ANALYSIS OF LANGEVIN DIFFUSIONS

ROHAN BULUSWAR

Abstract. This expository paper presents a construction of the Langevin diffusion process and both analytic and geometric perspectives on its convergence. First, it introduces stochastic calculus, culminating with Itô processes and Itô's formula. Second, it introduces Markov semigroup theory and several prominent functional inequalities, which serve to quantify the rate of convergence to the stationary distribution. Finally, it introduces optimal transport theory and Otto calculus to give an alternate interpretation of Langevin diffusion in Wasserstein space and help illuminate the 'curvature-dimension' condition in Bakry-Émery theory.

Contents

1. Introduction

This paper presents an overview and analysis of the Langevin diffusion process, defining it from first principles as a stochastic differential equation, obtaining results

about long-term behavior using Markov semigroup theory, and showing how it can also be viewed as a gradient flow in a pseudo-manifold of probability distributions.

Section 2 focuses on rigorously defining the Langevin SDE (stochastic differential equation). Familiarity with the basic notions of stochastic processes is expected - readers unfamiliar with any of these can consult Appendix A Stochastic Processes. We begin with an informal construction of Brownian motion, then move to stochastic integration and Itô's formula. Finally, we define the Langevin SDE.

Section 3 presents Markov semigroup theory, which views stochastic processes from the perspective of functional analysis. To begin, we define the important operators associated with a Markov semigroup and show how we obtain exact results about the evolution of the law of continuous-time Markov processes. Finally, we describe two important inequalities bounding the mixing times of Markov processes, and the conditions under which they hold.

Finally, Section 4 presents a geometric perspective on the Langevin diffusion. Some familiarity with Riemannian manifolds and Ricci curvature is expected, but a brief introduction can be found in Appendix B Riemannian Manifolds and Ricci Curvature. First, we relate curvature and dimension to the inequalities on mixing times of Markov processes and show the conditions under which they hold for the Langevin diffusion. Next, we introduce optimal transport theory and explain how it is used to construct a pseudo-manifold of probability distributions, leading to the creation of Otto calculus. We present proofs about functional equalities that utilize Otto calculus, and finally, describe how the Langevin diffusion can be viewed as a gradient flow in the pseudomanifold of distributions.

2. Stochastic Calculus and the Langevin SDE

2.1. **Markov Processes and Brownian Motion.** The focus of this paper is on a class of stochastic process called Markov processes. These are the 'memoryless' processes, in which future values of X_t depend only on the current value, and not past values, as captured in the following definition [BGL14, 8].

Definition 2.1 (**Markov property**)**.** Given a stochastic process (X_t), we say it has the Markov property if for any $A \in \mathcal{F}$ and any $s > t$, $\mathbb{P}(X_s \in A|\Sigma_t) = \mathbb{P}(X_s \in A|X_t)$.

The technical details of conditional probability and conditional expectation can be found in [Lawler23, 1.1]. In other words, the law of X_s conditioned on Σ_t is the same as conditioned on X_t. Moreover, we call the process time-homogenous if the law of X_s conditioned on X_t is the same as the law of X_{s-t} conditioned on X_0 for any s, t. The Markov processes studied in this paper are generally time-homogenous. Critically, we can then fully capture the information of the Markov process in the transition kernels (probability measures) $p_t(x, dy)$ where $\mathbb{P}(X_t \in A|X_0 = x) = \int_A 1p_t(x, dy)$. This allows for another interpretation of the Markov property [BGL14, 8].

Proposition 2.2 (**Weak Markov property**)**.** *If the process X_t satisfies the Markov property, then for any times $0 < t_1 \leq t_2 \leq ... \leq t_k$, the law of the random vector $(X_{t_1}, .., X_{t_k})$ given that $X_0 = x$ is*

$$p_{t_1}(x, dy_1)p_{t_2-t_1}(y_1, dy_2)...p_{t_k-tk-1}(y_{k-1}, dy_k).$$

The proof is left to the reader, as this is a fairly standard result. In the case of discrete Markov processes, we can only take the t_i to be natural numbers and this is in fact equivalent to the Markov property [Berestycki23, 10].

One basic example of a Markov property is the random walk. We could take a random walk on many different structures, but let us begin with $\mathbb{Z}$.

Definition 2.3. The simple random walk on $\mathbb{Z}$ is a discrete-time stochastic process $(X_n)_{n\in\mathbb{N}}$, given by $X_n = \sum_{i=1}^{n} S_i$, where the S_i are i.i.d random variables with $\mathbb{P}(S_i = 1) = \frac{1}{2} = \mathbb{P}(S_i = -1)$.

Proposition 2.4. *The simple random walk on $\mathbb{Z}$ is a Markov process.*

Proof. Using Definition 2.1, it suffices to show that for any integer z and time $t \in \mathbb{N}$,

$$\mathbb{P}(X_t = z|X_1, ..., X_{t-1}) = \mathbb{P}(X_t = z|X_{t-1})$$

There are two cases. First, suppose that $X_{t-1} \notin \{z-1, z+1\}$. Then it is impossible so that $X_t = z$, since $X_t - X_{t-1} \in \{-1, 1\}$ by construction. Hence, both the LHS and RHS are equal to 0. Next, suppose that $X_{t-1} \in \{z-1, z+1\}$, and let us suppose WLOG that $z - X_{t-1} = 1$. By construction, the RHS is equal to $\mathbb{P}(X_t = z|X_{t-1} = z-1) = \mathbb{P}(S_t = 1) = \frac{1}{2}$. For the LHS, we again exactly require $S_t = 1$, and S_t is assumed to be independent of all the X_i, so we can write

$$\frac{1}{2} = \mathbb{P}(S_t = 1) = \mathbb{P}(S_t = 1|X_1, ..., X_{t-2}) = \mathbb{P}(S_t = 1|X_1, ..., X_{t-2}, X_{t-1} = z-1)$$

$$\frac{1}{2} = \mathbb{P}(X_t = z|X_1, ..., X_{t-1} = z-1)$$

Hence, in either case, the two sides of the equation are the same, and the simple random walk X_t is a Markov process. □

This proof captures the intuitive notion that randomly moving particles, or colloquially drunkards, do not remember where they have been. The idea of the proof is not specific to $\mathbb{Z}$ - one can also show that a random walk on a graph, for example, is also a Markov chain. Random walks are useful for modeling real-world phenomena, from the fluctions of the prices of equities and their options to the movement of particles in a fluid impacted by many nearly-random collisions. Moreover, they are used to define the core class of processes on which we perform stochastic calculus.

However, for the purposes of modeling real-world systems, we prefer that our random walk be continuous in space and time. We will thus begin an informal construction of such an object, called Brownian motion or a Wiener process.

While the simple random walk takes values in $\mathbb{Z}$, we can just as well view it as taking values in $\mathbb{R}$, with a step size of 1. This step size is an arbitrary positive real, and will be denoted by Δx. Moreover, while it is discrete in time, we may assign a time increment of Δt, initially assumed to be 1. In this way, the simple random walk is akin to taking snapshots of some continuous-time process every Δt seconds. Hence, as we send $\Delta x \to 0$ and $\Delta t \to 0$, we expect to recover that continuous-time process: a continuous random walk with time values in $\mathbb{R}_{\geq 0}$ and position values in $\mathbb{R}$. The method of taking limits here must be done with some care, and the technical details are difficult. In fact, it is a priori unclear if such a stochastic process even exists. However, let us begin by describing the properties we require it to have, in order to be a good model of a continuous random walk.

Definition 2.5 (**Standard Brownian motion**). Standard Brownian motion (in one dimension) is a continuous-time stochastic process B_t taking values in $\mathbb{R}$ satisfying the following properties:

- $B_0 = 0$
- With probability one, the function $B(t) = B_t$ is continuous
- For any $0 < t_1 < t_2 < ... < t_k$, the random variables $B_{t_1}, B_{t_2} - B_{t_1}, ...,$ and $B_{t_k} - B_{t_{k-1}}$ are independent of each other
- For any $0 < t_1 \leq t_2$, $B_{t_2} - B_{t_1} \sim N(0, t_2 - t_1)$

The first property, that B_t starts at the origin, is mainly for convenience. We can always add some constant to obtain the process $c + B_t$, a continuous random walk from a different origin. The second property, that $B(t)$ is continuous, is the necessary level of niceness to perform stochastic calculus, and makes intuitive sense for the purposes of modeling. The third property, called independent increments, captures the notion that the steps at different times are independent of each other. In the case of the simple random walk, this was the assumption that the S_i are independent.

The fourth property, describing the distribution of the increments, has two implications. First is that increments of the same size have the same distribution, since $N(0, t_2 - t_1)$ only depends on the difference between t_2 and t_1, not their individual values. In the case of the simple random walk, this was the assumption that the S_i have the same distribution. Second, it prescribes that the distributions of the increments are normal with mean zero and linearly increasing variance. In fact, the previous properties are enough to conclude that the increments satisfy $B_{t_2} - B_{t_1} \sim N(\mu(t_2 - t_1), \sigma^2(t_2 - t_1))$ for some $\mu \in \mathbb{R}$, $\sigma^2 \geq 0$, but we do not provide a proof here [Lawler23, 44]. In the case of the standard Brownian motion, we have $\mu = 0$ and $\sigma^2 = 1$. To recover the general form, if B_t is the standard Brownian motion, we can consider $X_t = \sigma B_t + \mu t$. (The process X_t will still satisfy properties 2-4.) μ is called the drift and σ^2 is called the variance. This distribution of increments should make sense from the properties of the simple random walk. Because the S_i are independent, if they have mean μ and variance σ^2, $X_n = \sum_{i=1}^{n} S_i$ will have $\mathbb{E}(X_n) = n\mu$ and $\mathrm{Var}(X_n) = n\sigma^2$. Thus, the mean and variance of X_n grow linearly in time. Moreover, in considering the limit of random walks with finer step size, we are in fact taking an average of an increasing number of S_i, sampled from the same distribution. By the Central Limit Theorem, we expect the result to be normally distributed.

Although we have a precise definition of standard Brownian motion, the challenge of constructing such a stochastic process remains. We have informally thought of it as a limit of random walks as $\Delta x, \Delta t \to 0$, but much clarification is needed. A complete proof would be onerous, but there are important ideas to outline. First, we need to understand how Δx and Δt move together as they approach zero. Recall that for standard Brownian motion, we want $B_1 \sim N(0, 1)$. Let us assume that Δt is chosen so that $1 = N\Delta t$ for some natural N, and X_N occurs at $t = 1$. Then let us consider $X_N = \sum_{i=1}^{N} S_i$, where S_i takes on the values Δx and $-\Delta x$ with equal probability. No matter how Δx is chosen, $\mathbb{E}(S_i) = 0$, so $\mathbb{E}(X_N) = N\mathbb{E}(S_i) = 0$. However, $\mathrm{Var}(X_N) = N\mathrm{Var}(S_i) = N(\Delta x)^2$, and we require $\Delta x = \sqrt{\frac{1}{N}} = \sqrt{\Delta t}$. This property has consequences for the most important and computationally useful formulas of stochastic calculus. Notably, any approximation in time at the level of

Δt must be accurate at the level of $(\Delta x)^2$. We will thus find second-order Taylor expansions where ordinary calculus would only have one. As another consequence, $\frac{\Delta B_t}{\Delta t} \approx \frac{1}{\Delta t}$, which tends to infinity as $\Delta t \to 0$. One implication is that with probability one, Brownian motion is nowhere differentiable [Lawler23, 51].

Theorem 2.6. *With probability one, the function $B(t) = B_t$ is nowhere differentiable.*

Next, we give a brief outline of the method used to construct standard Brownian motion (the details will take too much space, but the idea is generally useful in studying stochastic process), as presented in detail in [Lawler23, 2.5]. We will in fact only construct B_t for $t \in [0, 1]$, but by repeating these one after another, beginning where the previous one ends, we obtain B_t for all $t \geq 0$. We begin with a countable set of independent standard normal random variables $\{Z_i\}$. We denote by $\mathcal{D}_n$ the set of rationals in $[0, 1]$ with denominator 2^n. The union $\mathcal{D} = \bigcup_{n\in\mathbb{N}} \mathcal{D}_n$ is the set of dyadic rationals in $[0, 1]$. $\mathcal{D}$ is countable, so for each $t \in \mathcal{D}$, we associate it with one of our standard normal variables Z_t. We can use these random variables to define B_t, our standard Brownian motion, for $t \in \mathcal{D}$, recursively in n for each $\mathcal{D}_n$. By beginning with $\mathcal{D}_1$ and then defining B_t for each $t \in \mathcal{D}_{n+1} \setminus \mathcal{D}_n$ for all $n \in \mathbb{N}$, we associate a unique random variable B_t to each $t \in \mathcal{D}$. This process $B_t, t \in \mathcal{D}$ satisfies every property of Brownian motion except continuity. However, because $\mathcal{D}$ is dense in $[0, 1]$, knowing that the full B_t ought to be continuous, we should have enough information. We can prove that the function $B(t) = B_t$ defined intially only on $\mathcal{D}$ is (almost surely) uniformly continuous. Hence, for each $t \in [0, 1]$, we can take $t_n \to t$ where $t_n \in \mathcal{D}$, and define $B_t = \lim_{n\to\infty} B_{t_n}$. Finally, one can check that this definition of B_t satisfies all of the necessary properties.

