1 Diffusion Models

Diffusion models [SDWMG15, HJA20] have become the most successful way in recent past to sample from a given distribution. One of the most notable implementations is DALL · E from Open AI, which generates images based on a prompt. In mathematical terms, the problem is to sample from an underlying distribution from which we only have samples. Our starting point is a given dataset

$$\{X_i\}_{i=1}^N \text{ which are independent samples from an underlying distribution } \rho_* \in \mathcal{P}(\mathbb{R}^L). \tag{DATA}$$

We should note that practitioner consider the distribution to be an encoded version of the actual data set whose ambient dimension is much larger $K \gg L$ and hence more complicated to manage. See section 1.3 for details on encoding and the underlying manifold hypothesis.

As a basis we take the standard normal distribution $\mu = \mathcal{N}(0,1) \in \mathcal{P}(\mathbb{R}^L)$. Our objective is to find a smooth an easy to evaluate generator function $G : \mathbb{R}^L \to \mathbb{R}^L$, such that $d(G\#\mu, \rho_*)$ is small for some meaningful metric d. A standard choice for the metric is to fix the discussion is the 2-Wasserstein distance, which is better suited than relative entropy for distributions whose support are lower dimensional, see the discussion in [ACB17] and section 1.3. Stable Diffusions constructs the mapping G, by inverting the Ornstein-Ulhembeck (OU) flow from $\rho_* \to \mu$. Namely, independent of the initial data ρ_* we know that the solution $\rho : [0,\infty) \times \mathbb{R}^L \to \mathbb{R}^L$ of linear parabolic equation

$$\begin{cases} \partial_t \rho = \Delta \rho + \nabla(x\rho) & (0,\infty) \times \mathbb{R}^L \\ \rho_0 = \rho_* \end{cases} \tag{OU}$$

satisfies that for most metrics the flow converges exponentially. For example, χ^2, relative entropy and 2-Wasserstein distance $d_2^2(\rho_t, \mu) \lesssim e^{-t}$. This convergence can be refined if the mean and the variance of ρ_* match those of μ, then we have the convergences improves to e^{-3t}. In lines with the dynamic interpretation of the 2-Wasserstein distance of Benamou-Brenier [BB00], the equation (OU) can be interpreted as a continuity equation. Namely, we can consider ρ_t to be the solution of

$$\begin{cases} \partial_t \rho + \nabla \cdot (V\rho) = 0 & (t,x) \in (0,\infty) \times \mathbb{R}^L \\ \rho_0 = \rho_* & x \in \mathbb{R}^L, \end{cases} \tag{Cont}$$

where the velocity field $V : (0,\infty) \times \mathbb{R}^L \to \mathbb{R}^L$ is given by $V = -\nabla \log \rho - x$. This equation can be solved by characteristics.

$$\begin{cases} \frac{d}{dt} X_t(x) = -\nabla \log \rho_t(X_t(x)) - X_t(x) & t \in (0,\infty) \\ X_0(x) = x. \end{cases} \tag{ODE}$$

Namely, we have that the solution to (OU) is given by $\rho_t = X_t\#\rho_*$. We should note that we need to know $\log \rho_t$ is smooth for (ODE) to make strict sense. This is directly related to the smoothness of ρ_* as well as a bound from below. See (Tweedies), for the inversion of the (SDE) versus the (ODE), which allows for lower regularity. See also the discussion before the objective, about well-posedness. The inversion of the flow is then given by inverting the characteristic flow in time (ODE). Namely, picking a time horizon $T \in (0, \infty)$, we consider

$$\begin{cases} \frac{d}{dt} Z_t^T(z) = -\nabla \log \rho_t(Z_t^T(z)) - Z_t^T(z) & t \in (0,T) \\ Z_T^T(z) = z, \end{cases} \tag{ODE^{-1}}$$

which is equipped with a boundary condition at the end point. Using that $d(\rho_T, \mu) \lesssim e^{-T}$, which is small for T large enough, we can consider sampling the terminal condition from the Gaussian $z \sim \mu$. In particular, this is equivalent to considering the continuity equation (Cont) with boundary condition at $t = T$:

$$\begin{cases} \partial_t \nu^T + \nabla \cdot (V \nu^T) = 0 & (t,z) \in (0,T) \times \mathbb{R}^L \\ \nu_T^T = \mu & z \in \mathbb{R}^L. \end{cases} \tag{Cont^{-1}}$$

