Diffusion-based generative modeling

For statisticians and probabilists

Eddie Aamari

April 11, 2025

The goal of these notes is to construct a generative method to "sample approximately" from an unknown distribution $p^*(x)dx$ from which we have observed an i.i.d. n-sample $X_1, \dots, X_n$.

We will present a family of methods often referred to as "diffusion models". Their popularity have blown up over the Summer of 2022, with the release of *Stable diffusions* for images. Although pioneering works can be found in physics before that, the rise of such methods can be dated back to [SSDK+20], which catch phrase is:

> Creating noise from data is easy; creating data from noise is generative modeling.

The global idea is to add noise to data incrementally while learning to denoise at each step, and then reverse the whole process (see Figure 1).

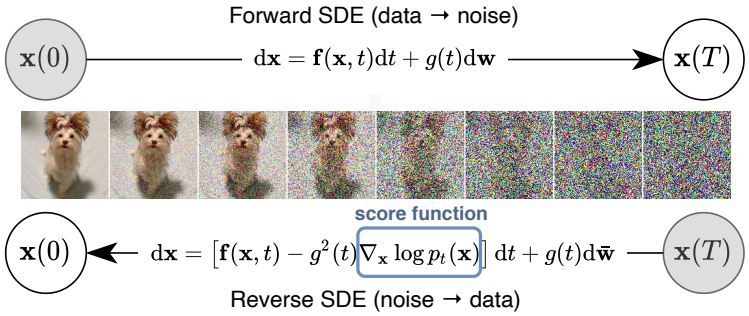

Figure 1: Big picture of generative modeling (taken from [SSDK+20]).

Contents

1 Stochastic calculus survival kit

There are two main ways to formalize diffusions generative models for quantitative data. One uses discrete time increments [HJA20] and requires knowledge on Markov chains only, but it does not yield a clear mathematical framework. We opt for the other way, which uses continuous time increments [SSDK+20] and requires tools from stochastic calculus. It will yield a quite unified functional framework to hold on to.

This section gives a minimal overview of stochastic calculus. To make the presentation lighter, we purposely leave all the convergence and measurability issues under the carpet. If you feel scammed, you shall find all the necessary mathematical details in Jean-François Le-Gall's book [LG16].

1.1 Brownian motion

1.1.1 Definition

Given a measurable space $(E, \mathcal{E})$ and an arbitrary index set $\mathcal{T}$, a *random process indexed by $\mathcal{T}$ with values in E* is a collection $(X_t)_{t\in\mathcal{T}}$ of random variables with values in E. Among such processes, we will focus on those with Gaussian marginals.

Definition 1.1 (Gaussian process)**.** *A (real-valued) random process is called a (centered) Gaussian process if any finite linear combination of the variables $(X_t)_{t\in\mathcal{T}}$ is centered Gaussian.*

The distribution of a centered Gaussian process is fully determined by its covariance kernel

$$K(s,t) := \mathbb{E}\left[X_s X_t\right], \quad \text{for all } s,t \in \mathcal{T}.$$

The main building block of stochastic calculus is the so-called Brownian motion, which we first present in dimension $d = 1$.

Definition 1.2 (Brownian motion)**.** *There exists a process $(B_t)_{t\geq 0}$ called Brownian motion, which is a centered Gaussian process over $\mathcal{T} = \mathbb{R}_+$ with continuous sample paths $t \mapsto B_t$ and such that any of the following equivalent properties holds.*

- *$B_0 = 0$ a.s., and for all $0 \leq s < t$, the random variable $B_t - B_s$ is independent of the σ-field $\mathcal{F}_s := \sigma\left(B_r, r \leq s\right)$ and distributed according to $\mathcal{N}(0, t-s)$.*
- *$B_0 = 0$ a.s., and for all $0 \leq t_0 < t_1 < \ldots < t_p$, the random variables $(B_{t_j} - B_{t_{j-1}})_j$ are independent and distributed according to $\mathcal{N}(0, t_j - t_{j-1})$.*
- *For all $s, t \geq 0$, $K(s,t) = s \wedge t$.*

Proof. See [LG16, Proposition 2.3] for the equivalences, and [LG16, Exercise 1.18] for Lévy's construction. A more geometric construction on $\mathcal{T} = [0,1]$ uses Donsker's invariance principle. It is based on a iid sequence $(X_i)_{i\in\mathbb{N}}$ of centered real random variables with unit variance. Define the piecewise-linear continuous process

$$Z_n(u) := \sum_{i=1}^{\lceil u \rceil} U_i \psi(u-i), u \in [0,1],$$

where $\psi(v) := \min\{1, \max\{0, v\}$. Then $(B_t)_{t\in[0,1]}$ can be constructed as the limit in distribution of the sequence of processes $\left(\frac{1}{\sqrt{n}} Z_n(nt)\right)_{t\in[0,1]}$.

□

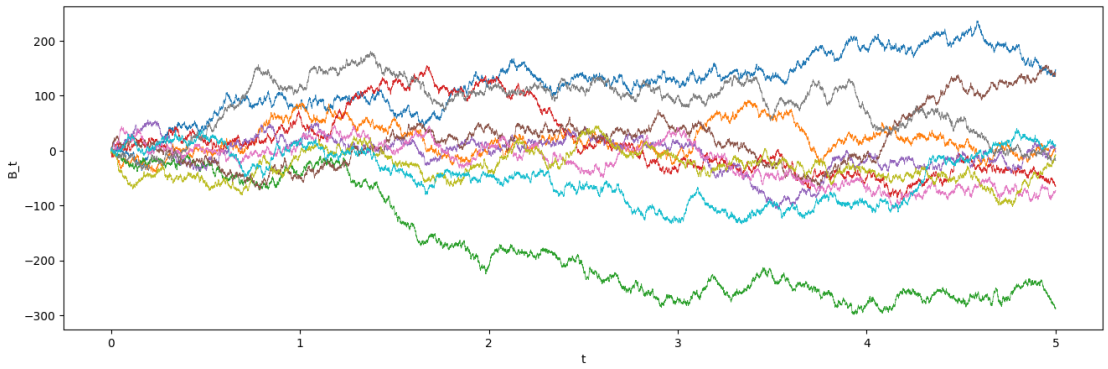

Figure 2: Ten trajectories of a Brownian motion.

See Figure 1.1.1 for an illustration of sample paths of $(B_t)_t$. As Definition 1.2 suggests, we will be dealing with measurability of random variables with respect to σ-fields indexed by (time) $t \in \mathcal{T}$. Hence, some vocabulary is in order.

Definition 1.3 (Filtration, adapted process)**.**

- *A* filtration *over $\mathcal{T} \subset \mathbb{R}$ is an increasing family $(\mathcal{F}_t)_{t\in\mathcal{T}}$ of σ-fields, i.e. $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$ with $s, t \in \mathcal{T}$.*
- *A stochastic process $(X_t)_{t\in\mathcal{T}}$ is said to be* adapted *to a filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ if for all $s \in \mathcal{T}$, X_t is $\mathcal{F}_t$-measurable.*

1.1.2 Regularity properties of the Brownian motion

Among the many nice properties that the Brownian motion exhibits, let us point out three of the most important ones.

- *(Martingale property)* The first characterization of Definition 1.2 yields that the Brownian motion is a martingale adapted to the filtration $\left(\mathcal{F}_s := \sigma(X_r, r \leq s)\right)_{s\geq 0}$, since for all $0 \leq s \leq t$,

$$\begin{aligned}\mathbb{E}\left[B_t \mid \mathcal{F}_s\right] &= \mathbb{E}\left[B_s \mid \mathcal{F}_s\right] + \mathbb{E}\left[B_t - B_s \mid \mathcal{F}_s\right] \\ &= B_s + \mathbb{E}\left[B_t - B_s\right] \\ &= B_s.\end{aligned}$$

- *(Hölder smoothness)* By definition, a Brownian motion has sample paths $t \mapsto B_t(\omega)$ that are continuous for almost all ω. In fact, they can be shown to be more regular. They are locally Hölder continuous with exponent $1/2 - \delta$ for all $0 < \delta < 1/2$, in the sense that $|B_t - B_s| \lesssim |t-s|^{1/2-\delta}$ a.s. (see [LG16, Corollary 2.11]). This essentially comes from the fact that for all $t \geq s \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{B_t - B_s}{\sqrt{t-s}}\right)^2\right] &= \frac{\mathbb{E}\left(B_t - B_s\right)^2}{t-s} \\ &= \frac{K(t,t) + K(s,s) - 2K(s,t)}{t-s} \\ &= 1.\end{aligned}$$

One can also show that this Hölder exponent is optimal, in the sense that for all $\delta > 0$, $(B_t)_t$ is a.s. *not* Hölder continuous with exponent $1/2 + \delta$, even locally.