Consequently, we know that Brownian motions, or random walks that are continuous in space and time, can exist. Like the simple random walk, Brownian motion is a Markov process, due to having independent and identically distributed increments. Finally, with a grasp of Brownian motion in one dimension, we can extend it to taking values in $\mathbb{R}^d$, its full generality. We have the following definition [Lawler23, 67].

Definition 2.7 (General Brownian motion)**.** Consider $\mu \in \mathbb{R}^d$, and let $\Gamma \in \mathbb{R}^d \times \mathbb{R}^d$ be a symmetric positive semi-definite matrix. Then a Brownian motion in $\mathbb{R}^d$ with drift μ and covariance Γ is a stochastic process $B_t = (B_t^1, B_t^2, ..., B_t^d)$ satisfying the following properties:

- $B_0 = 0$
- With probability one, the function $B(t) = B_t$ is continuous
- For any $0 < t_1 < t_2 < ... < t_k$, the random vectors $B_{t_1}, B_{t_2} - B_{t_1}, ..., B_{t_k} - B_{t_{k-1}}$ are independent of each other
- For any $0 < t_1 \leq t_2$, $B_{t_2} - B_{t_1}$ is normally distributed with mean $(t_2 - t_1)\mu$ and covariance $(t_2 - t_1)\Gamma$

The definition is essentially the same as in one dimension, adapted to the definition of the normal distribution in multiple dimensions. In fact, there is no difficulty in constructing multidimensional Brownian motion because its components are themselves Brownian motions, though they are not necessarily independent.

2.2. **Stochastic Integration.** By constructing Brownian motion we have laid the foundations of a more general class of 'random walk' stochastic processes, in which

the drift and variance may vary in time and space. The end goal, however, is to make sense of stochastic differential equations, and in particular, the Langevin equation. But how can we have differential equations when Brownian motion is almost certainly differentiable nowhere? The answer is that SDEs are defined in terms of integrals, and in particular, integrals against a Brownian motion. Hence, we focus next on the development of stochastic integration.

The goal of stochastic integration is to make sense of $\int X_t\, dB_t$, the integration of a continuous-time stochastic process against Brownian motion. If the Brownian motion B_t represents the random movements of a stock and X_t represents bets placed on the movement of that stock, then the integral $\int X_t\, dB_t$ is representative of the total profit made from these bets. To clarify this idea, let us return to the simpler case of 'discrete stochastic integration' with respect to a simple random walk. To begin, we introduce the following definition [Lawler23, 26].

Definition 2.8. Let X_n be discrete-time stochastic process with its natural filtration $\{\Sigma_n\}$. A sequence of random variables Y_n is called predictable if for all $n \in \mathbb{N}$, Y_n is measurable with respect to Σ_{n-1}.

This is a reasonable requirement because we would like the value of the integral $\int_0^T X_t dB_t$ to depend only on the values of B_t up to time T, and the same should be true for the discrete scenario. Now, we can formulate discrete stochastic integration.

Definition 2.9 (**Discrete stochastic integration**)**.** Let $S_1, S_2, ...$ be independent and identically distributed random variables with mean zero and variance σ^2. We then consider the stochastic process X_n by $X_n = \sum_{i=1}^n S_i$, a random walk, and its natural filtration. Moreover, let Y_n be a sequence of random variables that is predictable and satisfies $\mathbb{E}[Y_n^2] < \infty$ for all $n \in \mathbb{N}$. The discrete stochastic integral is defined by the process Z_n, where

$$Z_n = \sum_{i=1}^{n} Y_i S_i$$

If we assume that the S_i take on values of -1 and 1 with $\frac{1}{2}$ probability each, then X_n is exactly the simple random walk discussed earlier. However, we could also assume that the S_i are normally distributed with equal variance. In this case, X_n looks more like snapshots of a Brownian motion at equal time intervals. This intuition, like with constructing Brownian motion, will aid in constructing the full stochastic integral. First, however, it is worth noting some interesting properties of the discrete stochastic integral. To do so, it is worth introducing a new definition, for another important class of stochastic processes.

Definition 2.10 (Martingales)**.** Let X_t be a stochastic process with respect to a filtration $\{\Sigma_t\}$. Suppose that for all t, $\mathbb{E}[X_t^2] < \infty$. Then X_t is called a Martingale if for all pairs $t_1 < t_2$, $\mathbb{E}[X_{t_2}|\Sigma_{t_1}] = X_{t_1}$

This definition is valid for both discrete-time and continuous-time Martingales. It can be reformulated as $\mathbb{E}[X_{t_2} - X_{t_1}|\Sigma_{t_1}] = 0$. In other words, we require that absent relevant information, any jumps in X_t have expected value equal to 0. Hence, Martingales are also called 'fair games'. In fact, the discrete stochastic integral as defined above is a typical example of a Martingale:

Proposition 2.11. *Consider X_n, Y_n, and Z_n as in Definition 2.9. The discrete integral Z_n satisfies the following properties:*

- *Linearity*
- *Z_n is a Martingale*
- *$Var[Z_n] = \sigma^2 \sum_{i=1}^n \mathbb{E}[Y_i^2]$*

Proof. First, linearity of the integral follows immediately from linearity of expectation. Second, we can show that Z_n is a Martingale. To begin, note that $Z_n^2 = \sum_{i=1}^n \sum_{j=1}^n Y_i S_i Y_j S_j$. For any $i < j$, S_j is independent of S_i, Y_i, and Y_j, so that $\mathbb{E}[Y_i S_i Y_j S_j] = \mathbb{E}[S_j]\mathbb{E}[Y_i S_i Y_j] = 0\mathbb{E}[Y_i S_i Y_j] = 0$. ($Y_j$ is measurable w.r.t Σ_{j-1}, which is in turn independent from S_j, by definition of a random walk.) Hence, we are left with $\mathbb{E}[Z_n^2] = \mathbb{E}[\sum_{i=1}^n Y_i^2 S_i^2] = \sum \mathbb{E}[Y_i^2 S_i^2]$. Similarly, Y_i and S_i are independent, so this reduces to

$$\sum_{i=1}^n \mathbb{E}[Y_i^2]\mathbb{E}[S_i^2] = \sigma^2 \sum_{i=1} \mathbb{E}[Y_i^2]$$

By assumption, $\mathbb{E}[Y_k^2]$ is finite for each k, so $\mathbb{E}[Z_n^2]$ is also finite. Next, for any $i < j$ we need to show that

$$\mathbb{E}[Z_j - Z_i|\Sigma_i] = 0$$

To begin, we can write

$$\mathbb{E}[Z_j - Z_i|\Sigma_i] = \mathbb{E}[\sum_{k=i+1}^{j} Y_k S_k|\Sigma_i] = \sum_{k=i+1}^{j} \mathbb{E}[Y_k S_k|\Sigma_i]$$

As before, Y_k and S_k are independent, so $\mathbb{E}[Y_k S_k|\Sigma_i] = \mathbb{E}[Y_k|\Sigma_i]\mathbb{E}[S_k|\Sigma_i]$. Because $k > i$, S_k is independent of $S_1, S_2, ..., S_i$, and thus independent of Σ_i. Therefore, $\mathbb{E}[S_k|\Sigma_i] = \mathbb{E}[S_k] = 0$. Finally, we know that

$$\mathbb{E}[Z_j - Z_i|\Sigma_i] = \sum_{k=i+1}^{j} \mathbb{E}[Y_k|\Sigma_i]\mathbb{E}[S_k|\Sigma_i] = 0$$

Therefore, Z_n satisfies both of the requisite properties of a Martingale. Note that $Z_0 = 0$ because it is an empty sum, so for any n, $\mathbb{E}[Z_n] = \mathbb{E}[Z_n|\Sigma_0] = \mathbb{E}[Z_0|\Sigma_0] = 0$. ($\Sigma_0$ represents no information.) As a result, any betting strategy that cannot look into the future is expected to make no money trading stocks! Third, we can compute

$$\text{Var}[Z_n] = \mathbb{E}[Z_n^2] - \mathbb{E}[Z_n]^2 = \sigma^2 \sum_{i=1}^n \mathbb{E}[Y_i^2] - 0 = \sigma^2 \sum_{i=1}^n \mathbb{E}[Y_i^2]$$

where the first equality follows from our earlier calculations. □

While the proof itself relied on working in discrete-time, once we construct the continuous stochastic integral, we will find that many of the same properties still hold. The full technical details of the construction are beyond the scope of this paper, but we will give a brief summary. In the beginning, defining the stochastic integral is much like defining the Lebesgue integral. Akin to simple functions, we have the notion of a simple process: processes which are constant except for jumps at some fixed (finite in number) times $t_1, ..., t_n$. It is fairly straightforward to define the integral for a simple process, as it is essentially a discrete-time process with time intervals inserted: the details can be found in [Lawler23, 3.2]. Then, we consider well-behaved processes, and take an approximating sequence of simple processes that converges in the L^2 sense. We can define the integral of the original process

to be the limit of the integral of the simple processes, which can be proven to exist with probability one. With that definition, one can prove the following proposition [Lawler23, 90-91].

Proposition 2.12. *Let B_t be a standard Brownian motion with its natural filtration $\{\Sigma_t\}$. Consider processes A_t and C_t such that: for all t, A_t and C_t are measurable w.r.t Σ_t (i.e. they are adapted to $\{\Sigma_t\}$); A_t and C_t have continuous paths with probability one; there exists a constant M such that with probability one, $|A_t|, |C_t| \leq M$ for all t. If the stochastic integral is denoted by $Z_t = \int_0^t A_s dB_s$, then the following hold:*

- *Linearity: $\int_0^t (aA_s + cC_s)dB_s = a\int_0^t A_s dB_s + c\int_0^t C_s dB_s$ and $\int_0^t A_s dB_s = \int_0^r A_s dB_s + \int_r^t A_s dB_s$*
- *Martingale: Z_t is a Martingale, and in particular $\mathbb{E}[Z_t] = 0$ for all t*
- *Continuous paths: with probability one, the function $f(t) = Z_t$ is continuous*
- *Itô Isometry: $\mathbb{E}[(\int_0^t A_s dB_s)^2] = \int_0^t \mathbb{E}[A_s^2]ds$*

Note that the Itô Isometry is the equivalent of the variance calculations from the earlier proposition about discrete stochastic integration. While the continuity property is unique to continuous-time integration, the other two properties also carry over from before, as promised. Next, we relax the assumptions on A_t to define the stochastic integral for a larger class of processes. For example, A_t does not need to be bounded, although Z_t may no longer be a Martingale [Lawler23, 93]. Moreover, A_t may only be piecewise continuous. More details and broader definitions of the stochastic integral can be found in [Lawler23, 3.1-3.3].

2.3. **Itô's Formula.** Now, equipped with a notion of stochastic integration, we can begin to make sense of stochastic differential equations, and their solutions. There is, however, a more pressing question: given that the definition of stochastic integration is somewhat abstruse, how does one actually compute these integrals? For example, what is $\int_0^t B_s dB_s$? The main tool for answering these questions, analogous to Taylor's theorem or the fundamental theorem of calculus, is Itô's formula.

To begin, we consider a particular class of stochastic processes called Itô processes [Chewi24, 8].

Definition 2.13 (**Itô processes**)**.** Let B_s be a standard Brownian motion in $\mathbb{R}^N$. A stochastic process X_t with values in $\mathbb{R}^d$ is called an Itô process if it can be written as

$$X_t = \int_0^t b_s\, ds + \int_0^t \sigma_s\, dB_s$$

where b_s is a vector-valued stochastic process in $\mathbb{R}^d$ and σ_s is a matrix-valued stochastic process in $\mathbb{R}^{d\times N}$. (We assume that b_s and σ_s are such that the integrals are well-defined.) Alternatively, we may formally write the stochastic differential equation

$$dX_t = b_t dt + \sigma_t dB_t$$

In other words, an Itô process moves like a Brownian motion with changing mean (b_t) and variance (σ_t). These coefficients are called the drift and diffusion, respectively. Itô's formula (also called Itô's lemma) allows us to compute differentials of a function of an Itô process as follows [Chewi24, 9].

Theorem 2.14 (Itô's formula). *Let $f(t,X):\mathbb{R}_{\geq 0}\times\mathbb{R}^d\to\mathbb{R}$ be a function that is C^1 in t and C^2 in X. Moreover, let X_t be an Itô process satisfying $dX_t = b_t dt + \sigma_t dB_t$, as in Definition 2.13. Then $f(t,X_t)$ is also an Itô process, and it satisfies*

$$f(t,X_t)-f(0,X_0) = \int_0^t \partial_s f(s,X_s)+\langle\nabla f(s,X_s),b_s\rangle+\frac{1}{2}\langle\nabla^2 f(s,X_s),\sigma_s\sigma_s^T\rangle\,ds+\int_0^t\langle\sigma_s^T\nabla f(s,X_s),dB_s\rangle$$

A brief aside on notation: for matrices A and B of the same dimension, we define $\langle A,B\rangle := \mathrm{Tr}(AB^T)$. Moreover, $\nabla^2 f$ is used to denote the Hessian of f. Finally, if $v_s=(v_s^1,...,v_s^N)$ is a vector-valued stochastic process of dimension N (as $\sigma_s^T\nabla f(s,X_s)$ is) and $B_s=(B_s^1,...B_s^N)$ is a standard Brownian motion in $\mathbb{R}^N$, then we define $\int_0^t\langle v_s,dB_s\rangle=\sum_{i=1}^N\int_0^t v_s^i dB_s^i$.