Hence, an approximation of the original measure is given by the following mapping $\rho_* \sim \nu_0^T = Z_0^T \# \mu$. In particular, the generator function would be given by $G = Z_0^T$. Namely, if Z is a sample from μ, then a sample of ρ_* can be approximated by integrating (ODE^{-1}) numerically. Hence, in this way the problem of sampling has been transformed into a supervised learning problem in which we are interested in finding an approximation of the score function

$$\log \rho_t. \tag{Score Function}$$

Similarly, we can try to invert the (SDE) associated to (OU). The mathematical foundation for the inversion of the SDE is given in [And82]. In a nutshell, Brownian motion can be inverted path by path, if we have access to (Score Function). Here, we show how we can recover the backward process, using Bayes' Theorem. To be clear, we consider an Stochastic Particle associated to the (OU) process

$$dX = -X + \sqrt{2} dB. \tag{SDE}$$

If we try to infer the past given the future, we obtain Tweedies' formula which accounts for the likelihood of the path of the particle. For $s < t$, applying Bayes' theorem, we get

$$\begin{aligned} \mathbb{E}[e^{s-t} X_s | X_t = y] &= \int_{\mathbb{R}^L} e^{(s-t)} x P(X_s = x | X_t = y)\, dx \\ &= y + \int_{\mathbb{R}^L} (e^{s-t} x - y) P(X_s = x | X_t = y)\, dx \\ &= y + \frac{1}{P(X_t = y)} \int_{\mathbb{R}^L} (e^{(t-s)} x - y) P(X_t = y | X_s = x) P(X_s = x)\, dx \\ &= y - (1 - e^{-2(t-s)}) \nabla \log \rho_t(y), \end{aligned} \tag{Tweedies}$$

where in the last step we notice that the expression coincides with the logarithm of the (Score Function), up to a multiplicative constant. Similarly, evaluating at a general smooth function ϕ, we can differentiate (Tweedies) to obtain

$$-\frac{d}{ds} \mathbb{E}[\phi(X_s) | X_t = y] \Big|_{s = t^-} = \nabla \phi(y) y + 2 \nabla \phi(y) \nabla \log \rho_t(y) + \Delta \phi(y) =: \mathcal{L}_{\rho_t} \phi(y). \tag{1}$$

The operator $\mathcal{L}_{\rho_t}$ is the generator of a unique stochastic process moving backwards in time:

$$\begin{cases} d\hat{Z}_t^T = -\hat{Z}_t^T - 2 \nabla \log \rho_t(\hat{Z}_t^T) + \sqrt{2} dB_t \\ \hat{Z}_T^T(z) = z \end{cases} \tag{SDE^{-1}}$$

Similar to (Cont^{-1}), we can furnish the standard normal μ as the boundary conditions at $t = T$ to obtain the Fokker-Planck equation

$$\begin{cases} \partial_t \hat{\nu}_t - \nabla \cdot ((z + 2\nabla \log \rho_t)\hat{\nu}_t) = -\Delta \hat{\nu} \\ \hat{\nu}_T^T = \mu. \end{cases} \tag{OU^{-1}}$$

We should note (OU^{-1}) is well posed, even if there is a negative $\Delta\hat{\nu}$ in the right hand, as we are flowing backwards in time. Again having access to the score function, numerically integrating (SDE^{-1}) we have a another approximation of the original measure $\rho_* \sim \hat{\nu}_0^T = \hat{Z}_0^T \# \mu$.

