IIIS 2023 Spring: ATCS - Selected Topics in Learning, prediction and Optimization
Lecture date: 20th, February, 2023
Instructor: Jian Li **Scribe: Jiawei Xu**

1 Gaussian process

Definition 1 *A set of random variables $\{X_t\}_{t\in T}$ is called a Gaussian process (GP) if for any finite subset $\{t_1, t_2, \cdots, t_k\}$, $\{X_{t_1}, X_{t_2}, \cdots, X_{t_k}\}$ follows a jointly Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^k, \Sigma \in \mathbb{R}^{k\times k}$. Note that $|T|$ may be infinite and T may have its own structure, e.g., $T = \mathbb{R}$.*

A Gaussian Process can be described by its expectation function $\mu(t) = \mathbb{E}[X_t], t \in T$, and the covariance function (kernel function) $k(\cdot,\cdot)$ with $k(s,t) = \mathrm{Cov}(X_s, X_t)$, where $s, t \in T$. Here $k(.,.)$ should be a positive semidefinite (psd) function (i.e., for any finite subset $\{t_1, t_2, \cdots, t_k\}$, the matrix $\{k(t_i, t_j)\}_{i,j\in[k]}$ is a psd matrix).

Here are some examples of GPs with different kernel functions (depicted in Fig. 1).

1. The Ornstein-Uhlenbeck (OU) kernel: $k(x, x') = \exp\left(-\frac{|x-x'|}{l}\right)$.
2. The min kernel: $k(x, x') = \min(x, x')$ (this is the covariance function for Brownian motion).
3. The linear kernel: $k(x, x') = x^\top x'$.
4. The quadratic kernel: $k(x, x') = \left(x^\top x' + c\right)^2$.
5. The Radial Basis Function (RBF) kernel: $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right)$.
6. The periodic kernel (Exp-Sine-Squared): $k(x, x') = \sigma^2 \exp\left(\frac{-\sin^2(\|x-x'\|/p)}{l^2}\right)$.

2 Basics of Wiener process

Wiener process is named in honor of American mathematician Norbert Wiener for his investigations on the mathematical properties of the one-dimensional Brownian motion. In this lecture, the names Wiener process and Brownian motion are used interchangeably. Typically, people use $\{W_t\}_{t\geq 0}$ or $\{B_t\}_{t\geq 0}$ to denote the Wiener process.

Definition 2 $\{W_t\}_{t\geq 0}$ *is a Wiener process if*

1. $W_0 = 0$
2. *W_t is continuous in t a.s..*
3. *For any $s > t \geq 0$ we have $W_s - W_t \sim \mathcal{N}(0, s-t)$ and $W_s - W_t \perp W_t$. Here $X \perp Y$ means that X and Y are independent.*