We will generally focus on functions f that are constant in time, so that the term $\partial_s f(s,X_s)$ disappears, but the others remain. It is notable that in normal calculus, this result would only involve integrating the derivative, a first order approximation. Here, however, we have the Hessian $\nabla^2 f(s,X_s)$. This is because we know that in time δt, a Brownian motion experiences a movement on the order of $\delta x\approx\sqrt{\delta t}$. Hence, $(\delta x)^2\approx\delta t$, and second order terms are nonnegligible on the scale of δt. We will not prove the full extent of Itô's formula, but can develop some intuition by giving a rough argument for a less general version.

The following argument is presented in [Lawler23, 100-102]. We consider the case of $X_t=B_t$ (a standard Brownian motion in one dimension) and a function f that is constant in time. We claim that

$$f(B_t)-f(B_0)=\int_0^t f'(B_s)dB_s+\frac{1}{2}\int_0^t f''(B_s)ds$$

Without loss of generality, it suffices to consider $t=1$. Then we can write, for arbitrarily large $n\in\mathbb{N}$:

$$f(B_1)-f(B_0)=\sum_{i=1}^n f(B_{\frac{i}{n}})-f(B_{\frac{i-1}{n}})$$

By ordinary Taylor expansion, we can approximate

$$f(B_{\frac{i}{n}})-f(B_{\frac{i-1}{n}})=f'(B_{\frac{i-1}{n}})(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})+\frac{1}{2}f''(B_{\frac{i}{n}})(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})^2+o((B_{\frac{i}{n}}-B_{\frac{i-1}{n}})^2)$$

Summing up and taking the limit, we find that

$$f(B_1)-f(B_0)=\lim_{n\to\infty}\sum_{i=1}^n f'(B_{\frac{i-1}{n}})(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})+\lim_{n\to\infty}\sum_{i=1}^n\frac{1}{2}f''(B_{\frac{i}{n}})(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})^2+\lim_{n\to\infty}\sum_{i=1}^n o((B_{\frac{i}{n}}-B_{\frac{i-1}{n}})^2)$$

First, recall that $\delta x=(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})\approx\sqrt{\delta t}=\sqrt{\frac{1}{n}}$, so $(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})^2\approx\frac{1}{n}$. Thus, we can substitute to see that

$$f(B_1)-f(B_0)=\lim_{n\to\infty}\sum_{i=1}^n f'(B_{\frac{i-1}{n}})(B_{\frac{i}{n}}-B_{\frac{i-1}{n}})+\lim_{n\to\infty}\sum_{i=1}^n\frac{1}{2}f''(B_{\frac{i}{n}})\frac{1}{n}+\lim_{n\to\infty}\sum_{i=1}^n o(\frac{1}{n})$$

The last term is a summation of n terms of order smaller than $\frac{1}{n}$, so in the limit it is equal to 0. The second term is a Riemann sum approximation to the integral $\frac{1}{2}\int_0^1 f''(B_t)dt$, and the limit is precisely this integral. Finally, the first term is an approximation by a simple process to the stochastic integral $\int_0^1 f'(B_t)dB_t$. These

substitutions must be rigorously justified, but if the reader will believe them for now, we immediately recover this particular case of Itô's formula:

$$f(B_1) - f(B_0) = \int_0^1 f'(B_t)dB_t + \frac{1}{2}\int_0^1 f''(B_t)dt$$

Finally, we can return to computing $\int_0^t B_s dB_s$. Let us consider $f(x) = x^2$, and apply the simplified Itô's formula:

$$B_t^2 = \int_0^t 2B_s dB_s + \frac{1}{2}\int_0^t 2ds$$
$$2\int_0^t B_s dB_s = B_t^2 - t$$
$$\int_0^t B_s\, dB_s = \frac{1}{2}(B_t^2 - t)$$

Like in the argument for the simplified Itô's formula, we can use the intuition we have built to do unrigorous (formal) computations. The key fact is again that $(\Delta B_t)^2 \approx \Delta t$. Thus, Δt, ΔB_t, and $(\Delta B_t)^2$ are all relevant on the order of Δt, but $(\Delta t)^2$ and $(\Delta t)(\Delta B_t)$ are not, and can be treated as zero, much like $(\Delta x)^2$, $(\Delta x)^3$, etc. when computing derivatives in ordinary calculus. For example, suppose that X_t is a one-dimensional Itô process satisfying $dX_t = b_t dt + \sigma_t dB_t$ for $b_t \in \mathbb{R}$, $\sigma_t \in \mathbb{R}_{\geq 0}$. We may want to compute $\int_0^t (dX_t)^2$, which can be formalized as quadratic variation [Lawler23, 63].

Definition 2.15 (**Quadratic variation**)**.** Let X_t be a stochastic process. The quadratic variation of X is defined by

$$\langle X\rangle_t = \lim_{n\to\infty} \sum_{j=1}^{j\leq tn} [X_{\frac{j}{n}} - X_{\frac{j-1}{n}}]^2$$

We can write:

$$(dX_t)^2 = (b_t dt + \sigma_t dB_t)^2 = (b_t)^2(dt)^2 + 2b_t\sigma_t(dtdB_t) + (\sigma_t)^2(dB_t)^2$$

By substituting $(dt)^2 = 0$, $dtdB_t = 0$, and $(dB_t)^2 = dt$, we obtain

$$\int_0^t (dX_t)^2 = \int_0^t \sigma_s^2 ds$$

There are several ways in which Itô's formula can be used to do important calculations, as we will discover in the study of Markov semigroups.

2.4. **The Langevin SDE.** Finally, we may define the central object of study: the Langevin SDE. To begin, we consider a particularly important class of Itô processes called diffusion processes [Lawler23, 111].

Definition 2.16 (**Diffusion processes**)**.** A stochastic process X_t is called a diffusion process if it satisfies

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t$$

In other words, the movement of the process depends only on the time t and its current location. Consequently, diffusion processes have the Markov property. In particular, we study time-homogeneous processes, in which the functions μ and σ do not depend on t. Within diffusion processes, we can consider Langevin diffusions [Chewi24, ix-x].

Definition 2.17 (**Langevin diffusions**)**.** Let $V : \mathbb{R}^n \to \mathbb{R}$ be a C^2, strongly convex function. The Langevin diffusion is the stochastic process X_t taking values in $\mathbb{R}^n$ and satisfying

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$$

Intuitively, we think of X_t as a gradient flow with some noise (of constant variance). Although Langevin's work was motivated by random movements of particles in fluids, Langevin diffusion processes have found powerful applications in the realm of sampling algorithms. To access those applications, we must take an alternative view of stochastic processes as a whole, using functional analysis and in particular Markov semigroups, the focus of the next section.

3. Markov Semigroup Theory

3.1. **The Markov Semigroup and Related Operators.** Although formally a time-indexed collection of random variables, we often think about a stochastic process X_t as a particle moving through some state space over time. There is, however, another equally valid view. We can think about the probability density functions π_t of the random variables X_t, and how these functions evolve in time. Do they converge as t approaches infinity? If so, to what limit, and how fast is the convergence? Because the tools we have already seen (like Itô's formula) were constructed from the point of view of the particle, to answer these questions, we need to develop a new theory: that of Markov semigroups. As the name implies, the rest of this section, we only consider Markov processes. We begin with the following definition [Chewi24, 10].

Definition 3.1 (**Markov semigroups**)**.** Let X_t be a Markov process. The semigroup of operators $(P_t)_{t\geq 0}$ acts on measurable real-valued functions f (whose domain is the state space of X_t) and satisfies

$$P_t f(x) = \mathbb{E}[f(X_t)|X_0 = x]$$

Because Markov processes are fully defined by transition kernels, though it is somewhat sloppy language, we can talk about one Markov process X_t that may start at any X_0. Note that $P_0 = Id$ by definition, and the semigroup is abelian because $P_t P_s = P_s P_t = P_{t+s}$ [Chewi24, 10]. This fact can be demonstrated directly from the Markov property. As before, if we fix some x, we can assume that X_t begins at $X_0 = x$. First, by definition,

$$P_{t+s}f(x) = \mathbb{E}[f(X_{t+s})]$$

By the law of total expectation, we also know that

$$\mathbb{E}[f(X_{t+s})] = \mathbb{E}[\mathbb{E}[f(X_{t+s})|\Sigma_t]]$$

Substituting, we find

$$P_{t+s}f(x) = \mathbb{E}[\mathbb{E}[f(X_{t+s})|\Sigma_t]]$$

Because X_t is a Markov process we can simplify to

$$P_{t+s}f(x) = \mathbb{E}[\mathbb{E}[f(X_{t+s})|X_t]]$$

By definition of P_s, we have

$$P_{t+s}f(x) = \mathbb{E}[P_s f(X_t)]$$

Finally, by definition of P_t, this is just

$$P_{t+s}f(x) = P_t(P_s f)(x)$$

Hence, the two operators P_{t+s} and P_tP_s agree for any function f and input value x. Moreover, $P_{t+s} = P_{s+t}$ so the semigroup is indeed abelian.

In fact, we can give a definition of a Markov semigroup that does not make reference to any previously defined Markov process.

Definition 3.2 (**Markov semigroups, II**)**.** Let $(P_t)_{t\in\mathbb{R}_{\geq 0}}$ be a family of operators acting on real-valued, bounded, measurable functions f with domain of the measurable space $(E, \mathcal{F})$. Moreover, let μ be a σ-finite measure on E, and suppose that the following hold:

- For every bounded positive measurable function $f : E \to \mathbb{R}$ and every $t \geq 0$, $\int_E P_t f d\mu = \int_E f d\mu$ (μ is invariant)
- For every $t \geq 0$, P_t is a linear operator mapping bounded measurable functions to bounded measurable functions
- $P_0 = \mathrm{Id}$
- $P_t(1) = 1$, where $f = 1$ is a constant function
- If $f \geq 0$ then $P_t f \geq 0$
- For all $t, s \geq 0$, $P_{t+s} = P_t \circ P_s$
- For every $f \in L^2(\mu)$, $P_t f \to f$ converges in the L^2 norm

Then (P_t) is called a Markov semigroup of operators.

Hence, study can be done of Markov semigroups without reference to Markov processes: one is not subordinate to the other. In fact, when a Markov semigroup is derived from Markov process, it fully captures the original information [Chewi24, 11]. For the rest of this paper, however, we will only be concerned with Markov semigroups as a tool to analyze Markov processes.

Next, we move to another important operator in the theory of Markov semigroups, which allows us to take derivatives [Chewi24, 11].

Definition 3.3 (**Infinitesimal generator**)**.** Let (P_t) be the semigroup associated with some Markov process, and let f be a function (such that the following limit exists). The infinitesimal generator $\mathscr{L}$ is defined by

$$\mathscr{L} f = \lim_{t \searrow 0} \frac{P_t f - f}{t}$$

There are some technical details regarding the domain of this operator, but for now it suffices to note that if we take $L^p(\mu)$ to be the domain of the semigroup P_t (for some reference measure μ and $1 \leq p < \infty$) then it admits a dense linear subspace on which $\mathscr{L}$ is defined and is a linear operator [BGL14, 18]. As promised, the definition of $\mathscr{L}$ has the structure of a time derivative of the semigroup. For the sake of example, we can compute the generator of a Langevin process [Chewi24, 11]. Let

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$$

For any test function f (which is assumed to be constant in time), by Theorem 2.14, we can compute

$$f(X_t) - f(X_0) = \int_0^t \langle \nabla f(X_s), -\nabla V(X_s)\rangle + \frac{1}{2}\langle \nabla^2 f(X_s), \sqrt{2}I_n\sqrt{2}I_n^T\rangle ds + \int_0^t \langle \sqrt{2}I_n^T \nabla f(X_s), dB_s\rangle$$

$$f(X_t) - f(X_0) = \int_0^t -\langle \nabla f(X_s), \nabla V(X_s)\rangle + \mathrm{Tr}(\nabla^2 f(X_s))ds + \sqrt{2}\int_0^t \langle \nabla f(X_s), dB_s\rangle$$

Because B_s is a standard Brownian motion in $\mathbb{R}^n$, dB_s is mean-zero. It is also independent from X_s and thus $\nabla f(X_s)$, so that

$$\mathbb{E}[\langle \nabla f(X_s), dB_s \rangle] = 0$$

and thus

$$\mathbb{E}[\sqrt{2}\int_0^t \langle \nabla f(X_s), dB_s \rangle] = 0$$

Because we are computing a first time-derivative, approximation up to error $o(t)$ is sufficient. We can use ordinary Taylor expansion to write:

$$\int_0^t -\langle \nabla f(X_s), \nabla V(X_s) \rangle + \mathrm{Tr}(\nabla^2 f(X_s)) ds = t[-\langle \nabla f(X_0), \nabla V(X_0) \rangle + \mathrm{Tr}(\nabla^2 f(X_0))] + o(t)$$

Conditioned on $X_0 = x$, the LHS is entirely constants. Thus, we can write

$$\begin{aligned}
\mathbb{E}[f(X_t)|X_0 = x] - f(x) &= t[-\langle \nabla f(x), \nabla V(x) \rangle + \Delta f(x)] + o(t) \\
\frac{P_t f(x) - f(x)}{t} &= -\langle \nabla f(x), \nabla V(x) \rangle + \Delta f(x) + \frac{o(t)}{t} \\
\lim_{t \searrow 0} \frac{P_t f(x) - f(x)}{t} &= -\langle \nabla f(x), \nabla V(x) \rangle + \Delta f(x) + 0
\end{aligned}$$

$$\mathscr{L} f(x) = -\langle \nabla f(x), \nabla V(x) \rangle + \Delta f(x) \tag{3.4}$$

This result should feel intuitively reasonable. $\mathscr{L} f(x)$ represents the instantaneous change in $P_t f(x)$ at $t = 0$. In other words, it answers the question of how $\mathbb{E}[f(X_t)]$ is likely to change for small Δt, given that $X_0 = x$. We know (X_t) has drift $-\nabla V(x)$ at $t = 0$, and moving in this direction, by definition of gradient, results in a change to f of $\langle \nabla f(x), -\nabla V(x) \rangle$. Simultaneously, (X_t) has a random movement proportional to a standard Brownian motion. Because this Brownian motion is uniform in direction, it contributes an averaging effect of f among points in a spherical shell around x, as captured by the term $\Delta f(x)$.

