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

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

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 $t_1, \ldots, t_n \in [0,T]$, the random vector*

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

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 $Y_1, Y_2, \ldots, Y_n$ which are assumed to be noisy measurements of some true function $f(t)$:

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

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,

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

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 $m \to \infty$ 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:

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

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

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

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

$$\Sigma = \Phi^T \Lambda \Phi \tag{8}$$

where $\Lambda = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots)$ 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^2_T

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

$$\mathcal{K} : L^2_T \to L^2_T, \tag{9}$$

where

$$L^2_T = \Big\{ f : [0,T] \to \mathbb{R} : \int_0^T f(t)^2 \, dt < \infty \Big\}. \tag{10}$$

This operator is given by

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

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

- H it is a vector space in the sense that the function can be combined and multiplied by scalers just like $\mathbb{R}^n$
- There is an inner product $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ just like the vector inner product. In particular $\langle f, g \rangle = \langle g, f \rangle$. In L^2_T,
$$\langle f, g \rangle = \int_0^T f(t) g(t) dt \tag{12}$$
- 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* $\{X_t\}_{t\in[0,T]}$

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

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

  Proof.

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

$$\le \int_0^T \int_0^T K(s,t)^2 \, dt ds \left(\int_0^T f(t)^2 \, dt \right) \quad \text{(by Cauchy-Schwarz)} \tag{15}$$

By Assumption 1,

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

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

$$= \int_0^T \mathbb{E}[X_t^2] \, dt \int_0^T \mathbb{E}[X_t^2] \, dt = C^2 \tag{18}$$

□

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

  Proof. We have

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

$$= \mathbb{E}\Big[\Big(\int_0^T f(t) X_t \, dt \Big)^2 \Big] \geq 0. \tag{20}$$

□

- $\mathcal{K}$ 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 $\lambda_k \to 0$ but λ_k may not tend to any other value) and finally, that $\mathcal{K}$ has an eigenfunction expansion

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

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

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

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

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

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

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

convergen uniformly, i.e.

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

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*

$$X_t = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k \, \phi_k(t), \quad \xi_k \underset{\text{iid}}{\sim} \text{Normal}(0,1) \tag{25}$$

and the series converges in the sense that for

$$X_t^N = \sum_{k=1}^{N} \sqrt{\lambda_k}\xi_k \, \phi_k(t) \tag{26}$$

we have

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

Remark 1. *The convergence in Eq. 27 can be understood as being in the space*

$$L_t^\infty L_\omega^2 \coloneqq \Big\{ X : \Omega \times [0,T] \to \mathbb{R} \ : \ \sup_{t\in[0,T]} \mathbb{E}\big[X_t^2\big] < \infty \Big\}, \tag{28}$$

which has the norm

$$\|X\|_{L_t^\infty L_\omega^2} = \sup_{t\in[0,T]} \mathbb{E}[X_t^2]^{1/2}. \tag{3}$$

In-fact, this is a complete space (a Banach space) and therefore convergent sequences converge to limits in this space.

Remark 2. *If the eigenfunctions in the KL expansion are continuous, then each partial sum is a continuous function of t. Under the convergence guaranteed by the KL theorem, X_t^N converges uniformly (along a subsequence) to X_t, so X_t admits a version with continuous sample paths.*

However, even if each ϕ_k is differentiable, it is generally not *true that X_t is differentiable. Differentiability would require convergence of the derivative series*

$$\sum_{k=1}^{\infty} \beta_k \phi_k'(t),$$

which typically fails. In this sense, differentiation amplifies the high–frequency components of the expansion, and Gaussian process sample paths are generically continuous but nowhere differentiable (we will see examples shortly).

Proof. By Mercer's theorem,

$$\sum_{k=1}^{N} \lambda_k \phi_k(s)\phi_k(t) \;\longrightarrow\; K(s,t) \quad \text{uniformly on } [0,T]\times[0,T].$$

Then for $N > M$,

$$\mathbb{E}[(X_t^N - X_t^M)^2] = \sum_{k=M+1}^{\infty} \lambda_k \phi_k(t)^2 \;\longrightarrow\; 0$$

uniformly in $t \in [0,T]$. Thus $\{X^N\}$ is Cauchy in $L_t^\infty L_\omega^2$, and we denote its limit by X_t.

