CS886: Diffusion Models

Lec 00: Introduction

Yaoliang Yu

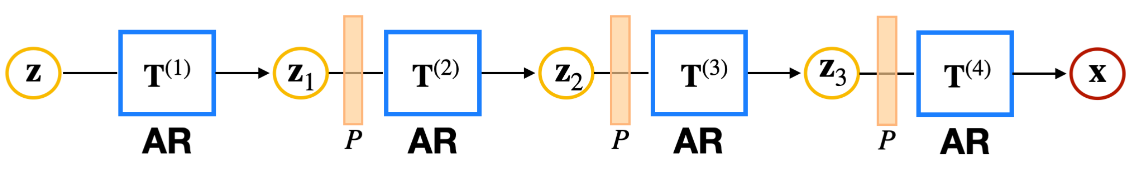
UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

January 9, 2024

Course Logistics

- Course web: https://cs.uwaterloo.ca/~y328yu/mycourses/886
- Learn: https://learn.uwaterloo.ca/d2l/home/982259
- Piazza: https://piazza.com/uwaterloo.ca/winter2024/cs886yu
- Part I: Necessary technical background (me)
- Part II: Paper presentation and discussion (you)
- Evaluation: presentation 40% + project report 60%
- Office hour: Thursday 1:30 - 2:30, DC3617

Auto-Regressive (AR) Flow Recalled



$$(\mathbf{T}_{\#} r)(\mathbf{x}) = r(\mathbf{z}) \;/\; \det\left(\nabla \mathbf{T}^{(1)} \mathbf{z}\right) \;/\; \det\left(\nabla \mathbf{T}^{(2)} \mathbf{z}_1\right) \;/\; \det\left(\nabla \mathbf{T}^{(3)} \mathbf{z}_2\right) \;/\; \det\left(\nabla \mathbf{T}^{(4)} \mathbf{z}_3\right)$$

$$x_j = z_j \cdot \exp\left(\alpha_j(z_1, \ldots, z_{j-1})\right) + \mu_j(z_1, \ldots, z_{j-1}) =: T_j(z_1, \ldots, z_{j-1}, z_j)$$

Now let the number of layers approach ∞!

Neural Ordinary Differential Equations (ODE)

$$\mathbf{x}_{t+1} \approx \mathbf{x}_t + \eta_t \cdot \mathbf{f}_t(\mathbf{x}_t) =: \mathbf{T}_t(\mathbf{x}_t)$$
$$\mathrm{d}\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t)\,\mathrm{d}t$$

- Suppose $\mathbf{x}_t \sim p_t$
- Apply change-of-variable-formula we know $\mathbf{x}_{t+1} \sim p_{t+1}$, where

$$\begin{aligned}\log p_{t+1}(\mathbf{x}_{t+1}) &= \log p_t(\mathbf{x}_t) - \log|\det \partial_{\mathbf{x}} \mathbf{T}_t(\mathbf{x}_t)| \\ &= \log p_t(\mathbf{x}_t) - \log|\det[\mathrm{Id} + \eta_t \cdot \partial_{\mathbf{x}} \mathbf{f}_t(\mathbf{x}_t)]| \\ &\approx \log p_t(\mathbf{x}_t) - \eta_t \cdot \langle \partial_{\mathbf{x}}, \mathbf{f}_t(\mathbf{x}_t) \rangle\end{aligned}$$

- Continuous change-of-variable formula:

$$\boxed{\frac{\mathrm{d}\log p_t(\mathbf{x}_t)}{\mathrm{d}t} = -\langle \partial_{\mathbf{x}}, \mathbf{f}_t(\mathbf{x}_t) \rangle}$$

T. Q. Chen et al. "Neural ordinary differential equations". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6572–6583.

Stochastic Differential Equations (SDE)

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

$$\mathbf{x}_{t+1} \approx \mathbf{x}_t + \eta_t \cdot \mathbf{f}_t(\mathbf{x}_t) + \mathbf{g}_t(\mathbf{x}_t), \quad \text{where} \quad \mathbf{g}_t(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{0}, \eta_t^2 G_t(\mathbf{x}_t) G_t(\mathbf{x}_t)^\top)$$

- $\mathbf{x}_{t+1}$ is now a noisy version of $\mathbf{x}_t$
- Suppose $\mathbf{x}_t \sim p_t$
- Kolmogorov forward equation (a.k.a. Fokker-Planck equation):

$$\boxed{\partial_t p_t = -\langle \partial_\mathbf{x}, p_t \mathbf{f}_t \rangle + \tfrac{1}{2} \langle \partial_\mathbf{x} \partial_\mathbf{x}^\top, p_t G_t G_t^\top \rangle}$$

- Kolmogorov backward equation (with fixed end time $t > s$):

$$-\partial_s p_s = \langle \mathbf{f}_s, \partial_\mathbf{x} p_s \rangle + \tfrac{1}{2} \langle G_s G_s^\top, \partial_\mathbf{x} \partial_\mathbf{x}^\top p_s \rangle$$

ODE $\Leftrightarrow$ SDE

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

- Any ODE is a (trivial) SDE with $G_t \equiv \mathbf{0}$
- Conversely, any SDE is equivalent to an ODE:

$$\boxed{\mathbf{f}_t \leftarrow \mathbf{f}_t - \tfrac{1}{2} G_t G_t^\top \partial_{\mathbf{x}} - \tfrac{1}{2} G_t G_t^\top \partial_{\mathbf{x}} \log p_t}$$

- The score function plays an important role:

$$\boxed{\mathbf{s}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x}) := \partial_{\mathbf{x}} \log p(\mathbf{x})}$$

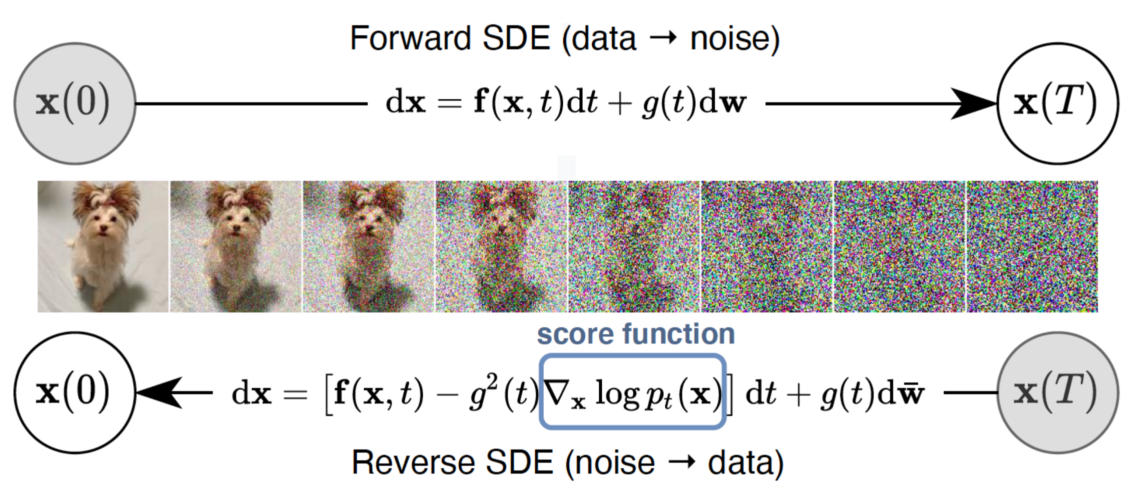
Reverse-time SDE

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

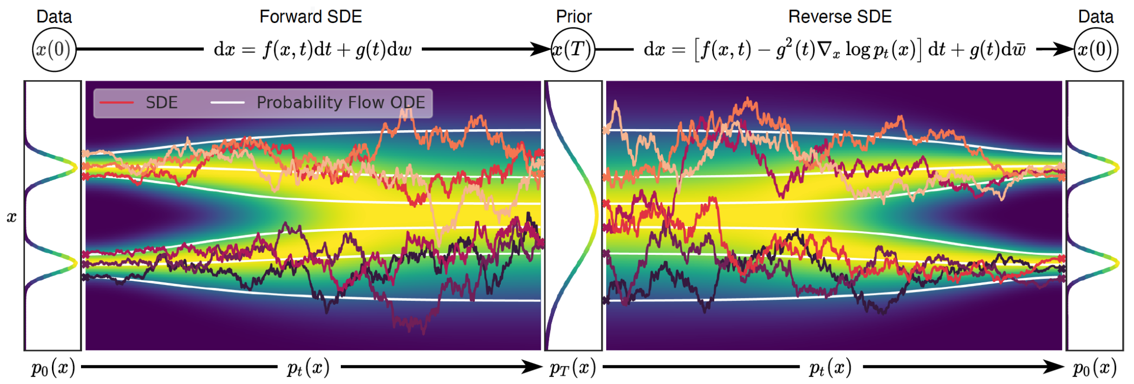
$$\mathrm{d}\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{f}}_t(\bar{\mathbf{x}}_t)\,\mathrm{d}t + G_t(\bar{\mathbf{x}}_t)\,\mathrm{d}\bar{\mathbf{n}}_t, \quad \text{where} \quad \boxed{\bar{\mathbf{f}}_t = \mathbf{f}_t - G_t G_t^\top \partial_{\mathbf{x}} - G_t G_t^\top \partial_{\mathbf{x}} \log p_t}$$