We can use the infitesimal generator to define two more important operators, beginning with the 'carré du champ' [BGL14, 20]:

Definition 3.5 (**Carré du champ**)**.** Let $\mathcal{A}$ be a vector subspace of the domain of $\mathscr{L}$ that is closed under products (i.e. is an algebra). Then for any $f, g \in \mathcal{A}$, we can define the operator

$$\Gamma(f, g) = \frac{1}{2}[\mathscr{L}(fg) - f\mathscr{L}(g) - g\mathscr{L}(f)]$$

As before, the construction of the domain of Γ is not our primary concern. Recalling that $\mathscr{L}$ acts as a differential operator, the expression inside the brackets should be familiar as the product rule from ordinary calculus. In fact, if $\mathscr{L}$ exactly satisfied the product rule, then we would simply have $\Gamma \equiv 0$. Hence, we can think of Γ as capturing how far $\mathscr{L}$ is from acting like familiar differential operators. We can again consider the case of the Langevin diffusion:

$$\begin{aligned}
\Gamma(f, g) &= \frac{1}{2}[(\Delta fg - \langle \nabla fg, \nabla V \rangle) - f(\Delta g - \langle \nabla g, \nabla V \rangle) - g(\Delta f - \langle \nabla f, \nabla V \rangle)] \\
&= \frac{1}{2}[\Delta fg - f\Delta g - g\Delta f + \langle \nabla V, -\nabla fg + f\nabla g + g\nabla f \rangle]
\end{aligned}$$

But by the product rule,

$$\nabla fg = f\nabla g + g\nabla f$$

so we can simplify to

$$= \frac{1}{2}[\Delta fg - g\Delta f - g\Delta f]$$

Applying the product rule twice also yields

$$\Delta(fg) = g\Delta f + 2\nabla f \cdot \nabla g + f\Delta g$$
$$\Gamma(f,g) = \langle \nabla f, \nabla g \rangle$$

Finally, we use the carré du champ to define one more operator that will be useful in the near future [Chewi24, 14]:

Definition 3.6 (**Dirichlet energy**)**.** Let (P_t) be a Markov semigroup with reference measure μ as in Definition 3.2, generator $\mathscr{L}$, and carré du champ Γ. The Dirichlet energy is then defined by

$$\mathscr{E}(f,g) = \int \Gamma(f,g) d\mu$$

Remarkably, in certain cases, when the semigroup is derived from a Markov process (X_t), the laws of the random variables exactly evolve in time to minimize the Dirichlet energy. This is the first, but not only, way in which we can view Langevin diffusions as pure gradient flows.

3.2. **Kolmogorov's Equations and Reversible Processes.** One of the most fundamental questions one can ask about a stochastic process (X_t), especially when described by a stochastic differential equation, is: given an initial probability distribution of X_0 and a time t, what is the distribution of X_t? Markov semigroups provide the answer to this question, and yield further insights into the long-term behavior of such processes.

To begin, recall that any Markov process (X_t) (here assumed to have state space $\mathbb{R}^n$) is fully specified by an initial location X_0 and a collection of transition kernels $p_t(x, dy)$, representing the probability measure of the distribution of X_t given that $X_0 = x$.

As we saw, the semigroup property, $P_{t+s} = P_t \circ P_s$, reflects the definition of a Markov process. The property is also reflected in the Chapman-Kolmogorov equation, which relates these transition kernels [BGL14, 16]:

$$p_{t+s}(x, dz) = \int_{y \in \mathbb{R}^n} p_t(y, dz) p_s(x, dy) \tag{3.7}$$

If we instead consider $p_t(x, y)$, the density of $p_t(x, dy)$ with respect to a reference measure (here the natural choice is the Lebesgue), then this equation is simply

$$p_{t+s}(x, z) = \int_{y \in \mathbb{R}^n} p_t(y, z) p_s(x, y) d\lambda(y) \tag{3.8}$$

These equations express that to find the probability of moving from x to z in time $t + s$, we can consider all the cases of the location at time s. For each y, the probability (density) of moving from x to y in time s is by definition $p_s(x, y)$, at which point, due to the Markov property, the probability of ending at z is $p_t(y, z)$. These cases are mutually exclusive so we integrate to find the total density $p_{t+s}(x, z)$.

As promised, we move to the first major use of Markov semigroups towards analyzing the Langevin diffusion: Kolmogorov's forward and backward equations.

To begin, we can use the densities of the transition kernels to give a more explicit formulation of the semigroup [BGL14, 54]:

$$P_t f(x) = \int_{y \in \mathbb{R}^n} f(y) p_t(x,y) d\lambda(y)$$

We can recall that by definition, $\partial_t P_t f(x)|_{t=0} = \mathscr{L} f(x)$. But what if we want to evaluate at an arbitrary time t? We have the following formula [Chewi24, 12]:

Proposition 3.9 (**Kolmogorov's backward equation**). *For any $t \geq 0$, $\partial_t P_t f = \mathscr{L} P_t f = P_t \mathscr{L} f$.*

Proof.

$$\partial_t P_t f = \lim_{h \searrow 0} \frac{P_{t+h} f - P_t f}{h} = \lim_{h \searrow 0} \frac{P_h - Id}{h} P_t f = \mathscr{L} P_t f$$

Because the semigroup is commutative, we can repeat the same calculation on the left.

$$\partial_t P_t f = \lim_{h \searrow 0} \frac{P_{t+h} f - P_t f}{h} = P_t \lim_{h \searrow 0} \frac{P_h - Id}{h} f = P_t \mathscr{L} f$$

□

We have seen that at any time t, the derivative of $P_t(f)$ is just $\mathscr{L}(P_t f)$, so we may formally write $P_t = \exp(t\mathscr{L})$. In fact, when working with Markov chains on finite spaces, so that $\mathscr{L}$ is a matrix, this formula is exactly true [BGL14, 35]. We can alternatively view Kolmogorov's backward equation as a statement about the kernel densities $p_t(x,y)$. The following calculations lack some rigor but should demonstrate the intuition:

$$P_t f(x) = \int_{y \in \mathbb{R}^n} f(y) p_t(x,y) d\lambda(y)$$

To emphasize that $\mathscr{L}$ is acting as a differential operator in x, we use the subscript $\mathscr{L}_x$. Using Kolmogorov's equation and pulling the time-derivative through the integral in y, we have

$$\mathscr{L}_x P_t f(x) = \partial_t P_t f(x) = \partial_t \int_{y \in \mathbb{R}^n} f(y) p_t(x,y) d\lambda(y) = \int_{y \in \mathbb{R}^n} \partial_t [f(y) p_t(x,y)] d\lambda(y)$$

$$\mathscr{L}_x \int_{y \in \mathbb{R}^n} f(y) p_t(x,y) d\lambda(y) = \mathscr{L}_x P_t f(x) = \int_{y \in \mathbb{R}^n} f(y) \partial_t p_t(x,y) d\lambda(y)$$

Finally, pulling the $\mathscr{L}_x$ through the integral in y and function of y, we conclude that

$$\int_{y \in \mathbb{R}^n} f(y) \mathscr{L}_x p_t(x,y) d\lambda(y) = \int_{y \in \mathbb{R}^n} f(y) \partial_t p_t(x,y) d\lambda(y)$$

$$\mathscr{L}_x p_t(x,y) = \partial_t p_t(x,y)$$

Hence, Kolmogorov's backward equation describes how the transition probability $p_t(x,y)$ changes in time for a fixed y, in terms of a spatial derivative in x. From this perspective, we expect a similar relation with $\mathscr{L}_y$ to calculate $\partial_t p_t(x,y)$ for a fixed x. To address this expectation we have Kolmogorov's forward equation. To begin, we introduce the dual of the Markov semigroup, P_t^*.

Let π_t denote the law of the random variable X_t, which has some starting distribution π_0. Then for any test function f, we can compute $\mathbb{E}[f(X_t)] = \int_{x \in \mathbb{R}^n} P_t f(x) d\pi_0(x)$ - we are just taking a weighted average over $P_t f(x)$ according to the distribution

π_0. Because this integral is bilinear in f and (the density function of) π_0, we can treat it as an inner product and thus define the adjoint operator P_t^* to satisfy

$$\int P_t f(x) d\pi_0(x) = \int f(x) dP_t^* \pi_0(x)$$

But at the same time, $\mathbb{E}[f(X_t)] = \int f(x) d\pi_t(x)$, so that $P_t^* \pi_0 = \pi_t$. By differentiating in time, we have

$$\partial_t \int f(x) dP_t^* \pi_0(x) = \int f(x) \partial_t dP_t^* \pi_0(x)$$

(Note that the above may be used as a definition of the time derivative of a measure.) Simultaneously, by the previous equality and Kolmogorov's backward equation,

$$\partial_t \int f(x) dP_t^* \pi_0(x) = \partial_t \int P_t f(x) d\pi_0(x) = \int P_t(\mathscr{L} f(x)) d\pi_0(x)$$

By definition of the dual,

$$\int P_t(\mathscr{L} f) d\pi_0(x) = \int \mathscr{L} f(x) dP_t^* \pi_0(x)$$

Defining the dual $\mathscr{L}^*$ in a similar manner yields

$$\int P_t(\mathscr{L} f) d\pi_0(x) = \int f(x) d\mathscr{L}^* P_t^* \pi_0(x)$$

Hence for any function f, we have

$$\partial_t \int f(x) dP_t^* \pi_0(x) = \int f(x) d\mathscr{L}^* P_t^* \pi_0(x)$$

Thus, we conclude that $\partial_t P_t^* \pi_0 = \mathscr{L}^* P_t^* \pi_0$. By similar calculations, one can show that as before, $\mathscr{L}^*$ and P_t^* commute. Hence, Kolmogorov's forward equation, also called the Fokker-Planck equation (when applied to the densities of the measures π_t), states that [Chewi24, 12]

$$\partial_t P_t^* \pi_0 = \mathscr{L}^* P_t^* \pi_0 = P_t^* \mathscr{L}^* \pi_0 \tag{3.10}$$

Alternatively,

$$\partial_t \pi_t = \mathscr{L}^* \pi_t \tag{3.11}$$

As with the backward equation, we can characterize this statement with respect to the kernel densities. By a similar calculation (left as an exercise to the reader), one can show that $\partial_t p_t(x, y) = \mathscr{L}_y^* p_t(x, y)$.

Equipped with these two equations, powerful tools for analyzing stochastic processes, we can return to one of the original questions of this section: what is the long-term behavior of the probability measures π_t, the laws of X_t? First, we need a definition for a steady-state of the process.

Definition 3.12. Let X_t be a Markov process in $\mathbb{R}^n$. Let π be a probability distribution on $\mathbb{R}^n$, and let $X_0 \sim \pi$. We call π a stationary (or invariant) distribution of X_t if for all $t \geq 0$, $X_t \sim \pi$.

In other words, even though the variable X_t itself may not reach a steady-state, its law may. According to (3.11), π is stationary exactly when $\mathscr{L}^* \pi = 0$. We earlier calculated the generator for the Langevin diffusion process. By integration, one can show that $\mathscr{L}^* f = \Delta f + \operatorname{div}(f \nabla V)$ [Chewi24, 13]. Hence, the density of a

stationary distribution, π, must satisfy

$$0 = \Delta\pi + \text{div}(\pi\nabla V) = \text{div}(\pi(\nabla \ln \pi + \nabla V))$$

The solution is of the form $\nabla(\ln \pi + V) = 0$, so that $\ln \pi + V = C$ for some constant C, and $\pi = A\text{exp}(-V)$. It is from this fact that the Langevin diffusion derives so much value in sampling algorithms: desired distributions often take the form of being proportional to $\exp(-V)$ for some function V, but with a constant A that is difficult to calculate. Now, equipped with the knowledge of what the steady-state distribution of the Langevin diffusion is, we move to the problem of convergence. This problem, the different ways in which we can define convergence, and the bounds on it we can achieve, are the focus of the remainder of the paper.

Now, with a notion of a stationary distribution, we consider together a Markov process X_t, its invariant distribution π as the reference measure, and its semigroup P_t, which acts on $L^2(\pi)$. Given these, we can focus on a special class of Markov processes [BGL14, 25]:

Definition 3.13 (**Reversible processes**)**.** Let P_t be the Markov semigroup of a process X_t, as above. The semigroup is called symmetric (with respect to π) if for any $f, g \in L^2(\pi)$ and any $t \geq 0$, $\int fP_t g d\pi = \int gP_t f d\pi$.

