Diffusion processes

with applications in diffusion networks

W. Genuist[1,2], É. Savin[1,3]
{wilfried.genuist,eric.savin}@centralesupelec.fr

[1]Computer Science Dept.
ONERA, France

[2]Laboratoire de Mécanique Paris-Saclay
CentraleSupélec, France

[3]Mechanical and Environmental Engineering Dept.
CentraleSupélec, France

December 21, 2024

Outline

1. Stochastic differential equations (SDE)
   - First-order stochastic systems driven by noise
   - Stochastic integrals
   - Diffusion processes
2. Numerical simulations of SDE
   - Stochastic modeling with SDE
   - Numerical schemes
3. Generative Adversarial Networks (GAN)
4. Variational Auto-Encoders (VAE)
5. Diffusion models in ML
   - Continuous setting
   - Denoising Diffusion Probabilistic Modeling (DDPM)
   - Score Matching with Langevin Dynamics (SMLD)
   - Guidance
   - Consistency models

Outline

First-order stochastic differential equation

A general first-order stochastic differential equation for the process $\boldsymbol{U}$ indexed on $\mathbb{R}_+$ with values in $\mathbb{R}^q$:

$$\dot{\boldsymbol{U}}(t) = \underline{\boldsymbol{b}}(\boldsymbol{U},t) + \boldsymbol{\sigma}(\boldsymbol{U},t)\boldsymbol{F}(t)\,, \quad \boldsymbol{U}(0) = \boldsymbol{U}_0\,,$$

with the data:

- $\boldsymbol{u},t \mapsto \underline{\boldsymbol{b}}(\boldsymbol{u},t) : \mathbb{R}^q \times \mathbb{R}_+ \to \mathbb{R}^q$ the *drift* function;
- $\boldsymbol{u},t \mapsto \boldsymbol{\sigma}(\boldsymbol{u},t) : \mathbb{R}^q \times \mathbb{R}_+ \to \mathbb{R}^{q\times p}$ the *diffusion* (or *scattering*) operator;
- $\boldsymbol{U}_0$ is an r.v. in $\mathbb{R}^q$ with known marginal PDF $\pi_0(\boldsymbol{u}_0)$;
- $\boldsymbol{F}(t) = (F_1(t),\dots F_p(t))$ is a second-order Gaussian random process indexed on $\mathbb{R}$ with values in $\mathbb{R}^p$, also centered, stationary, such that $F_1(t),\dots F_p(t)$ are mutually independent and mean-square continuous, with:

$$\boldsymbol{S}_{\boldsymbol{F}}(\omega) = S_0 \mathbf{1}_{[-B,B]}(\omega)[\boldsymbol{I}_p]\,, \quad S_0 > 0\,, \quad B > 0\,.$$

First-order systems driven by noise

- $B < +\infty$: colored noise, hot topic!

$$\boldsymbol{U}(t) = \boldsymbol{U}_0 + \int_0^t \underline{\boldsymbol{b}}(\boldsymbol{U}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s)\boldsymbol{F}(s)\mathrm{d}s\,.$$

- $B \to +\infty$: $\boldsymbol{F} \to \dot{\boldsymbol{W}}$ the *normalized Gaussian white noise*, and the solution of the first-oder SDE holds as a "stochastic integral":

$$\boldsymbol{U}(t) = \boldsymbol{U}_0 + \int_0^t \underline{\boldsymbol{b}}(\boldsymbol{U}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s) \circ \mathrm{d}\boldsymbol{W}(s)\,.$$

- *Causality*: the family of r.v. $\{\boldsymbol{U}(s), 0 \leqslant s \leqslant t\}$ is independent of the family of r.v. $\{\boldsymbol{F}(\tau), \tau > t\}$ or $\{\mathrm{d}\boldsymbol{W}(\tau), \tau > t\}$.

Definition

- *The normalized Gaussian white noise $\boldsymbol{B}(t) \equiv \dot{\boldsymbol{W}}(t)$ with values in $\mathbb{R}^p$ is the Gaussian stochastic process indexed on $\mathbb{R}$, centered, stationary, with the spectral density matrix:*

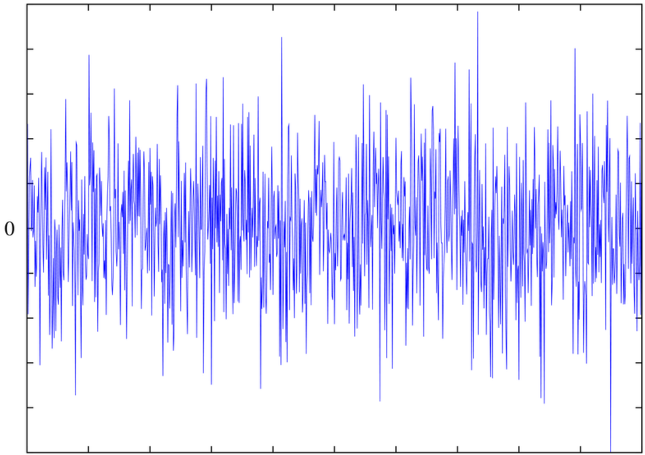
$$\boldsymbol{S}_{\boldsymbol{B}}(\omega) = \frac{1}{2\pi}\boldsymbol{I}_p \,.$$

- *Since $B_1(t), \dots B_p(t)$ are uncorrelated and jointly Gaussian, they are mutually independent.*
- *$\boldsymbol{B}$ is not second order $|||\boldsymbol{B}(t)|||^2 = \int \mathrm{Tr}\, \boldsymbol{S}_{\boldsymbol{B}}(\omega)\mathrm{d}\omega = +\infty$.*

This definition holds in the sense of generalized stochastic processes $\varphi \mapsto \boldsymbol{B}(\varphi) : \mathscr{D}(T) \to L^2(\Omega, \mathbb{R}^p)$ where $\mathscr{D}(T)$ is the set of $\mathscr{C}^\infty$ functions having a compact support within $T \subseteq \mathbb{R}$.

White noise

Definition



White noise.

Wiener process

Definition

The white noise is the (generalized) derivative of the Wiener process, or *Brownian motion*.

Definition

The (normalized) Wiener process $\boldsymbol{W}(t)$ *with values in* $\mathbb{R}^p$ *is the stochastic process indexed on* $\mathbb{R}_+$*, such that:*

- $W_1(t), \dots W_p(t)$ *are mutually independent;*
- $\boldsymbol{W}(0) = \boldsymbol{0}$ *almost surely (a.s.);*
- *If* $0 \leqslant s < t < +\infty$ *let* $\Delta\boldsymbol{W}(s,t) = \boldsymbol{W}(t) - \boldsymbol{W}(s)$*, then:*
  - $\forall m$ *and* $0 < t_1 < t_2 < \cdots < t_m < +\infty$, $\boldsymbol{W}(0)$, $\Delta\boldsymbol{W}(0,t_1)$, $\Delta\boldsymbol{W}(t_1,t_2)$, ... $\Delta\boldsymbol{W}(t_{m-1},t_m)$ *are mutually independent r.v. (independent increments);*
  - $\Delta\boldsymbol{W}(s,t)$ *is a Gaussian, centered, second-order r.v. with* $\boldsymbol{C}_{\Delta\boldsymbol{W}}(s,t) = (t-s)\boldsymbol{I}_p$.

Wiener process

Characterization

Consequently it can be shown that:

- $\boldsymbol{W}(t)$ is a second-order Gaussian, centered, mean-square continuous, non stationary stochastic process;
- the covariance and conditional PDF for $0 \leqslant t, s < +\infty$:

$$\boldsymbol{C}_{\boldsymbol{W}}(t,s) = \mathrm{Min}(t,s)\boldsymbol{I}_p \,,$$

$$\overline{\pi}(\boldsymbol{v}'; t+s | \boldsymbol{v}; t) = (2\pi s)^{-\frac{p}{2}} \mathrm{e}^{-\frac{\|\boldsymbol{v}'-\boldsymbol{v}\|^2}{2s}} \,;$$

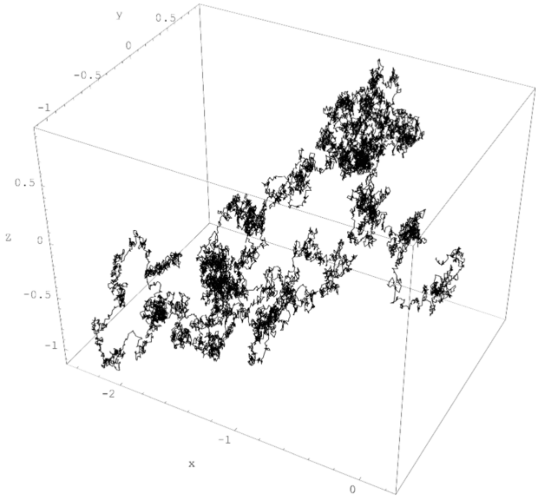
- $\boldsymbol{W}(t)$ has a.s. continuous sample paths;
- sample paths $t \mapsto \boldsymbol{W}(t,\theta)$, $\theta \in \Omega_\theta$, are non differentiable a.s.

As a generalized derivative with $\mathrm{d}\boldsymbol{W} = (\mathrm{d}W_1, \dots \mathrm{d}W_p)$:

$$\mathrm{d}\boldsymbol{W}(\varphi) = \boldsymbol{B}(-\dot{\varphi}) \,, \quad \forall \varphi \in \mathscr{D}(\mathbb{R}) \,.$$

Wiener process

Characterization



Wiener process in $\mathbb{R}^3$.

- Let $\boldsymbol{X}(t)$ be a stochastic process indexed by $\mathbb{R}_+$ with a.s. continuous sample paths.
- Assume the r.v. $\{\boldsymbol{X}(s), 0 \leqslant s \leqslant t\}$ are independent of the r.v. $\{\Delta\boldsymbol{W}(t,\tau), \tau > t\}$: a *non anticipative* process, then

$$\int_0^t \boldsymbol{X}(s)\mathrm{d}_\lambda \boldsymbol{W}(s)$$

$$= \underset{K\to+\infty}{\mathrm{l.\,i.\,p.}} \sum_{k=1}^{K} \left[(1-\lambda)\boldsymbol{X}(t_k) + \lambda\boldsymbol{X}(t_{k+1})\right] \Delta\boldsymbol{W}(t_k, t_{k+1})\,,$$

for any partition $0 = t_1 < t_2 < \cdots < t_{K+1} = t$ of $[0,t]$ with $\max_{1\leqslant k\leqslant K}(t_{k+1}-t_k) \underset{K\to+\infty}{\longrightarrow} 0$.

Stochastic integrals

Application to stochastic differential calculus

- A simple example–remind $\Delta W \propto \Delta t^{\frac{1}{2}}$ for the real-valued Wiener process W:

$$\int_0^t W(s)\mathrm{d}_\lambda W(s) = \frac{1}{2}W(t)^2 + \left(\lambda - \frac{1}{2}\right)t\,,$$

from which one deduces the stochastic differential:

$$\mathrm{d}_\lambda(W(t)^2) = 2W(t)\mathrm{d}_\lambda W(t) + (1-2\lambda)\mathrm{d}t\,.$$

- More generally ($\lambda = 0$ is called the *Itō formula*):

$$\mathrm{d}_\lambda(f(W(t)) = f'(W(t))\mathrm{d}_\lambda W(t) + \left(\frac{1}{2} - \lambda\right) f''(W(t))\mathrm{d}t\,.$$

Setting $W(t_k) = W_k$:

$$\begin{aligned}
\int_0^t W(s)\mathrm{d}_\lambda W(s) &= \operatorname*{l.\,i.\,p.}_{K\to+\infty} \sum_{k=1}^{K} \left[(1-\lambda)W_k + \lambda W_{k+1}\right](W_{k+1} - W_k) \\
&= \operatorname*{l.\,i.\,p.}_{K\to+\infty} \sum_{k=1}^{K} \left[(1-2\lambda)W_k W_{k+1} - \lambda W_{k+1}^2 - (1-\lambda)W_k^2\right] \\
&= \operatorname*{l.\,i.\,p.}_{K\to+\infty} \sum_{k=1}^{K} \frac{1}{2}\left[(2\lambda-1)(W_{k+1} - W_k)^2 + W_{k+1}^2 - W_k^2\right] \\
&= \frac{1}{2}(2\lambda-1) \operatorname*{l.\,i.\,p.}_{K\to+\infty} \sum_{k=1}^{K} (W_{k+1} - W_k)^2 + \frac{1}{2}(W_{K+1}^2 - W_1^2) \\
&= \frac{1}{2}(2\lambda-1)t + \frac{1}{2}W(t)^2\,,
\end{aligned}$$

since $W_1 = W(0) = 0$ and $W_{K+1} = W(t)$. Besides:

$$\mathbb{E}\left\{\int_0^t W(s)\mathrm{d}_\lambda W(s)\right\} = \lambda t\,.$$

- If $\lambda = \frac{1}{2}$ the usual differential calculus applies, and the solution of SDE holds as a *Stratonovich integral* (1966):

$$\boldsymbol{U}(t) = \boldsymbol{U}_0 + \int_0^t \underline{\boldsymbol{b}}(\boldsymbol{U}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s) \circ \mathrm{d}\boldsymbol{W}(s)\,.$$

- If $\lambda = 0$, its solution holds as an *Itō integral* (1944):

$$\boldsymbol{U}(t) = \boldsymbol{U}_0 + \int_0^t \boldsymbol{b}(\boldsymbol{U}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s)\mathrm{d}\boldsymbol{W}(s)\,,$$

where:

$$\boldsymbol{b} = \underline{\boldsymbol{b}} + \frac{1}{2}(\boldsymbol{D}_{\boldsymbol{u}}\boldsymbol{\sigma})\boldsymbol{\sigma}^{\mathsf{T}}\,.$$

- $\boldsymbol{U}(t)$ is a *Markov process.*

- Let $\boldsymbol{U}(t) \in \mathbb{R}^q$ be the solution of the ISDE:

$$\boldsymbol{U}(t) = \boldsymbol{U}_0 + \int_0^t \boldsymbol{b}(\boldsymbol{U}(s), s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s)\mathrm{d}\boldsymbol{W}(s)\,.$$

- Let $\phi : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$ be a smooth function. Then *Itō's formula* states that:

$$\begin{aligned}\phi(\boldsymbol{U}(t), t) = \phi(\boldsymbol{U}_0, 0) &+ \int_0^t \frac{\partial \phi}{\partial t}(\boldsymbol{U}(s), s)\mathrm{d}s \\ &+ \int_0^t \boldsymbol{\nabla}_{\boldsymbol{u}}\phi(\boldsymbol{U}(s), s) \cdot \mathrm{d}\boldsymbol{U}(s) \\ +\frac{1}{2}\int_0^t \boldsymbol{\sigma}(\boldsymbol{U}(s), s)\boldsymbol{\sigma}(\boldsymbol{U}(s), s)^{\mathsf{T}} &: \boldsymbol{\nabla}_{\boldsymbol{u}} \otimes \boldsymbol{\nabla}_{\boldsymbol{u}}\phi(\boldsymbol{U}(s), s)\mathrm{d}s\,,\end{aligned}$$

where $\mathrm{d}\boldsymbol{U}(s) = \boldsymbol{b}(\boldsymbol{U}(s), s)\mathrm{d}s + \boldsymbol{\sigma}(\boldsymbol{U}(s), s)\mathrm{d}\boldsymbol{W}(s)$.