- Time flows backwards for the bar $\overline{\text{quantities}}$
- Forward SDE: diffuses date into noise
- Reverse SDE: molds noise into data
- $\mathbf{f}_t$ and G_t together specify $\bar{\mathbf{f}}_t$: key is to estimate the score $\partial_{\mathbf{x}} \log p_t$

B. D. O. Anderson. "Reverse-time diffusion equation models". *Stochastic Processes and their Applications*, vol. 12, no. 3 (1982), pp. 313–326.



Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.



Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

Score Matching

$$\begin{aligned}\mathbb{F}(p\|q) &:= \tfrac{1}{2}\mathbb{E}_{\mathsf{X}\sim q}\|\partial_{\mathbf{x}}\log p(\mathsf{X}) - \partial_{\mathbf{x}}\log q(\mathsf{X})\|_2^2 \\ &= \mathbb{E}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathbf{s}_p(\mathsf{X})\|_2^2 + \langle\partial_{\mathbf{x}}, \mathbf{s}_p(\mathsf{X})\rangle + \tfrac{1}{2}\|\mathbf{s}_q(\mathsf{X})\|_2^2\right] \\ &\approx \hat{\mathbb{E}}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathbf{s}_p(\mathsf{X})\|_2^2 + \langle\partial_{\mathbf{x}}, \mathbf{s}_p(\mathsf{X})\rangle\right]\end{aligned}$$

- Under mild conditions, $\mathbb{F}(p\|q) = 0 \iff p \propto q$
- A Convenient way to estimate the score $\mathbf{s}_q$ and hence the density q
- The model score function $\mathbf{s}_p$ can be chosen as any NN

A. Hyvärinen. "Estimation of Non-Normalized Statistical Models by Score Matching". *Journal of Machine Learning Research*, vol. 6, no. 24 (2005), pp. 695–709.

Score Matching for Exponential Family

$$\min_{\boldsymbol{\theta}} \ \hat{\mathbb{E}}_{\mathsf{X}\sim q}\left[\tfrac{1}{2}\|\mathbf{s}(\mathsf{X};\boldsymbol{\theta})\|_2^2 + \langle \partial_{\mathbf{x}}, \mathbf{s}(\mathsf{X};\boldsymbol{\theta})\rangle\right]$$

- If the model density p is in the exponential family:

$$\mathbf{s}(\mathbf{x};\boldsymbol{\theta}) = \partial_{\mathbf{x}}\left\langle \mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle = \left[\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})\right]^{\top}\boldsymbol{\theta}$$
$$\langle \partial_{\mathbf{x}}, \mathbf{s}(\mathbf{x};\boldsymbol{\theta})\rangle = \langle \partial_{\mathbf{x}}, \partial_{\mathbf{x}}\left\langle \mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle\rangle = \left\langle \partial_{\mathbf{x}}^2\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}\right\rangle$$

- Can solve $\boldsymbol{\theta}$ in closed-form by simply setting the derivative w.r.t. $\boldsymbol{\theta}$ to $\mathbf{0}$:

$$\boldsymbol{\theta} = -\left\{\hat{\mathbb{E}}_{\mathsf{X}\sim q}[\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})]^{\top}[\partial_{\mathbf{x}}\mathbf{T}(\mathbf{x})]\right\}^{-1} \cdot \hat{\mathbb{E}}_{\mathsf{X}\sim q}[\partial_{\mathbf{x}}^2\mathbf{T}(\mathbf{x})]$$

- For multivariate Gaussian, $\boldsymbol{\theta} = (S^{-1}, S^{-1}\boldsymbol{\mu})$, $\mathbf{T}(\mathbf{x}) = (-\frac{1}{2}\mathbf{x}\mathbf{x}^{\top}, \mathbf{x})$ and

$$\min_{\boldsymbol{\mu}, S} \ \hat{\mathbb{E}}_{\mathsf{X}\sim q} \ \tfrac{1}{2}\|S^{-1}(\mathbf{x}-\boldsymbol{\mu})\|_2^2 - \mathrm{tr}(S^{-1})$$

Denoising Auto-Encoder

- Suppose also have a latent variable Z with joint density $q(\mathbf{x}, \mathbf{z})$
- Exchage differentiation with integration we obtain:

$$
\begin{aligned}
\mathbb{F}(p\|q) &:= \tfrac{1}{2}\mathbb{E}_{\mathsf{X}\sim q}\|\partial_{\mathbf{x}} \log p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X})\|_2^2 \\
&= \tfrac{1}{2}\mathbb{E}_{(\mathsf{X},\mathsf{Z})\sim q}[\|\mathbf{s}_p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2 + \|\mathbf{s}_q(\mathsf{X})\|_2^2 - \|\partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2] \\
&\approx \tfrac{1}{2}\hat{\mathbb{E}}_{(\mathsf{X},\mathsf{Z})\sim q}\|\mathbf{s}_p(\mathsf{X}) - \partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})\|_2^2
\end{aligned}
$$

- Useful when the conditional density $\partial_{\mathbf{x}} \log q(\mathsf{X}|\mathsf{Z})$ is easy to obtain

P. Vincent. "A Connection Between Score Matching and Denoising Autoencoders". *Neural Computation*, vol. 23, no. 7 (2011), pp. 1661–1674.

Score-based Diffusion Generative Models

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

$$\mathbf{x}_{t+1} \approx \mathbf{x}_t + \eta_t \cdot \mathbf{f}_t(\mathbf{x}_t) + \mathbf{g}_t(\mathbf{x}_t), \quad \text{where} \quad \mathbf{g}_t(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{0}, \eta_t^2 G_t(\mathbf{x}_t) G_t(\mathbf{x}_t)^\top)$$

- Key is to estimate the score $\mathbf{s}_t(\mathbf{x}) = \partial_{\mathbf{x}} \log p_t$
- Apply denoising auto-encoder score matching:

$$\min_{\boldsymbol{\theta}} \hat{\mathbb{E}}_{t\sim\mu,(\mathsf{X}_t,\mathsf{X}_0)\sim q(\mathbf{x}_t,\mathbf{x}_0)} \lambda_t \|\mathbf{s}_t(\mathsf{X}_t;\boldsymbol{\theta}) - \partial_{\mathbf{x}} \log q(\mathsf{X}_t|\mathsf{X}_0)\|_2^2$$

  - $\mathsf{X}_0 \sim q(\mathbf{x})$, the data density
  - $q(\mathbf{x}_t|\mathbf{x}_0)$ can be derived from the forward SDE, in closed-form if $\mathbf{f}_t$ is affine