We should note that for the stochastic case (SDE^{-1}), there are some references already in the literature analyzing this situation [CCL$^+$23, DB23b, DB23a, LLT23, CDS23, LWCC23, WHT23], mostly with the restrictive assumption that the density of the distribution ρ_* is Lipschitz. Only [CDS23] reduces the hypothesis to finite Fisher relative information, for the stochastic case (OU^{-1}). None of them actually cover the relevant case of what happens when the underlying measure is not absolutely continuous, or understand what is the output of the algorithm in this case. In fact, this is the case where the algorithm is being implemented, see section 1.3. A first objective, is to understand an implicit biases that the algorithm might be introducing in the low regularity setting.

Objective 1:. Estimate quantitatively the distance $d(\rho_*, \nu_0^T)$ and $d(\rho_*, \hat{\nu}_0^T)$, studying particularly the sensitivity to approximations of the Score Function. The main objective is to obtain results in the 2-Wasserstein distance, that are independent of the regularity of ρ_*.

Here is where I want to leverage my experience dealing with low regularity objects in Geometric Measure Theory [DMMN18, DM19, DW24] as well as Wasserstein gradient flow related problems [DYY22, CDD$^+$19, CDFL22, CDM16] for measure valued evolutions. See section 1, for a glimpse into the regularization procedure the algorithm introduces to empirical measures. Moreover, my experience with the related Wasserstein Generative Adversarial Networks algorithm [DSC24], allows me to compare with a classical benchmark in the literature. From the mathematical perspective the most general conditions to have well posedness of transport equation (Cont^{-1}) are the ones associated to the Di Perna-Lions theory of renormalized solutions [DL89], which was then extended by Ambrosio [Amb04] to include BV vector fields with bounded divergence. For the Fokker-Planck equation (OU^{-1}), the most general is to require the vector fields to have bounded divergence, and belong to L^2 in space and time, see Le Bris and Lions [BL08]. We should note, that we if don't assume regularity in the initial distribution, the score function will never satisfy the conditions in [Amb04] or [BL08]. With respect to my own work, in terms of making sense of vector fields with negative regularity in critical spaces and non-local diffusion, see [DS18].

Relationship with the JKO scheme. Discretizing (ODE^{-1}), and applying Forward Euler we obtain the mapping $(I - \Delta t(x + \nabla \log \rho))$, appears naturally in the approximation of the (OU) by JKO algorithm, which was was originally proposed in the seminal paper [JKO98]. Namely, we can approximate the (OU) equation, considering $\rho_{n+1} = \arg\min_\rho \frac{d_2^2(\rho_n,\rho)}{2} + \Delta t \mathcal{H}(\rho|\mu)$ where $\mathcal{H}(\rho|\mu) = \begin{cases} \int_{\mathbb{R}^K} \frac{d\rho}{d\mu} \log \frac{d\rho}{d\mu}\, d\mu & \rho \ll \mu \\ +\infty & \text{otherwise} \end{cases}$ is the relative entropy with respect to the Gaussian. The sufficient condition to be the minimizer in JKO step is given by $(I-\Delta t(x+\nabla \log \rho_{n+1}))\#\rho_{n+1} = \rho_n$, just like backward Euler for ODE's, this is a implicit characterization. Using convexity, we know that this mapping is almost contractive in the d_2 distance, see [AGS08, Section 10.1.1].

Empirical measures Although a usual hypothesis is that the data points $\{X_i\}_{i=1}^N$ are being sampled from an underlying distribution ρ_*, this distribution will never be smooth. Hence, it is enlightening to consider the case that the data distribution is exactly an empirical measure $\rho_* = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$.

$$\rho_{proxy} = (I - \Delta t \nabla \log \rho_{\Delta t})\# e^{\Delta t(-\Delta)}\rho_*. \qquad \text{(proxy)}$$