Definition

The conditional probability given $t_0 < \cdots < t_m < t$:

$$\overline{\pi}(\boldsymbol{u};t|\boldsymbol{u}_0,\dots\boldsymbol{u}_m;t_0,\dots t_m) = \frac{\pi(\boldsymbol{u}_0,\dots\boldsymbol{u}_m,\boldsymbol{u};t_0,\dots t_m,t)}{\pi(\boldsymbol{u}_0,\dots\boldsymbol{u}_m;t_0,\dots t_m)}\,.$$

Definition

Let $\boldsymbol{U}(t)$ be a stochastic process defined on $(\Omega,\mathcal{E},P)$ and indexed on $\mathbb{R}_+$ with values in $\mathbb{R}^q$. It is a Markov process if:

- *for all $0 \leqslant t_1 < \cdots < t_m < t$ and $\boldsymbol{u}_1,\dots\boldsymbol{u}_m,\boldsymbol{u}$ in $\mathbb{R}^q$*

  $$\overline{\pi}(\boldsymbol{u};t|\boldsymbol{u}_0,\dots\boldsymbol{u}_m;t_0,\dots t_m) = \overline{\pi}(\boldsymbol{u};t|\boldsymbol{u}_m;t_m)\,;$$

- *the marginal PDF $\pi_0(\boldsymbol{u}_0)$ of $\boldsymbol{U}(0)$ can be any PDF.*

SDE

W. Genuist, É. Savin

SDE
First-order systems
Stochastic integrals
Diffusion processes
Numerical solutions
Stochastic modeling
Numerical schemes
GAN
VAE
Diffusion & ML
Continuous case
DDPM
SMLD
Guidance
CM

- A Markov process is fully characterized by:
  - its marginal PDF $\pi(\boldsymbol{u};t)$,
  - and its *transition PDF* $\overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s)$, $0 \leqslant s < t < +\infty$, with

$$\pi(\boldsymbol{u};t) = \int_{\mathbb{R}^q} \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s)\pi(\boldsymbol{v};s)\mathrm{d}\boldsymbol{v}\,.$$

- $\overline{\pi}$ satisfies the *Chapman-Kolmogorov equation*:

$$\overline{\pi}(\boldsymbol{u};t|\boldsymbol{u}';t') = \int_{\mathbb{R}^q} \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s)\overline{\pi}(\boldsymbol{v};s|\boldsymbol{u}';t')\mathrm{d}\boldsymbol{v}\,, \quad t' < s < t\,.$$

- Homogeneous Markov process:

$$\overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) = \overline{\pi}(\boldsymbol{u};t-s|\boldsymbol{v};0)\,, \quad 0 \leqslant s < t < +\infty\,.$$

- The Brownian motion is a Markov process.

Diffusion processes

Definition

Definition

The $\mathbb{R}^q$–valued continuous-time Markov process $\boldsymbol{U}(t)$ with a.s. continuous sample paths and transition PDF $\overline{\pi}(\boldsymbol{v};s|\boldsymbol{u};t)$ is a diffusion process if $\forall\epsilon>0$ (but not necessarily small), $\forall\boldsymbol{u}\in\mathbb{R}^q$ the first moments of its increments are such that for $h>0$:

$$\int_{\|\boldsymbol{v}-\boldsymbol{u}\|\geqslant\epsilon}\overline{\pi}(\mathrm{d}\boldsymbol{v};t+h|\boldsymbol{u};t)=\mathrm{o}(h)\,,$$

$$\int_{\|\boldsymbol{v}-\boldsymbol{u}\|<\epsilon}(\boldsymbol{v}-\boldsymbol{u})\overline{\pi}(\mathrm{d}\boldsymbol{v};t+h|\boldsymbol{u};t)=h\boldsymbol{b}(\boldsymbol{u},t)+\mathrm{o}(h)\,,$$

$$\int_{\|\boldsymbol{v}-\boldsymbol{u}\|<\epsilon}(\boldsymbol{v}-\boldsymbol{u})\otimes(\boldsymbol{v}-\boldsymbol{u})\overline{\pi}(\mathrm{d}\boldsymbol{v};t+h|\boldsymbol{u};t)=h\boldsymbol{a}(\boldsymbol{u},t)+\mathrm{o}(h)\,,$$

where $\boldsymbol{b}\in\mathbb{R}^q$ and $\boldsymbol{a}\in\mathbb{R}^{q\times q}$ symmetric, positive.

Diffusion processes

Interpretation

- *Continuity*: particles moving on sample paths of a diffusion process only make small jumps, or the probability of moving a distance ϵ goes to zero as h goes to zero no matter how small ϵ is.
- *Drift*: those particles can have a net mean velocity $\boldsymbol{b}$.
- *Diffusion*: particles spread as time increases with the rate $\operatorname{Tr}\boldsymbol{a}$. Entropy increases while the phase space contracts, thus some information (energy) gets lost.

$$\boldsymbol{U}(t+h)-\boldsymbol{U}(t)\approx h\boldsymbol{b}(\boldsymbol{U}(t),t)+\boldsymbol{a}^{\frac{1}{2}}(\boldsymbol{U}(t),t)\Delta\boldsymbol{W}(t,t+h)\,.$$

Diffusion processes

Infinitesimal generator

- *Infinitesimal generator*: for a time-continuous process $\boldsymbol{U}(t)$ and a suitably regular function ϕ, the infinitesimal generator $\mathcal{A}_t$ is

$$\mathcal{A}_t\phi(\boldsymbol{u},t) = \lim_{h\downarrow 0}\frac{\mathbb{E}\{\phi(\boldsymbol{U}(t+h),t+h)|\boldsymbol{U}(t)=\boldsymbol{u}\} - \phi(\boldsymbol{u},t)}{h}\,.$$

- For a diffusion process:

$$\mathcal{A}_t\phi(\boldsymbol{u},t) = \partial_t\phi + \boldsymbol{b}(\boldsymbol{u},t)\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\phi + \frac{1}{2}\boldsymbol{a}(\boldsymbol{u},t):\boldsymbol{\nabla}_{\boldsymbol{u}}\otimes\boldsymbol{\nabla}_{\boldsymbol{u}}\phi\,,$$

with (formal) adjoint operator:

$$\mathcal{A}_t^*\phi(\boldsymbol{u},t) = -\partial_t\phi - \boldsymbol{\nabla}_{\boldsymbol{u}}\cdot(\boldsymbol{b}(\boldsymbol{u},t)\phi) + \frac{1}{2}\boldsymbol{\nabla}_{\boldsymbol{u}}\otimes\boldsymbol{\nabla}_{\boldsymbol{u}}:(\boldsymbol{a}(\boldsymbol{u},t)\phi)\,.$$

Fokker-Planck equation

The marginal PDF π and transition PDF $\overline{\pi}$ of a diffusion process satisfy the *Fokker-Planck equation* $\mathcal{A}_t^*\pi = 0$, or:

$$\partial_t \pi + \boldsymbol{\nabla}_{\boldsymbol{u}} \cdot \left(\pi \boldsymbol{b} - \frac{1}{2} \boldsymbol{\nabla}_{\boldsymbol{u}} \cdot (\pi \boldsymbol{a}) \right) = 0 \,,$$

with $\pi(\boldsymbol{u}_0;0) = \pi_0(\boldsymbol{u}_0)$ and $\lim_{h\downarrow 0} \overline{\pi}(\boldsymbol{u};t+h|\boldsymbol{v};t) = \delta(\boldsymbol{u}-\boldsymbol{v})$.

$$\begin{aligned}
\int_{\mathbb{R}^q} f(\boldsymbol{u}) \partial_t \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) \mathrm{d}\boldsymbol{u} &= \lim_{h\downarrow 0} \frac{1}{h} \int_{\mathbb{R}^q} f(\boldsymbol{u}) \left(\overline{\pi}(\boldsymbol{u};t+h|\boldsymbol{v};s) - \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) \right) \mathrm{d}\boldsymbol{u} \\
&= \lim_{h\downarrow 0} \frac{1}{h} \int_{\mathbb{R}^q} \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) \left[\int_{\mathbb{R}^q} f(\boldsymbol{u}') \overline{\pi}(\boldsymbol{u}';t+h|\boldsymbol{u};t) \mathrm{d}\boldsymbol{u}' - f(\boldsymbol{u}) \right] \mathrm{d}\boldsymbol{u} \quad \text{(C-K)} \\
&= \lim_{h\downarrow 0} \frac{1}{h} \int_{\mathbb{R}^q} \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) \int_{\mathbb{R}^q} \left(f(\boldsymbol{u}') - f(\boldsymbol{u}) \right) \overline{\pi}(\boldsymbol{u}';t+h|\boldsymbol{u};t) \mathrm{d}\boldsymbol{u}' \mathrm{d}\boldsymbol{u} \quad \text{(norm.)} \\
&= \lim_{h\downarrow 0} \frac{1}{h} \int_{\mathbb{R}^q} \overline{\pi}(\boldsymbol{u};t|\boldsymbol{v};s) \int_{\|\boldsymbol{u}'-\boldsymbol{u}\|<\epsilon} \left(f(\boldsymbol{u}') - f(\boldsymbol{u}) \right) \overline{\pi}(\boldsymbol{u}';t+h|\boldsymbol{u};t) \mathrm{d}\boldsymbol{u}' \mathrm{d}\boldsymbol{u} \,, \quad \forall f \in \mathscr{C}_0^2 \,.
\end{aligned}$$

Then use a Taylor expansion for f, definitions of drift and diffusion, and integrate by parts.

Backward Kolmogorov equation

The transition PDF $\overline{\pi}(\cdot|\boldsymbol{v};s)$ of a diffusion process also satisfies the *Backward Kolmogorov equation* $\mathcal{A}_s\overline{\pi}=0$, or:

$$\begin{aligned}\partial_s\overline{\pi}(\cdot|\boldsymbol{v};s)+\boldsymbol{b}(\boldsymbol{v},s)\cdot\boldsymbol{\nabla}_{\boldsymbol{v}}\overline{\pi}(\cdot|\boldsymbol{v};s)&\\+\frac{1}{2}\boldsymbol{a}(\boldsymbol{v},s):\boldsymbol{\nabla}_{\boldsymbol{v}}\otimes\boldsymbol{\nabla}_{\boldsymbol{v}}\overline{\pi}(\cdot|\boldsymbol{v};s)&=0\,,\end{aligned}$$

with $\lim_{h\downarrow 0}\overline{\pi}(\boldsymbol{u};s|\boldsymbol{v};s-h)=\delta(\boldsymbol{v}-\boldsymbol{u})$.

Itō's stochastic differential equations (ISDE)

Solutions

$$\mathrm{d}\boldsymbol{U} = \boldsymbol{b}(\boldsymbol{U},t)\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{U},t)\mathrm{d}\boldsymbol{W}\,,\quad \boldsymbol{U}(0) = \boldsymbol{U}_0\,,$$

with the regularity assumptions:

$$\begin{aligned}\|\boldsymbol{b}(\boldsymbol{u},t)\| + \|\boldsymbol{\sigma}(\boldsymbol{u},t)\| &\leqslant K(1+\|\boldsymbol{u}\|)\,,\\ \left\|\boldsymbol{b}(\boldsymbol{u}',t)-\boldsymbol{b}(\boldsymbol{u},t)\right\| + \left\|\boldsymbol{\sigma}(\boldsymbol{u}',t)-\boldsymbol{\sigma}(\boldsymbol{u},t)\right\| &\leqslant K\left\|\boldsymbol{u}'-\boldsymbol{u}\right\|\,.\end{aligned}$$

1. Then the SDE has a unique solution, with a.s. continuous sample paths. If in addition $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ are independent of t, $\boldsymbol{U}(t)$ is homogeneous.
2. If $t \mapsto \boldsymbol{b}(\boldsymbol{u},t) \in \mathbb{R}^q$ and $t \mapsto \boldsymbol{\sigma}(\boldsymbol{u},t) \in \mathbb{R}^{q\times p}$ are continuous, $\boldsymbol{U}(t)$ is also a diffusion process with $\boldsymbol{a} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}}$.

Itō's stochastic differential equations (ISDE)

Example: Black-Scholes[1] model (or geometric Brownian motion)

- The relative variation of a stock $U(t)$ with constant (annualized) drift rate μ and volatility σ:

$$\frac{\mathrm{d}U}{U} = \mu \mathrm{d}t + \sigma \mathrm{d}W \,, \quad U(0) = U_0 \,.$$

- Transformation to a Stratonovich SDE:

$$\frac{\mathrm{d}U}{U} = \left(\mu - \frac{\sigma^2}{2}\right) \mathrm{d}t + \sigma \circ \mathrm{d}W \,, \quad U(0) = U_0 \,,$$

for which "normal rules of integration" apply:

$$U(t) = U_0 \mathrm{e}^{(\mu - \frac{\sigma^2}{2})t + \sigma W(t)} \,.$$

- The Fokker-Planck equation:

$$\partial_t \pi + \mu \partial_u (\pi u) - \frac{\sigma^2}{2} \partial_u^2 (\pi u^2) = 0 \,, \quad \pi(u;0) = \pi_0(u) \,.$$

[1] Fischer Black (1938–1995), Myron Scholes (1941–): American financial economists. M. Scholes received the Sveriges Riksbank Prize in Economic Sciences in Memory of A. Nobel in 1997 for this model for valuing options, together with Robert Merton (1944–).

Time reversal of diffusions[2]

- Let $T > 0$:

$$\overline{\boldsymbol{b}}(\boldsymbol{u},t) = -\boldsymbol{b}(\boldsymbol{u},T-t) + \frac{\boldsymbol{\nabla}_{\boldsymbol{u}} \cdot (\pi(\boldsymbol{u},T-t)\boldsymbol{a}(\boldsymbol{u},T-t))}{\pi(\boldsymbol{u},T-t)}\,,$$

$$\overline{\boldsymbol{\sigma}}(\boldsymbol{u},t) = \boldsymbol{\sigma}(\boldsymbol{u},T-t)\,,$$

and:

$$\overline{\mathcal{A}}_t\phi(\boldsymbol{u}) = \overline{\boldsymbol{b}}(\boldsymbol{u},t)\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\phi + \frac{1}{2}\overline{\boldsymbol{a}}(\boldsymbol{u},t) : \boldsymbol{\nabla}_{\boldsymbol{u}}\otimes\boldsymbol{\nabla}_{\boldsymbol{u}}\phi\,,$$

with $\overline{\boldsymbol{a}} = \overline{\boldsymbol{\sigma}}\,\overline{\boldsymbol{\sigma}}^{\mathsf{T}}$.

- Then $\overline{\boldsymbol{U}}(t) = \boldsymbol{U}(T-t)$ is a Markov diffusion process with infinitesimal generator $\overline{\mathcal{A}}_t$, such that:

$$\mathrm{d}\overline{\boldsymbol{U}} = \overline{\boldsymbol{b}}(\overline{\boldsymbol{U}},t)\mathrm{d}t + \overline{\boldsymbol{\sigma}}(\overline{\boldsymbol{U}},t)\mathrm{d}\boldsymbol{W}\,,\quad 0\leqslant t<T\,.$$

[2] U. G. Haussmann, É. Pardoux. Time reversal of diffusions. *Ann. Probab.* **14**(4), 1188-1205 (1986).