Inference After Learning

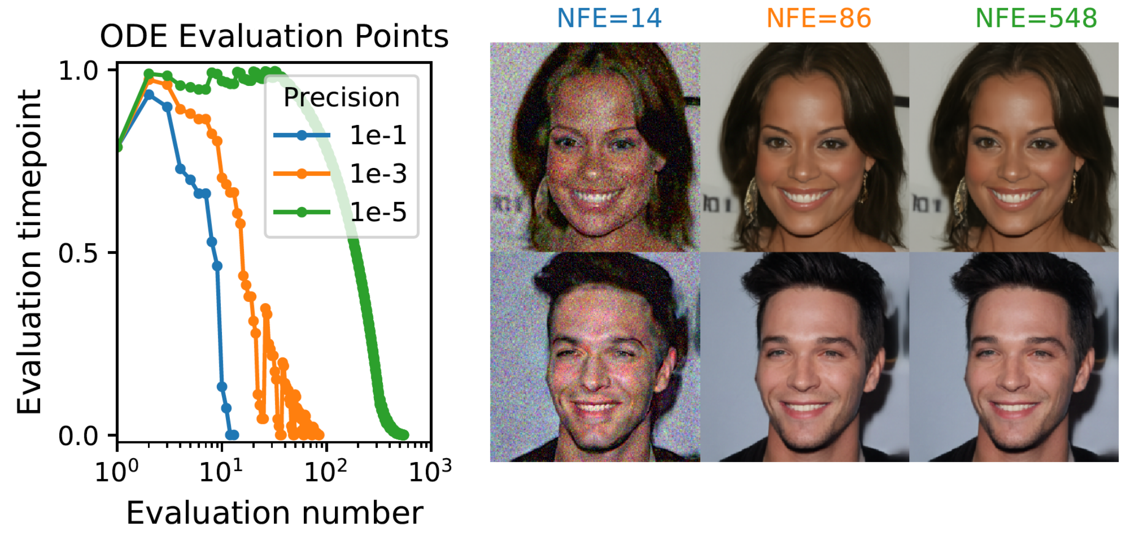
$$\mathrm{d}\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t)\,\mathrm{d}t + G_t(\mathbf{x}_t)\,\mathrm{d}\mathbf{n}_t$$

$$\mathrm{d}\bar{\mathbf{x}}_{t+1} = \boxed{\mathbf{f}_t - G_t G_t^\top \partial_{\mathbf{x}} - G_t G_t^\top \mathbf{s}_t(\bar{\mathbf{x}}_t; \boldsymbol{\theta})}\,\mathrm{d}t + G_t(\bar{\mathbf{x}}_t)\,\mathrm{d}\bar{\mathbf{n}}_t$$

$$\mathrm{d}\mathbf{x}_{t+1} = \boxed{\mathbf{f}_t - \tfrac{1}{2} G_t G_t^\top \partial_{\mathbf{x}} - \tfrac{1}{2} G_t G_t^\top \mathbf{s}_t(\mathbf{x}_t; \boldsymbol{\theta})}\,\mathrm{d}t$$

- Run the reverse SDE or the equivalent ODE
  - sample $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$
  - apply numerical SDE or ODE solver (e.g. Euler-Maruyama)

D. J. Higham. "An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations". *SIAM Review*, vol. 43, no. 3 (2001), pp. 525–546.

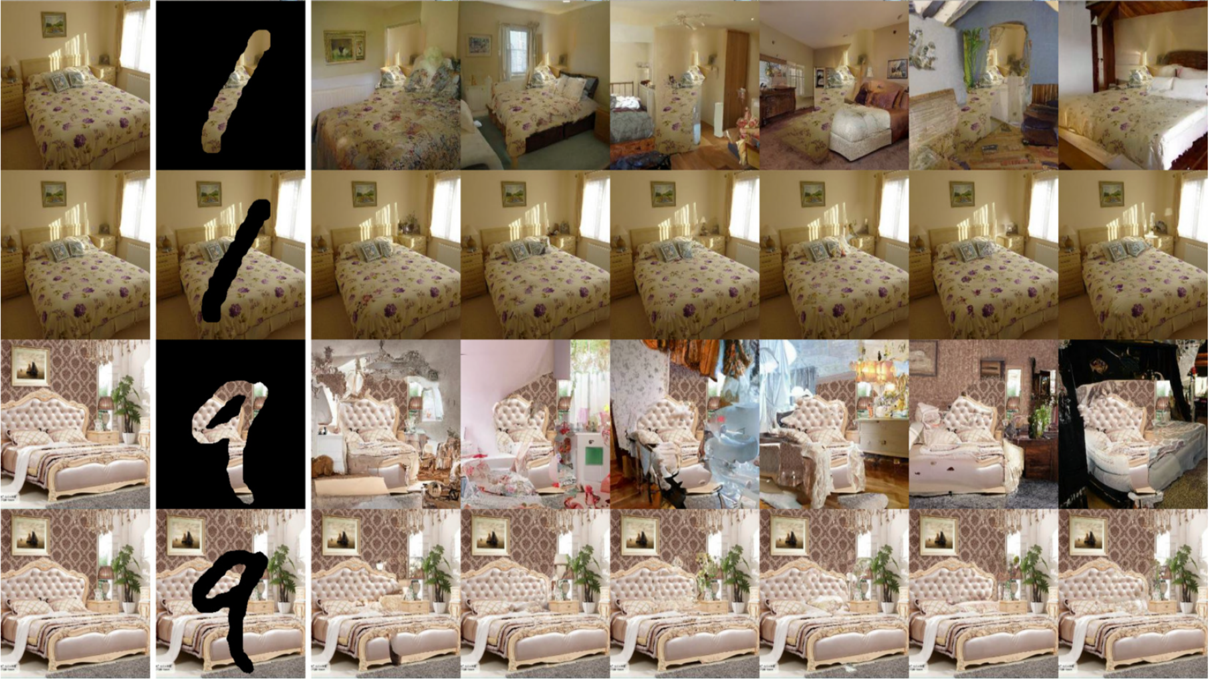


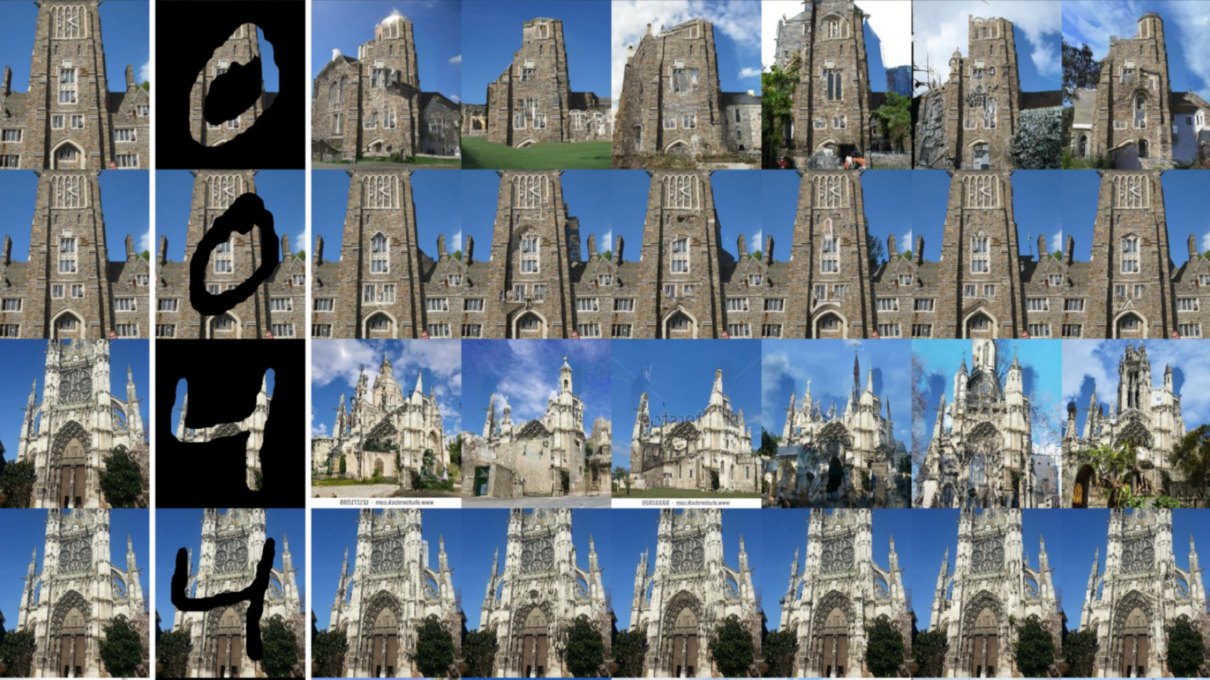
Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

