CS 886 Winter 2024:
Lecture Notes

Lecture notes taken, unless otherwise specified, by myself during the Winter 2024 offering of CS 886, taught by Yaoliang Yu.

I am not a graduate student. This is going to be a disaster.

Lectures

Chapter 1

Introduction

Lecture 1
Jan 9

A diffusion model is basically a limit of infinite auto-regressive models. We can construct this $\mathbf{x}_{t+1} \approx \mathbf{x}_t + \eta_t \cdot \mathbf{f}_t(\mathbf{x}_t)$ as an ODE and take the limit:

$$\mathrm{d}\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t)$$

Make a stochastic differential equation: add a perturbation $G_t(\mathbf{x}_t)$.

$$\mathrm{d}\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t)\,\mathrm{d}t + G_t(\mathbf{x}_t)\,\mathrm{d}\mathsf{B}_t$$

Clearly, an ODE is an SDE with $G_t \equiv \mathbf{0}$. But an SDE can be integrated to recover an ODE using the score function.

Given an SDE, we can reverse it. That is, if the forward SDE goes from data to noise, a backwards SDE can generate data from noise. The key is estimating that score function.

If we have some normalized distribution $p(x) = \frac{\exp(E(x))}{\int \exp(E(x))}$ for some energy function E, then $\log p(x) = E(x) - \log \int \cdots$ which gives us an ignorable integral constant. Then, we can define a loss function.

This is all very vague.

1.1 Probability basics

Definition 1.1.1 (random variable)

Fix a sample space Ω equipped with a σ-algebra $\mathcal{F}$ where $\mu : \mathcal{F} \to [0,1]$ assigns probability.

A random variable is a function $\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{S}, \mathcal{B}, \mathsf{X}_{\#}\mu)$ to the state space $\mathbb{S}$ (in this course, always $\mathbb{R}$).

The distribution of X, notated $\mathsf{X}_{\#}\mu$, is a probability measure on $\mathcal{B} \subseteq 2^{\mathbb{S}}$, i.e.,

$$(\mathsf{X}_{\#}\mu)(S) := \mu(\{\omega : \mathsf{X}(\omega) \in S\}) = \mu(\mathsf{X}^{-1}(S))$$

so long as $\mathsf{X}^{-1}(\mathcal{B}) \subseteq \mathcal{F}$. If $\omega \simeq \mu$, then $\mathsf{X}(\omega) \simeq \mathsf{X}_{\#}\mu$.

Example 1.1.2. Say $\mathsf{X} \simeq \mathcal{N}(0,1)$.

That formally means $\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$ and $\mathsf{X}_{\#}\mu = \mathcal{N}(0,1)$.

Example 1.1.3. Say $\mathsf{Y} \simeq \chi^2_1$.

That formally means $\mathsf{Y} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$ and $\mathsf{Y}_{\#}\mu = \chi^2_1$.

Consider $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ where $x \mapsto x^2$. Compose $f(\mathsf{X}) : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$. Then, $f(\mathsf{X}) \simeq \chi^2_1$ by definition of the χ^2 distribution.

Observe that the distribution of $f(\mathsf{X})$ is $(f \circ \mathsf{X})_{\#}\mu = f_{\#}[\mathsf{X}_{\#}\mu]$.

We want to go the other direction, to solve the inverse problem going from distributions to functions.

Problem 1.1.4

Given distributions P and Q, find f such that $f_{\#}P = Q$.

In a generative model, $P = \mathcal{N}(\mathbf{0}, I)$ is noise, and Q is the data distribution. We want to find f such that if we draw $\mathsf{X} \simeq P$, then we can apply $f(\mathsf{X}) \simeq Q$.

Theorem 1.1.5 (Representation through Push-forward)

Let P be any continuous distribution on $\mathbb{R}^m$. For any distribution Q on $\mathbb{R}^d$, there exist push-forward maps $f : \mathbb{R}^m \to \mathbb{R}^d$ such that $\mathsf{Z} \simeq P \implies f(\mathsf{Z}) \simeq Q$ (equivalently $f_{\#}P = Q$).

We need P to be continuous so that we can send it to anything.

In practice, we never have a continuous Q, since our data is discrete. Instead, we only have an approximated empirical $\hat{Q}$.

Example 1.1.6. Let $X \simeq \mathcal{N}(0,1)$ and $Y \simeq \chi^2_1$.

Solution. By definition, $f(x) = x^2$ works.

But we can also consider the CDF Φ of $\mathcal{N}(0,1)$. If we apply $\Phi(\mathsf{X})$:

$$\Pr[\Phi(\mathsf{X}) \leq u] = \Pr[\mathsf{X} \leq \Phi^{-1}(u)] = \Phi(\Phi^{-1}(u)) = u$$

we get a uniform distribution.

Then, apply the inverse CDF Ψ^{-1} of χ^2:

$$\Pr[\Psi^{-1}(\Phi(\mathsf{X})) \leq t] = \Pr[\Phi(\mathsf{X}) \leq \Psi(t)] = \Psi(t)$$

since $\Psi(\mathsf{X})$ is uniform, which means that $\Psi^{-1}(\Phi(\mathsf{X}))$ has CDF Ψ.