For any $t_1, \ldots, t_m \in [0,T]$, the vector $(X_{t_1}^N, \ldots, X_{t_m}^N)$ is Gaussian with mean zero and covariance

$$\mathbb{E}[X_{t_i}^N X_{t_j}^N] = \sum_{k=1}^{N} \lambda_k \phi_k(t_i)\phi_k(t_j) \;\longrightarrow\; K(t_i,t_j).$$

Since $X^N \to X$ in $L_t^\infty L_\omega^2$, we have $(X_{t_1}^N, \ldots, X_{t_m}^N) \to (X_{t_1}, \ldots, X_{t_m})$ in mean square. Therefore $(X_{t_1}, \ldots, X_{t_m})$ is Gaussian with covariance $K(t_i,t_j)$, and X is a Gaussian process with covariance function K. □

The KL theorem says that $\beta_k \sim \text{Normal}(0,\lambda_k)$, or equivalently $\sigma_k^2 = \lambda_k$ in Eq. (6). This also says that even though a Gaussian process on $[0,T]$ is apparently a very complex object from a probabilistic perspective, the randomness is actually generated by a simple iid sequence of Normal random variables. To be precise, the underlying probability space is $\Omega = \mathbb{R}^{\mathbb{N}}$ and $\mathcal{F} = \sigma(\xi_k; k \in \mathbb{N})$ where the coordinate maps ξ_k are iid.

If we think about the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the probability distribution $\mathbb{P}$ is therefore the probability distribution induced by an infinite product of iid Gaussian measures, often written heuristically as

$$\prod_{k=1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\xi_k^2/2},$$

although in reality this infinite product density is not well-defined. What is well-defined is that every finite collection $(\xi_1, \ldots, \xi_n)$ has the usual n–dimensional standard Gaussian distribution.

Conditional distributions

In most applications of a GP, we are interested in the distribution of a GP conditional on some points (e.g. in interpolation we condition on the observations). Typically those observations are noisy but let's first consider conditioning on noiseless observations. That is, let $\{X_t\}_{t\in[0,T]}$ be a GP with kernel K and consider the distribution of

$$X_T | X_{t_1}, \ldots, X_{t_k} \tag{29}$$

where T is the time we want to make a prediction at. This is Normal and can simply be computed by usual Gaussian conditioning formula starting from the joint distribution of $(X_T, X_{t_1}, \ldots, X_{t_k})$. To describe this, we introduce some notation $\mathbf{t} = (t_1, \ldots, t_k)$ and use $K(\mathbf{t},\mathbf{t})$ to denote the matrix with

entries $K(\mathbf{t},\mathbf{t})_{i,j} = K(X_{t_i}, X_{t_j})$. Then write $\mathbf{k}(T,\mathbf{t}) = (K(T,t_1), K(T,t_2), \ldots, K(T,t_k))$. Finally let $\mathbf{X}(\mathbf{t}) = (X_{t_1}, \ldots, X_{t_k})^T$.

Now note that the joint distribution is

$$(X_T, X_{t_1}, \ldots, X_{t_k}) \sim \text{Normal}(0, \Sigma) \tag{30}$$

where

$$\Sigma = \begin{bmatrix} K(0,0) & \mathbf{k}(T,\mathbf{t})^T \\ \mathbf{k}(T,\mathbf{t}) & K(\mathbf{t},\mathbf{t}) \end{bmatrix} \tag{31}$$

The usual Gaussian conditioning formula (Exercise 1.8) then yields

$$X_T|\mathbf{X}(\mathbf{t}) \sim \text{Normal}(\mathbf{k}(T,\mathbf{t})^T K(\mathbf{t},\mathbf{t})^{-1}\mathbf{X}(\mathbf{t}), K(0,0) - \mathbf{k}(T,\mathbf{t})^T K(\mathbf{t},\mathbf{t})^{-1}\mathbf{k}(T,\mathbf{t})) \tag{32}$$

The exact same idea can be generalized to predict the joint distribution at multiple times. Essentially, conditioning on some points creates a new Gaussian process which is not mean zero and interpolates between the points we are conditioning on. This procedure is best visualized and you can find many nice visualizations online.

Examples

Exponential kernel (Ornstein–Uhlenbeck process)

Consider the process $\{X_t\}_{t\in\mathbb{R}}$ with kernel

$$K(t,s) = Ae^{-\gamma|t-s|}. \tag{33}$$

As the distance between points grows, the correlations decay, so we expect a process whose fluctuations are in some sense controlled. This is also an example of a *stationary kernel*, since the covariance depends only on the time difference: we can write $K(t,s) = \kappa(|t-s|)$ with $\kappa(h) = Ae^{-\gamma h}$. The Gaussian process generated by the exponential kernel is called the *Ornstein–Uhlenbeck process*. It plays a central role in the study of stochastic differential equations, since it is both a Gaussian process and the solution of a linear SDE.

