A stochastic control view on Score-Based Generative Models

Giovanni Conforti and Giacomo Greco

March 9, 2024

Let $\mu^\star \in \mathcal{P}(\mathbb{R}^d)$ be a probability distribution on $\mathbb{R}^d$ and imagine you would like to sample from this distribution. If $\mu^\star$ is in Gibbs form and the energy functional is known, *i.e.*, if $\mu^\star(\mathrm{d}x) = e^{-U(x)}\mathrm{d}x$, and we have access to U, then a popular choice for sampling from $\mu^\star$ consists in Monte Carlo methods. For example, one could use the overdamped Langevin diffusion

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,,$$

whose equilibrium measures coincides with $\mu^\star \propto \exp(-U)$. Therefore in order to sample from $\mu^\star$ is enough considering the random variable X_T (for $T \gg 1$ large enough).

However, in most applications, an analytical expression for U in not available and the above Monte Carlo approach is not feasible. On the contrary, in most applications samples from $\mu^\star$ can be obtained. These samples might be obtained at a high cost, however this is a cost that we won't pay anymore once our generative model is built.

1 Score-based diffusion models

The models we will analyze in this lecture notes are the Score-Based Diffusion Models (SBDM) which provide with an algorithmic framework for the following principle

"*Creating noise from data is easy; creating data from noise is generative modeling.*"[1]

The general idea is that in order to "create noise from data" one can use an ergodic forward Markov process started from the data distribution, which converges to its invariant distribution (which is typically a Gaussian law with independent components). This invariant distribution plays the role of the "noise" which we can easily created from the data by considering the forward evolution of this ergodic Markov process. Then, creating "data from noise" is done by considering the time-reversal of the forward process, called the backward process. More precisely, one can sample from the noisy invariant distribution (from which sampling can be easily done) and then consider the backward dynamics started in these noisy samples. Since we have considered the time-reversal process, at time $t = 0$ the backward process ideally provides random variables distributed according to our starting $\mu^\star$ data distribution.

However, considering the time-reversal Markov process is a non trivial task since the generator of the time reversal depends on the data distribution. To give an example, when the forward process, sometimes called the noising process, is a diffusion process, then the generator of the backward

[1] This quote is taken from the abstract of [Song et al., 2021].

generator depends on the score function of the forward process, *i.e.*, it depends on the law of the forward process (and hence from $\mu^\star$ itself). At the heart of the success of SBDM is the fact that the score has a numerically tractable expression as a conditional expectation, as we will see later.

1.1 Description of the algorithm

In this section we provide a more precise mathematical description of the generative model. For exposition's clarity, we use upper arrow in order to explicit whether we are considering the forward process $\overrightarrow{X}$ or the backward process $\overleftarrow{X}$ and similarly for the corresponding laws on the path space $\Omega = C([0,T);\mathbb{R}^d)$.

Forward process The forward Markov process we consider is an Ornstein-Uhlenbeck process started in the data distribution, that is

$$\begin{cases} \mathrm{d}\overrightarrow{X}_t = -\overrightarrow{X}_t\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \overrightarrow{X}_0 \sim \mu^\star. \end{cases} \tag{1}$$

Let us denote by $\overrightarrow{\mathrm{P}}$ its law on Ω and by $(\overrightarrow{p}_t)_{t\geq 0}$ its marginal flow, $\overrightarrow{p}_t = \mathcal{L}(\overrightarrow{X}_t)$. It is well known that $\overrightarrow{p}_t$ converges[2] to the standard Gaussian distribution $\gamma(\mathrm{d}x) = (2\pi)^{-d/2}\exp(-|x|^2/2))\mathrm{d}x$ as $t\uparrow\infty$ (which means that $\overrightarrow{p}_T \approx \gamma$ for $T \gg 1$ large enough).

The backward process For a fixed (large) time horizon $T \gg 1$, we may now consider the backward process $\overleftarrow{X}_t$ defined as

$$\overleftarrow{X}_t = \overrightarrow{X}_{T-t}, \quad t\in[0,T]$$

and denote by $\overleftarrow{\mathrm{P}}$ the law of the backward process on $\Omega = C([0,T);\mathbb{R}^d)$ and by $(\overleftarrow{p}_t)_{t\geq 0}$ its marginal flow, *i.e.*, $\overleftarrow{p}_t = \mathcal{L}(\overleftarrow{X}_t)$. Clearly $\overleftarrow{p}_t = \overrightarrow{p}_{T-t}$, so that indeed the backward process at time $t=T$ can be used for sampling from the data distribution, since $\overleftarrow{p}_T = \overrightarrow{p}_0 = \mu^\star$.

Naturally, the definition of time reversal depends on the choice of T, but we omit this in the notation.

A classical result in the literature about time-reversal for diffusion processes (see for instance [Cattiaux et al., 2023]) states that under suitable regularity assumptions on $\mu^\star$, the law of the time reversal is a weak solution to the stochastic differential equation

$$\begin{cases} \mathrm{d}\overleftarrow{X}_t = +\overleftarrow{X}_t\mathrm{d}t + 2\nabla\log\overrightarrow{p}_{T-t}(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \overleftarrow{X}_T \sim \mu^\star. \end{cases}$$

Note that we could have written $\overleftarrow{p}_t$ instead of $\overrightarrow{p}_{T-t}$ in the above, as these two laws coincide. In Machine-Learning and statistics literature, the time-dependent vector field

$$(t,x)\mapsto \nabla\log\overrightarrow{p}_t(x)$$

is often referred to as the *score* function associated to the diffusion process (1), since it corresponds to the gradient of the log-likelihood function. For later convenience, let us also introduce the *relative*

[2]This convergence is exponentially fast in $\mathbf{W}_2$-Wasserstein distance, see [Bakry et al., 2013, Theorem 9.7.2].

score as the function

$$(t, y) \mapsto \nabla \log \frac{d\overrightarrow{p}_t}{d\gamma}(y)\,.$$

Considering the relative score in the above backward dynamics has the advantage of considering the same sign for drift term $-\overleftarrow{X}$ as in the forward dynamics. Indeed a direct computation shows that we can then rewrite the dynamics of the time reversal as

$$\begin{cases} \mathrm{d}\overleftarrow{X}_t = -\overleftarrow{X}_t \mathrm{d}t + 2\nabla \log \frac{d\overrightarrow{p}_{T-t}}{d\gamma}(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \overleftarrow{X}_T \sim \mu^\star. \end{cases} \tag{2}$$

1.2 Score as a conditional expectation

At the core of the success of SBDM stems the observation that the score function can be interpreted as a conditional expectation. In order to prove that, firstly not that the transition probability of the Ornstein-Uhlenbeck process (1) is explicitly known and equals

$$\overrightarrow{p}_t(x, y) = (2\pi\sigma_t^2)^{-d/2} \exp\Big(- \frac{|y - \mu_t x|^2}{2\sigma_t^2}\Big), \quad \text{where } \mu_t = e^{-t}, \text{ and } \sigma_t^2 = (1 - e^{-2t})\,.$$

Therefore, from the identity

$$\overrightarrow{p}_t(y) = \int p_t(x, y)\mu^\star(\mathrm{d}x)$$

we obtain by differentiating under the integral sign that

$$\nabla_y \overrightarrow{p}_t(y) = - \int \sigma_t^{-2}(y - \mu_t x)\overrightarrow{p}_t(x, y)\mu^\star(\mathrm{d}x)\,.$$

By defining, via Bayes Theorem, the conditional distribution (of $\overrightarrow{X}_0$ given $\overrightarrow{X}_t$) as

$$\overrightarrow{p}_{0|t}(\mathrm{d}x, y) = \frac{\mu^\star(\mathrm{d}x)\overrightarrow{p}_t(x, y)}{\overrightarrow{p}_t(y)}$$

we then obtain the score representation

$$\nabla_y \log \overrightarrow{p}_t(y) = - \int \sigma_t^{-2}(y - \mu_t x)\overrightarrow{p}_{0|t}(\mathrm{d}x, y)\,,$$

which equivalently reads as

$$2\,\nabla \log \overrightarrow{p}_t(y) = \sinh(t)^{-1}\,\mathbb{E}[\overrightarrow{X}_0 - e^t\,\overrightarrow{X}_t | \overrightarrow{X}_t = y] \tag{3}$$

Relying on the representation of conditional expectation as L^2-projection (see [Klenke, 2014, Corollary 8.17]) we obtain that the score can be seen as the optimizer in

$$\min_{\mathrm{s}:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}[|\mathrm{s}(\overrightarrow{X}_t) - \sinh(t)^{-1}\,(\overrightarrow{X}_0 - e^t\,\overrightarrow{X}_t)|^2]\,.$$

This identification with the conditional expectation is at the core of the success of score-based generative modeling since it allows in to approximate the score function with a neural network.