That is, $f = \Psi^{-1} \circ \Phi$ works as well. In fact, we know this is a distinct solution, since this function is always increasing, but x^2 is not. □

Remark 1.1.7. If we add the condition that f is monotonically increasing, the only f is $\Psi^{-1} \circ \Phi$. We call this the optimal transformation.

We can consider $\mathsf{X} \xrightarrow{g} \mu$ as a composition of functions $\mathsf{X} \xrightarrow{g_1} \mathsf{X}_1 \xrightarrow{g_2} \mathsf{X}_2 \xrightarrow{g_3} \cdots \xrightarrow{g_n} \mathsf{X}_n \approx \mu$.

Then, to go backwards, $Y \xleftarrow{h} \mu$ ...

This does not work because [reason]

Definition 1.1.8 (stochastic process)

Consider "time" $t \in \mathbb{T}$. Equivalently:

- A collection of random variables $\mathsf{X} : \mathbb{T} \to \mathbb{R}^{\Omega} : t \mapsto \mathsf{X}(t, \cdot)$
- A random function to paths $\mathsf{X} : \Omega \to \mathbb{R}^{\mathbb{T}} : \omega \mapsto \mathsf{X}(\cdot, \omega)$
- A bivariate function $\mathsf{X}(t, \omega) : \mathbb{T} \times \Omega \to \mathbb{R}$

depending on if we fix the time or state.

We can equivalently write $\mathsf{X}(t)$, $\mathsf{X}(t, \omega)$, or X_t.

1.2 Brownian motion

Definition 1.2.1 (Brownian motion)

A stochastic process $\{\mathsf{B}_t : t \geq 0\}$ such that:

- $\mathsf{B}_0 \equiv 0$
- Each increment $B_{t_1} - B_{t_0}$ is independent of all others
- All increments $\mathsf{B}_t - \mathsf{B}_s \simeq \mathsf{B}_{t-s}$
- $\mathbf{B}_t \simeq \mathcal{N}(0, t)$
- The function $t \mapsto \mathsf{B}_t(\omega)$ is continuous for all ω

Brownian motion is a Gaussian process with covariance kernel $\kappa(s, t) := \mathbb{E}[\mathsf{B}_s \mathsf{B}_t] = s \wedge t$ (i.e., the minimum of s and t). We can simulate this (discretely) now, since we can simulate a Gaussian process.

Side note: "white noise" is typically defined as the derivative B'_t of a Brownian motion, since $\kappa' := \partial_{12} \kappa$ is also a kernel.

However, we still need the continuity condition. Luckily, we cheat and cite some theorem:

Theorem 1.2.2 (Continuity condition of stochastic processes)
Let X_t be a stochastic process with index $t \in \mathbb{R}^m$. If for some $\alpha, \beta, \mathsf{L} > 0$,

$$\mathbb{E}[\|\mathsf{X}_s - \mathsf{X}_t\|^\alpha] \leq \mathsf{L}\|t-s\|^{m+\beta}$$

for all s and t, then there exists a modification $\tilde{\mathsf{X}}_t$ that is locally Hölder continuous of order $\gamma < \beta/\alpha$.

To be Hölder continuous at s of order γ means that for all t around s, $\|\mathsf{X}_s - \mathsf{X}_t\| \leq c \cdot \|s-t\|^\gamma$.

We can show this condition holds for Brownian motion.

Kolmogorov's construction For finitely many $t_1, \ldots, t_n$, define $\mathsf{B}_{1:n} := (B_{t_i}) \simeq \mathcal{N}(\mathbf{0}, K_n)$ where $K_n(t_i, t_j) = t_i \wedge t_j$. Then, by the Kolmogorov extension theorem, a continuous process B_t exists that satisfies all conditions except continuity.

We know that

$$\begin{aligned}
\mathbb{E}|\mathsf{B}_s - \mathsf{B}_t|^{2k} &= \mathbb{E}|\mathsf{B}_{s-t}|^{2k} && \text{stationary} \\
&= \mathbb{E}\left|\sqrt{t-s} \cdot \mathsf{B}_1\right|^{2k} && \text{Gaussian} \\
&= |t-s|^k \cdot \mathbb{E}|B_1|^k
\end{aligned}$$

by properties of Brownian motion. With $\alpha = 2k$, $m = 1$, $\beta = k - m = k - 1$, we have continuity of order $\gamma < \frac{k-1}{2k}$ which maxes out to $\frac{1}{2}$.

Theorem 1.2.3 (Irregularity)
Brownian motion is nowhere Hölder continuous of order $\gamma > \frac{1}{2}$.

That is, it is differentiable nowhere.

We can make a heuristic argument as well. At some time τ, Brownian motion is independent of the information up to τ. That is, it is a memoryless Markov process. But then the left and right derivatives can never agree, since they are independently generated.