Equivalently, we say that π is reversible for P_t. This terminology arises because if π is a reversible distribution for the semigroup P_t and $p_t(x, y)$ are the densities of the transition kernels with respect to π, then p_t is symmetric: $p_t(x, y) = p_t(y, x)$. This fact can be easily derived from the definition as follows:

$$\int f(x)P_t g(x)d\pi(x) = \int g(x)P_t f(x)d\pi(x)$$

$$\int_x f(x) \int_y g(y)p_t(x, y)d\pi(y)d\pi(x) = \int_x g(x) \int_y f(y)p_t(x, y)d\pi(y)d\pi(x)$$

$$\int_x \int_y f(x)g(y)p_t(x, y)d\pi(y)d\pi(x) = \int_x \int_y f(y)g(x)p_t(x, y)d\pi(y)d\pi(x)$$

Because x and y are just dummy variables in the right side, we can rewrite it as

$$\int_y \int_x f(x)g(y)p_t(y, x)d\pi(x)d\pi(y)$$

This is again equal to $\int_x \int_y f(x)g(y)p_t(x, y)d\pi(y)d\pi(x)$, and this equality holds for any test functions f, g, so we must have $p_t(x, y) = p_t(y, x)$. In other words, under the reversible distribution, the probability of moving from x to y in time t is the same as moving from y to x. At the same time, the definition is exactly the statement that P_t is a symmetric operator. The definition is also equivalent to $\int f\mathscr{L}gd\pi = \int g\mathscr{L}fd\pi$, so it also states that $\mathscr{L}$ is a symmetric operator [BGL14, 25-26]. To see why the carré-du-champ and Dirichlet energy are useful, we have the following theorem [Chewi24, 14]:

Theorem 3.14 (**Integration by parts**)**.** *Let P_t be a Markov semigroup with its associated operators that is symmetric w.r.t π. For any functions f, g, we have*

$$\int f(-\mathscr{L})gd\pi = \int g(-\mathscr{L})fd\pi = \int \Gamma(f, g)d\pi = \mathscr{E}(f, g)$$

In fact, this is an equivalent condition to symmetry, via the first equality. This theorem allows us to show that $-\mathscr{L}$ is a positive semi-definite operator [BGL14,

28]. It is also a (nontrivial) fact that in the case of symmetric semigroups, $\mathscr{L}$ is not just symmetric, but self-adjoint. These qualifiers on $\mathscr{L}$ allow for deeper analysis, since it is easier to work with positive self-adjoint operators. Despite this fact, and that the Langevin diffusion is symmetric w.r.t its invariant distribution (which can be checked using Theorem 3.14 and is left as an exercise), in working with Kolmogorov's equations, we introduced both $\mathscr{L}$ and $\mathscr{L}^*$. In that calculation, we were taking the adjoint in $L^2(\lambda)$ instead of $L^2(\pi)$, where they are the same. Hence, if we take f_t to be the density of π_t w.r.t π, Kolmogorov's forward equation reads:

$$\partial_t f_t = \mathscr{L} f_t \tag{3.15}$$

Because it holds for the Langevin diffusion, moving forward, we will generally assume that π is reversible. Finally, we return to the question of Dirichlet energy and gradient flow. Let us consider a Markov semigroup P_t with a reversible invariant distribution π. Let f_t be the densities of the laws of X_t w.r.t π, so that $t \mapsto f_t$ is a curve in $L^2(\pi)$. We can also consider the functional $\mathscr{E}(f_t) = \mathscr{E}(f_t, f_t)$ and its $L^2(\pi)$ gradient, $\nabla \mathscr{E}(f_t)$. As usual, the gradient is defined such that for any curve $t \mapsto f_t$ and its velocity $v_t = \partial_s f_s|_{s=t}$,

$$\partial_t \mathscr{E}(f_t)|_{t=0} = \langle \partial_t f_t|_{t=0}, \nabla \mathscr{E}(f_t)\rangle = \int v_0 \nabla \mathscr{E}(f_0) d\pi$$

Then we can use Theorem 3.14 to show [Chewi24, 15]:

$$\mathscr{E}(f_t) = \int f_t(-\mathscr{L}) f_t d\pi$$

$$\partial_t \mathscr{E}(f_t) = \partial_t \int f_t(-\mathscr{L}) f_t d\pi = \int \partial_t [f_t(-\mathscr{L}) f_t] d\pi$$

By the product rule,

$$\partial_t \mathscr{E}(f_t) = \int (\partial_t f_t)(-\mathscr{L} f_t) + f_t(\partial_t(-\mathscr{L} f_t)) d\pi$$

Because $\mathscr{L}$ is a linear operator with no dependence on t, we can interchange again to find

$$\partial_t \mathscr{E}(f_t) = \int v_t(-\mathscr{L}) f_t + f_t(-\mathscr{L}) v_t d\pi$$

Next, by symmetry of $-\mathscr{L}$, we can combine to see

$$\partial_t \mathscr{E}(f_t) = 2\int v_t(-\mathscr{L}) f_t$$

$$\int v_0 \nabla \mathscr{E}(f_0) d\pi = \partial_t \mathscr{E}(f_t)|_{t=0} = 2\int v_0(-\mathscr{L}) f_0 d\pi$$

This equality holds for any chosen f_0, v_0, so we must conclude $\nabla \mathscr{E}(f_0) = -2\mathscr{L} f_0$. A gradient flow of $\mathscr{E}$ will thus take the form $\partial_t f_t = -\nabla \mathscr{E}(f_t) = 2\mathscr{L} f_t$. But we also assumed that f_t is the law of X_t, so that by (3.15), $\partial_t f_t = \mathscr{L} f_t$. Hence, up to a difference in units of time, the laws of X_t evolve to minimize the Dirichlet energy.

3.3. **Two Important Functional Inequalities.** In this section, we investigate two functional inequalities that, when satisfied, give desired convergence results. Ultimately, we want to show that if π_t is the law of X_t and π is the invariant distribution, then $\pi_t \to \pi$ as $t \to \infty$, in some appropriate sense. When this occurs, we say that the Markov process mixes, because it forgets its initial condition. Thus, these kinds of results are also called mixing time results. To begin, however, we

take an alternate approach. If X_t mixes, then for large t, we would expect $P_t f(x)$ to converge to a constant function, as the law of X_t will not depend on X_0. That constant must be the true mean of f, $\int f d\pi$.

To bound the rate of this convergence, we turn to the negative generator, $-\mathscr{L}$, and its spectrum. Recall that when π is reversible, $-\mathscr{L}$ is positive semi-definite and self-adjoint, and therefore has real nonnegative eigenvalues. It will always have 0 as an eigenvalue because for any constant function c, $P_t(c)$ is also the constant function c, so $\partial_t P_t(c) = \mathscr{L} P_t(c) = \mathscr{L}(c) = 0$. However, let us suppose that the eigen values of $-L$ exist in $\{0\} \cup [C, \infty)$. (To avoid the case of $\lambda = 0$, we only consider functions with mean zero, i.e. in the orthogonal complement of the subspace of constant functions in $L^2(\pi)$.) This is appropriately called a spectral gap for $-\mathscr{L}$. Recall that because $\partial_t P_t f = \mathscr{L} P_t f$, we can formally write $P_t = \exp(t\mathscr{L})$. Hence, if $\mathscr{L}$ is bounded above by $-C$, uniformly in time, then we would expect (negative) exponential convergence of $P_t f$ to 0, with constant proportional to $-C$. To formalize this convergence, we have the Poincaré inequality [BGL14, 181].

Definition 3.16 (**Poincaré inequality**)**.** Let P_t be a Markov semigroup with invariant distribution π. We say that π satisfies a Poincaré inequality with constant C if for any function f in the domain of $\mathscr{E}$

$$\mathrm{Var}_\pi(f) \leq C\mathscr{E}(f)$$

Variance is defined as usual in probability theory: $\int f^2 d\pi - (\int f d\pi)^2$. One immediate consequence of the Poincaré inequality, abbreviated PI(C), is that if $\mathscr{E}(f) = 0$ then $\mathrm{Var}_\pi(f) = 0$, i.e. f is constant almost everywhere. Hence, the convergence of $\mathscr{E}(f_t)$ downwards (where f_t is the density of π_t w.r.t π), as proven earlier, goes hand-in-hand with convergence of f_t to a constant. This constant must be 1, and $f_t = 1$ implies that $\pi_t = \pi$.

Returning to the spectrum of $-\mathscr{L}$, let f be an eigenfunction with eigenvalue $\lambda > 0$. Then if the Poincaré inequality holds with constant C and f has mean zero, by Theorem 3.14

$$\int f^2 d\pi \leq C\mathscr{E}(f) = C \int f(-\mathscr{L}) f d\pi = C\lambda \int f^2 d\pi$$

$$C\lambda \geq 1 \implies \lambda \geq \frac{1}{C}$$

We thus derive an upper bound on $\frac{1}{C}$, or a lower bound on C, from the spectral gap of $-\mathscr{L}$. As promised, we have a theorem linking the Poincaré inequality to exponential decay of $P_t f$ [BGL14, 182].

Theorem 3.17. *Let P_t be a Markov semigroup with invariant distribution π. The following statements are equivalent:*

(1) π satisfies PI(C)
(2) For any $f \in L^2(\pi)$ and $t \geq 0$,

$$Var_\pi(P_t f) \leq exp(-\frac{2t}{C}) Var_\pi(f)$$

(3) For any f in $L^2(\pi)$, there exists a constant $c(f) > 0$ such that for any $t \geq 0$,

$$Var_\pi(P_t f) \leq c(f) exp(-\frac{2t}{C})$$

The proof hinges on showing that $\partial_t \text{Var}(P_t f) = -2\mathscr{E}(P_t f)$. Finally, we can use the Poincaré inequality to show convergence of $\pi_t \to \pi$. Again, let f_t be the density of π_t with respect to π. By Kolmogorov's forward equation and symmetry of $\mathbb{P}_t$, $P_t f_0 = f_t$. As $\pi_t \to \pi$, $f_t \to 1$, so we are interested in the convergence of $||f_t - 1|| \to 0$. We formalize this through χ^2 divergence [Chewi24, 17]:

Definition 3.18 (χ^2 **divergence**)**.** Let π' and π be probability measures on the same space. We define the χ^2 divergence of π' w.r.t π by:

$$\chi^2(\pi'||\pi) = ||\frac{d\pi'}{d\pi} - 1||_{L^2(\pi)}$$

when $\pi' << \pi$ and ∞ otherwise.

Substituting π_t for π' yields $\chi^2(\pi_t||\pi) = \text{Var}(f_t)$, which is exactly what Theorem 3.17 allows us to bound. Thus, we have another consequence of the Poincaré inequality [Chewi24, 17]. If P_t satisfies PI(C), then for any π_0 and $t \geq 0$,

$$\chi^2(\pi_t||\pi) \leq \exp(-\frac{2t}{C})\chi^2(\pi_0||\pi)$$

We thus obtain our first convergence result of $\pi_t \to \pi$. However, the question remains of how to prove the existence of a Poincaré inequality. This question will be explored more in the next section, but for now, we have one result specifically useful for the case of Langevin diffusions [BGL14, 203]:

Theorem 3.19 (**Kannan-Lovász-Simonovits-Bobkov**)**.** *Let π be a probability measure on $\mathbb{R}^n$ given by $d\pi = e^{-V}dx$ for some smooth convex function V. Then with respect to $\Gamma = |\nabla f|^2$, π satisfies a Poincaré inequality.*

The proof of this theorem is rooted in geometry, and while it will not be covered here, is the first glimpse at the geometry underlying the analysis of Markov processes and semigroups. Note that taking V to be the potential function, $e^{-V}dx$ is exactly the stationary distribution of the Langevin diffusion. Moreover, $|\nabla f|^2$ is its carré du champ. Then, all we need is for V to be smooth and convex. In this case, we call π log-concave. In fact, for the purposes of optimization it is often reasonable to assume that V is convex, and for purposes of sampling it is reasonable to assume that π is log-concave [Chewi24, x].

In considering the convergence of $P_t f$ towards the constant function $\int f d\pi$, variance was one possible way to measure how far a function is from being constant. Another comes in the form of entropy [BGL14, 236]:

Definition 3.20 (**Entropy**)**.** For any probability measure μ, we define the entropy of f as

$$\text{Ent}_\mu(f) = \int f \log(f) d\mu - (\int f d\mu) \log(\int f d\mu)$$

As with variance, the entropy of f is zero iff f is constant a.e. Moreover, it is only defined for nonnegative functions f, and $0 \log(0) = 0$. Just as bounds on variance give rise to the Poincaré inequality, entropy allows us to define a new inequality.