Indeed we might restrict the minimization to a (rich enough) parametric family $\{\mathrm{s}^\theta : \theta \in \Theta\}$ and replace the above objective function with its empirical version, which can be computed by considering the forward evolution of our starting data samples. This leads to considering the minimization

$$\min_{\theta\in\Theta} \frac{1}{M}\sum_{i=1}^{M} |\mathrm{s}^\theta(\overrightarrow{X}^i_t) - \sinh(t)^{-1}\,(\overrightarrow{X}^i_0 - e^t\,\overrightarrow{X}^i_t)|^2,$$

where $(\overrightarrow{X}^i_0, \overrightarrow{X}^i_t)_{i\leq M}$ are i.i.d. samples of $(\overrightarrow{X}_0, \overrightarrow{X}_t)$. One can use stochastic gradient descent methods to solve (4).

Finally, in view of the choice that we made of using the relative score in the backward dynamics (2), let us also give the parametric minimization problem associated to the relative score $2\,\nabla \log \overrightarrow{p}_t/\gamma$. This can be easily obtained by noticing that (3) implies

$$2\nabla \log \frac{\overrightarrow{p}_t}{\mathrm{d}\gamma}(y) = \sinh(t)^{-1}\,\mathbb{E}[\overrightarrow{X}_0 - e^{-t}\,\overrightarrow{X}_t | \overrightarrow{X}_t = y]\,.$$

Relying again on the representation of conditional expectation as L^2-projection, we obtain that the relative score can be seen as the optimizer in

$$\min_{\mathrm{s}:\mathbb{R}^d\to\mathbb{R}^d} \mathbb{E}[|\mathrm{s}(\overrightarrow{X}_t) - \sinh(t)^{-1}\,(\overrightarrow{X}_0 - e^{-t}\,\overrightarrow{X}_t)|^2]\,,$$

problem that can be empirically trained as

$$\min_{\theta\in\Theta} \frac{1}{M}\sum_{i=1}^{M} |\mathrm{s}^\theta(\overrightarrow{X}^i_t) - \sinh(t)^{-1}\,(\overrightarrow{X}^i_0 - e^{-t}\,\overrightarrow{X}^i_t)|^2, \tag{4}$$

for a rich enough parametric family $\{\mathrm{s}^\theta : \theta \in \Theta\}$.

1.3 Implementation and limits of SBDM

For a given time horizon $T > 0$, a partition $0 = t_0 < t_1 < ... < t_N = T$ and a choice of approximation $\mathrm{s}^{\theta^\star}_{T-t_k}$ in (4) at time $t = T - t_k$ with $k = 1, \ldots, N$ for the relative score, we can define a score-based diffusion model $X^\star$ recursively as follows

$$X^\star_0 \sim \gamma, \quad \mathrm{d}X^\star_t = [-X^\star_t + \mathrm{s}^{\theta^\star}_{T-t_k}(X^\star_{t_k})]\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t \in [t_k, t_{k+1}] \tag{5}$$

This algorithm can be considered as an Euler scheme for the backward process (2), where

- we have replaced the score with its parametric approximation approximation $X^\star$;
- instead of considering $-X^\star_{t_k}$ in the drift, we have kept $-X^\star_t$ since (5) can be explicitly solved;
- the initial distribution is set to be the Gaussian law γ.

The main question is to bound, in some distance, the distance between the law of $X^\star_T$ (which approximates $\overleftarrow{X}_T \sim \mu^\star$) and the data distribution $\mu^\star$. In order to do so, one should take into account three sources of error

1) The initialization error: we start our score-based diffusion model $X_0^\star \sim \gamma$, whereas for the backward process we have $\overleftarrow{p}_0 = \overrightarrow{p}_T \neq \gamma$. Clearly the greater our time horizon $T \gg 1$ is and the smaller is the mismatch between $\overrightarrow{p}_T$ and γ. Then one should analyze how this approximation error propagates along our diffusion model (5).

2) The score-approximation error: in (5) we replace the true score relative function $\nabla \log \frac{\mathrm{d}\overrightarrow{p}_t}{\mathrm{d}\gamma}$ with a parametric approximation $\mathrm{s}^{\theta^\star}_{T-t}$.

3) The time-discretization error, which is the most challenging source of error to handle since it involves also the approximate score function.

Below we provide the pseudo-code of SBDM algorithm.

Algorithm 1: Score Based Diffusion Model

Input: M samples $X^i \sim \mu^\star$ from our data distribution, time horizon $T \gg 1$ large enough, partition $0 = t_0 < t_1 < \cdots < t_N = T$, parametrized vector field family $\{\mathrm{s}^\theta : \theta \in \Theta\}$

for $i = 1, \dots, M$ **do**

`/* Forward OU evolution */`

Initialize $\overrightarrow{X}^i_0 = X^i$

Sample the forward OU process $(\overrightarrow{X}^i_{t_k})_{k=1,\dots,N}$ using (1)

end

Learn $\theta^\star$ optimizer in (4)

`/* Sample backward evolution, using the parametrized score` $(\mathrm{s}^{\theta^\star}_{T-t_k})_{k=0,\dots,N}$ `*/`

Initialize $X_0^\star = X_0^1$; `// or any` $X_0^\star \sim \mu$

Sample backward process $(X^\star_{t_k})_{k=1,\dots,N}$ via (5)

Output: $X_T^\star$

Let us recall that when sampling from the forward and backward SDEs (1)and (5) one can either solve them exactly, or just use Euler-Maruyama scheme.

1.4 Convergence analysis of SBDM

In this section we are interested in understanding the performance of SBDM algorithm, *i.e.*, in providing quantitative error estimates between the distribution $\mathcal{L}(X_T^\star)$ and our data distribution $\mu^\star$. The result presented in this section are based on [Conforti et al., 2023a]. We will measure the distance between this two probability measure via relative entropy $\mathscr{H}$ (also known as KL-divergence) which is defined for any two probability measure for any two probability measures $\mathfrak{p}, \mathfrak{q}$ as the quantity

$$\mathscr{H}(\mathfrak{p}|\mathfrak{q}) := \begin{cases} \mathbb{E}_{\mathfrak{p}}\left[\log \frac{\mathrm{d}\mathfrak{p}}{\mathrm{d}\mathfrak{q}}\right] & \text{if } \mathfrak{p} \ll \mathfrak{q}, \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

For exposition's sake, here we are going to consider constant step-sizes $t_{k+1} - t_k = h := T/N$. We will assume that we are able to learn the parametric relative score $(\mathrm{s}^{\theta^\star}_{T-t_k})_{k=0,\dots,N}$ such that

A 1. *There exists* $\varepsilon^2 > 0$ *such that*

$$\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}[|\mathrm{s}^{\theta^\star}_{T-t_k}(\overrightarrow{X}_{T-t_k}) - 2\nabla \log \frac{\mathrm{d}\overrightarrow{p}_{T-t_k}}{\mathrm{d}\gamma}(\overrightarrow{X}_{T-t_k})|^2] \leq \varepsilon^2 \,.$$

Assumptions like this, where one assumes that we are able to learn the (relative) score up to an L^2 mistake of order ε^2 are common when analyzing score-based generative model [Chen et al., 2023].

