”Do what you can, with what you have, where you are.”

— Theodore Roosevelt

To my family and friends.

Contents

1 Introduction

In the recent years the Artificial Intelligence (AI) field has experienced an incredible rise in progress. One of the most important (and we should define the notion of importance) algorithms, or most publicized perhaps, has been ChatGPT which being a Large Language Model (LLM) is able to generate data, such as text and images. An algorithm with such an ability is called ”generative” because by ’learning’ the data distribution it is able to produce synthetic data, that is to sample from the learnt probability distribution which is close, in some sense, to the data distribution. Most of the times the probability distribution we want to sample from is high-dimensional and most likely non-log-concave, that is multimodal and not well-behaved. Sampling models are based on the construction of stochastic processes having as steady-state distribution the target one; the very first models we study are the ones evolving in continuous-time, for which there are plenty of articles and books studying their theoretical convergence guarantees. Then, due to the impossibility, in most cases, of simulating a continuous-time process, we apply some discretization schemes, thus introducing a first approximation. Here as well the literature is very flourishing; however, concerning the setting of non-convex sampling, there still is a lot of work to do given the difficulties in obtaining algorithms capable of sampling in acceptable times. Finally, there is the new setting identified by generative algorithms, such as Diffusion Models. The amount of research in this direction is big due to its current academic and private appeal. Notwithstanding, there is a long way to go. The goal of this thesis is to provide a not exhaustive overview of theoretical convergence guarantees for, first, continuous-time processes, second, for classical sampling algorithms and finally, for Diffusion Models. In order to be able to give some convergence guarantees we introduce a set of assumptions about the underlying setting. Then, use these assumptions to derive bounds on the distance/divergence (e.g. 2-wasserstein) between the probability distribution we are effectively sampling from and the data distribution, assuming there exists one. For the second and third cases, these bounds will depend on the approximations introduced during implementation, which are influenced by the model parameters.

But how do these algorithms work? Well, there are several available models and a lot depends on what kind of data we are working with. In the context of image-data, video

and audio-data, the state-of-the-art performances are achieved by the so-called Diffusion Models, that is probabilistic generative models that progressively add noise to the data so that all the structure is lost and consequently learn to reverse this process in order to sample new data (see Figure (1)). The

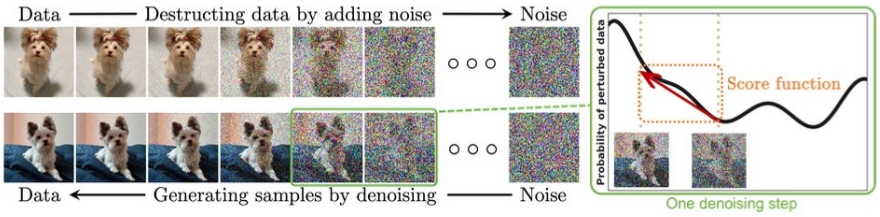

Figure 1: Generation of a new sample (on the left); illustration of score function role in one denoising step (on the right).

The general formulation of Diffusion Models was introduced by [30], where Stochastic Differential Equations (SDEs) are used. In this setting we need to define a forward SDE of the form,

$$\mathrm{d}X_t = f(X_t, t)\mathrm{d}t + g(t)\mathrm{d}B_t, \quad X_0 \sim p_{data}, \tag{1.1}$$

so that the solution to such an equation is a diffusion process $(X_t)_{t\geq 0}$, defined over a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ such that the filtration is complete and the process is adapted to such a filtration and taking values in $\mathbb{R}^d$. Here $f(\cdot,\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is the drift coefficient of the process (the deterministic component) and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is the diffusion coefficient. Then, $(X_t)_{t\geq 0}$ satisfies the integral equation,

$$X_t = X_0 + \int_0^t f(X_s, s)\mathrm{d}s + \int_0^t g(s)\mathrm{d}B_s, \quad \forall t \geq 0. \tag{1.2}$$

In real-world applications we have a dataset of i.i.d. samples from p_{data} to carry out the training procedure (i.e. estimating a sequence of *Stein* scores as we will see in a bit). Usually we design this forward SDE so that it converges to some easy-to-sample parametric distribution (e.g. a standard d-dimensional Gaussian, γ^d). We know that a diffusion process governed by [1.1] is such that its marginal probability distribution, namely $p_t := \mathrm{law}(X_t)$, satisfies the *Kolmogorov's forward equation*, also known as the Fokker-Planck equation, given by,

$$\partial_t p_t = \frac{1}{2} g(t)^2 \Delta p_t - \nabla \cdot (p_t f(X_t, t)) \,. \tag{1.3}$$

Thus the SDE allows us to model a diffusion process at a microscopic level (giving insight about single random trajectories) while the related PDE model the system at a macroscopic level (giving a smooth, deterministic time-evolution of the marginal probability measures). Then, [1] it can be shown that the process defined as $\overleftarrow{X}_t := X_{T-t}$, for $t \in [0, T]$, satisfies the SDE

$$\mathrm{d}\overleftarrow{X}_t = \left\{ -f(\overleftarrow{X}_t, T-t) + g(T-t)^2 \nabla \ln p_{T-t}(\overleftarrow{X}_t) \right\} \mathrm{d}t + g(T-t)\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim p_T, \tag{1.4}$$

where now $(B_t)_{t\in[0,T]}$ is the reverse Brownian Motion (for ease of notation we keep it as the forward one), and $\nabla_{\overleftarrow{x}_t} \ln p_{T-t}$ is the *Stein score* or *score function* and it is a vector field pointing to the direction of data with higher likelihood and less noise. However, the score function is not know given that, in order to compute $p_t, \ \forall t \geq 0$, we should know p_{data}. Thus we need to estimate a sequence of score functions along the time interval $[0, T]$. Following the work of [15] and [33] we can estimate the score functions by training a time-dependent score-based model $s_\theta(X_t, t)$ (parametrized by a deep neural network) so to solve the following minimization problem known as denoising score matching,

$$\min_\theta \mathbb{E}_{t\sim U[0,T]} \left\{ \lambda(t) \mathbb{E}_{X_0 \sim p_{data}} \mathbb{E}_{X_t|X_0 \sim p_{t|0}} \left[\| s_\theta(X_t, t) - \nabla_{x_t} \ln p(X_t|X_0) \|^2 \right] \right\}, \tag{1.5}$$

where $p_{data} \equiv p_0$. It was demonstrated that this minimization problem guarantees that the estimated score functions equals $\nabla_{x_t} \ln p(X_t)$ almost surely.

Now, for the sake of completeness, we will briefly introduce an instance of the above setting, namely the *denoising diffusion probabilistic model* (DDPM), following the idea of [28] and [14]. A DDPM consists in defining two discrete-time Markov chains: the first one, defined forward in time, add noise to the structured data and it is designed such that it converges to a known parametric easy-to-sample prior distribution (e.g. standard Gaussian) while the second one, defined backward in time, reverse this process by learning the transition kernels by means of deep neural networks architectures. Formally, we have our data $X_0 \in \mathbb{R}^d$ which is distributed according to an unknown distribution p_{data}. Then, as we know, in order to define a Markov chain we just need a set of transition kernels so that the joint distribution

has the following form: $p(X_0, ..., X_T) = p_{data}(X_0) \prod_{t=1}^{T} p_{t|t-1}(X_t|X_{t-1})$, $\forall T \in \mathbb{N}$. Thus we design T transition kernels such that our process converge to a stationary standard Gaussian distribution; we define $p_{t|t-1}(X_t|X_{t-1}) = \mathcal{N}\left(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I_d\right)$, with $\beta_t \in (0,1)$ being an hyperparameter. Then, given that $p(X_1, ..., X_T|X_0)$ is Gaussian, each marginal is still Gaussian. Therefore for each $t \in \{1, ..., T\}$, we obtain $p_{t|0}(X_t|X_0) = \mathcal{N}\left(X_t; \sqrt{\overline{\alpha_t}}X_0, (1-\overline{\alpha_t})I_d\right)$, where $\alpha_t := 1 - \beta_t$ and $\overline{\alpha_t} := \prod_{s=0}^{t} \alpha_s$. As $T \to \infty$ we have $p_T(X_T) := \int p(X_T|X_0)p_{data}(X_0)dX_0 \approx \mathcal{N}(X_T; 0, I_d)$, given the choice of the positive noise scales β_t, $t \in \{1, ..., T\}$.

Then, we need to reverse the process through a learnable Markov chain. Here we start from $\gamma^d = \mathcal{N}(X_T; 0, I_d)$; therefore, we are producing a first approximation in our model given that $p_T(X_T) \approx \mathcal{N}(X_T; 0, I_d)$. Then we have a sequence of learnable transition kernels of the form $p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \Sigma_\theta(X_t, t))$, where $\theta \in \Theta$ is the parameters vector of the deep neural network used to estimate the moments of the transition kernels.

In order to train the neural network we need to define a loss function. As in [14] and [36], we can define the following loss,

$$\mathcal{L} = \mathbb{E}_{t\sim U[1,T], X_0\sim p_{data}, \epsilon\sim\mathcal{N}(0,I_d)} \left[\lambda(t)||\epsilon - \epsilon_\theta(X_t, t)||^2\right], \tag{1.6}$$

where $\lambda(t)$ is a positive weighting function, $X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\epsilon$ and ϵ_θ is a deep neural network that predicts the noise vector added to X_0. Equation [1.5] can be reduced to this form as well. As we said when introducing diffusion models based on SDEs, the DDPM model specification can be seen as a discretization of an SDE of the form:

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t,$$

which is also called Variance Preserving (VP) SDE, meaning that the process has bounded variance along the trajectories. Consequently, the reverse SDE is given by

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)\left(X_t - \nabla_{x_t}\log p(X_t)\right)\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\overline{B}_t.$$

Therefore the DDPM formulation is just a discretization of the above SDEs, and the

set of estimated backward transition kernels identifies a numerical SDE solver for the backward SDE.

Thus, how to generate new samples from the target probability distribution? Well, there are several numerical solvers which can approximate SDE trajectories, e.g. the Euler-Maruyama. For example, in the context of DDPM we use ancestral sampling which corresponds to a specific discretization of the backward VP SDE. More generally, the approach used for sampling is the following: we adopt a predictor-corrector (PC) approach, meaning we use a predictor, namely any numerical solver with a fixed discretization strategy (e.g. ancestral sampling), to obtain a first estimate of the next sample and then we adopt a corrector, namely any gradient-based MCMC (e.g. annealed Langevin dynamics), in order to correct the initial sample from the numerical solver so that we manage to sample from the right marginal.

2 Markov Processes

In this chapter we introduce some theory behind Markov processes. Typically, stochastic processes indexed by a continuous variable (e.g., time) are defined on spaces of functions (serving as sample spaces) since, under suitable regularity conditions, one may assume the existence of continuous modifications of the trajectories. Then, a probability measure defined on these functions is also known as a path measure (the most famous being the Wiener measure which, in some sense, is the analogue in infinite dimensions of the Lebesgue measure), and it is such that its finite-dimensional marginals satisfy the necessary consistency conditions to yield a well-defined probabilistic object. In this framework, the Kolmogorov extension theorem guarantees the existence of such a measure. Moreover, due to the fact that stochastic processes are defined on infinite-dimensional spaces, we might want to introduce some analytical tools from functional analysis to rigorously study such mathematical objects. Consequently, we present the Feller semigroup with the corresponding infinitesimal generator. Then, we explore the use of functional inequalities as guarantees for an exponentially fast convergence of the curve of probability measures, defined by the stochastic process dynamics, to the equilibrium distribution. Finally, we will specialize the preceding to the subsection identified by diffusion processes.

2.1 Theoretical Formulation

The theory of continuous-time Markov Processes is fundamental when studying key properties of diffusion processes we will consider due to the fact that the latter are instances of such processes. Let's present some key concepts following mainly the work of [18], [2] and [26].

Let's assume there exists a Markov Process $(X_t)_{t\geq 0}$ defined on an underlying filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F})_{t\geq 0}, \mathbb{P})$, taking values on a measurable space $(E, \mathscr{E})$. Then, a *Markovian transition kernel*, from E into E, is a mapping $P : E \times \mathscr{E} \to [0, 1]$ satisfying the following two properties:

(i) $\forall x \in E$, the mapping $\mathscr{E} \ni A \mapsto P(x, A)$ is a probability measure on $(E, \mathscr{E})$,

(ii) $\forall A \in \mathscr{E}$, the mapping $E \ni x \mapsto P(x, A)$ is $\mathscr{E}$-measurable.

This object, i.e. the Markovian transition kernel, is a generalization of the transition matrix used for Markov Chains, and therefore it evolves the distribution of the process over time given that $\pi_t(A) = \int_E P_t(x, A)\pi_0(\mathrm{d}x)$. Related to it, there is another important mathematical object: the Markov semigroup.

Definition 1. *A Markov Process $(X_t)_{t\geq 0}$ is associated with its own Markov transition Semigroup $\boldsymbol{P} = (P_t)_{t\geq 0}$ which satisfies the following properties:*

(i) $\forall x \in E, P_0(x, \mathrm{d}y) = \delta_x(\mathrm{d}y)$ *meaning P_0 is the identity operator,*

(ii) $\forall s, t \geq 0, A \in \mathscr{E}$*,*

$$P_{t+s}(x, A) = \int_E P_s(x, \mathrm{d}y)P_t(y, A),$$

meaning $P_{s+t} = P_t \circ P_s$ (it is a convolution), which is known as the Chapman-Kolmogorov identity,

(iii) $\forall A \in \mathscr{E}$*, the mapping $(t, x) \mapsto P_t(x, A)$ is measurable with respect to the product sigma-algebra $\mathcal{B}(\mathbb{R}_+) \otimes \mathscr{E}$.*

Then, given $f \in B(E) = \{f : E \to R$ such that f is bounded and measurable$\}$, for which the sup-norm, $||f|| = \sup_{x\in E} |f(x)|$, is considered, we have,

$$\begin{aligned} P_t f(x) &:= \mathbb{E}\left[f(X_t)|X_0 = x\right], \\ &= \int_E f(y)P_t(x, \mathrm{d}y), \\ &= \int_E f(y)p_t(x, y)\mathrm{d}y, \end{aligned}$$

where $p_t(x, y)$ is the density kernel of the probability measure $P_t(x, \mathrm{d}y)$ with respect to the Lebesgue measure. Furthermore, the mapping $B(E) \ni f \mapsto P_t f$ is a contraction of $B(E)$, namely $||P_t f|| \leq ||f||$. It is clear from property (i) of Definition [1], that we have $\lim_{t\downarrow 0} P_t f(x) = x\ (= P_0 x)$. Notice that, if we choose f to be the indicator function, that is $f = \mathbb{1}_A$ with $A \in \mathscr{E}$, the expression of $P_t f(x)$ becomes $P(X_{s+t} \in A|X_r, 0 \leq r \leq s) = P_t(X_s, A)$, highlighting the so called *Markov property*. Consequently, if we let the initial distribution of the process be $\pi_0 := \mathrm{law}(X_0)$, then if we take $0 < t_1 < t_2 < ... < t_p$ and

$A_0, ..., A_p \in \mathscr{E}$, $p \in \mathbb{N}$, we have,

$$P(X_0 \in A_0, X_{t_1} \in A_1, ..., X_{t_p} \in A_p) = \int_{A_0} \pi_0(\mathrm{d}x_0) \int_{A_1} P_{t_1}(x_0, \mathrm{d}x_1) \times \\ \times \int_{A_2} \cdots \int_{A_p} P_{t_p - t_{p-1}}(x_{p-1}, \mathrm{d}x_p),$$

and this also applies to other functions satisfying the same regularity conditions. Therefore, a Markov process with transition semigroup **P** is an $(\mathscr{F}_t)$-adapted process $(X_t)_{t\geq 0}$ with values in E such that for each $s,\ t \geq 0\ s.t.\ s < t$ and $f \in B(E)$, $\mathbb{E}[f(X_{s+t})|\mathscr{F}_s] = P_t f(X_s)$. However, given a semigroup, does a Markov process with that semigroup exist? Is it unique? The first step to prove this is to state a specific version of the famous *Kolmogorov extension theorem*. First, let's define the sample space, which consists, as custom, in a functional space, i.e. $\Omega^* = E^{\mathbb{R}_+}$; thus, $w : \mathbb{R}_+ \to E$, $w \in \Omega^*$. Then, we equip the sample space with the σ-algebra $\mathscr{F}^*$ for which the coordinate mappings $w \mapsto w(t)$ are measurable for $t \in \mathbb{R}_+$. With $C(\mathbb{R}_+)$ we denote the collection of all finite subsets of the positive real line, and for every $U \in C(\mathbb{R}_+)$, let $w_U : \Omega^* \to E^U$ be the mapping associating to each function w its restriction to U. At the same time, given $U,\ V \in C(\mathbb{R}_+)$ with $U \subset V$, then we can write $w_U^V : E^V \to E^U$. Now, we can state the theorem.