Definition 3.21 (**log-Sobolev inequality**)**.** The Markov semigroup P_t with invariant distribution π is said to satisfy a log-Sobolev inequality with constant C (or LSI(C)) if for all functions f in the domain of $\mathscr{E}$,

$$\text{Ent}_\pi(f^2) \leq 2C\mathscr{E}(f)$$

We can again rephrase the LSI as a statement about probability distributions, as opposed to functions. To do so, we introduce two new definitions [Chewi24, 17]:

Definition 3.22 (**Kullback-Leibler divergence**)**.** If π' and π are two probability measures such that $\frac{d\pi'}{d\pi} = f$, then the Kullback-Leibler divergence of π' w.r.t π, also called the relative entropy, is defined by

$$KL(\pi'||\pi) = \int f \log(f) d\pi = \int \log(f) d\pi' = \mathrm{Ent}_\pi(f)$$

(As before, if π' is not absolutely continuous w.r.t π then $KL(\pi'||\pi) = \infty$.) Second, [BGL14, 237]:

Definition 3.23 (**Fisher information**)**.** If π' and π are two probability measures such that $\frac{d\pi'}{d\pi} = f$, then the Fisher information of π' w.r.t π is defined by

$$I(\pi'||\pi) := \mathscr{E}(f, \log(f)) = \int \frac{\Gamma(f)}{f} d\pi$$

We can use Fisher information to give an equivalent formulation of LSI(C) [Chewi24, 18]: for all density functions f, we require that

$$KL(f\pi||\pi) \leq \frac{C}{2} I(f\pi||\pi) \tag{3.24}$$

To make use of this formulation, note that one can show that if π_t is the law of X_t and $f_t = \frac{d\pi_t}{d\pi}$, then $\partial_t KL(\pi_t||\pi) = -I(\pi_t||\pi)$ [Chewi24, 17]. We may thus conclude exponential convergence of $KL(\pi_t||\pi)$ to zero as $t \to \infty$, just as with $\chi^2(\pi_t||\pi)$ under a Poincaré inequality. Specifically we have [Chewi24, 18]:

Theorem 3.25. *For a Markov semigroup P_t with invariant distribution π, the following two conditions are equivalent:*

(1) π satisfies LSI(C)
(2) For any π_0 and $t \geq 0$,

$$KL(\pi_t||\pi) \leq exp(-\frac{2t}{C}) KL(\pi_0||\pi)$$

Log-Sobolev inequalities are stronger than Poincaré inequalities. Moreover, they are often more useful because $\chi^2(\pi_0||\pi)$ is likely to be much larger than $KL(\pi_0||\pi)$ [Chewi24, 37,38]. Hence, moving into the next section, we study the conditions under which an LSI holds.

4. Geometry of Markov Semigroups, Optimal Transport, and Otto Calculus

4.1. **Curvature-Dimension Conditions.** To begin, we address the problem of proving the existence of a log-Sobolev inequality. First, we must define the iterated carré du champ operator by [Chewi24, 18]:

$$\Gamma_2(f, g) := \frac{1}{2}(\mathscr{L}\Gamma(f, g) - \Gamma(f, \mathscr{L}g) - \Gamma(g, \mathscr{L}f)) \tag{4.1}$$

The iterated carré du champ is so named because its formula is analogous to that of Γ, but with Γ replacing multiplication of functions. Γ_2 is the key to unlocking log-Sobolev inequalities, per the following fundamental theorem [Chewi24, 19]:

Theorem 4.2 (**Bakry-Émery Theorem**). *Let P_t be a Markov semigroup such that for some constant α, for all functions f,*

$$\Gamma_2(f,f) \geq \alpha\Gamma(f,f)$$

Then P_t satisfies LSI(C) for some $C \leq \frac{1}{\alpha}$.

Hence, for any $\alpha > 0$, we can conclude the existence of an LSI, and it has a stronger guaranteed constant for larger α. The condition in this theorem is referred as a curvature condition and is denoted by $\mathrm{CD}(\alpha,\infty)$. To make use of this theorem, we examine when it is satisfied in the case of Langevin diffusions.

Substituting the case of the Langevin diffusion into the definition of Γ_2 we have

$$\Gamma_2(f,f) = \frac{1}{2}[\mathscr{L}(||\nabla f||^2) - 2\langle\nabla f, \nabla\mathscr{L}f\rangle]$$

Then by the Bochner identity for Euclidean space [Chewi24, 50],

$$\frac{1}{2}\Delta(||\nabla f||^2) = \langle\nabla\Delta f, \nabla f\rangle + ||\nabla^2 f||_{HS}^2$$

where $||\nabla^2 f||_{HS}^2 = \mathrm{Tr}(\nabla^2 f(\nabla^2 f)^T)$ and by the equation [Chewi24, 51]

$$\nabla\mathscr{L}f - \mathscr{L}\nabla f = -\nabla^2 V\nabla f,$$

one can show that

$$\Gamma_2(f,f) = ||\nabla^2 f||_{HS}^2 + \langle\nabla f, \nabla^2 V\nabla f\rangle$$

We therefore have the following theorem [Chewi24, 19]:

Theorem 4.3. *A Langevin diffusion with potential V satisfies CD(α,∞) iff V is α-strongly convex, i.e. for any vector w, $\langle w, \nabla^2 V w\rangle \geq \alpha$.*

Now, let us consider a more general kind of Langevin diffusion, taking place on an arbitrary manifold M instead of only $\mathbb{R}^n$, with potential V. The manifold admits a volume measure μ, and we then define a reference measure π by $\frac{d\pi}{d\mu} = C\exp(-V)$ for some constant C. The diffusion has the same infinitesimal generator and carré du champ, but on a manifold, where the Bochner identity takes the form of

$$\frac{1}{2}\Delta(||\nabla f||^2) = \langle\nabla\Delta f, \nabla f\rangle + ||\nabla^2 f||_{HS}^2 + Ric(\nabla f, \nabla f)$$

where $Ric(\cdot,\cdot)$ is the Ricci curvature tensor [Chewi24, 81]. Hence, we conclude that instead

$$\Gamma_2(f,f) = ||\nabla^2 f||_{HS}^2 + \langle\nabla f, (\nabla^2 V + Ric)\nabla f\rangle,$$

so that $Ric + \nabla^2 V$ is taking the place of $\nabla^2 V$. The first term captures the geometry of the space and the second of the potential, and thus the reference measure. In this case, to conclude $\mathrm{CD}(\alpha,\infty)$ (and therefore an LSI with $C \leq \frac{1}{\alpha}$, by Theorem 4.2), it suffices to show that for any vector field X, $Ric(X,X) + \langle X, \nabla^2 V X\rangle \geq \alpha||X||^2$ everywhere [Chewi24, 81]. In fact, the curvature condition can be further refined to a curvature-dimension condition $\mathrm{CD}(\alpha,d)$:

Definition 4.4 (**Curvature-dimension condition**). The Markov semigroup P_t satisfies $\mathrm{CD}(\alpha,d)$ if for any function f,

$$\Gamma_2(f,f) \geq \alpha\Gamma(f,f) + \frac{1}{d}(\mathscr{L}f)^2$$

This condition is so-named because if we take a Brownian motion on a manifold M (i.e. Langevin diffusion with constant V), the condition is equivalent to M

having Ricci curvature at least α everywhere, in all directions, and dimension at most d [Chewi24, 82]. In summary, we have the following result [BGL14, 215,270]:

Theorem 4.5. *If P_t is a Markov semigroup satisfying CD(α, d) for $\alpha > 0$ and $d > 1$, then it also satisfies PI($\frac{d-1}{\alpha d}$) and LSI($\frac{d-1}{\alpha d}$).*

4.2. **Optimal Transport Theory and Wasserstein Space.** One issue with the previous approach is that in considering the tensor $Ric + \nabla^2 V$, the properties of the manifold M in which we are working and the diffusion process (i.e. choice of V) become intertwined. In order to separate them, we present another way to view the evolution of Langevin diffusions through the theory of optimal transport.

Consider separable metric spaces X and Y, and let $P(X), P(Y)$ represent the space of probability measures on X and Y respectively. Then for any cost function $c : X, Y \to \mathbb{R}_{\geq 0}$ we define the optimal transport cost as follows [Chewi24, 20]:

Definition 4.6 (**Optimal transport cost**)**.** For $\mu_x \in P(X)$ and $\mu_y \in P(Y)$, the optimal transport cost is given by

$$T(\mu_x, \mu_y) := \inf_{\pi \in \Pi(\mu_x, \mu_y)} \int_{X \times Y} c(x, y) d\pi(x, y)$$

where $\Pi(\mu_x, \mu_y) \subseteq P(X \times Y)$ is the subset of probability measures on $X \times Y$ having first and second marginal distributions of μ_x and μ_y, respectively. In other words, $\pi \in \Pi(\mu_x, \mu_y)$ satisfies that for any bounded continuous functions f and g,

$$\int_{X \times Y} (f(x) + g(y)) d\pi(x, y) = \int_X f(x) d\mu_x + \int_Y g(y) d\mu_y$$

A minimizing $\pi \in \Pi(\mu_x, \mu_y)$ is called an optimal transport plan and always exists when c is lower semicontinuous [Chewi24, 20]. In order to define a distance between probability measures on the same space, we consider a particular kind of optimal transport problem [OV99, 362]:

Definition 4.7 (**Wasserstein distance**)**.** If M is a manifold and μ, ν are two probability measures, we define their Wasserstein distance by

$$W_2(\mu, \nu)^2 = \inf_{\pi \in \Pi(\mu, \nu)} \int_{M \times M} d(x, y)^2 d\pi(x, y)$$

When computing Wasserstein distance, we always assume that μ and ν have finite second moment. That is, for any chosen reference point $p \in M$,

$$\int_M d(p, x)^2 d\mu(x) < \infty$$

We denote the set of all such measures μ by $P_2(M)$. Equivalently, we can write [OV99, 362]:

$$W_2(\mu, \nu)^2 = \inf\{\mathbb{E}[d(X, Y)^2] : X \sim \mu, Y \sim \nu\}$$

where X and Y are random variables on M with laws μ and ν respectively. Finally, using duality, we can give one more equivalent formulation of the problem [Chewi24, 21]:

$$\frac{1}{2} W_2(\mu, \nu)^2 = \sup\{\int f d\mu + \int g d\nu : f, g \in \mathcal{D}(\mu, \nu)\} \tag{4.8}$$

where

$$\mathcal{D}(\mu,\nu) = \{f, g \in L^1(\mu) \times L^1(\nu) : f(x) + g(y) \leq \frac{||x-y||^2}{2} \text{a.e.}\}$$

The purpose of introducing optimal transport is to think about the set of probability measures itself on a manifold M as having a manifold-like structure. It will not be locally homeomorphic to Euclidean space, but we can endow it with a tangent space and local inner product, and from there, do calculus on it.

First, we must specify the set that forms the manifold of probability measures on M. We only work with measures that are absolutely continuous w.r.t the standard volume measure on M, and we again want to assume that they have finite second moment. That set, denoted by $P_{2,ac}(M)$, along with W_2 distance, is a complete and separable metric space [Chewi24, 26]. Defining the tangent space is slightly trickier.

Consider any curve in $P_{2,ac}(M)$: $\mu_t, t \in (-\varepsilon, \varepsilon)$, with $\mu_0 = \mu$. We want to define the velocity of this curve at $t = 0$, i.e. $\frac{\partial \mu_t}{\partial_t}|_{t=0}$. It is possible to prove that for any such curve, there exists a function Φ satisfying [OV99, 371]:

$$-\nabla \cdot (\mu \nabla \Phi) = \frac{\partial \mu_t}{\partial_t}|_{t=0} \tag{4.9}$$

which means that for any test function f,

$$\int \nabla f \cdot \nabla \Phi d\mu = \frac{d}{dt} \int f d\mu_t \tag{4.10}$$

We can picture the mass of the probability distribution μ_t as moving according to the gradient of Φ. For any point $p \in M$, the original mass there, $d\mu$, moves in the direction $\nabla\Phi$, resulting to a change of f of $\langle \nabla f, \nabla \Phi \rangle$, by definition of gradient. Integrating over the whole space must therefore yield the time-derivative of $\int f d\mu_t$. On the other hand, given a function Φ and an initial distribution μ_0, we can generate the corresponding curve using the previous differential equation, (4.9). There is therefore a one-to-one correspondence between velocities of curves through μ and gradient fields $\nabla\Phi$ (up to adding a constant to Φ). Hence, we set $T_\mu P$ to be the space of L^2 functions on M up to a constant difference and we define the local inner product by $\langle \Phi, \Psi \rangle = \int_M \langle \nabla\Phi, \nabla\Psi \rangle d\mu$ [OV99, 371]. A tangent space and local inner product allows us to construct geodesics, and therefore define a metric on our manifold of probability measures, even though we already began with the notion of W_2 distance. It therefore behooves us to check that the induced geodesic distance is exactly $W_2(\mu,\nu)$: luckily, using the dual version of the optimal transport problem, this exactly holds [OV99, 371-373]. Therefore, the choice of tangent space and inner product was correct. Now, with a more robust notion of this psuedo-manifold of probability measures, henceforth referred to as Wasserstein space and denoted by $P(M)$, we can perform calculus: this calculus is called 'Otto calculus' for Felix Otto, the mathematician who pioneered it.