Next, we are going to assume some regularity for our data distribution $\mu^\star$

A 2. *We assume that $\mu^\star$ has finite Fisher information $\mathcal{I}_\gamma(\mu^\star) < +\infty$ (w.r.t. γ), i.e., that*

$$\mathcal{I}_\gamma(\mu^\star) \coloneqq \left\| \nabla \log \frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma} \right\|^2_{\mathrm{L}^2(\mu)} < +\infty \,.$$

Under the following assumptions we will prove that

Theorem 1 (Theorem 1 in [Conforti et al., 2023a])**.** *Assume **A**1 and **A**2. Then SBDM (with constant step-size $h = T/N$) satisfies the following bound*

$$\mathscr{H}(\mu^\star | \mathcal{L}(X_T^\star)) \lesssim e^{-2T}\, \mathscr{H}(\mu^\star | \gamma) + \varepsilon^2\, T + h\, \mathcal{I}_\gamma(\mu^\star) \,,$$

where $\lesssim$ indicates that the inequality holds up to a positive numerical constant (independent from our parameters and from $\mu^\star$).

We will sketch the proof of this result at the end of the current section, after some preliminary results.

The relative score process The proof strategy relies on the study of the behjaviour of the (relative) score process, which is defined as

$$Y_t \coloneqq 2\, \nabla \log \frac{\mathrm{d}\overrightarrow{p}_{T-t}}{\mathrm{d}\gamma}(\overleftarrow{X}_t) \,,$$

where we recall $(\overleftarrow{X}_t)_{t\in[0,T]}$ is the backward process solving (2). In the next result we show that this relative score process satisfies a SDE, that can be interpreted as the adjoint equation in the stochastic maximum principle.

Lemma 1.1. *For any fixed $\delta > 0$, on the time interval $[0, T-\delta]$ the relative score process satisfies*

$$\begin{cases} \mathrm{d}Y_t = Y_t\, \mathrm{d}t + \sqrt{2}\, Z_t \cdot \mathrm{d}B_t \\ \text{with } Z_t = 2\, \nabla^2 \log \frac{\mathrm{d}\overrightarrow{p}_{T-t}}{\mathrm{d}\gamma}(\overleftarrow{X}_t) \,, \end{cases}$$

and by definition we have $Y_T = 2\, \nabla \log \frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma}(\overleftarrow{X}_T)$.

Proof. Firstly, set $f_t \coloneqq \mathrm{d}\overrightarrow{p}_{T-t}/\mathrm{d}\gamma$ and notice that then it satisfies the Kolmogorov equation

$$\begin{cases} \partial_t f_t(x) + \Delta f_t(x) - \langle x,\, \nabla f_t(x) \rangle = 0 \\ f_T = \frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma} \,. \end{cases}$$

Then if we consider $\psi_t \coloneqq \log f_t$ it clearly solves the Hamilton-Jacobi-Bellman equation

$$\begin{cases} \partial_t \psi_t(x) + \Delta \psi_t(x) + |\nabla \psi_t|^2(x) - \langle x,\, \nabla \psi_t(x) \rangle = 0 \\ \psi_T = \log \frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma} \,. \end{cases} \tag{7}$$

Since $Y_t = 2\nabla\psi_t(\overleftarrow{X}_t)$, from Ito's formula[3] we deduce that

$$\begin{aligned}
\mathrm{d}Y_t &= 2\partial_t\nabla\psi_t(\overleftarrow{X}_t)\mathrm{d}t + 2\nabla^2\psi_t\cdot\mathrm{d}\overleftarrow{X}_t + 2\,\Delta\,\nabla\psi_t(\overleftarrow{X}_t)\,\mathrm{d}t\\
&= 2\,[\partial_t\nabla\psi_t(\overleftarrow{X}_t) + \nabla^2\psi_t(\overleftarrow{X}_t)(-\overleftarrow{X}_t - 2\nabla\psi_t(\overleftarrow{X}_t)) + \Delta\,\nabla\psi_t(\overleftarrow{X}_t)]\,\mathrm{d}t + 2\sqrt{2}\,\nabla^2\psi_t(\overleftarrow{X}_t)\cdot\mathrm{d}B_t\\
&= 2\,\nabla[\partial_t\psi_t + \Delta\psi_t + |\nabla\psi_t|^2](\overleftarrow{X}_t)\,\mathrm{d}t - 2\nabla^2\psi_t(\overleftarrow{X}_t)\,\overleftarrow{X}_t\,\mathrm{d}t + \sqrt{2}\,Z_t\cdot\mathrm{d}B_t\\
&\overset{(7)}{=} 2\left(\nabla\langle\overleftarrow{X}_t,\,\nabla\psi_t(\overleftarrow{X}_t)\rangle - \nabla^2\psi_t(\overleftarrow{X}_t)\,\overleftarrow{X}_t\right)\mathrm{d}t + \sqrt{2}\,Z_t\cdot\mathrm{d}B_t = 2\,\nabla\psi_t(\overleftarrow{X}_t)\,\mathrm{d}t + \sqrt{2}\,Z_t\cdot\mathrm{d}B_t\,,
\end{aligned}$$

which concludes our proof. □

Corollary 1.2. *If $\mu^\star$ has finite eight-moments, for any $0\le s\le t\le T$ it holds*

$$\mathbb{E}[|Y_t|^2] - \mathbb{E}[|Y_s|^2] = \int_s^t 2\,\mathbb{E}\Big[|Y_u|^2 + \|Z_u\|_{\mathrm{FR}}^2\Big]\,\mathrm{d}u\,,$$

where $\|Z_t\|_{\mathrm{FR}}$ is the Frobenius norm of the matrix Z_t. As a consequence we obtain that

$$\mathbb{E}[|Y_s|^2] \le e^{-2(t-s)}\,\mathbb{E}[|Y_t|^2]\,.$$

Proof. By Ito's formula and from the previous lemma we immediately deduce that

$$\mathrm{d}|Y_t|^2 = 2|Y_t|^2\,\mathrm{d}t + 2\sqrt{2}\,Y_t\cdot Z_t\cdot\mathrm{d}B_t + 2\|Z_t\|_{\mathrm{FR}}^2\,\mathrm{d}t\,.$$

The finite eight-moments assumption guarantees that $(Y_t\cdot Z_t\cdot\mathrm{d}B_t)_{t\in[0,T)}$ is a martingale[4] and therefore the validity of our first claim. From the non-negativity of the latter term, by taking expectation we then deduce that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[|Y_t|^2] \ge 2\,\mathbb{E}[|Y_t|^2]\,,$$

which combined with (a backward in time) Gronwall's lemma yields to the second claim. □

Proof of main result The proof of our entropy bound will be based on a Girsanov Theorem where the drift at time t may depend on the whole trajectory up to time t. Particularly we are going to prove that

Lemma 1.3. *Take two probability measures $\mu^1,\mu^2\in\mathbb{R}^d$ and for $i=1,2$ let X^i be the solution of*

$$\begin{cases}\mathrm{d}X_t^i = b_t^i(X_{[0,t]}^i)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\\ X_0^i\sim\mu^i\,,\end{cases}$$

where $X_{[0,t]}^i$ stands for the time-evolution of the process up to time t. Then if we set $\mathrm{P}^i := \mathcal{L}(X_{[0,T]}^i)\in\mathcal{P}(\Omega)$ it holds

$$\mathscr{H}(\mathrm{P}^2|\mathrm{P}^1) = \mathscr{H}(\mu^2|\mu^1) + \mathbb{E}\int_0^T |b_t^2 - b_t^1|^2(X_{[0,t]}^2)\,\mathrm{d}t\,.$$

[3] We restrict ourselves in the time interval $[0,T-\delta]$ for any arbitrary small $\delta>0$ in order to guarantee enough regularity for ψ_t and therefore for applying Ito's formula here. We refer the reader to [Conforti et al., 2023a, Proposition 1].

[4] We refer the interested reader to [Conforti et al., 2023a, Lemma 1] for this technical detail.