Interpolation

Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

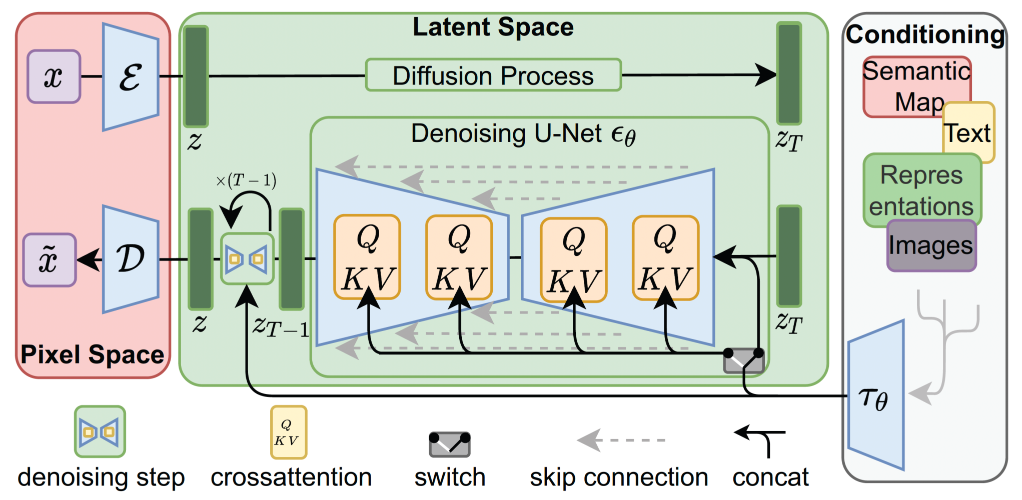
L00






shutterstock
www.shutterstock.com · 31170427

Stable Diffusion



R. Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10674–10685.

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

- What is a Brownian motion B_t?
- What is the integral $\int_0^t G_t(\mathbf{x}_t)\,\mathrm{dB}_t$?
- What is a stochastic differential equation (SDE)?
  - Fokker-Planck-Kolmogorov equation
- What is a reverse SDE?
- What is a score function and how to estimate it?
- How to numerically simulate an SDE or its reverse?

What Is a Random Variable?

- We fix a standard sample space $(\Omega, \mathcal{F}, \mu)$
  - $\mathcal{F} \subseteq 2^{\Omega}$, $\mu : \mathcal{F} \to [0,1]$ assigns probability
  - e.g. $\Omega = [0,1]$ and μ the Lebesgue measure
- A random variable (r.v.) is a *function* $\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{S}, \mathcal{B})$
  - $\mathbb{S}$ is the state space (range), e.g., $\mathbb{S} = \mathbb{R}$
  - the distribution of X is a probability measure on $\mathcal{B} \subseteq 2^{\mathbb{S}}$:

$$\forall S \in \mathcal{B}, \quad (\mathsf{X}_{\#}\mu)(S) := \mu(\{\omega : \mathsf{X}(\omega) \in S\}) = \mu(\mathsf{X}^{-1}(S))$$

  - for this to always make sense, need $\mathsf{X}^{-1}(\mathcal{B}) \subseteq \mathcal{F}$ (so-called measurability)

Push-forward

$$\forall S \in \mathcal{B}, \quad (\mathsf{X}_{\#}\mu)(S) := \mu(\{\omega : \mathsf{X}(\omega) \in S\}) = \boxed{\mu(\mathsf{X}^{-1}(S))}$$

- The function (r.v.) X "pushes" the probability μ forward to the state space

$$\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{S}, \mathcal{B}, \mathsf{X}_{\#}\mu)$$

  - by pulling the computation on $\mathcal{B}$ back to the sample space $(\Omega, \mathcal{F}, \mu)$ through X^{-1}

- In particular, if $\omega \simeq \mu$, then $\mathsf{X}(\omega) \simeq \mathsf{X}_{\#}\mu$

- This is one of the main ideas behind generative modeling

Example

- $\mathsf{X} \simeq \mathcal{N}(0,1)$, meaning, $\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$ and $\mathsf{X}_{\#}\mu = \mathcal{N}(0,1)$
  - $\omega \simeq \mu \implies \mathsf{X}(\omega) \simeq \mathcal{N}(0,1)$
- $\mathsf{Y} \simeq \chi_1^2$, meaning, $\mathsf{Y} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$ and $\mathsf{Y}_{\#}\mu = \chi_1^2$
  - $\omega \simeq \mu \implies \mathsf{Y}(\omega) \simeq \chi_1^2$
- Consider the function $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}), \quad x \mapsto x^2$
- Then, the composition $f(\mathsf{X}) : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$
- What is the distribution of $f(\mathsf{X})$?

Function $\longrightarrow$ Distribution

$$\mu[f(\mathsf{X}) \in S] = \mu[\mathsf{X} \in f^{-1}(S)] = \mu[\mathsf{X}^{-1}\big(f^{-1}(S)\big)] = [\mathsf{X}_{\#}\mu]\big(f^{-1}(S)\big) = [f_{\#}(\mathsf{X}_{\#}\mu)](S)$$

- The distribution of $f(\mathsf{X})$ is thus $(f \circ \mathsf{X})_{\#}\mu = f_{\#}[\mathsf{X}_{\#}\mu]$
  - recall that $\mathsf{X}_{\#}\mu$ is the distribution of X
- For our choice of f, we know $f(\mathsf{X}) \simeq \mathsf{Y}$, i.e., $f_{\#}[\mathsf{X}_{\#}\mu] = \mathsf{Y}_{\#}\mu$
  - in other words, $\mathsf{X} \simeq \mathsf{X}_{\#}\mu \implies f(\mathsf{X}) \simeq \mathsf{Y}_{\#}\mu$
- Two equivalent views
  - $f \circ \mathsf{X}$ as composition: $\mathsf{X} : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B})$, $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$
  - abstract 1 layer away: $f : (\mathbb{R}, \mathcal{B}, \mathcal{N}(0, 1)) \to (\mathbb{R}, \mathcal{B})$
- From one probability μ on Ω, each function f induces a distribution $f_{\#}\mu$ on $\mathbb{S}$

Function $\longleftarrow$ Distribution

Inverse problem: given distributions P and Q, find f such that $f_{\#}P = Q$

- In generative models, $P = \mathcal{N}(\mathbf{0}, I)$ is pure noise, Q data distribution
- If possible, draw $\mathsf{X} \simeq P$ to get $f(\mathsf{X}) \simeq Q$ (a.k.a. sampling or inference)

Theorem: 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

$$\mathsf{Z} \simeq P \implies f(\mathsf{Z}) \simeq Q, \quad i.e., \quad f_{\#}P = Q.$$

- In reality, only have empirical $\hat{Q}$, which is discrete

Example: $\mathsf{X} \simeq \mathcal{N}(0,1)$, $\mathsf{Y} \simeq \chi^2_1$

- By definition, $\mathsf{Y} \simeq \mathsf{X}^2$ hence $f(x) = x^2$ works
- What is the distribution of $\Phi(\mathsf{X})$, Φ being c.d.f. of $\mathcal{N}(0,1)$?

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

- What is the distribution of $\Psi^{-1}(\Phi(\mathsf{X}))$, Ψ being c.d.f. of χ^2?

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

- Thus, $f(x) = [\Psi^{-1} \circ \Phi](x)$ also works

Generalization

- Forward: $\mathsf{X} \xrightarrow{g} \mu$
  - $\mathsf{X} \xrightarrow{g_1} \mathsf{X}_1 \xrightarrow{g_2} \mathsf{X}_2 \xrightarrow{g_3} \cdots \xrightarrow{g_n} \mathsf{X}_n \approx \mu$
- Backward: $\mathsf{Y} \xleftarrow{h} \mu$
  - $\mathsf{Y} \approx \mathsf{Y}_n \xleftarrow{h_n} \cdots \xleftarrow{h_3} \mathsf{Y}_2 \xleftarrow{h_2} \mathsf{Y}_1 \xleftarrow{h_1} \mu$
- $f = h \circ g$ brings X to Y; will stretch $n \to \infty$
- Difficulty?

What Is a Stochastic Process?

