). \quad (5)$$

The β_k can then be found with least squares by minimizing $\|\mathbf{Y} - \mathbf{f}\|_2^2$ where \mathbf{f} and \mathbf{Y} are respectively vectors of measurements and predicted values of the function. Here, we view $f(t)$ as a deterministic function for each set of β_k .

As $m \rightarrow \infty$ we need to regularize and in the Bayesian setting this is done by introducing a prior distribution on the β . The natural choice is for β_k to be Normal and independent between the k :

$$\beta_k \sim \text{Normal}(0, \sigma_k^2). \quad (6)$$

We now can view $f(t)$ as a random function – in-fact, a GP. We can in principle perform Bayesian inference to obtain the posterior distribution of the β from our data; however, we may not care about the β since our end goal is actually to perform interpolation, meaning make predictions about the function values at intermediate t . The idea of Gaussian processes regression is to perform Bayesian inference directly on the function space. In other words, we treat the random functions $f(t)$ as our parameters. Let's go back to our usual notation for stochastic processes and write $X_t = f(t)$.

Eigenfunction expansion

Linear algebra approach

The goal of this section is to connect the kernel view to the series expansion/prior distributions (Eq. 5 and Eq. 6) and to do this rigorously requires a bit of background on linear operators on Hilbert spaces. However the basic ideas can be understood with more elementary linear algebra calculations. Using that the β are independent, we have

$$K(t, s) = \mathbb{E}[X_t X_s] = \sum_{k=1}^{\infty} \sum_{k'=1}^{\infty} \mathbb{E}[\beta_k \beta_{k'}'] \phi_k(t) \phi_{k'}(s) = \sum_{k=1}^{\infty} \sigma_k^2 \phi_k(t) \phi_k(s) \quad (7)$$

Now take a finite set of times (t_1, \dots, t_k) and let Σ be the covariance matrix of $(X_{t_1}, \dots, X_{t_k})$. Eq. 7 says that

$$\Sigma = \Phi \Lambda \Phi^T \quad (8)$$

where $\Lambda = \text{diag}(\sigma_1^2, \sigma_2^2, \dots)$ and $\Phi_{k,i} = \phi_k(t_i)$. Note that Φ and Λ are infinite dimensional matrices but linear algebra still works. Therefore, to go from the kernel view of a GP to the series expansion, the basis functions and weight variances are obtained as the eigenfunctions and eigenvalues of the covariance operator respectively.

The kernel as a linear operator on L_T^2

We are now going to see how the same idea works, but working with linear operators on a function space rather than the covariance matrix. This will then be used to justify the infinite dimensional matrix equation 8. For a kernel K on $[0, T]$, introduce the linear operator

$$\mathcal{K} : L_T^2 \rightarrow L_T^2, \quad (9)$$

where

$$L_T^2 = \left\{ f : [0, T] \rightarrow \mathbb{R} : \int_0^T f(t)^2 dt < \infty \right\}. \quad (10)$$

This operator is given by

$$(\mathcal{K}f)(s) = \int_0^T K(s, t) f(t) dt. \quad (11)$$

It is important to understand some properties of L_T^2 . In particular, that this is a Hilbert space $H = L_T^2$. Recall this means that

- H it is a vector space in the sense that the function can be combined and multiplied by scalars just like \mathbb{R}^n
- There is an inner product $\langle \cdot, \cdot \rangle : H \times H \rightarrow \mathbb{R}$ just like the vector inner product. In particular $\langle f, g \rangle = \langle g, f \rangle$. In L_T^2 ,

$$\langle f, g \rangle = \int_0^T f(t)g(t)dt \quad (12)$$

- The inner product induces a norm, $\|f\| = \sqrt{\langle f, f \rangle}$ under which H is complete, meaning sequences $\{f_n\}_n$ which converge in this norm converge to elements of H .

We make the following assumption, which ensures \mathcal{K} has nice behavior. In the book this appears in T5.10.

Assumption 1. For the GP $\{X_t\}_{t \in [0, T]}$

$$\int_0^T \mathbb{E}[X_s^2] ds = C < \infty \quad (13)$$

for some constant C and $K(s, t)$ is continuous in s and t .

The operator \mathcal{K} associated with a process satisfying Assumption 1 has the following properties.

- First, \mathcal{K} is symmetric in L_T^2 (this does not require Assumption 1) and simply follows from the definition of K .
- the linear operator \mathcal{K} is *bounded*, meaning $\|\mathcal{K}f\| \leq C\|f\|$.