4.3. **Applications of Otto Calculus.** Now, we explore some applications of Otto calculus. In this section, we do not present rigorous proofs (recall that Wasserstein space is not a bona fide manifold), but rather formal arguments that should yield value for the intuition behind the results. Let us consider a manifold M, with volume measure V and a reference probability measure ν having density proportional

to $\exp(-V)$. First, recall from (3.24) that ν satisfying LSI(C) is equivalent to having that for any $\mu = f\nu$, $KL(\mu||\nu) \leq \frac{C}{2} I(\mu||\nu)$. We are now prepared to present a formal proof of the following theorem, analogous to the prior result about LSI for measures on manifolds, but focused on the probability measures independent of any Markov semigroup, as presented in [OV99, 366]:

Theorem 4.11. *If ν is a probability measure as defined above such that $\nabla^2 V + Ric \geq C(Id)$ as bilinear operators, then ν satisfies LSI($\frac{1}{C}$).*

Proof. For the sake of brevity, we give an outline of the proof. We can begin by defining the functional $E(\mu) = KL(\mu||\nu)$. To begin doing formal calculus, we need to compute the gradient of E. Fortunately, as my good friend and fellow aspiring mathematician Otto put it, "differentiation follows by triviality". To be precise, we can consider a geodesic μ_t beginning at $\mu \in P$ with corresponding velocity Φ_t. Then by definition,

$$\frac{d}{dt} E(\mu_t) = \langle \nabla E(\mu_t), \Phi_t \rangle$$

From Definition 3.22, we can write

$$E(\mu_t) = \int \frac{d\mu_t}{d\nu} \log(\frac{d\mu_t}{d\nu}) d\nu = \int \log(\frac{d\mu_t}{d\nu}) d\mu_t$$

Thus, using (4.10), we can write

$$\frac{d}{dt} E(\mu_t) = \int \nabla \log(\frac{d\mu_t}{d\nu}) \cdot \nabla \Phi_t d\mu_t$$

Hence, in general,

$$\langle \nabla E(\mu), \Phi \rangle = \int \nabla \log(\frac{d\mu}{d\nu}) \cdot \nabla \Phi d\mu$$

Consequently, we also know that

$$||\nabla E(\mu)||^2 = \langle \nabla E(\mu), E(\mu) \rangle = \int ||\nabla \log(\frac{d\mu}{d\nu})||^2 d\mu$$

In fact, in the case of Langevin diffusions, this quantity is equal to the Fisher information I($\mu||\nu$) [Chewi24, 18]. Finally, using similar methods (and again applying the Bochner identity), one can show that the Hessian can be given by [OV99, 374]:

$$\langle \nabla^2 E(\mu)\Phi, \Phi \rangle = \int ||\nabla^2 \Phi||_{HS} + \nabla \Phi \cdot (Ric + \nabla^2 V) \nabla \Phi d\mu$$

This result parallels our computations of $\Gamma_2(f, f)$ in section 4.1. With this, we are ready to prove the theorem. We know by hypothesis that $Ric + \nabla^2 V \geq C\text{Id}$. Because the first term is a norm, and thus nonnegative, we also know that $\nabla^2 E(\mu) \geq C\text{Id}$. Substituting for KL($\mu||\nu$) and I($\mu||\nu$), we need to show that for any μ that is absolutely continuous w.r.t ν, $E(\mu) \leq \frac{1}{2C} ||\nabla E(\mu)||^2$. The key insight is to consider the curve μ_t where $\mu_0 = \mu$ and μ_t follows the gradient flow of E. Hence,

$$\frac{d}{dt} E(\mu_t) = \langle \nabla E(\mu_t), -\nabla E(\mu_t) \rangle = -||\nabla E(\mu_t)||^2$$

Moreover, we have the following, by the product rule:

$$\frac{d}{dt} ||\nabla E(\mu_t)||^2 = 2\langle \nabla E(\mu_t), \frac{d}{dt} \nabla E(\mu_t) \rangle$$

$$\frac{d}{dt} ||\nabla E(\mu_t)||^2 = 2\langle \nabla E(\mu_t), \nabla^2 E(\mu_t) \frac{d\mu_t}{dt} \rangle$$

$$\frac{d}{dt}||\nabla E(\mu_t)||^2 = 2\langle \nabla E(\mu_t), \nabla^2 E(\mu_t)(-\nabla E(\mu_t))\rangle$$

We know from the hypothesis that $\nabla^2 E \geq C\text{Id}$, so

$$\frac{d}{dt}||\nabla E(\mu_t)||^2 \leq -2C||\nabla E(\mu_t)||^2$$

Hence, we have exponential decay, and

$$||\nabla E(\mu_t)||^2 \leq \exp(-2Ct)||\nabla E(\mu)||^2$$

Thus, as $t \to \infty$, $||\nabla E(\mu_t)||^2 \to 0$. We know that the relative entropy $E(\mu)$ is only zero when $\frac{d\mu}{d\nu}$ is constant, but because they are both probability measures, that would mean $\mu = \nu$. ν is thus the only minimum of $E(\mu)$, so we must have $\mu_t \to \nu$, in the sense of Wasserstein distance. Now, we compute

$$E(\mu_t) - E(\mu) = \int_0^t \frac{d}{dt}E(\mu_t) = \int_0^t -||\nabla E(\mu_t)||^2\, dt$$

$$E(\mu) - E(\mu_t) = \int_0^t ||\nabla E(\mu_t)||^2\, dt \leq ||\nabla E(\mu)||^2 \int_0^t e^{-2Ct}\, dt$$

$$E(\mu) - E(\mu_t) \leq ||\nabla E(\mu)||^2(\frac{1}{2C})(1 - e^{-2Ct}) \leq \frac{1}{2C}||\nabla E(\mu)||^2$$

Because $\mu_t \to \nu$ as $t \to \infty$, we also have $E(\mu_t) \to 0$. Because this holds for all t, we must have

$$E(\mu) \leq \frac{1}{2C}||\nabla E(\mu)||^2$$

This completes the proof. □

4.4. **Langevin Diffusion as Gradient Flow.** Finally, we have one more incredible result that demonstrates the utility of Otto calculus and Wasserstein space.

The proof comes from a seminal paper by Jordan, Kinderlehrer and, Otto. To begin, we consider a Langevin diffusion (X_t) in $\mathbb{R}^n$ with potential V. As before, let π_t denote the law of X_t. Recall that if f_t denotes the density of π_t w.r.t λ, then by the Fokker-Planck equation (3.11), we have $\partial_t f_t = \mathscr{L}^* f_t$. We will work in $L^2(\lambda)$, so we cannot assume symmetry of $\mathscr{L}$. We can then compute the adjoint of the generator to find [JKO98, 3]:

$$\frac{df_t}{dt} = \text{div}(f_t \nabla V) + \Delta f_t$$

We have already seen that the stationary distribution of the Langevin diffusion, and thus the stationary solution of this PDE, has f proportional to $\exp(-V)$ (with a constant factor such that it integrates to 1, since it is a probability measure).

The authors are interested in finding a discrete-time scheme to approximate the solution to this PDE, and define it as follows. Define a functional by

$$F(f_t) = \int_{\mathbb{R}^n} V(x) f_t(x) d\lambda(x) + \int_{\mathbb{R}^n} f_t(x) \log(f_t(x)) d\lambda(x)$$

This is in fact just the Kullback-Leibler divergence, $\text{KL}(\pi_t||\pi)$, where π is the stationary distribution. This can be seen as follows:

$$KL(\pi_t||\pi) = \int \frac{d\pi_t}{d\pi} \log(\frac{d\pi_t}{d\pi}) d\pi$$

$$KL(\pi_t||\pi) = \int \frac{d\pi_t}{d\lambda} \log(\frac{d\pi_t}{d\pi}) d\lambda$$

Because $\frac{d\pi_t}{d\pi} = \frac{\frac{d\pi_t}{d\lambda}}{\frac{d\pi}{d\lambda}}$

$$KL(\pi_t||\pi) = \int f_t \log(\frac{f_t}{\exp(-V)})d\lambda$$

$$KL(\pi_t||\pi) = \int f_t(\log(f_t) - \log(\exp(-V)))d\lambda = \int f_t V d\lambda + \int f_t \log(f_t)d\lambda$$

As before, they consider the set of probability measures K with finite second moment, i.e. $\int |x|^2 f_t(x)dx < \infty$. Moreover, let d represent the Wasserstein distance. Then for a given initial distribution f_0 and a step size h, they iteratively define f_{k+1} to minimize

$$\frac{1}{2}d(f_{k+1}, f_k)^2 + hF(f_{k+1})$$

over all $f_{k+1} \in K$. This is a proximal method for discretizing gradient flow, which is in general equivalent to the backwards Euler method, defined by $\frac{f_{k+1}-f_k}{h} = -\nabla F(f_{k+1})$ [PB13, 145]. (Because Otto calculus had not yet been developed, ∇F did not yet have any meaning, but the scheme captures the same idea as discretizing gradient flow.) We will not give the details of the proof (the full paper is worth reading, but too long to include here), but the first step is to show that, with some fairly weak additional assumptions on V, the scheme is well-defined. After that, the main result is as follows:

Theorem 4.12. *Suppose f_0 satisfies $F(f_0) < \infty$ and for fixed $h > 0$, let $(f_{k,h})_k$ be the sequence of density functions solving the discretization scheme. Then we define the continuous-time interpolation $f_h(t, x) : (0, \infty) \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ by $f_h(t, x) = f_{k,h}(x)$ for all $t \in [kh, (k + 1)h)$ where $k \in \mathbb{N}_0$. Moreover, let $g_t(x)$ be the unique solution to*

$$\frac{dg_t}{dt} = div(g_t \nabla V) + \Delta g_t$$

such that $g_t \to f_0$ strongly in $L^1(\mathbb{R}^n)$, as $t \to 0$. Then as $h \to 0$, $f_h(t) \to g_t$ weakly in $L^1(\mathbb{R}^n)$ for all $t \in (0, \infty)$ and more generally $f_h \to g$ strongly in $L^1((0, T) \times \mathbb{R}^n)$ for any $T < \infty$.

In other words, for a Langevin diffusion, discretely approximating a gradient flow of KL divergence, in the limit of finer step size, is exactly how the law of X_t evolves. Thus, while X_t is defined as a 'noisy gradient flow', π_t follows an exact gradient flow. While this proof did not construct Wasserstein space as a pseudo-manifold of probability measures, it did use W_2 distance in its discretization scheme. In doing so, it laid the foundations for performing formal calculations in Wasserstein space, where the gradient of the KL divergence is well-defined in a certain sense.

Acknowledgments

First, I would like to thank my mentor, Antares Chen, for introducing me to the Langevin diffusion, optimal transport theory, and Otto calculus. I am grateful for his guidance as I worked through difficult computations to deepen my understanding, and for his advice on this paper. Second, I would like to thank Peter May for running the REU program and ensuring that it is open to anyone who is interested in mathematics research. Finally, I would like to thank my parents for their support of my interest in math, before, during, and after the program.

5. References

(1) [BGL14] D. Bakry, I. Gentil, M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, Cham, 2014.

(2) [Lawler23] G. Lawler, *Stochastic Calculus: An Introduction with Applications*, online notes. Available at https://www.math.uchicago.edu/~lawler/finbook.pdf.

(3) [Berestycki23] N. Berestycki, *Stochastic processes*, online notes. Available at https://homepage.univie.ac.at/nathanael.berestycki/wp-content/uploads/2023/03/StochasticProcesses.pdf.

(4) [Chewi24] S. Chewi, *Log-Concave Sampling*, unpublished book draft. Available at https://chewisinho.github.io/main.pdf.

(5) [Ollivier13] Y. Ollivier. "A visual introduction to Riemann curvatures and some discrete generalizations". In: *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the 50th Séminaire de Mathématiques Supérieures (SMS), Montréal, 2011*. Ed. by Galia Dafni, Robert McCann, and Alina Stancu. AMS, 2013, pp. 197– 219. Available at https://hal.science/hal-00858008.

(6) [OV99] F. Otto, C. Villani, *Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality*, Journal of Functional Analysis **173** (2000), 2, 361-400, DOI 10.1006/jfan.1999.3557.

(7) [JKO98] R. Jordan, D. Kinderlehrer, F. Otto, *The variational formulation of the Fokker-Planck equation*, SIAM Journal on Mathematical Analysis **29** (1998), 1, DOI 10.1137/S0036141096303359.

(8) [PB13] N. Parikh, N. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization **1** (2013), 3, 123-231.

Appendices

Appendix A Stochastic Processes. To make sense of stochastic differential equations, we will begin by understanding what a stochastic process is. To begin formalizing randomness, we need to work in a probability space.

Definition 5.1 (**Probability space**)**.** Let Ω be a set, and let $\Sigma \subseteq \mathcal{P}(\Omega)$ be a σ-algebra . If $\mathbb{P}$ is a (nonnegative) measure such that $\mathbb{P}(\Omega) = 1$, then we call the triple $(\Omega, \Sigma, \mathbb{P})$ a probability space.

The key idea is the *random variable*. Intuitively, a random variable is exactly what it sounds like: a variable that takes on values non-deterministically. We use them to model a variety of real-world processes, like the movements of a stock's price of a particle undergoing motion with random 'noisy' forces. Formally, we say the following [BGL14, 7].

Definition 5.2 (**Random variable**)**.** Given a probability space $(\Omega, \Sigma, \mathbb{P})$ and a measurable space $(E, \mathcal{F})$ (i.e. a set E with a σ-algebra $\mathcal{F}$), a random variable X is a function $X : \Omega \to E$ such that for any $A \in \mathcal{F}$, $X^{-1}(A) \in \Sigma$.

Most often, we will consider the case where $E = \mathbb{R}^n$ and $\mathcal{F}$ is the Borel algebra generated by the Euclidean topology. There are some assumptions about E necessary to proceed, but the examples we deal with are nice enough that we do not need to worry about these technical details. For example, it suffices to assume that

E is a Polish space, or that it is separable and metrizable by a metric in which it is complete, and that $\mathcal{F}$ is its Borel algebra [BGL14, 7].

The essential information about a random variable is often not literally the function $X : \Omega \to E$, but rather, the probabilities with which it takes on various values in E. We might informally refer to this information as its distribution, and formalize it as follows.