The first important feature we discuss is that the process generated by this kernel is not differentiable, a consequence of the fact that the kernel is not differentiable at zero. Let us see what this implies for X_t. To see this, note that

$$\mathbb{E}\left[\left(\frac{X_{t+h}-X_t}{h}\right)^2\right] = \frac{1}{h^2}\,\mathbb{E}\big[(X_{t+h}-X_t)^2\big] \tag{34}$$

$$= \frac{1}{h^2}\Big(\mathbb{E}[X_{t+h}^2] + \mathbb{E}[X_t^2] - 2\,\mathbb{E}[X_{t+h}X_t]\Big) \tag{35}$$

$$= \frac{1}{h^2}\Big(2K(0) - 2K(h)\Big) = \frac{2A}{h^2}\big(1 - e^{-\gamma|h|}\big). \tag{36}$$

Using the expansion $1 - e^{-\gamma|h|} \sim \gamma|h|$ as $h \to 0$, we find

$$\mathbb{E}\left[\left(\frac{X_{t+h}-X_t}{h}\right)^2\right] \sim \frac{2A\gamma}{|h|} \xrightarrow[h\to 0]{} \infty. \tag{37}$$

Thus the mean–square derivative of X_t does not exist: the process is not differentiable in the mean–square sense. Hence the trajectories of X_t are very jagged, a generic feature of continuous Markov processes.

Now consider the conditional distributions, which can be understood just by considering (X_{t_1}, X_{t_2}, X_T) with $T > t_2 > t_1$. Let's consider the regression equation

$$X_T = aX_{t_1} + bX_{t_2} + \epsilon$$

for constants a, b where ϵ must be independent of X_{t_1} and X_{t_2}. Therefore, a and b are determined by the orthogonality conditions

$$\mathbb{E}[(X_T - aX_{t_1} - bX_{t_2})X_{t_i}] = 0, \qquad i = 1, 2.$$

Using the OU covariance, these equations become

$$e^{-\gamma(T-t_1)} = a + b\, e^{-\gamma(t_2-t_1)}, \tag{38}$$

$$e^{-\gamma(T-t_2)} = a\, e^{-\gamma(t_2-t_1)} + b. \tag{39}$$

Equivalently, in matrix form,

$$\begin{bmatrix} e^{-\gamma(T-t_1)} \\ e^{-\gamma(T-t_2)} \end{bmatrix} = \begin{bmatrix} 1 & e^{-\gamma(t_2-t_1)} \\ e^{-\gamma(t_2-t_1)} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{40}$$

Solving this linear system gives

$$a = 0, \qquad b = e^{-\gamma(T-t_2)}.$$

This tells us that the OU process is actually a Markov process (although we haven't defined this in general yet).

Squared exponential kernel (Gaussian kernel)

Consider now the process $\{X_t\}_{t\in\mathbb{R}}$ with the squared exponential (or Gaussian) kernel

$$K(t, s) = A \exp\left(-\frac{(t-s)^2}{2\ell^2}\right), \tag{41}$$

where $\ell > 0$ is a characteristic length scale. As in the exponential case, correlations decay as $|t - s|$ increases, but here the decay is much faster than exponential. This is again a *stationary kernel*, since the covariance depends only on the difference $t - s$.

A key difference from the exponential kernel is that $K(t, s)$ is smooth (in fact, real analytic) at $t = s$. This has strong consequences for the regularity of the sample paths. To see this, compute

$$\mathbb{E}\left[\left(\frac{X_{t+h} - X_t}{h}\right)^2\right] = \frac{1}{h^2}\Big(2K(0) - 2K(h)\Big) \tag{42}$$

$$= \frac{2A}{h^2}\left(1 - \exp\left(-\frac{h^2}{2\ell^2}\right)\right). \tag{43}$$

Using the expansion

$$1 - \exp\left(-\frac{h^2}{2\ell^2}\right) \sim \frac{h^2}{2\ell^2} \quad \text{as } h \to 0,$$

we find

$$\mathbb{E}\left[\left(\frac{X_{t+h} - X_t}{h}\right)^2\right] \longrightarrow \frac{A}{\ell^2}, \qquad h \to 0. \tag{44}$$

Thus the mean–square derivative of X_t exists. In fact, one can show that sample paths of a GP with squared exponential kernel are almost surely infinitely differentiable. This should be contrasted with the exponential kernel, where the lack of differentiability of K at the origin leads to very rough sample paths.