- *(Quadratic variation)* Samples paths of $(B_t)_t$ being not more than 1/2-Hölder everywhere, they do not have finite length. In fact, for all sequence of subdivisions $0 = t_0^n < t_1^n < \ldots < t_{p_n}^n = t$ of $[0, t]$ whose maximal spacing $\max_{1 \leq j \leq p_n} |t_j - t_{j-1}|$ tends to zero as $n \to \infty$, we have

$$\sum_{j=1}^{p_n} |B_{t_j^n} - B_{t_{j-1}^n}| \xrightarrow[n\to\infty]{a.s.} \infty.$$

  We say that $(B_t)_t$ has infinite *first variation*. However, we can show that its *quadratic variation* is always well defined and deterministic. More precisely, we have

$$\sum_{j=1}^{p_n} (B_{t_j^n} - B_{t_{j-1}^n})^2 \xrightarrow[n\to\infty]{L^2} t.$$

1.2 Itô stochastic integral

Since $(B_t)_t$ exhibits infinite *first* variation, it is not possible to define the integral $\int_s^t \phi(u) \mathrm{d}B_u$ of a (smooth enough) function $\phi : \mathbb{R} \to \mathbb{R}$ as a special case of the usual Stieltjes integral. For $(F_t)_t$ of finite first variation [LG16, Section 4.1.1], this integral is characterized by the fact that it satisfies the fundamental theorem of calculus asserting that for all $\Phi \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$,

$$\Phi(F_t) = \Phi(F_s) + \int_s^t \Phi'(F_u) \underbrace{\mathrm{d}F_u}_{F_u' \mathrm{d}u}.$$

Equivalently, it is not straightforward to define a notion of differential $\mathrm{d}B_t$, which would satisfy a similar chain rule as $\mathrm{d}\Phi(F_t) = \Phi'(F_t)\mathrm{d}F_t$.

However, we can give this integral a meaning through the fact that its *quadratic* variation is finite. This will yield a tweaked fundamental theorem of calculus called *Itô's formula* (see Theorem 1.14). The standard construction of this integral goes through the following elementary processes, which play the role of *simple functions* in Lebesgue's integral.

Definition 1.4 (Elementary stochastic process)**.** *A stochastic process $(X_t)_{t \in [a,b)}$ is said to be* elementary *if there exist deterministic values $a = t_0 < t_1 < \ldots < t_p = b$ and random variables $(X_j)_{0 \leq j \leq p-1}$ such that for all $t \in [a, b)$,*

$$X_t = \sum_{j=1}^{p} X_{j-1} \mathbb{1}_{[t_{j-1}, t_j)}(t).$$

Said otherwise, an elementary process is a piecewise constant random process. With the above convention of notation, we have $X_{t_j} = X_j$ for all $j < p$. The integral against the Brownian motion is naturally defined as the weighted increments on each of its constant pieces.

Definition 1.5 (Itô integral of an elementary process)**.** *If $(X_t)_t$ is an elementary process as in Definition 1.4, define*

$$\int_a^b X_t \mathrm{d}B_t := \sum_{j=1}^{p} X_{t_{j-1}} (B_{t_j} - B_{t_{j-1}}).$$

As a first elementary remark, let us point out that $\int_a^b \mathrm{d}B_t = B_b - B_a$, which motivates notation $\mathrm{d}B_t$. In fact, the above proto-integral fulfills a few desirable properties that an actual integral should satisfy.

Proposition 1.6. *Let $(X_t)_t$ and $(Y_t)_t$ be elementary processes indexed by $[a,b]$, adapted to the natural filtration $(\sigma(B_r, r \leq s))_s$ of the Brownian motion.*

- *(Linearity) For all $\lambda, \mu \in \mathbb{R}$,*

$$\int_a^b \lambda X_t + \mu Y_t \mathrm{d}B_t = \lambda \int_a^b X_t \mathrm{d}B_t + \mu \int_a^b Y_t \mathrm{d}B_t.$$

- *(Centering) If $\mathbb{E}[|X_t|] < \infty$ for all $t \in [a,b]$, then $\mathbb{E}\left[\left|\int_a^b X_t \mathrm{d}B_t\right|\right] < \infty$, and*

$$\mathbb{E}\left[\int_a^b X_t \mathrm{d}B_t\right] = 0.$$

- *(Square integrability and isometry) If $\mathbb{E}[X_t^2] < \infty$ for all $t \in [a,b]$, then $\mathbb{E}\left[\left(\int_a^b X_t \mathrm{d}B_t\right)^2\right] < \infty$. If furthermore $\mathbb{E}[Y_t^2] < \infty$ for all $t \in [a,b]$, then*

$$\mathbb{E}\left[\left(\int_a^b X_t \mathrm{d}B_t\right)\left(\int_a^b Y_t \mathrm{d}B_t\right)\right] = \int_a^b \mathbb{E}\left[X_t Y_t\right] \mathrm{d}t.$$

Proof. Left as an exercise. □

The last property asserts that the map

$$\begin{aligned} L^2([0,T]\times\Omega) \supset \mathcal{M}^2 &\longrightarrow L^2(\Omega) \\ (X_t)_{0\leq t\leq T} &\longmapsto \int_0^T X_t \mathrm{d}B_t \end{aligned}$$

is an isometry. At this point in the construction, this map is only defined on the subspace of $L^2([0,T]\times\Omega)$ generated by the adapted elementary processes. Similarly as for Lebesgue's integral, the idea is to extend its definition by continuity onto the larger subspace $\mathcal{M}^2 \subset L^2([0,T]\times\Omega)$ of adapted processes approximable by elementary processes [1].

Definition 1.7 (Itô integral against the Brownian motion)**.** *For all stochastic processes in $\mathcal{M}^2$, define*

$$\int_a^b X_t \mathrm{d}B_t := \lim_{n\to\infty} \sum_{j=1}^{p_n} X_{t_{j-1}^n}(B_{t_j^n} - B_{t_{j-1}^n}),$$

where the limit is in $L^2(\Omega)$, and (t_i^n) is any sequence of subdivisions of $[a,b]$ with $\max_j |t_j^n - t_{j-1}^n| \xrightarrow[n\to\infty]{} 0$.

Proposition 1.8. *All the properties of Proposition 1.6 are still valid when $(X_t)_t$ is a "nice enough" stochastic process.*

[1] Lots of measurability issues purposely left under the carpet here. See [LG16, Chapter 4].

Example 1.9 ($\int_0^T B_t \mathrm{d}B_t$)**.** *Let us consider the stochastic integral* $\int_0^T B_t \mathrm{d}B_t$*. This quantity makes sense, because the stochastic process* $X_t = B_t$ *is in* $\mathcal{M}^2$*: it is adapted with continuous trajectories and finite integrated second moment*

$$\int_0^T \mathbb{E}[B_t^2]\mathrm{d}t = \int_0^T t\mathrm{d}t = T^2/2.$$

Let (t_i^n) *be a sequence of subdivisions of* $[0,T]$ *with* $\max_i |t_{i+1}^n - t_j^n| \xrightarrow[n\to\infty]{} 0$*. Write*

$$B_t^{(n)} := \sum_{j=1}^{p_n} B_{t_{j-1}^n} \mathbb{1}_{[t_{j-1}, t_j)}(t)$$

for the associated elementary process approximating $(B_t)_t$*. By definition, we have*

$$\begin{aligned}
\int_0^T B_t \mathrm{d}B_t &= \lim_{n\to\infty} \int_0^T B_t^{(n)} \mathrm{d}B_t \\
&= \lim_{n\to\infty} \sum_{j=1}^{p_n} B_{t_{j-1}^n}(B_{t_j^n} - B_{t_{j-1}^n}) \\
&= \lim_{n\to\infty} \left\{ \frac{1}{2}\sum_{j=1}^{p_n}(B_{t_j^n}^2 - B_{t_{j-1}^n}^2) - \frac{1}{2}\sum_{j=1}^{p_n}(B_{t_j^n} - B_{t_{j-1}^n})^2 \right\} \\
&= \frac{1}{2}(B_T^2 - B_0^2) - \lim_{n\to\infty} \frac{1}{2}\sum_{j=1}^{p_n}(B_{t_j^n} - B_{t_{j-1}^n})^2 \\
&= \frac{1}{2}(B_T^2 - T),
\end{aligned}$$

where the last line uses the formula for the quadratic variation of the Brownian motion. At the end of the day, we recognize a similar structure as for $\int_0^T F_t \mathrm{d}F_t = \frac{1}{2}F_T^2$ *when* F *is* $\mathcal{C}^1$ *and* $F_0 = 0$*, but with an extra additive compensator to center the process.*

Example 1.10 (Law of $\int_0^T f_t \mathrm{d}B_t$)**.** *Let* $f : [0,T] \to \mathbb{R}$ *be a continuous deterministic function, and consider* $\int_0^T f_t \mathrm{d}B_t$*. By Definition 1.7, it is the limit in* L^2 *of Gaussian, so it Gaussian. Furthermore, from Proposition 1.8, it has mean zero and variance*

$$\mathrm{Var}\left(\int_0^T f_t \mathrm{d}B_t\right) = \int_0^T f_t^2 \mathrm{d}t.$$

Hence, $\int_0^T f_t \mathrm{d}B_t \sim \mathcal{N}\big(0, \int_0^T f_t^2 \mathrm{d}t\big)$*.*

1.3 A notion of stochastic differential: Itô stochastic calculus

The above construction of Itô integral extends to more general process than the Brownian motion. We will limit ourselves to the following class of processes.

Definition 1.11 (Itô process, stochastic differential)**.** *An* Itô process *(or* stochastic integral *is a stochastic process* $(X_t)_t$ *adapted to* $(\mathcal{F}_t)_t$ *which can be written as*

$$X_t = X_0 + \int_0^t a_t \mathrm{d}t + \int_0^t b_t \mathrm{d}B_t,$$

where a_t, b_t are continuous stochastic processes in L^1 and L^2 respectively. If so, the stochastic differential *of $(X_t)_t$ is defined as*

$$\mathrm{d}X_t := a_t \mathrm{d}t + b_t \mathrm{d}B_t.$$

If so, a_t is called the drift *and b_t the* diffusion term *(or* volatility*) of $(X_t)_t$.*

Here, a_t and b_t may depend (implicitly or explicitly) of the process $(X_s)_{s\leq t}$ itself. Let us emphasize that the stochastic differential is only a shorthand notation for the equality between stochastic integrals above. However, as we shall expect, one easily checks that if F_t is a $\mathcal{C}^1$ process, we recover the classical notion of differential through $\mathrm{d}F_t = F_t' \mathrm{d}t$. This case corresponds to a zero diffusion term $b_t = 0$.