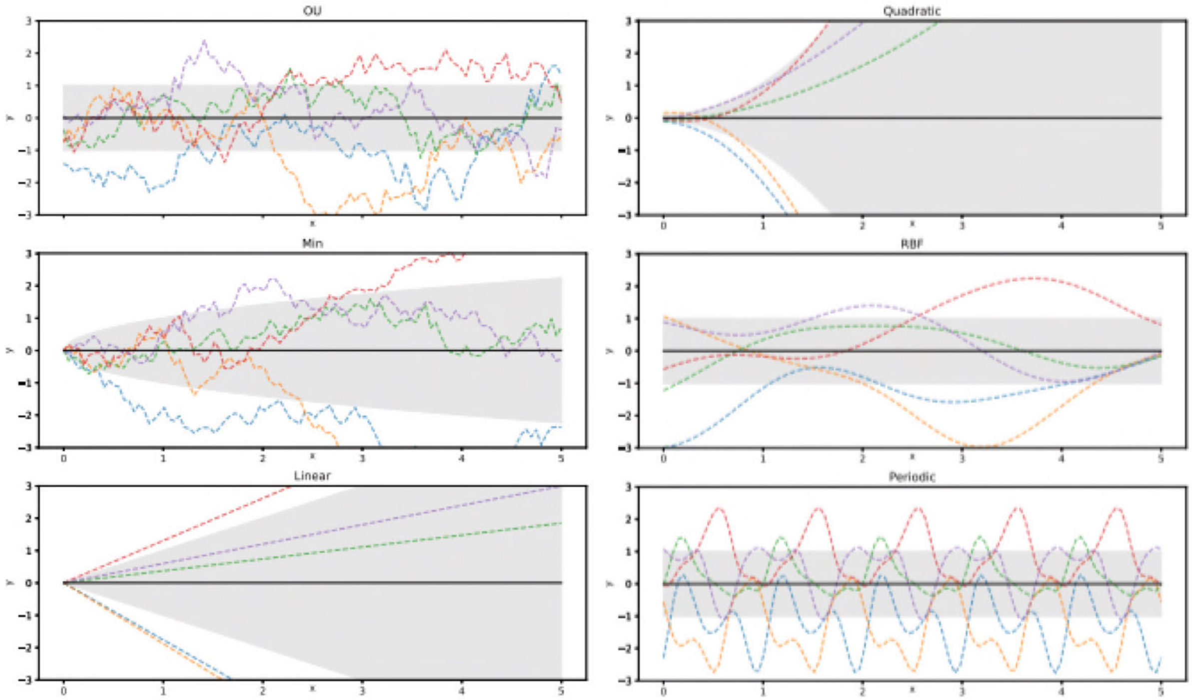

Figure 1: Examples of GPs with different kernels.

Remark: The above definition is not the most basic definition of Wiener process. If we want to define Wiener process completely rigorously, we need measure theory and it is nontrivial to show the existence of a process with the above properties. There are many excellent textbooks on this topic (see e.g., [4]).

We will study calculus involving Wiener process. There are two types of integrals: one is like $\int_0^T W_t dt$ (W_t appears in the integrand) and the other is like $\int_0^T W_t dW_t$ (W_t appears in the differential). The later is called *Ito's integral* and we will study it later. The former is in fact just the Riemann integral. The only difference is that the integrand is a random variable and hence the integral is also a random variable.

Example 3 *The integral $Y := \int_0^T W_t dt$ is also a r.v. following the Gaussian distribution. In fact we have $Y \sim \mathcal{N}(0, \frac{1}{3}T^3)$.*

Proof: By definition of Riemann integrals, we have

$$Y = \int_0^T W_t dt = \lim_{\Delta t \to 0} \sum_{k=0}^{n-1} \Delta t_k W_{t_k}$$

where $0 = t_0 < t_1 < \cdots < t_n = T$, $\Delta t_k = t_{k+1} - t_k$ and $\Delta t = \max(\Delta t_k)$. Hence, it is easy to see that $\mathbb{E}[Y] = 0$ since it is a sum of zero-mean Gaussians. Now, we analyze its variance:

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}\left[\lim_{\Delta t \to 0}\left(\left(\Delta t \sum_{j=0}^{n-1} W_{t_j}\right)\left(\Delta t \sum_{k=0}^{n-1} W_{t_k}\right)\right)\right] \\
&= \lim_{\Delta t \to 0}\left((\Delta t)^2 \sum_{j,k}^{n-1} \mathbb{E}[W_{t_j} W_{t_k}]\right) \\
&= \lim_{\Delta t \to 0}\left((\Delta t)^2 \sum_{j,k}^{n-1} \min(t_j, t_k)\right) \\
&= \int_0^T \int_0^T \min(s,t)\mathrm{d}t\mathrm{d}s = \frac{1}{3}T^3.
\end{aligned}$$

The third equation uses the covariance formula for Wiener process: $\mathbb{E}[W_t W_s] = \min(s,t)$. □

Proposition 4 *A few properties of Wiener processes:*

1. *(Quadratic variation) Let* $0 = t_0 < t_1 < \cdots < t_N = t$, $\Delta t = \max_j(t_j - t_{j-1})$, *and* $\Delta W_j = W_{t_j} - W_{t_{j-1}} \sim (0, t_j - t_{j-1})$ [1]. *Then, with probability 1, the quadratic variation*

$$\mathsf{QV}(W) = \lim_{\Delta t \to 0} \sum_{j=1}^{N} (\Delta W_j)^2 = T.$$

   Note that for a differentiable function, the quadratic variation is 0.