Definition 1.2.4 (Brownian bridge)
A stochastic process $\{\mathsf{B}^\circ_t : t \in [0,1]\}$ where:

- $\mathsf{B}^\circ_0 \equiv \mathsf{B}^\circ_1 \equiv 0$
- Each increment $B^\circ_{t_1} - B^\circ_{t_0}$ is independent of all others
- All increments $\mathsf{B}^\circ_t - \mathsf{B}^\circ_s \simeq \mathsf{B}^\circ_{t-s}$
- $\mathbf{B}^\circ_t \simeq \mathcal{N}(0, t(1-t))$
- The function $t \mapsto \mathsf{B}^\circ_t(\omega)$ is continuous for all ω

We can go back and forth between Brownian motions and bridges. If t is restricted to $[0,1]$, $\mathsf{B}^\circ_t \simeq \mathsf{B}_t - t\mathsf{B}_1$ and $\mathsf{B}_t \simeq \mathsf{B}^\circ_t + t\mathsf{Z}$ with perturbation $\mathsf{Z} \simeq \mathcal{N}(0,1) \perp \mathsf{B}^\circ_t$.

Since t is in the variance, $\frac{1}{\sqrt{c}}\mathsf{B}_{ct}$ is also Brownian, as is $t\mathsf{B}_{1/t}$.

[aside: Poisson and Lévy processes]

Now, consider a bunch more ways to construct Brownian motion.

Wiener's construction Let G_n be i.i.d. Gaussian variables. Then, based on a Fourier process, take

$$\mathsf{B}_t = t\mathsf{G}_0 + \sum_{n=1}^{\infty} \frac{\sin(n\pi t)}{nt}\mathsf{G}_n$$

We can apply this by truncating the series to get a Brownian path.

Ciecielski's construction For Haar wavelets (square waves) $\phi_{k/2^n}(t)$, notice that we can write:

$$\int_0^1 \mathsf{B}'_t \cdot \phi_{k/2^n}(t)\mathrm{d}t \simeq \mathsf{G}_{k/2^n}$$

which means we can go the other way to get

$$\mathsf{B}_t = \int \mathsf{B}'_t \,\mathrm{d}t = t\mathsf{G}_0 + \sum \mathsf{G}_{k/2^n} \int \phi_{k/2^n}(s)\,\mathrm{d}s$$

which is calculable, I guess.

Lévy's construction (which is actually simple-ish) Initialize points $\mathsf{B}_0 = 0$, $\mathsf{B}_1 \simeq \mathcal{N}(0,1)$. Recursively interpolate between points, adding a Gaussian perturbation to each point before saving it. In fact, it's possible to prove this is the same as Ciecielski's method, but without the weird shenanigans.

Donsker's construction Take n i.i.d. random variables $\xi_i \simeq F$ each with mean 0 and unit variance from literally any distribution. Let $S_n = \sum_{i=1}^{n} \xi_i$ be the cumulative sum. Then, $\mathsf{X}_t^n := \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$ will eventually converge to Brownian motion even though there's no Gaussian. In other words, Brownian motion is a sort of limiting behaviour of a random walk.

In fact, this is a weird stronger statement of the central limit theorem. The CLT says that at S_n, the sum converges to a Gaussian. Donsker says the whole path to get to S_n converges.

Lecture 2
Jan 16

Chapter 2

The stochastic integral

Recall: Brownian motion can be considered as a function of ω, t but it's easier to think of it as either:

1. A distribution $\Omega \to C(\mathbb{R}_+)$ from the sample space to the space of continuous functions of t
2. A collection of random variables $B(\cdot, t)$ for each time point t

It is hard to define Brownian motion if we just say

$$\mathrm{d}\mathsf{X}_t = \mathbf{f}_t(\mathsf{X}_t)\,\mathrm{d}t + G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

since Brownian motion is nowhere differentiable, i.e., $\mathrm{d}\mathsf{B}_t$ does not exist. Instead, write the integral

$$\mathsf{X}_T = \mathsf{X}_0 + \int_0^T \mathbf{f}_T(\mathsf{X}_t)\,\mathrm{d}t + \int_0^T G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$$

where the first term is pretty normal but the second is funny.

How can we define this?

2.1 Integration, writ large

An integral has an integrand $G_t(\mathsf{X}_t)$ living in some space and integrator B_t living in some other space. The integral is just a pair $\langle G_t(\mathsf{X}_t); \mathsf{B}_t \rangle$ that respects some properties.

An integral should have:

- a bilinear form: $\int(\alpha f + \beta g)\,\mathrm{d}\mathsf{B} = \alpha \int f\,\mathrm{d}\mathsf{B} + \beta \int g\,\mathrm{d}\mathsf{B}$ and $\int f\,\mathrm{d}(\alpha\mathsf{B} + \beta\mathsf{M}) = \alpha \int f\,\mathrm{d}\mathsf{B} + \beta \int f\,\mathrm{d}\mathsf{M}$
- continuity with respect to integrand and integrator
- generality to a big class of integrators/integrands
- some sort of change of variable formula
- computability

Definition 2.1.1 (Weiner integral)

Let $g : [0, T] \to \mathbb{R}$ be of bounded variation where $g(0) = g(T) = 0$. We define the integral

$$\int_0^T g(t)\,\mathrm{d}\mathsf{X}_t = \cancel{g(t)\mathsf{B}_t|_0^T} - \int_0^T \mathsf{X}_t\,\mathrm{d}g(t)$$

as the integration by parts.