For a given $y \sim \rho_{\Delta t}$, we map it to the function $\overline{X}_{\Delta t}(y)$ which is a weighted average of X_i by the exponential of the distance to y:

$$\overline{X}_{\Delta t}(y) = \sum_{i=1}^N X_i \nu^i_{\Delta t}(y) \qquad \text{with} \qquad \nu^i_{\Delta t}(y) = \frac{e^{-\frac{|y-X_i|^2}{2\Delta t}}}{\sum_{j=1}^N e^{-\frac{|y-X_j|^2}{2\Delta t}}}.$$

Applying Laplace's principle we readily obtain that

$$\overline{X}_{\Delta t}(y) \to_{\Delta t \to 0} \arg\min_{X_i} |y - X_i|.$$

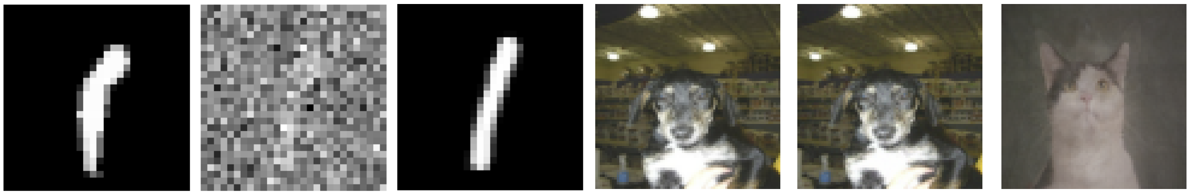


Figure 1: From left to right, the samples X_r, y and $\overline{X}_{\Delta t}(y)$, for the datasets MNIST and Cats and Dogs dataset, where we tookout the bases image X_r to obtain a more diverse sample.

Objective 2:. Understand the regularization process of empirical measures, in 1-step of the diffusion model algorithm. Check under which topology this regularization is continuous. Clarify if there are any hidden biases introduced by the regularization.

1.1 Control for sub-riemmanian manifolds

Sub-Riemannian Manifolds. We consider M a smooth Riemannian manifold, with TM its usual vector bundle. Due to inherent constraints of the modeled control system, we consider that for a given $p \in M$, we can only actuate in subspace $D_p \subset TM_p$ of all possible tangent directions. More formally, we consider that the distribution D is a subbundle of TM.

Control Systems on Sub-Riemannian Manifolds. We consider the simplest system of control on a sub-Riemannian manifold given by a *driftless system*. That is to say, that the system is described by the following differential equations:

$$\dot{p} = \sum_{i=1}^{m} u_i X_i(p), \qquad \text{(driftless system)}$$

where $p \in M$ is the state of the system (a point on the manifold), $X_1, \ldots, X_m$ are the vector fields that span the distribution D at each point of the manifold (i.e. $\mathrm{span}\{X_1(p), ..., X_m(p)\} = D_p$), and $u_1, \ldots, u_m \in \mathbb{R}$ are the control inputs (scalar controls applied to the vector fields $X_1, \ldots, X_m$). The trajectory of the system is constrained to move within the subspace defined by D, meaning that it is restricted to a set of accessible directions at each point.

Example: Dynamics of the Car-like Robot. We consider the simplest relatable system of trying to park a car. The state $p = (x, y, \theta, \varphi)$, represents the the position $(x, y) \in \mathbb{R}^2$, the orientation of the car $\theta \in \mathbb{R}$ and the steering $\varphi \in (-\pi/4, \pi/4)$ which is bounded to avoid destabilization. The deterministic dynamics of the car-like robot are:

$$\dot{x} = u_1 \cos\theta, \quad \dot{y} = u_1 \sin\theta, \quad \dot{\theta} = \frac{u_1}{L} \tan\varphi, \quad \dot{\varphi} = u_2,$$

where L is the wheelbase of the car, u_1 determines the forward velocity, and u_2 the rate of steering.

Controllability. The Chow-Rashevskii theorem is a key result in sub-Riemannian geometry. The theorem states that if the Lie algebra generated by the control vector fields $X_1, \ldots, X_m$ spans the tangent space T_pM at each point $p \in M$, then the system is controllable, see [CR39, Ras40, Sus95].