Time reversal of diffusions

Discretized forward diffusion $\boldsymbol{U}_{i+1}=\boldsymbol{U}_i+\boldsymbol{b}_i\Delta t+\sigma_i\sqrt{\Delta t}\boldsymbol{Z}$, $\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, from which one deduces:

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{u}_i|\boldsymbol{u}_{i+1}) &= \frac{\mathbb{P}(\boldsymbol{u}_{i+1}|\boldsymbol{u}_i)\mathbb{P}(\boldsymbol{u}_i)}{\mathbb{P}(\boldsymbol{u}_{i+1})} \\
&= \frac{\mathbb{P}(\boldsymbol{u}_i)}{\mathbb{P}(\boldsymbol{u}_{i+1})}\mathcal{N}(\boldsymbol{u}_{i+1};\boldsymbol{u}_i+\boldsymbol{b}_i\Delta t,\sigma_i^2\Delta t)\,;
\end{aligned}
$$

But:

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{u}_i) &= \mathbb{P}(\boldsymbol{u}_{i+1})+(\boldsymbol{u}_i-\boldsymbol{u}_{i+1})\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\mathbb{P}(\boldsymbol{u}_{i+1})+\cdots \\
\frac{\mathbb{P}(\boldsymbol{u}_i)}{\mathbb{P}(\boldsymbol{u}_{i+1})} &= 1+(\boldsymbol{u}_i-\boldsymbol{u}_{i+1})\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\log\mathbb{P}(\boldsymbol{u}_{i+1})+\cdots \\
&\simeq \exp\left[(\boldsymbol{u}_i-\boldsymbol{u}_{i+1})\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\log\mathbb{P}(\boldsymbol{u}_{i+1})\right]\,,
\end{aligned}
$$

from which one deduces:

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{u}_i|\boldsymbol{u}_{i+1}) &\propto \exp\left[(\boldsymbol{u}_i-\boldsymbol{u}_{i+1})\cdot\boldsymbol{\nabla}_{\boldsymbol{u}}\log\mathbb{P}(\boldsymbol{u}_{i+1})-\frac{\|\boldsymbol{u}_{i+1}-\boldsymbol{u}_i-\boldsymbol{b}_i\Delta t\|^2}{2\sigma_i^2\Delta t}\right] \\
&= \exp\left[-\frac{\left\|\boldsymbol{u}_i-(\boldsymbol{u}_{i+1}-\boldsymbol{b}_i\Delta t+\sigma_i^2\boldsymbol{\nabla}_{\boldsymbol{u}}\log\mathbb{P}(\boldsymbol{u}_{i+1})\Delta t)\right\|^2}{2\sigma_i^2\Delta t}\right]\,.
\end{aligned}
$$

Probability flow ODE

- Since $\boldsymbol{\nabla}_{\boldsymbol{u}} \cdot (\pi \boldsymbol{a}) = \pi \boldsymbol{\nabla}_{\boldsymbol{u}} \cdot \boldsymbol{a} + \boldsymbol{a} \boldsymbol{\nabla}_{\boldsymbol{u}} \pi$, the Fokker-Planck equation also reads:

$$\partial_t \pi + \boldsymbol{\nabla}_{\boldsymbol{u}} \cdot \left(\pi \boldsymbol{b}^\dagger\right) = 0\,,$$

  with:

$$\boldsymbol{b}^\dagger(\boldsymbol{u},t) = \boldsymbol{b}(\boldsymbol{u},t) - \frac{1}{2}\frac{\boldsymbol{\nabla}_{\boldsymbol{u}} \cdot (\pi(\boldsymbol{u},t)\boldsymbol{a}(\boldsymbol{u},t))}{\pi(\boldsymbol{u},t)}\,.$$

- It is the Fokker-Planck equation associated to the SDE:

$$\mathrm{d}\boldsymbol{U}^\dagger = \boldsymbol{b}^\dagger(\boldsymbol{U}^\dagger,t)\mathrm{d}t\,,\quad \boldsymbol{U}^\dagger(0) = \boldsymbol{U}_0\,,$$

  which is coined "probability flow ODE" (Ordinary Differential Equation).
- $\boldsymbol{U}$ and $\boldsymbol{U}^\dagger$ have the same marginals, but they don't have the same distribution (as random processes).

Feynman-Kac formula

- Consider the *backward* PDE with final condition at $T>0$ and (scalar) solution $\boldsymbol{x},t\mapsto u(\boldsymbol{x},t)$:

$$\partial_t u+\boldsymbol{b}(\boldsymbol{x})\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}u+\frac{1}{2}\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\sigma}(\boldsymbol{x})^{\mathsf{T}}:\boldsymbol{\nabla}_{\boldsymbol{x}}\otimes\boldsymbol{\nabla}_{\boldsymbol{x}}u=f(\boldsymbol{x})\,,$$
$$u(\boldsymbol{x},T)=\phi(\boldsymbol{x})\,;$$

- Consider the process $\boldsymbol{X}_t$ on $[\tau,T]$ solving:

$$\mathrm{d}\boldsymbol{X}_t=\boldsymbol{b}(\boldsymbol{X}_t)\mathrm{d}t+\boldsymbol{\sigma}(\boldsymbol{X}_t)\mathrm{d}\boldsymbol{W}_t\,;$$

- Then:

$$\boxed{u(\boldsymbol{\xi},\tau)=\mathbb{E}\left\{\phi(\boldsymbol{X}_T)-\int_\tau^T f(\boldsymbol{X}_t)\mathrm{d}t\,\middle|\,\boldsymbol{X}_\tau=\boldsymbol{\xi}\right\}}\,.$$

Feynman-Kac formula

- Consider the *forward* PDE with initial condition at $t=0$ and (scalar) solution $\boldsymbol{x},t \mapsto u(\boldsymbol{x},t)$:

$$\begin{aligned}\partial_t u &= \boldsymbol{b}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} u + \frac{1}{2}\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\sigma}(\boldsymbol{x})^{\mathsf{T}} : \boldsymbol{\nabla}_{\boldsymbol{x}} \otimes \boldsymbol{\nabla}_{\boldsymbol{x}} u\,, \\ u(\boldsymbol{x},0) &= \phi(\boldsymbol{x})\,;\end{aligned}$$

- Consider the process $\boldsymbol{X}_t$ on $[0,\tau]$ solving:

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{X}_t)\mathrm{d}\boldsymbol{W}_t\,;$$

- Then:

$$\boxed{u(\boldsymbol{\xi},\tau) = \mathbb{E}\{\phi(\boldsymbol{X}_\tau)|\boldsymbol{X}_0 = \boldsymbol{\xi}\}}\,.$$

Feynman-Kac formula

Backward case. Using Itō's formula for $u(\boldsymbol{X}_t, t)$:

$$\begin{aligned}
\mathrm{d}u &= (\partial_t u)\mathrm{d}t + \boldsymbol{\nabla}_{\boldsymbol{x}} u \cdot \mathrm{d}\boldsymbol{X}_t + \frac{1}{2}\boldsymbol{\nabla}_{\boldsymbol{x}} \otimes \boldsymbol{\nabla}_{\boldsymbol{x}} u : \mathrm{d}\boldsymbol{X}_t \otimes \mathrm{d}\boldsymbol{X}_t \\
&= \left(\partial_t u + \boldsymbol{\nabla}_{\boldsymbol{x}} u \cdot \boldsymbol{b}(\boldsymbol{X}_t) + \frac{1}{2}\boldsymbol{\nabla}_{\boldsymbol{x}} \otimes \boldsymbol{\nabla}_{\boldsymbol{x}} u : \boldsymbol{\sigma}(\boldsymbol{X}_t)\boldsymbol{\sigma}(\boldsymbol{X}_t)^{\mathsf{T}}\right)\mathrm{d}t + \boldsymbol{\nabla}_{\boldsymbol{x}} u \cdot \boldsymbol{\sigma}(\boldsymbol{X}_t)\mathrm{d}\boldsymbol{W}_t \\
&= f(\boldsymbol{X}_t)\mathrm{d}t + \boldsymbol{\nabla}_{\boldsymbol{x}} u \cdot \boldsymbol{\sigma}(\boldsymbol{X}_t)\mathrm{d}\boldsymbol{W}_t\,;
\end{aligned}$$

Integrating between τ and T:

$$u(\boldsymbol{X}_T, T) - u(\boldsymbol{X}_\tau, \tau) = \phi(\boldsymbol{X}_T) - u(\boldsymbol{X}_\tau, \tau) = \int_\tau^T f(\boldsymbol{X}_s)\mathrm{d}s + \int_\tau^T \boldsymbol{\nabla}_{\boldsymbol{x}} u \cdot \boldsymbol{\sigma}(\boldsymbol{X}_s)\mathrm{d}\boldsymbol{W}_s\,;$$

Taking the expectation the martingale part vanishes and then:

$$u(\boldsymbol{\xi}, \tau) = \mathbb{E}\left\{\phi(\boldsymbol{X}_T) - \int_\tau^T f(\boldsymbol{X}_s)\mathrm{d}s \,\middle|\, \boldsymbol{X}_\tau = \boldsymbol{\xi}\right\}.$$

Forward case. Change of variable $t \to T - t$.

Feynman-Kac formula

- *Example*: heat equation in $\mathbb{R}^d$ with $\boldsymbol{b} = \boldsymbol{0}$ and $\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} = \boldsymbol{I}$,

$$\begin{aligned} \partial_t T &= \frac{1}{2}\Delta T\,, \quad t > 0\,, \\ T(\boldsymbol{x}, 0) &= T_0(\boldsymbol{x})\,; \end{aligned}$$

- Then:

$$\boxed{T(\boldsymbol{x}, t) = \mathbb{E}\{T_0(\boldsymbol{x} + \boldsymbol{W}_t)\}}\,.$$

Outline

First-order stochastic differential equation

A general first-order stochastic differential equation for the process $\boldsymbol{U}$ indexed on $\mathbb{R}_+$ with values in $\mathbb{R}^q$:

$$\begin{cases} \dot{\boldsymbol{U}}(t) = \boldsymbol{b}(\boldsymbol{U},t) + \boldsymbol{\sigma}(\boldsymbol{U},t)\boldsymbol{F}(t)\,, & t > 0\,, \\ \boldsymbol{U}(0) = \boldsymbol{U}_0\,, \end{cases}$$

with the data:

- $\boldsymbol{u}, t \mapsto \boldsymbol{b}(\boldsymbol{u},t) : \mathbb{R}^q \times \mathbb{R}_+ \to \mathbb{R}^q$ the *drift* function;
- $\boldsymbol{u}, t \mapsto \boldsymbol{\sigma}(\boldsymbol{u},t) : \mathbb{R}^q \times \mathbb{R}_+ \to \mathbb{M}_{q,p}(\mathbb{R})$ the *scattering* operator;
- $\boldsymbol{U}_0$ is an r.v. in $\mathbb{R}^q$ with known marginal PDF $\pi_0(\boldsymbol{u}_0)$;
- $\boldsymbol{F}(t) = (F_1(t), \dots F_p(t))$ is a second-order Gaussian random process indexed on $\mathbb{R}^+$ with values in $\mathbb{R}^p$, centered, mean-square continuous.

Definition

$\boldsymbol{F}(t)$ indexed on $\mathbb{R}^+$ with values in $\mathbb{R}^p$, second-order, Gaussian, centered and mean-square continuous admits a Markovian realization if:

$$\begin{cases} \boldsymbol{F}(t) = \boldsymbol{H}\boldsymbol{V}(t)\,, & t \geqslant 0\,, \\ \dot{\boldsymbol{V}}(t) = \boldsymbol{P}\boldsymbol{V}(t) + \boldsymbol{Q}\boldsymbol{B}(t)\,, & t > 0\,, \\ \boldsymbol{V}(0) = \boldsymbol{V}_0 & a.s. \end{cases}$$

where $\boldsymbol{V}_0$ is a Gaussian r.v. in $\mathbb{R}^n$, $\boldsymbol{V}(t)$ is a diffusion process indexed on $\mathbb{R}_+$ with values in $\mathbb{R}^n$, $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{M}_n(\mathbb{R})$, $\boldsymbol{H} \in \mathbb{M}_{p,n}(\mathbb{R})$, $\Re\mathrm{e}\{\lambda_j(\boldsymbol{P})\} < 0$.

- This is equivalent to a linear Itō stochastic differential equation.
- $\boldsymbol{V}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0 = \int_0^{+\infty} \mathrm{e}^{\tau\boldsymbol{P}}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\mathrm{e}^{\tau\boldsymbol{P}^\mathsf{T}}\,\mathrm{d}\tau$.

Physically realizable process

Definition

Definition

$\boldsymbol{F}(t)$ indexed on $\mathbb{R}$ with values in $\mathbb{R}^p$, second-order, mean-square stationary and continuous, centered, is physically realizable if $\exists \mathbb{H} \in L^2(\mathbb{R})$, $\operatorname{supp} \mathbb{H} \subseteq \mathbb{R}_+$, such that:

$$\boldsymbol{F}(t) = \int_{-\infty}^{t} \mathbb{H}(t-\tau)\boldsymbol{B}(\tau)\mathrm{d}\tau \,,$$

or equivalently $\boldsymbol{S_F}(\omega) = \frac{1}{2\pi}\widehat{\mathbb{H}}(\omega)\widehat{\mathbb{H}}(\omega)^*$, $\forall \omega \in \mathbb{R}$.

A necessary and sufficient condition (Rozanov 1967):

$$\int_{\mathbb{R}} \frac{\ln(\det \boldsymbol{S_F}(\omega))}{1+\omega^2}\mathrm{d}\omega > -\infty \,.$$

Markovian realization

Existence for a physically realizable process

Theorem

A necessary and sufficient condition:

$$\boldsymbol{S}_{\boldsymbol{F}}(\omega) = \frac{\boldsymbol{R}(\mathrm{i}\omega)\boldsymbol{R}(\mathrm{i}\omega)^*}{2\pi|P(\mathrm{i}\omega)|^2}\,, \quad \textit{or}\ \mathbb{H}(\omega) = \frac{\boldsymbol{R}(\mathrm{i}\omega)}{P(\mathrm{i}\omega)}\,,$$

where:

- $P(z)$ *is a polynomial of degree* d *on* $\mathbb{C}$ *with real coefficients and roots in the half-plane* $\Re\mathrm{e}(z) < 0$,
- $\boldsymbol{R}(z)$ *is a polynomial on* $\mathbb{C}$ *with coefficients in* $\mathbb{M}_{p,n}(\mathbb{R})$ *and degree* $r < n$.

The Markovian realization always exists in infinite dimension $n = +\infty$.