This works nicely since X_t is continuous with respect to t and if we have sufficiently nice $g(t)$, this works. But diffusion models won't have nice $g(t)$.

What about $\int_0^T \mathsf{B}_t\,\mathrm{d}\mathsf{B}_t$, which is just…cursed.

Let's try a Riemann sum from MATH 137.

Definition 2.1.2 (Riemann-Stieltjes approximation)

Let $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$ with $\delta_n := \max_{0\le k\le n} |t_{k+1} - t_k| \to 0$.

Also pick $\tau_k := (1-\lambda)t_k + \lambda t_{k+1}$ for some $\lambda \in [0, 1]$.

Then, define the approximation R of $\int_0^1 \mathsf{B}_t\,\mathrm{d}\mathsf{B}_t$

$$\begin{aligned} R := R(\delta_n, \lambda) &= \sum_{k=0}^{n} \mathsf{B}_{\tau_k}[\mathsf{B}_{t_{k+1}} - \mathsf{B}_{t_k}] \\ &= \frac{\mathsf{B}_1^2}{2} - \underbrace{\frac{1}{2}\sum_{k=0}^{n}[\mathsf{B}_{t_{k+1}} - \mathsf{B}_{t_k}]^2}_{S_n(1)} + \underbrace{\sum_{k=0}^{n}[\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k}]^2}_{S_n(\lambda)} + \underbrace{\sum_{k=0}^{n}[\mathsf{B}_{t_{k+1}} - \mathsf{B}_{\tau_k}][\mathsf{B}_{\tau_k} - \mathsf{B}_{t_k}]}_{\text{some Gaussians with mean 0}} \\ &= \frac{\mathsf{B}_1^2}{2} - \frac{1}{2} + \lambda \end{aligned}$$

Claim that $S_n(\lambda) \to \lambda$.

Now, depending on what we choose for λ, i.e., what point we pick in the Riemann slice, we get different integrals. Itô picks $\lambda = 0 \implies R = (\mathsf{B}_1^2 - 1)/2$ and Stratonovich picks $\lambda = \frac{1}{2} \implies R = (\mathsf{B}_1^2)/2$. For this course, we pick Itô's integral.

2.2 Itô's integral

We construct Itô's integral by just wanting simple properties, e.g., the integral over a union should be the sum, integration should preserve convergence (i.e., continuity), etc.

Recall X_t is actually $\mathsf{X}(t, \omega)$ a function of two variables. Define an indicator $\mathsf{X}(t, \omega) = \mathbf{1}_{(\varsigma,\tau]}(t) \cdot \mathbf{1}_A(\omega)$ for some $A \subseteq \Omega$ that depends only on information up to time ς.

Now, define $\int \mathsf{X}_t\,\mathrm{d}\mathsf{B}_t := [\mathsf{B}_\tau - \mathsf{B}_\varsigma] \cdot 1_A$. This gives us linearity (good!) and integrates out t (better!).

For more complex $\mathsf{X}(t, \omega) = \sum_k c_k \mathbf{1}_{(\varsigma_k,\tau_k]} \cdot A_k$, define $\int \mathsf{X}_t\,\mathrm{d}\mathsf{B}_t := \sum_k c_k[\mathsf{B}_{\tau_k} - \mathsf{B}_{\varsigma_k}]$ to get more linearity.

Suppose $\mathsf{X}_t \in \mathcal{F}_t = \sigma(\{\mathsf{B}_s : 0 \le s \le t\})$ is an arbitrary non-anticipating function depending on information up to t. Also suppose $t \mapsto \mathsf{X}_t$ is left continuous. Then, we can discretize X_t by taking rectangles from the left point:

$$\mathsf{X}_t^n := \mathsf{X}_{(\lceil t2^n \rceil - 1)/2^n} = \sum_{k=0}^{\infty} \mathsf{X}_{k/2^n} [\![k/2^n < t \le (k+1)/2^n]\!]$$

We can approximate this as $\mathsf{X}_{k/2^n}(\omega) \approx \sum_i \frac{i}{2^m} \mathbf{1}_{A_i}(\omega)$ using indicator functions on the sets $A_i := \{(i-1)/2^m < \mathsf{X}_{k/2^n} \le i/2^m\}$.

Finally, define $\int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t = \lim_n \int \mathsf{X}_t^n \, \mathrm{d}\mathsf{B}_t$. This is the same as defining a Lebesgue integral.

We now try to define an isometry. Recall $\mathsf{X}(t, \omega) : \mathbb{R}_+ \times \Omega \to \mathbb{R}$.