Theorem 1. *Let E be a Polish space equipped with its Borel σ-algebra $\mathscr{E}$. For every $U \in C(\mathbb{R}_+)$, let P_U be a probability measure on E^U. Then, assume that the collection $(P_U, U \in C(\mathbb{R}_+))$ is consistent in the following sense: given $U,\ V \in C(\mathbb{R}_+)$, such that $U \subset V$, then P_U is the image of P_V under w_U^V. If this holds true then there exists a unique probability measure P on $(\Omega^*, \mathscr{F}^*)$ such that $w_U(P) = P_U$ for every $U \in C(\mathbb{R}_+)$.*

It should be clear that the theorem states that, given a collection of finite-dimensional distributions, satisfying the consistency condition, then we can construct a probability measure having as finite-dimensional marginals the previous measures. Consequently, we can specialize the preceding theorem to the case of interest: Markov processes.

Corollary 1. *Assume that E is Polish and that $(P_t)_{t\geq 0}$ is a transition semigroup on E. Let π_0 be a probability measure on E. Then there exists a unique probability measure P on Ω^* under which the canonical process $(X_t)_{t\geq 0}$ is a Markov process with transition semigroup $(P_t)_{t\geq 0}$ and $X_0 \sim \pi_0$.*

Notice that the adjective 'canonical' refers to the construction of the process, where $X_t(w) = w(t)$, meaning that the value of the process at time t is just the evaluation, at time t, of the trajectory function. Hence, we showed that given a specific Markovian semigroup, there exists a unique Markov process associated to it.
An important clarification is needed. The Markov semigroup we defined and will work with, even if some properties have not been given yet (e.g. the conservative one), is often referred to as a *Feller semigroup* and the same name holds for the process. This is necessary if we desire to define a process which can have a modification with càdlàg paths, the strong Markov property and an infinitesimal generator, concept that we are about to introduce. Finally, we use the notation $(X_t^x)_{t\geq 0}$ to indicate the process started at $X_0 = x$.
So, going back to the semigroup, one aspect we might be interested in studying concerns its infinitesimal behavior, that is how the semigroup evolves functions over an infinitesimal time interval. Here we will not go into the mathematical details needed in order to precisely defined the functional space on which such a linear operator is defined. Following this consideration, we can define the *infinitesimal generator* $\mathscr{L}$ of the Markov semigroup **P**.

Definition 2. *Given a Markov semigroup $(P_t)_{t\geq 0}$ the infinitesimal generator $\mathscr{L}$ is defined via,*

$$\mathscr{L}f := \lim_{t\downarrow 0} \frac{P_t f - f}{t}, \tag{2.1}$$

for all functions for which the above limit exists

The set containing all such functions is the domain of the generator, that we denote as $\mathcal{D}(\mathscr{L})$. For what concerns here, the infinitesimal generator uniquely identifies the Markov semigroup and therefore the Markov process (a probability measure on the space of either right-continuous or continuous functions having values in E). If f satisfies certain regularity conditions such that the above limit exists, then also $P_t f$ satisfies these conditions; thus we can apply the infinitesimal generator to it as well, obtaining $\mathscr{L}(P_t f) = P_t(\mathscr{L}f)$. Furthermore,

$$P_t f = f + \int_0^t P_s(\mathscr{L}f)ds = f + \int_0^t \mathscr{L}(P_s f)\mathrm{d}s, \tag{2.2}$$

meaning that $\forall x \in E$, the map $t \mapsto P_t f(x)$ is continuously differentiable on $\mathbb{R}_+$ and

satisfies the differential equation $\frac{d}{dt}P_t = \mathscr{L}P_t$, whose solution is $P_t = e^{t\mathscr{L}}$, highlighting the positivity property of **P**. More generally, for all $t > 0$ and $x \in E$, we have that $(P_t f(x))_{t>0,x\in E}$ is a solution to the *Kolmogorov's backward equation*,

$$\frac{\partial}{\partial t}P_t f = \mathscr{L}(P_t f), \quad P_0 f = f, \tag{2.3}$$

which is an abstract heat equation. By arguments from functional analysis we have that the law of X_t, namely π_t, is given by $\pi_t = P_t^*\pi_0$, where P_t^* is the adjoint operator of P_t which form the dual semigroup $(P_t^*)_{t\geq 0}$. Then it can be shown that π_t satisfies the Fokker-Planck equation, given by,

$$\frac{\partial}{\partial t}\pi_t = \mathscr{L}^*\pi_t = P_t^*\mathscr{L}^*\pi_0, \tag{2.4}$$

where $\mathscr{L}^*$ is the adjoint of $\mathscr{L}$ with respect to some reference measure m (e.g. the invariant measure) in the $\mathbb{L}^2(m)$-space, meaning that the following equivalence of inner products is satisfied $\int_E (\mathscr{L}f)g dm = \int_E f(\mathscr{L}^*g)dm$, for $f, g \in \mathcal{D}(\mathscr{L})$.
The Fokker-Planck can be used to obtain the σ-finite measure π on $(E, \mathscr{E})$, known as the stationary/invariant measure of the process, which for every bounded positive measurable function $f : E \to \mathbb{R}$ and $t \geq 0$, satisfies,

$$\int_E P_t f d\pi = \int_E f d\pi, \tag{2.5}$$

meaning that $P_t^*\pi = \pi$, namely π is a fixed point for the adjoint operator, which is unique if the contraction inequality is strict. In fact, π solves $\mathscr{L}^*\pi = 0$. Thus, once the process has converged to its stationary distribution, in the probabilistic language, the distribution of each observation will continue to be π, hence giving a sample of i.d. random variables; consequently, given that the process does not exhibit an evolution of the marginal probability measures anymore, the equality $\mathbb{E}_\pi \mathscr{L}f = 0$ holds true for all functions f in $\mathcal{D}(\mathscr{L})$.
When working with a Markov semigroup $(P_t)_{t\geq 0}$ with infinitesimal generator $\mathscr{L}$ and stationary distribution π, the usual space of functions taken into consideration is the Hilbert space of real-valued functions which are measurable and square-integrable with respect to the invariant measure, namely $\mathbb{L}^2(\pi)$. Then, an important and easier to

study setting is identified by a symmetric Markov semigroup, as given by the following definition.

Definition 3. *The Markov semigroup $(P_t)_{t\geq 0}$ is symmetric, therefore reversible, with respect to the invariant measure π if for all $f, g \in \mathbb{L}^2(\pi)$ and $t \geq 0$ we have,*

$$\int_E (P_t f) g d\pi = \int_E f (P_t g) d\pi. \tag{2.6}$$

The above equality assumes that the process is already at stationarity, namely $X_0 \sim \pi$. A Markov Process being π-reversible is π-stationary and time-reversible; in fact, if we take, for example, two indicator functions, $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$ for measurable sets A, B, then

$$\begin{aligned}
\int_E (P_t f) g d\pi &= \int_B P(X_t \in A | X_0 \in B) d\pi, \\
&= P(X_t \in A, X_0 \in B), \\
&= P(X_0 \in A, X_t \in B), \\
&= \int_E f (P_t g) d\pi.
\end{aligned}$$

When working within the Hilbert space $\mathbb{L}^2(\pi)$ we have that the semigroup as well as its infinitesimal generator are self-adjoint (they have real spectrum), i.e. $P_t = P_t^*$ and $\mathscr{L} = \mathscr{L}^*$. Then it is natural to consider the density $\rho_t := \pi_t / \pi$, which satisfies the Fokker-Planck $\partial_t \rho_t = \mathscr{L} \rho_t$. Then the process will converge, with some velocity, to the stationary measure π, resulting in $\partial_t \rho_t = 0$.

But how fast does the process converge to stationarity? Well, this is quite an intriguing question whose answer is kind of intricate due to its dependence on several factors, such as the geometry of the space as well as the shape of the invariant distribution, e.g. it is easier to converge to a unimodal and strong log-concave distribution such as a Gaussian than to a multimodal distribution. A process, in order to achieve stationarity, must mix well, in the sense that it needs to explore the state space as fast as possible, taking into account the relative 'importance' of regions of the distribution. In mathematical terms, we might define the mixing time as the first time that the distribution of the process is ϵ close to the invariant measure, that is, $t_{\text{mix}}(\epsilon, \pi_0) = \inf\{t \in \mathbb{R} | d_{TV}(\pi_t, \pi) \leq \epsilon\}$, where TV stands for Total Variation, a distance

that will be introduced later on.

We are interested in the behavior of $P_t f$ as $t \to \infty$. We know that the spectral decomposition of $P_t f$ is given by $P_t f = \int_0^\infty e^{-\lambda t} \mathrm{d}E_\lambda f$, for any $f \in \mathbb{L}^2(\pi)$. From this we get that $\lim_{t\to\infty} P_t f = 0,\ \forall \lambda > 0$, namely $P_t f$ converges in $\mathbb{L}^2(\pi)$ to the projection of f on the space of functions satisfying $\mathscr{L} f = 0$, that is to the kernel of the infinitesimal generator. The spectral components with $\lambda > 0$ do not matter anymore, because they vanish exponentially as $t \to \infty$. Hence, we have that, in terms of spectral decomposition, the functions f such that $\mathscr{L} f = 0$, are associated with $\lambda = 0$, meaning that $\lim_{t\to\infty} P_t f = \mathrm{d}E_0 f$, which corresponds to the projection of f onto $Ker(\mathscr{L})$. In order to be able to formally state the previous result, we need to define the concept/property of *ergodicity*, which implies that a process is irreducible, positive recurrent, and non-explosive.

Definition 4. *The infinitesimal generator $\mathscr{L}$ is said to be ergodic if every $f \in \mathcal{D}(\mathscr{L})$ such that $\mathscr{L} f = 0$ is constant. In particular, $f = 0$ if π has infinite mass (e.g. Lebesgue).*

The further requirement for the case where π has infinite mass is due to the fact that $\mathcal{D}(\mathscr{L}) \subset \mathbb{L}^2(\pi)$, so that we need $\int_E f^2 \mathrm{d}\pi < \infty$; when π has infinite mass over E the only way for f to satisfy the above constraint is to be constantly equal to 0. Then, another important property that the semigroup must satisfy is the *mass conservation* one. This is stated in the following definition.

Definition 5. *The semigroup $(P_t)_{t\geq 0}$ is said to be conservative if for every $t \geq 0$, $P_t(\mathbb{1}) = \mathbb{1}$.*

This property is crucial to have a well-defined Markov process. Combining the last two notions, we obtain the following proposition, as in [2].

Proposition 1. *Assume that ergodicity holds. Then it holds:*

- *If $\pi(E) = \infty$ then, for any $f \in \mathbb{L}^2(\pi)$, $\lim_{t\to\infty} P_t f = 0$ in $\mathbb{L}^2(\pi)$.*
- *If $\pi(E) < \infty$ and the semigroup is conservative, for any $f \in \mathbb{L}^2(\pi)$,*

$$\lim_{t\to\infty} P_t f = \int_E f \mathrm{d}\pi, \quad \text{in } \mathbb{L}^2(\pi).$$

An example of a process having $\pi(E) < \infty$ with a conservative semigroup is the Ornstein-Uhlenbeck process, that will play an important role in the diffusion models setting. Its SDE formulation is,

$$\mathrm{d}X_t = \theta(x^* - X_t)\mathrm{d}t + \sigma \mathrm{d}B_t,$$

we have the deterministic mean reversion component, i.e. $\theta(x^* - X_t)\mathrm{d}t$, which pushes the process to its stationary distribution, namely the d-dimensional Gaussian having x^* as mean and $\sigma^2/2\theta$ as variance.
Finally, let us introduce the *Dirichlet energy* whose concept is key in defining Markov processes due to the fact that usually we start by defining this functional and then, by also using the Riesz Representation Theorem, we prove the existence of the corresponding infinitesimal generator. It will appear in the two functional inequalities that we are going to treat in the next sub-section.

Definition 6. *Assume that the Markov semigroup $(P_t)_{t\geq 0}$ is reversible with self-adjoint generator $\mathscr{L}$ and stationary distribution π. Then, the Dirichlet energy associated with $\mathscr{L}$ is the bilinear form,*

$$\mathcal{E}(f,g) := \langle f, (-\mathscr{L})g\rangle_{\mathbb{L}^2(\pi)} = \langle (-\mathscr{L})f, g\rangle_{\mathbb{L}^2(\pi)}, \tag{2.7}$$

for functions f, g in its domain, $\mathcal{D}(\mathcal{E})$.

In particular, we have $\mathcal{E}(f,f) \geq 0,\ \forall f \in \mathcal{D}(\mathcal{E})$, showing that the operator $-\mathscr{L}$ is non-negative (it has non-negative real spectrum). This functional measures the smoothness of suitable functions, weighting for the stationary measure.

2.2 Functional Inequalities

The case we are interested in is the second one, i.e. $\pi(E) < \infty$; thus, we notice that the conditional expected value $P_t f(x) = \mathbb{E}(f(X_t)|X_0 = \cdot)$ converges to the unconditional (stationary) one $\mathbb{E}(f(X_t))$, because stationarity implies that the dynamics of the process do not change its marginal distribution over time. But how fast does the process convergence to its stationary distribution? Well, there are some functional inequalities that help us quantify the velocity with which stochastic processes converge.

2.2.1 Poincaré inequality

The first functional inequality we treat is the *Poincaré* or *Spectral Gap inequality*. It is the simplest, if we are allowed to say so, inequality which quantifies ergodicity, giving results about the convergence of the semigroup, $\mathbf{P} = (P_t)_{t\geq 0} \to \pi$. Given the representation $P_t f(x) = \int_E f(y) p_t(x, \mathrm{d}y),\ t \geq 0, x \in E$, where we can also assume the transition kernel to have density with respect to the Lebesgue measure, convergence means $p_t(x, \mathrm{d}y) \to \mathrm{d}\pi(y),\ x \in E,\ as\ t \to \infty$. Let's first state the Poincaré inequality and then discuss it.

Definition 7. *A Markov Process $(X^x)_{t\geq 0}$ is said to satisfy a Poincaré - Spectral Gap inequality P(C_{PI}) with constant $C_{PI} > 0$, if for all smooth and compactly supported functions $f : E \to \mathbb{R}$ we have,*

$$\mathrm{Var}_\pi f \leq C_{PI} \mathcal{E}(f). \tag{2.8}$$

Then, the least constant $C_{PI} > 0$ such that the above inequality is satisfied is called *Poincaré constant*.