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

Brownian Motion

A stochastic process $\{B_t : t \geq 0\}$ is called Brownian motion if

- Initialization: $B_0 \equiv 0$
- Independent increment: $\forall n, \forall t_0 \leq t_1 \leq \cdots \leq t_n, B_{t_1} - B_{t_0} \perp \cdots \perp B_{t_n} - B_{t_{n-1}}$
- Stationary increment: $\forall s \leq t, B_t - B_s \simeq B_{t-s} - B_0$
- Gaussian: $B_t \simeq \mathcal{N}(0, t)$
- Continuous sample path: for (almost) all ω, $t \mapsto B_t(\omega)$ is continuous

Brownian motion is a (continuous) Gaussian process with covariance kernel

$$\kappa(s,t) := \mathbb{E}(B_s B_t) = s \wedge t$$

Derivative Kernel

- Let $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a (reproducing) kernel
  - $\forall n$, $\forall \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{X}$, let $K_{ij} := \kappa(\mathbf{x}_i, \mathbf{x}_j)$, we have $K \succeq \mathbf{0}$
  - equivalently, $\exists \varphi : \mathbb{X} \to \mathbb{H}$ such that $\kappa(\mathbf{x}, \mathbf{z}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{z}) \rangle$
- The partial derivative $\kappa' := \partial_{12}\kappa$ is also a kernel
- Derivative kernel of the Brownian motion kernel $\kappa(s,t) = s \wedge t$?
- "White noise" B'_t as derivative of Brownian motion B_t

"There are in this world optimists who feel that any symbol that starts off with an integral sign must necessarily denote something that will have every property that they should like an integral to possess. This of course is quite annoying to us rigorous mathematicians; what is even more annoying is that by doing so they often come up with the right answer."

— *Edward J. McShane*

E. J. McShane. "Integrals devised for special purposes". *Bulletin of the American Mathematical Society*, vol. 69 (1963), pp. 597–627.

Continuity

Theorem: Continuity condition of stochastic processes

Let X_t be a stochastic process with index $t \in \mathbb{R}^m$. If for some $\alpha, \beta, \mathsf{L} > 0$,

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

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

Hölder continuous at s of order γ:

$$\forall t \text{ around } s, \quad \|\mathsf{X}_s - \mathsf{X}_t\| \leq c \cdot \|s - t\|^\gamma$$

- Gaussian kernel: $\kappa(s,t) = \kappa(s-t) = \exp(-(s-t)^2)$
- Laplacian kernel: $\kappa(s,t) = \kappa(s-t) = \exp(-|s-t|)$

Kolmogorov's Construction of Brownian Motion

- For any finitely many $t_1, \ldots, t_n$,

$$\mathsf{B}_{1:n} := (\mathsf{B}_{t_1}, \ldots, \mathsf{B}_{t_n}) \simeq \mathcal{N}(\mathbf{0}, K_n)$$

  where $K_n(t_i, t_j) = t_i \wedge t_j$.

- Kolmogorov extension theorem $\implies$ Gaussian process B_t exists
- Moment: $\mathbb{E}|\mathsf{B}_s - \mathsf{B}_t|^{2k} = \mathbb{E}|\sqrt{t-s} \cdot \mathsf{B}_1|^{2k} = |t-s|^k \cdot \mathbb{E}|\mathsf{B}_1|^k$
- Identifying $\alpha = 2k, m = 1, \beta = k - m = k - 1 \implies \gamma < \frac{k-1}{2k}$

Brownian motion is locally Hölder continuous of order $\gamma < \frac{1}{2}$

Nondifferentiable Almost Everywhere (a.e.)

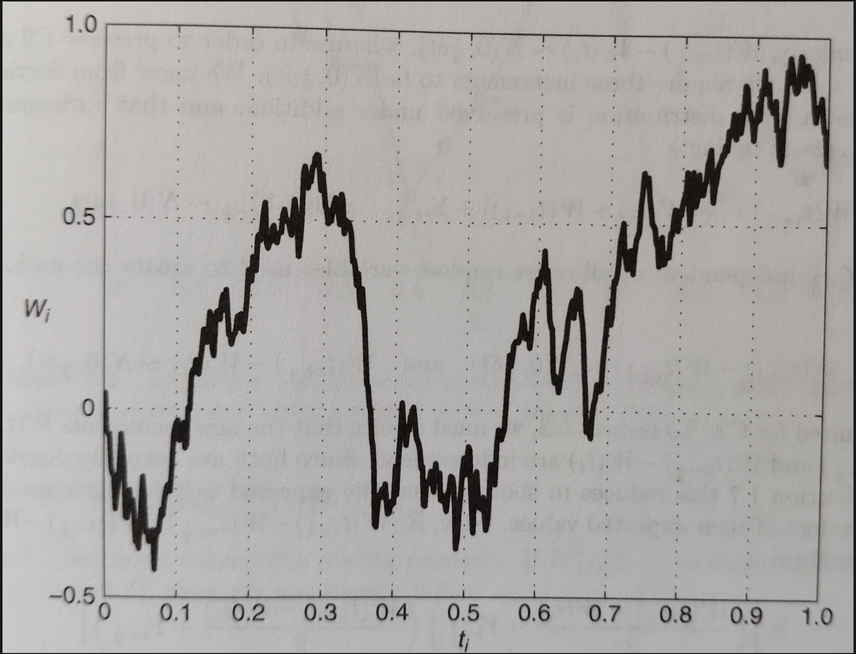
Theorem: Irregularity

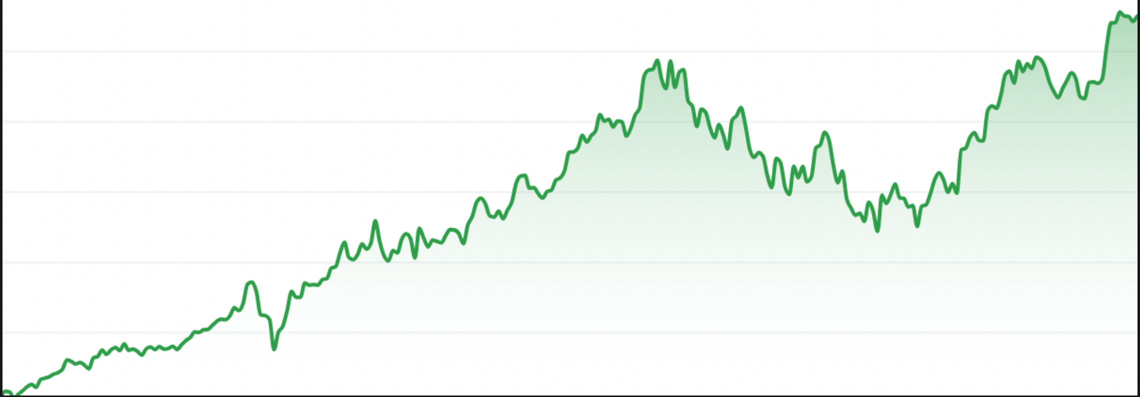
Brownian motion is nowhere Hölder continuous of order $\gamma > \frac{1}{2}$.

- Sample path of Brownian motion is of infinite variation over any (nonempty) interval
- With a bit more work, it can be proved that

$$\Pr\left(\limsup_{h\to 0} \frac{\sup_{0\le t\le 1-h} |\mathsf{B}_{t+h} - \mathsf{B}_t|}{\sqrt{2h|\log h|}} = 1\right) = 1,$$

thus Brownian motion is not Hölder continuous of order $\frac{1}{2}$ (at some point t).