Proof. Recall that $\omega_\cdot$ denotes the canonical process on Ω. From Girsanov Theorem we have

$$\frac{\mathrm{dP}^2}{\mathrm{dP}^1}(\omega) = \frac{\mathrm{d}\mu^2}{\mathrm{d}\mu^1}(\omega_0)\, \exp\left[2\int_0^T (b_t^2 - b_t^1)(\omega_{[0,t]})\,\mathrm{d}\omega_t - \int_0^t (|b_t^2|^2 - |b_t^1|^2)(\omega_{[0,t]})\,\mathrm{d}t\right].$$

Therefore, the relative entropy reads as

$$\begin{aligned}
\mathscr{H}(\mathrm{P}^2|\mathrm{P}^1) &= \mathscr{H}(\mu^2|\mu^1) + \mathbb{E}_{\mathrm{P}^2}\left[2\int_0^T (b_t^2 - b_t^1)(\omega_{[0,t]})\,\mathrm{d}\omega_t - \int_0^t (|b_t^2|^2 - |b_t^1|^2)(\omega_{[0,t]})\,\mathrm{d}t\right] \\
&= \mathscr{H}(\mu^2|\mu^1) + \mathbb{E}\left[2\int_0^T (b_t^2 - b_t^1)(X^2_{[0,t]})\,\mathrm{d}X_t^2 - \int_0^t (|b_t^2|^2 - |b_t^1|^2)(X^2_{[0,t]})\,\mathrm{d}t\right] \\
&= \mathscr{H}(\mu^2|\mu^1) + \mathbb{E}\left[2\int_0^T (b_t^2 - b_t^1)(X^2_{[0,t]})\, b_t^2(X^2_{[0,t]}) - |b_t^2|^2(X^2_{[0,t]}) + |b_t^1|^2)(X^2_{[0,t]})\,\mathrm{d}t\right] \\
&= \mathscr{H}(\mu^2|\mu^1) + \mathbb{E}\left[2\int_0^T |b_t^2 - b_t^1|^2(X^2_{[0,t]})\,\mathrm{d}t\right]
\end{aligned}$$

□

Sketch of the proof of Theorem 1. We will prove this theorem under the extra assumption the $\mu^\star$ has finite eight-moments.[5]

Let us start by noticing that, since $\mu^\star = \mathcal{L}(\overleftarrow{X}_T)$ we have

$$\mathscr{H}(\mu^\star|\mathcal{L}(X_T^\star)) = \mathscr{H}(\mathcal{L}(\overleftarrow{X}_T|\mathcal{L}(X_T^\star)) \le \mathscr{H}(\mathcal{L}(\overleftarrow{X}_\cdot)|\mathcal{L}(X_\cdot^\star))\,,$$

where the probabilities appearing in the last term are probability measures on the path space Ω and the last inequality is known as *Data processing* inequality for relative entropy [Nutz, 2021, Lemma 1.6] (in this specific case can be deduced from Lemma 1.3 applied to the time reversal of these measures). From Lemma 1.3 we then deduce that

$$\begin{aligned}
\mathscr{H}(\mu^\star|\mathcal{L}(X_T^\star)) &\le \mathscr{H}(\mathcal{L}(\overleftarrow{X}_0)|\mathcal{L}(X_0^\star)) + 2\sum_{k=0}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} \left|2\nabla \log \frac{\mathrm{d}\overrightarrow{p}_{T-t}}{\mathrm{d}\gamma}(\overleftarrow{X}_t) - \mathrm{s}^{\theta^\star}_{T-t_k}(\overleftarrow{X}_{t_k})\right|^2 \mathrm{d}t \\
&= \mathscr{H}(\mathcal{L}(\overrightarrow{X}_T)|\gamma) + 2\sum_{k=0}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} \left|Y_t - \mathrm{s}^{\theta^\star}_{T-t_k}(\overleftarrow{X}_{t_k})\right|^2 \mathrm{d}t \\
&\lesssim \mathscr{H}(\mathcal{L}(\overrightarrow{X}_T)|\gamma) + 2\sum_{k=0}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} |Y_t - Y_{t_k}|^2\,\mathrm{d}t + h\sum_{k=0}^{N-1}\mathbb{E}[|Y_{t_k} - \mathrm{s}^{\theta^\star}_{T-t_k}(\overleftarrow{X}_{t_k})|^2] \\
&\le e^{-2\,T}\mathscr{H}(\mu^\star|\gamma) + 2\sum_{k=0}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} |Y_t - Y_{t_k}|^2\,\mathrm{d}t + \varepsilon^2\,T\,,
\end{aligned}$$

where the last step follows from **A**1 and from the fact that the forward Ornstein-Uhlenbeck process invariant measure is the Gaussian γ which satisfies a log-Sobolev inequality [Bakry et al., 2013, Proposition 5.5.1] and therefore an exponential entropy decay [Bakry et al., 2013, Theorem 5.2.1].

[5] We refer the reader to the proof of Lemma 2 in [Conforti et al., 2023a] for an approximating argument that removes this assumption.

In order to conclude we should bound the term in the middle. Notice that if we integrate from t_k to $t \in [t_k, t_{k+1}]$, from Lemma 1.1 we may deduce that

$$\mathbb{E}|Y_t - Y_{t_k}|^2 \leq 2\,\mathbb{E}\left[\left|\int_{t_k}^t Y_s \mathrm{d}s\right|^2\right] + 4\,\mathbb{E}\left[\left|\int_{t_k}^t Z_s \cdot \mathrm{d}B_s\right|^2\right].$$

From Jensen's inequality and Ito isometry we then have

$$\begin{aligned}\mathbb{E}|Y_t - Y_{t_k}|^2 \leq 2(t-t_k)\mathbb{E}\int_{t_k}^t |Y_s|^2 \mathrm{d}s + 4\mathbb{E}\int_{t_k}^t \|Z_s\|_{\mathrm{FR}}^2 \mathrm{d}s &\lesssim \int_{t_k}^{t_{k+1}} 2\,\mathbb{E}\left[|Y_s|^2 + \|Z_s\|_{\mathrm{FR}}^2\right] \mathrm{d}s \\ &= \mathbb{E}[|Y_{t_{k+1}}|^2] - \mathbb{E}[|Y_{t_k}|^2]\,,\end{aligned}$$

where the last step follows from Corollary 1.2. In conclusion we have shown that

$$\begin{aligned}\mathscr{H}(\mu^\star|\mathcal{L}(X_T^\star)) &\lesssim e^{-2\,T}\mathscr{H}(\mu^\star|\gamma) + 2\,h\sum_{k=0}^{N-1}(\mathbb{E}[|Y_{t_{k+1}}|^2] - \mathbb{E}[|Y_{t_k}|^2]) + \varepsilon^2\,T \\ &= e^{-2\,T}\mathscr{H}(\mu^\star|\gamma) + 2\,h\,(\mathbb{E}[|Y_T|^2] - \mathbb{E}[|Y_0|^2]) + \varepsilon^2\,T \\ &\leq e^{-2\,T}\mathscr{H}(\mu^\star|\gamma) + 2\,h\,\mathbb{E}[|Y_T|^2] + \varepsilon^2\,T\end{aligned}$$

which concludes our proof since

$$\mathbb{E}[|Y_T|^2] = \mathbb{E}\left[\left|\nabla \log \frac{\mathrm{d}\overrightarrow{p}_0}{\mathrm{d}\gamma}(\overleftarrow{X}_T)\right|^2\right] = \mathbb{E}\left[\left|\nabla \log \frac{\mathrm{d}\mu^\star}{\mathrm{d}\gamma}(\overrightarrow{X}_0)\right|^2\right] = \mathcal{I}_\gamma(\mu^\star)\,.$$

□

2 Diffusion Flow Matching

SBDM interpolate between an easy-to-sample from distribution (the Gaussian distribution) and the data distribution in infinite time horizon. This means that in practice in order to have good quality samples from SBDM we need to choose a very large time horizon $T \gg 1$, so that $\overrightarrow{p}_T \approx \gamma$. However considering a wide time horizon is computationally costly. In order to avoid this extra error source and in order to make the algorithm computationally lighter we would like to sample

- a pair of *arbitrary* data distributions
- in *finite* time (let us say $T = 1$).

In other words, we want to design an efficient algorithm that builds a "bridge" between any two distributions. Very recently, a new family of algorithms called *stochastic interpolants* has been recently proposed [Albergo and Vanden-Eijnden, 2023, Shi et al., 2023b, Peluchetti, 2023] to address this problem. In this notes, we focus on a specific instance of this class, namely Diffusion Flow Matching (DFM), which however contains all the main ideas of the theory.