Suppose that f is an eigenfunction of the positive operator $-\mathscr{L}$, so that $-\mathscr{L} f = \lambda f$ for some eigenvalue $\lambda > 0$. Applying the Spectral Gap inequality to such a function shows that, if $\lambda \neq 0$ then $C_{PI}\lambda \geq 1$ or $\lambda \geq 1/C_{PI}$. In fact, the expected value of f is such that $\int_E f \mathrm{d}\pi = -\frac{1}{\lambda}\int_E \mathscr{L} f \mathrm{d}\pi = 0$ by invariance, giving us the following P(C_{PI}): $\int_E f^2 \mathrm{d}\pi \leq C_{PI}\mathcal{E}(f) = C_{PI}\int_E f(-\mathscr{L} f)\mathrm{d}\pi = C_{PI}\lambda \int_E f^2 \mathrm{d}\pi$. Therefore, under P($C_{PI}$), every non-zero eigenvalue of $-\mathscr{L}$ is greater or equal to $1/C_{PI}$. This holds true also when the spectrum is not just discrete, meaning that the spectrum of $-\mathscr{L}$ will be $\{0\} \cup [1/C_{PI}, +\infty)$, so that P($C_{PI}$) describes a gap in its spectrum, and the higher the first $\lambda > 0$ the faster the convergence of the process, as it will be clear by reading the following theorem.

The main consequence of the Poincaré inequality is that the Markov semigroup **P** converges exponentially to equilibrium, in the $\mathbb{L}^2(\pi)$- sense, as stated by the following theorem.

Theorem 2. *The following are equivalent.*

1. *The Markov process satisfies a Poincaré inequality with constant C_{PI}.*

2. *For all* $f \in \mathbb{L}^2(\pi)$ *with* $\int_E f \mathrm{d}\pi = 0$ *and all* $t \geq 0$*,*

$$\|P_t f\|^2_{\mathbb{L}^2(\pi)} \leq e^{-\frac{2t}{C_{PI}}} \|f\|^2_{\mathbb{L}^2(\pi)}. \tag{2.9}$$

The assumption about the stationary mean of f is just for simplicity. If we did not make this assumption, then the inequality would be,

$$\mathrm{Var}_\pi(P_t f) \leq e^{-2t/C_{PI}} Var_\pi(f), \ \ \forall f \in \mathbb{L}^2(\pi), t \geq 0,$$

which shows, even more clearly, the exponential decay in variance as the process evolves over time, given the assumption that a Poincaré inequality holds.
In order to use such results for the convergence of the entire distribution, namely of $\pi_t = \text{law}(X_t)$ to π, then we can just consider $f = \rho_0 - 1$ in the above setting, where $\rho_t := \pi_t/\pi$ and, as we already saw, $\rho_t = P_t \rho_0$ by the self-adjoint property of the semigroup. Hence, we obtain the so-called *chi-squared divergence*, which is not a metric as the name suggests; it has the following expression: $\chi^2(\pi_t \| \pi) := \|\mathrm{d}\pi_t/\mathrm{d}\pi - 1\|^2_{\mathbb{L}^2(\pi)} = var_\pi \mathrm{d}\pi_t/\mathrm{d}\pi$ if $\pi_t \ll \pi$, $\chi^2(\pi_t \| \pi) := \infty$ otherwise . Then, we have the following theorem.

Theorem 3. *The following are equivalent.*

1. *The Markov process satisfies a Poincaré inequality with constant* C_{PI}*.*

2. *For any initial distribution* $\pi_0 = \text{law}(X_0)$ *and all* $t \geq 0$*,*

$$\chi^2(\pi_t \| \pi) \leq e^{-\frac{2t}{C_{PI}}} \chi^2(\pi_0 \| \pi).$$

Therefore, given an initial $\mathbb{L}^2$-distance between radon-nikodym derivatives, we have that this distance decrease exponentially fast as a function of time. At the same time, the lower the Poincaré constant C_{PI} (i.e. the higher the first eigenvalue greater than zero of $-\mathscr{L}$), the faster the convergence, as can also be understood by looking at the spectral decomposition.

2.2.2 Log-Sobolev inequality

Another useful functional inequality which is frequently used is the so-called *log-Sobolev inequality (LSI)*, which contains more information than the Poincaré inequality due to

the fact that it does provide an upper bound on the entropy of a density function and not on its variance. Furthermore, differently from what is required for the Poincaré inequality to hold, here the diffusion property is necessary.

Given a measure space $(E, \mathscr{E}, \nu)$ with ν possibly infinite, we define for all positive and integrable functions f such that $\int_E f|\log f|d\nu < \infty$ the entropy functional,

$$\mathcal{E}nt_\nu(f) = \int_E f \log f \mathsf{d}\nu - \int_E f \mathsf{d}\nu \log\big(\int_E f \mathsf{d}\nu\big).$$

When ν is a probability measure then we have $\mathcal{E}nt_\nu(f) \geq 0$ and equality holds only if f is constant. In fact, if ν is a probability measure and f is not a density, then the above entropy functional corresponds to the Kullback-Leibler (KL) divergence between $\tilde{\nu}(\mathsf{d}x) = \dfrac{f(x)}{\int_E f \mathsf{d}\nu}\nu(\mathsf{d}x)$ and ν. Now, we can define the log-Sobolev inequality.

Definition 8. *A Markov Process $(X^x)_{t\geq 0}$ is said to satisfy a tight log-Sobolev inequality with constant $C_{LSI} > 0$ if for all smooth and compactly supported functions $f : E \to \mathbb{R}$ we have,*

$$\mathcal{E}nt_\pi(f^2) \leq 2C_{LSI}\mathcal{E}(f). \tag{2.10}$$

Given the properties of the carré du champ operator we can rewrite the LSI as $\mathcal{E}nt_\pi(f) \leq \frac{C_{LSI}}{2}\int_E \dfrac{\Gamma(f)}{f}\mathsf{d}\pi$. When the function f represents a density with respect to the probability measure π then the entropy functional of f is the KL divergence between $\mathsf{d}\nu = f\mathsf{d}\pi$ and π, namely,

$$\mathrm{KL}(\nu\|\pi) = \mathcal{E}nt_\pi(f) = \int_E \log f \mathsf{d}\nu.$$

At the same time, we can define the *Fisher information* of ν with respect to π as

$$\mathrm{I}(\nu|\pi) := \int_E \frac{\Gamma(f)}{f}\mathsf{d}\pi,$$

which for $E = \mathbb{R}^d$ specializes to $\mathrm{I}(\nu|\pi) = \int_{\mathbb{R}^d} f|\nabla \log f|^2 \mathsf{d}\pi$. Therefore, a LSI ends up being the following,

$$\mathrm{KL}(\nu\|\pi) \leq \frac{C_{LSI}}{2}\,\mathrm{I}(\nu|\pi).$$

Just as the Poincaré inequality implies an exponential decay in variance, the Log-

Sobolev inequality implies an exponential decay of entropy, as stated by the following theorem.

Theorem 4. *The log-Sobolev inequality with constant C_{LSI} for the probability measure π is equivalent to saying that for every positive function $f \in L^1(\pi)$ such that $\int_E f|\log f|\mathrm{d}\pi < \infty$,*

$$\mathcal{E}nt_\pi(P_t f) \leq e^{-2t/C_{LSI}}\mathcal{E}nt_\pi(f), \quad \forall t \geq 0. \tag{2.11}$$

This expression can be made clearer if we consider, once again, the KL divergence formulation. In fact, in the context of Markov processes we want to establish some kind of inequality which assures us about the speed of convergence of such a process. As we consider $\mathrm{d}\pi_t = P_t f \mathrm{d}\pi$, $t \geq 0$, that is $f = \mathrm{d}\pi_0/\mathrm{d}\pi$, the entropy decay equation becomes,

$$\mathrm{KL}(\pi_t\|\pi) \leq e^{-2t/C_{LSI}}\,\mathrm{KL}(\pi_0\|\pi), \tag{2.12}$$

meaning that as the time goes on the process exhibits an exponential decay of the KL divergence between the probability measure of the process at time t, π_t, and the equilibrium one, π. The fact that relative entropy decays along the process trajectories is validated by noticing that the time derivative of the relative entropy is the opposite of the relative Fisher information; given that $\mathrm{I}_\pi(\cdot) \geq 0$, it is always true that along the sample paths the relative entropy decreases. We can formalize what we said by stating the following proposition, known as *de Bruijn's identity*.

Proposition 2. *For every positive f which is smooth and compactly supported, it holds,*

$$\frac{d}{dt}\mathcal{E}nt_\pi(P_t f) \leq -\,\mathrm{I}_\pi(P_t f). \tag{2.13}$$

Furthermore, convergence in KL divergence implies convergence in total variation. Let's recall its definition.

Definition 9. *The total variation (TV) distance between π_t and π is given by,*

$$\|\pi_t - \pi\|_{TV} := \sup_{A\in\mathscr{F}} |\pi_t(A) - \pi(A)|, \tag{2.14}$$

which becomes $\|\pi_t - \pi\|_{TV} = \frac{1}{2}\int |\pi_t(x) - \pi(x)|\mathrm{d}x$, assuming π_t and π are dominated by the Lebesgue measure (we use π_t and π to indicate the densities as well as the

probability measure). It turns out that the TV distance satisfies the *Pinsker-Csizsár-Kullback inequality*, which reads,

$$\|\pi_t - \pi\|_{TV}^2 \leq \frac{1}{2} \mathrm{KL}(\pi_t \| \pi). \tag{2.15}$$

Thus, if we control $\mathrm{KL}(\pi_0 \| \pi)$, then the previous theorem implies the stronger convergence of the process toward equilibrium in total variation.
What about the properties the distributions need to satisfy in order to satisfy the two functional inequalities we talked about? Well, if the distribution is strongly long-concave then it satisfies both inequalities, as stated by the following Lemma.

Lemma 1. *Let π be a distribution on $E = \mathbb{R}^d$. Then it holds the following.*

1. *(Bakry-Emery-theorem) If π is α-strongly log-concave, then it satisfies a log-Sobolev inequality with constant $C_{LSI} \leq 1/\alpha$.*

2. *If π satisfies a log-Sobolev inequality with constant C_{LSI}, then it also satisfies a Poincaré inequality with constant $C_{PI} \leq C_{LSI}$.*

Therefore, if a distribution is α-strongly log-concave then the Markov process having it as equilibrium distribution will converge to it exponentially fast, given what we discussed about the Poincaré and log-Sobolev inequalities. However, it is important to remember that for the LSI we need the initial distribution to be dominated by the invariant meausure, while for the PI we need the initial distribution to be square integrable with respect to π. An example of α-strongly log-concave measure is the Gaussian measure, where $\alpha = 1/\sigma^2$; the more concentrated the Gaussian is (the lower the variance), the higher the α, the lower the LSI and PI constants, and, finally, the faster the convergence to equilibrium.
Up to now we have presented results about the convergence of a continuous-time process with respect to two divergences and stated that convergence in relative entropy implies convergence in total variation distance; however, total variation distance does not consider the metric structure and the geometry of the space of probability measures, as it only measures the maximum possible difference in probability mass over all measurable sets without considering how far such a mass has to be moved in order to transform one into the other. Consequently, we might want to assess convergence

results in some metric having the just mentioned properties. The ideal candidate is the 2-Wasserstein distance. Given the Polish space $(\mathbb{R}^d, \|\cdot\|_2)$, we denote the space of measures over the d-dimensional Euclidean space having finite second moment by $\mathcal{P}_2(\mathbb{R}^d)$. Then, for any measures $p_1, p_2 \in \mathcal{P}_2(\mathbb{R}^d)$, we let $\mathcal{C}(p_1, p_2)$ be the set of couplings between the two such that if $(X, Y) \sim q$, with $q \in \mathcal{C}(p_1, p_2)$, then $X \sim p_1$ and $Y \sim p_2$. Finally, we have the following definition.

Definition 10. *Given two probability measures $p_1, p_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance between p_1 and p_2 is,*

$$\mathcal{W}_2(p_1, p_2)^2 := \inf_{q \in \mathcal{C}(p_1, p_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 dq(x, y). \tag{2.16}$$

Due to the work of [22], we can establish the so-called *Talagrand's (T_2) inequality*, that is, a transportation inequality comparing the $\mathcal{W}_2$-metric from optimal transport with the relative entropy, highlighting the contraction property of the $\mathcal{W}_2$-metric along the Fokker-Planck trajectory.

Lemma 2. *Otto-Villani Theorem Suppose that π is a distribution on $\mathbb{R}^d$ satisfying an LSI with constant C_{LSI}. Then, for all distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2(\mu, \pi) \leq 2C_{LSI}\,\mathrm{KL}(\mu\|\pi). \tag{2.17}$$

Thus, along the diffusion path the $\mathcal{W}_2$-metric is controlled by the relative entropy, if the stationary distribution satisfies a log-Sobolev inequality. In this way we can guarantee an exponential convergence to equilibrium also under a more structured metric. There exists an equivalent inequality when π satisfies the weaker condition imposed by a Poincaré inequality; we just have to replace the KL divergence with the chi-squared one and C_{LSI} with C_{PI}.

2.3 Diffusion Processes

One instance of Markov processes is given by the family of diffusion processes which are an important modeling tool in several fields. Some of the theory is treated in [25], [24], [17] and [23].

2.3.1 Stochastic Differential Equations

Diffusion processes are solutions to stochastic differential equations, where the random component is typically modeled by the Brownian Motion (BM) or, more generally, by stochastic integrals with respect to Brownian Motion. As a matter of fact, Brownian Motion is the first and most fundamental diffusion process ever studied. In 1817, the Scottish botanist Robert Brown observed under a microscope the "rapid oscillatory motion" of microscopic particles within pollen grains suspended in water. The mathematical formulation of this phenomenon was first provided by A. Einstein and later refined by N. Wiener, establishing Brownian Motion as the most famous stochastic process characterized by independent and normally distributed increments, which are also stationary. The Brownian Motion $(B_t)_{t\geq}$ in $\mathbb{R}^d$ consists of d independent standard real Brownian Motions. Assuming $B_0 = 0$ then $B_t \sim \mathcal{N}(0, tI_d)$. Its infinitesimal generator is $\mathscr{L} = \frac{1}{2}\Delta$, but usually we consider the time rescaling version $(\tilde{B}_t)_{t\geq 0} = (B_t)_{t\geq 0}$ so that the generator is the laplacian. The invariant and reversible measure of such a process is the Lebesgue one (dx) given that as $t \to \infty$ the Gaussian distribution flattens out, so that the dynamics visits the whole state-space; this makes the Brownian Motion not ergodic. The associated Markov (Brownian) semigroup $(P_t)_{t\geq 0}$ admits kernel densities with respect to the Lebesgue, $p_t(x, y) = (4\pi t)^{-d/2} \exp(-|x - y|^2/4t),\ t > 0,\ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. The kernels satisfy the parabolic heat equation $\partial_t p_t = \Delta p_t$; this is why the Brownian semigroup is also called Heat semigroup.

Informally, the term *diffusion* is associated to Markov processes having continuous trajectories and that can be characterized through their infinitesimal generators. As a matter of fact, there are two ways of studying diffusion processes. The first approach is the probabilistic one, where we work with a SDE and consider each of its possible sample paths; the second approach is the analytical one, where we consider the Fokker-Planck PDE and its description of the density evolution. Actually, the two methods are not mutually independent and are linked by the Markov semigroup and its infinitesimal generator.

The general formulation of a stochastic differential equation is given by,

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t, \quad X_0 = x, \tag{2.18}$$

where $X_t, B_t, b(\cdot) \in \mathbb{R}^d$ and $\sigma(\cdot) \in \mathbb{R}^{d\times d}$. As standard practice, we can not divide by dt due to the non-differentiability of the Brownian Motion. The most proper notation for such a dynamics is given by the stochastic integral equation,

$$X_t = X_0 + \int_0^t b(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}B_s, \quad t \geq 0. \tag{2.19}$$

An important issue is related with existence and uniqueness of a stochastic process solving such equations. For what concerns existence we must require a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ on which we define a d-dimensional Brownian Motion $(B_t)_{t\geq 0}$. On the other hand, uniqueness might be a more delicate issue. We say that $(X_t^x)_{t\geq 0}$ is a solution to (2.18) if it satisfies the equation path-wise, namely for almost every $\omega \in \Omega$. Then uniqueness implies that, for any initial value x, two solutions X and X' of (2.18) are such that the maps $t \mapsto X_t$ and $t \mapsto X'_t$ are equal almost everywhere. The following theorem formalizes the existence and uniqueness of solutions to (2.18).