Example 1.12. *From Example 1.9, $B_t^2 = t + \int_0^t 2B_s \mathrm{d}B_s$. Therefore, we have $\mathrm{d}B_t^2 = \mathrm{d}t + 2B_t \mathrm{d}B_t$.*

In the above example, notice the fundamental difference with the regular differential of a $\mathcal{C}^1$ function which yields $\mathrm{d}(F_t)^2 = 2F_t \mathrm{d}F_t$. The extra term comes from the fact that the Brownian motion has finite quadratic variation. This property naturally transfers to Itô processes

Proposition 1.13 (Quadratic variation of an Itô process)**.** *If $(X_t)_t$ is an Itô process as in Definition 1.11, then it has finite quadratic variation*

$$\langle X \rangle_t := \lim_{n\to\infty} \sum_{j=1}^{p_n} (X_{t_j^n} - X_{t_{j-1}^n})^2.$$

Because it is a continuous non-decreasing process, $(\langle X\rangle_t)_t$ has finite first variation, and $\mathrm{d}\langle X\rangle_t = b_t^2 \mathrm{d}t$.

The quadratic variation of an Itô process appears explicitly in the aforementioned tweaked chain rule called *Itô formula*.

Theorem 1.14 (Itô formula)**.** *Let $(X_t)_{0\leq t\leq T}$ be a Itô process and $\Phi \in \mathcal{C}^{2,1}(\mathbb{R}\times\mathbb{R}_+, \mathbb{R})$ be a function of space-time variable (x,t). Then $\big(\Phi(X_t,t)\big)_{0\leq t\leq T}$ is a Itô process with stochastic differential*

$$\mathrm{d}\Phi(X_t,t) = \partial_t\Phi(X_t,t)\mathrm{d}t + \partial_x\Phi(X_t,t)\mathrm{d}X_t + \frac{1}{2}\partial^2_{x,x}\Phi(X_t,t)\mathrm{d}\langle X\rangle_t.$$

Note that if $\mathrm{d}X_t = a_t\mathrm{d}t + b_t\mathrm{d}B_t$, Itô formula rewrites as

$$\mathrm{d}\Phi(X_t,t) = \big(\partial_t\Phi(X_t,t) + a_t\partial_x\Phi(X_t,t)\big)\mathrm{d}t + \big(b_t\partial_x\Phi(X_t,t) + \frac{b_t^2}{2}\partial^2_{x,x}\Phi(X_t,t)\big)\mathrm{d}B_t.$$

Sketch of proof. Let us consider the simpler case where $\Phi(x,t) = \Phi(x)$ is homogeneous in time. In this case, the integral form of Itô formula to be shown is

$$\Phi(X_t) = \Phi(X_0) + \int_0^t \Phi'(X_s)\mathrm{d}X_s + \frac{1}{2}\int_0^t \Phi''(X_s)\mathrm{d}\langle X\rangle_s.$$

To prove it, come back to Definition 1.7 of the Itô integral. Given an arbitrarily fine partition of $[0,t]$, consider the telescopic sum

$$\begin{aligned}\Phi(X_t) &= \Phi(X_0) + \sum_{j=1}^{p} \Phi(X_{t_j}) - \Phi(X_{t_{j-1}}) \\ &= \Phi(X_0) + \sum_{j=1}^{p} \Phi'(X_{t_{j-1}})(X_{t_j} - X_{t_{j-1}}) + \frac{1}{2}\sum_{j=1}^{p} \Phi''(X_{t^*_{j-1}})(X_{t_j} - X_{t_{j-1}})^2,\end{aligned}$$

where the second equality comes from Taylor-Lagrange formula and $t^*_{j-1} \in [t_{j-1}, t_j]$. Dealing with each sum separately, we get that

$$\sum_{j=1}^{p} \Phi'(X_{t_{j-1}})(X_{t_j} - X_{t_{j-1}}) \xrightarrow[p\to\infty]{} \int_0^t \Phi'(X_s)\mathrm{d}X_s$$

by the definition of the stochastic integral, and by uniform continuity of $(X_t)_t$,

$$\begin{aligned}\sum_{j=1}^{p} \Phi''(X_{t^*_{j-1}})(X_{t_j} - X_{t_{j-1}})^2 &\simeq \sum_{j=1}^{p} \Phi''(X_{t_{j-1}})(X_{t_j} - X_{t_{j-1}})^2 \\ &\xrightarrow[p\to\infty]{} \int_0^t \Phi''(X_s)\mathrm{d}\langle X\rangle_s,\end{aligned}$$

which concludes the proof. □

Remark 1.15 (Sanity check for $\mathcal{C}^1$ processes)**.** *The Itô formula does not contradict the classical fundamental theorem of calculus*[2]*. Indeed, replacing X_t by a $\mathcal{C}^1$ process F_t, the second term is zero because in this case, F_t has finite first variation $V(F)_t$, and hence quadratic variation equal to zero. Indeed, from Hölder inequality,*

$$\begin{aligned}\langle F\rangle_t &= \lim_{p\to\infty} \sum_{j=1}^{p} (F_{t_j} - F_{t_{j-1}})^2 \\ &\leq \lim_{n\to\infty} \max_{1\leq j\leq p} |F_{t_j} - F_{t_{j-1}})| \underbrace{\sum_{j=1}^{p_n} |F_{t_j} - F_{t_{j-1}})|}_{\to V(F)_t} \\ &\leq \lim_{n\to\infty} \max_{1\leq j\leq p} \|F'\|_\infty |t_j - t_{j-1}| V(F)_t \\ &= 0.\end{aligned}$$

Exercise 1.16. *Revisit the proof of Example 1.9 using Itô formula.*

1.4 Multidimensional stochastic calculus

All the above can be generalized to random processes with values in $\mathbb{R}^d$. Everything is then defined component-wise. That is, the Brownian motion $(B_t)_{t\geq 0}$ is a Gaussian process with independent coordinates being real-valued Brownian motions. The integral and stochastic differential are defined accordingly. Finally, Itô's formula writes as follows.

Theorem 1.17 (Multidimensional Itô formula)**.** *Let $(X_t)_{0\leq t\leq T}$ be a Itô process in $\mathbb{R}^d$ and $\Phi \in \mathcal{C}^{2,1}(\mathbb{R}^d \times \mathbb{R}_+, \mathbb{R}^k)$ be a function of space-time variable (x,t). Then $\big(\Phi(X_t,t)\big)_{0\leq t\leq T}$ is a Itô process in $\mathbb{R}$ with stochastic differential*

$$\mathrm{d}\Phi(X_t,t) = \partial_t\Phi(X_t,t)\mathrm{d}t + \sum_{k=1}^{d} \partial_{x_k}\Phi(X_t,t)\mathrm{d}X_t^{(k)} + \frac{1}{2}\sum_{k,\ell=1}^{d} \partial^2_{x_k,x_\ell}\Phi(X_t,t)\mathrm{d}\langle X^{(k)}, X^{(\ell)}\rangle_t,$$

where $X_t = (X_t^{(1)}, \dots, X_t^{(d)})$, and $\langle U, V\rangle_t := \lim_{n\to\infty} \sum_{j=1}^{p_n} (U_{t_j^n} - U_{t_{j-1}^n})(V_{t_j^n} - V_{t_{j-1}^n})$.

[2] Fortunately, these lecture notes are not completely nonsense.

Exercise 1.18 (Product rule and value of $\int_0^T f_t \mathrm{d}B_t$)**.** *Use the Theorem 1.17 to prove that if $(X_t)_t$ and $(Y_t)_t$ are independent centered Itô processes, then*

$$X_t Y_t = X_0 Y_0 + \int_0^t X_s \mathrm{d}Y_s + \int_0^t Y_s \mathrm{d}X_s.$$

If $f : [0,T] \to \mathbb{R}$ is a $\mathcal{C}^1$ deterministic function, show that $\int_0^T f_t \mathrm{d}B_t = f_T B_T - \int_0^T B_t \mathrm{d}f_t$. Compare with Example 1.10.

2 Diffusion from a distribution and back

2.1 Ornstein–Uhlenbeck process

Now equipped with a notion of stochastic differential, one may wonder how to solve stochastic differential *equations*. Historically, one of the most central one in diffusion-based generative models is the following.