2.1 Building bridges between probability measures

In this section we are interested in building bridges between probability distributions on $\mathbb{R}^d$.

Brownian bridges Let us start by recalling here how Brownian motions can be used in order to build bridges between two points $x, y \in \mathbb{R}^d$ (or equivalently between the Dirac probability measures δ_x and δ_y). Let us denote by R the Wiener measure, *i.e.*, the law on $\Omega = C([0,1];\mathbb{R}^d)$ of the reversible Brownian motion on $[0,1]$ (that is a Brownian motion whose law at any time $t \in [0,1]$ coincides with its equilibrium measure, *i.e.*, the Lebesgue measure)[6]. For any $x, y \in \mathbb{R}^d$ the Brownian bridge R^{xy} from x to y, is defined as the conditional probability measure that for any Borel set $\mathsf{A} \subseteq \Omega$ reads as

$$\mathrm{R}^{xy}[\mathsf{A}] = \mathrm{R}[\mathsf{A} | \, \omega_0 = x, \omega_1 = y] \,,$$

where $(\omega_t)_{t\in[0,1]}$ denotes the canonical process on Ω. The Brownian bridge can be represented also through a stochastic differential equation. Namely, R^{xy} is the law that solves the SDE[7]

$$\begin{cases} \mathrm{d}X_t^{xy} = \frac{(y - X^{xy})}{1-t} \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t \\ X_0^{xy} = x \,. \end{cases} \tag{8}$$

A simple way to build bridges From Brownian bridges between points we can now build bridges between distributions by just considering mixtures of Brownian bridges.

Given any two distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ define the set of couplings of μ and ν as the set of probability measures on the product space $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν respectively, that is the set

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi(A \times \mathbb{R}^d) = \mu(A) \, , \pi(\mathbb{R}^d \times A) = \nu(A) \ \forall A \in \mathcal{B}(\mathbb{R}^d)\}.$$

Then a natural way of building bridges between μ and ν consists in considering a coupling π, sample a *travel plan*, *i.e.*, an initial and final point and then just connect them via a Brownian bridge. In formulas, we consider

$$(X_0^{\mathrm{I}}, X_1^{\mathrm{I}}) \sim \pi, \quad X_t^{\mathrm{I}} = X_t^{X_0^{\mathrm{I}}, X_1^{\mathrm{I}}} \quad \forall t \in [0,1] \,,$$

where for any $x, y \in \mathbb{R}^d$ the process $(X_t^{xy})_{t\in[0,1]}$ is the Brownian bridge from x to y solving (8). We call the stochastic process $(X_t^I)_{t\in[0,1]}$ the stochastic interpolant between μ and ν. Equivalently, if we define P^{I} as the law of the interpolant $(X_t^I)_{t\in[0,1]}$, we have that for any measurable $\mathsf{A} \subseteq \Omega$

$$\mathrm{P}^{\mathrm{I}}[\mathsf{A}] = \mathbb{P}[X_\cdot^{\mathrm{I}} \in \mathsf{A}] = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathrm{R}^{xy}[\mathsf{A}] \, \pi(\mathrm{d}x \, \mathrm{d}y) \,,$$

which indeed is equivalent to saying that P^{I} is a π-mixture of Brownian bridges.

It is straightforward to verify that the marginal flow of P^{I}, which we denote $(p_t^{\mathrm{I}})_{t\in[0,1]}$, is such that $p_0^{\mathrm{I}} = \mu, p_1^{\mathrm{I}} = \nu$. Therefore, for any coupling $\pi \in \Pi(\mu, \nu)$ between μ and ν we have defined a stochastic interpolant process and its corresponding flow, which connects μ to ν.

Reciprocal classes A fundamental observation is that the stochastic interpolant $(X_t^{\mathrm{I}})_{t\in[0,1]}$ is *not* a Markov process in general. However, by construction, it shares the same bridges of the Wiener measure R. More precisely,

$$(\mathrm{P}^{\mathrm{I}})^{xy} = \mathrm{R}^{xy} \quad \forall x, y \in \mathbb{R}^d,$$

[6] This choice guarantees the reversibility and stationarity of the process

[7] Let us draw a connection with the previous section and remark here that this bridge SDE can be thought as the backward dynamics associated to a Brownian motion. Indeed the drift term in the above SDE coincides with the score $2 \nabla \log \overrightarrow{p}_{1-t}(y)$ associated to a forward Brownian diffusion started in $\overrightarrow{X}_0 = y$.

where $(\mathrm{P}^{\mathrm{I}})^{xy}$ is the xy-bridge of P^{I}, that is to say, the law of P^{I} conditionally on $\{\omega_0 = x, \omega_1 = y\}$. Otherwise said, P^{I} belongs to the *reciprocal class* of the Wiener measure R, defined as

$$\mathcal{R} = \{\mathrm{P} : \mathrm{P}[\mathsf{A}] = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathrm{R}^{xy}[\mathsf{A}]\, \mathrm{P}_{0,1}(\mathrm{d}x\, \mathrm{d}y) \quad \forall \mathsf{A} \in \mathcal{B}(\Omega)\}$$

where in the above we denoted by A a general measurable subset of Ω and P_{01} is the joint law at times $0, 1$ of P, *i.e.*, $\mathrm{P}_{0,1} = (\omega_0, \omega_1)_{\#}\mathrm{P}$. Therefore, despite not being a Markov process $(X_t^{\mathrm{I}})_{t\in[0,1]}$, its law is in the reciprocal class of the Brownian motion, *i.e.*, $\mathrm{P}^{\mathrm{I}} \in \mathcal{R}$.

2.2 Markovian projections and DFM algorithm

So far we have shown how to build bridges between two marginal distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ in a simple way (*i.e.*, as a mixture of Brownian bridges). However in order to implement the sampling in an algorithm it would be convenient that P^{I} solves an SDE so that we could sample from P^{I} via Euler-Maruyama scheme. Unfortunately, as we already mentioned P^{I} generally doesn't solve any SDE since it is not Markov. The core idea behind Diffusion Flow Matching (DFM) algorithm is then trying to compensate this lack of Markovianity by computing the Markovian projection of P^{I}. Otherwise said, we are interested in finding a measure P which is Markov and that "mimics" the marginal flow of the stochastic interpolant P^{I}, *i.e.*, such that $(\omega_t)_{\#}\mathrm{P} = p_t^{\mathrm{I}}$ for any fixed $t \in [0,1]$. Equivalently, our aim is learning a drift field

$$b_t : [0,1] \times \mathbb{R}^d \longrightarrow \mathbb{R}^d \quad (t,x) \mapsto b_t(x)$$

such that the solution $X^{\mu,b}$ of the SDE

$$\begin{cases} \mathrm{d}X_t^{\mu,b} = b_t(X_t^{\mu,b})\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \\ X_0^{\mu,b} \sim \mu \end{cases} \tag{9}$$

satisfies

$$\mathcal{L}(X_t^{\mu,b}) = p_t^{\mathrm{I}} \quad \forall t \in [0,1]\,. \tag{10}$$

Notice that, since $p_0^{\mathrm{I}} = \mu$ and $p_1^{\mathrm{I}} = \nu$, we are guaranteed that $X_0^{\mu,b} \sim \mu$ and $X_1^{\mu,b} \sim \nu$, or equivalently that $\mathcal{L}(X_0^{\mu,b}, X_1^{\mu,b}) \in \Pi(\mu,\nu)$ is a coupling between our two marginals. The process $(X_t^{\mu,b})_{t\in[0,1]}$ is often referred to as the Markovian projection of $(p_t^{\mathrm{I}})_{t\in[0,1]}$, or also Gyöngy projection of $(p_t^{\mathrm{I}})_{t\in[0,1]}$, in view of the contribution of [Gyöngy, 1986].