Theorem 5. *Assume that the maps* $\mathbb{R}^d \ni x \mapsto \sigma(x)$ *and* $\mathbb{R}^d \ni x \mapsto b(x)$ *have a linear growth,*

$$\|\sigma(x)\| + \|b(x)\| \leq D(1 + \|x\|), \tag{2.20}$$

for some finite constant D and all $x \in \mathbb{R}^d$ *and are locally Lipschitz, meaning that for any compact set* $K \subset \mathbb{R}^d$*, there exists a constant* $C_K > 0$ *such that,*

$$\|\sigma(x) - \sigma(y)\| + \|b(x) - b(y)\| \leq C_K\|x - y\|, \tag{2.21}$$

for all $x, y \in K$*. Then, there exists a unique solution* $(X_t^x)_{t\geq 0}$ *to* (2.18) *defined on the time interval* $[0, \infty)$*.*

Then the unique solution to (2.18), i.e. $(X_t^x)_{t\geq 0}$, will be a Markov process with continuous trajectories, adapted to the filtration $(\mathscr{F}_t)_{t\geq 0}$ and will satisfy certain properties such as the strong Markov property. Moreover it is associated with its own Markov semigroup $(P_t)_{t\geq 0}$ defined by $P_t f(x) = \mathbb{E}[f(X_t^x)]$. Furthermore, the infinitesimal generator is given by the second order operator,

$$\mathscr{L} f = \frac{1}{2}\sum_{i,j=1}^{d} (\sigma\sigma^{\mathrm{T}})_{ij}\partial_{ij}^2 f + \sum_{i=1}^{d} b_i \partial_i f. \tag{2.22}$$

A natural question, then, is to ask whether the process $(f(X_t^x))_{t\geq 0}$ satifies a SDE as well. The answer is yes and it is given by the famous Itô's formula.

Theorem 6. *Let us consider the process $(f(X_t^x))_{t\geq 0}$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function, at least belonging to $\mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$, and $(X_t^x)_{t\geq 0}$ is a solution to* (2.18)*. Then, we have,*

$$\mathbf{d}f(X_t) = \sum_{i,j=1}^{d} \sigma_i^j(X_t)\partial_i f(X_t)\mathbf{d}B_t^i + \mathscr{L}f(X_t)\mathbf{d}t, \tag{2.23}$$

$$= \mathbf{d}M_t^f + \mathscr{L}f(X_t)\mathbf{d}t, \qquad X_0 = x, \tag{2.24}$$

where the operator $\mathscr{L}$ is as in (2.22) *and $(M_t^f)_{t\geq 0}$ is a local martingale.*

It is then straightforward to understand that the generator models the infinitesimal deterministic evolution of the expected value of a smooth function f applied to the process, given the present state. The stochastic integral equation corresponding to (2.24) reads,

$$f(X_t^x) = M_t^f + \int_0^t \mathscr{L}f(X_s^x)\mathrm{d}s, \;\; t \geq 0. \tag{2.25}$$

This shows that diffusion processes solving SDEs are always *semi-martingales*. Assuming the local martingale to be a true martingale, we have, for every $t \geq 0$,

$$\mathbb{E}[f(X_t^x)] = f(x) + \frac{1}{2}\int_0^t \mathbb{E}[\mathscr{L}f(X_s^x)]\mathrm{d}s, \tag{2.26}$$

where the first term is given by the martingale equality condition, i.e. $\mathbb{E}[M_t^f|\mathscr{F}_s] = M_s^f$. Equation (2.26) tells us that the process $(X_t^x)_{t\geq 0}$ is a Diffusion process with generator $\frac{1}{2}\mathscr{L}$. In fact, setting $P_t f(x) = \mathbb{E}[f(X_t^x)] = \mathbb{E}[f(X_t)|X_0 = x], \;\; \forall t \geq 0$, it holds that,

$$P_t f(x) = f(x) + \frac{1}{2}\int_0^t P_s\mathscr{L}f(x)\mathrm{d}s, \;\; t \geq 0. \tag{2.27}$$

This coincides with equation (2.2), derived previously. And as we already highlighted, $P_t f$ is the solution of the heat equation $\partial_t P_t f = \frac{1}{2}\mathscr{L}P_t f$ starting from f.

2.3.2 Langevin Diffusion

A key role in sampling theory is played by the so-called *Overdamped Langevin diffusion*, which is a general diffusion process used in physics, for example, to model the dynamics of particles undergoing diffusion due to both deterministic (here identified by the potential field) and stochastic (a simple stochastic integral with respect to the Brownian Motion) forces. The Langevin diffusion, together with its associated Fokker-Planck equation, describe a physical system from both a microscopic and a macroscopic point of view.

Definition 11. *A stochastic process $(X_t)_{t\geq 0}$ is a Langevin diffusion with potential field $V \in \mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$, if it solves the following stochastic differential equation (SDE),*

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 \sim \pi_0, \tag{2.28}$$

where $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian Motion.

Does the Langevin SDE have a unique strong solution? Well, as a result of theorem (5), if ∇V is Lipschitz, that is V is smooth, then there is a unique strong solution, namely, given a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F})_{t\geq 0}, \mathbb{P})$ and a d-dimensional standard Brownian Motion defined on it then there is a unique adapted process $(X_t)_{t\geq 0}$ having a modification with continuous trajectories and satisfying the integral equation $X_t = X_0 - \int_0^t \nabla V(X_s)\mathrm{d}s + \sqrt{2}B_t, \ \ \forall t \geq 0.$
The drift term of the Langevin SDE is a pure gradient flow,

$$\dot{X}_t = -\nabla V(X_t),$$

which is an ODE targeting a minimizer of f. Consequently, Langevin diffusion can be interpreted as "a gradient flow + noise". However, such a dynamics converges in law to the stationary distribution of the process, $\pi \propto e^{-V(x)}$, which is unique and associated to the following *Fokker-Planck* equation,

$$\frac{\partial}{\partial t}\pi_t = \Delta\pi_t + \nabla\cdot(\pi_t \nabla V(x)), \quad X_0 \sim \pi_0, \tag{2.29}$$

$$= \mathscr{L}^*\pi_t. \tag{2.30}$$

Notice that the infinitesimal generator for the Langevin diffusion is $\mathscr{L} = \Delta - \nabla V \cdot \nabla$ and $\mathscr{L} = \mathscr{L}^*$ only on the space $\mathbb{L}^2(\pi)$. Note that $-\nabla V = \nabla \log \pi$, so that the process uses information about the the shape/geometry of the distribution to move toward higher density regions of π. Then, the noise term is essential to prevent the process from getting stuck in modes and to ensure proper exploration of the entire support of π.

As we said, Langevin diffusion is a general family of stochastic processes; if, for example, we set $V = \frac{1}{2}||\cdot||^2$ then we would have the *Ornstein–Uhlenbeck (OU)* process having as stationary measure the d-dimensional standard Gaussian, γ^d.

Going back to the optimization-sampling connection, [16] showed that if we consider the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, where $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability measures with finite second moment defined over the d-dimensional Euclidean space and $\mathcal{W}_2$ is the 2-Wasserstein distance, then the Fokker-Planck dynamics, namely, the curve of measures $t \mapsto \pi_t$, can be interpreted as the gradient flow of the *Kullback–Leibler divergence*, $\mathrm{KL}(\cdot||\pi)$, with respect to 2-Wasserstein distance. Therefore, sampling consists in an optimization problem in a suitable space of measures; in fact, Langevin dynamics is the steepest descent flow that transports any initial distribution of the system to the invariant measure defined by the potential field.

Previously, we introduced two functional inequalities, namely the Poincaré inequality and the log-Sobolev one, in order to characterize properties satisfied by the stationary measure which guarantee an exponential convergence of the curve $t \mapsto \pi_t$ to π. For a Langevin diffusion $(X_t)_{t\geq 0}$ defined on $\mathbb{R}^d$ the PI reads,

$$\mathrm{Var}_\pi f \leq C_{PI}\mathbb{E}_\pi \left[\|\nabla f\|^2\right], \tag{2.31}$$

for all functions $f : \mathbb{R}^d \to \mathbb{R}$. For $E = \mathbb{R}^d$ it is even more explicit the fact that the Dirichlet form measures the smoothness of a function. Instead, the LSI reads,

$$\mathcal{E}nt_\pi(f^2) \leq 2C_{LSI}\mathbb{E}_\pi \left[\|\nabla f\|^2\right], \quad \text{or equivalently} \tag{2.32}$$

$$\mathsf{KL}(\pi_t\|\pi) \leq 2C_{LSI}\mathbb{E}_\pi \left[\|\nabla\sqrt{\rho_t}\|^2\right], \tag{2.33}$$

where we recall that $\rho_t = \mathrm{d}\pi_t/\mathrm{d}\pi$ is the Radon-Nikodym derivative of $\pi_t = \pi_0 P_t$ with respect to π. Notice that **P** is self-adjoint due to the reversibility property of the Langevin

diffusion with respect to its stationary distribution.

The class of measures satisfying LSI and/or PI is quite large, including all strongly log-concave measures (due the Bakry–Emery criterion) and, for the PI, all log-concave measures. In particular, LSI and Poincaré inequality are preserved under bounded perturbation and Lipschitz mapping whereas logconcavity would be destroyed. Given these properties, it is easy to exhibit examples of non-logconcave distributions satisfying LSI or Poincaré inequality; one possible example is obtained by taking a convex body and apply small perturbations to it so that it is not convex anymore but still satisfies isoperimetry. Take a look at Figure (2), taken from [32].

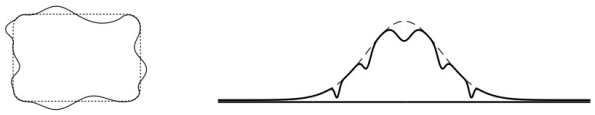

Figure 2: Illustrations of non-logconcave distributions satisfying LSI or Poincar´e inequality: the uniform distribution on a nonconvex set (left), and a small perturbation of a logconcave distribution, e.g., Gaussian (right)).

Consider one last observation. Processes in continuous-time will be discretized in order to be applicable, as we will see in the next chapter. The way in which the continuous-time result affects its discretization counterpart operates through two channels. In fact, if the potential V is α-strongly convex then, if we consider two different Langevin diffusions, so that the corresponding Fokker-Planck describe the evolution of the maps $t \mapsto \nu_t$ and $t \mapsto \pi_t$ with starting point ν_0 and π_0, then we have a contraction of the $\mathcal{W}_2$ distance between the two processes, $\mathcal{W}_2^2(\nu_t, \pi_t) \leq e^{-2\alpha t}\mathcal{W}_2^2(\nu_0, \pi_0)$. At the same time, if the equilibrium distribution, π, of the Langevin diffusion satisfies a LSI with constant $C_{LSI} = 1/\alpha$, then along the dynamics $t \mapsto \pi_t$ we have the exponential decrease in relative entropy, $\mathsf{KL}(\pi_t\|\pi) \leq e^{-2\alpha t}\mathsf{KL}(\pi_0\|\pi)$. Finally, if π satisfies the PI then the previous result holds for the chi-square divergence. Then, these results affect the discretization analysis because at initialization we typically have, for the first two cases, $\mathcal{W}_2^2(\pi_0, \pi), \mathsf{KL}(\pi_t\|\pi) = O(d)$ so that the Langevin dynamics reaches a ϵ-error in $O(\log(d/\epsilon))$ time. For the PI, instead, we have $\chi^2(\pi_0\|\pi) = \exp(O(d))$ so that it takes $O(d \vee \log(1/\epsilon))$ time to reach the same amount of error. Secondly, if the result in continuous-time is given using a specific metric/divergence then it is likely to have the same metric/divergence for the discretization analysis.

3 Log-Concave Sampling

A crucial task in probabilistic modeling is to be able to sample from some target distribution in order to, for example, approximate quantities of interest, e.g. approximate the expected value of a function of the distribution with Monte Carlo methods; we can think of the bayesian context where we would like to sample from some posterior distribution $p(\theta|X_{1:n}) \propto p(\theta)p(X_{1:n}|\theta)$ for which the normalizing constant, $p(X_{1:n}) = \int_\Theta p(\theta)p(X_{1:n}|\theta)d\theta$, is difficult to compute numerically due to the high-dimensional integral ($\theta \in \Theta \subseteq \mathbb{R}^d$ with d being a big number). Hence, this line of research is key to all those fields using probabilistic modeling (e.g. physics, machine learning, finance). We might say that sampling is a very important and fascinating field of research due to its fundamental importance in statistical applications given that without good samples most analyses cannot be conducted. The adjective 'good', referred to 'samples', means that sometimes we are not able to precisely sample from the target distribution but from an approximation of it; well, in order to have good samples, then we need to reduce, at an acceptable level, the distance/divergence between the target probability distribution and the one from which we effectively sample from.

Thus, the general goal is to be able to generate random variables whose law is the closest possible (e.g. in relative entropy rather than in 2-wasserstein distance or Total Variation) to $\pi \propto e^{-V(x)}$. It is crucial to come up with fast and precise models/samplers. In this thesis two settings will be presented. The first one concerns situations in which we know π up to a normalization constant and we assume to have a first-order oracle for π, meaning that given a query point $x \in \mathbb{R}^d$, the oracle returns $V(x) - V(0)$ and $\nabla V(x)$. The second setting is the one characterizing Diffusion Models, where we do not know the analytical expression of π, including any type of information about it, but we have a collection of i.i.d. samples from it which can be used to estimate the gradient appearing in the time-reversal SDE.

Given the interpretation of Langevin diffusion (which appears in different forms in several sampling algorithms) as a gradient flow in a suitable space of measures, it follows that there exists a deep connection between sampling and optimization techniques. That is why many ideas and results in sampling theory are taken from and have a corresponding in the optimization field.

3.1 Langevin Monte Carlo

As shown in the previous chapter, when we are interested in sampling from a distribution, π, we can check whether this distribution satisfies some functional isoperimetric inequalities and if so, we can design a Langevin process converging to such a distribution of interest. In that case, we can obtain samples from the target probability measure quite fast; in fact, if π satisfies a Poincaré inequality then the continuous-time dynamics converges exponentially fast in chi-squared divergence ($\mathbb{L}^2(\pi)$), while if it satisfies a log-Sobolev inequality then it converges exponentially fast in KL divergence. However, in practice, in order to implement the Langevin diffusion,

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

where as usual $X_t, B_t \in \mathbb{R}^d$, we need to discretize it. The most famous and basic discretization method is the *Euler-Maruyama* discretization which is known as *Langevin Monte Carlo (LMC)* or *Unadjusted Langevin Algorithm (ULA)*. The equation reads,

$$X_{(k+1)h} := X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}), \tag{3.1}$$

assuming constant step-size $h > 0$. Naturally, as $h \downarrow 0$ we obtain the corresponding continuous-time dynamics. Let p_{kh} be the probability distribution of the k-th step of the ULA, that is $X_k \sim p_{kh}$. It turns out that the stationary distribution of the ULA, p_∞, is biased, that is, $p_\infty \neq \pi$. Consequently, differently from what happens for the Langevin diffusion, the relative entropy $KL(p_k\|\pi)$ does not tend to 0 along the ULA trajectory because of the bias term $KL(p_\infty\|\pi) > 0$. Therefore, when implementing such an algorithm we are approximately sampling from π; here is why we would like to obtain some non-asymptotic bounds for the distance (usually the 2-Wasserstein one) between the algorithm distribution and the target one. However, there are several settings in which we can apply the ULA; this means that based on the setting assumptions we will obtain different non-asymptotic bounds for the convergence. The algorithm works as follows.

1. Initialize $X_0 \sim p_0$; choose N and set $k = 0$

2. For $k = 0, ..., N-1$:

$$\text{sample } Z_{(k+1)h} \sim \mathcal{N}(0, 2hI_d)$$
$$\text{set } X_{(k+1)h} \leftarrow X_{kh} - h\nabla V(X_{kh}) + Z_{(k+1)h}$$

3. Output $(X_h, ..., X_{Nh})$.

Then, the output is a simulation of a trajectory from an approximation to a π-invariant process. This means that, being the ULA not ergodic, we can not use statistics of the sample to estimate quantities of interest, e.g. the temporal mean for the expected value.