First-order SDE (cont'd)

A non linear first-order stochastic differential equation for the process $\boldsymbol{Z}(t) = (\boldsymbol{U}(t), \boldsymbol{V}(t))$ indexed on $\mathbb{R}_+$ with values in $\mathbb{R}^\nu$, $\nu = q + n$:

$$\begin{cases} \mathrm{d}\boldsymbol{Z}(t) = \boldsymbol{b}_z(\boldsymbol{Z}, t)\mathrm{d}t + \boldsymbol{\sigma}_z \mathrm{d}\boldsymbol{W}\,, \quad t > 0\,, \\ \boldsymbol{Z}(0) = \boldsymbol{Z}_0\,, \end{cases}$$

where $\boldsymbol{Z}_0 = (\boldsymbol{U}_0, \boldsymbol{V}_0)$,

$$\boldsymbol{b}_z(\boldsymbol{u}, \boldsymbol{v}, t) = \begin{bmatrix} \boldsymbol{b}(\boldsymbol{u}, t) + \boldsymbol{\sigma}(\boldsymbol{u}, t)\boldsymbol{H}\boldsymbol{v} \\ \boldsymbol{P}\boldsymbol{v} \end{bmatrix}, \quad \boldsymbol{\sigma}_z = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q} \end{bmatrix},$$

and $\boldsymbol{W}(t)$ is the Wiener process in $\mathbb{R}^\nu$.

Numerical integration of SDE

Strong convergence

$$
\left\{
\begin{array}{ll}
\mathrm{d}U(t) = b(U,t)\mathrm{d}t + \sigma(U,t)\mathrm{d}W(t)\,, & t > 0\,, \\
U(0) = U_0 & \text{a.s.}
\end{array}
\right.
$$

Definition

An approximation $(\tilde{U}_j)_j$ *converges with strong order* $k > 0$ *if* $\exists K_j > 0$*:*

$$
\mathbb{E}\left\{\left|U(j\Delta t) - \tilde{U}_j\right|\right\} \leqslant K_j(\Delta t)^k\,.
$$

The sample paths of the approximation $\tilde{U}$ should be close to those of the Itō process.

$$\begin{cases} \mathrm{d}U(t) = b(U,t)\mathrm{d}t + \sigma(U,t)\mathrm{d}W(t)\,, & t > 0\,, \\ U(0) = U_0 & \text{a.s.} \end{cases}$$

Definition

An approximation $(\tilde{U}_j)_j$ converges with weak order $k > 0$ if for any polynomial g $\exists K_{g,j} > 0$:

$$\left| \mathbb{E}\{g(U(j\Delta t))\} - \mathbb{E}\{g(\tilde{U}_j)\} \right| \leqslant K_{g,j}(\Delta t)^k\,.$$

The probability distribution of the approximation should be close to that of the Itō process in order to get a good estimate of the expectation ($g(u) = u$) or the variance ($g(u) = u^2$), for example.

Time discrete approximations

Explicit 0.5–order methods

- Assume that v and a are independent of time t (thus $U(t)$ is a diffusion process), and let $t_j = j\Delta t$, $b_j = b(\tilde{U}_j)$, $\sigma_j = \sigma(\tilde{U}_j)$, $U_0 \sim \pi_0(\mathrm{d}u_0)$, $G \sim \mathcal{N}(0,1)$.
- Itō SDE: the *Euler-Maruyama scheme* (1955),

$$\begin{aligned} \tilde{U}_{j+1} &= \tilde{U}_j + b_j\Delta t + \sigma_j\sqrt{\Delta t}\,G\,, \\ \tilde{U}_0 &= U_0\,. \end{aligned}$$

- Stratonovich SDE: the *Euler-Heun scheme* (1982),

$$\begin{aligned} \tilde{U}_{j+1} &= \tilde{U}_j + b_j\Delta t + \tilde{\sigma}_j\sqrt{\Delta t}\,G\,, \\ \tilde{\sigma}_j &= \frac{1}{2}\left[\sigma_j + \sigma\left(\tilde{U}_j + \sigma_j\sqrt{\Delta t}\,G\right)\right]\,, \\ \tilde{U}_0 &= U_0\,. \end{aligned}$$

- Both have a strong order $k = \frac{1}{2}$ (vs. $k = 1$ for ordinary differential equations) and a weak order $k = 1$.

Time discrete approximations
Explicit 1–order methods

- The *Milstein scheme* (1974):

$$\tilde{U}_{j+1} = \tilde{U}_j + b_{\lambda,j}\Delta t + \sigma_j\sqrt{\Delta t}\,G + \frac{1}{2}\sigma_j\sigma_j'\Delta t(G^2 + 2\lambda - 1)\,,$$
$$\tilde{U}_0 = U_0\,,$$

where $\lambda = 0$ (Itō SDE) or $\lambda = \frac{1}{2}$ (Stratonovich SDE).

- The *Runge-Kutta Milstein scheme* (1984):

$$\tilde{U}_{j+1} = \tilde{U}_j + b_{\lambda,j}\Delta t + \sigma_j\sqrt{\Delta t}\,G + \frac{1}{2}\sigma_j\tilde{\sigma}_j'\Delta t(G^2 + 2\lambda - 1)\,,$$
$$\sigma_j\tilde{\sigma}_j' = (\Delta t)^{-\frac{1}{2}}\left[\sigma\left(\tilde{U}_j + \sigma_j\sqrt{\Delta t}\right) - \sigma_j\right]\,,$$
$$\tilde{U}_0 = U_0\,.$$

- Both have strong and weak orders $k = 1$ (under mild conditions on b and σ).

Time discrete approximations

Stochastic Taylor approximations

- Higher-order schemes may be derived using stochastic Taylor expansions:

$$U_{j+1} - U_j = \int_{t_j}^{t_{j+1}} b(U)\mathrm{d}t + \int_{t_j}^{t_{j+1}} \sigma(U)\mathrm{d}W$$

$$\simeq \int_{t_j}^{t_{j+1}} \big(b(U_j) + b'(U_j)\Delta U_j\big)\,\mathrm{d}t + \int_{t_j}^{t_{j+1}} \big(\sigma(U_j) + \sigma'(U_j)\Delta U_j\big)\,\mathrm{d}W\,,$$

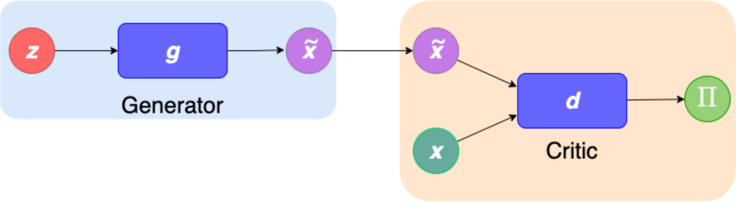
  where $\Delta U_j = \int_{t_j}^{t} b(U)\mathrm{d}\tau + \int_{t_j}^{t} \sigma(U)\mathrm{d}W$.

- Then $\int_{t_j}^{t_{j+1}} \int_{t_j}^{t} \mathrm{d}_\lambda W(s)\mathrm{d}_\lambda W(t) = \frac{1}{2}(\Delta W)^2 + (\lambda - \frac{1}{2})\Delta t$.
- Higher-order expansions involve additional r.v. $\Delta Z_j = \int_{t_j}^{t_{j+1}} \int_{t_j}^{t} \mathrm{d}W\mathrm{d}t$ with $\mathbb{E}\{(\Delta Z_j)^2\} \propto \Delta t^3$ *etc.*
- Weak Taylor approximations $U_0 \sim \hat{U}_0$, $\Delta W \sim \Delta\hat{W}$, $\Delta Z_j \sim \Delta\hat{Z}_j$ with approximately the same moment properties.

Outline

Setting

- *Unsupervised learning*: infer $\boldsymbol{X}^* \sim \mathbb{P}(\boldsymbol{x})$ where $\mathbb{P}(\boldsymbol{x})$ is only partially known through the dataset $\{\boldsymbol{x}_n\}_{n=1}^N$;
- *Discriminator*: $\mathbb{R}^q \to \mathbb{R} : \boldsymbol{x} \mapsto d_{\boldsymbol{\phi}}(\boldsymbol{x})$;
- *Generator*: $\mathbb{R}^p \to \mathbb{R}^q : \boldsymbol{z} \mapsto \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$ with $p \ll q$.



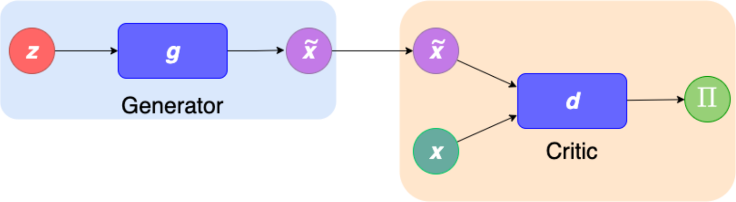
Training and inference

- **Loss function:** $\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, say, and

$$\boldsymbol{\theta}^*, \boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \frac{1}{N} \sum_{n=1}^{N} [d_{\boldsymbol{\phi}}(\boldsymbol{x}_i) - d_{\boldsymbol{\phi}}(\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{Z}_i))]$$

$$(+\text{ gradient penalty})\,;$$

- **Inference:** $\boldsymbol{X}^* = \boldsymbol{g}_{\boldsymbol{\theta}^*}(\boldsymbol{Z})$ where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, say.



Outline

Setting

- *Unsupervised learning*: infer $\boldsymbol{X}^* \sim \mathbb{P}(\boldsymbol{x})$ where $\mathbb{P}(\boldsymbol{x})$ is only partially known through the dataset $\{\boldsymbol{x}_n\}_{n=1}^N$;
- *Encoder* (forward relationship $\mathbb{R}^q \to \mathbb{R}^p$ with $p \ll q$):

$$\boldsymbol{Z} = \text{encoder}_{\boldsymbol{\phi}}(\boldsymbol{X}) \sim \mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \approx \mathbb{P}(\boldsymbol{z}|\boldsymbol{x})\,;$$

- *Decoder* (backward relationship $\mathbb{R}^p \to \mathbb{R}^q$):

$$\boldsymbol{X} = \text{decoder}_{\boldsymbol{\theta}}(\boldsymbol{Z}) \sim \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \approx \mathbb{P}(\boldsymbol{x}|\boldsymbol{z})\,;$$

- The *latent variable* is *e.g.* $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
- *Example*: in `jpeg` the encoder is the discrete cosine transform, the decoder is its inverse, and $\boldsymbol{z}$ are the projection coefficients.

Evidence

- Straightforward inference is $\boldsymbol{X}^* = \arg\max \mathbb{P}(\boldsymbol{x})$ but:

$$\begin{aligned}
\log \mathbb{P}(\boldsymbol{x}) &= \log \frac{\mathbb{P}(\boldsymbol{x},\boldsymbol{z})}{\mathbb{P}(\boldsymbol{z}|\boldsymbol{x})} \\
&= \log \frac{\mathbb{P}(\boldsymbol{x}|\boldsymbol{z})\mathbb{P}(\boldsymbol{z})\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{\mathbb{P}(\boldsymbol{z}|\boldsymbol{x})\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \\
&= \log \mathbb{P}(\boldsymbol{x}|\boldsymbol{z}) - \log \frac{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{\mathbb{P}(\boldsymbol{z})} + \log \frac{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{\mathbb{P}(\boldsymbol{z}|\boldsymbol{x})}
\end{aligned}$$

  where $\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ could actually be anything else.

- Then (remind $\mathbb{D}_{\mathrm{KL}}(\mathbb{P}||\mathbb{Q}) \overset{\text{def}}{=} \mathbb{E}_{\mathbb{P}}\{\log \frac{\mathbb{P}}{\mathbb{Q}}\} \geqslant 0$):

$$\begin{aligned}
\log \mathbb{P}(\boldsymbol{x}) &= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log \mathbb{P}(\boldsymbol{x})\} \\
&= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log \mathbb{P}(\boldsymbol{x}|\boldsymbol{Z})\} - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z})) + \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z}|\boldsymbol{x})) \\
&\geqslant \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log \mathbb{P}(\boldsymbol{x}|\boldsymbol{Z})\} - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z})) \\
&\simeq \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{Z})\} - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z}))\,.
\end{aligned}$$

Evidence Lower BOund (ELBO)

- The *Evidence Lower BOund* (ELBO), or *variational bound*, is:

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) &\stackrel{\text{def}}{=} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left\{\log\frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{Z})}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{Z}|\boldsymbol{x})}\right\} \\ &= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{Z})\} - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z}))\,;\end{aligned}$$

- Sampling $\boldsymbol{Z}\sim\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and maximizing $\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$, we almost do maximize $\log\mathbb{P}(\boldsymbol{x})$ on average;
- *Prior matching*: $\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z}))$ tells how good the encoder is *vs.* a prior belief held over latent variables, and has to be minimized to maximize the ELBO;
- *Reconstruction*: $\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\{\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{Z})\}$ tells how good the decoder is, and has to be maximized to maximize the ELBO.

Training and inference

- **Encoder**: we choose $\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x}),\sigma^2_{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{I})$, where $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x})$ and $\sigma^2_{\boldsymbol{\phi}}(\boldsymbol{x})$ are Neural Networks;
- **Decoder**: we choose $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z}),\sigma^2\boldsymbol{I})$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z})$ is a Neural Network and σ is a parameter;
- **Loss function**:

$$\boldsymbol{\theta}^*,\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \frac{1}{N}\sum_{n=1}^{N}\left(\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_n|\boldsymbol{z}_n) - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_n)||\mathbb{P}(\boldsymbol{z}))\right)\,,$$

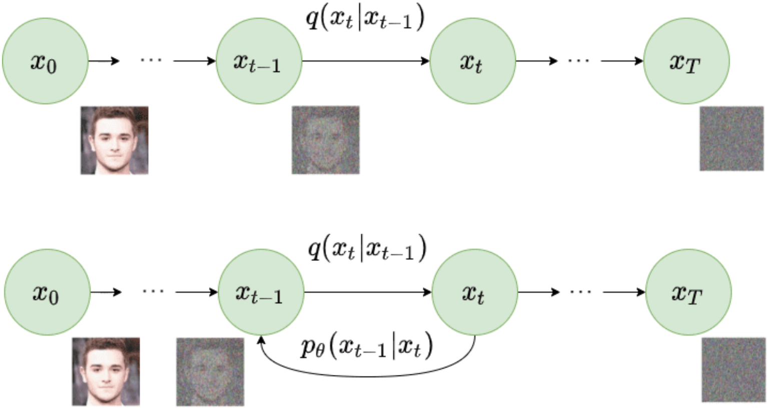
where $\boldsymbol{z}_n \sim \mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_n)$, and for *e.g.* $\mathbb{P}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$ in a latent space of dimension p:

$$\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||\mathbb{P}(\boldsymbol{z})) = \frac{1}{2}\left(\left\|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sigma^{2p}_{\boldsymbol{\phi}}(\boldsymbol{x}) - p\log\sigma^2_{\boldsymbol{\phi}}(\boldsymbol{x})\right)\,;$$

- **Inference**: $\boldsymbol{X}^* = \boldsymbol{\mu}_{\boldsymbol{\theta}^*}(\boldsymbol{Z})$ where $\boldsymbol{Z} \sim \mathbb{P}(\boldsymbol{z})$.