In other words, if

$$\mathrm{span}\{X_1(p), X_2(p), \ldots, X_m(p)\} = T_pM \quad \text{for all } p \in M, \qquad \text{(Lie Algebra)}$$

then for any pair of points $p_0, p_1 \in M$, there exists a control input $u(t)$ such that the trajectory of (driftless system) starting from p_0 reaches p_1. In terms of the Car-like robot, the everyday example is to parallel park.

For the PDE community, the condition (Lie Algebra) is usually associated to Hörmander's hypoellipticity [Hör67, Hör71], on the regularization properties of the associated hypo-elliptic operator

$$\Delta_{\mathcal{D}} = \sum_{i=1}^{m} X_i^2 + (\nabla_\omega \cdot X_i) X_i, \tag{2}$$

where $\nabla_\omega \cdot$ is the divergence associated to the intrinsic volume form ω. For a newer account of Hörmander's theorem using Maliavan calculus, see [Hai11]. For an introduction to heat equation associated to (2), see [ABB19a, Chapter 21].

Finding a control using Stable Diffusion. Although the Chow-Rashevskii theorem shows the existence of a control, finding that control is not necessarily an easy task [AS92, Jea14, ABB19b]. A way of constructing such control can be informed by Stable diffusions, see [EGP24, GHS24].

To illustrate this point we use the Car-like robot example, and we look to find a control that drives the vehicle from point p_0 to point p_1. We consider the time evolution of replacing the controls u_1 and u_2 with derivatives of Brownian motion dW_1 and dW_2, staring from the desired target location p_1. That is to say we consider the evolution of SDE for some time horizon $[0, T]$:

$$dx = \cos\theta \, dW_1, \quad dy = \sin\theta \, dW_1, \quad d\theta = \frac{1}{L} \tan\varphi \, dW_1, \quad d\varphi = dW_2,$$

where W_2 is reflected through the boundary of $[-\pi/4, \pi/4]$, see [FKDB$^+$24] for the stable diffusion algorithm in a bounded domain. To invert time, we first re-write this SDE in the more familiar form

$$\begin{cases} dz = \sigma(z)dW \\ z_0 = p_1 \end{cases}, \qquad \text{with} \qquad \sigma(z) = \begin{pmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ \frac{1}{L}\tan\varphi & 0 \\ 0 & 1 \end{pmatrix}. \tag{3}$$

Up to adding boundaries for (x, y) and periodicing the orientation θ, we know that that as $t \to \infty$ the $\rho_t = \mathrm{Law}(z_t)$ converges to the volume element in M. Following the guiding principle of Stable Diffusions, we consider a time horizon $T \in (0, \infty)$ and apply the results in Anderson [And82], to consider inversion of the SDE (3):

$$\begin{cases} dq = -\sigma(q)\left(\sigma^t(q)\nabla \log \rho(q) + d\tilde{W}\right) \\ q_T = p_0, \end{cases} \tag{4}$$

where $\tilde{W}$ is a standard Brownian motion moving backward in time. To be able to invert the system, [And82] requires that the law is smooth, which is assured by Hörmander theorem using the condition on the (Lie Algebra). Hence, if want to find a control that brings the point p_0 to the point p_1: We flow the SDE (3) forward in time, and record $\nabla \log \rho_t$ the gradient of the (Score Function), and we can create a time dependent feedback control, which has the form

$$\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \sigma^t(q)\nabla \log \rho_{T-t}(q) + d\tilde{W}. \tag{5}$$

With probability 1 with repect to the Brownian Motion, the stochastic controller (5) will drive the particle from p_0 to p_1.

Objectives. One of the first questions we are interested in is safety of this controller. Namely, can small perturbations of the gradient of the (Score Function) create undesirable behavior? This construction also raises the question of how is this controller (5) related to an optimal controller, which gives rise to the geodesic in the sub-Riemmanian structure. A possibility to recover a deterministic control is to consider the small noise regime, by introducing a small parameter $\varepsilon \to 0^+$ and taking it to zero. This is akin to recovering the Wasserstein distance from the Schrödinger bridge problem, see [Léo13] or more generally [GT20]. In applied math, this connection is usually known through Sinkhorn's algorithm which uses an entropic regularization to recover the Wasserstein distance as $\varepsilon \to 0^+$, see [PC$^+$19].