Definition 2.2.1 (Doléans measure)
$\lambda = \lambda_{\mathsf{B}^2}$ as follows:

$$\begin{aligned}
\lambda\{(s,t] \times \mathbf{1}_A\} &:= \mathbb{E}[(\mathsf{B}_t^2 - \mathsf{B}_s^2)\mathbf{1}_A] = \mathbb{E}[(\mathsf{B}_t - \mathsf{B}_s)^2 \mathbf{1}_A] \\
&= \mathbb{E}[\mathbb{E}[(\mathsf{B}_t^2 - 2(\mathsf{B}_t - B_s)\mathsf{B}_s - \mathsf{B}_s^2)\mathbf{1}_A \mid \mathcal{F}_s]] \\
&= \mathbb{E}[\mathbb{E}[(\mathsf{B}_t^2 - 0) \mid \mathcal{F}_s]\mathbf{1}_A - \mathsf{B}_s^2 \mathbf{1}_A] \\
&= \mathbb{E}[\mathsf{B}_t^2 - \mathsf{B}_s^2]\mu(A) \\
&= (t-s) \cdot \mu(A)
\end{aligned}$$

We can now treat $\mathsf{X}(t, \omega)$ as a random variable from $\mathcal{L}^2(\mathbb{R}_+ \times \Omega, \mathcal{P}, \lambda)$ to $\mathbb{R}$. Then, the integral $\int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t$ is a linear map that sends $\mathsf{X} \in \mathcal{L}^2(\mathbb{R}_+ \times \Omega, \mathcal{P}, \lambda)$ to $\mathcal{L}^2(\Omega, \mu)$.

Also, we have the isometry

$$\|X\|^2_{L^2(\mathbb{R}_+ \times \Omega, \mathcal{P}, \lambda)} = \left\| \int \mathsf{X}_t \, \mathrm{d}\mathsf{B}_t \right\|_{\mathcal{L}^2(\Omega, \mu)}$$

This is nice because if we know $\|\mathsf{X}^n(t,\omega) - \mathsf{X}(t,\omega)\|^2 \to 0$, then $\|\mathsf{X}^n(t,\omega) - \mathsf{X}^m(t,\omega)\|^2 \to 0$. That is, given convergence of X^n, we get Cauchy. Then, by the definition, the integral exists.

2.3 Martingales

Define $\mathsf{M}_t = \int_0^t \mathsf{X}_\tau \, \mathrm{d}\mathsf{B}_\tau = \int \mathbf{1}_{(0,t]} \cdot \mathsf{X}_\tau \, \mathrm{d}\mathsf{B}_\tau$. We get these results:

- If $t \to \mathsf{X}_t$ is continuous, so is $t \to \mathsf{M}_t$
- Defining $\mathsf{M}_0 := 0$, we get $\mathbb{E}[\mathsf{M}_t] = 0$.
- The usual $\int_s^t \mathsf{X}_\tau \, \mathrm{d}\mathsf{B}_\tau = \int_0^t \mathsf{X}_\tau \, \mathrm{d}\mathsf{B}_\tau - \int_0^s \mathsf{X}_\tau \, \mathrm{d}\mathsf{B}_\tau$
- $\mathbb{E}[\mathsf{M}_t \mid \mathcal{F}_s] = \mathbb{E}[(\mathsf{M}_t - \mathsf{M}_s) + \mathsf{M}_s \mid \mathcal{F}_s] = \mathsf{M}_s$
- $\mathbb{E}[\mathsf{M}_t^2 - \mathsf{M}_s^2] = \mathbb{E}[\mathsf{M}_t^2 - 2(\mathsf{M}_t - \mathsf{M}_s)\mathsf{M}_s + \mathsf{M}_s^2] = \mathbb{E}[\mathsf{M}_t - \mathsf{M}_s]^2$

Notice that we never used any properties of Brownian motion other than the martingale property. That is, we have defined

$$\int \mathsf{Z}_t \, \mathrm{d}\mathsf{M}_t$$

for *any* $t \mapsto \mathsf{Z}_t$ left-continuous/non-anticipating and M_t continuous martingale. Then, we can define

$$\int_0^t \mathsf{X}\mathsf{Z} \, \mathrm{d}\mathsf{B} = \int_0^t \mathsf{Z} \, \mathrm{d}\mathsf{M} \quad \text{where} \quad \mathsf{M} = \int_0^t \mathsf{X} \, \mathrm{d}\mathsf{B}$$

as the integral with respect to the integral since the first integral was a martingale.

2.4 Differentials

Define the quadratic variation $[\mathsf{M}]_t = \mathsf{S}_t^n := \sum_{k=0}^n [\mathsf{M}_{t_{k+1}} - \mathsf{M}_{t_k}]^2$. As the sum of non-negative terms, it is increasing. Also, for Brownian motion $[\mathsf{B}]_t = t$.

Theorem 2.4.1 (Itô's formula)

For all functions $f(\mathsf{M}_t, \mathsf{V}_t)$ where M_t is a continuous martingale and V_t is continuously differentiable in t,

$$\mathrm{d}f(\mathsf{M}_t, \mathsf{V}_t) = f_x(\mathsf{M}_t, \mathsf{V}_t) \, \mathrm{d}\mathsf{M}_t + f_y(\mathsf{M}_t, \mathsf{V}_t) \, \mathrm{d}\mathsf{V}_t + \frac{1}{2} f_{xx}(\mathsf{M}_t, \mathsf{V}_t) \, \mathrm{d}[\mathsf{M}]_t$$

Then, we can calculate $[\int_0^t \mathsf{X} \, \mathrm{d}\mathsf{M}] = \int_0^t \mathsf{X}^2 \, \mathrm{d}[\mathsf{M}]$