Outline

Unsupervised learning by noising/denoising



Langevin diffusion

- Langevin diffusion:

$$\mathrm{d}\boldsymbol{X}_t = -\boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{U}(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{W}_t\,, \quad \mathbb{R}^p \ni \boldsymbol{X}_0 \sim \pi_0\,.$$

- Regardless of π_0, $\boldsymbol{X}_t$ converges in law towards a density $\propto \mathrm{e}^{-\mathcal{U}(\boldsymbol{x})}$ as $t \to \infty$. The Fokker-Planck equation reads:

$$\partial_t \pi = \boldsymbol{\nabla}_{\boldsymbol{x}} \cdot (\pi \boldsymbol{\nabla}_{\boldsymbol{x}}\mathcal{U}) + \Delta\pi\,.$$

- **Example**: Ornstein-Uhlenbeck process, $\mathcal{U}(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{x}\right\|^2$, thus

$$\mathrm{d}\boldsymbol{X}_t = -\boldsymbol{X}_t\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{W}_t\,, \quad \boldsymbol{X}_0 \sim \pi_0\,,$$

and

$$\partial_t \pi = \boldsymbol{\nabla}_{\boldsymbol{x}} \cdot (\pi\boldsymbol{x}) + \Delta\pi\,.$$

Ornstein-Uhlenbeck process

- In integral form:

$$\boldsymbol{X}_t = \mathrm{e}^{-t}\left(\boldsymbol{X}_0 + \sqrt{2}\int_0^t \mathrm{e}^s \mathrm{d}\boldsymbol{W}_s\right)\,,$$

with ground-truth density $\pi = \pi_0(\cdot/\mathrm{e}^t) \star \mathcal{N}(0, 1-\mathrm{e}^{-2t})$.

Let $\boldsymbol{Y}_t = \mathrm{e}^t\boldsymbol{X}_t$. Then by Itō's formula with $\boldsymbol{\phi}(\boldsymbol{x},t) = \mathrm{e}^t\boldsymbol{x}$:

$$\begin{aligned}
\mathrm{d}\boldsymbol{Y}_t &= (\partial_t\boldsymbol{\phi})\mathrm{d}t + \boldsymbol{D}_{\boldsymbol{x}}\boldsymbol{\phi}\cdot\mathrm{d}\boldsymbol{X}_t + \frac{1}{2}\boldsymbol{D}^2_{\boldsymbol{x}}\boldsymbol{\phi} : \mathrm{d}\boldsymbol{X}_t\otimes\mathrm{d}\boldsymbol{X}_t \\
&= \mathrm{e}^t\boldsymbol{X}_t\mathrm{d}t + \mathrm{e}^t\mathrm{d}\boldsymbol{X}_t \qquad\qquad (\partial_t\boldsymbol{\phi} = \boldsymbol{\phi}\,,\ \boldsymbol{D}_{\boldsymbol{x}}\boldsymbol{\phi} = \mathrm{e}^t\boldsymbol{I}\,,\ \boldsymbol{D}^2_{\boldsymbol{x}}\boldsymbol{\phi} = \boldsymbol{0}) \\
&= \mathrm{e}^t\boldsymbol{X}_t\mathrm{d}t + \mathrm{e}^t(-\boldsymbol{X}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{W}_t) \\
&= \sqrt{2}\mathrm{e}^t\mathrm{d}\boldsymbol{W}_t\,.
\end{aligned}$$

Thus (the second equality is integration by parts):

$$\boldsymbol{Y}_t = \boldsymbol{X}_0 + \sqrt{2}\int_0^t \mathrm{e}^s\mathrm{d}\boldsymbol{W}_s = \boldsymbol{X}_0 + \sqrt{2}\left(\mathrm{e}^t\boldsymbol{W}_t - \int_0^t \mathrm{e}^s\boldsymbol{W}_s\mathrm{d}s\right)\,.$$

Ornstein-Uhlenbeck process

- It is a Gaussian process with conditional moments:

$$\mathbb{E}\{\boldsymbol{X}_t|\boldsymbol{X}_0\} = \mathrm{e}^{-t}\boldsymbol{X}_0\,,$$
$$\mathbb{V}\{\boldsymbol{X}_t|\boldsymbol{X}_0\} = (1-\mathrm{e}^{-2t})\boldsymbol{I}\,.$$

$$\mathbb{E}\{(\boldsymbol{X}_t-\mathbb{E}\{\boldsymbol{X}_t\})\otimes(\boldsymbol{X}_{t'}-\mathbb{E}\{\boldsymbol{X}_{t'}\})|\boldsymbol{X}_0\} = 2\mathrm{e}^{-t-t'}\mathbb{E}\left\{\left(\int_0^t \mathrm{e}^s\mathrm{d}\boldsymbol{W}_s\right)\otimes\left(\int_0^{t'} \mathrm{e}^s\mathrm{d}\boldsymbol{W}_s\right)\right\}\,.$$

Then apply Itō's isometry formula:

$$\mathbb{E}\left\{\left(\int_0^t Y_s\mathrm{d}\boldsymbol{W}_s\right)\otimes\left(\int_0^t Z_s\mathrm{d}\boldsymbol{W}_s\right)\right\} = \mathbb{E}\left\{\int_0^t Y_sZ_s\mathrm{d}s\right\}\boldsymbol{I}$$

to obtain:

$$\mathbb{E}\left\{\left(\int_0^t \mathrm{e}^s\mathrm{d}\boldsymbol{W}_s\right)\otimes\left(\int_0^{t'} \mathrm{e}^s\mathrm{d}\boldsymbol{W}_s\right)\right\} = \mathbb{E}\left\{\int_0^{\min(t,t')} \mathrm{e}^{2s}\mathrm{d}s\right\} = \frac{1}{2}\left(\mathrm{e}^{2\min(t,t')}-1\right)\,.$$

Backward Ornstein-Uhlenbeck process

- Let $\rho(\boldsymbol{x},t)=\pi(\boldsymbol{x},T-t)$ starting from some large enough $T\gg 0$, it satisfies the Fokker-Planck equation:

$$\partial_t\rho=-\partial_t\pi=-\boldsymbol{\nabla}_{\boldsymbol{x}}\cdot(\rho\boldsymbol{x})-\Delta\rho\,.$$

- But for any $\alpha\geqslant 0$:

$$-\Delta\rho=\alpha\Delta\rho-(1+\alpha)\boldsymbol{\nabla}\cdot(\rho\boldsymbol{\nabla}\log\rho)\,,$$

such that the Fokker-Planck equation also reads:

$$\partial_t\rho=-\boldsymbol{\nabla}_{\boldsymbol{x}}\cdot(\rho\boldsymbol{x}+(1+\alpha)\rho\boldsymbol{\nabla}_{\boldsymbol{x}}\log\rho)+\alpha\Delta\rho\,.$$

- This is the law density of the process $\boldsymbol{X}_{T-t}$ that follows backward Langevin diffusion:

$$\mathrm{d}\boldsymbol{X}_{T-t}=(\boldsymbol{X}_{T-t}+(1+\alpha)\boldsymbol{\nabla}_{\boldsymbol{x}}\log\rho(\boldsymbol{X}_{T-t},t))\mathrm{d}t+\sqrt{2\alpha}\,\mathrm{d}\boldsymbol{W}_t\,,$$

starting from $\boldsymbol{X}_T\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$.

Score function

- *Stein's score function*:

$$\boldsymbol{s}(\boldsymbol{x},t)=\boldsymbol{\nabla}_{\boldsymbol{x}}\log\pi(\boldsymbol{x},t)=\boldsymbol{\nabla}_{\boldsymbol{x}}\log\rho(\boldsymbol{x},T-t)$$

  is approximated from samples of $\boldsymbol{X}_t$ from the forward flow, $\boldsymbol{s}(\boldsymbol{x},t)\approx\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$.

- For Ornstein-Uhlenbeck $\boldsymbol{X}_t\sim\mathrm{e}^{-t}\boldsymbol{X}_0+\sqrt{1-\mathrm{e}^{-2t}}\boldsymbol{Z}$, $\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, then setting $\mathbb{P}_{\boldsymbol{X}_t|\boldsymbol{X}_0}(\boldsymbol{x}|\boldsymbol{x}_0)\stackrel{\text{def}}{=}\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)$:

$$\begin{aligned}\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)&=-\frac{\boldsymbol{x}-\mathrm{e}^{-t}\boldsymbol{x}_0}{1-\mathrm{e}^{-2t}}\qquad(t\neq0)\\&=\boldsymbol{\nabla}_{\boldsymbol{x}}\log\frac{\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)}{\mathbb{P}(\boldsymbol{x}_0)}\\&=\frac{\boldsymbol{\nabla}_{\boldsymbol{x}}\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)}{\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)}\,.\end{aligned}$$

Denoising score matching

- Since $\mathbb{P}(\boldsymbol{x};t)=\int\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)\mathrm{d}\boldsymbol{x}_0$ one has:

$$
\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x};t) &= \frac{\boldsymbol{\nabla}_{\boldsymbol{x}}\mathbb{P}(\boldsymbol{x};t)}{\mathbb{P}(\boldsymbol{x};t)} \\
&= \frac{1}{\mathbb{P}(\boldsymbol{x};t)}\int\boldsymbol{\nabla}_{\boldsymbol{x}}\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)\mathrm{d}\boldsymbol{x}_0 \\
&= \frac{1}{\mathbb{P}(\boldsymbol{x};t)}\int\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)\mathrm{d}\boldsymbol{x}_0 \quad \text{(previous slide)} \\
&= \int\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)\mathbb{P}(\mathrm{d}\boldsymbol{x}_0|\boldsymbol{x};t) \quad \left(\frac{\mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0;t)}{\mathbb{P}(\boldsymbol{x};t)}\overset{\text{def}}{=}\mathbb{P}(\boldsymbol{x}_0|\boldsymbol{x};t)\right) \\
&= \mathbb{E}_{\boldsymbol{X}_0|\boldsymbol{X}_t}\{\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)\}\,. \quad \left(\mathbb{E}_{\boldsymbol{X}_0|\boldsymbol{X}_t}\{\boldsymbol{z}\}\overset{\text{def}}{=}\int\boldsymbol{z}\mathrm{d}\mathbb{P}(\boldsymbol{x}_0|\boldsymbol{x};t)\right)
\end{aligned}
$$

- The score function is $\boldsymbol{s}(\boldsymbol{x},t)\approx\boldsymbol{s}_{\boldsymbol{\theta}^*}(\boldsymbol{x},t)$ where:

$$
\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{X}_0}\{\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}\{\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)-\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)\|^2\}\}\,,
$$

averaging vs. $\boldsymbol{X}_t$ ($\mathbb{E}_{\boldsymbol{X}_t}\mathbb{E}_{\boldsymbol{X}_0|\boldsymbol{X}_t}=\mathbb{E}_{\boldsymbol{X}_0}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}=\mathbb{E}_{\boldsymbol{X}_0,\boldsymbol{X}_t}$).

Denoising score matching with diffusion

- The score function is $\boldsymbol{s}(\boldsymbol{x},t)\approx\boldsymbol{s}_{\boldsymbol{\theta}^*}(\boldsymbol{x},t)$ where:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \int_0^T \mathbb{E}_{\boldsymbol{X}_0,\boldsymbol{X}_t}\left\{\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \frac{\mathrm{e}^{-t}\boldsymbol{x}_0-\boldsymbol{x}}{1-\mathrm{e}^{-2t}}\right\|^2\right\}\lambda(\mathrm{d}t)\,,$$

  with a weighting scheme λ w.r.t. time s.t. $\lambda(\{0\})=0$ (because $\lim_{t\to 0}\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0;t)=\infty$); a common choice is $\lambda(t)=\sqrt{t}$.
- The function $\boldsymbol{s}_{\boldsymbol{\theta}}(\cdot,t):\mathbb{R}^p\to\mathbb{R}^p$ is typically a neural network from $\mathbb{R}^p$ into himself, *e.g.* a U-net for images.

Remark #1: implicit score matching[3]

- *Explicit score matching* reads:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} J_{\mathrm{ESM}}(\boldsymbol{\theta})\,,$$

where:

$$\begin{aligned}
J_{\mathrm{ESM}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{X})\right\|^2\right\} \\
&= \int\left(\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|^2 + \frac{1}{2}\left\|\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x})\right\|^2 - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}\log\mathbb{P}(\boldsymbol{x})\right)\mathbb{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int\left(\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|^2 - \frac{\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}\mathbb{P}(\boldsymbol{x})}{\mathbb{P}(\boldsymbol{x})}\right)\mathbb{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + C \\
&= \int\left(\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|^2 + \boldsymbol{\nabla}_{\boldsymbol{x}}\cdot\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\mathbb{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + C \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X})\right\|^2 + \boldsymbol{\nabla}_{\boldsymbol{x}}\cdot\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X})\right\} + C\,;
\end{aligned}$$

- The blue term is called *implicit score matching* $J_{\mathrm{ISM}}(\boldsymbol{\theta})$.

[3] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6**(4), 695-709 (2005).

Remark #2: denoising score matching[4]

- *Denoising score matching* reads:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} J_{\mathrm{DSM}}(\boldsymbol{\theta})\,,$$

where:

$$\begin{aligned}
J_{\mathrm{DSM}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}') - \boldsymbol{\nabla}_{\boldsymbol{x}'}\log\mathbb{P}(\boldsymbol{X}'|\boldsymbol{X})\right\|^2\right\} \\
&= \iint\left(\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}')\right\|^2 - \frac{\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}_{\boldsymbol{x}'}\mathbb{P}(\boldsymbol{x}'|\boldsymbol{x})}{\mathbb{P}(\boldsymbol{x}'|\boldsymbol{x})}\right)\mathbb{P}(\boldsymbol{x},\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' + C \\
&= \mathbb{E}_{\boldsymbol{X}'}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}')\right\|\right\} - \iint \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}_{\boldsymbol{x}'}\mathbb{P}(\boldsymbol{x}'|\boldsymbol{x})\mathbb{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' + C \\
&= \mathbb{E}_{\boldsymbol{X}'}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}')\right\|\right\} - \iint \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}')\cdot\boldsymbol{\nabla}_{\boldsymbol{x}'}\mathbb{P}(\boldsymbol{x},\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' + C \\
&= \mathbb{E}_{\boldsymbol{X}'}\left\{\frac{1}{2}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}')\right\| - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}')\cdot\boldsymbol{\nabla}_{\boldsymbol{x}'}\log\mathbb{P}(\boldsymbol{X}')\right\} + C \\
&= J_{\mathrm{ESM}}(\boldsymbol{\theta}) + C'\,.
\end{aligned}$$

[4] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.* **23**(7), 1661-1674 (2011).