Definition 2.1 (Ornstein-Uhlenbeck process)**.** *An Ornstein-Uhlenbeck process with parameters $\lambda, \sigma > 0$ starting at $x \in \mathbb{R}^d$ driven by a d-dimensional Brownian motion is a stochastic process on $T = \mathbb{R}_+$ satisfying*

$$\begin{cases} \mathrm{d}X_t = -\lambda X_t \mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t, \\ X_0 = x. \end{cases}$$

To try and solve such a stochastic differential equation (SDE), note that its integral form

$$X_t = x - \int_0^t \lambda X_t \mathrm{d}t + \sqrt{2}\sigma B_t,$$

yields that the mean $m(t) := \mathbb{E}[X_t]$ of X_t satisfies $m'(t) = -\lambda m(t)$ with $m(0) = x$, so that $m(t) = e^{-\lambda t}x$. Hence, let us introduce the renormalized process $Y_t := e^{\lambda t} X_t$. By applying Itô formula (Theorem 1.14) to $\Phi(x,t) := e^{\lambda t}x$, we get

$$\begin{aligned} \mathrm{d}Y_t &= \underbrace{\partial_t \Phi(X_t,t)}_{=\lambda Y_t} \mathrm{d}t + \underbrace{\partial_x \Phi(X_t,t)}_{=e^{\lambda t}} \mathrm{d}X_t + \frac{1}{2} \underbrace{\partial^2_{x,x} \Phi(X_t,t)}_{=0} \mathrm{d}(\sqrt{2}\sigma)^2 \mathrm{d}t \\ &= (\lambda Y_t - \lambda e^{\lambda t} X_t)\mathrm{d}t + e^{\lambda t}\sqrt{2}\sigma \mathrm{d}B_t \\ &= \sqrt{2}\sigma e^{\lambda t} \mathrm{d}B_t. \end{aligned}$$

This means that $Y_t = Y_0 + \int_0^t \sqrt{2}\sigma e^{\lambda s} \mathrm{d}B_s$, or equivalently,

$$X_t = x e^{-\lambda t} + \int_0^t \sqrt{2}\sigma e^{\lambda(s-t)} \mathrm{d}B_s.$$

If x is deterministic, we obtain that (see Example 1.10)

$$X_t \sim \mathcal{N}\big(x e^{-\lambda t}, \frac{\sigma^2}{\lambda}(1 - e^{-2\lambda t})\big).$$

As a result, $X_t \xrightarrow{t\to\infty} \mathcal{N}(0, \sigma^2/\lambda)$ in distribution. See Figure 3 for an illustration.

All this derivation easily generalizes to parameters $\lambda = \lambda_t$ and $\sigma = \sigma_t$ depending on time.

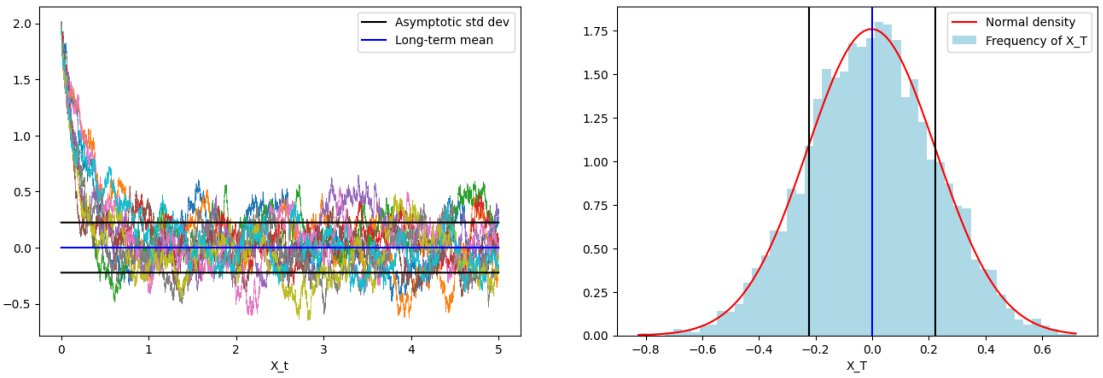

Figure 3: Ten trajectories of an homogeneous Ornstein-Uhlenbeck (Definition 2.1) starting from $X_0 = 2$ with $\lambda = 5$ and $\sigma = 1/2$, all stopped at time $T = 5$ (left). Histogram of X_T on $N = 5000$ draws compared to the limiting normal (right).

Proposition 2.2 (Time-inhomogenous Ornstein-Uhlenbeck process)**.** *The generalized Ornstein-Uhlenbeck equation*

$$\begin{cases} \mathrm{d}X_t = -\lambda_t X_t \mathrm{d}t + \sqrt{2}\sigma_t \mathrm{d}B_t, \\ X_0 = x. \end{cases}$$

admits for unique solution

$$X_t = xe^{-\mu_t} + \int_0^t \sqrt{2}\sigma_s e^{\mu_s - \mu_t} \mathrm{d}B_s,$$

where $\mu_t := \int_0^t \lambda_s \mathrm{d}s$.

Proof. Left as an exercise. □

If now X_0 has a non-deterministic distribution, we obtain the distribution of X_t straightforwardly.

Proposition 2.3. *If* $X_0 \sim p_0(x)\mathrm{d}x$ *and* $(X_t)_{t\geq 0}$ *is given by the generalized Ornstein-Ulhenbeck process of Proposition 2.2, then* $X_t \sim p_t(x)\mathrm{d}x$ *has the distribution of*

$$X_0 e^{-\mu_t} + \sqrt{\left(\int_0^t 2\sigma_s^2 e^{2(\mu_t - \mu_s)} \mathrm{d}s\right)} Z,$$

where $Z \sim \mathcal{N}(0,1)$ *is independent from* X_0*, and* $\mu_t := \int_0^t \lambda_s \mathrm{d}s$.

See Figure 4 for an illustration of Proposition 2.3. From there, the core idea of diffusion generative models can be summarized as follows. Starting from an unknown sample distribution $X_0 \sim p_{\text{data}}$ and gradually adding noise to X_0 (i.e. letting an Ornstein-Uhlenbeck process run from starting point X_0), we converge towards a known Gaussian distribution $\mathcal{N}(0, \sigma^2/\lambda)$ at $t = \infty$. If we know how to reverse this dynamics, then starting from a (easy to generate) fresh random variable with distribution $\mathcal{N}(0, \sigma^2/\lambda)$, we will obtain a fresh sample with distribution (close to) p_{data}.

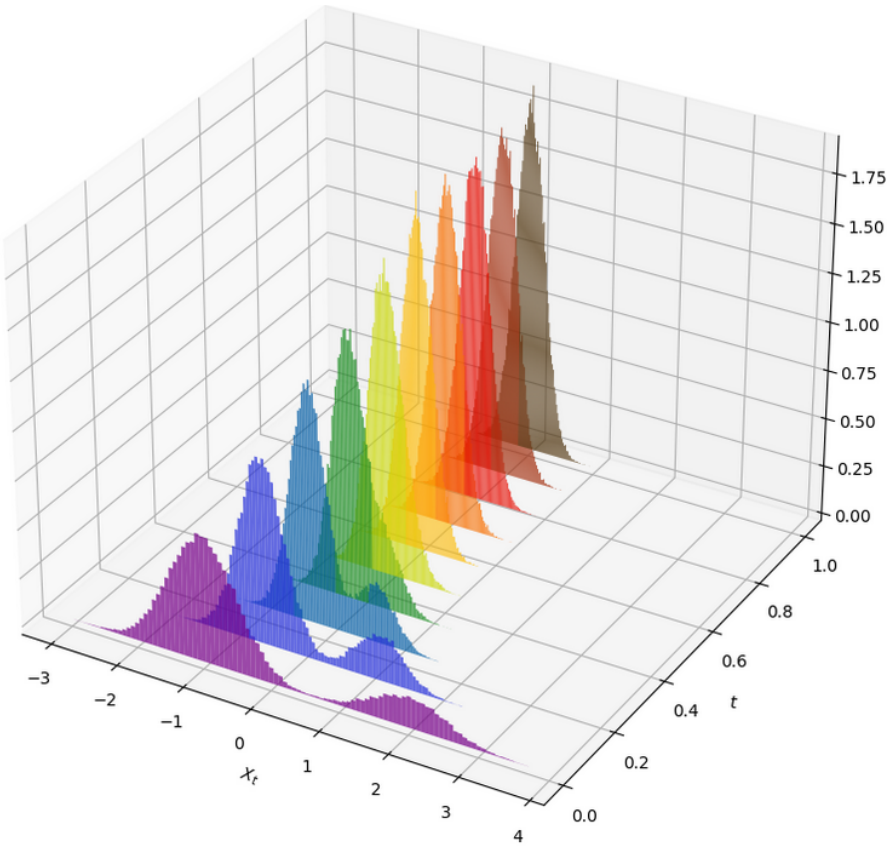

Figure 4: Exemplifying Proposition 2.3 with histograms of Ornstein-Uhlenbeck processes stopped at $T = 1$ starting from X_0 with mixture distribution $p_{\text{data}} = 0.8\mathcal{N}(-1, 1/2) + 0.2\mathcal{N}(-2, 1/2)$. Diffusion parameters are as in Figure 3. Histograms are computed over $N = 50000$ trajectories.

Remark 2.4 (Applicability of the above theory)**.** • *All the above generalizes to higher dimensions $d > 1$ (see Section 1.4), making this idea actually applicable for high-dimensional data*

- *In practice, simulating an Itô process with* known and computable *drift a_t and diffusion term b_t can be done approximately by time discretization. The simplest algorithm for this is called the* Euler scheme*, used to generate the figures of these notes. It uses the very Definition 1.7 of an Itô integral.*

2.2 Fokker-Planck equation

2.2.1 Diffusion processes and PDEs

To formalize how to *reverse time* in stochastic differential equations properly, one has to turn towards the theory of Partial Differential Equations (PDEs) [And82]. Given a smooth enough function $f : \mathbb{R}^d \to \mathbb{R}$ and vector field $V : \mathbb{R}^d \to \mathbb{R}^d$, we denote by

- $\nabla f := (\partial_{x_1} f, \dots, \partial_{x_d} f)$ the *gradient* of f,
- $\nabla \cdot V := \sum_{k=1}^d \partial_{x_k} V_k$ the *divergence* of V,
- $\Delta f := \nabla \cdot \nabla f = \sum_{k=1}^d \partial^2_{x_k, x_k} f$ the *Laplacian* of f.