Chapter 3

Stochastic differential equations

Lecture 3
Jan 23

Chapter 4

Fokker–Planck–Kolmogorov equation

Lecture 4
Jan 30

4.1 Deriving FPK

Let $p(s, \mathbf{x}, t, y)$ be a Markov kernel such that $p(s, \mathbf{x}, t, y) \geq 0$ and $\int p(s, \mathbf{x}, t, \mathbf{y})\, \mathrm{d}\mathbf{y} = 1$. We also enforce the Chapman-Kolmogorov equation

$$p(s, \mathbf{x}, t, y) = \int p(s, \mathbf{x}, \tau, z) p(\tau, \mathbf{z}, t, \mathbf{y})\, \mathrm{d}\mathbf{z}$$

for all $s < \tau < t$. Also, define an operator $T_{s,t}$ such that

$$T_{s,t} f(x) = [T_{s,t}(f)](x) = \int p(s, \mathbf{x}, t, \mathbf{y}) f(\mathbf{y})\, \mathrm{d}\mathbf{y}$$

to integrate out $\mathbf{y}$.

For example, let $f(\mathbf{y}) = \mathbb{1}_{\mathsf{B}}(\mathbf{y})$. Then, $T_{s,t} f(\mathbf{x}) = \int p(s, \mathbf{x}, t\mathbf{y}) \mathbb{1}_{\mathsf{B}}(y)\, \mathrm{d}y = \Pr[s, \mathbf{x}, t, \mathsf{B}]$.

Now, using T notation, we can write

$$\begin{aligned} T_{s,t} &= T_{s,\tau} T_{\tau,t} \\ &= \iint p(s, \mathbf{x}, \tau, \mathbf{z}) p(\tau, \mathbf{z}, t, \mathbf{y}) f(\mathbf{y})\, \mathrm{d}z\, \mathrm{d}y \\ &= \int p(s, \mathbf{x}, \tau, \mathbf{z}) T_{\tau,t} f(\mathbf{z})\, \mathrm{d}z \\ &= T_{s,\tau} T_{\tau,t} f(\mathbf{x}) \end{aligned}$$

Can we use this to define a derivative? What happens if $s \to t$? Define the <u>generator</u> of the Markov process

$$L_t = \lim_{h \downarrow 0} \frac{T_{t,t+h} - T_{t,t}}{h} = \lim_{h \downarrow 0} \frac{T_{t,t+h} - \mathrm{id}}{h}$$

and claim it exists because proving it rigorously is painful.

Consider a generator that can be written as

$$L_t f(\mathbf{x}) = \frac{1}{2} A(t, \mathbf{x}) \cdot \nabla^2 f(\mathbf{x}) + \mathbf{b}(t, \mathbf{x}) \cdot \nabla f(\mathbf{x})$$

which all generators from SDEs can be. This holds when

1. $\lim_{h\downarrow 0} p(t, \mathbf{x}, t+h, B_\varepsilon(\mathbf{x})) = 0$. That is, for a small timestep h, the chance of leaving a small ball ε around $\mathbf{x}$ is also very small. Intuitively, we have "continuity".
2. $\lim_{h\downarrow 0} \int_{B_\varepsilon(\mathbf{x})} (\mathbf{y}-\mathbf{x}) p(t, \mathbf{x}, t+h, \mathbf{y})\, \mathrm{d}\mathbf{y} = \mathbf{b}(t, x)$. That is, $\mathbf{b}(t, \mathbf{x})$ is the expected drift (average movement).
3. $\lim_{h\downarrow 0} \int_{B_\varepsilon(\mathbf{x})} (\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^\top p(t, \mathbf{x}, t+h, \mathbf{y})\, \mathrm{d}\mathbf{y} = A(t, x)$ That is, $A(t, \mathbf{x})$ is the expected covariance. Since the matrix $(\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^\top$ is rank 1 and positive semi-definite, $A(t, x)$ must also be positive semi-definite.

Equivalently, we can express this by saying

1. $\Pr[\|\mathbf{x}_{t+h} - \mathbf{x}_t\| \geq \varepsilon \mid \mathbf{x}_t] \to 0$ as $h \downarrow 0$.
2. $\frac{\mathbb{E}[\mathbf{x}_{t+h} - \mathbf{x}_t | \mathbf{x}_t = \mathbf{x}]}{h} \to \mathbf{b}(t, \mathbf{x})$
3. $\frac{\mathbb{E}[(\mathbf{x}_{t+h} - \mathbf{x}_t)(\mathbf{x}_{t+h} - \mathbf{x}_t)^\top | \mathbf{x}_t = \mathbf{x}]}{h} \to A(t, \mathbf{x})$

Now, let us write $\langle L_t f; p\rangle$ for the integral $\int L_t f \cdot p\, \mathrm{d}\mathbf{y}$. We have a useful property that

$$\langle L_t f; p\rangle = \langle f; L_t^* p\rangle$$

where L_t^* is the adjoint of L_t. For example, with our generator,

$$\begin{aligned}
\left\langle \frac{1}{2} A \cdot \nabla^2 f + \mathbf{b}\nabla \cdot f; p \right\rangle &= \left\langle \nabla^2 f; \frac{1}{2} pA \right\rangle + \langle \nabla f; p\mathbf{b}\rangle \\
&= \left\langle f; \nabla^2 \frac{1}{2} pA \right\rangle - \langle f; \nabla p\mathbf{b}\rangle \\
&= \left\langle f; \nabla^2 \cdot \frac{1}{2} pA - \nabla \cdot p\mathbf{b} \right\rangle \\
L_t^* p &= \nabla^2 \frac{1}{2} A - \nabla \mathbf{b}
\end{aligned}$$

using the integration by parts formula $\langle f'; g\rangle = -\langle f; g'\rangle$.