Remark #2: denoising score matching

- If $\boldsymbol{X}'$ is a noised version of $\boldsymbol{X}$ s.t. $\boldsymbol{X}' = \boldsymbol{X} + \sigma \boldsymbol{Z}$, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then $\mathbb{P}(\boldsymbol{X}'|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}, \sigma^2 \boldsymbol{I})$ and:

$$J_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \left\{ \frac{1}{2} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X} + \sigma \boldsymbol{Z}) + \frac{\boldsymbol{Z}}{\sigma} \right\|^2 \right\} .$$

- Noise Conditional Score Network (NCSN)[5] considers different noise levels $\sigma_1, \sigma_2, \dots \sigma_T$:

$$J_{\mathrm{NCSN}}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \left\{ \frac{\lambda_i}{2} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X} + \sigma_i \boldsymbol{Z}) + \frac{\boldsymbol{Z}}{\sigma_i} \right\|^2 \right\} ,$$

  with weights $\lambda_1, \lambda_2, \dots \lambda_T$ (a common choice is $\lambda_i = \sigma_i^2$ and $\frac{\sigma_1}{\sigma_2} = \cdots = \frac{\sigma_{T-1}}{\sigma_T} > 1$).

[5] Y. Song, S. Ermon. Generative modeling by estimating gradients of the data distribution. arXiv:1907.05600 (2019).

DDPM in continuous setting[6]

- Forward Ornstein-Uhlenbeck process with time-varying noise $\beta(t)>0$:

$$\mathrm{d}\boldsymbol{X}_t=-\frac{1}{2}\beta(t)\boldsymbol{X}_t\mathrm{d}t+\sqrt{\beta(t)}\mathrm{d}\boldsymbol{W}_t\,,\quad \boldsymbol{X}_0\sim\pi_0\,.$$

- The solution reads:

$$\boldsymbol{X}_t=\mathrm{e}^{-\frac{1}{2}\int_0^t\beta(s)\mathrm{d}s}\left(\boldsymbol{X}_0+\int_0^t\mathrm{e}^{\frac{1}{2}\int_0^s\beta(s')\mathrm{d}s'}\sqrt{\beta(s)}\mathrm{d}\boldsymbol{W}_s\right)\,,$$

with moments:

$$\mathbb{E}\{\boldsymbol{X}_t|\boldsymbol{X}_0\}=\mathrm{e}^{-\frac{1}{2}\int_0^t\beta(s)\mathrm{d}s}\boldsymbol{X}_0\,,$$
$$\mathbb{V}\{\boldsymbol{X}_t|\boldsymbol{X}_0\}=\left(1-\mathrm{e}^{-\int_0^t\beta(s)\mathrm{d}s}\right)\boldsymbol{I}\,.$$

Apply again Itō's isometry formula with $Y_s=Z_s=\mathrm{e}^{\frac{1}{2}\int_0^s\beta(s')\mathrm{d}s'}\sqrt{\beta(s)}$.

[6] Called Variance Preserving (VP) SDE in: Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. arXiv:2011.13456 (2020).

DDPM in continuous setting

- Backward Ornstein-Uhlenbeck process $\overleftarrow{\boldsymbol{X}}_t \equiv \boldsymbol{X}_{T-t}$:

$$\mathrm{d}\overleftarrow{\boldsymbol{X}}_t = +\frac{1}{2}\beta(t)\left(\overleftarrow{\boldsymbol{X}}_t + 2\boldsymbol{s}(\overleftarrow{\boldsymbol{X}}_t, T-t)\right)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{W}_t\,,$$

starting from $\overleftarrow{\boldsymbol{X}}_0 = \boldsymbol{X}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

DDPM in discrete setting

- Forward process: apply *e.g.* Euler-Maruyama scheme,

$$\begin{aligned}\boldsymbol{X}_{i+1} &= \left(1-\frac{1}{2}\beta(t_i)\Delta t\right)\boldsymbol{X}_i + \sqrt{\beta(t_i)\Delta t}\,\boldsymbol{Z}\,, \quad \boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})\,,\\ &\simeq \sqrt{1-\beta_i}\,\boldsymbol{X}_i + \sqrt{\beta_i}\,\boldsymbol{Z}\\ &= \sqrt{\alpha_i}\,\boldsymbol{X}_i + \sqrt{1-\alpha_i}\,\boldsymbol{Z}\end{aligned}$$

setting $\beta(t_i)\Delta t = \beta_i = 1-\alpha_i$ such that $\beta_i \ll 1$.

- Backward process:

$$\begin{aligned}\boldsymbol{X}_{i-1} &= \left(1+\frac{1}{2}\beta_i\right)\boldsymbol{X}_i + \beta_i\boldsymbol{s}(\boldsymbol{X}_i,t_i) + \sqrt{\beta_i}\,\boldsymbol{Z}\\ &\simeq \left(1+\frac{1}{2}\beta_i\right)(\boldsymbol{X}_i + \beta_i\boldsymbol{s}(\boldsymbol{X}_i,t_i)) + \sqrt{\beta_i}\,\boldsymbol{Z}\\ &\simeq \frac{1}{\sqrt{1-\beta_i}}(\boldsymbol{X}_i + \beta_i\boldsymbol{s}(\boldsymbol{X}_i,t_i)) + \sqrt{\beta_i}\,\boldsymbol{Z}\\ &= \frac{1}{\sqrt{\alpha_i}}(\boldsymbol{X}_i + (1-\alpha_i)\boldsymbol{s}(\boldsymbol{X}_i,t_i)) + \sqrt{1-\alpha_i}\,\boldsymbol{Z}\,.\end{aligned}$$

DDPM encoder

- Given a *noise schedule* $\boldsymbol{\phi} = (\alpha_1, \alpha_2, \dots \alpha_T)$, the *encoder* is the Markov chain $\boldsymbol{X}_{1:T}$:

$$\boldsymbol{X}_i = \sqrt{\alpha_i}\,\boldsymbol{X}_{i-1} + \sqrt{1-\alpha_i}\,\boldsymbol{Z}\,, \quad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\,,$$

starting from $\boldsymbol{X}_0 \sim \pi_0$, with transition probabilities:

$$\begin{aligned}\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) &= \mathcal{N}(\boldsymbol{x}_i; \sqrt{\alpha_i}\,\boldsymbol{x}_{i-1}, (1-\alpha_i)\boldsymbol{I})\,,\\ \mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0) &= \mathcal{N}(\boldsymbol{x}_i; \sqrt{A_i}\,\boldsymbol{x}_0, (1-A_i)\boldsymbol{I})\,,\end{aligned}$$

where $A_i = \prod_{j=1}^{i} \alpha_j$.

- Choose $1 \geqslant \alpha_1 > \cdots > \alpha_T > 0$ so that the terminal transition probability $\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_T|\boldsymbol{x}_0) \approx \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$ independently of $\boldsymbol{x}_0$.

DDPM encoder

- $\boldsymbol{X}_{1:T}$ also has reverse transition probability, often called *forward process posterior*:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i,\boldsymbol{x}_0)=\mathcal{N}\left(\boldsymbol{x}_{i-1};\boldsymbol{\mu}_i(\boldsymbol{x}_i,\boldsymbol{x}_0),\tau_i^2\boldsymbol{I}\right)\,,$$

where:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i,\boldsymbol{x}_0)=\tau_i^2\left(\frac{\sqrt{\alpha_i}}{1-\alpha_i}\boldsymbol{x}_i+\frac{\sqrt{A_{i-1}}}{1-A_{i-1}}\boldsymbol{x}_0\right)\,,\;\tau_i^2=\frac{(1-\alpha_i)(1-A_{i-1})}{1-A_i}\,.$$

Apply Markov property and Bayes rule:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})=\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_{i-1},\boldsymbol{x}_0)=\frac{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i,\boldsymbol{x}_0)\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_0)}$$

$$\Rightarrow\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i,\boldsymbol{x}_0)=\frac{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_0)}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}$$

$$=\frac{\mathcal{N}(\boldsymbol{x}_i;\sqrt{\alpha_i}\boldsymbol{x}_{i-1},(1-\alpha_i)\boldsymbol{I})\mathcal{N}(\boldsymbol{x}_{i-1};\sqrt{A_{i-1}}\boldsymbol{x}_0,(1-A_{i-1})\boldsymbol{I})}{\mathcal{N}(\boldsymbol{x}_i;\sqrt{A_i}\boldsymbol{x}_0,(1-A_i)\boldsymbol{I})}\,.$$

ELBO

- *Unsupervised learning*: infer $\boldsymbol{X}_0^* \sim \pi_0 = \mathrm{d}\mathbb{P}(\boldsymbol{x}_0)$ using $\boldsymbol{X}_{1:T}$ as latent variables,

$$\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi}) \overset{\text{def}}{=} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left\{\log\frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_0,\boldsymbol{X}_{1:T})}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{X}_{1:T}|\boldsymbol{x}_0)}\right\}.$$

- Sampling $\boldsymbol{X}_{1:T} \sim \mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ and maximizing $\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi})$, we almost do maximize $\log\mathbb{P}(\boldsymbol{x}_0)$ on average.

$$\begin{aligned}\log\mathbb{P}(\boldsymbol{x}_0) &= \log\int\mathbb{P}(\boldsymbol{x}_0,\mathrm{d}\boldsymbol{x}_{1:T})\\ &= \log\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left\{\frac{\mathbb{P}(\boldsymbol{x}_0,\boldsymbol{X}_{1:T})}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{X}_{1:T}|\boldsymbol{x}_0)}\right\}\\ &\geqslant \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left\{\log\frac{\mathbb{P}(\boldsymbol{x}_0,\boldsymbol{X}_{1:T})}{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{X}_{1:T}|\boldsymbol{x}_0)}\right\},\end{aligned}$$

using Jensen's inequality $f(\mathbb{E}\{X\}) \geqslant \mathbb{E}\{f(X)\}$ with $f(x) = \log x$.

ELBO

- The ELBO also reads:

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_1|\boldsymbol{x}_0)}\{\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{X}_1)\} - \mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_T|\boldsymbol{x}_0)||\mathbb{P}(\boldsymbol{x}_T)) \\ - \sum_{i=2}^{T}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}\{\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i,\boldsymbol{x}_0)||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i))\}\,.\end{aligned}$$

- *Prior matching*: $\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_T|\boldsymbol{x}_0)||\mathbb{P}(\boldsymbol{x}_T))$ tells how good the encoder is *vs.* a prior distribution of the latent $\boldsymbol{X}_T$, and has to be minimized to maximize the ELBO;
- *Reconstruction*: $\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_1|\boldsymbol{x}_0)}\{\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{X}_1)\}$ tells how good the *decoder* (denoiser) is, and has to be maximized to maximize the ELBO;
- *Diffusion loss*: $\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i,\boldsymbol{x}_0)||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i))$, $2\leqslant i\leqslant T$, tell how much forward/backward transitions to latent variables are consistent, and have to be minimized to maximize the ELBO.

Training

- **Encoder**: choose the noise schedule $\boldsymbol{\phi}$, in which case there is nothing left to learn in $\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_T|\boldsymbol{x}_0)||\mathbb{P}(\boldsymbol{x}_T))$;
- **Decoder**: choose $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_{i-1};\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i),\tau_i^2\boldsymbol{I})$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is a Neural Network;
- **Loss function**: then the consistency $\mathbb{D}_{\mathrm{KL}}$'s and reconstruction term read

$$\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i,\boldsymbol{x}_0)||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i)) = \frac{1}{2\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2 ,$$

$$\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{X}_1) = -\frac{1}{2\tau_1^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_1) - \boldsymbol{x}_0\right\|^2 + C ,$$

such that (since $\boldsymbol{\mu}_1(\boldsymbol{X}_1,\boldsymbol{x}_0) = \boldsymbol{x}_0$ with $\alpha_0 = 1$)

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi}) &= -\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}\left\{\frac{1}{\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2\right\} \\ &\simeq -\frac{T}{2}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0),i\sim\mathcal{U}(1,T)}\left\{\frac{1}{\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2\right\} .\end{aligned}$$

DDPM Inference

- **Inference**: with optimal parameters $\boldsymbol{\theta}^*$ of the Neural Network $\boldsymbol{\mu}_{\boldsymbol{\theta}}$,

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}_0}\left\{\mathrm{ELBO}(\boldsymbol{X}_0;\boldsymbol{\theta},\boldsymbol{\phi})\right\}\\ &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\boldsymbol{x}_i),i\sim\mathcal{U}(1,T)}\left\{\frac{1}{2\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{X}_0)\right\|^2\right\},\end{aligned}$$

for $i = T, T-1, \dots 1$ do

$$\boldsymbol{X}_{i-1} = \boldsymbol{\mu}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i) + \tau_i \boldsymbol{Z}\,, \quad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})\,,$$

starting with $\boldsymbol{X}_T \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$.

Alternative training #1
"Predict image"

- Since:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i,\boldsymbol{x}_0)=\tau_i^2\left(\frac{\sqrt{\alpha_i}}{1-\alpha_i}\,\boldsymbol{x}_i+\frac{\sqrt{A_{i-1}}}{1-A_{i-1}}\,\boldsymbol{x}_0\right)\,,$$

  choose the reparameterization:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)=\boldsymbol{\mu}_i(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i))=\tau_i^2\left(\frac{\sqrt{\alpha_i}}{1-\alpha_i}\,\boldsymbol{x}_i+\frac{\sqrt{A_{i-1}}}{1-A_{i-1}}\,\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)\,;$$

- Then:

$$\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2=\frac{(1-\alpha_i)A_{i-1}}{(1-A_{i-1})(1-A_i)}\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{x}_0\right\|^2\,,$$

  and with $\Delta_i=\mathrm{SNR}_{i-1}-\mathrm{SNR}_i>0$, $\mathrm{SNR}_i=\frac{A_i}{1-A_i}$:

$$\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi})=-\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}\left\{\Delta_i\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{x}_0\right\|^2\right\}\,.$$

Alternative training[7] #1

Learning the noise schedule

- The *signal-to-noise ratio* SNR_i of $\boldsymbol{X}_i \sim \mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)$ is represented by a monotonically increasing Neural Network $\omega_{\boldsymbol{\phi}}(t)$:

$$\mathrm{SNR}_i \stackrel{\text{def}}{=} \frac{A_i}{1-A_i} = \mathrm{e}^{-\omega_{\boldsymbol{\phi}}(t_i)}\,,$$

  such that $A_i = \sigma(-\omega_{\boldsymbol{\phi}}(t_i))$ and $1 - A_i = \sigma(\omega_{\boldsymbol{\phi}}(t_i))$, where:

$$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$$

  is the usual sigmoid, or logistic activation function;
- Then $\boldsymbol{x} \to \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $t \to \omega_{\boldsymbol{\phi}}(t)$ are learnt altogether.

[7] D. P. Kingma, T. Salimans, B. Poole, J. Ho. Variational diffusion models. arXiv:2107.00630 (2021).

Alternative training[8] #2

"Predict noise"

- Since $\boldsymbol{X}_i = \sqrt{A_i}\boldsymbol{X}_0 + \sqrt{1-A_i}\boldsymbol{Z}$:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) \to \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{z}) = \frac{1}{\sqrt{\alpha_i}}\left(\boldsymbol{x}_i - \frac{1-\alpha_i}{\sqrt{1-A_i}}\,\boldsymbol{z}\right),$$

  choose the reparameterization:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \frac{1}{\sqrt{\alpha_i}}\left(\boldsymbol{x}_i - \frac{1-\alpha_i}{\sqrt{1-A_i}}\,\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right);$$

- Then:

$$\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i, \boldsymbol{z})\right\|^2 = \frac{(1-\alpha_i)^2}{\alpha_i(1-A_i)}\left\|\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{z}\right\|^2,$$

  and with $\lambda_i = \frac{\Delta_i}{\mathrm{SNR}_i} = \mathrm{e}^{\omega_{\boldsymbol{\phi}}(t_i)-\omega_{\boldsymbol{\phi}}(t_{i-1})} - 1$:

$$\mathrm{ELBO}(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\boldsymbol{Z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left\{\lambda_i\left\|\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}\left(\sqrt{A_i}\boldsymbol{x}_0 + \sqrt{1-A_i}\boldsymbol{Z}\right) - \boldsymbol{Z}\right\|^2\right\}.$$

[8] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. `arXiv:2006.11239` (2020); coined "reparameterization trick" in ML literature.