With these operators, integrations by parts write as

$$\int_{\mathbb{R}^d} f(x) \nabla \cdot V(x) \mathrm{d}x = - \int_{\mathbb{R}^d} \langle \nabla f(x), V(x) \rangle \mathrm{d}x,$$

so that

$$\int_{\mathbb{R}^d} f(x) \Delta g(x) \mathrm{d}x = - \int_{\mathbb{R}^d} \langle \nabla f(x), \nabla g(x) \rangle \mathrm{d}x = \int_{\mathbb{R}^d} \Delta f(x) g(x) \mathrm{d}x,$$

Proposition 2.5 (Fokker-Planck characterization of the dynamic)**.** *Let $(X_t)_t$ be the solution of the SDE*

$$\mathrm{d}X_t = a_t(X_t)\mathrm{d}t + \sqrt{2}\sigma_t(X_t)\mathrm{d}B_t,$$

with initial condition $X_0 \sim p_0(x)\mathrm{d}x$ having a smooth density with respect to the Lebesgue measure in $\mathbb{R}^d$. Then for all $t \geq 0$, X_t has a density p_t with respect to the Lebesgue measure, and this density satisfies the Fokker-Planck *equation*

$$\partial_t p_t = -\nabla \cdot \big(a_t p_t\big) + \Delta\big(\sigma_t^2 p_t\big).$$

Proof. Write $\Phi(x, t) = \Phi_t(x)$ for an arbitrary test function in $\mathcal{C}^{2,1}(\mathbb{R}^d \times [0, T], \mathbb{R})$. Then from Theorem 1.17,

$$\begin{aligned}
\mathrm{d}\Phi_t(X_t) &= \partial_t \Phi_t(X_t)\mathrm{d}t + \sum_{k=1}^d \partial_{x_k} \Phi_t(X_t)\mathrm{d}X_t^{(k)} + \frac{1}{2}\sum_{k,\ell=1}^d \partial^2_{x_k, x_\ell} \Phi_t(X_t)\mathrm{d}\langle X^{(k)}, X^{(\ell)}\rangle_t \\
&= \partial_t \Phi_t(X_t)\mathrm{d}t + \langle \nabla \Phi_t(X_t), \mathrm{d}X_t\rangle + \sigma_t^2 \Delta \Phi_t(X_t)\mathrm{d}t,
\end{aligned}$$

where we used that $\mathrm{d}\langle B^{(k)}, B^{(\ell)}\rangle_t = \delta_{k,\ell}\mathrm{d}t$ by independence of the components of the Brownian motion. This expression simplifies to

$$\mathrm{d}\Phi_t(X_t) = \big(\partial_t \Phi_t(X_t) + \langle \nabla \Phi_t(X_t), a_t\rangle + \sigma_t^2 \Delta \Phi_t(X_t)\big)\mathrm{d}t + \sqrt{2}\sigma_t \langle \nabla \Phi_t(X_t), \mathrm{d}B_t\rangle.$$

From the centering property of Proposition 1.8, we get that $\mathbb{E}\big[\sqrt{2}\sigma_t\langle\nabla\Phi_t(X_t), \mathrm{d}B_t\rangle\big] = 0$. Now writing the above expression in integral form and taking its expectation with respect to $X_t \sim p_t(x)\mathrm{d}x$, we get

$$\begin{aligned}
\mathbb{E}\big[\Phi_t(X_t) - \Phi_0(X_0)\big] &= \int_0^t \mathbb{E}\left[\big(\partial_t\Phi_s(X_s) + \langle\nabla\Phi_s(X_s), a_s\rangle + \sigma_s^2\Delta\Phi_s(X_s)\big)\right]\mathrm{d}s \\
&= \int_0^t \int_{\mathbb{R}^d} \big(\partial_t\Phi_s(x) + \langle\nabla\Phi_s(x), a_s(x)\rangle + \sigma_s^2(x)\Delta\Phi_s(x)\big)p_s(x)\mathrm{d}x\mathrm{d}s \\
&= \int_0^t \int_{\mathbb{R}^d} \big(-\partial_t p_s(x) - \nabla\cdot(p_s(x)a_s(x)) + \Delta(p_s(x)\sigma_s^2(x))\big)\Phi_s(x)\mathrm{d}x\mathrm{d}s,
\end{aligned}$$

where we used an integration by parts for each term. Since this integral equation is true for all smooth enough Φ, we obtain the result. □

2.2.2 Diffusion processes and ODEs

The Fokker–Planck equation can be seen as describing the evolution of the probability density $p_t(x)$ of the position of the particle X_t under the influence of a drift force $a_t(X_t)\mathrm{d}t$ and random forces $\sqrt{2}\sigma_t(X_t)\mathrm{d}B_t$. As such, it is linked with transport of the mass p_0 through time.

Proposition 2.6. *The Fokker-Planck equation for* $\mathrm{d}X_t = a_t(X_t)\mathrm{d}t + \sqrt{2}\sigma_t(X_t)\mathrm{d}B_t$ *can be recast as the non-linear* transport equation

$$\begin{aligned}
\partial_t p_t(x) &= -\nabla\cdot\big(v_t(x)p_t(x)\big) \\
\textit{with velocity field} \quad v_t &:= a_t - \sigma_t^2\nabla\log p_t - \nabla\sigma_t^2.
\end{aligned}$$

Proof. Starting from the Fokker-Planck equation

$$\begin{aligned}
\partial_t p_t &= -\nabla\cdot\big(a_t(x)p_t\big) + \Delta\big(\sigma_t^2 p_t\big) \\
&= -\nabla\cdot\big(a_t(x)p_t - \nabla\big(\sigma_t^2 p_t\big)\big) \\
&= -\nabla\cdot\big(\{a_t(x) - \nabla\big(\sigma_t^2 p_t\big)/p_t\}p_t\big),
\end{aligned}$$

the proof follows by noticing that

$$\begin{aligned}
\nabla\big(\sigma_t^2 p_t\big)/p_t &= \sigma_t^2\nabla p_t/p_t + \nabla\sigma_t^2 \\
&= \sigma_t^2\nabla\log p_t + \nabla\sigma_t^2.
\end{aligned}$$

□

The above transport equation can be seen as the evolution of marginals of a deterministic ODE with a random initialization, as the following result shows.

Proposition 2.7. *If we consider the solution trajectories of the ordinary differential equation*

$$\begin{cases}
\mathrm{d}x_t = v_t(x_t)\mathrm{d}t, \\
x_0 \sim p_0(x)\mathrm{d}x,
\end{cases}$$

then for all $t \geq 0$, $x_t \sim p_t(x)\mathrm{d}x$ *where* p_t *is given by the Fokker-Planck equation of Proposition 2.6.*

Proof. Writing $x_t \sim q_t(x)\mathrm{d}x$, then for all test function Φ,

$$\begin{aligned}\int_{\mathbb{R}^d} \Phi(x)\partial_t q_t(x)\mathrm{d}x &= \partial_t \,\mathbb{E}\left[\Phi(x_t)\right] \\ &= \mathbb{E}\left[\partial_t \Phi(x_t)\right] \\ &= \mathbb{E}\left[\langle \nabla\Phi(x_t), x_t'\rangle\right] \\ &= \int_{\mathbb{R}^d} \langle \nabla\Phi(x), v_t(x)\rangle q_t(x)\mathrm{d}x \\ &= -\int_{\mathbb{R}^d} \Phi(x)\nabla\cdot\left(v_t(x)q_t(x)\right)\mathrm{d}x,\end{aligned}$$

and hence q_t satisfies Fokker-Planck. Since $q_0 = p_0$, we get the result provided that Fokker-Planck has a unique solution. □

At this point, we have constructed two very different continuous random processes, but with identical marginal probability densities p_t:

- $(X_t)_t$ is nowhere differentiable. It satisfies a stochastic differential equation (Proposition 2.5).
- $(x_t)_t$ is smooth. It satisfies an ordinary differential equation (Proposition 2.7).

In fact, both point of view shall provide generative strategies. Overall, the key ingredients for a diffusion-like generative model to be operable are

- *(Interpolation)* The family of distributions $(p_t)_t$ connects $p_0 = p_{\text{data}}$ and $p_T \simeq \mathcal{N}(0,1)$;
- *(Samplability)* The marginals p_t are easy to sample starting from $X_0 \sim p_{\text{data}}$;
- *(Reversibility)* One can learn a way to reverse the time dynamic of $(p_t)_t$.

2.3 Backward process

As above, let us consider the Itô process $\mathrm{d}X_t = a_t(X_t)\mathrm{d}t + \sqrt{2}\sigma_t(X_t)\mathrm{d}B_t$. For instance, we have seen that the Ornstein-Ulhenbeck process provides an easy way to generate random variables $X_T \sim x_T \sim p_T(x)\mathrm{d}x \simeq \mathcal{N}(0, \sigma^2/\lambda)$ from a seed random variable $X_0 \sim p_{\text{data}}$ and the resolution of an ODE (for x_T) or a SDE (for X_T), usually done numerically with a Euler scheme. We now want to reverse time, and try to build a backward process, meaning that for all $t \in [0,T]$,

$$\overleftarrow{\mathbf{x}}_t \sim X_{T-t}.$$

2.3.1 Ordinary time-reversal

A first idea to reverse the dynamic is to use the ODE formulation of Proposition 2.7.