We can now write the (forward) FPK equation

$$\partial_t p(s, \mathbf{x}, t, \mathbf{y}) = L_t^* p(s, \mathbf{x}, t, \mathbf{y})$$

and the backward FPK equation

$$\partial_s p(s, \mathbf{x}, t, \mathbf{y}) = -L_s^* p(s, \mathbf{x}, t, \mathbf{y})$$

Proof. To derive this, take a function f and consider the derivative of the integral

$$\begin{aligned}
\partial_t \langle f; p(s, \mathbf{x}, t, \mathbf{y})\rangle &= \partial_t T_{s,t} f(\mathbf{x}) \\
&= \lim_{h\downarrow 0} \frac{T_{s,t+h} - T_{s,t}}{h} f(\mathbf{x}) \\
&= \lim_{h\downarrow 0} T_{s,t} \frac{T_{t,t+h} - T_{t,t}}{h} f(\mathbf{x}) \\
&= T_{s,t} L_t f(\mathbf{x}) \\
&= \langle L_t f; p(s, \mathbf{x}, t, \mathbf{y})\rangle \\
&= \langle f; L_t^* p(s, \mathbf{x}, t, \mathbf{y})\rangle
\end{aligned}$$

Then, if $\langle f; \partial_t p\rangle = \langle f; L_t^* p\rangle$ for all f, we must have $\partial_t p = L_t^* p$. □

4.2 Applying to the SDE

Recall the SDE $\mathrm{d}\mathsf{X}_t = \mathbf{b}_t(\mathsf{X}_t)\,\mathrm{d}t + G_t(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t$.

Claim that it has a generator $L_t f = \mathbf{b}_t \cdot \nabla f + \frac{1}{2}[G_t G_t^\top] \cdot \nabla^2 f$.

Then, the forward FPK equation is $\partial_t p = L_t^* p = -\nabla \cdot (p\mathbf{b}_t) + \frac{1}{2}\nabla^2 \cdot (pG_t G_t^\top)$

Proof. Take a function f. By Îto's formula,

$$\begin{aligned}
\mathrm{d}f(\mathsf{X}_t) &= \nabla f(\mathsf{X}_t)\,\mathrm{d}\mathsf{X}_t + \frac{1}{2}\nabla^2 f(\mathsf{X}_t) \cdot G_t G_t^\top\,\mathrm{d}t \\
&= \left[\mathbf{b}_t \cdot \nabla f + \frac{1}{2}\nabla^2 f + G_t G_t^\top\right]\mathrm{d}t + \nabla f \cdot G_t\,\mathrm{d}B_t \\
\mathrm{d}\left\langle f; p(s,\mathbf{x},t,\cdot)\right\rangle &= \left\langle L_t f; p(s,\mathbf{x},t,\cdot)\right\rangle \mathrm{d}t + 0 \\
\left\langle f; \partial_t p(s,\mathbf{x},t,\cdot)\right\rangle &= \left\langle f; L_t^* p(s,\mathbf{x},t,\cdot)\right\rangle \mathrm{d}t
\end{aligned}$$

and as above, since the formula must hold for all f, we have what we want. □

Now, suppose $G = 0$ so that $\mathrm{d}\mathsf{X}_t = \mathbf{b}_t(\mathsf{X}_t)\,\mathrm{d}t$. Then, $\partial_t p = -\nabla \cdot (pb)$. This is the continuity equation.

Recall the heat equation $\partial_t p = \Delta p = \sum \partial_i^2 p$. We recover the heat equation when $G = \sqrt{2}I$. In other words, the heat equation is the SDE $\mathrm{d}\mathsf{X}_t = \sqrt{2}\,\mathrm{d}\mathsf{B}_t$.

To solve this, try transforming the PDE into an ODE. Let $p(t,x) = p(t,\cdot)$ and consider $p : t \mapsto p(t,\cdot)$ as a function of just t. Then, we have $\frac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla f(p_t)$ with the energy function $f(p) = \frac{1}{2}\|\nabla p\|_2^2$.

Running gradient descent on f will find us the solution. Consider the derivative

$$\begin{aligned}
\left.\frac{\mathrm{d}f(p+\varepsilon q)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} &= \frac{\mathrm{d}\frac{1}{2}\|\nabla(p+\varepsilon q)\|}{\mathrm{d}\varepsilon} \\
&= \frac{\mathrm{d}\frac{1}{2}\int\langle \nabla p + \varepsilon\nabla q, \nabla p + \varepsilon\nabla q\rangle}{\mathrm{d}\varepsilon} \\
&= \int \langle \nabla p, \nabla q\rangle\,\mathrm{d}x \\
&= \langle \nabla p; \nabla q\rangle \\
&= -\langle \Delta p; q\rangle
\end{aligned}$$

Then, $\frac{\mathrm{d}p_t}{\mathrm{d}t} = \Delta p_t$.