Finally, consider conditional distributions. As before, fix $T > t_2 > t_1$ and consider the regression

$$X_T = aX_{t_1} + bX_{t_2} + \epsilon,$$

with ϵ independent of (X_{t_1}, X_{t_2}). The coefficients are again determined by the orthogonality conditions

$$\mathbb{E}[(X_T - aX_{t_1} - bX_{t_2})X_{t_i}] = 0, \qquad i = 1, 2.$$

Using the squared exponential covariance, these become

$$\exp\left(-\frac{(T-t_1)^2}{2\ell^2}\right) = a + b\,\exp\left(-\frac{(t_2-t_1)^2}{2\ell^2}\right), \tag{45}$$

$$\exp\left(-\frac{(T-t_2)^2}{2\ell^2}\right) = a\,\exp\left(-\frac{(t_2-t_1)^2}{2\ell^2}\right) + b. \tag{46}$$

In general, the solution of this system has both $a \neq 0$ and $b \neq 0$. In particular, conditioning on X_{t_2} alone is not sufficient to determine the conditional law of X_T.

This shows that, unlike the Ornstein–Uhlenbeck process, the Gaussian process with squared exponential kernel is *not* Markov: information from earlier times cannot be discarded without loss. The price paid for smoothness of the sample paths is long–range memory in time.

Brownian motion kernel

Consider the process $\{X_t\}_{t\in[0,T]}$ with kernel

$$K(t,s) = t \wedge s = \min(t,s). \tag{47}$$

For $h \neq 0$, using $K(t+h, t+h) = t+h$, $K(t,t) = t$, and $K(t+h,t) = t$, we have

$$\mathbb{E}\left[\left(\frac{X_{t+h} - X_t}{h}\right)^2\right] = \frac{1}{h^2}\,\mathbb{E}\big[(X_{t+h} - X_t)^2\big] \tag{48}$$

$$= \frac{1}{h^2}\Big(K(t+h,t+h) + K(t,t) - 2K(t+h,t)\Big) \tag{49}$$

$$= \frac{1}{h^2}\big((t+h) + t - 2t\big) = \frac{|h|}{h^2} = \frac{1}{|h|}. \tag{50}$$

Thus

$$\mathbb{E}\left[\left(\frac{X_{t+h} - X_t}{h}\right)^2\right] = \frac{1}{|h|} \xrightarrow[h\to 0]{} \infty,$$

so the mean–square derivative of X_t does not exist, and sample paths are very jagged (as for the OU process).

Fix $T > t_2 > t_1 \geq 0$ and consider the regression

$$X_T = aX_{t_1} + bX_{t_2} + \varepsilon,$$

where ε is independent of (X_{t_1}, X_{t_2}). As before, the coefficients a, b are determined by the orthogonality conditions

$$\mathbb{E}[(X_T - aX_{t_1} - bX_{t_2})X_{t_i}] = 0, \qquad i = 1, 2.$$

Using the Brownian covariance $K(s,t) = s \wedge t$, we have

$$\mathrm{Var}(X_{t_1}) = t_1, \quad \mathrm{Var}(X_{t_2}) = t_2, \quad \mathrm{Cov}(X_{t_1}, X_{t_2}) = t_1, \quad \mathrm{Cov}(X_{t_1}, X_T) = t_1, \quad \mathrm{Cov}(X_{t_2}, X_T) = t_2.$$

The orthogonality equations become

$$t_1 = a\, t_1 + b\, t_1, \tag{51}$$

$$t_2 = a\, t_1 + b\, t_2. \tag{52}$$

Equivalently,

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} t_1 & t_1 \\ t_1 & t_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{53}$$

From the first equation $t_1 = t_1(a + b)$, and since $t_1 > 0$ this gives $a + b = 1$. Substituting into the second equation,

$$t_2 = at_1 + bt_2 = (1 - b)t_1 + bt_2,$$

so

$$t_2 - bt_2 = t_1 - bt_1 \implies (1 - b)(t_2 - t_1) = 0.$$

As $t_2 \neq t_1$, we conclude $1 - b = 0$, hence $b = 1$ and $a = 0$.