Theorem 2.8 (Backward deterministic dynamic)**.** *If the solution to* $\mathrm{d}x_t = v_t(x_t)\mathrm{d}t$ *has density* $x_t \sim p_t(x)\mathrm{d}x$*, then the solution to*

$$\begin{cases}\mathrm{d}\overleftarrow{x}_t = -v_t(\overleftarrow{x}_t)\mathrm{d}t \\ \overleftarrow{x}_0 \sim p_T(x)\mathrm{d}x\end{cases}$$

satisfies $\overleftarrow{x}_t \sim x_{T-t}$ *for all* $t \in [0,T]$.

Proof. Straightforward from Proposition 2.7. □

Remark 2.9 (Time-reversal is an improper term)**.** *Despite the catchy naming, we have not actually reversed $(x_t)_{0\le t\le T}$ as a stochastic process. Indeed, the trajectories of $\overleftarrow{x}_t$ have differentiable trajectories, while X_{T-t} has $\mathcal{C}^{1/2-}$ trajectories for $\sigma_t \neq 0$. In fact, we have only constructed a process $\overleftarrow{x}_t$ that has the same* marginals *as x_{T-t}*

This result explicitly displays the requirements to simulate the backward process:

- Sample $\overleftarrow{X}_0$ from p_T, (supposedly easy for large T if we chose the forward diffusion well)
- Run a ODE solver with velocity field $v_t(x) = -a_t + \sigma_t^2 \nabla \log p_t(x)$.

Note that here, the drift is a priori unknown because it depends on the *score* $\nabla \log p_{T-t}$.

2.3.2 Stochastic time-reversal(s)

Given user-defined *noise schedule* $(\overleftarrow{\sigma_t})_{0\le t\le T}$, one can reinterpret the Fokker-Planck equation driving the dynamic at the level of probability densities.

Theorem 2.10 (Backward stochastic dynamic)**.** *If the solution to $\mathrm{d}X_t = a_t(X_t)\mathrm{d}t + \sqrt{2}\sigma_t(X_t)\mathrm{d}B_t$ has density $X_t \sim p_t(x)\mathrm{d}x$, then the solution to*

$$\begin{cases} \mathrm{d}\overleftarrow{X}_t = \overleftarrow{a}_t(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{2}\overleftarrow{\sigma_t}\mathrm{d}B_t \\ \overleftarrow{X}_0 \sim p_T(x)\mathrm{d}x \end{cases}$$

with $\overleftarrow{a}_t := -a_{T-t} + \nabla(\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2) + (\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2)\nabla \log p_{T-t}$ satisfies

$$\overleftarrow{X}_t \sim X_{T-t}.$$

As pointed out in Remark 2.9, let us insist on the fact that this result only states that the *marginals* distributions X_{T-t} and $\overleftarrow{X}_t$ are the same. It does not conclude anything about the full stochastic processes $(X_{T-t})_t$ and $(\overleftarrow{X}_t)_t$.

For the choice $\overleftarrow{\sigma_t} := 0$, we recover exactly the result of Theorem 2.8. Another very common choice is $\overleftarrow{\sigma_t} := \sigma_{T-t}$, yielding a dynamic of exact same diffusive type as the forward one.

Proof. From Proposition 2.5, the Fokker-Planck equation associated to the forward process is

$$0 = -\partial_t p_t - \nabla \cdot \big(a_t p_t\big) + \Delta(\sigma_t^2 p_t).$$

In distribution, reversing the dynamic amounts to consider $t \mapsto p_{T-t}$ instead of $t \mapsto p_t$, which reverses the sign of the time derivative and leaves the spatial ones unchanged. Therefore,

$$0 = +\partial_t p_{T-t} - \nabla \cdot \big(a_{T-t} p_{T-t}\big) + \Delta(\sigma_{T-t}^2 p_{T-t}).$$

To recognize an instance of the Fokker-Planck equation with diffusive term $\overleftarrow{\sigma_t}^2$, we then write

$$\begin{aligned} & 0 = -\partial_t p_{T-t} + \nabla \cdot \big(a_{T-t} p_{T-t}\big) - \Delta(\sigma_{T-t}^2 p_{T-t}) \\ \iff & 0 = -\partial_t p_{T-t} + \big(\nabla \cdot \big(a_{T-t} p_{T-t}\big) - \Delta(\{\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2\} p_{T-t})\big) + \Delta(\overleftarrow{\sigma}_t^2 p_{T-t}). \end{aligned}$$

In the middle term, as in the proof of Proposition 2.6, we use the fact that

$$\Delta(\sigma^2 p) = \nabla \cdot \nabla(\sigma^2 p) = \nabla \cdot (\{\sigma^2 \nabla \log p + \nabla \sigma^2\} p)$$

to get the equivalent equation

$$0 = -\partial_t p_{T-t} - \nabla \cdot \left(\overleftarrow{a}_{T-t} p_{T-t} \right) + \overleftarrow{\sigma}_t^2 \Delta p_{T-t},$$

where

$$\overleftarrow{a}_t := -a_{T-t} + \nabla(\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2) + (\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2) \nabla \log p_{T-t}.$$

At the end of the day, we recognize this Fokker-Planck equation as characterizing the announced backward stochastic dynamic.

□

This result explicitly displays the requirements to simulate the backward process:

- Sample $\overleftarrow{X}_0$ from p_T, (supposedly easy for large T if we chose the forward diffusion well)
- Run a SDE solver with
  - diffusion coefficient $\overleftarrow{\sigma}_t$, which we choose;
  - drift $\overleftarrow{a}_t(x) := -a_{T-t}(x) + (\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2) \nabla \log p_{T-t}(x)$, which unfortunately depends on the unknown distribution p_{T-t} again.

3 Score-based generative models

Let us present a couple ways to estimate the *score* function $(x, t) \mapsto \nabla \log p_t(x)$. *Score matching* is the standard terminology to refer to this part. As will become clear in Section 4, the loss we consider is very adapted to generative modeling. It is the so-called *Fisher divergence* is given by

$$\begin{aligned} \text{Fisher}(p \mid \widehat{p}) &:= \int_{\mathbb{R}^d} \|\nabla \log p(x) - \nabla \log \widehat{p}(x)\|^2 p(x) \mathrm{d}x \\ &= \mathbb{E}_{X \sim p}[\|\nabla p(X) - \nabla \widehat{p}(X)\|^2], \end{aligned}$$

and for which the L^2 structure allows for drastic simplifications when optimizing over $s(x) := \nabla \log \widehat{p}(x)$, see below. Indeed, at this point, $\text{Fisher}(p \mid \widehat{p})$ cannot be trivially estimated from sample because of the dependence in $\nabla \log p$ in the expectation.

3.1 Vanilla score matching

The main trick for score matching dates back to [HD05]. It is based on the following simple result.

Proposition 3.1 (Vanilla score trick)**.** *For all smooth density $p : \mathbb{R}^d \to \mathbb{R}_+$, there exists $c_p \geq 0$ such that the following holds. For all smooth $s : \mathbb{R}^d \to \mathbb{R}^d$ decaying sufficiently fast at infinity,*

$$\mathbb{E}_{X \sim p} \left[\|\nabla \log p(X) - s(X)\|^2 \right] = c_p + \mathbb{E}_{X \sim p} \left[2\nabla \cdot s(X) + \|s(X)\|^2 \right].$$

Proof. We simply develop the left-hand side to get

$$\begin{aligned} &\mathbb{E}_{X \sim p} \left[\|\nabla \log p(X) - s(X)\|^2 \right] \\ &\quad = \mathbb{E}_{X \sim p} \left[\|\nabla \log p(X)\|^2 \right] - \mathbb{E}_{X \sim p} \left[2\langle \nabla \log p(X), s(X) \rangle \right] + \mathbb{E}_{X \sim p} \left[\|s(X)\|^2 \right]. \end{aligned}$$

The first term does not depend on s and the last one is just as desired. The middle one can be integrated by parts through

$$\begin{aligned}\int_{\mathbb{R}^d} \langle \nabla \log p(x), s(x) \rangle p(x) \mathrm{d}x &= \int_{\mathbb{R}^d} \langle \nabla p(x), s(x) \rangle \mathrm{d}x \\ &= -\int_{\mathbb{R}^d} p(x) \nabla \cdot s(x) \mathrm{d}x,\end{aligned}$$

which yields the result. □

From there, one can fit a parametric set of functions $(s_\theta)_{\theta \in \Theta}$ (typically neural networks) to learn the score $\nabla \log p_t(x)$ via the empirical risk minimization

$$\theta_t \in \operatorname*{argmin}_\theta \mathbb{E}_{X_t \sim p_t} \left[2\nabla \cdot s_\theta(X_t) + \|s_\theta(X_t)\|^2 \right]. \tag{1}$$

Note that an empirical version of the above expectation is indeed available to us, from simulations of the *forward* process.