Definition 5.3. Given a probability space $(\Omega, \Sigma, \mathbb{P})$ and an output space $(E, \mathcal{F})$, we define the law of the random variable $X : \Omega \to E$ by the measure μ on $\mathcal{F}$ satisfying $\mu(F) = \mathbb{P}(X^{-1}(F))$, for any $F \in \mathcal{F}$.

Note that because by definition X is a measurable function, $X^{-1}(F) \in \Sigma$ for any $F \in \mathcal{F}$, so this is always well-defined.

Our primary object of interest is a particular stochastic process. We can think of a stochastic process as time-series data of random variables. Most broadly, a stochastic process is simply a collection of random variables on the same probability space indexed by some (infinite) set.

Definition 5.4 (**Stochastic process**)**.** Suppose we have a probability space $(\Omega, \Sigma, \mathbb{P})$, an output space $(E, \mathcal{F})$, and a set I of infinite cardinality. A stochastic process is a collection $\{X_t | t \in I\}$ where each $X_t : \Omega \to E$ is a random variable.

When I is countable, we identify it with $\mathbb{N}$ and have a discrete-time stochastic process. When I has the cardinality of $\mathbb{R}$, we identify it with $\mathbb{R}_{\geq 0}$ and we have a continuous-time process. We will thus always assume that $I = \mathbb{N}$ or $I = \mathbb{R}_{\geq 0}$, and we do not consider any I with cardinality greater than $\mathbb{R}$. Note that while we initially define a stochastic process to be a time-indexed collection of random variables, it can also be seen as a random path in E. Indeed, we can consider the function $X : \Omega \times I \to E$ given by $X(\omega, t) = X_t(\omega)$, which fully characterizes the stochastic process. For any $\omega \in \Omega$, we thus have a function $X_\omega(t) : I \to E$, which can thus be seen as a randomly chosen function from I to E, or a random path in E. A stochastic process is therefore a random variable taking values in the space of functions from I to E. This dual view will often be just as interesting and useful as the original.

Any stochastic process is equipped with a filtration, which loosely speaking represents the information contained in the process up to time t. We can begin with the case of $I = \mathbb{N}$.

Definition 5.5. A discrete-time filtration is a sequence of σ-algebras $(\Sigma_n)_{n \in \mathbb{N}}$ such that for any $m, n \in \mathbb{N}$, if $m < n$, then $\Sigma_m \subseteq \Sigma_n$. Moreover, $\Sigma_0 = \{\emptyset\}$.

In the context of stochastic processes, we use a filtration of the σ-algebra Σ of the probability space, not of the output space. Moreover, we always assume that for any $n \in \mathbb{N}$, $\Sigma_n \subseteq \Sigma$. We can analogously define continuous-time filtrations, but will need some extra assumptions [Lawler23, 55-56].

Definition 5.6. A continuous-time filtration is a collection of σ-algebras $\{\Sigma_t | t \in \mathbb{R}_{\geq 0}\}$ such that for any $s < t$, $\Sigma_s \subseteq \Sigma_t$. We assume that it has the following properties:

- Right-continuity: For any $t \in \mathbb{R}_{\geq 0}$, $\Sigma_t = \bigcap_{s > t} \Sigma_s$
- Strong completeness: we assume that for any t, Σ_t contains all null sets of Σ. Recall that a null set A satisfies $A \subseteq B$ for $B \in \Sigma$, $\mathbb{P}(B) = 0$.

These technical details do not appear in any proofs presented in this paper, but they are important for the rigor of the foundations of stochastic processes and are thus worth reading. Once again, for any t, $\Sigma_t \subseteq \Sigma$. We think about Σ_t as representing all of the information contained in the stochastic process up to time t. Specifically, in both the discrete and continuous cases, for any $t \in I$, the random variable X_t is always (in this paper) assumed to be measurable with respect to Σ_t. To match the intuition, we will always assume that we are using the natural filtration, where Σ_t is the σ-algebra generated by $\{X_s : s \leq t\}$, or more precisely,

$$\Sigma_t = \sigma\{X_s^{-1}(A) | s \leq t, A \in \mathcal{F}\}.$$

Appendix B Riemannian Manifolds and Ricci Curvature. The work of the remaining sections takes place in the world of Riemannian manifolds, and requires some understanding of the different measures of curvature in these manifolds. We therefore begin with a brief informal introduction to what manifolds are, and other useful definitions and identities.

To begin, a manifold M of dimension n, for $n \in \mathbb{N}$, is a topological space that at each point $p \in M$ is locally homeomorphic to an open subset of $\mathbb{R}^n$. The space is also assumed to be Hausdorff and have a countable base for its topology. We often view the manifold as a subset of $\mathbb{R}^k$ for some natural $k \geq n$. A differentiable manifold is one that is equipped at each point p with a tangent space T_pM. T_pM is a linear subspace of $\mathbb{R}^k$, of dimension n, and is comprised of all possible velocities at p of curves in M passing through p. Finally, a Riemannian manifold is one equipped at each point p with an inner product defined on its tangent space, $\langle \cdot, \cdot \rangle_p : T_pM \times T_pM \to \mathbb{R}$. These inner products vary smoothly in p.

The local inner products induce a norm, so that we can measure the lengths of tangent vectors. If $f : [0,1] \to M$ is a curve, then for all t, $f'(t) \in T_{f(t)}M$ and we can measure the length of the curve by integration: $\int_0^1 ||f'(t)||_{f(t)} dt$. The $f(t)$ in the subscript indicates that we take the norm w.r.t the inner product at $p = f(t)$, but is implicit henceforth. We can then define the distance between points $p_0, p_1 \in M$ by:

$$d(p_0, p_1) := \inf\{\int_0^1 ||f'(t)|| dt : f(0) = p_0, f(1) = p_1\}$$

In words, we simply take the length of the shortest possible path between the two points. With this distance function, the manifold is also a metric space, and is assumed to be connected and complete [Oll10]. A geodesic is a curve such that locally, for any two points, the geodesic realizes their distance (i.e. is the shortest path). They are thus analogous to straight lines in Euclidean space. So that the curve specifies the function, we often assume that geodesics have constant (not necessarily unit) speed. Moreover, for any point p and vector $v \in T_pM$, there exists a unique geodesic beginning at p with initial velocity v. If that geodesic is given by $f(t)$, we define the endpoint of v to be $f(1)$ and the map from p to the endpoint is denoted $\exp_p(v)$ [Oll10]. We can also define the inverse map $\log_p(v)$, where $\log_p(p') = v \iff \exp_p(v) = p'$.

Next, we consider the differential structure of manifolds. The overall goal is to understand what a gradient flow on a manifold is and how differential operators inform us about curvature. To begin, consider a smooth function $f : M \to \mathbb{R}$. We can define its differential df at a point p to map $T_pM \to \mathbb{R}$ by taking a curve beginning at p with velocity v and setting $(df)_p(v) = (\partial_t f(p_t))|_{t=0}$ [Chewi24, 77].

The differential $(df)_p$ is thus an element of the dual space of T_pM. Moreover, we define the gradient of f at p in the usual way: for any vector $v \in T_pM$, $\nabla f(p)$ satisfies $(df)_p(v) = \langle \nabla f(p), v \rangle_p$. The gradient flow of f is again the curve $p(t)$ whose velocity is always $-\nabla f(p(t))$ [Chewi24, 77].

Finally, we have the question of curvature. First, we consider parallel transport, which is a way to associate a vector $v \in T_pM$ to $v' \in T_qM$ for nearby p, q. Given $p \in M$ and $q \in M$, because we assume that $d(p, q)$ is small, we can assume that $q = \exp_p(w)$ for some $w \in T_pM$ [Ollivier13, 2]. Then, consider any $v \in T_pM$ such that is orthogonal to w. We then define v' by considering all vectors in T_qM that are orthogonal to the geodesic from p to q, and choosing one, v', that minimizes the distance between $\exp_p(v)$ and $\exp_q(v')$. The map from v to v' is called parallel transport and can be extended to the full domain T_pM by requiring that it is linear [Ollivier13, 2]. Moreover, for farther points, one can parallel transport smaller distances along a geodesic between the two.

This is an intuitive approach, but a formal definition can be given using more differential operators. Let us first define a vector field to be a map X associating to each point p a tangent vector $X(p) \in T_pM$. Given a function f, a vector field produces a new function Xf by $Xf(p) = (df)_p(X(p)) = \langle \nabla f(p), X(p) \rangle_p$ [Chewi24, 77-78]. Vector fields can thus be seen as differential operators on functions, but also on other vector fields via the Levi-Civita connection. The Levi-Civita connection is a map that takes in vector fields X, Y and outputs another vector field $\nabla_X Y$.

One important property of the Levi-Civita connection (and of affine connections more broadly) is that for vector fields X_1, X_2 and a smooth function f, $\nabla_{fX_1+X_2} Y = f\nabla_{X_1} Y + \nabla_{X_2} Y$, where f just scales the vector fields $X_1(p)$ and $\nabla_{X_1} Y(p)$ by $f(p)$. As a result of this property, one can prove the following lemma:

Lemma 5.7. *If $U \subseteq M$ is open and X_1, X_2, Y are vector fields such that $X_1|_U \equiv X_2|_U$ then $(\nabla_{X_1} Y)|_U \equiv (\nabla_{X_2} Y)|_U$.*

In other words, for fixed Y, local values of $\nabla_X Y$ only depend on local values of X. Finally, for any curve $f : \mathbb{R} \to M$, we can define the covariant derivative, the generalization of directional derivatives. (We assume f is not self-intersecting, as in the case of geodesics.) Consider the velocity field $f'(t)$ (a vector field on the subset $f(\mathbb{R}) \subseteq M$), and extend it smoothly to a vector field X defined on all of M. By the localization property in the above lemma, one can show that how we choose X does not matter. Now, for any vector field Y, we define the covariant derivative of Y along $f(t)$ by [Chewi24, 78]:

$$D_f Y(t) := (\nabla_X Y)(f(t))$$

Because the extension to X is arbitrary, we also write it as $(\nabla_{f'(t)} Y)(f(t))$. Finally, we can define the parallel transport of a vector $v \in T_{f(0)}M$ along the curve $f(t)$ by the vector field V such that $D_f V(t) \equiv 0$.

With a robust notion of parallel transport, we can begin defining curvature. To begin, we have sectional curvature [Ollivier13, 3-4]:

Definition 5.8 (**Sectional curvature**)**.** Let M be a Riemannian manifold, and consider some $x \in M$. Moreover, consider $v, w \in T_pM$ with unit length and $\varepsilon, \delta > 0$. Let $y = \exp_x(\delta v)$ and let w' be the parallel transport of w from x to y. Then if

$p = \exp_x(\varepsilon w)$ and $q = \exp_y(\varepsilon w')$, we have for some K

$$d(p,q) = \delta(1 - \frac{\varepsilon^2}{2}K(v,w) + O(\varepsilon^3 + \varepsilon^2\delta))$$

$K(v,w)$ is called the sectional curvature at x.

In the case of Euclidean space, parallel transport creates a rectangle so the distance between p and q is the same as between x and y, i.e. δ. We would thus correctly find that $K \equiv 0$. But in other settings, that is not the case. Note that while the estimate of δ is correct up to first order in δ, ε, the curvature provides a necessary second-order correction. With sectional curvature, we may define Ricci curvature, which is ultimately our object of interest [Ollivier13, 4]:

Definition 5.9 (**Ricci curvature**)**.** Consider an n-dimensional manifold M and $x \in M$. If $v \in T_xM$ and $||v|| = 1$, we define the Ricci curvature $Ric(v)$ as n times mean of $K(v,w)$ for $w \in T_xM$, $||w|| = 1$.

We can also give an alternate characterization that is easier to visualize [Ollivier13, 4]:

Proposition 5.10. *Consider a point $x \in M$ (again of dimension n), a unit vector $v \in T_xM$, and $\varepsilon, \delta > 0$. Moreover, let $y = exp_x(\delta v)$, $S_x = \{w \in T_xM : ||w|| = \varepsilon\}$, and $S_y = \{w \in T_yM : ||w|| = \varepsilon\}$. If S_x is mapped to S_y via parallel transport, the average distance travelled by a point in S_x is given by*

$$\delta(1 - \frac{\varepsilon^2}{2n}Ric(v) + O(\varepsilon^3 + \varepsilon^2\delta)$$

Positive Ricci curvature is thus characterized by the phrase "balls are closer than their centers are" [Ollivier13, 4]. Ricci curvature can also be derived from the more general notion of the Riemann curvature tensor. Given vector fields W, X, Y, and Z, the Riemann curvature tensor is defined, using the Levi-Civita connection, by [Chewi24, 79]:

$$Riem(W,X,Y,Z) = \langle \nabla_X \nabla_W Y - \nabla_W \nabla_X Y + \nabla_{[W,X]} Y, Z \rangle$$

where the Lie bracket $[W,X]$ is a vector field implicitly defined by $[W,X](f) = W(X(f)) - X(W(f))$. Notably, when evaluating at a point p, the output depends only on $W(p), X(p), Y(p)$, and $Z(p)$, so the following is well-defined. For a point p and fixed $v, w \in T_pM$, the Ricci curvature tensor is defined by $Ric(v,w) = \mathrm{Tr} Riem(u, \cdot, v, \cdot)$. What we previously denoted by $Ric(v)$ was more properly $Ric(v,v)$, since the tensor takes two vector inputs. Positive Ricci curvature can have remarkable consequences, which are explored section 4.1.