2. *The Wiener process is not differentiable.*

In fact, the first item can be seen easily using the law of large number. The second is a corollary of the first: indeed, if a function f is differentiable, its quadratic variation is 0. This can be easily seen as follows: suppose f' is the derivative of f.

$$\mathsf{QV}(f) = \lim_{\Delta t \to 0} \sum_{j=1}^{N} (\Delta f_j)^2 \leq \max_j |\Delta f_j| \lim_{\Delta t \to 0} \sum_{j=1}^{N} |\Delta f| \leq \max(f')^2 \lim_{\Delta t \to 0} \Delta t \sum_{j=1}^{N} |\Delta t_j| = 0.$$

3 Itô Integral

Next, let us study the Itô integral of a Wiener process. In particular, we would like to provide a rigorous definition for notions like $\int_0^T f(t, W_t)\mathrm{d}W_t$. However, since W_t is a stochastic process, we need to be really careful about what f we can integrate. The following definition is very important.

[1] From this we have $\mathbb{E}[(\Delta W_j)^2] = t_j - t_{j-1}$, and intuitively we'll have $(dW_t)^2 = dt, \int (dW_t)^2 = T$, which might be helpful in understanding the Itô integral afterwards.

Definition 5 (Non-anticipativity) *A function $f(t)$ is non-anticipative w.r.t. a Wiener process $\{W_t\}$ if $f(t) = f(\{W_s\}_{s \le t}, t)$ for some function f, i.e., the value of $f(t)$ does not depend on W_s for $s > t$.*[2]

Definition 6 (Itô integral) *Suppose f is non-anticipative w.r.t. a Wiener process $\{W_t\}$ such that $\mathbb{E}\left[\int_0^T |f(t)|^2 dt\right] < \infty$. Then, the Itô integral is defined as*

$$\int_0^T f(t)dW_t = \lim_{\Delta t \to 0} \sum_{j=1}^{n} f(t_{j-1})(W(t_j) - W(t_{j-1})).^3$$

Note that the integral may not follow a Gaussian distribution, unless f is independent of W_t. Also, note that the non-anticipativity ensures that $f(t_{j-1})$ and $W(t_j) - W(t_{j-1})$ are independent. Without non-anticipativity, the limit may not be well defined, as shown in the following example (Example 9). Specifically, we present a sequence of functions f_k (which can adapt to the future change of W_t), such that $\lim_{n\to\infty} f_k \to 0$ uniformly over $[0, 1]$, but $\sum_{j=1}^{n} f_k(t_{j-1})(W(t_j) - W(t_{j-1}))$ diverges to $+\infty$ as $k \to +\infty$.

Definition 7 *The total variation of a function $g : [0, 1] \to \mathbb{R}$ is defined as the following limit:*

$$\mathsf{TV}(g) = \lim_{\Delta t \to 0} \sum_{j=0}^{n-1} |g(t_{j+1}) - g(t_j)|.$$

Exercise 8 *It is an easy exercise to show that*

$$\mathsf{TV}(W_t) = +\infty.$$

Example 9 *Consider a function $w(t) : [0, 1] \to \mathbb{R}$ such that $\mathsf{TV}(w) = \infty$. There exists a sequence of piece-wise constant functions $\{f_k\}_{k\in\mathbb{N}}$ such that $\{f_k\}_{k\in\mathbb{N}}$ converges uniformly to 0 but $I(f_k) = \sum_{j=1}^{n} f_k(t_{j-1})(w(t_j) - w(t_{j-1}))$ diverges.*

Proof: By definition, there exist partitions $\{\Pi_n = (t_0, \dots, t_n)\}_{n\in\mathbb{N}^+}$ of $[0, 1]$ s.t. the variation of g on Π_n diverges. Let $h_k(t_i) = \mathrm{sgn}(g(t_{i+1}) - g(t_i))$. Obviously $I(h_k) \to \infty$ as $k \to \infty$. Let $f_k = I(h_k)^{-1/2} h_k$, which converges to 0 uniformly as $k \to \infty$. But we have $I(f_k) = I(h_k)^{1/2} \to \infty$. $\square$

Since the total variation of a Wiener process is obviously infinity, this counter-example illustrates the reason why we need non-anticipativity in Definition 6.