Alternative training #3

"Predict score"

- Since $\boldsymbol{X}_i|\boldsymbol{X}_0 \sim \mathcal{N}(\sqrt{A_i}\boldsymbol{X}_0, (1-A_i)\boldsymbol{I})$, *Tweedie's formula* yields $\mathbb{E}\{\sqrt{A_i}\boldsymbol{X}_0|\boldsymbol{X}_i\} = \boldsymbol{X}_i + (1-A_i)\boldsymbol{s}(\boldsymbol{X}_i, t_i)$ such that the reverse transition mean can be written:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) \to \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{s}) = \frac{1}{\sqrt{\alpha_i}}\left(\boldsymbol{x}_i + (1-\alpha_i)\boldsymbol{s}(\boldsymbol{x}_i, t_i)\right) ,$$

choose the reparameterization:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \frac{1}{\sqrt{\alpha_i}}\left(\boldsymbol{x}_i + (1-\alpha_i)\,\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right) ;$$

- Then (since $\boldsymbol{s} \equiv -\frac{\boldsymbol{z}}{\sqrt{1-A_i}}$):

$$\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i, \boldsymbol{z})\right\|^2 = \frac{(1-\alpha_i)^2}{\alpha_i}\left\|\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) + \frac{\boldsymbol{z}}{\sqrt{1-A_i}}\right\|^2 ,$$

and with $\lambda_i' = (1-\alpha_i)\frac{\mathrm{SNR}_{i-1}}{\mathrm{SNR}_i}$:

$$\mathrm{ELBO}(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left\{\lambda_i'\left\|\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}\left(\sqrt{A_i}\boldsymbol{x}_0 + \sqrt{1-A_i}\boldsymbol{Z}\right) + \frac{\boldsymbol{Z}}{\sqrt{1-A_i}}\right\|^2\right\}.$$

Tweedie's formula (or empirical Bayes)

- Let $\boldsymbol{Y} = \boldsymbol{X} + \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{X} \sim \pi_{\boldsymbol{X}}$;
- Then $\pi_{\boldsymbol{Y}}(\boldsymbol{y}) \propto \int \pi_{\boldsymbol{X}}(\boldsymbol{x}) \exp(-\frac{\|\boldsymbol{y}-\boldsymbol{x}\|^2}{2\sigma^2}) \mathrm{d}\boldsymbol{x}$ and:

$$\begin{aligned}
\frac{\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}\} - \boldsymbol{y}}{\sigma^2} &= \frac{\int \boldsymbol{x} \frac{\pi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}}{\pi_{\boldsymbol{Y}}(\boldsymbol{y})} - \boldsymbol{y}}{\sigma^2} \\
&= \frac{\int (\frac{\boldsymbol{x}-\boldsymbol{y}}{\sigma^2}) \pi_{\boldsymbol{X}}(\boldsymbol{x}) \exp(-\frac{\|\boldsymbol{y}-\boldsymbol{x}\|^2}{2\sigma^2}) \mathrm{d}\boldsymbol{x}}{\int \pi_{\boldsymbol{X}}(\boldsymbol{x}) \exp(-\frac{\|\boldsymbol{y}-\boldsymbol{x}\|^2}{2\sigma^2}) \mathrm{d}\boldsymbol{x}} \\
&= \boldsymbol{\nabla}_{\boldsymbol{y}} \log \int \pi_{\boldsymbol{X}}(\boldsymbol{x}) \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{x}\|^2}{2\sigma^2}\right) \mathrm{d}\boldsymbol{x} \\
&= \boldsymbol{\nabla}_{\boldsymbol{y}} \log \pi_{\boldsymbol{Y}}(\boldsymbol{y}) \\
&= \boldsymbol{s}_{\boldsymbol{Y}}(\boldsymbol{y});
\end{aligned}$$

- Thus $\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} = \boldsymbol{Y} + \sigma^2 \boldsymbol{s}_{\boldsymbol{Y}}(\boldsymbol{Y})$, which allows to estimate the score $\boldsymbol{s}_{\boldsymbol{Y}}$ given the data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{n=1}^N$.

Alternative inference

- For $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\mathbf{A}\boldsymbol{x} + \mathbf{b}, \mathbf{S})$, one has $\mathbb{P}(\boldsymbol{y}) = \int \mathbb{P}(\boldsymbol{y}|\boldsymbol{x})\mathbb{P}(\mathrm{d}\boldsymbol{x}) = \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{S} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathsf{T}})$.
- Therefore starting from:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_i; \sqrt{A_i}\boldsymbol{x}_0, (1 - A_i)\boldsymbol{I}) \,,$$
$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{i-1}; \sqrt{A_{i-1}}\boldsymbol{x}_0, (1 - A_{i-1})\boldsymbol{I}) \,,$$

  with a noise schedule $\boldsymbol{\phi} = (\alpha_1, \alpha_2, \dots \alpha_T)$ where $A_i = \prod_{j=1}^{i} \alpha_j$, one may *choose* $\mathbf{A}$ and $\mathbf{b}$ arbitrarily such that:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{i-1}; \mathbf{A}\boldsymbol{x}_i + \mathbf{b}, \boldsymbol{\Sigma}_i)$$

  with a variance $\boldsymbol{\Sigma}_i$ independent of $\boldsymbol{\phi}$.

DDIM inference[9]

- In Denoising Diffusion Implici Models (DDIM) one chooses $\mathbf{A} = \mathrm{A}\boldsymbol{I}$, $\mathbf{b} = \mathrm{b}\boldsymbol{x}_0$, and $\boldsymbol{\Sigma}_i = \sigma_i^2\boldsymbol{I}$ with:

$$\mathrm{A} = \sqrt{\frac{1 - A_{i-1} - \sigma_i^2}{1 - A_i}}\,, \quad \mathrm{b} = \sqrt{A_{i-1}} - \mathrm{A}\sqrt{A_i}\,,$$

such that $\mathbb{Q}_\phi(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{i-1}; \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0), \sigma_i^2\boldsymbol{I})$ where:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) = \sqrt{\frac{1 - A_{i-1} - \sigma_i^2}{1 - A_i}}\boldsymbol{x}_i + \left(\sqrt{A_{i-1}} - \sqrt{\frac{A_i(1 - A_{i-1} - \sigma_i^2)}{1 - A_i}}\right)\boldsymbol{x}_0\,.$$

- The underlying forward process is no longer Markovian (since $\boldsymbol{x}_i$ could depend on both $\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_0$), but denoising levels can be controlled through the independent parameters σ_i^2.

[9] J. Song, C. Meng, S. Ermon. Denoising diffusion implicit models. `arXiv:2010.02502` (2020).

DDIM inference

- Since $\boldsymbol{X}_i = \sqrt{A_i}\boldsymbol{X}_0 + \sqrt{1-A_i}\boldsymbol{Z}$:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) \to \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{z}) = \sqrt{\frac{A_{i-1}}{A_i}}\left(\boldsymbol{x}_i - \sqrt{1-A_i}\boldsymbol{z}\right) + \sqrt{1-A_{i-1}-\sigma_i^2}\,\boldsymbol{z}\,;$$

- Then DDIM inference reads:

$$\begin{aligned}\boldsymbol{X}_{i-1} &= \sqrt{\frac{A_{i-1}}{A_i}}\left(\boldsymbol{X}_i - \sqrt{1-A_i}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i)\right) \\ &\quad + \sqrt{1-A_{i-1}-\sigma_i^2}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i) + \sigma_i\boldsymbol{Z}\,,\end{aligned}$$

whereas DDPM inference was:

$$\boldsymbol{X}_{i-1} = \frac{1}{\sqrt{\alpha_i}}\left(\boldsymbol{X}_i - \frac{1-\alpha_i}{\sqrt{1-A_i}}\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i)\right) + \tau_i\boldsymbol{Z}\,,$$

where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{\theta}^*$ are optimal parameters for the neural network $\boldsymbol{z}_{\boldsymbol{\theta}}$ predicting the noise.

Accelerating DDIM inference

- Intermediate steps $i-1, i-2\dots$ up to an index $j<i$ can be skipped:

$$\boldsymbol{X}_j = \sqrt{\frac{A_j}{A_i}}\left(\boldsymbol{X}_i - \sqrt{1-A_i}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i)\right) + \sqrt{1-A_j-\sigma_i^2}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i) + \sigma_i \boldsymbol{Z}\,;$$

- Choosing $\sigma_i = 0$ makes DDIM deterministic:

$$\boldsymbol{X}_{i-1} = \sqrt{\frac{A_{i-1}}{A_i}}\left(\boldsymbol{X}_i - \sqrt{1-A_i}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i)\right) + \sqrt{1-A_{i-1}}\,\boldsymbol{z}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i)\,.$$

SMLD in continuous setting[10]

- Forward process with time-varying noise $\beta(t) > 0$:

$$\mathrm{d}\boldsymbol{X}_t = \sqrt{\beta(t)}\,\mathrm{d}\boldsymbol{W}_t\,, \quad \boldsymbol{X}_0 \sim \pi_0\,.$$

- The solution reads:

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \sqrt{\beta(s)}\mathrm{d}\boldsymbol{W}_s\,,$$

with moments:

$$\mathbb{E}\{\boldsymbol{X}_t|\boldsymbol{X}_0\} = \boldsymbol{X}_0\,,$$
$$\mathbb{V}\{\boldsymbol{X}_t|\boldsymbol{X}_0\} = \left(\int_0^t \beta(s)\mathrm{d}s\right)\boldsymbol{I}\,.$$

[10] Called Variance Exploding (VE) SDE in: Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. arXiv:2011.13456 (2020).

SMLD in continuous setting

- Backward process $\overleftarrow{\boldsymbol{X}}_t \equiv \boldsymbol{X}_{T-t}$:

$$\mathrm{d}\overleftarrow{\boldsymbol{X}}_t = +\beta(t)\boldsymbol{s}(\overleftarrow{\boldsymbol{X}}_t, T-t)\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\boldsymbol{W}_t\,, \quad \overleftarrow{\boldsymbol{X}}_0 = \boldsymbol{X}_T\,,$$

  which looks like (stochastic) gradient descent on $\log\pi$ with learning rate $\beta(t)$ since Stein's score function reads $\boldsymbol{s}(\boldsymbol{x},t) = \boldsymbol{\nabla}_{\boldsymbol{x}} \log\pi(\boldsymbol{x},t)$.

SMLD encoder

- Given a *variance schedule* $\boldsymbol{\phi} = (\sigma_1^2, \sigma_2^2, \dots \sigma_T^2)$ such that $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_T^2$, the *encoder* is the Markov chain $\boldsymbol{X}_{1:T}$:

$$\boldsymbol{X}_i = \boldsymbol{X}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\,\boldsymbol{Z}\,, \quad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\,,$$

starting from $\boldsymbol{X}_0 \sim \pi_0$, with transition probabilities:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{x}_{i-1}, (\sigma_i^2 - \sigma_{i-1}^2)\boldsymbol{I})\,,$$
$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{x}_0, \sigma_i^2\boldsymbol{I})\,;$$

- $\boldsymbol{X}_{1:T}$ also has reverse transition probability:

$$\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{i-1}; \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0), \tau_i^2\boldsymbol{I}\right)\,,$$

where:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) = \frac{1}{\sigma_i^2}\left(\sigma_{i-1}^2\,\boldsymbol{x}_i + (\sigma_i^2 - \sigma_{i-1}^2)\,\boldsymbol{x}_0\right)\,, \quad \tau_i^2 = \frac{\sigma_{i-1}^2}{\sigma_i^2}(\sigma_i^2 - \sigma_{i-1}^2)\,.$$

Training

- **Encoder**: choose the variance schedule $\boldsymbol{\phi}$, so there is nothing left to learn in $\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_T|\boldsymbol{x}_0)||\mathbb{P}(\boldsymbol{x}_T))$;
- **Decoder**: choose $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_{i-1};\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i),\tau_i^2\boldsymbol{I})$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is a Neural Network;
- **Loss function**: then the consistency $\mathbb{D}_{\mathrm{KL}}$'s and reconstruction term read

$$\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i,\boldsymbol{x}_0)||\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1}|\boldsymbol{X}_i)) = \frac{1}{2\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2\,,$$

$$\log\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{X}_1) = -\frac{1}{2\tau_1^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_1)-\boldsymbol{x}_0\right\|^2 + C\,,$$

such that (since $\boldsymbol{\mu}_1(\boldsymbol{X}_1,\boldsymbol{x}_0)=\boldsymbol{x}_0$ with $\sigma_0=0$)

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi}) &= -\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}\left\{\frac{1}{\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2\right\} \\ &\simeq -\frac{T}{2}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0),i\sim\mathcal{U}(1,T)}\left\{\frac{1}{\tau_i^2}\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\right\|^2\right\}.\end{aligned}$$

Alternative training #1

"Predict image"

- Since:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i,\boldsymbol{x}_0)=\frac{1}{\sigma_i^2}\left(\sigma_{i-1}^2\,\boldsymbol{x}_i+(\sigma_i^2-\sigma_{i-1}^2)\,\boldsymbol{x}_0\right)\,,$$

choose the reparameterization:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)=\boldsymbol{\mu}_i(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i))=\frac{1}{\sigma_i^2}\left(\sigma_{i-1}^2\,\boldsymbol{x}_i+(\sigma_i^2-\sigma_{i-1}^2)\,\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)\,;$$

- Then:

$$\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{\mu}_i(\boldsymbol{X}_i,\boldsymbol{x}_0)\|=\frac{\sigma_i^2-\sigma_{i-1}^2}{\sigma_i^2}\,\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{x}_0\|\,,$$

and with $\Delta_i=1/\sigma_{i-1}^2-1/\sigma_i^2$:

$$\mathrm{ELBO}(\boldsymbol{x}_0;\boldsymbol{\theta},\boldsymbol{\phi})=-\frac{1}{2}\sum_{i=1}^{T}\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\phi}}(\boldsymbol{x}_i|\boldsymbol{x}_0)}\{\Delta_i\,\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i)-\boldsymbol{x}_0\|^2\}\,.$$

Alternative training[11] #2

"Predict noise"

SDE

W. Genuist, É. Savin

SDE
First-order systems
Stochastic integrals
Diffusion processes

Numerical solutions
Stochastic modeling
Numerical schemes

GAN

VAE

Diffusion & ML
Continuous case
DDPM
SMLD
Guidance
CM

- Since $\boldsymbol{X}_i = \boldsymbol{X}_0 + \sigma_i \boldsymbol{Z}$:

$$\boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{x}_0) \to \boldsymbol{\mu}_i(\boldsymbol{x}_i, \boldsymbol{z}) = \boldsymbol{x}_i - \frac{\sigma_i^2 - \sigma_{i-1}^2}{\sigma_i}\, \boldsymbol{z}\,,$$

  choose the reparameterization:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \boldsymbol{x}_i - \frac{\sigma_i^2 - \sigma_{i-1}^2}{\sigma_i}\, \hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\,;$$

- Then:

$$\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{\mu}_i(\boldsymbol{X}_i, \boldsymbol{z})\right\| = \frac{\sigma_i^2 - \sigma_{i-1}^2}{\sigma_i} \left\|\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{X}_i) - \boldsymbol{z}\right\|\,,$$

  and with $\lambda_i = \sigma_i^2/\sigma_{i-1}^2 - 1$:

$$\mathrm{ELBO}(\boldsymbol{x}_0; \boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\sum_{i=1}^{T} \mathbb{E}_{\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})} \left\{\lambda_i \left\|\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_0 + \sigma_i \boldsymbol{Z}\right) - \boldsymbol{Z}\right\|^2\right\}\,.$$

[11] Y. Song, S. Ermon. Generative modeling by estimating gradients of the data distribution. arXiv:1907.05600 (2019).