4.3 Score matching

Consider two densities p and q. We define the Fisher divergence

$$F(p \parallel q) = \mathbb{E}_{\mathsf{X}\sim q}\|\nabla \log p(x) - \nabla \log q(x)\|_2^2 = \mathbb{E}_{\mathsf{X}\sim q}\left\|\mathbf{s}_p(x) - \mathbf{s}_q(x)\right\|^2$$

Also, define noisy versions $p_t = p * \mathcal{N}(0, 2t)$ and $q_t = q * \mathcal{N}(0, 2t)$. Then, the KL divergence

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(p_t \parallel q_t) = \frac{\mathrm{d}}{\mathrm{d}t}\int p_t \log\frac{p_t}{q_t}\,\mathrm{d}x = -F(p_t \parallel q_t)$$

Proof. Recall that the standard Gaussian $\mathcal{N}(0, 2t)$ will satisfy the heat equation $\partial_t p = \Delta p$. Then, since it is a convolution, p_t also satisfies the heat equation. That is,

$$\partial_t(p * \mathcal{N}(0, 2t)) = p * \partial_t \mathcal{N}(0, 2t) = p * \Delta\mathcal{N}(0, 2t) = \Delta(p * N(0, 2t))$$

Now,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(p_t \parallel q_t) \\
&= \int \frac{dp_t}{t} \cdot \log\frac{p_t}{q_t}\,\mathrm{d}x + \cancel{\int p_t \cdot \frac{\partial_t p_t}{p_t}\,\mathrm{d}x} - \int p_t \frac{\partial_t q_t}{q_t}\,\mathrm{d}x \\
&= \int \Delta p_t \cdot \log p_t\,\mathrm{d}x - \int \Delta p_t \log q_t\,\mathrm{d}x - \int p_t \frac{\Delta q_t}{q_t}\,\mathrm{d}x
\end{aligned}$$

The first term

$$\int \Delta p_t \cdot \log p_t\,\mathrm{d}x = \langle \Delta p_t; \log p_t\rangle = \left\langle p_t \frac{\Delta p_t}{p_t}; \log p_t \right\rangle = -\int p_t \|\mathbf{s}_{p_t}\|^2\,\mathrm{d}x$$

and the second term

$$-\int \Delta p_t \log q_t\,\mathrm{d}x = \int p_t \left\langle \mathbf{s}_{p_t}, \mathbf{s}_{q_t}\right\rangle \mathrm{d}x$$

and the third term

$$\begin{aligned}
-\int p_t \frac{\Delta q_t}{q_t}\,\mathrm{d}x &= -\left\langle \frac{p_t}{q_t}; \Delta q_t\right\rangle \\
&= \left\langle \nabla\frac{p_t}{q_t}; \nabla q_t\right\rangle \\
&= \int \frac{\nabla p_t \cdot q_t - p_t \cdot \nabla q_t}{q_t^2} \cdot \nabla q_t\,\mathrm{d}x \\
&= \int \frac{\nabla p_t}{p_t} \cdot \frac{\nabla q_t}{q_t} p_t - p_t\|\mathbf{s}_{q_t}\|^2\,\mathrm{d}x \\
&= \int \left\langle \mathbf{s}_{p_t}, \mathbf{s}_{q_t}\right\rangle \cdot p_t - p_t\|\mathbf{s}_{q_t}\|^2\,\mathrm{d}x
\end{aligned}$$

Then, we have

$$-\int p_t\|\mathbf{s}_{p_t}\|^2\,\mathrm{d}x + 2\int p_t \left\langle \mathbf{s}_{p_t}, \mathbf{s}_{q_t}\right\rangle \mathrm{d}x - \int p_t\|\mathbf{s}_{q_t}\|^2\,\mathrm{d}x = -\int p_t\|\mathbf{s}_{p_t} - \mathbf{s}_{q_t}\|^2\,\mathrm{d}x = -F(p_t \parallel q_t)$$

as desired. □

Now, given some distributions p and q, we want to find T such that $q = T_\#^{-1} p$ (i.e., $p = T_\# q$).

For example, if x is noise, we want to have $T^{-1}(x)$ be data. Mathematically, this is the same as taking data and pushing it to noise. If we can do one, we can do the other.

We build T over time. Let $p_t := (T_t^{-1})_\# p$ and $q_t := (T_t)_\# q$.

We can build p_t as the change in variables formula $p_t(x) = p(T_t x) \cdot |\det T_t'(x)|$.

To train, we want to minimize KL-divergence

$$\min_{T_t} \mathrm{KL}(q \parallel p_t) = \min_{T_t} \int q \log\frac{q}{p_t} = \min_{T_t} -\int q \cdot \log p_t \approx \max_{T_t} \sum_i \log p_t \mathbf{x}_i$$

we end up with a maximum log-likelihood estimator.