Definition 10 *(Martingale) A martingale is a stochastic process for which, at a particular time, the conditional expectation of the next value in the sequence is equal to the present value, regardless of all prior values. For discrete time stochastic process $\{Y_1, Y_2, \dots\}$ such that $\mathbb{E}(|Y_n|) < \infty$, a martingale means the following holds:*

$$\mathbf{E}(Y_{n+1} \mid Y_1, \dots, Y_n) = Y_n.$$

[2] For example, $f(t) := W_t$ is non-anticipative (w.r.t. $\{W_t\}$); $f(t) := \max_{t\le s\le T} W_s$ is not non-anticipative; $f(t) := \begin{cases} 0 & \text{if } \max_{t\le s\le T} f(s,) \\ 1 & \text{otherwise.} \end{cases}$ is not non-anticipative.

[3] Here we use $f(t_{j-1})$ instead of $f(t_j)$ to make sure that $W(t_j) - W(t_{j-1})$ is independent with $f(t_{j-1})$.

For a continuous time stochastic process $\{Y_t\}_t$ such that $\mathbb{E}(|Y_t|) < \infty$, a martingale means the following holds: for any $t > s$,

$$\mathbf{E}(Y_t \mid \mathcal{F}_s) = Y_s.$$

Here $\mathcal{F}_s$ is the filtration up to time s. If you are not family with the language of filtration, you can understand $\mathcal{F}_s$ as the history up to time s.

Exercise 11 *It is an easy exercise to show that $\{W_t\}_t$ and the Ito integral $\int_0^T f(t)dW_t$ are both martingales.*

Remark 12 *There are also two different integral similar to Itô integral that you may find in the literature. They are useful in certain technical context. We do not cover them in detail. Here we use the same notations in Definition 6.*

1. *(Backward Itô integral) We define*

$$\int_0^T f(t)\,dW_t = \lim_{\Delta t \to 0} \sum_{j=1}^{n} f(t_j)(W(t_j) - W(t_{j-1})).$$

Note that in the Backward Itô integral we change t_{j-1} to t_j. This integral is particularly useful when one to reverse the time of the process.

2. *(Fisk-Stratonovich integral) We define*

$$\int_0^T f(t)\,dW_t = \lim_{\Delta t \to 0} \sum_{j=1}^{n} f(s_j)(W(t_j) - W(t_{j-1}))$$

where $s_j = \frac{t_{j-1}+t_j}{2}$. A nice feature of Fisk-Stratonovich integral is that we can still use the ordinary chain rule (as in ordinary calculus). Sometimes this is useful.

However, both the above integrals have the problem that the two product terms at right-hand-side are non-independent. So the expectations of them may not be 0, hence they are typically not martingale. So in most cases, we use Ito's integral, unless specified explicitly.

4 Itô Process and Itô Calculus

Now, let us consider how to compute an Itô integral. We first need to define the concept of the Itô process.

Definition 13 (Itô process) *A stochastic process $X(t)$ is called an Itô process if it satisfies the following*

$$X(t) = X(0) + \int_0^t \delta(s)\,dW_s + \int_0^t b(s)\,ds$$

where W_t is a Wiener process. This is often written in the differential form for short:

$$dX(t) = \delta(t)\,dW_t + b(t)\,dt.$$

Very often the differential form is given. But we should keep in mind that its rigorous meaning is the integral equation. This definition illustrates that an infinitesimal increment of $X(t)$ follows $\mathcal{N}(b(t), \delta^2(t))$, but $X(t)$ may not be a Gaussian random variable. Here $\delta(t)\mathrm{d}W_t$ is called the diffusion term, and $b(t)\mathrm{d}t$ is called the drift term.

It is important to remember that we **cannot** do the normal differentiate process in Itô calculus, that is, assume that F is the primitive function of f ($F' = f$), then we cannot have $dF(W_t) = f(W_t)dW_t$ or $\int_0^t f(W_t)dW_t = F(W_t) - F(W_0)$. Consider the following illuminating example.