Finally, in order to implement the ULA, we need to be able to sample from a Gaussian which we can easily do and to compute the gradient of the potential field for which we assumed to have an oracle.

3.1.1 Wasserstein Coupling

One of the first papers to introduce the similarities between gradient descent and the ULA was [10], where he provides some theoretical convergence guarantees for the Langevin Monte Carlo method in the usual $\mathcal{W}_2$ metric, assuming smooth and strongly log-concave distributions. A nice analysis of ULA, given the preceding assumptions, is based on coupling together the continuous-time process with the corresponding discretization. Then, the coupling is used to measure the approximation error accumulated along the trajectories.

Theorem 7. *For $k \in \mathbb{N}$, let p_{kh} be the probability distribution of the k-th iteration of the ULA with step size $h > 0$. Assume that the target $\pi \propto e^{-V}$ is α-strongly log-concave and β-smooth, i.e. $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Then, assuming $h \lesssim \frac{1}{\beta\kappa}$, for all $N \in \mathbb{N}$,*

$$\mathcal{W}_2(p_{Nh}, \pi) \leq e^{-\frac{\alpha Nh}{2}} \mathcal{W}_2(p_0, \pi) + O\left(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\right). \tag{3.2}$$

If we set $h \asymp \frac{\epsilon^2}{\beta\kappa d}$, then for any $\epsilon \in [0, \sqrt{d}]$ we obtain $\sqrt{\alpha}\mathcal{W}_2(p_{Nh}, \pi) \leq \epsilon$ after

$$N = O\left(\frac{\kappa^2 d}{\epsilon^2} \log \frac{\sqrt{\alpha}\mathcal{W}_2(p_0, \pi)}{\epsilon}\right) \quad \textit{iterations.}$$

The term $\kappa := \beta/\alpha$ is the condition number and $h \asymp \frac{\epsilon^2}{\beta\kappa d}$ means that there exists a positive constant C such that $h \approx C\frac{\epsilon^2}{\beta\kappa d}$. Equation (3.2) shows that along the dynamics (3.1) we have an exponential decrease of the initial $\mathcal{W}_2$-distance because the ULA converges exponentially fast to its stationary distribution p_∞, plus a term which represents an upper bound on the bias $\mathcal{W}_2(p_\infty, \pi)$, that scales polynomially with respect to the dimension of the state-space and step-size. This makes sense given that the higher the step-size the more the trajectories from the ULA and from the continuous-time process differ due non-exhaustive state-space exploration; then, this error is proportional to the number of dimensions in which the trajectory evolves. Furthermore, the condition number affects negatively the convergence because the higher it is the more different curvature is along different directions of the potential function, or equivalently, of the (negative) log-likelihood of the target distribution. Finally, note that we might initialize the algorithm at the mode (minimizer), x_*, of π (V) given the less complexity required to find it (less with respect to sampling). Then $p_0 = \delta_{x_*}$ would imply $\sqrt{\alpha}\mathcal{W}_2(p_0, \pi) \leq \sqrt{d}$, given the assumptions on π. It is desirable to state results for $\sqrt{\alpha}\mathcal{W}_2$ because in this metric is scale-invariant. We now give the proof of (7), as in [26].

Proof of Theorem 7. We assume that both the Langevin diffusion and the ULA are initialized at the same measure p_0. We couple the two processes by taking $X_0 = Z_0$ and using the same Brownian Motion,

$$X_h = Z_0 - h\nabla V(Z_0) + \sqrt{2}B_h,$$
$$Z_h = Z_0 - \int_0^h \nabla V(Z_t)\mathrm{d}t + \sqrt{2}B_h.$$

Coupling means to introduce some kind of dependence structure (same starting point and same Brownian Motion here) between the two processes while maintaining the original marginal distributions. Then we have,

$$\mathcal{W}^2(p_h, \pi_h) \leq \mathbb{E}[\|X_h - Z_h\|^2] \leq \mathbb{E}\left[\left\|\int_0^h \nabla V(Z_t)\mathrm{d}t - h\nabla V(Z_0)\right\|^2\right]$$
$$\leq h\int_0^h \mathbb{E}[\|\nabla V(Z_t) - \nabla V(Z_0)\|^2]\mathrm{d}t.$$

Now, given that V is β-smooth, i.e. $\|\nabla V(Z_t) - \nabla V(Z_0)\|^2 \leq \beta^2\|Z_t - Z_0\|^2$, we need to

bound the movement of the Langevin diffusion in time t, namely $\|Z_t - Z_0\| = -\int_0^t \nabla V(Zs)\mathrm{d}s + \sqrt{2}B_t$. As a result of stochastic calculus theory, if ∇V is β-Lipschitz and $t \leq 1/3\beta$ then,

$$\mathbb{E}[\|Z_t - Z_0\|^2] \leq 8t^2\mathbb{E}\left[\|\nabla V(Z_0)\|^2\right] + 8dt$$

and,

$$\mathcal{W}_2^2(p_h, \pi_h) \leq \beta^2 h \int_0^h \mathbb{E}\left[\|Z_t - Z_0\|^2\right] \mathrm{d}t \leq 3\beta^2 h^4 \mathbb{E}\left[\|\nabla V(Z_0)\|^2\right] + 4\beta^2 dh^3.$$

Therefore, we managed to bound the discretization error for one step of the ULA.
Now we produce a coupling of $p_{(k+1)h}$ and π. Assume $X_{kh} \sim p_{kh}$ and $Z_{kh} \sim \pi$ to be optimally coupled, that is such that they minimize their $\mathcal{W}_2$-distance. Using the same Brownian Motion we set,

$$\begin{aligned} X_{(k+1)h} &:= X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}), \\ Z_t &:= Z_{kh} - \int_{kh}^t \nabla V(Z_s)\mathrm{d}s + \sqrt{2}(B_t - B_{kh}), \quad for\ t \in [kh, (k+1)h]. \end{aligned}$$

Then, $X_{(k+1)h} \sim p_{(k+1)h}$ and $Z_t \sim \pi$ for $t \geq kh$ due to stationarity. We define an auxiliary process,

$$\bar{X}_t := X_{kh} - \int_{kh}^t \nabla V(\bar{X}_s)\mathrm{d}s + \sqrt{2}(B_t - B_{kh}), \quad t \in [kh, (k+1)h],$$

which denotes the Langevin diffusion started at X_{kh}. Then, we bound,

$$\begin{aligned} \mathcal{W}_2(p_{(k+1)h}, \pi) &\leq \sqrt{\mathbb{E}\left[\|X_{(k+1)h} - Z_{(k+1)h}\|^2\right]} \\ &\leq \sqrt{\mathbb{E}\left[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2\right]} + \sqrt{\mathbb{E}\left[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2\right]}. \end{aligned}$$

For the first term, which consists of two Langevin diffusion with an α-strongly convex potential, we have the following contraction inequality,

$$\begin{aligned} \mathbb{E}\left[\|\bar{X}_{(k+1)h} - Z_{(k+1)h}\|^2\right] &\leq e^{-2\alpha h}\mathbb{E}\left[\|X_{kh} - Z_{kh}\|^2\right] \\ &= e^{-2\alpha h}\mathcal{W}_2^2(p_{kh}, \pi). \end{aligned}$$

For the second term we are back to the setting we analyzed for the one step discretization bound. Thus,

$$\begin{aligned}\mathbb{E}\left[\|X_{(k+1)h} - \bar{X}_{(k+1)h}\|^2\right] &\leq 3\beta^2 h^4 \mathbb{E}\left[\|\nabla V(X_{kh})\|^2\right] + 4\beta^2 dh^3 \\ &\lesssim \beta^4 h^4 \mathbb{E}\left[\|X_{kh} - Z_{kh}\|^2\right] + \beta^2 h^4 \mathbb{E}_\pi\left[\|\nabla V\|^2\right] + \beta^2 dh^3 \\ &\leq \beta^4 h^4 \mathcal{W}_2^2(p_{kh}, \pi) + \beta^3 dh^4 + \beta^2 dh^3,\end{aligned}$$

where we used the fact that if V is β-smooth then $\mathbb{E}_\pi\left[\|\nabla V\|^2\right] \leq \beta d$. Dropping the second term given the assumption $h \lesssim 1/\beta\kappa$ and using the relation $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$, we obtain,

$$\mathcal{W}_2(p_{(k+1)h}, \pi) \leq e^{-\alpha h}\mathcal{W}_2(p_{kh}, \pi) + O\left(\beta^2 h^2 \mathcal{W}_2(p_{kh}, \pi) + \beta d^{1/2} h^{3/2}\right).$$

Provided that $h \lesssim 1/\beta\kappa$ we have $e^{-\alpha h} + O(\beta^2 h^2) \leq e^{-\alpha h/2}$, so that,

$$\mathcal{W}_2(p_{(k+1)h}, \pi) \leq e^{-\alpha h/2}\mathcal{W}_2(p_{kh}, \pi) + O\left(\beta d^{1/2} h^{3/2}\right).$$

Then, we can iterate this construction to obtain,

$$\mathcal{W}_2(p_{Nh}, \pi) \leq e^{-\alpha Nh/2}\mathcal{W}_2(p_0, \pi) + O\left(\frac{\beta d^{1/2} h^{1/2}}{\alpha}\right),$$

completing the proof. □

This is a nice and intuitive proof, which relies on strong log-concavity of the stationary measure given the use of the contraction property of the Langevin gradient flow in the Wasserstein metric. However, this proof does not hold under weaker assumptions such as LSI and, at the same time, it does not provide sharp, tight non-asymptotic bounds. The next proof gives tight dependence of LMC on the condition number κ of the equilibrium distribution π.

3.1.2 Convex Optimization

This second proof is the result of considering the ULA as a first-order optimization algorithm in the space of measures and use tools from convex optimization. A first treatment of this approach is given in [34]; the finest results are then obtained in [12].

Here we will assume $\pi = \exp(-V)$ thus V incorporates the normalizing constant as well. The starting point is to decompose the relative divergence in two components,

$$\mathsf{KL}(\mu\|\pi) := \int_{\mathbb{R}^d} V \mathsf{d}\mu + \int_{\mathbb{R}^d} \mu \ln \mu,$$

where the first component, $\mathbb{E}_\mu[V]$, is usually called the *energy functional* and the second one is the usual negative entropy functional. Then, π satisfies a variational principle, namely it minimizes, over all probability measures on $E = \mathbb{R}^d$, the above functional, $\mathsf{KL}(\cdot\|\pi)$, which is also called *free energy functional*. This is a consequence of what we already pointed out in the previous chapters, namely that Langevin diffusion is the gradient flow of the free energy functional in the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. Correspondingly, we divide the ULA dynamics in two parts,

$$X^+_{kh} := X_{kh} - h\nabla V(X_{kh}),$$
$$X_{(k+1)h} := X^+_{kh} + \sqrt{2}(B_{(k+1)h} - B_{kh}).$$

The first step is a simple gradient descent on the potential V. It can be shown that $p^+_{kh} := \mathsf{law}(X^+_{kh})$ is obtained, in the space of measures, from p_{kh} by taking a gradient step of the energy functional $\mathcal{E}(\mu) := \int_{\mathbb{R}^d} V \mathsf{d}\mu$ with respect to the geometry induced by the 2-Wasserstein distance. The second step is the heat flow (the name comes from the fact the Fokker-Planck of the Brownian Motion is the heat equation) that is a Wasserstein gradient flow for the entropy functional $\mathcal{H}(\mu) := \int_{\mathbb{R}^d} \mu \ln \mu$, in the space of measures. This way of dividing the dynamics into its components (linearly here) and then discretize both of them is known as a splitting scheme. The LMC algorithm is also called the "forward-flow" algorithm due to the fact that the gradient descent step is called the forward method in optimization while the second step is the (heat)-flow. Then, this scheme aims to obtain sharp bounds by showing that the forward step dissolves the energy, $\mathcal{E}$, without increasing substantially the entropy, $\mathcal{H}$ and that the flow step does the same but in the other direction. Let's state the results.

Lemma 3. *Let $\pi = \exp(-V)$ be the target distribution satisfying $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. Let $(p_{kh})_{k\in\mathbb{N}}$ denote the probabilities measures defined by the ULA with step-size $h \in$*

$[0, \beta^{-1}]$. *Then,*

$$\mathrm{KL}(p_{(k+1)h} \| \pi) \leq \frac{1 - \alpha h}{2h} \mathcal{W}_2^2(p_{kh}, \pi) - \mathcal{W}_2^2(p_{(k+1)h}, \pi) + 2\beta d h^2. \tag{3.3}$$

From this lemma we obtain the following theorem, due to [12].

Theorem 8. *Let* $\pi = \exp(-V)$ *be the target distribution satisfying* $0 \preceq \alpha I_d \preceq \nabla^2 V \preceq \beta I_d$. *Let* $(p_{kh})_{k \in \mathbb{N}}$ *denote the probability measures defined by the ULA. Then:*

1. *(*V *convex,* $\alpha = 0$*) For any* $\epsilon \in [0, \sqrt{d}]$, *if we set* $h \asymp \epsilon^2 / \beta d$, *then for the mixture distribution* $\bar{p}_{Nh} := N^{-1} \sum_{k=1}^{N} p_{kh}$ *it holds that,*

$$\sqrt{\mathrm{KL}(\bar{p}_{Nh} \| \pi)} \leq \epsilon, \quad \textit{with mixing time} \tag{3.4}$$

$$t_{\textit{mix}}(\epsilon, p_0) = O\left(\frac{\beta d \, \mathcal{W}_2^2(p_0, \pi)}{\epsilon^4} \right). \tag{3.5}$$

2. *(*V *strongly-convex,* $\alpha > 0$*) Then, for any* $\epsilon \in [0, \sqrt{d}]$, *if we set* $h \asymp \epsilon^2 / \beta d$, *we obtain* $\sqrt{\alpha} \mathcal{W}_2(p_{Nh}, \pi) \leq \epsilon$ *and* $\sqrt{\mathrm{KL}(\bar{p}_{Nh,2Nh} \| \pi)} \leq \epsilon$ *with mixing time*

$$t_{\textit{mix}}(\epsilon, p_0) = O\left(\frac{\kappa d}{\epsilon^2} \log\left(\frac{\sqrt{\alpha} \mathcal{W}_2(p_0, \pi)}{\epsilon} \right) \right), \tag{3.6}$$

where $\bar{p}_{Nh,2Nh} := N^{-1} \sum_{k=N+1}^{2N} p_{kh}$.

Thus, considering the strongly log-concave case, we managed to reduce the computational cost, that is the mixing time, by a factor equal to the condition number κ, with respect to the proof we gave with the 2-wasserstein coupling.

The LMC algorithm can be applied in several contexts having different assumptions. For example, joining the work of [32] and [8] we can obtain, under the assumptions that π satisfies a LSI and V is β-smooth, sampling guarantees in relative entropy that scale polynomially with respect to the parameters, e.g. $N = O\left(\frac{C_{LSI}^2 \beta^2 d}{\epsilon^2} \log \frac{\mathrm{KL}(p_0 \| \pi)}{\epsilon^2} \right)$. This setting covers cases where π is not log-concave as long as the above assumptions are met.

3.2 Metropolis-Hastings

The Langevin Monte Carlo algorithm is the most basic sampling algorithm given that it only consists in a discretization of the corresponding continuous-time dynamics. Furthermore, given the presence of the bias introduced by the discretization, the algorithm scales as poly$(1/\epsilon)$ meaning that the more precise we want to be in approximating π the higher, as a polynomial function of $1/\epsilon$, the computational cost (that is the number of iterations needed by the algorithm) we have. This is why these kinds of samplers are called *low-accuracy samplers*.

In order to obtain algorithms that scale as polylog$(1/\epsilon)$, i.e. they scale polynomially in the logarithm of the argument, we need to remove the bias; this is achieved by *high-accuracy samplers*, e.g. by applying the Metropolis–Hastings filter to the ULA and obtain the *Metropolis-adjusted Langevin algorithm (MALA)*, which was introduce by [3].