It turns out that such drift always exist and admits a tractable expression in terms of a conditional expectation

Theorem 2. *The drift field*

$$b_t(x) = (1-t)^{-1}\mathbb{E}[X_1^{\mathrm{I}} - X_t^{\mathrm{I}} | X_t^{\mathrm{I}} = x] \tag{11}$$

mimics the marginal flow of P^{I}*, i.e.* (10) *holds.*

Proof. Define $p_t(x,y)$ as the heat kernel[8]

$$p_t(x,y) = (4\pi t)^{-d/2}\, \exp(-|y-x|^2/_{4t}).$$

[8]Pay attention that since we have $\sqrt{2}$ in front of the Brownian motion in the SDE (9), here we have the factor 4 in the heat kernel. This ensure that p_t solves the heat equation $\partial_t p_t = \Delta p_t$ (with the Laplacian Δ and not $\Delta/2$).

Moreover, let us introduce $\tilde{\pi}(\mathrm{d}x\mathrm{d}y) = p_1^{-1}(x,y)\pi(\mathrm{d}x\mathrm{d}y)$, where we recall $\pi \in \Pi(\mu,\nu)$ being here the coupling fixed when building the stochastic interpolant P^{I} (so in particular $\pi = (\omega_0,\omega_1)_{\#}\mathrm{P}^{\mathrm{I}}$). Then, from the definition of P^{I} as π-mixture of Brownian bridges and since R^{xy} and from the Markov property it follows that for any $A \subset \mathbb{R}^d$ it holds

$$\begin{aligned}
p_t^{\mathrm{I}}[A] = \mathbb{P}[X_t^{\mathrm{I}} \in A] &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbb{P}[X_t^{x_0x_1} \in A]\,\pi(\mathrm{d}x_0\,\mathrm{d}x_1) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathrm{R}^{x_0x_1}[\omega_t \in A]\;\pi(\mathrm{d}x_0\,\mathrm{d}x_1)\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathrm{R}^{x_0}[\omega_t \in A|\omega_1 = x_1]\,\pi(\mathrm{d}x_0\,\mathrm{d}x_1)\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\mathrm{R}^{x_0}[\omega_t \in A, \omega_1 = x_1]}{p_1(x_0,x_1)}\,\pi(\mathrm{d}x_0\,\mathrm{d}x_1)\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\mathrm{R}^{x_0}[\omega_t \in A]\,\mathrm{R}^{x_0}[\omega_1 = x_1|\omega_t \in A]}{p_1(x_0,x_1)}\,\pi(\mathrm{d}x_0\,\mathrm{d}x_1)\\
&= \int_A \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{p_t(x_0,x)\,p_{1-t}(x,x_1)}{p_1(x_0,x_1)}\,\pi(\mathrm{d}x_0\,\mathrm{d}x_1)\,\mathrm{d}x
\end{aligned}$$

where R^{x_0} denotes the law of a Brownian motion started in x_0. Therefore we have shown that

$$p_t^{\mathrm{I}}(x) = \int \frac{p_t(x_0,x)\,p_{1-t}(x,x_1)}{p_1(x_0,x_1)}\,\pi(\mathrm{d}x_0,\mathrm{d}x_1) = \int p_t(x_0,x)p_{1-t}(x,x_1)\,\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\,. \tag{12}$$

Combining the above identity with the heat equation $\partial_t p_t(x,y) = \Delta_y p_t(x,y) = \Delta_x p_t(x,y)$ we find

$$\begin{aligned}
\partial_t p_t^{\mathrm{I}}(x) &= \int [\Delta_x p_t(x_0,x)p_{1-t}(x,x_1) - p_t(x_0,x)\Delta_x p_{1-t}(x,x_1)]\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\\
&= \int [(\Delta_x p_t(x_0,x)p_{1-t}(x,x_1) + \nabla_x p_t(x_0,x)\cdot\nabla_x p_{1-t}(x,x_1))\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\\
&- \int (p_t(x_0,x)\Delta_x p_{1-t}(x,x_1) + \nabla_x p_t(x_0,x)\cdot\nabla_x p_{1-t}(x,x_1))]\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\\
&= \nabla_x \cdot \int \left[\nabla_x p_t(x_0,x)p_{1-t}(x,x_1) - p_t(x_0,x)\nabla_x p_{1-t}(x,x_1)\right]\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\\
&= \nabla_x \cdot \int \left(\nabla_x \log p_t(x_0,x) - \nabla_x \log p_{1-t}(x,x_1)\right) p_t(x_0,x)p_{1-t}(x,x_1)\,\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)
\end{aligned}$$

Therefore, p_t^{I} satisfies the continuity equation

$$\begin{aligned}
&\partial_t p_t^{\mathrm{I}} + \nabla_x\cdot(v_t p_t^{\mathrm{I}}) = 0\,,\\
&v_t(x) := (p_t^{\mathrm{I}})^{-1}(x)\int \left(\nabla_x \log p_{1-t}(x,x_1) - \nabla_x \log p_t(x_0,x)\right) p_t(x_0,x)p_{1-t}(x,x_1)\,\tilde{\pi}(\mathrm{d}x_0,\mathrm{d}x_1)\,.
\end{aligned} \tag{13}$$

It then follows, that upon defining

$$b_t(x) := v_t(x) + \nabla_x \log p_t^{\mathrm{I}}(x), \tag{14}$$

we find that p_t^{I} satisfies the Fokker-Planck-equation

$$\partial_t p_t^{\mathrm{I}} + \nabla_x\cdot(b_t p_t^{\mathrm{I}}) - \Delta p_t^{\mathrm{I}} = 0$$

Therefore the drift b_t as defined in (14) is a mimicking drift. It remains to show that definition (14) coincides with (11). To see this, observe that we get directly from (12) that

$$\nabla_x \log p_t^{\mathrm{I}}(x) = (p_t^{\mathrm{I}})^{-1}(x) \int \Big(\nabla_x \log p_t(x_0, x) + \nabla_x \log p_{1-t}(x, x_1) \Big) p_t(x_0, x) p_{1-t}(x, x_1) \tilde{\pi}(\mathrm{d}x_0, \mathrm{d}x_1)$$

Plugging this result into (14), using (13) , recalling (12) and noticing that

$$\nabla_x \log p_{1-t}(x, x_1) = \frac{x_1 - x}{2(1-t)},$$

yield to

$$\begin{aligned} b_t(x) &= (p_t^{\mathrm{I}})^{-1}(x) \int 2\, \nabla_x \log p_{1-t}(x, x_1)\, p_t(x_0, x) p_{1-t}(x, x_1) \tilde{\pi}(\mathrm{d}x_0, \mathrm{d}x_1) \\ &= 2\, \mathbb{E}[\nabla_x \log p_{1-t}(X_t, X_1) | X_t = x] = (1-t)^{-1}\, \mathbb{E}[X_1^{\mathrm{I}} - X_t^{\mathrm{I}} | X_t^{\mathrm{I}} = x]\,. \end{aligned}$$

□

Since the mimicking drift b_t is a conditional expectation, we can again rely on the representation of conditional expectation as L^2-projection (see [Klenke, 2014, Corollary 8.17]) and see the mimicking drift as the optimizer in

$$\min_{\mathrm{s}:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}[|\mathrm{s}(X_t^{\mathrm{I}}) - (1-t)^{-1} (X_1^{\mathrm{I}} - X_t^{\mathrm{I}})|^2]\,.$$

Therefore from an algorithmic point of view, we can again solve this problem via neural networks, *i.e.*, given a rich enough class of parameterized vector fields $\{\mathrm{s}^\theta : \theta \in \Theta\}$ we can approximate the mimicking drift via solving

$$\min_{\theta \in \Theta} \frac{1}{M} \sum_{i=1}^{M} |\mathrm{s}^\theta((X_t^{\mathrm{I}})^i) - (1-t)^{-1}((X_1^{\mathrm{I}})^i - (X_t^{\mathrm{I}})^i)|^2, \tag{15}$$

where $(X_t^{\mathrm{I}}, X_1^{\mathrm{I}})^i$ are i.i.d. samples from $(X_t^{\mathrm{I}}, X_1^{\mathrm{I}})$. If $\theta^\star$ denotes the optimal parameter for the above problem, then $\mathrm{s}^{\theta^\star}$ will be our approximated mimicking drift.