Example 14 *For example, for a given $n \in \mathbb{N}^+, T \geq 0$ we denote $\mathbf{W}_j = W_{\frac{jT}{n}}$ for $j \in [n]$. If we let $F(x) = \frac{x^2}{2}, f(x) = x$, then we cannot have $\int_0^T f(W_t)dW_t = \frac{W_T^2}{2}$ because clearly the expectation of left-hand-side is 0, but right-hand-side isn't. In fact, by straight calculations we have*

$$
\begin{aligned}
\int_0^T f(W_t)dW_t &= \lim_{n\to\infty} \sum_{j=1}^n \mathbf{W}_{j-1} d\mathbf{W}_j = \lim_{n\to\infty} \sum_{j=1}^n \mathbf{W}_{j-1}(\mathbf{W}_j - \mathbf{W}_{j-1}) \\
&= \lim_{n\to\infty} \left(-\sum_{j=1}^n (\mathbf{W}_j - \mathbf{W}_{j-1})^2 + \sum_{j=1}^n \mathbf{W}_j(\mathbf{W}_j - \mathbf{W}_{j-1}) \right) \\
&= \lim_{n\to\infty} \left(-\sum_{j=1}^n (\mathbf{W}_j - \mathbf{W}_{j-1})^2 + \sum_{j=1}^n (\mathbf{W}_j^2 - \mathbf{W}_{j-1}^2) - \sum_{j=1}^n \mathbf{W}_{j-1}(\mathbf{W}_j - \mathbf{W}_{j-1}) \right) \\
&= -T + W_T^2 - \int_0^T W_t dW_t.
\end{aligned}
$$

Therefore, we have $\int_0^T f(W_t)dW_t = \frac{W_T^2}{2} - \frac{T}{2}$ (this is the correct answer). Intuitively, the additional term $-T/2$ comes from the quadratic variation $\sum_j (d\mathbf{W}_j)^2$.

To correctly calculate the differentiate process in Itô calculus, we need to introduce the Itô's lemma below.

Lemma 15 (Itô's lemma) *Assume that $X(t)$ is an Itô process, $F(X(t), t)$ is a smooth function, then $Y(t) := F(X(t), t)$ is also an Itô process. It has the form $\Sigma(t)dW_t + B(t)dt$ with*

$$
\begin{aligned}
\Sigma(t) &= \frac{\partial F(X(t), t)}{\partial X}\delta(t), \\
B(t) &= \frac{\partial F(X(t), t)}{\partial t} + \frac{1}{2}\delta^2(t)\frac{\partial^2 F(X(t), t)}{\partial X^2} + b(t)\frac{\partial F(X(t), t)}{\partial X}.
\end{aligned}
$$

Proof Sketch. We only provide a high level proof sketch. As we will see, Itô's lemma is nothing but the Taylor expansion up to 2nd order terms. In particular, assume we perturb t to t' with difference $\Delta t = t - t'$. Let $\Delta W = W_t - W_{t'}$ and $\Delta X = X(t) - X(t')$. Since F is smooth (in fact $F \in C^2$ is enough), by Taylor expansion we have

$$
F(X(t), t) - F(X(t'), t') = \frac{\partial F}{\partial t}\Delta t + \frac{\partial F}{\partial X}\Delta X + \frac{1}{2}\frac{\partial^2 F}{\partial X^2}(\Delta X)^2 + \cdots
$$

We then substitute ΔX by $\delta(t)\Delta W + b(t)\Delta t$, and we see that the terms involving $\Delta W \Delta t, (\Delta t)^2$ or larger order terms will vanish when $\Delta t \to 0$ (after the whole expression is divided by Δt). [1] The only interesting term that remains is $(\Delta W_t)^2$. By quadratic variation of Wiener process, we know that the limit of $\sum(\Delta W_t)^2$ (or $\int_0^T (\mathrm{d}W_t)^2$) is T. Hence, $(\Delta W_t)^2$ can be replaced by Δt (or $(\mathrm{d}W_t)^2$ is replaced by $\mathrm{d}t$). Adding the remaining terms will obtain the Itô's lemma. □