The class of Metropolis-Hastings (MH) algorithms is an established family of Markov Chain Monte Carlo (MCMC) algorithms used to sample from a target distribution π by constructing a π-reversible Markov Chain. A first formulation was developed in 1953 by the physicist Nicholas Metropolis and then generalized by the statistician W.K. Hastings. An instance of such computational models is given by the Gibbs sampler, which is suitable for sampling from a multivariate distribution by iteratively sampling from its univariate full conditional distributions. These algorithms apply an idea similar to Rejection Sampling, that is given the current state of the chain they propose a potential new sample in the chain from a proposal kernel Q and then either accept or reject the new move with a well-defined probability which ensures the chain to be π-stationary. Let Q be a kernel on the product space $\mathbb{R}^d \times \mathbb{R}^d$ so that $Q(x, \cdot)$ is a probability measure on the d-dimensional Euclidean space, for all $x \in \mathbb{R}^d$. Usually, we choose kernels such that $Q(x, \cdot)$ is absolutely continuous with respect to the Lebesgue measure and will write $Q(x, y)$ to indicate the corresponding density evaluated at $y \in \mathbb{R}^d$. Thus, we start from $X \in \mathbb{R}^d$ and propose a new sample $Y \sim Q(X, \cdot)$; then we accept Y with an acceptance probability equal to $a(X, Y)$. Usually, the choice for such an acceptance probability is given by the Metropolis-Hastings filter,

$$a(X, Y) := \min \left\{ 1, \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right\}. \tag{3.7}$$

Then, the transition kernel $P : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ of the resulting Markov Chain $(X_n)_{n\geq 0}$ is given by,

$$P(x, \mathsf{d}y) = Q(x, \mathsf{d}y)A(x, y) + \left(1 - \int_{\mathbb{R}^d} Q(x, \mathsf{d}y')A(x, y')\right)\delta_x(\mathsf{d}y), \tag{3.8}$$

where the integral computes the total probability of accepting a new move so that inside the parenthesis we have the rejection probability and $\delta_x(\mathsf{d}y)$ is the Dirac measure. Then, the Chapman-Kolmogorov equations allow us to compute the k-th step transition kernel as $P^k(x, \mathsf{d}y) = \int_{z\in\mathbb{R}^d} P^{k-1}(x, \mathsf{d}z)P(z, \mathsf{d}y)$. It turns out that the Markov chain we just defined through its transition kernel P, which depends on the proposal kernel Q and on the acceptance probability $a(\cdot, \cdot)$, is π-reversible, that is,

$$\forall x, y, \in \mathbb{R}^d, x \neq y : \quad \pi(x)P(x, y) = \pi(y)P(y, x),$$

which is a stronger condition than π-stationarity, which can be obtained by summing both sides,

$$\begin{aligned}\pi(x) &= \sum_{y\in\mathbb{R}^d} \pi(x)P(x, y) \\ &= \sum_{y\in\mathbb{R}^d} \pi(y)P(y, x).\end{aligned}$$

Furthermore, a π-reversible Markov chain is also time-reversible, that is the process defined as $Y_N = X_0, ..., Y_0 = X_N$ is a Markov chain with the same transition probabilities. Finally, we can define the transition operator $\mathcal{T}_P$ as,

$$\mathcal{T}_P(p_n)(S) := \int_{y\in\mathbb{R}^d} \mathsf{d}p_n(y)P(y, A), \quad \forall A \in \mathcal{B}(\mathbb{R}^d),$$

so that if p_n is the distribution of the current state then $\mathcal{T}_P(p_n) = p_{n+1}$ and, recursively, $\mathcal{T}_{P^k}(p_n) = p_{n+k}$.

Therefore, the Metropolis-Hastings filter (consisting on the acceptance-rejection step) makes the sampling algorithm unbiased, no matter the choice of the proposal kernel. The proposal kernel will influence the velocity at which the algorithm converges to stationarity. One important thing is that here we are assuming the chain to be irreducible, recurrent and aperiodic, that is ergodic, so that the stationary distribution is unique and

we can apply the ergodic theorem (after a burn-in period) which can be interpreted as a law of large numbers for Markov chains. Now, in order to implement the MH algorithm we need to be able to easily sample from $Q(x, \cdot)$ and to evaluate its density for the acceptance probability.

The choice of the proposal, as we said, allows for some flexibility. Among the options we can find: the Independent MH algorithm where the proposal is independent on the current value of the chain, the Metropolized random walk (MRW) where the proposal is a d-dimensional Gaussian centered at the current state with diagonal covariance matrix, Metropolized Hamiltonian Monte Carlo (MHMC) and the Metropolis-adjusted Langevin algorithm (MALA) where the proposal is given by the LMC algorithm, namely $Q(x, \cdot) = \mathcal{N}(\cdot; x - h\nabla V(x), 2hI_d)$.

3.2.1 Metropolis-adjusted Langevin algorithm

The general assumptions are always that the potential field, V, of $\pi \propto e^{-V}$, is α-strongly convex and β-smooth, i.e. $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$.

Then, the procedure of the MALA is as follows. Choose the step-size $h > 0$ (this choice depends on several variables), the total number of iterations N and set $n = 0$. Initialize the algorithm at $X_0 \sim p_0$. Then, for $n = 0, .., N - 1$, repeat:

1. Proposal step: sample $Y_{n+1} \sim Q(X_n, \cdot)$, where

$$Q(X_n, \cdot) := (4\pi h)^{-d/2} \exp\left(-\frac{\| \cdot - X_n + h\nabla V(X_n)\|^2}{4h}\right),$$

2. Accept-reject step: with probability $\min\left\{1, \frac{\pi(Y_{n+1})Q(Y_{n+1}, X_n)}{\pi(X_n)Q(X_n, Y_{n+1})}\right\}$ set $X_{n+1} \leftarrow Y_{n+1}$, else set $X_{n+1} \leftarrow X_n$,

3. Output: $(X_1, ..., X_N)$.

Now, a delicate role is played by the starting point, $X_0 \sim p_0$. There are two options: a feasible start and a warm one. Basically, the difference is given by how close the initial distribution is to the stationary one. Usually, we take $p_0 = normal(0, \beta^{-1}I_d)$ with the minimizer of V being 0, $\nabla V(0) = 0$. For the warm start, instead, we assume to have found p_0 such that $\chi^2(p_0\|\pi) = O(1)$ (remember what we said at the end of chapter 2 about the usual initial complexity under a PI, $\chi^2(p_0\|\pi) = \exp(O(d))$, so that this

is indeed an improvement). We are using the chi-square divergence because Markov chains are usually analyzed via spectral theory which is related to the PI - Spectral Gap inequality.
Now, we can state the sampling guarantees for both the situations. For the feasible start setting you can take a look at [13].

Theorem 9. *Consider the target distribution $\pi \propto e^{-V}$ satisfying the assumptions previously stated. Then, given $X_0 \sim \mathcal{N}(0, \beta^{-1} I_d)$ and appropriate step-size $h > 0$, the MALA returns a measure p_N satisfying,*

$$\sqrt{\chi^2 \left(p_N \| \pi\right)} \leq \epsilon \quad \textit{after} \tag{3.9}$$

$$N = \tilde{O}\left(\kappa d \ \operatorname{polylog} \frac{1}{\epsilon}\right) \quad \textit{iterations}. \tag{3.10}$$

For the result with a warm start you can consult [9], [35] and [6].

Theorem 10. *Consider the target distribution $\pi \propto e^{-V}$ satisfying the assumptions previously stated. Then, given an initial distribution p_0 satisfying $\chi^2(p_0 \| \pi) = O(1)$ and appropriate step-size $h > 0$, the MALA returns a measure p_N satisfying,*

$$\|p_N - \pi\|_{TV} \leq \epsilon \quad \textit{after} \tag{3.11}$$

$$N = \tilde{O}\left(\kappa d^{1/2} \ \operatorname{polylog} \frac{1}{\epsilon}\right) \quad \textit{iterations}. \tag{3.12}$$

Therefore, then we have an improvement of magnitude $d^{1/2}$ in the scaling of the computational cost that, for high-dimensional distributions, can implicate remarkable differences.

3.3 More Computational Models

The computational models presented so far are not the only options available to carry out sampling, but there are a lot of them, the literature is really ample. For example, we might adopt a different discretization scheme such as the randomized midpoint

discretization that instead of approximating the integral $\int_{kh}^{(k+1)h} \nabla V(X_t)\mathrm{d}t$ (appearing in the Langevin diffusion in the time interval $[kh,(k+1)h]$) with $-h\nabla V(X_{kh})$ it adopts an unbiased estimator, i.e. $-h\nabla V(X_{(k+u_k)h})$ with $u_k \sim \mathrm{uniform}([0,1])$. However, this term is not known so that we must approximate it with an Euler-Maruyama step, to finally obtain the RM-LMC scheme:

$$X_{(k+1)h} := X_{kh} - h\nabla V(X_{(k+u_k)h}) + \sqrt{2}\left(B_{(k+1)h} - B_{kh}\right),$$
$$X_{(k+u_k)h} := X_{kh} - u_k h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+u_k)h} - B_{kh}),$$

with $(u_k)_{k\in\mathbb{N}} \overset{i.i.d.}{\sim} \mathrm{uniform}([0,1])$ and independent of X_0 and the Brownian Motion. To sample the Gaussian noise we first sample from the uniform and then from $\mathrm{normal}(0,u_k h)$. Then, using the same assumptions on π, i.e. $\alpha I_d \preceq \nabla^2 V \preceq \beta I_d$, and similar parameters, this discretization scheme reach $\sqrt{\alpha}\mathcal{W}_2(p_{Nh},\pi) \leq \epsilon$ after $N = \tilde{O}\left(\frac{\kappa d^{1/2}}{\epsilon} \vee \frac{\kappa^{4/3} d^{1/3}}{\epsilon^{2/3}}\right)$ iterations. Thus, the RM-LMC scales faster, in all the parameters, than the corresponding the LMC that uses Euler-Maruyama approximation scheme.

Then, we have two famous alternatives that change the underlying process. The first one is the Hamiltonian Monte Carlo algorithm (whose integration within the MH framework is particularly known) which takes inspiration from the Hamiltonian formulation of classical mechanics where the laws of motion do not follow Newton's laws but a system of coupled first-order ODEs. The other one is the *Underdamped* Langevin diffusion (or *kinetic* Langevin) that is an extension of the Hamiltonian Monte Carlo dynamics and its dynamics is governed by a system of two SDEs, one modeling the position and the other the velocity of the system. Please, refer to [21] and [7] for an introduction to these topics.

Finally, as already said, the literature regarding sampling algorithms based on stochastic processes is really vast and it includes models that were not cited here, such as Mirror Langevin, the Proximal Sampler but also non-reversible MCMC algorithms based on Piecewise Deterministic Markov Processes, and all of them can output different non-asymptotic bounds depending on the setting. The instance involving non log-concave measures not satisfying functional inequalities such as LSI and PI is particularly complicated at the moment as there are not algorithms scaling in acceptable times. Diffusion models might be an alternative paradigm capable of solving, at least partially,

this problem. As we will see in the next chapter, [5], [20] and [4] have obtained algorithms scaling polynomially in all the parameters. But maybe some assumptions are not realistic/flexible enough.

4 Diffusion Models

Diffusion Models (DMs) are deep generative models exhibiting state-of-the-art performances in many applications, e.g. image synthesis, video generation and molecule design. They combine a fascinating theoretical foundation, such as the theory of diffusion processes, with an increasing interest from researchers due to its applicability in many businesses. As a consequence of this there has been a significant growth in the amount of research papers published in this field; thus, it is quite difficult to stay abreast with the continuous developments made by the community. However, despite this, there still is a lot of work to do toward both theoretical and empirical directions.
These models operate by defining a forward-in-time SDE that progressively destroys the structure of the data by injecting noise, ultimately transforming a sample into pure noise. Under certain regularity conditions, there exists a corresponding reverse-in-time SDE that can reconstruct the data structure, generating new samples from noise and thus enabling generative modeling. However, some approximations are necessary, so, in practice, we sample from a distribution that is as close as possible to p_{data}.

4.1 Overview

Usually, the forward SDE, whose role is to destroy the data structure, is given by the Ornstein-Uhlenbeck (OU) process $\{X_t^x; t \geq 0, x \in \mathbb{R}^d\}$, which is governed by the equation,

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 = x, \tag{4.1}$$

where $(B_t)_{t\geq 0}$ is a standard d-dimensional Brownian Motion and the initial value is a sample from the data distribution, $X_0 \sim p_{data}$. This is the unique time-homogeneous Markov process which is also a Gaussian process. Furthermore, it is an instance of the Langevin diffusion with unique and reversible stationary distribution the standard d-dimensional Gaussian, $\mathrm{d}\gamma^d(x) = (2\pi)^{-d/2} e^{-|x|^2/2} \mathrm{d}x$. Its generator is $\mathscr{L} = \Delta - x \cdot \nabla$ and it is self-adjoint in the $\mathbb{L}^2(\gamma^d)$-space. The OU process is one of the few examples of stochastic processes having a closed-form solution; in this case, it is given by,

$$X_t^x = e^{-t}\left(x + \sqrt{2}\int_0^t e^s \mathrm{d}B_s\right),$$

where we have the Itô stochastic integral $\int_0^t e^s \mathrm{d}B_s \sim \mathcal{N}\left(0, \int_0^t e^{2s}\mathrm{d}sI_d\right)$, with the variance given by the Itô isometry. Therefore, $X_t^x \sim \mathcal{N}\Big(e^{-t}x, (1-e^{-2t})I_d\Big)$, so that the semigroup $(P_t)_{t\geq 0}$ has an explicit representation, given by,

$$P_t f(x) = \mathbb{E}\left(f(X_t^x)\right) = \int_{\mathbb{R}^d} f\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) \mathrm{d}\gamma^d(y), \quad t \geq 0, x \in \mathbb{R}^d.$$

It is standard knowledge that the OU process (better, the standard Gaussian) satisfies both the log-Sobolev and the Poincaré inequalities with constants equal to 1.

The stationary measure of the forward process is called, in the context of DMs, prior distribution because the generative procedure starts by sampling from it; sampling from the d-dimensional standard Gaussian is easy, that is why the OU process has been, up to now, the default choice. As shown in [1] through a rearrangement of the Fokker-Planck, equation (4.1) admits a reverse-time SDE which has the following expression,

$$\mathrm{d}\overleftarrow{X}_t = \left\{\overleftarrow{X}_t + 2\nabla_{\overleftarrow{x}_t} \ln p_{T-t}(\overleftarrow{X}_t)\right\}\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim p_T, \tag{4.2}$$

where $\overleftarrow{X}_t := X_{T-t}$ for $t \in [0,T]$ is the backward diffusion process. Note that the two processes share the same marginals distributions. The *score function* or *Stein score*, $\nabla \ln p_t$, is a vector field which points towards regions of the space where the high-dimensional vector $\overleftarrow{X}_t$ has higher density, namely towards the modes of the distribution, as Langevin dynamics does. As derived in [30], the reverse-SDE (4.2) is associated with a corresponding ODE, called *Probability Flow ODE*, which has the same marginal distributions along the trajectory and has the following dynamics,

$$\dot{X}_t = X_{T-t} + \nabla \ln p_{T-t}(X_{T-t}). \tag{4.3}$$

Therefore, this is an alternative way of carrying out the generation of new samples (through ODE solvers). However, the absence of stochastic noise might prevent it to discover low probability regions of distributions.