Diffusion Flow Matching algorithm What we have seen so far, is that DFM gives us a way of sampling in finite time a target distribution ν from an initial distribution μ, given a fixed coupling $\pi \in \Pi(\mu, \nu)$. Indeed, given a a partition $0 = t_0 < t_1 < \cdots < t_N = 1$, once the parametric mimicking drift $(\mathrm{s}_{t_k}^{\theta^\star})_{k \leq N-1}$ is learned via (15), we can sample from the approximated mimicking SDE

$$\begin{cases} \mathrm{d}X_t^\star = \mathrm{s}_{t_k}^{\theta^\star}(X_{t_k}^\star)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \text{for } t \in [t_k, t_{k+1}] \\ X_0^\star \sim \mu \end{cases} \tag{16}$$

Then a natural question to address is to upper bound the distance between the law of $X_1^\star$ and ν, since we have $\mathcal{L}(X_1^{\mu,b}) = \nu$ and $X_1^\star$ is an approximated version of the random variable $X_1^{\mu,b}$.

Below we provide the pseudo-code for DFM algorithm, where we employ Euler-Maruyama scheme for solving both the Brownian bridge SDE (8) and the parametrized Markovian projection SDE (16).

Algorithm 2: Diffusion Flow Matching

Input: M samples $(X_0, X_1)^i \sim \pi \in \Pi(\mu, \nu)$ from a tractable distribution, partition $0 = t_0 < t_1 < \cdots < t_N = 1$, parametrized vector field family $\{\mathrm{s}^\theta : \theta \in \Theta\}$

for $i = 1, \ldots, M$ **do**

  `/* Sample stochastic interpolant */`

  Initialize $(X_0^{\mathrm{I}})^i = X_0^i$

  Sample $(X_{t_k}^{\mathrm{I}})_{k=1,\ldots,N}^i$ using the Brownian bridge (8) from $x = X_0^i$ to $y = X_1^i$

end

Learn $\theta^\star$ optimizer in (15)

`/* Sample Markovian projection, using parametrized mimicking drift` $\mathrm{s}_{t_k}^{\theta^\star}$ `*/`

Initialize $X_0^\star = X_0^1$; `// or any` $X_0^\star \sim \mu$

Sample $(X_{t_k}^\star)_{k=1,\ldots,N}$ via (16)

Output: $X_1^\star$

2.3 Connection with entropic optimal transport

In a nutshell, DFM takes as input a coupling $\pi^0 \in \Pi(\mu, \nu)$, then first builds the corresponding stochastic interpolant $(X_t^{\mathrm{I}})_{t\in[0,1]}$, and finally computes a mimicking drift. The corresponding generative model is then given by the process $X_\cdot^\star$. This process is an approximation of the process $X_\cdot^{\mu,b}$, defined by

$$\begin{cases} \mathrm{d}X_t^{\mu,b} = b_t(X_t^{\mu,b})\mathrm{d}t + \mathrm{d}B_t, \\ X_0^{\mu,b} \sim \mu \end{cases}$$

Diffusion Schrödinger Bridge Matching In general, there is no reason to expect that the joint law $\mathcal{L}(X_0^{\mu,b}, X_1^{\mu,b})$ of initial and terminal time of the Markovian projection coincides with the initial coupling, though its marginal distributions do, *i.e.*, $\pi^1 := \mathcal{L}(X_0^{\mu,b}, X_1^{\mu,b}) \in \Pi(\mu, \nu)$. Thus, neglecting the approximation errors produced by replacing $X_\cdot^{\mu,b}$ with $X_\cdot^\star$, we may view DFM as an algorithm that takes as input a given coupling $\pi^0 \in \Pi(\mu, \nu)$ and produces a new coupling $\pi^1 \in \Pi(\mu, \nu)$. It is very tempting to implement this procedure as a recursive algorithm, called Diffusion Schrödinger Bridge Matching (DSBM) [Shi et al., 2023a], which computes a sequence of couplings $(\pi^n)_{n\geq 1} \in \Pi(\mu, \nu)$ according to the following scheme:

Algorithm 3: Diffusion Schrödinger Bridge Matching [Shi et al., 2023b]

Input: $\pi^0 \in \Pi(\mu, \nu)$ tractable distribution (*i.e.*, I can sample from it), N number of iterations

for $i = 0, \ldots, N-1$ **do**

  Define the next stochastic interpolant by $X_\cdot^{\mathrm{I},n} = X_\cdot^{X_0^n, X_1^n}$, with $(X_0^n, X_1^n) \sim \pi^n$

  Compute the mimicking drift $b_t^{n+1}(x) = (1-t)^{-1}\mathbb{E}[X_1^{\mathrm{I},n} - X_t^{\mathrm{I},n} | X_t^{\mathrm{I},n} = x]$

  Define the new coupling as $\pi^{n+1} := \mathcal{L}(X_0^{\mu,b^{n+1}}, X_1^{\mu,b^{n+1}})$

end

Output: π^N

Let us remark here, that from a practical point of view the mimicking drift will be learned as done in DFM Algorithm 2 and that at each step instead of defining π^{n+1} we will actually produce

samples that are distributed according to it, which can be used in the following step as starting sample's set.

Let us forget for now on the practical implementation of DSBM and focus on the sequence of couplings $(\pi^n)_{n\geq 1}$ that generates. A natural question is then: does this sequence converge? What is its limit?

Answering this question draws a bridge between DSBM and Entropic Optimal Transport theory.

Entropic Optimal Transport and the Schrödinger problem Define the probability measure

$$\mathrm{R}^{\mu}_{0,T}(\mathrm{d}x\mathrm{d}y) = \mu(\mathrm{d}x)\exp(-|y-x|^2/_{4T})\mathrm{d}y\,,$$

that is the joint law at time 0 and $2T$ of a Brownian motion started in $B_0 \sim \mu$.

Then the Schrödinger problem (SP) with time horizon $T > 0$ (or regularization T) is defined as

$$\inf_{\pi\in\Pi(\mu,\nu)} \mathscr{H}(\pi|\mathrm{R}^{\mu}_{0,T})\,,$$

where $\mathscr{H}$ is the relative entropy functional, which we recall is defined at (6). This problem dates back to the two seminal papers [Schrödinger, 1931, Schrödinger, 1932] where Schrödinger was interested in finding the best approximation of the joint law of a Brownian motion at times $0, T$ in relative entropy sense among all couplings of two given fixed marginals μ, ν. Remarkably, this problem is equivalent to a regularized version of the classical quadratic Monge-Kantorovich optimal transport problem, known as Entropic Optimal Transport problem (EOT), which reads as

$$\inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2\mathrm{d}\pi + 4T\mathscr{H}(\pi|\mu\otimes\nu)\,.$$

From the above expression is immediate to guess that indeed in the asymptotic regime $T \downarrow 0$ SP/EOT converges to the Monge-Kantorovich Optimal transport problem.

A more remarkable thing is that SP admits a unique solution (the relative entropy is a strictly convex functional and we consider the convex subset $\Pi(\mu,\nu)$), which we will denote as $\pi^\star$ and this solution is incredibly regular. In this notes, the most relevant property of $\pi^\star$ is the Markovianity. To be more precise, set $T = 1$ and let us we denote by $\mathrm{P}^\star$ the $\pi^\star$-mixture of Brownian bridges, *i.e.*, the stochastic interpolant built from $\pi^\star$

$$\mathrm{P}^\star = \mathcal{L}(X_\cdot^{X_0,X_1}) \quad \text{with } (X_0,X_1)\sim\pi^\star\,.$$

Then $\mathrm{P}^\star$ is Markov and $(X_t^{X_0,X_1})_{t\in[0,1]}$ is a Markov process [Léonard, 2014]. This particularly implies that $\pi^\star$ is a fixed point for the iteration of DSBM since the stochastic interpolant is already a Markov process and therefore its Markovian projection will equal the interpolant itself.