Objective 3:. Similar to standard stable diffusion, see objective 1, study the sensitivity of the controller (5) with respect to the score function and the initial and final points, p_0 and p_1. Use the insights from the convergence of Schrödinger bridges to Wasserstein geodesics [GT20], to show that as $\varepsilon \to 0^+$ the measure over paths of trajectories which is induced by the stochastic controller (5) has a Laplace principle with respect to the geodesic in the sub-Riemannian structure. Namely, almost surely as $\varepsilon \to 0^+$ the trajectory controlled trajectory converges to the geodesic.

1.2 Conditional Sampling implementation

Another main advantage of diffusion models is their flexibility to sample from conditional distributions. More specifically, we consider a possibly stochastic observation operator $\mathcal{A} : \mathbb{R}^L \to \mathbb{R}^m$ with $m \leq L$. We want to sample from the conditional distribution of having a given observation

$$y = \mathcal{A}(x_0) \in \mathbb{R}^m.$$

Using Bayes' theorem, we have that we can decouple the gradient of the associated score function

$$\nabla \log \rho_t(x|y) = \nabla \log \rho_t(y|x) + \nabla \log \rho_t(x).$$

So we can invert the (OU) flow associated to $\rho_t(x|y)$, if we have access to $\rho_t(y|x)$. This can be done using extra training, by creating a new dataset that tracks the evolution of the observation operator through (OU), see [SSDK+21, Appendix I]. Although this is theoretically possible, any particular problem requires to generate a new tailored dataset, and more computing time for extra training. Instead, a popular, and comparatively much easier alternative, is to use Diffusion Posterior Sampling (DPS), which was proposed in [CKM+23]. We start by re-writing the conditional probability over all possible initial conditions of the diffusion process

$$P(y|X_t) = \mathbb{E}_{X_0 \sim P(X_0|X_t)}[P(y|X_0, X_t)] = \mathbb{E}_{X_0 \sim P(X_0|X_t)}[P(y|X_0)].$$

The next step is to approximate this expectation by using (Tweedies) formula

$$\hat{X}_0(X_t) = \mathbb{E}_{X_0 \sim P(X_0|X_t)}[X_0] = e^t \left(X_t - (1 - e^{-2t}) \nabla \log \rho_t(X_t) \right).$$

We approximate the conditional probability by

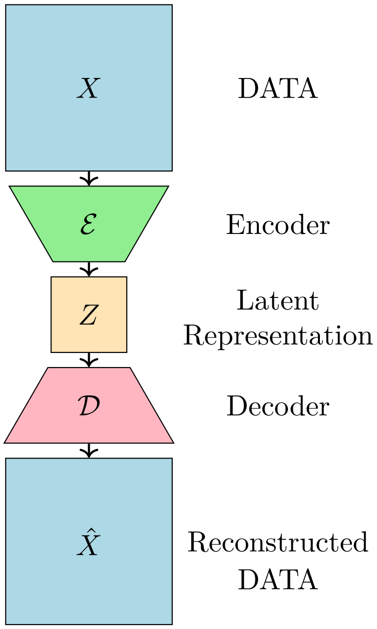
$$P(y|X_t) = \mathbb{E}_{X_0 \sim P(X_0|X_t)}[P(y|X_0, X_t)] \sim P(y|\mathbb{E}_{X_0 \sim P(X_0|X_t)}[X_0]) = P(y|\hat{X}_0(X_t)).$$

This expression can then be differentiated in terms of X_t as the expression for $\hat{X}_0(\cdot)$ and the observation mapping $\mathcal{A}(\cdot)$ are considered to be known explicitly. Although this approximation gives a really fast and easy implementation, little is known about what is the inherent bias of such an approximation to the sampling.