However, not knowing the data distribution we cannot marginalize the conditional distribution of the OU process; therefore, we need to estimate it using a deep neural network, thus introducing an approximation error. The loss function used in the training procedure is given by equation (1.5); in practice, that loss is replaced with the empiri-

cal counterpart that uses the set of i.i.d. samples from p_{data}, namely $\left(X_0^{(1)}, ..., X_0^{(n)}\right) \overset{iid}{\sim} p_{data}$. Then, a set of score functions is estimated; in fact, let $h > 0$ be the step-size of the discretization, so that the time interval $[0, T]$ is divided into N sub-intervals indexed by $k \in \{0, ..., N\}$ such that $T = Nh$. Let's assume that during the training procedure we have trained a deep neural network to estimate the score function, $s_\theta(\overleftarrow{X}_{kh}, T - kh) \approx \nabla \ln p_{T-kh}$, for each time $k \in \{0, ..., N\}$. The treatment of Noise Conditional Score Networks can be found in [29]. Then, between the discretization steps, $t \in [kh, (k+1)h]$, we freeze the score, so that we obtain the following SDE,

$$\mathrm{d}\overleftarrow{X}_t = \left[\overleftarrow{X}_t + 2s_\theta\left(\overleftarrow{X}_{kh}, T - kh\right)\right]\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t \in [kh, (k+1)h], \tag{4.4}$$

which represents the so-called exponential integrator scheme (see [37]). This is again a linear SDE, thus it has a closed form solution, namely $\overleftarrow{X}_{(k+1)h} | \overleftarrow{X}_{kh}$ is normally distributed. We just introduced the second approximation of the algorithm, the discretization scheme. Finally, in order to start the sampling procedure we sample from the standard d-dimensional Gaussian, γ^d, because we do not know the marginal p_T which would be available if we knew p_{data}. Fortunately, the process (4.1) converges exponentially fast, that is $\mathrm{KL}\left(p_T \| \gamma^d\right) \leq e^{-2T}\mathrm{KL}\left(p_{data} \| \gamma^d\right)$, as a result of the LSI. Therefore we introduce a third approximation in our algorithm, $p_T \approx \gamma^d$. Then, denote with $q_t := \mathrm{law}(\overleftarrow{X}_t)$ the law of the algorithm at time t.

4.2 Convergence

It is of great interest to study theoretical convergence guarantees of such models, that is, given a set of assumptions (e.g. finite second moment of p_{data}) about the model, state upper bounds on the distance between the measure we sample from, q_T, and the target one, p_{data}, as a function of the approximations that were implemented and of the assumptions made.

Most of the times the distributions from which we want to generate new sample are really high-dimensional (e.g. images can have millions of dimensions), probably they have several modes and are non log-concave. This prevents the assumption that p_{data} satisfies a LSI, as in [19]. There are also some papers which present results (bounds) that scale exponentially in the dimension of the data and some parameters, as in [11].

One of the main problems is whether the data distribution is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$ or if, instead, it satisfies the manifold hypothesis, that is, p_{data} is supported on a lower-dimensional manifold embedded in the d-dimensional euclidean space. This might sound like a reasonable hypothesis since, as in the case of images, it is quite unlikely that we need millions of dimensions to characterize a single sample. If you want to know how to estimate the intrinsic data dimension take a look at [31]. As a consequence of the acceptance of the manifold hypothesis we encounter several problems: the score function is not well-defined and will explode when $t \downarrow 0$ due to the fact that $\nabla \log p_t = \nabla p_t / p_t$ with the denominator being zero outside of the manifold. Another problem is that $\mathrm{KL}(p_{data} \| \gamma^d) = \infty$ if $p_{data} << \gamma^d$ does not hold, so that we can not describe the convergence of the forward process to equilibrium. However, polynomial guarantees have been obtained also under the manifold hypothesis.
Let's consider the results in [5], with similar results obtained in [20] and [4],. They make three assumptions about the data distribution, p_{data}.

1. The score function is L-Lipschitz for each $t \geq 0$:

$$\|\nabla \ln p_t(x) - \nabla \ln p_t(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^d. \tag{4.5}$$

2. The data distribution has finite second moment, $\mathsf{m}_2^2 := \mathbb{E}_{p_{data}}[\|\cdot\|^2] < \infty$.

3. For all $k = 1, ..., N$ we have an L^2-bound on the score estimation error,

$$\mathbb{E}_{p_{kh}}\left[\|s_\theta(\cdot, kh) - \nabla \ln p_{kh}(\cdot)\|^2\right] \leq \epsilon_{score}^2. \tag{4.6}$$

The first assumption assures us that $p_{data} << \mathrm{d}x$ so that it does not lie on a submanifold; this is crucial in order to state the first theorem. The second, requiring finite variance, is a quite standard assumption in order to guarantee a well-posed model. The last assumption is crucial for the results of this paper, given the statement there reported: "*Given an L2-accurate score estimate, diffusion models can sample from (essentially) any data distribution.*". Deep neural networks take in input high-dimensional data and are parametrized by an impressive number of weights and biases; furthermore, the nonlinearities introduce by the so-called activation functions make the loss

landscape non-convex and highly complex, so that the optimization procedure is mostly empirical and not theoretically understandable yet. Consequently, it is not possible to give very precise guarantees for the estimation error. Therefore, they reduced the problem of sampling from a high-dimensional multimodal non log-concave distribution to the task of being able to estimate with L^2-accuracy the score functions, $\nabla \ln p_t$. This requirements is more reasonable than the one of L^∞-accuracy introduced in previous works due to the fact that the neural network might not be able to generalize super-well in low density regions of the distributions because we might have few samples, coming from those regions, for the training procedure. However, empirically, these architectures seem to be quite good in estimating the score functions.
The forward and reverse SDEs are defined as in (4.1) and (4.2). We can state the main theorem of the paper.

Theorem 11. *Suppose that assumptions 1, 2 and 3 hold. Let q_T be the output of equation* (4.4) *started with a sample from γ^d, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$ where $L \geq 1$. Then we have the following bound*

$$\|q_T - p_{data}\|_{TV} \lesssim e^{-T}\sqrt{\mathrm{KL}(p_{data}\|\gamma^d)} + (L\sqrt{dh} + L\mathsf{m}_2 h)\sqrt{T} + \epsilon_{score}\sqrt{T}. \tag{4.7}$$

Then, assuming $\mathrm{KL}(p_{data}\|\gamma^d) \leq \mathrm{poly}(d)$ (which is standard assumption) and $\mathsf{m}_2 \leq d$ (meaning that the distribution does not have high variance), choosing $T \asymp \log(\mathrm{KL}(q\|\gamma^d)/\epsilon)$ and $h \asymp \epsilon^2/L^2 d$, we have

$$\|q_T - p_{data}\|_{TV} \leq \tilde{O}(\epsilon + \epsilon_{score}), \quad in$$
$$N = \tilde{\Theta}\left(\frac{L^2 d}{\epsilon^2}\right) \text{ iterations.}$$

First of all, this computational complexity is very good given that it scales polynomially in all the parameters of interest. Therefore, assuming an L^2-accurate score estimate (for each time along the discretization scheme), we can sample with polynomial complexity even when p_{data} is non log-concave, but satisfies the other two assumptions, i.e. it has full support and has finite second moment. This sampling guarantee matches state of the art mixing time of the LMC algorithm under a LSI assumption, as in [32] and [8].

The bound (4.18) has three terms on the right-hand side. The first one concerns the convergence of the forward SDE, $p_T \approx \gamma^d$; we know that the OU process with γ^d as stationary measure converges exponentially fast (at rate $\exp(-2t)$) in relative entropy as implied by the LSI. The second term is about the discretization error of the trajectory and the last term due to the score function error. Thus, these are three approximation used to bound the convergence of the model.

It turns out that the best way to study Markov diffusions is via relative entropy using the Girsanov's theorem to bound the divergence between path measures. Let's give a sketch of the proof, but first some notation:

- The data distribution is $p_{data} = p_0$.
- The forward process is denoted $(X_t)_{t\in[0,T]}$, and $X_t \sim p_t$.
- The reverse process is denoted $(\overleftarrow{X}_t)_{t\in[0,T]}$, where $\overleftarrow{X}_t := X_{T-t} \sim p_{T-t}$.
- The sampling algorithm is denoted $(\overleftarrow{X}_t)_{t\in[0,T]}$, where $\overleftarrow{X}_t \sim q_t$ Recall that we initialize at $q_0 = \gamma^d$, the standard Gaussian measure.
- The process $(\overleftarrow{X}_t^{p_T})_{t\in[0,T]}$, where is the same as $(\overleftarrow{X}_t)_{t\in[0,T]}$, except that we initialize this process at p_T rather than at γ^d. We write $\overleftarrow{X}_t^{p_T} \sim q_t^{p_T}$.

It is useful to consider a single stochastic process, $(X_t)_{t\in[0,T]}$, and two different measures over the path space $C([0,T];\mathbb{R}^d)$. The considered path measures are:

- $\overleftarrow{P}_T$, under which $(X_t)_{t\in[0,T]}$ has the law of the reverse process (4.2);
- $Q_T^{p_T}$, under which $(X_t)_{t\in[0,T]}$ has the law of the SGM algorithm initialized at p_T (corresponding to the process $(\overleftarrow{X}_t^{p_T})_{t\in[0,T]}$).

Let's proceed with the sketch of the proof.

Sketch of Proof of Theorem 11. Let's recall a consequence of Girsanov's theorem.

Theorem 12. *For $t \in [0,T]$, let $\mathcal{L}_t = \int_0^t b_s dB_s$ where B is a P-Brownian motion. Assume $\mathbb{E}_P \int_0^T \|b_s\|^2 s < \infty$. Then, $\mathcal{L}$ is a P-martingale in the Hilbert $L^2(P)$. Then, if*

$$\mathbb{E}_P \mathcal{E}(\mathcal{L})_T = 1\,, \quad \textit{where} \quad \mathcal{E}(\mathcal{L})_t := \exp\Big(\int_0^t b_s\, dB_s - \frac{1}{2}\int_0^t \|b_s\|^2\, ds\Big)\,, \tag{4.8}$$

then $\mathcal{E}(\mathcal{L})$ is also a P-martingale and the process

$$t \mapsto B_t - \int_0^t b_s \, ds \tag{4.9}$$

is a Brownian motion under $Q := \mathcal{E}(\mathcal{L})_T \, P$, the probability distribution with density $\mathcal{E}(\mathcal{L})_T$ w.r.t. P.

The condition (4.8) is needed in order to apply the Girsanov's theorem. If it holds, then we can apply the theorem to $P = \overleftarrow{P_T}$ and

$$b_t = \sqrt{2}\left(s_\theta(X_{kh}, T - kh) - \nabla \ln p_{T-t}(X_t)\right), \tag{4.10}$$

where $t \in [kh, (k+1)h]$. This tells us that under $Q = \mathcal{E}(\mathcal{L})_T \overleftarrow{P_T}$, there exists a Brownian motion $(\tilde{B}_t)_{t\in[0,T]}$ such that

$$\mathsf{d}B_t = \sqrt{2}\left(s_\theta(X_{kh}, T - kh) - \nabla \ln p_{T-t}(X_t)\right) \mathsf{d}t + \mathsf{d}\tilde{B}_t \,. \tag{4.11}$$

Recall that under $\overleftarrow{P_T}$ we have a.s.

$$\mathsf{d}X_t = \{X_t + 2\,\nabla \ln p_{T-t}(X_t)\}\,\mathsf{d}t + \sqrt{2}\,\mathsf{d}B_t \,, \qquad X_0 \sim p_T \,. \tag{4.12}$$

The equation above still holds Q-a.s. since $P \ll \overleftarrow{P}_T$ (even if B is no longer a Q-Brownian motion). Substituting (4.11) into (4.12) we have Q-a.s.,

$$\mathsf{d}X_t = \{X_t + 2\,s_\theta(X_{kh}, T - kh)\}\;\mathsf{d}t + \sqrt{2}\,\mathsf{d}\tilde{B}_t \,, \qquad X_0 \sim p_T \,. \tag{4.13}$$

This translates in the fact that, under Q, the distribution of X is the sampling algorithm started at p_T, i.e., $Q = Q_T^{p_T} = \mathcal{E}(\mathcal{L})_T \overleftarrow{P}_T$. Therefore,

$$\mathrm{KL}(\overleftarrow{P}_T \| Q_T^{p_T}) = \mathbb{E}_{\overleftarrow{P}_T} \ln \frac{\mathsf{d}\overleftarrow{P}_T}{\mathsf{d}Q_T^{p_T}} = \mathbb{E}_{\overleftarrow{P}_T} \ln \mathcal{E}(\mathcal{L})_T^{-1} \tag{4.14}$$

$$= \sum_{k=0}^{N-1} \mathbb{E}_{\overleftarrow{P}_T} \int_{kh}^{(k+1)h} \|s_\theta(X_{kh}, T - kh) - \nabla \ln p_{T-t}(X_t)\|^2, \mathsf{d}t \,, \tag{4.15}$$

where we applied the martingale equality $\mathbb{E}_{\overleftarrow{P}_T} \mathcal{L}_t = 0$.

The equality (4.14) allows us to bound the relative entropy between the reverse process and sampling algorithm started at the same distribution, p_T.

Then, in order to conclude the proof we need to prove that

$$\|Q_T^{p_T} - \overleftarrow{P_T}\|_{TV}^2 \leq \mathrm{KL}(\overleftarrow{P_T}\|Q_T^{p_T}) \lesssim (\epsilon_{score}^2 + L^2 dh + L^2 \mathsf{m}_2^2 h^2)T, \tag{4.16}$$

where the first inequality is the Pinsker-Csizsar- Kullback inequality.

Finally, by data processing inequality,

$$\|q_T - p_{data}\|_{TV} \leq \|Q_T - Q_T^{p_T}\|_{TV} + \|Q_T^{p_T} - \overleftarrow{P_T}\|_{TV} \leq \|p_T - \gamma^d\|_{TV} + \|Q_T^{p_T} - \overleftarrow{P_T}\|_{TV}, \tag{4.17}$$

where the first term on the right-hand-side is bounded by the convergence of the forward OU process while the second by (4.16), so to obtain

$$\|q_T - p_{data}\|_{TV} \lesssim e^{-T}\sqrt{\mathrm{KL}(p_{data}\|\gamma^d)} + (L\sqrt{dh} + L\mathsf{m}_2 h)\sqrt{T} + \epsilon_{score}\sqrt{T}. \tag{4.18}$$

□

Notwithstanding, we can obtain similar results if we assume that the manifold hypothesis holds true and that p_{data} has bounded support; in this case, the new target becomes $p_t,\ t > 0$ so that the probability distribution from which we effectively sample is q_{T-t}. Then, it can be shown that $\|q_{T-t} - p_t\|_{TV} \leq \epsilon_{TV}$ and $\mathcal{W}_2(q_{T-t}, p_{data}) \leq \epsilon_{\mathcal{W}_2}$ given specific choices of the parameters. The early stopping technique is quite used to sample from a distribution having full support, in order to not have an exploding behavior with the Stein score at time $t \approx 0$.

5 Conclusions

This work has presented a theoretical investigation into the convergence properties of stochastic processes used in generative modeling, spanning Markov processes, classical sampling algorithms, and Diffusion Models. We have demonstrated how functional inequalities, ergodicity conditions, and spectral analysis contribute to establishing rigorous guarantees on the convergence of probability distributions to their target measures. Our analysis of classical sampling algorithms reveals the trade-offs between computational efficiency and theoretical guarantees, particularly in high-dimensional and non-log-concave settings. The discussion on Diffusion Models illustrates how their forward-backward stochastic dynamics enable generative modeling, albeit with necessary approximations that impact theoretical bounds on convergence. Despite significant advancements, challenges remain, particularly in understanding the role of deep neural networks in score-based sampling and the implications of the manifold hypothesis. Future research directions should focus on refining theoretical bounds, exploring alternative training paradigms, and improving the robustness of generative models under realistic constraints.

References

[1] Brian DO Anderson. “Reverse-time diffusion equation models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326.

[2] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*. Vol. 103. Springer, 2014.

[3] Julian Besag et al. “Bayesian computation and stochastic systems”. In: *Statistical science* (1995), pp. 3–41.

[4] Hongrui Chen, Holden Lee, and Jianfeng Lu. “Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 4735–4763.