Inference

- **Inference**: with optimal parameters $\boldsymbol{\theta}^*$ of the Neural Network $\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}$,

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}_0}\left\{\mathrm{ELBO}(\boldsymbol{X}_0;\boldsymbol{\theta},\boldsymbol{\phi})\right\} \\ &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}_0,\boldsymbol{Z},i\sim\mathcal{U}(1,T)}\left\{\lambda_i \left\|\hat{\boldsymbol{z}}_{\boldsymbol{\theta}}\left(\boldsymbol{X}_0+\sigma_i\boldsymbol{Z}\right)-\boldsymbol{Z}\right\|^2\right\},\end{aligned}$$

for $i = T, T-1, \dots 1$ do

$$\boldsymbol{X}_{i-1} = \boldsymbol{X}_i - \frac{\sigma_i^2-\sigma_{i-1}^2}{\sigma_i}\,\hat{\boldsymbol{z}}_{\boldsymbol{\theta}^*}(\boldsymbol{X}_i) + \tau_i\boldsymbol{Z}\,, \quad \boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})\,,$$

starting with $\boldsymbol{X}_T \sim \mathcal{N}(\boldsymbol{0},\sigma_T^2\boldsymbol{I})$.

Conditional generation

- In practical applications conditional generation is required $\boldsymbol{X}_0 \sim \pi_0(\boldsymbol{x}_0|\boldsymbol{Y})$ for some constraints $\boldsymbol{Y}$.
- The noising and denoising processes are unchanged, provided the score function is evaluated as the conditional score $\boldsymbol{s}(\boldsymbol{x},t|\boldsymbol{y}) = \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{x},t|\boldsymbol{y})$ such that (since $\mathbb{P}(\boldsymbol{X},\boldsymbol{Y}) = \mathbb{P}(\boldsymbol{X}|\boldsymbol{Y})\mathbb{P}(\boldsymbol{Y}) = \mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})\mathbb{P}(\boldsymbol{X})$):

$$
\begin{aligned}
\boldsymbol{s}(\boldsymbol{x},t|\boldsymbol{y}) &= \boldsymbol{s}(\boldsymbol{x},t) + \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{y}|\boldsymbol{x},t) \\
&= \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{x},t,\boldsymbol{y})\,.
\end{aligned}
$$

- The adversarial gradient $\boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{y}|\boldsymbol{x},t)$ is then scaled to leverage the level to which the constraints $\boldsymbol{y}$ are adhered to in so-called classifier and classifier-free guidances.

Classifier guidance[12]

- In *classifier guidance* a guidance scale $\lambda > 0$ is introduced to make sure that the constraints $\boldsymbol{y}$ are more closely fulfilled:

$$\boldsymbol{s}_\lambda(\boldsymbol{x}, t|\boldsymbol{y}) = \boldsymbol{s}(\boldsymbol{x}, t) + \lambda \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{y}|\boldsymbol{x}, t)\,,$$

  where the posterior $\pi(\boldsymbol{y}|\boldsymbol{X}_t)$ is trained beforehand–a classifier for example.

- The classifier may become ineffective for strongly noised images, hence a classifier-free guidance shall be preferred.

[12] P. Dhariwal, A. Nichol. Diffusion models beat GANs on image synthesis. `arXiv:2105.05233` (2021).

Classifier-free guidance[13]

- In *classifier-free guidance* the joint distribution $\mathbb{P}(\boldsymbol{X}_t, \boldsymbol{Y})$ is directly trained using pairs $(\boldsymbol{X}_i^{(j)}, \boldsymbol{Y}^{(j)})_{j=1}^N$, $0 < i \leqslant T$, and the conditional score reads:

$$\boldsymbol{s}_\lambda(\boldsymbol{x}, t|\boldsymbol{y}) = (1-\lambda)\boldsymbol{s}(\boldsymbol{x}, t) + \lambda \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{x}, t, \boldsymbol{y}) \,,$$

where $\boldsymbol{s}(\boldsymbol{x}, t) \equiv \boldsymbol{\nabla}_{\boldsymbol{x}} \log \pi(\boldsymbol{x}, t, \boldsymbol{y} = \varnothing)$ (*i.e.* perform random dropout on the conditioning information to learn both scores altogether).

[13] J. Ho, T. Salimans. Classifier-free diffusion guidance. arXiv:2207.12598 (2022).

Conditional Ornstein-Uhlenbeck process

- Forward Ornstein-Uhlenbeck process with mean constraint $\boldsymbol{\mu}(\boldsymbol{Y})$ and variance constraint $\sigma(\boldsymbol{Y})$:

$$\mathrm{d}\boldsymbol{X}_t = -\frac{1}{2}\beta(t)(\boldsymbol{X}_t - \boldsymbol{\mu}(\boldsymbol{Y}))\mathrm{d}t + \sqrt{\sigma^2(\boldsymbol{Y})\beta(t)}\,\mathrm{d}\boldsymbol{W}_t\,.$$

- The solution reads:

$$\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{Y}) + \mathrm{e}^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{X}_0 - \boldsymbol{\mu}(\boldsymbol{Y})) + \sigma(\boldsymbol{Y})\int_0^t \mathrm{e}^{-\frac{1}{2}\int_s^t \beta(s')\mathrm{d}s'}\sqrt{\beta(s)}\mathrm{d}\boldsymbol{W}_s\,,$$

with moments:

$$\mathbb{E}\{\boldsymbol{X}_t|\boldsymbol{X}_0,\boldsymbol{Y}\} = \boldsymbol{\mu}(\boldsymbol{Y}) + \mathrm{e}^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{X}_0 - \boldsymbol{\mu}(\boldsymbol{Y}))\,,$$

$$\mathbb{V}\{\boldsymbol{X}_t|\boldsymbol{X}_0,\boldsymbol{Y}\} = \sigma^2(\boldsymbol{Y})(1 - \mathrm{e}^{-\int_0^t \beta(s)\mathrm{d}s})\boldsymbol{I}\,.$$

Conditional Ornstein-Uhlenbeck process

- Backward Ornstein-Uhlenbeck process $\overleftarrow{\boldsymbol{X}}_t \equiv \boldsymbol{X}_{T-t}$:

$$\mathrm{d}\overleftarrow{\boldsymbol{X}}_t = +\frac{1}{2}\beta(t)\left(\overleftarrow{\boldsymbol{X}}_t - \boldsymbol{\mu}(\boldsymbol{Y}) + 2\sigma(\boldsymbol{Y})\boldsymbol{s}(\overleftarrow{\boldsymbol{X}}_t, T-t|\boldsymbol{Y})\right)\mathrm{d}t + \sqrt{\sigma^2(\boldsymbol{Y})\beta(t)}\mathrm{d}\boldsymbol{W}_t\,,$$

  starting from $\overleftarrow{\boldsymbol{X}}_0|\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{Y}), \sigma^2(\boldsymbol{Y})\boldsymbol{I})$.
- The conditional score $\boldsymbol{s}(\boldsymbol{x}, t|\boldsymbol{y}) \simeq \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t|\boldsymbol{y})$ can be trained by denoising score matching:

$$J_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}_0,\boldsymbol{Y},\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),t\sim\mathcal{U}(0,T)}\left\{\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{\mu}_t(\boldsymbol{X}_0,\boldsymbol{Y}) + \sigma_t(\boldsymbol{Y})\boldsymbol{Z}, t|\boldsymbol{Y}\right) + \frac{\boldsymbol{Z}}{\sigma_t(\boldsymbol{Y})}\right\|^2\right\}\,,$$

  where:

$$\boldsymbol{\mu}_t(\boldsymbol{x}_0, \boldsymbol{y}) = \boldsymbol{\mu}(\boldsymbol{y}) + \mathrm{e}^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{x}_0 - \boldsymbol{\mu}(\boldsymbol{y}))\,,$$

$$\sigma_t(\boldsymbol{y}) = \sigma(\boldsymbol{y})\sqrt{1 - \mathrm{e}^{-\int_0^t \beta(s)\mathrm{d}s}}\,.$$

Consistency models[14]

- Recall backward Langevin diffusion setting $\boldsymbol{Y}_t = \boldsymbol{X}_{T-t}$:

$$\mathrm{d}\boldsymbol{Y}_t = (\boldsymbol{Y}_t + (1+\alpha)\boldsymbol{s}(\boldsymbol{Y}_t, T-t))\mathrm{d}t + \sqrt{2\alpha}\,\mathrm{d}\boldsymbol{W}_t\,,$$

  starting from $\boldsymbol{Y}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

- The associated probability flow ODE thus reads:

$$\mathrm{d}\boldsymbol{Y}_t = (\boldsymbol{Y}_t + \boldsymbol{s}(\boldsymbol{Y}_t, T-t))\mathrm{d}t\,.$$

- There exists an ODE flow $\boldsymbol{F} : \mathbb{R}_+ \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ such that:

$$\boldsymbol{Y}_t = \boldsymbol{F}(t-s, \boldsymbol{Y}_s)\,, \quad 0 \leqslant s < t \leqslant T\,.$$

- The goal is now to learn the flow $\boldsymbol{F}(t,\boldsymbol{y}) \simeq \boldsymbol{F}_{\boldsymbol{\Theta}}(t,\boldsymbol{y})$ so that $\boldsymbol{Y}_T = \boldsymbol{F}_{\boldsymbol{\Theta}}(T, \boldsymbol{Y}_0)$ can be sampled in one step.

[14] Y. Song, P. Dhariwal, M. Chen, I. Sutskever. Consistency models. `arXiv:2303.01469` (2023).

Training the ODE flow

- Take $t \sim \mathcal{U}(0,T)$ and sample $\boldsymbol{Y}_+ \overset{d}{=} \boldsymbol{X}_t$ of the forward process from $\boldsymbol{X}_0 \sim \pi_0$;
- For some small $\delta_t > 0$:
  - *Consistency training*: set $\boldsymbol{Y}_- \overset{d}{=} \boldsymbol{X}_{t-\delta_t}$ which is sampled independently from $\boldsymbol{Y}_+$;
  - *Consistency distillation*: from the probability flow ODE set
  $$\boldsymbol{Y}_- = (1+\delta_t)\boldsymbol{Y}_+ + \delta_t \boldsymbol{s}(\boldsymbol{Y}_+, T-t)\,,$$
  which requires a pretrained score but performs better;
- But $\boldsymbol{Y}_T = \boldsymbol{F}_{\boldsymbol{\Theta}}(t, \boldsymbol{Y}_{T-t}) = \boldsymbol{F}_{\boldsymbol{\Theta}}(t-\delta_t, \boldsymbol{Y}_{T-t+\delta_t})$, thus the training loss $\mathcal{L}$ to be minimized is chosen as:

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{X}_0, t\sim\mathcal{U}(0,T)}\left\{\lambda(t)\left\|\boldsymbol{F}_{\boldsymbol{\Theta}}(t,\boldsymbol{Y}_+) - \boldsymbol{F}_{\boldsymbol{\Theta}}(t-\delta_t,\boldsymbol{Y}_-)\right\|^2\right\}\,,$$

where $\lambda(t)$ is a weighting function.

To be continued...

- Boundary conditions (Maarten's question);
- (Schrödinger) bridges;
- Manifolds...

Further reading...

Stochastic differential equations:

- A. Friedman: *Stochastic Differential Equations and Applications*, vol. I & II, Academic Press (1975);
- P. E. Kloeden, E. Platen: *Numerical Solution of Stochastic Differential Equations*, 3rd ed., Springer (1999);
- J.-F. Le Gall: *Mouvement Brownien, Martingales et Calcul Stochastique*, Springer (2013);
- B. K. Øksendal: *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer (2003);
- S. Särkkä, A. Solin: *Applied Stochastic Differential Equations*, Cambridge University Press (2019);
- C. Soize: *The Fokker-Planck Equation for Stochastic Dynamical Systems and its Explicit Steady State Solutions*, World Scientific (1994);
- D. Talay: Simulation of stochastic differential systems. In *Probabilistic Methods in Applied Physics* (P. Krée & W. Wedig, eds.), pp. 54-96. Lecture Notes in Physics **451**, Springer (1995).

Further reading...

Variational diffusion models:

- S. H. Chan. Tutorial on diffusion models for imaging and vision. `arXiv:2403.18103` (2024);
- V. de Bortoli. Generative modeling. `https://vdeborto.github.io/project/generative_modeling/` (2023);
- H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, S. Z. Li. A survey on generative diffusion models, *IEEE Trans. Knowl. Data Eng.* **36**(7), 2814-2830 (2024);
- Z. Chang, G. A. Koulieris, H. P. H. Shum. On the design fundamentals of diffusion models: A Survey. `arXiv:2306.04542` (2023);
- M. Chen, S. Mei, J. Fan, M. Wang. An overview of diffusion models: applications, guided generation, statistical rates and optimization. `arXiv:2404.07771` (2024);
- T. Duan. Diffusion models from scratch. `http://www.tonyduan.com/diffusion/` (2023);
- T. Karras, M. Aittala, T. Aila, S. Laine. Elucidating the design space of diffusion-based generative models. `arXiv:2206.00364` (2022);
- C. Luo. Understanding diffusion models: A unified perspective. `arXiv:2208.11970` (2022);

Further reading...

Variational diffusion models (cont'd):

- G. Peyré. Denoising diffusion models. Mathematical Tours (2023); jupyter notebook;
- Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. `arXiv:2011.13456` (2020);
- F. de Souza Ribeiro, B. Glocker. Demystifying variational diffusion models. `arXiv:2401:06281` (2024);
- W. Tang, H. Zhao. Score-based diffusion models via stochastic differential equations – a technical tutorial. `arXiv:2402.07487` (2024);
- R. E. Turner, C.-D. Diaconu, S. Markou, A. Shysheya, A. Y. K. Foong, B. Mlodozeniec. Denoising diffusion probabilistic models in six simple steps. `arXiv:2402.04384` (2024);
- L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. `arXiv:2209.00796` (2022); *ACM Comput. Surv.* **56**(4), 105 (2023).