Remark 3.2 (But... In practice?)**.** • *Equation* (1) *needs to be solved globally for* $t \in [0, T]$*. We could discretize* $0 = t_0 < \ldots < t_p = T$ *and fit* p *scores* $s_{\theta_{t_0}}, \ldots, s_{\theta_{t_p}}$ *in parallel. However, it appears that learning the whole function* $(x, t) \mapsto \nabla \log p_t(x)$ *globally in space* and *time is more efficient. This fact follows the intuition, since closeby* t_j *should result in closeby* $s_{\theta_{t_j}}$*. Therefore, practitioners tend fit one single space-time neural net with the time-integrated loss*

$$\theta \in \operatorname*{argmin}_\theta \int_0^T w(t) \mathbb{E}_{X_t \sim p_t} \left[2\nabla \cdot s_\theta(X_t, t) + \|s_\theta(X_t, t)\|^2 \right] \mathrm{d}t,$$

with w *being a weight function chosen by the user (typically decreasing).*

- *Overall, the loss function to minimize has the form*

$$\ell(\theta) := \sum_{j=0}^{p} w(t_j) \sum_{i=1}^{n} \left(2\nabla \cdot s_\theta(X_{t_j,i}, t_j) + \|s_\theta(X_{t_j,i}, t_j)\|^2 \right), \tag{2}$$

where sample batches $(X_{t_0,i})_{i \leq n}, \ldots, (X_{t_p,i})_{i \leq n}$ *are obtained by SDE simulations starting from data* $X_1, \ldots, X_n \sim p_0$*. Even with these simulated sample taken as granted, note that performing gradient descent on* (2) *requires to evaluate second order gradients* $\nabla_\theta \nabla_x s_\theta(x)$*, which is very costly.*

3.2 Denoising score matching

3.2.1 General principle

One way to avoid the general numerical limitations described in Remark 3.2 is to take advantage of the *convolutional* structure of the noising process [Vin11]. Writing $p * g(x) := \int_{\mathbb{R}^d} p(y) g(x-y) \mathrm{d}y$ for the convolution of densities $p, g : \mathbb{R}^d \to \mathbb{R}_+$, we can build upon the following result.

Proposition 3.3 (Denoising score trick)**.** *If* $X \sim p(x)\mathrm{d}x$ *and* $\varepsilon \sim g(x)\mathrm{d}x$ *are independent, then* $X_\varepsilon := X + \varepsilon \sim (p * g)(x)\mathrm{d}x$*. Furthermore, there exists* $c'_{p,g}$ *such that for all smooth* $s : \mathbb{R}^d \to \mathbb{R}^d$*,*

$$\mathbb{E}_{X_\varepsilon \sim p*g} \left[\|\nabla \log(p * g)(X_\varepsilon) - s(X_\varepsilon)\|^2 \right] = c'_{p,g} + \mathbb{E}_{(X,\varepsilon) \sim p \otimes g} \left[\|\nabla \log g(\varepsilon) - s(X + \varepsilon)\|^2 \right].$$

Proof. By properties of the convolution, $p_g := p * g$ is smooth as soon as p *or* g is smooth. From Proposition 3.1 applied to $X_\varepsilon \sim p_g$, we have

$$\mathbb{E}_{X_\varepsilon \sim p*g}\left[\|\nabla \log(p*g)(X_\varepsilon) - s(X_\varepsilon)\|^2\right] = c_{p*g} + \mathbb{E}_{(X,\varepsilon)\sim p\otimes g}\left[2\nabla \cdot s(X+\varepsilon) + \|s(X+\varepsilon)\|^2\right].$$

Furthermore, the second term of the last display writes as

$$\begin{aligned} 2\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \nabla \cdot s(x+y)g(y)\mathrm{d}y\right)p(x)\mathrm{d}x &= -2\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \langle \nabla \log g(y), s(x+y)\rangle g(y)\mathrm{d}y\right)p(x)\mathrm{d}x \\ &= -2\mathbb{E}_{(X,\varepsilon)\sim p\otimes g}\left[\langle \nabla \log g(\varepsilon), s(X+\varepsilon)\rangle\right]. \end{aligned}$$

The proof is then complete by recognizing the square

$$\mathbb{E}\left[-2\langle \nabla \log g(\varepsilon), s(X+\varepsilon)\rangle + \|s(X+\varepsilon)\|^2\right] = \mathbb{E}\left[\|\nabla \log g(\varepsilon) - s(X+\varepsilon)\|^2\right] - \mathbb{E}\left[\|\nabla \log g(\varepsilon)\|^2\right],$$

with $\mathbb{E}\left[\|\nabla \log g(\varepsilon)\|^2\right]$ depending only on g. □

As desired, the expression given by Proposition 3.3 does not involve any derivative of the candidate score s. Instead, the derivative is undertaken by the score $\nabla \log g$ of the chosen noise.

3.2.2 Ornstein-Uhlenbeck denoising trick

To see the denoising trick in action, we now apply Proposition 3.3 to the density $p = p_t$ associated to the Ornstein-Ulhenbeck process $X_t \sim e^{-\lambda t}X_0 + \varepsilon_t$ at time t as defined in Section 2.1. Its distribution does write as a convolution $p_t = q_t * g_t$ with:

- the scaled distribution $q_t(x) = e^{\lambda t}p(e^{\lambda t}x)$ of the drifted signal $e^{-\lambda t}X_0$;
- the noise distribution $g_t(x) = (2\pi\Sigma_t^2)^{-d/2}\exp\left(-\|x\|^2/(2\Sigma_t^2)\right)$ of the Gaussian $\varepsilon_t \sim \mathcal{N}(0, \Sigma_t)$ with
$$\Sigma_t := \frac{\sigma^2}{\lambda}(1 - e^{-2\lambda t}).$$
Hence, $\nabla \log g_t(x) = -x/\Sigma_t^2$.

The time-integrated loss minimization becomes equivalent to

$$\begin{aligned} \theta &\in \operatorname*{argmin}_\theta \int_0^T w(t)\mathbb{E}\left[\|\nabla \log g_t(\varepsilon_t) - s_\theta(X_t, t)\|^2\right]\mathrm{d}t \\ &= \operatorname*{argmin}_\theta \int_0^T w(t)\mathbb{E}\left[\|\nabla \log g_t(\varepsilon_t) - s_\theta(e^{-\lambda t}X_0 + \varepsilon_t, t)\|^2\right]\mathrm{d}t \\ &= \operatorname*{argmin}_\theta \int_0^T w(t)\mathbb{E}\left[\left\|-\frac{\varepsilon_t}{\Sigma_t^2} - s_\theta(e^{-\lambda t}X_0 + \varepsilon_t, t)\right\|^2\right]\mathrm{d}t. \end{aligned}$$

If $\xi \sim \mathcal{N}(0, I_{d\times d})$ is independent from X_0, this leads to the concise expresion

$$\theta \in \operatorname*{argmin}_\theta \int_0^T w(t)\mathbb{E}\left[\left\|\left(\frac{-\xi}{\Sigma_t}\right) - s_\theta(e^{-\lambda t}X_0 + \Sigma_t\xi, t)\right\|^2\right]\mathrm{d}t.$$

In practice, the above minimization is discretized over the time interval $[0, T]$. The expected value is approximated via the sample mean for the X_0 part. Integration with respect to ξ can either be approximated by Monte Carlo methods or computed exactly. In the latter case, given

data $\mathbb{X}_n = \{X^{(1)}, \dots, X^{(n)}\}$ and used-defined time steps $0 \leq t_1 \leq \dots \leq t_k \leq T$, we end up with the loss

$$L_{\mathbb{X}_n}(\theta) = \sum_{k=1}^{k} w(t_k) \sum_{i=1}^{n} \mathbb{E}_\xi \left[\left\| \left(\frac{-\xi}{\Sigma_{t_k}} \right) - s_\theta(e^{-\lambda t_k} X^{(i)} + \Sigma_{t_k} \xi, t_k) \right\|^2 \right].$$

Then, the backward process shall be discretized via an Euler-Maruyama method based on the same time steps $(t_k)_k$.

Remark 3.4 (Tweedie's formula: why score matching is about denoising)**.** *The last expression highlights why we sometimes say that s_θ "learns the noise". In the Gaussian case, a fitted score $s_\theta(\cdot, t)$ is actually meant to fit the opposite of rescaled noise $-\xi/\Sigma_t$ from observation X_t.*

In fact, looking again at Proposition 3.3 in full generality, we see that the minimizer $s^* : \mathbb{R}^d \to \mathbb{R}^d$ of

$$\operatorname*{argmin}_{s:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}_{X_\varepsilon \sim p*g} \left[\|\nabla \log(p * g)(X_\varepsilon) - s(X_\varepsilon)\|^2 \right] = \operatorname*{argmin}_{s:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}_{(X,\varepsilon)\sim p\otimes g} \left[\|\nabla \log g(\varepsilon) - s(X+\varepsilon)\|^2 \right]$$

is unique $\mathbb{P}_{X+\varepsilon}$-almost surely, and is characterized by the conditional expectation

$$s^*(X+\varepsilon) := \mathbb{E}\left[\nabla \log(\varepsilon) \mid X+\varepsilon\right].$$

4 Sampling from a learnt score

Given a learnt score function $s : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, meant to approximate $\nabla \log p_{T-t}$, let us now quantify the error of the output learnt probability distribution. Many natural discrepancies between probability measures could be considered. This includes information-theoretic ones such as the total variation distance and the Kullback-Leibler divergence, or transport-based ones such as Wasserstein distances or Integral Probability Metrics, much more relevant for high-dimensional data. For sake of simplicity, we will only consider The *Kullback-Leibler* divergence, defined as

$$\mathrm{KL}(p \mid q) = \int_{\mathbb{R}^d} \log\left(\frac{p(x)}{q(x)}\right) p(x) \mathrm{d}x.$$

4.1 Exact Kullback-Leibler dynamics

In the literature of score matching, the first bound to be obtained on the KL-divergence is attributed to [SSDK+20]. Its proof is based on *Girsanov's theorem*, which relies heavily on advanced stochastic calculus. Let us give an alternative one taken from Simon Coste's website, based on Proposition 2.6 and simple integral calculus.