Objective 4:. Understand what is the exact measure the DPS algorithm is actually sampling from, even under regularity assumptions. Compare and contrast with new alternatives of sampling, that consider higher order approximations of the posterior sampling [RCK+24].

The discussion above purposely ignores the encoding decoding steps associated to Latent Diffusion Models, see section 1.3. This introduces another layer of difficulty, as the decoding from latent space to pixel space is one-to-many, see [CCS+23] for an algorithm that considers this more in depth.

1.3 The Manifold Hypothesis: Latent Diffusion Models



Examples of datasets like (DATA) are high resolution images or audio clips. In practice, not every possible configuration of pixels forms a realistic image. Instead, meaningful images, like those of faces or landscapes, lie on some lower dimensional "manifold" that captures the patterns and structures inherent to real images. The hypothesis implies that the essential information about the content of a high-resolution image is compressed in fewer dimensions than the full pixel count would suggest. This idea underpins most of the generative algorithms right now like Autoencoders [HRW86, ERVO21] and GANs [GPAM+14, ACB17]. The loss function is typically adversarial, meaning that an auxiliary discriminator network is trained in parallel, to differentiate real samples X from reconstructed samples $\hat{X} = \mathcal{D}(\mathcal{E}(X))$.

Today, Latent Diffusion Models (LDM) [RBL+22] are the most successful implementations of diffusion models. The idea is to separate training into two different parts. Given an original data distribution in pixel space $\rho_* = \mathcal{P}(\mathbb{R}^K)$. The auto-encoder training step is done a-priori, so that the diffusion model is used only to recover the induced latent distribution $\hat{\rho}_* =: \mathcal{D}\#\rho_* \in \mathcal{P}(\mathbb{R}^L)$ with $L \ll K$. In this way, the sampling problem becomes a much lighter computational task.

References

[ABB19a] Andrei Agrachev, Davide Barilari, and Ugo Boscain. *A comprehensive introduction to sub-Riemannian geometry*, volume 181. Cambridge University Press, 2019.

[ABB19b] Andrei A. Agrachev, Davide Barilari, and Ugo Boscain. *A Comprehensive Introduction to Sub-Riemannian Geometry*, volume 181 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2019.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[Amb04] Luigi Ambrosio. Transport equation and cauchy problem for bv vector fields. *Inventiones mathematicae*, 158(2):227–260, 2004.

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

[AS92] Andrei A. Agrachev and Andrey V. Sarychev. Filtrations of a lie algebra of vector fields and the nilpotent approximation of controllable systems. *Soviet Mathematics Doklady*, 35:467–471, 1992.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[BL08] C Le Bris and P-L Lions. Existence and uniqueness of solutions to fokker–planck type equations with irregular coefficients. *Communications in Partial Differential Equations*, 33(7):1272–1317, 2008.

[CCL$^+$23] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023.

[CCS$^+$23] Hyungjin Chung, Sinho Chewi, Ricardo Silva, Simon Lacoste-Julien, and Arnaud Doucet. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *arXiv preprint arXiv:2307.00619*, 2023.

[CDD$^+$19] José A Carrillo, Matías G Delgadino, Jean Dolbeault, Rupert L Frank, and Franca Hoffmann. Reverse hardy–littlewood–sobolev inequalities. *Journal de Mathématiques Pures et Appliquées*, 132:133–165, 2019.

[CDFL22] José A Carrillo, Matias G Delgadino, Rupert L Frank, and Mathieu Lewin. Fast diffusion leads to partial mass concentration in keller–segel type stationary solutions. *Mathematical Models and Methods in Applied Sciences*, 32(04):831–850, 2022.

[CDM16] Jose A Carrillo, Matias G Delgadino, and Antoine Mellet. Regularity of local minimizers of the interaction energy via obstacle problems. *Communications in Mathematical Physics*, 343:747–781, 2016.

[CDS23] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: Finite fisher information is all you need. *arXiv preprint arXiv:2305.11234*, 2023.