1.0
0.5
0
−0.5
W_i
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
t_i



Brownian Motion is Markov

Theorem: (Strong) Markov property

$\{B_{t+\tau} - B_\tau\}_{t\geq 0}$ is a Brownian motion and independent of $\mathcal{F}_\tau$

- For $\tau = s$, $\mathcal{F}_s = \sigma(B_1, \ldots, B_s)$: information up to time τ
- For small $t > 0$, Brownian motion has forgotten how it went into B_τ
- It started afresh and hence cannot match the left and right derivatives at τ
- Brownian motion is a Markov process:

$$\Pr(B_{t+s} \in S|\mathcal{F}_s) = \Pr(B_{t+s} - B_s + B_s \in S|\mathcal{F}_s) = \Pr(B_{t+s} \in S|B_s)$$

Brownian Bridge

A stochastic process $\{B_t^\circ : t \in [0,1]\}$ is called a Brownian bridge if

- Initialization: $B_0^\circ = B_1^\circ \equiv 0$
- Independent increment: $\forall n, \forall t_0 \le t_1 \le \cdots \le t_n$, $B_{t_1}^\circ - B_{t_0}^\circ \perp \cdots \perp B_{t_n}^\circ - B_{t_{n-1}}^\circ$
- Stationary increment: $\forall 0 \le s \le t \le 1$, $B_t^\circ - B_s^\circ \simeq B_{t-s}^\circ - B_0^\circ$
- Gaussian: $B_t^\circ \simeq \mathcal{N}(0, t(1-t))$
- Continuous sample path: for (almost) all ω, $t \mapsto B_t^\circ(\omega)$ is continuous

Brownian bridge is a (continuous) Gaussian process with covariance kernel

$$\kappa(s,t) := s \wedge t - st$$

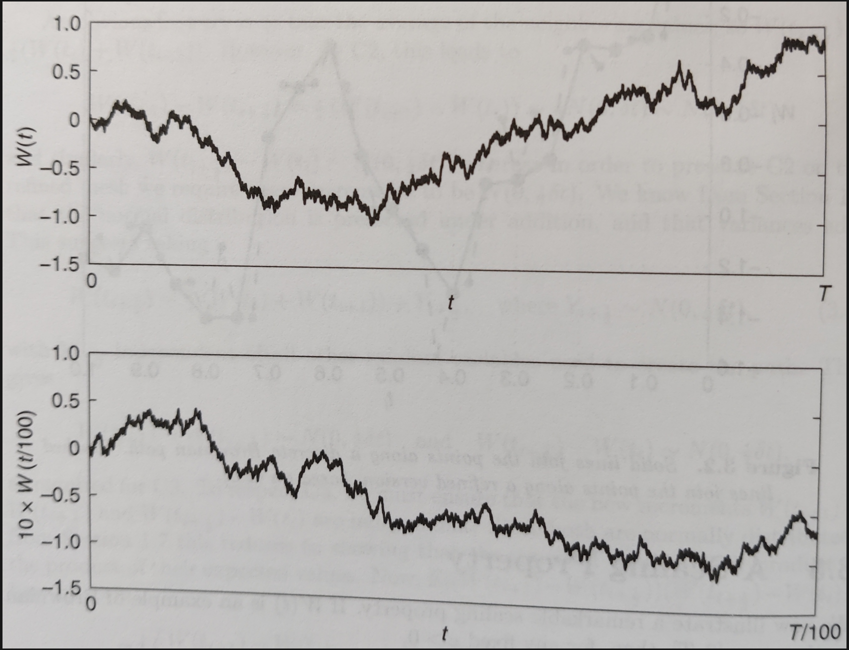
Some Calculus

Restricting t to $[0,1]$:

- $B_t^\circ \simeq B_t - tB_1$
- $B_t \simeq B_t^\circ + tZ$, where $Z \simeq \mathcal{N}(0,1) \perp B_t^\circ$

The following are Brownian motions:

- Change of time: $\frac{1}{\sqrt{c}} B_{ct}$
- Time inversion: $tB_{1/t}$ (what about $\frac{1}{t}B_t$?)
- Independent combination: $\sqrt{\lambda}B_t + \sqrt{1-\lambda}Z_t$ for $B_t \perp Z_t$


1.0
0.5
0
−0.5
−1.0
−1.5
W(t)
0
t
T
1.0
0.5
0
−0.5
−1.0
−1.5
10 × W (t/100)
0
t
T/100

Lévy Process

A stochastic process $\{X_t\}$ is called a Lévy process if

- Initialization: $X_0 \equiv 0$
- Independent increment: $\forall n, \forall t_0 \leq t_1 \leq \cdots \leq t_n, X_{t_1} - X_{t_0} \perp \cdots \perp X_{t_n} - X_{t_{n-1}}$
- Stationary increment: $\forall s \leq t, X_t - X_s \simeq X_{t-s} - X_0$
- Continuity in probability: $\lim_{t \downarrow 0} X_t \to X_0 = 0$ (i.p.)

Consequence of independent and stationary increment:

$$X_t = \sum_{i=1}^{n} \underbrace{[X_{it/n} - X_{(i-1)t/n}]}_{i.i.d. \sim X_{t/n}}$$

i.e., X_t is infinitely divisible. Continuity forces $X_t \simeq F^{(t)}$ for some distribution F.

Lévy-Khintchine Formula

Theorem: Lévy process representation

X_t is a Lévy process iff

$$\mathbb{E}\exp(iu\mathsf{X}_t) = \exp\left\{t\Big[\underbrace{iub}_{①}\underbrace{-\sigma^2u^2/2}_{②} + \int\underbrace{\left(e^{iux} - 1 - iux\,[\![x \le 1]\!]\right)}_{③}\mathrm{d}\nu(x)\Big]\right\},$$

where ν is a measure with $\nu(\{0\}) = 0$ and $\int (1 \wedge x^2)\,\mathrm{d}\nu(x) < \infty$.

① : Deterministic process $\mathsf{X}_t = bt$

② : Brownian motion $\sigma\mathsf{B}_t$

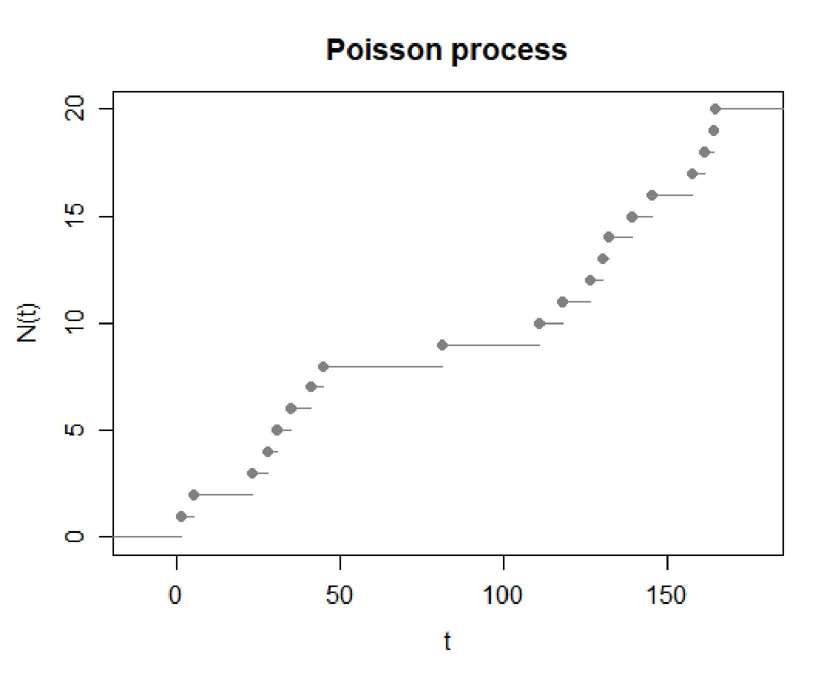
③ : Purely jump process

Poisson Process

A stochastic process N_t is called a Poisson process if

- Initialization: $\mathsf{N}_0 \equiv 0$
- Independent increment: $\forall n, \forall t_0 \leq t_1 \leq \cdots \leq t_n$, $\mathsf{N}_{t_1} - \mathsf{N}_{t_0} \perp \cdots \perp \mathsf{N}_{t_n} - \mathsf{N}_{t_{n-1}}$
- Stationary increment: $\forall s \leq t$, $\mathsf{N}_t - \mathsf{N}_s \simeq \mathsf{N}_{t-s} - \mathsf{N}_0$
- Poisson: $\mathsf{N}_t \simeq \mathrm{Pois}(\lambda t)$
- Right continuity: for (almost) all ω, $t \mapsto \mathsf{N}_t(\omega)$ is right continuous with left limit