One may then expect that, if the sequence of couplings produced by DSBM $(\pi^n)_{n\geq 1}$ converges, then we have

$$\lim_{n\to+\infty}\pi^n = \pi^\star\,,$$

since $\pi^\star$ is a fixed point. This is indeed the case, since for any fixed couple of marginals μ, ν there is a unique Markovian probability measure P on the path space such that $(\omega_0,\omega_1)_{\#}\mathrm{P}\in\Pi(\mu,\nu)$ and which is in the reciprocal class of the Wiener measure (*i.e.*, the conditional laws $\mathrm{P}^{xy} = \mathrm{R}^{xy}$ are Brownian bridges):

Theorem 3 (Proposition 5 in [Shi et al., 2023b]). *If* $\mathrm{P} \in \mathcal{P}(\Omega)$ *be a Markov probability measure, belongs to the reciprocal class of the Wiener measure (i.e.,* $\mathrm{P} \in \mathcal{R}$*) and if* $(\omega_0, \omega_1)_{\#}\mathrm{P} \in \Pi(\mu, \nu)$ *with* $(\omega_0, \omega_1)_{\#}\mathrm{P} \ll \mathrm{R}_{0,1}$[9]*, then* $\mathrm{P} = \mathrm{P}^\star$ *is the law of the Schrödinger bridge and* $(\omega_0, \omega_1)_{\#}\mathrm{P} = \pi^\star$ *is the solution to SP.*

This result follows from [Léonard, 2014, Theorem 2.12b] combined with [Léonard et al., 2014, Theorem 2.14].

Sinkhorn's algorithm We have seen that DSBM provides us with an algorithm that computes $\pi^\star \in \Pi(\mu, \nu)$, that is the solution of SP/EOT. Despite DSBM being quite recent, SP/EOT popularity dates back at least to [Cuturi, 2013] where it was shown that the solution of SP can computed in an efficient and fast way using an iterative algorithm known as Sinkhorn's algorithm (which is even older [Sinkhorn, 1964, Sinkhorn and Knopp, 1967]). The idea behind this iterative algorithm is fitting one marginal constraint at a time in the best possible way (*i.e.*, via entropic projections). To be more precise, if we introduce the subsets

$$\Pi(\mu, \star) \coloneqq \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \quad \text{s.t. } \pi(A \times \mathbb{R}^d) = \mu(A) \text{ for any } A \in \mathcal{B}(\mathbb{R}^d)\},$$
$$\Pi(\star, \nu) \coloneqq \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \quad \text{s.t. } \pi(\mathbb{R}^d \times A) = \nu(A) \text{ for any } A \in \mathcal{B}(\mathbb{R}^d)\},$$

then the algorithm reads as follows

Algorithm 4: Sinkhron's algorithm

Input: $\pi^{0,0} = \nu(\mathrm{d}y) \otimes \mathrm{R}_{0,T}(\mathrm{d}x, y) \in \Pi(\star, \nu)$, N number of iterations

for $i = 0, \dots, N-1$ **do**

  Define $\pi^{n+1,n} \coloneqq \arg\min_{\Pi(\mu,\star)} \mathscr{H}(\cdot|\pi^{n,n})$

  Define $\pi^{n+1,n+1} \coloneqq \arg\min_{\Pi(\star,\nu)} \mathscr{H}(\cdot|\pi^{n+1,n})$

end

Output: $\pi^{N,N-1} \in \Pi(\mu, \star)$ or $\pi^{N,N} \in \Pi(\star, \nu)$

The idea behind the algorithm is that, if the sequence $(\pi^{n+1,n}, \pi^{n+1,n+1})$ converges to a fixed point of the iteration $\pi^\star$, then this fixed point belongs to $\Pi(\mu, \star) \cap \Pi(\star, \nu) = \Pi(\mu, \nu)$ and therefore it is a coupling between my two marginals. Moreover, at each step of the algorithm we just impose one marginal constraint which makes easier the computation of the iterates (in discrete settings it translates in a simple matrix/vector manipulation [Peyré and Cuturi, 2019]). The exponential convergence of Sinkhorn's algorithm has been extensively studied both in discrete [Sinkhorn, 1964, Sinkhorn and Knopp, 1967, Peyré and Cuturi, 2019] and in compact/bounded settings [Chen et al., 2016, Deligiannidis et al., 2024, Berman, 2020]. Quite recently its exponential convergence has been established as well in unbounded settings [Conforti et al., 2023b, Eckstein, 2023].

Differences between DSBM and Sinkhorn Both algorithms alternate between some kind of entropic projections. DSBM alternates

- projections onto the reciprocal class $\mathcal{R}$. This corresponds to the construction of the stochastic interpolant,

[9] This last hypothesis is met along DSBM thanks to [Shi et al., 2023b, Lemma 6] as long as $\pi^0 \ll \mathrm{R}_{0,1}$.

- projections onto the set of Markovian processes. This corresponds to the computation of the mimicking drift.

Sinkhorn's algorithm alternates between

- projections on the set of probability measures fitting only the first marginal $\Pi(\mu, \star)$,
- projections on the set of probability measures fitting only the second marginal $\Pi(\star, \nu)$.

References

[Albergo and Vanden-Eijnden, 2023] Albergo, M. S. and Vanden-Eijnden, E. (2023). Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*.

[Bakry et al., 2013] Bakry, D., Gentil, I., and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media.

[Berman, 2020] Berman, R. J. (2020). The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations. *Numerische Mathematik*, 145(4):771–836.

[Cattiaux et al., 2023] Cattiaux, P., Conforti, G., Gentil, I., and Léonard, C. (2023). Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 59(4):1844–1881.

[Chen et al., 2023] Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. (2023). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*.

[Chen et al., 2016] Chen, Y., Georgiou, T., and Pavon, M. (2016). Entropic and Displacement Interpolation: A Computational Approach Using the Hilbert Metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396.

[Conforti et al., 2023a] Conforti, G., Durmus, A., and Silveri, M. G. (2023a). Score diffusion models without early stopping: finite Fisher information is all you need. *arXiv preprint arXiv:2308.12240*.

[Conforti et al., 2023b] Conforti, G., Durmus, A. O., and Greco, G. (2023b). Quantitative contraction rates for Sinkhorn algorithm: beyond bounded costs and compact marginals. *arXiv preprint arXiv:2304.04451*.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

[Deligiannidis et al., 2024] Deligiannidis, G., de Bortoli, V., and Doucet, A. (2024). Quantitative uniform stability of the iterative proportional fitting procedure. *The Annals of Applied Probability*, 34(1A):501 – 516.

[Eckstein, 2023] Eckstein, S. (2023). Hilbert's projective metric for functions of bounded growth and exponential convergence of Sinkhorn's algorithm. *arXiv preprint arXiv:2311.04041*.

[Gyöngy, 1986] Gyöngy, I. (1986). Mimicking the one-dimensional marginal distributions of processes having an ito differential. *Probability Theory and Related Fields*, 71(4):501–516.

[Klenke, 2014] Klenke, A. (2014). *Probability Theory: A Comprehensive Course*. Springer.

[Léonard, 2014] Léonard, C. (2014). A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574.

[Léonard et al., 2014] Léonard, C., Rœlly, S., and Zambrini, J.-C. (2014). Reciprocal processes. A measure-theoretical point of view. *Probability Surveys*, 11(none):237 – 269.

[Nutz, 2021] Nutz, M. (2021). Introduction to Entropic Optimal Transport. *http://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf*.

[Peluchetti, 2023] Peluchetti, S. (2023). Diffusion bridge mixture transports, schrödinger bridge problems and generative modeling.

[Peyré and Cuturi, 2019] Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

[Schrödinger, 1931] Schrödinger, E. (1931). Über die Umkehrung der Naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153.

[Schrödinger, 1932] Schrödinger, E. (1932). La théorie relativiste de l'électron et l' interprétation de la mécanique quantique. *Ann. Inst Henri Poincaré*, (2):269 – 310.

[Shi et al., 2023a] Shi, Y., Bortoli, V. D., Campbell, A., and Doucet, A. (2023a). Diffusion schrödinger bridge matching. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[Shi et al., 2023b] Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. (2023b). Diffusion Schrödinger bridge matching. *arXiv preprint arXiv:2303.16852*.

[Sinkhorn, 1964] Sinkhorn, R. (1964). A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.

[Sinkhorn and Knopp, 1967] Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

[Song et al., 2021] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.