[CKM$^+$23] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[CR39] Wei-Liang Chow and P Rashevsky. On the topology of submanifolds in euclidean space. *Mathematica (Cluj)*, 15(1):12–17, 1939.

[DB23a] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2305.12345*, 2023.

[DB23b] Valentin De Bortoli. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 2023.

[DL89] Ronald J DiPerna and Pierre-Louis Lions. Ordinary differential equations, transport theory and sobolev spaces. *Inventiones mathematicae*, 98(3):511–547, 1989.

[DM19] Matias Gonzalo Delgadino and Francesco Maggi. Alexandrov's theorem revisited. *Anal. PDE*, 12(6):1613–1642, 2019.

[DMMN18] Matias G Delgadino, Francesco Maggi, Cornelia Mihaila, and Robin Neumayer. Bubbling with l 2-almost constant mean curvature and an alexandrov-type theorem for crystals. *Archive for rational mechanics and analysis*, 230:1131–1177, 2018.

[DS18] Matías G Delgadino and Scott Smith. Hölder estimates for fractional parabolic equations with critical divergence free drifts. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, volume 35, pages 577–604. Elsevier, 2018.

[DSC24] MG Delgadino, Bruno B Suassuna, and Rene Cabrera. Gan: Dynamics. *arXiv e-prints*, pages arXiv–2405, 2024.

[DW24] Matias G Delgadino and Daniel Weser. A heintze-karcher inequality with free boundaries and applications to capillarity theory. *Journal of Functional Analysis*, 287(9):110584, 2024.

[DYY22] Matias G Delgadino, Xukai Yan, and Yao Yao. Uniqueness and nonuniqueness of steady states of aggregation-diffusion equations. *Communications on Pure and Applied Mathematics*, 75(1):3–59, 2022.

[EGP24] Karthik Elamvazhuthi, Darshan Gadginmath, and Fabio Pasqualetti. Denoising diffusion-based control of nonlinear systems. *arXiv preprint arXiv:2402.02297*, 2024.

[ERVO21] Patrick Esser, Robin Rombach, Arash Vahdat, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021.

[FKDB$^+$24] Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael John Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2024.

[GHS24] Erlend Grong, Karen Habermann, and Stefan Sommer. Score matching for subriemannian bridge sampling. *arXiv preprint arXiv:2404.15258*, 2024.

[GPAM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GT20] Nicola Gigli and Luca Tamanini. Benamou–brenier and duality formulas for the entropic cost on $rcd_*(k, n)$ spaces. *Probability Theory and Related Fields*, 176(1):1–34, 2020.

[Hai11] Martin Hairer. On malliavin's proof of hörmander's theorem. *Bulletin des sciences mathematiques*, 135(6-7):650–666, 2011.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[Hör67] Lars Hörmander. Hypoelliptic second order differential operators. *Acta Mathematica*, 119(1):147–171, 1967.

[Hör71] Lars Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1971.

[HRW86] Geoffrey E. Hinton, David E. Rumelhart, and Ronald J. Williams. Learning representations by back-propagating errors. In *Proceedings of the IEEE Conference on Neural Networks*, pages 1–6. IEEE, 1986.

[Jea14] Frédéric Jean. *Control of Nonholonomic Systems: from Sub-Riemannian Geometry to Motion Planning*. Springer Briefs in Mathematics. Springer, Cham, 2014.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[Léo13] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2013.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2304.09876*, 2023.

[LWCC23] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[Ras40] P Rashevsky. The problem of control in a smooth manifold. *Doklady Akademii Nauk SSSR*, 26:567–570, 1940.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[RCK+24] Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9472–9481, 2024.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 2256–2265. PMLR, 2015.

[SSDK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[Sus95] HJ Sussmann. The structure of control systems on sub-riemannian manifolds. *Mathematics of Control, Signals, and Systems*, 8(1):1–47, 1995.

[WHT23] Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. *arXiv preprint arXiv:2306.01798*, 2023.