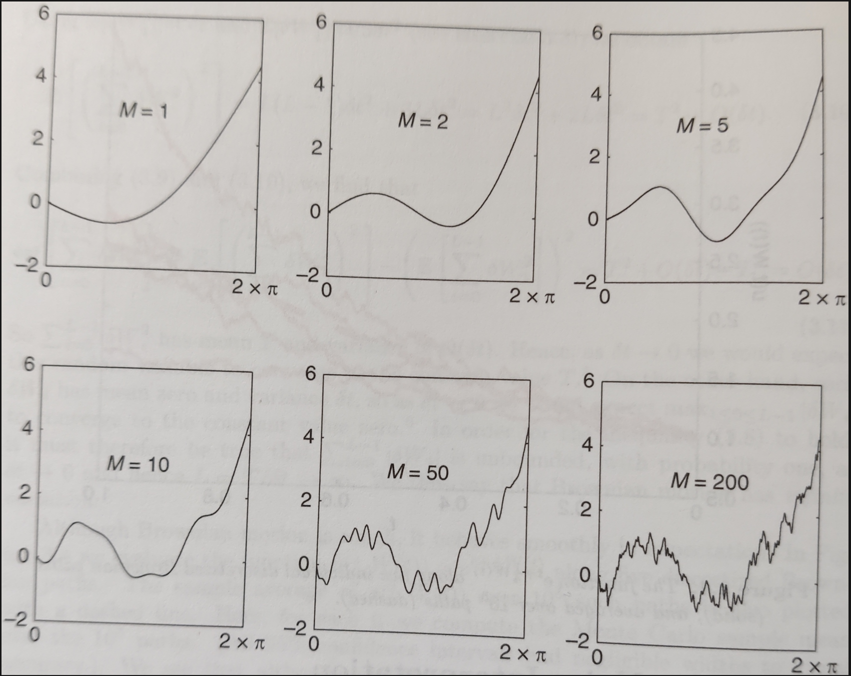
$\mathsf{N}_t \in \mathbb{Z}_+$, increasing, finitely many jumps of size 1 in finite time


Poisson process
N(t)
t
0
5
10
15
20
0
50
100
150

Wiener's Construction of Brownian Motion

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

- Trigonometric functions $\varphi_n(t) := \exp(in\pi t)$ as orthogomal basis in $L^2([0,1])$
- $\mathsf{G}_n \overset{i.i.d.}{\simeq} \mathcal{N}(0,1)$
- Truncating n leads to a discretized path


M = 1
M = 2
M = 5
M = 10
M = 50
M = 200
6
4
2
0
−2
0
2 × π

Ciesielski's Construction of Brownian Motion

- Haar wavelets: $\varphi_0(t) \equiv 1$, for $n \in \mathbb{N}$, $k = 1, 3, \ldots, 2^n - 1$,

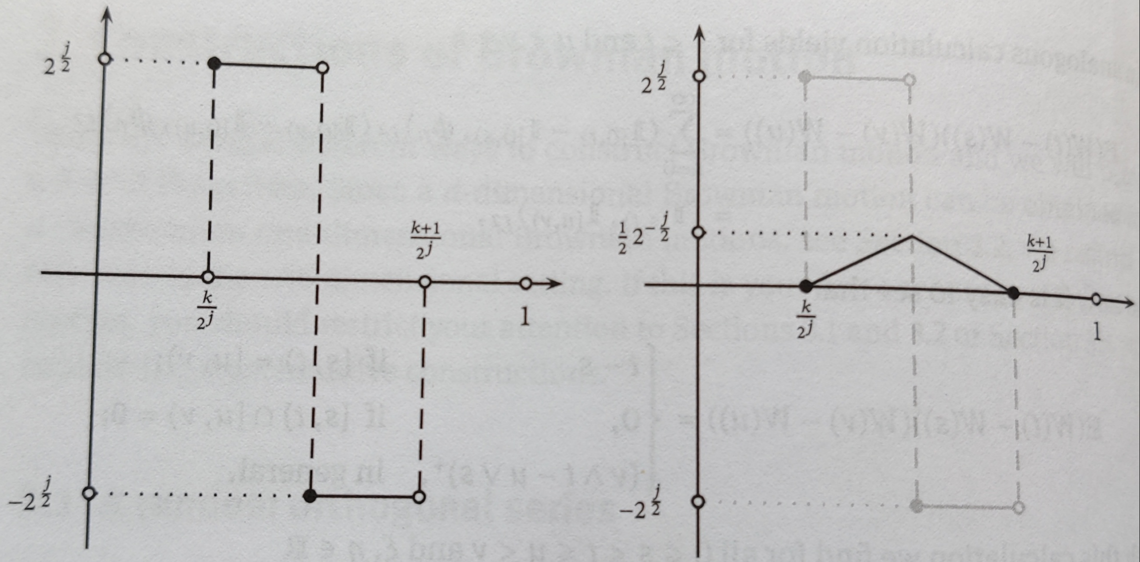
$$\varphi_{k/2^n}(t) = 2^{(n-1)/2} \cdot \Big([\![k - 1 < t2^n \leq k]\!] - [\![k < t2^n \leq k + 1]\!] \Big)$$

- Expand B'_t over the Haar wavelets:

$$\int_0^1 \mathsf{B}'_t \cdot \varphi_{k/2^n}(t)\,\mathrm{d}t = 2^{(n-1)/2} \cdot [(\mathsf{B}_{k/2^n} - \mathsf{B}_{(k-1)/2^n}) - (\mathsf{B}_{(k+1)/2^n} - \mathsf{B}_{k/2^n})] \simeq \mathsf{G}_{k/2^n}$$

- Reconstruct $\mathsf{B}'_t = \mathsf{G}_0\varphi_0 + \sum_n \mathsf{G}_{k/2^n} \cdot \varphi_{k/2^n}(t)$ and thus

$$\boxed{\mathsf{B}_t = \int_0^1 \mathsf{B}'_t\,\mathrm{d}t = t\mathsf{G}_0 + \sum_n \mathsf{G}_{k/2^n} \int_0^t \varphi_{k/2^n}(s)\,\mathrm{d}s}$$



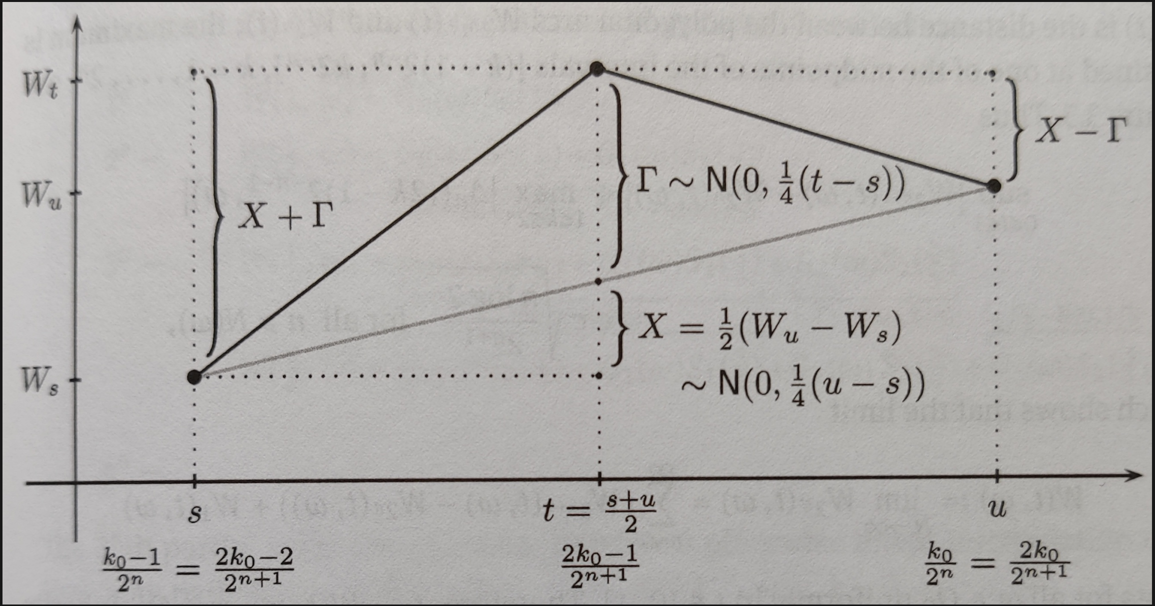
$2^{\frac{j}{2}}$
$-2^{\frac{j}{2}}$
$\frac{k}{2^j}$
$\frac{k+1}{2^j}$
1
$2^{\frac{j}{2}}$
$\frac{1}{2}2^{-\frac{j}{2}}$
$-2^{\frac{j}{2}}$
$\frac{k}{2^j}$
$\frac{k+1}{2^j}$
1