Proposition 4.1. *Let $(p_t)_{0\leq t\leq T}$ and $(q_t)_{0\leq t\leq T}$ be two families of smooth probability densities on $\mathbb{R}^d$, respectively driven by the transport equations*

$$\partial_t p_t(x) = -\nabla \cdot \big(v_t(x) p_t(x)\big) \qquad \text{and} \qquad \partial_t q_t(x) = -\nabla \cdot \big(u_t(x) q_t(x)\big),$$

with smooth enough velocity fields $v_t, u_t : \mathbb{R}^d \to \mathbb{R}^d$. Then we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t \mid q_t) = \int_{\mathbb{R}^d} \left\langle v_t(x) - u_t(x), \nabla \log\left(\frac{p_t(x)}{q_t(x)}\right) \right\rangle p_t(x) \mathrm{d}x.$$

Proof. Assuming that a time integral derivative inversion is legit, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t \mid q_t) &= \int_{\mathbb{R}^d} \partial_t \left\{p_t \log q_t - p_t \log p_t\right\} \\ &= \int_{\mathbb{R}^d} \partial_t p_t (1 + \log p_t) - \partial_t p_t \log q_t - \frac{p_t}{q_t}\partial q_t \\ &= \int_{\mathbb{R}^d} \partial_t p_t + \int_{\mathbb{R}^d} \partial_t p_t \log(p_t/q_t) - \int_{\mathbb{R}^d} \frac{p_t}{q_t}\partial q_t.\end{aligned}$$

The first term is zero because $\int_{\mathbb{R}^d} p_t = 1$ for all t. Because $(p_t)_t$ follows the transport equation with velocity v_t, the second term rewrites as

$$-\int_{\mathbb{R}^d} \nabla \cdot (v_t(x)p_t(x)) \log(p_t/q_t) = \int_{\mathbb{R}^d} \langle v_t(x)p_t(x), \nabla \log(p_t/q_t)\rangle \,,$$

where we use an integration by parts. Similarly, the third one is equal to

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{p_t}{q_t}\nabla \cdot (u_t q_t) &= -\int_{\mathbb{R}^d} \langle \nabla(p_t/q_t), u_t q_t\rangle \\ &= -\int_{\mathbb{R}^d} \langle \nabla \log(p_t/q_t), u_t p_t\rangle \,,\end{aligned}$$

where the last equality follows from $\nabla(p_t/q_t) = (p_t/q_t)\nabla \log(p_t/q_t)$. Putting everything together, we obtain the result. □

As a direct consequence, we can integrate this bound to get an explicit formula for the Kullback-Leibler divergence along the flow.

Corollary 4.2. *In the context of Proposition 4.1, we have*

$$\mathrm{KL}(p_T \mid q_T) = \mathrm{KL}(p_0 \mid q_0) + \int_0^T \int_{\mathbb{R}^d} \left\langle v_t(x) - u_t(x), \nabla \log\left(\frac{p_t(x)}{q_t(x)}\right)\right\rangle p_t(x)\mathrm{d}x\mathrm{d}t.$$

4.2 Application to flow matching

[Forward SDE] Let $p_0 : \mathbb{R}^d \to \mathbb{R}_+$ be some probability distribution of interest. Starting from $X_0 \sim p_0(x)\mathrm{d}x$, we run the forward SDE $\mathrm{d}X_t = a_t(X_t)\mathrm{d}t + \sqrt{2}\sigma_t(X_t)\mathrm{d}B_t$ over the time interval $[0, T]$. We denote by $(p_t)_{0\le t\le T}$ the associated distribution.
[True backward (S)DE] Given some backwards noise schedule $(\overleftarrow{\sigma_t})_{0\le t\le T}$ (possibly zero), Theorem 2.10 asserts that the solution to

$$\begin{cases}\mathrm{d}\overleftarrow{X}_t = \overleftarrow{a}_t(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{2}\overleftarrow{\sigma_t}(\overleftarrow{X}_t)\mathrm{d}B_t \\ \overleftarrow{X}_0 \sim p_T(x)\mathrm{d}x\end{cases}$$

with $\overleftarrow{a}_t := -a_{T-t} + \nabla(\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2) + (\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2)\nabla \log p_{T-t}$ satisfies $\overleftarrow{X}_t \sim X_{T-t}$, which we denote by $\overleftarrow{p}_t = p_{T-t}$.
[Approximated backward (S)DE] Given some score function $s : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ meant to approximate $\nabla \log p_{T-t}$, and an easy-to-sample user-defined density p_∞ we now run the SDE

$$\begin{cases}\mathrm{d}\hat{X}_t = \hat{a}_t(\hat{X}_t)\mathrm{d}t + \sqrt{2}\overleftarrow{\sigma_t}(\hat{X}_t)\mathrm{d}B_t \\ \hat{X}_0 \sim p_\infty(x)\mathrm{d}x\end{cases} \tag{3}$$

with $\hat{a}_t := -a_{T-t} + \nabla(\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2) + (\sigma_{T-t}^2 + \overleftarrow{\sigma_t}^2)s_t$. We write $\hat{X}_t \sim q_t(x)\mathrm{d}x$. After running the SDE until time T, the new *fake* sample that is output by the method is $\hat{X}_T$.

Proposition 4.3. *The final time of the stochastic process* (3) *satisfies*

$$\mathrm{KL}(p_0 \mid q_T) \leq \mathrm{KL}(p_T \mid p_\infty) + \int_0^T \int_{\mathbb{R}^d} \frac{(\sigma_t^2(x) + \overleftarrow{\sigma}_{T-t}^2(x))^2}{4\overleftarrow{\sigma}_{T-t}^2(x)} \left\|\nabla \log p_t(x) - s_{T-t}(x)\right\|^2 p_t(x)\mathrm{d}x$$

Proof. From Proposition 2.6, its family of densities $\overleftarrow{p}_t = p_{T-t}$ satisfy the transport equation

$$\begin{aligned} \partial_t \overleftarrow{p}_t &= -\nabla \cdot \left(v_t \overleftarrow{p}_t\right) \\ \text{with velocity field} \quad v_t &:= \overleftarrow{a}_t - \overleftarrow{\sigma}_t^2 \nabla \log \overleftarrow{p}_t - \nabla \overleftarrow{\sigma}_t^2 \\ &= -a_{T-t} + \sigma_{T-t}^2 \nabla \log p_{T-t} + \nabla \sigma_{T-t}^2. \end{aligned}$$

Similarly, the family of densities $(q_t)_{0 \leq t \leq T}$ satisfy the transport equation

$$\begin{aligned} \partial_t q_t &= -\nabla \cdot \left(u_t q_t\right) \\ \text{with velocity field} \quad u_t &:= \hat{a}_t - \overleftarrow{\sigma}_t^2 \nabla \log q_t - \nabla \overleftarrow{\sigma}_t^2 \\ &= -a_{T-t} + \sigma_{T-t}^2 s_t - \overleftarrow{\sigma}_t^2 (\nabla \log q_t - s_t) + \nabla \sigma_{T-t}^2. \end{aligned}$$

Hence, applying Corollary 4.2 to $(\overleftarrow{p}_t)_t = (p_{T-t})_t$ and $(q_t)_t$ on the time interval $[0, T]$ yields

$$\begin{aligned} &\mathrm{KL}(p_0 \mid q_T) - \mathrm{KL}(p_T \mid p_\infty) \\ &= \int_0^T \int_{\mathbb{R}^d} \left\langle \sigma_{T-t}^2 (\nabla \log p_{T-t} - s_t) + \overleftarrow{\sigma}_t^2 (\nabla \log q_t - s_t), \nabla \log \left(\frac{p_{T-t}}{q_t}\right) \right\rangle p_{T-t} \\ &= \int_0^T \int_{\mathbb{R}^d} \left\langle (\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2)(\nabla \log p_{T-t} - s_t) - \overleftarrow{\sigma}_t^2 \nabla \log \left(\frac{p_{T-t}}{q_t}\right), \nabla \log \left(\frac{p_{T-t}}{q_t}\right) \right\rangle p_{T-t} \end{aligned}$$

In the integrand, the inner product simplifies to

$$\begin{aligned} &-\overleftarrow{\sigma}_t^2 \left\| \nabla \log \left(\frac{p_{T-t}}{q_t}\right) \right\|^2 + \left\langle (\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2)(\nabla \log p_{T-t} - s_t), \nabla \log \left(\frac{p_{T-t}}{q_t}\right) \right\rangle \\ &\qquad \leq \frac{(\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2)^2}{4\overleftarrow{\sigma}_t^2} \left\|\nabla \log p_{T-t} - s_t\right\|^2, \end{aligned}$$

where the inequality follows from $\langle a, b \rangle \leq \|a\|^2/(4\lambda) + \lambda \|b\|^2$ with $a = (\sigma_{T-t}^2 + \overleftarrow{\sigma}_t^2)(\nabla \log p_{T-t} - s_t)$, $b = \nabla \log (p_{T-t}/q_t)$ and $\lambda = \overleftarrow{\sigma}_t^2$. The final result follows after the change of variable $t' = T - t$. □

References

[And82] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[HD05] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.