Proof.

$$\|\mathcal{K}f\|^2 = \int_0^T \left(\int_0^T K(s, t) f(t) dt \right)^2 ds \quad (14)$$

$$\leq \int_0^T \int_0^T K(s, t)^2 dt ds \left(\int_0^T f(t)^2 dt \right) \quad (15)$$

By Assumption 1,

$$\int_0^T \int_0^T K(s, t)^2 dt ds = \int_0^T \int_0^T \mathbb{E}[X_s X_t]^2 dt ds \quad (16)$$

$$\leq \int_0^T \int_0^T \mathbb{E}[X_s^2] \mathbb{E}[X_t^2] dt ds \quad (\text{by Cauchy-Schwarz}) \quad (17)$$

$$= \int_0^T \mathbb{E}[X_t^2] dt \int_0^T \mathbb{E}[X_t^2] dt = C^2 \quad (18)$$

□

- \mathcal{K} is non-negative, meaning for any $f \in L_T^2$, $\langle f, \mathcal{K}f \rangle \geq 0$.

Proof. We have

$$\langle f, \mathcal{K}f \rangle = \int_0^T \int_0^T f(s) \mathbb{E}[X_s X_t] f(t) ds dt \quad (19)$$

$$= \mathbb{E} \left[\left(\int_0^T f(t) X_t dt \right)^2 \right] \geq 0. \quad (20)$$

□

- \mathcal{K} is compact in the sense that the image of any bounded subset of L_T^2 has compact closure; equivalently, if $\{f_n\}$ is norm-bounded in L_T^2 , then $\{\mathcal{K}f_n\}$ has a norm-convergent subsequence. I won't prove this.

In summary, \mathcal{K} is a symmetric, bounded, compact, non-negative operator. There is a general result that tells us such operators behave like symmetric matrices. More specifically, the eigenvalues are real, the eigenvectors are orthogonal, \mathcal{K} has a pure point spectrum such that the only point where the eigenvalues can possibly accumulate is 0 (we can have $\lambda_k \rightarrow 0$ but not λ_k tending to any other value) and finally, that \mathcal{K} has an eigenfunction expansion

$$\mathcal{K}f = \sum_{k=1}^{\infty} \lambda_k \langle f, \phi_k \rangle \phi_k(t). \quad (21)$$

This should remind you of Eq. 8.

Main results: Mercer's and KL Theorem

A subtle but key point about Eq. 21 is that the sequence of operators \mathcal{K}_n defined by

$$\mathcal{K}_N f = \sum_{k=1}^N \lambda_k \langle f, \phi_k \rangle \phi_k(t) \quad (22)$$

converges to \mathcal{K} in L_T^2 . What this means is that

$$\|\mathcal{K}_N f - \mathcal{K}f\|^2 = \int_0^T ((\mathcal{K}_N f)(s) - (\mathcal{K}f)(s))^2 ds \rightarrow 0. \quad (23)$$

It turns out, for \mathcal{K} defined by Eq. 11, we can say something even stronger which rigorously justifies the infinite dimensional eigendecomposition of Σ given by Eq. 8. This is Mercer's Theorem.

Theorem 1 (Mercer (T5.12)). *In addition the properties stated above for any symmetric, bounded, compact, non-negative operator, K admits the uniformly and absolutely convergent expansion*

$$K(s, t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t).$$

This means that the kernels

$$K_N(s, t) = \sum_{k=1}^N \lambda_k \phi_k(s) \phi_k(t).$$

converge uniformly, i.e.

$$\sup_{(s,t) \in [0,T] \times [0,T]} |K(s, t) - K_N(s, t)| \rightarrow 0, \quad N \rightarrow \infty. \quad (24)$$

We now return to the question of how the process X_t itself can be represented in the series expansion. This is addressed by the following Theorem.

Theorem 2 (Karhunen–Loève (T5.13)). *Let $\{X_t\}_{t \in [0,T]}$ be a mean-zero GP satisfying Assumption 1. Let $\{\lambda_k\}$ and $\{\phi_k\}$ be the eigenvalues and eigenfunctions defined by Mercer’s Theorem. The process admits the expansion*

$$X_t = \sum_{k=1}^{\infty} \beta_k \phi_k(t), \quad \beta_k \sim \text{Normal}(0, \lambda_k) \quad (25)$$

and the series converges in the sense that for

$$X_t^N = \sum_{k=1}^N \beta_k \phi_k(t) \quad (26)$$

we have

$$\lim_{N \rightarrow \infty} \sup_{t \in [0,1]} \mathbb{E}[(X_t^N - X_t)^2] = 0 \quad (27)$$