[5] Sitan Chen et al. “Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions”. In: *arXiv preprint arXiv:2209.11215* (2022).

[6] Yuansi Chen and Khashayar Gatmiry. “A simple proof of the mixing of Metropolis-adjusted Langevin algorithm under smoothness and isoperimetry”. In: *arXiv preprint arXiv:2304.04095* (2023).

[7] Xiang Cheng et al. “Underdamped Langevin MCMC: A non-asymptotic analysis”. In: *Conference on learning theory*. PMLR. 2018, pp. 300–323.

[8] Sinho Chewi et al. “Analysis of langevin monte carlo from poincare to log-sobolev”. In: *Foundations of Computational Mathematics* (2024), pp. 1–51.

[9] Sinho Chewi et al. “Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm”. In: *Conference on Learning Theory*. PMLR. 2021, pp. 1260–1300.

[10] Arnak Dalalyan. “Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent”. In: *Conference on Learning Theory*. PMLR. 2017, pp. 678–689.

[11] Valentin De Bortoli. “Convergence of denoising diffusion models under the manifold hypothesis”. In: *arXiv preprint arXiv:2208.05314* (2022).

[12] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. “Analysis of Langevin Monte Carlo via convex optimization”. In: *Journal of Machine Learning Research* 20.73 (2019), pp. 1–46.

[13] Raaz Dwivedi et al. “Log-concave sampling: Metropolis-Hastings algorithms are fast”. In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[15] Aapo Hyvärinen and Peter Dayan. “Estimation of non-normalized statistical models by score matching.” In: *Journal of Machine Learning Research* 6.4 (2005).

[16] Richard Jordan, David Kinderlehrer, and Felix Otto. “The variational formulation of the Fokker–Planck equation”. In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17.

[17] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*. Vol. 113. Springer Science & Business Media, 1991.

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

[19] Holden Lee, Jianfeng Lu, and Yixin Tan. “Convergence for score-based generative modeling with polynomial complexity”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22870–22882.

[20] Holden Lee, Jianfeng Lu, and Yixin Tan. “Convergence of score-based generative modeling for general data distributions”. In: *International Conference on Algorithmic Learning Theory*. PMLR. 2023, pp. 946–985.

[21] Radford M Neal et al. “MCMC using Hamiltonian dynamics”. In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[22] Felix Otto and Cédric Villani. “Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality”. In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[23] Grigorios A Pavliotis. “Stochastic processes and applications”. In: *Texts in applied mathematics* 60 (2014).

[24] L Chris G Rogers and David Williams. *Diffusions, Markov processes, and martingales: Itô calculus*. Vol. 2. Cambridge university press, 2000.

[25] Leonard CG Rogers and David Williams. *Diffusions, markov processes, and martingales: Volume 1, foundations*. Cambridge university press, 2000.

[26] S.Chewi. *Log-Concave Sampling*. Lecture Notes, 2024.

[27] Adrien Saumard and Jon A Wellner. "Log-concavity and strong log-concavity: a review". In: *Statistics surveys* 8 (2014), p. 45.

[28] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[29] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in neural information processing systems* 32 (2019).

[30] Yang Song et al. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[31] Jan Stanczuk et al. "Your diffusion model secretly knows the dimension of the data manifold". In: *arXiv preprint arXiv:2212.12611* (2022).

[32] Santosh Vempala and Andre Wibisono. "Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices". In: *Advances in neural information processing systems* 32 (2019).

[33] Pascal Vincent. "A connection between score matching and denoising autoencoders". In: *Neural computation* 23.7 (2011), pp. 1661–1674.

[34] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Conference on Learning Theory*. PMLR. 2018, pp. 2093–3027.

[35] Keru Wu, Scott Schmidler, and Yuansi Chen. "Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling". In: *Journal of Machine Learning Research* 23.270 (2022), pp. 1–63.

[36] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *ACM Computing Surveys* 56.4 (2023), pp. 1–39.

[37] Qinsheng Zhang and Yongxin Chen. “Fast sampling of diffusion models with exponential integrator”. In: *arXiv preprint arXiv:2204.13902* (2022).

A Technical appendix

A.1 Convexity

For this subsection of the appendix we will mainly follow the review from [27].

Definition 12. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if, for all $\theta \in [0,1]$, it satisfies the following inequality*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \quad \forall x, y \in \mathbb{R}^d.$$

This definition of convexity is the most basic one and doesn't make any assumption about the curvature of the function. The following one is a stronger notion of convexity which, therefore, applies to a subset of the set of convex functions.

Definition 13. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex if, for $\alpha > 0$ and all $\theta \in [0,1]$, we have*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{1}{2}\alpha\theta(1-\theta)||x-y||^2, \quad \forall x, y \in \mathbb{R}^d.$$

Equivalently, it holds that $f(x) - \frac{1}{2}\alpha||x||^2$ is convex. The strong convexity assumption implies a lower bound for the function's curvature; in fact, if f is differentiable then the above inequality becomes $f(y) \geq f(x)\nabla f(x)^T(y-x) + \frac{1}{2}\alpha||y-x||^2$, where we can notice that the function will have positive curvature, meaning that the eigenvalues of the Hessian (if well defined) will be lower bounded by α, which can be written in terms of positive semi-definiteness as,

$$\nabla^2 f(x) \succeq \alpha I \iff \nabla^2 f(x) - \alpha I \succeq 0.$$

Furthermore, a function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is Lipschitz continuous with constant β, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x-y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Consequently, this is the same as requiring that the largest eigenvalue of $\nabla^2 f$ is at

most equal to β.

Now we can define the notion of log-concave distribution.

Definition 14. *A density function π with respect to the Lebesgue measure λ on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is log-concave if $\pi = e^{-\psi}$, where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a convex function.*

If we didn't have access to the normalizing constant then we may express it as $\pi \propto e^{-f}$, where f is convex.

Consequently, we have the following definition for a Probability Measure (PM).

Definition 15. *A probability measure P on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that the affine hull of $supp(P)$ has dimension d is a log-concave measure if and only if it has a log-concave density function $\pi = e^{-\psi}$ on $\mathbb{R}^d$; that is, we have*

$$P(A) = \int_A \pi d\lambda, \quad A \in \mathcal{B}(\mathbb{R}^d).$$

We can go on by defining the strong log-concavity of a density function π on $\mathbb{R}^d$.

Definition 16. *Let Σ be a $d \times d$ positive definite matrix and let $\mu \in \mathbb{R}^d$. Then a random vector X and its density π are strongly log-concave, and write $\pi \in SLC_2(\mu, \Sigma, d)$, if*

$$\pi(x) = g(x)\gamma_\Sigma(x - \mu), \quad x \in \mathbb{R}^d,$$

where g is a log-concave function and $\gamma_\Sigma(\cdot - \mu)$ denotes the density of a d-dimensional Gaussian.

Let's notice that for the special case identified by $\Sigma = \sigma^2 I$ we have that the negative log-likelihood, namely $-\log \pi(x)$, are strongly convex with $\alpha = 1/\sigma^2$; if we think about a Gaussian we know that as the variance increases the density becomes less concentrated with fatter tails. Between log-concave distributions and strongly log-concave ones there exists an intermediate class: log-concave distributions of order p.

Definition 17. *A log-concave distribution of order $p \geq 1$ is identified by a density function of the form $\pi(x) = x^{p-1}g(x)$, $x > 0$, where g is log-concave on $\mathbb{R}_+$.*

It can be shown that log-concave distributions of order $p > 1$ are more concentrated than simple log-concave ones. Examples of such distributions are: the Gamma with

$\alpha \geq 1$ degrees of freedom is log-concave of order α, while the Beta with parameter $\alpha \geq 1$ and $\beta \geq 1$ is still log-concave of order α.

A.2 Spectral Theory

In linear algebra particular emphasis is given to the spectral theorem, which is an instance of the Singular Value Decomposition. The theorem states that, if we have an Hermitian matrix A (or a symmetric matrix), then we can diagonalize it through a change of basis. Each element on the diagonal represents a real eigenvalue of A; the new basis is an orthonormal basis of eigenvectors of A. Thus, A can be written as a linear combination of pairwise orthogonal projections, called its spectral decomposition. Formally, given a symmetric matrix A, representing a linear map $f : V \to V$, with eigenvalues $\lambda_1, ..., \lambda_n$ and corresponding eigenvectors $v_1, ..., v_n$, we can write it as $A = \sum_{i=1}^{n} \lambda_i E_i$, where $E_i = v_i v_i^T$, with $||v_i|| = 1 \ \ \forall i$, is the projection matrix onto the eigenspace associated with λ_i; these eigenspaces are orthogonal, thus, we can decompose the vector space V on which the matrix A acts as $V = \oplus_{i=1}^{n} V_i$, where $V_i = \{v \in V : Av = \lambda_i v\}$. The properties of such projections matrices are the following: $E_i E_j = 0 \ \ \forall i \neq j$ due to orthogonality, $E_i^2 = E_i$ and $\sum_{i=1}^{n} E_i = Id$.
Therefore, when we apply a symmetric matrix A to some vector $v \in V$, we can decompose its effect as a linear combination of the effects of the single eigenvalues. Furthermore, diagonalizing a matrix offers several advantages, e.g. computing $A^k, \ e^A$. When moving to an infinite-dimensional setting, the theory gets a little bit more intricate. However, the general idea is the same. Instead of talking about diagonalization we talk about spectral decomposition.

Definition 18. *Given a linear operator A on a Banach space $\mathcal{B}$, the resolvent set $\rho(A)$ of A is the set of all complex values $\lambda \in \mathbb{C}$ such that the range of $\lambda Id - A$ is dense in $\mathcal{B}$ and such that $\lambda Id - A$ has a bounded inverse $R_\lambda = (\lambda Id - A)^{-1}$, called the resolvent operator. Then the spectrum $\sigma(A)$ of A is $\mathbb{C} \setminus \rho(A)$.*

This spectrum is made by three components: the *point spectrum*, which contains the eigenvalues of A, namely those values of λ such that there exists a non-zero solution $f \in \mathcal{B}$ to $Af = \lambda f$, where f is an eigenfunction; the *continuous spectrum*, made by those values of λ for which the range of $\lambda Id - A$ is dense in $\mathcal{B}$ but the inverse is not

bounded; the *residual spectrum* which contains the values of λ which do not satisfy neither of the above conditions. It turns out that for self-adjoint operators on separable Hilbert spaces the spectrum, namely the values of λ, is always real and the residual spectrum is empty. How do we represent the spectral decomposition of such an operator? Given a positive self-adjoint operator A on a real, separable Hilbert space $\mathcal{H}$, a spectral decomposition is an increasing family of closed linear subspaces of $\mathcal{H}$, namely $(\mathcal{H}_\lambda)_{\lambda\geq 0}$. This family is required to be right-continuous, i.e. $\bigcap_{\lambda'>\lambda}\mathcal{H}_{\lambda'} = \mathcal{H}_\lambda$ and $\bigcup_{\lambda\geq 0}\mathcal{H}_\lambda$ is dense in $\mathcal{H}$. Then, as we did in the finite-dimensional case, for every $\lambda \geq 0$ we consider the orthogonal projection P_λ onto $\mathcal{H}_\lambda$. The finite sum becomes an integral, as shown by the following theorem.

Theorem 13. *Given a positive self-adjoint operator A on $\mathcal{H}$, there exists a spectral decomposition $(E_\lambda)_{\lambda\geq 0}$ such that*

$$A = \int_0^\infty \lambda \mathbf{d}E_\lambda.$$

Furthermore, for any measurable function ψ on $\mathbb{R}_+$, one may define the operator $\psi(A)$ as

$$\psi(A) = \int_0^\infty \psi(\lambda)\mathbf{d}E_\lambda,$$

on functions $f \in \mathcal{H}$ such that $||\psi(A)f||^2 = \int_0^\infty \psi(\lambda)^2 \mathbf{d}\langle E_\lambda f, f\rangle < \infty$, where we have the notation $||E_\lambda f||^2 = \langle E_\lambda f, f\rangle$.

It is quite clear that this is just a generalization of the formula $A = \sum_{i=1}^n \lambda_i E_i$. The range of integration is positive due to the fact that a positive operator has positive spectrum.

How does this apply to diffusion processes we work with? Well, in our case the infinitesimal generator, L, is a negative operator; thus, its opposite, namely -L, is a positive self-adjoint operator on the Hilbert space $\mathbb{L}^2(\mu)$. Its spectral decomposition is $-L = \int_0^\infty \lambda \mathrm{d}E_\lambda$. As a consequence of the previous theorem we have that $P_t f = \int_0^\infty e^{-\lambda t}\mathrm{d}E_\lambda f$, for every $f \in \mathbb{L}^2(\mu)$ and $t \geq 0$. Finally, note that if the spectrum of A is discrete, namely made only by normal eigenvalues having finite-dimensional eigenspaces, then we can express the Hilbert space as $\mathbb{L}^2(\mu) = \bigoplus_{k\in\mathbb{N}} E_k$, where $(E_k)_k$ is a sequence of closed subspaces of $\mathbb{L}^2(\mu)$ such that E_k is the eigenspace

associated with $-\lambda_k$, $\lambda_k \geq 0$. Then we decompose $f \in \mathbb{L}^2(\mu)$ as $f = \sum_{k\in\mathbb{N}} f_k$ and $-Lf = \sum_{k\in\mathbb{N}} \lambda_k f_k$, $P_t f = \sum_{k\in\mathbb{N}} e^{-\lambda_k t} f_k$.

A.3 Stochastic Calculus

Stochastic processes in continuous-time are often defined on infinite-dimensional spaces. For example, the canonical definition of the Brownian Motion, $(B_t)_{t\geq 0}$ with $B_t \in \mathbb{R}^d$, is on the probability space $(\Omega, \mathscr{F}, \mathbb{P}) = (\mathcal{C}(\mathbb{R}_+;\mathbb{R}^d), \mathscr{B}(\mathcal{C}(\mathbb{R}_+;\mathbb{R}^d)), \mathbb{W})$, where $\mathbb{W}$ is the Wiener measure, a probability measure on the path space of continuous trajectories $\mathcal{C}(\mathbb{R}_+;\mathbb{R}^d)$, so that

$$\mathbb{W}(A) = \mathbb{P}(t \mapsto B_t(\omega) \in A), \ \ \forall A \in \mathscr{B}(\mathcal{C}(\mathbb{R}_+;\mathbb{R}^d)).$$

Then, we specialize the measurable sets to cylinder sets because they allow us to correctly define the Wiener measure via the finite-dimensional distributions of Brownian motion on such sets, thanks to the Kolmogorov extension theorem.
Girsanov's theorem enables us to define the Radon-Nikodym derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ between two path measures so that if we change the measure we obtain a new process with different drift. Let's state the theorem.

Theorem 14. *Let $(B_t)_{t\in[0,T]}$ be a (standard) Brownian Motion under the Wiener measure $\mathbb{W}$ and let $(b_t)_{t\in[0,T]}$ be a progressive process with finite variance, $\mathbb{E}^{\mathbb{W}}\left(\int_0^T \|b_s\|^2 ds\right) < \infty$. Let $M_t := \int_0^t \langle b_s, dB_s\rangle$ for $t \in [0,T]$ denote a continuous local martingale and let $[M,M]_t := \int_0^t \|b_s\|^2 ds$ denote its quadratic variation. Let's define the exponential martingale*

$$\mathcal{E}(M) := \exp\left(M - \frac{1}{2}[M,M]_t\right).$$

Assume that $\mathcal{E}(M)$ is a $\mathbb{W}$-martingale and define the path measure $\mathbb{P}$ as

$$\frac{d\mathbb{P}}{d\mathbb{W}} = \mathcal{E}(M)_T.$$

Then, under $\mathbb{P}$ the map

$$t \mapsto \tilde{B}_t := B_t - [B,M]_t = B_t - \int_0^t b_s ds,$$

is a (standard) Brownian Motion.

Therefore, if we change the path measure to $\mathbb{P}$ we introduce the drift component $\int_0^t b_s \mathrm{d}s$.