If you do not want to memorize Itô lemma, you can simply use the following formula.

$$\mathrm{d}Y_t = \frac{\partial F}{\partial t}\mathrm{d}t + \frac{\partial F}{\partial X}\mathrm{d}X_t + \frac{1}{2}\frac{\partial^2 F}{\partial X^2}(\mathrm{d}X_t)^2.$$

Then, we replace $\mathrm{d}X_t$ by $\delta(t)\mathrm{d}W_t + b(t)\mathrm{d}t$. For the product of the differentials, we use the following Itô's table. Basically, when we see $(\mathrm{d}W_t)^2$, we can replace it by $\mathrm{d}t$. In other cases, the product is simply zero.

Itô's table		
	$\mathrm{d}W_t$	$\mathrm{d}t$
$\mathrm{d}W_t$	$\mathrm{d}t$	0
$\mathrm{d}t$	0	0

Itô process can be also extended to the multi-dimensional setting.

Definition 16 *Let $\{W_t\}_{t\geq 0}$ be an m-dim Wiener process. $\{X_t\}_t$ is an n-dim Itô process if*

$$X_t = X_0 + \int_0^t F(s)\,ds + \int_0^t G(s)\,dW_s,$$

where $G(s) \in \mathbb{R}^{n\times m}$, F and G are non-anticipative w.r.t. $\{W_t\}_{t\geq 0}$. This is also written as

$$dX(t) = F(t)\,dt + G(t)\,dW_t.$$

Theorem 17 (Itô's rule for multi-dim processes) *Let $\{X_t\}$ be an m-dim Itô process, $F(X_t, t)$ be a smooth, twice differentiable function w.r.t. (X, t). We have*

$$dF(X_t, t) = \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial x}dX_t + \frac{1}{2}dX_t^\top \frac{\partial^2 F}{\partial x^2}dX_t.$$ [1]

Example 18 *We are given two Ito processes $dX_t^1 = a(t)dt + b(t)dW_t$, $dX_t^2 = c(t)dt + e(t)dW_t$. Compute $d(X_t^1 X_t^2)$ and present it in the form of Ito process. Note that the corresponding Hessian matrix of $X_t^1 X_t^2$ is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Using Ito's rule for multidimensional case, we have*

$$\begin{aligned} d(X_t^1 X_t^2) &= X_t^2(a(t)\,dt + b(t)\,dW_t) + X_t^1(c(t)\,dt + e(t)\,dW_t) \\ &\quad + (a(t)\,dt + b(t)\,dW_t)\cdot(c(t)\,dt + e(t)\,dW_t) \\ &= X_t^2(a(t)\,dt + b(t)\,dW_t) + X_t^1(c(t)\,dt + e(t)\,dW_t) + b(t)e(t)\,dt \\ &= (c(t)X_t^1 + a(t)X_t^2 + b(t)e(t))\,dt + (e(t)X_t^1 + b(t)X_t^2)\,dW_t. \end{aligned}$$

[1] In fact we have $\mathbb{E}[|\Delta W|^p(\Delta t)^q] \propto (\Delta t)^{\frac{p}{2}+q}$ for $p, q \in \mathbb{N}$.

[1] Note that $\frac{\partial F}{\partial t}, \mathrm{d}X_t$ are vectors, and $\frac{\partial F}{\partial x}, \frac{\partial^2 F}{\partial x^2}$ are matrices.

References

[1] Wikipedia contributors. (2023, January 9). Gaussian process. In Wikipedia, The Free Encyclopedia. Retrieved 08:53, February 23, 2023

[2] Wikipedia contributors. (2023, February 19). Wiener process. In Wikipedia, The Free Encyclopedia. Retrieved 08:54, February 23, 2023

[3] Wikipedia contributors. (2023, February 16). Itô's lemma. In Wikipedia, The Free Encyclopedia. Retrieved 08:55, February 23, 2023

[4] Van Handel, Ramon. "Stochastic calculus, filtering, and stochastic control." Course notes., URL http://www. princeton. edu/rvan/acm217/ACM217. pdf 14 (2007).