Lévy's Construction of Brownian Motion

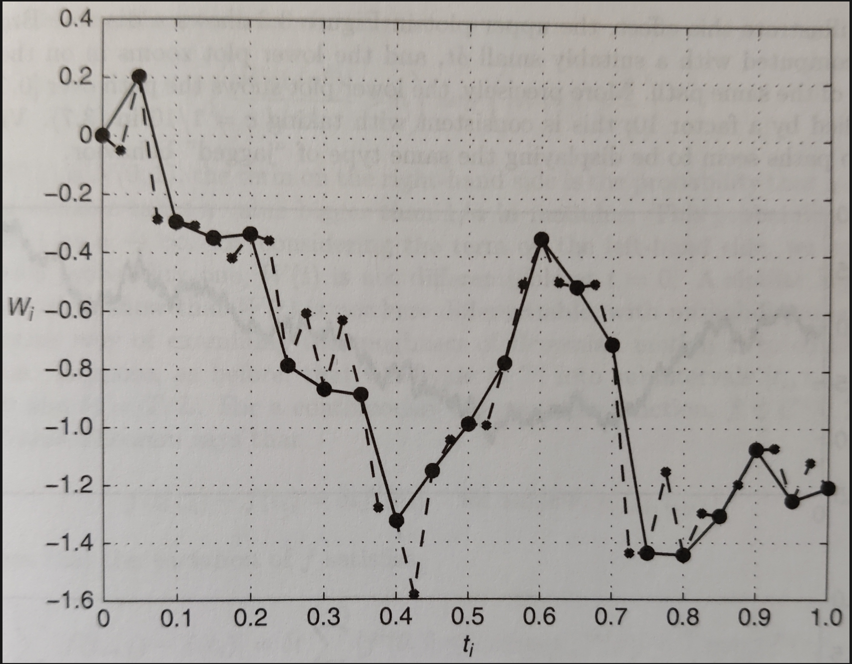
- Initialize $\mathsf{B}_0 = 0, \mathsf{B}_1 \simeq \mathcal{N}(0, 1)$
- Repeat for each $n = 0, 1, 2, \ldots$, $l = 1, 2, \ldots, 2^{n+1} - 1$

$$\mathsf{B}_{l/2^{n+1}} = \begin{cases} \mathsf{B}_{k/2^n}, & l = 2k \\ \frac{1}{2}[\mathsf{B}_{k/2^n} + \mathsf{B}_{(k+1)/2^n}] + 2^{-(n+2)/2}\mathsf{G}_{k/2^n}, & l = 2k + 1 \end{cases}$$

  - refine the grid by appending each middle point
  - linearly interpolate at the middle point
  - add scaled, independent, Gaussian perturbation to the middle point



W_t
W_u
W_s
$X + \Gamma$
$\Gamma \sim \mathsf{N}(0, \frac{1}{4}(t-s))$
$X - \Gamma$
$X = \frac{1}{2}(W_u - W_s)$
$\sim \mathsf{N}(0, \frac{1}{4}(u-s))$
s
$t = \frac{s+u}{2}$
u
$\frac{k_0-1}{2^n} = \frac{2k_0-2}{2^{n+1}}$
$\frac{2k_0-1}{2^{n+1}}$
$\frac{k_0}{2^n} = \frac{2k_0}{2^{n+1}}$


0.4
0.2
0
−0.2
−0.4
W_i
−0.6
−0.8
−1.0
−1.2
−1.4
−1.6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
t_i

Donsker's Construction of Brownian Motion

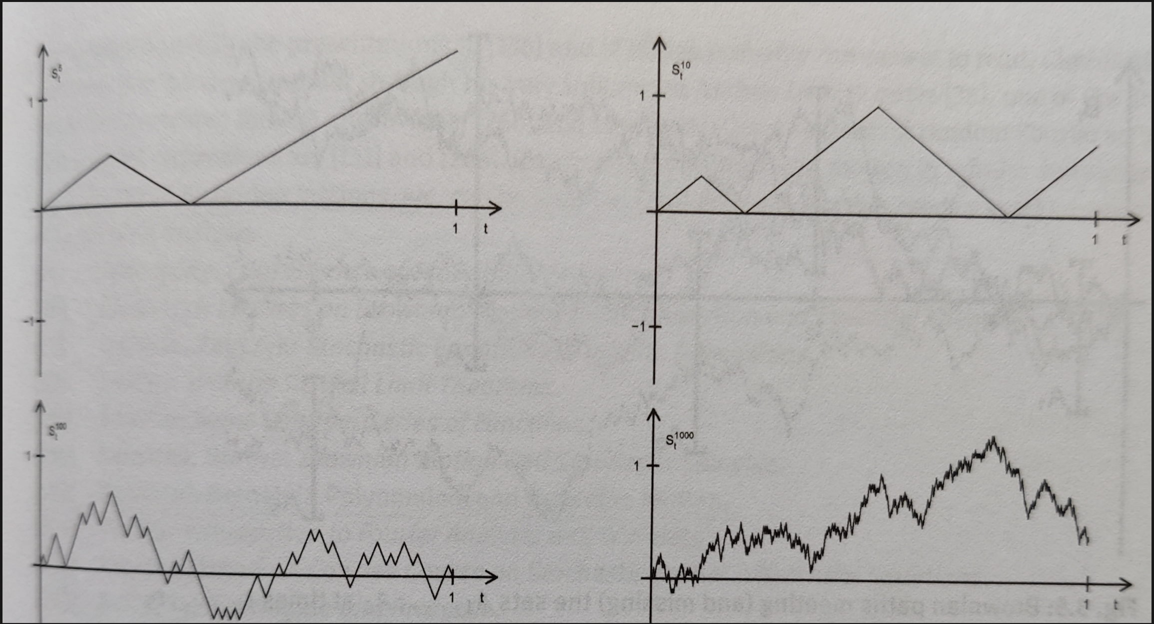
Theorem: Convergence in distribution

Suppose X_t^n converges to X_t for any finite section (i.e., for finitely many t), $\mathsf{X}_1 - \mathsf{X}_{1-\delta} \Rightarrow 0$ as $\delta \to 0$, and for any $r \le s \le t$ and $\lambda > 0$,

$$\Pr[|\mathsf{X}_s^n - \mathsf{X}_r^n| \wedge |\mathsf{X}_t^n - \mathsf{X}_s^n| \ge \lambda] \le \frac{1}{\lambda^{4\beta}}[h(t) - h(r)]^{2\alpha},$$

where $\beta \ge 0$, $\alpha > \frac{1}{2}$ and h is increasing continuous. Then, $\mathsf{X}^n \Rightarrow \mathsf{X}$.

- Let $\xi_i \overset{i.i.d.}{\simeq} F$ with 0 mean and unit variance
- Let $S_n = \sum_{i=1}^n \xi_i$ be the cumsum
- $\mathsf{X}_t^n := \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} \Rightarrow \mathsf{B}_t$ and $\tilde{\mathsf{X}}_t^n := \mathsf{X}_t^n + (nt - \lfloor nt \rfloor) \frac{1}{\sqrt{n}} \xi_{\lfloor nt \rfloor + 1} \Rightarrow \mathsf{B}_t$
- $\hat{Q}_t^n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \left([\![\xi_i \le t]\!] - F(t) \right) \Rightarrow \mathsf{B}_{F(t)}^\circ$; $\sup_t |\hat{Q}_t^n| \to \sup_t |\mathsf{B}_{F(t)}^\circ|$


S_t^5
1
-1
1
t
S_t^{10}
1
-1
1
t
S_t^{100}
1
1
t
S_t^{1000}
1
1
t

Integration

- Let $g : [0, T] \to \mathbb{R}$ be of bounded variation (e.g., continuously differentiable)
- $g(0) = g(T) = 0$
- Define the integral through integration by parts:

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

  - the rhs exists if $t \mapsto \mathsf{X}_t$ is continuous, a.k.a. Riemann-Stieltjes-integral

- What about $\int_0^T \mathsf{B}_t\, \mathrm{d}\mathsf{B}_t$?
  - need significantly new ideas