

CS 886 Winter 2024:

Lecture Notes

1	Introduction	2
1.1	Probability basics	2
1.2	Brownian motion	4

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

Lecture 1	Jan 9	2
-----------	-----------------	---

Chapter 1

Introduction

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:

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

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

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

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 $\log p(x) = E(x) - \log \int \dots$ 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 $\mu : \mathcal{F} \rightarrow [0, 1]$ assigns probability.

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

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

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

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

Lecture 1
Jan 9

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

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

Example 1.1.3. Say $Y \simeq \chi_1^2$.

That formally means $Y : (\Omega, \mathcal{F}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$ and $Y_{\#}\mu = \chi_1^2$.

Consider $f : (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$ where $x \mapsto x^2$. Compose $f(X) : (\Omega, \mathcal{F}, \mu) \rightarrow (\mathbb{R}, \mathcal{B})$. Then, $f(X) \simeq \chi_1^2$ by definition of the χ^2 distribution.

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

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 $X \simeq P$, then we can apply $f(X) \simeq Q$.

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 $f : \mathbb{R}^m \rightarrow \mathbb{R}^d$ such that $Z \simeq P \implies f(Z) \simeq Q$ (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 $Y \simeq \chi_1^2$.

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

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

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

we get a uniform distribution.

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

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

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

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 $X \xrightarrow{g} \mu$ as a composition of functions $X \xrightarrow{g_1} X_1 \xrightarrow{g_2} X_2 \xrightarrow{g_3} \dots \xrightarrow{g_n} X_n \approx \mu$.

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

This does not work because [reason]

Definition 1.1.8 (stochastic process)

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

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

depending on if we fix the time or state.

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

1.2 Brownian motion

Definition 1.2.1 (Brownian motion)

A stochastic process $\{B_t : t \geq 0\}$ such that:

- $B_0 \equiv 0$
- Each increment $B_{t_1} - B_{t_0}$ is independent of all others
- All increments $B_t - B_s \simeq B_{t-s}$
- $B_t \simeq \mathcal{N}(0, t)$
- The function $t \mapsto B_t(\omega)$ is continuous for all ω

Brownian motion is a Gaussian process with covariance kernel $\kappa(s, t) := \mathbb{E}[B_s B_t] = s \wedge t$ (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 \mathbf{X}_t be a stochastic process with index $t \in \mathbb{R}^m$. If for some $\alpha, \beta, L > 0$,

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

for all s and t , then there exists a modification $\tilde{\mathbf{X}}_t$ 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 , $\|\mathbf{X}_s - \mathbf{X}_t\| \leq c \cdot \|s - t\|^\gamma$.

We can show this condition holds for Brownian motion.

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

We know that

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

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 $\{\mathbf{B}_t^\circ : t \in [0, 1]\}$ where:

- $\mathbf{B}_0^\circ \equiv \mathbf{B}_1^\circ \equiv 0$
- Each increment $B_{t_1}^\circ - B_{t_0}^\circ$ is independent of all others
- All increments $\mathbf{B}_t^\circ - \mathbf{B}_s^\circ \simeq \mathbf{B}_{t-s}^\circ$
- $\mathbf{B}_t^\circ \simeq \mathcal{N}(0, t(1-t))$
- The function $t \mapsto \mathbf{B}_t^\circ(\omega)$ is continuous for all ω

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

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

[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

$$B_t = tG_0 + \sum_{n=1}^{\infty} \frac{\sin(n\pi t)}{nt} G_n$$

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:

$$\int_0^1 B'_t \cdot \phi_{k/2^n}(t) dt \simeq G_{k/2^n}$$

which means we can go the other way to get

$$B_t = \int B'_t dt = tG_0 + \sum G_{k/2^n} \int \phi_{k/2^n}(s) ds$$

which is calculable, I guess.

Lévy's construction (which is actually simple-ish) Initialize points $B_0 = 0$, $B_1 \simeq \mathcal{N}(0, 1)$. 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.

Donsker's construction Take 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, $X_t^n := \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$ 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.

Definition 1.2.5

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

$$\int_0^T g(t) dX_t = - \int_0^T X_t dg(t)$$

as the integration by parts.