CS395T: Continuous Algorithms, Part XVI
Stochastic calculus

Kevin Tian

1 Drift-diffusion processes

In this lecture, we introduce the tools required to study a framework for sampling algorithm design based on discretizing *drift-diffusion processes*, which is able to exploit different structural aspects of the target density compared to the previous few lectures (for example, an appropriate notion of well-conditionedness for densities). We adopt a dual perspective of such processes, viewing them as both a stochastic evolution in particle space (i.e., a particle in $\mathbb{R}^d$ following a stochastic trajectory), as well as a deterministic evolution in density space (i.e., an evolving measure in $\mathcal{P}(\mathbb{R}^d)$, the set of probability densities on $\mathbb{R}^d$ which are absolutely continuous with respect to the Lebesgue measure). This perspective lets us adopt a rich set of analytical tools on both $\mathbb{R}^d$ and $\mathcal{P}(\mathbb{R}^d)$ to analyze the convergence and bound the discretization error of the resulting algorithms.

We mention that this section is largely structured to briefly set up the mathematical tools to rigorously study stochastic calculus, and most of the formalism can be ignored on a first read in the later sections. We start by introducing Itô calculus, beginning with *Brownian motion*.

Definition 1 (Brownian motion)**.** *We define* Brownian motion *in* $\mathbb{R}^d$*, denoted by* $\{\boldsymbol{B}_t\}_{t\geq 0}$*, to be a* stochastic process*, i.e., a random sequence of points in* $\mathbb{R}^d$ *indexed by* $t \geq 0$ *(thought of as time), satisfying the following properties.*

1. $\boldsymbol{B}_0 = \mathbf{0}_d$.
2. $\{\boldsymbol{B}_t\}_{t\geq 0}$ *is continuous with probability* 1.
3. *For all* $k \in \mathbb{N}$ *and all* $\{t_i\}_{i=0}^k \subset \mathbb{R}_{\geq 0}$ *with* $t_0 = 0 < t_1 < \ldots < t_k$*, all of the random variables* $\boldsymbol{B}_{t_{i+1}} - \boldsymbol{B}_{t_i}$ *for* $0 \leq i \leq k-1$*, are mutually independent.*
4. *For all* $0 \leq s \leq t$*,* $\boldsymbol{B}_t - \boldsymbol{B}_s$ *is distributed as* $\mathcal{N}(\mathbf{0}_d, (t-s)\mathbf{I}_d)$.

For an existence proof of Brownian motion, see Chapter 7 of [Dur10]. The probability space that Brownian motion is defined on, as well as all stochastic processes we will study, is denoted $\{\mathcal{F}_t\}_{t\geq 0}$, a *filtration* satisfying $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $0 \leq s \leq t$. We say that Brownian motion is a stochastic process *adapted to* the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Informally, we can think of $\mathcal{F}_t$ as containing the information of the randomness used up to time t, i.e., the realizations of the random Brownian motion, so the inclusion condition means that more information is always available later in time.

To give some intuition for the properties of Brownian motion, we prove the reflection principle.

Lemma 1 (Reflection principle)**.** *Let* $\{B_t\}_{t\geq 0}$ *be Brownian motion in* $\mathbb{R}$*. For all* $t \geq 0$ *and* $a > 0$*,*

$$\Pr\left[\sup_{0\leq s\leq t} B_s \geq a\right] = 2\Pr\left[B_t \geq a\right].$$

Proof. Let $\tau \geq 0$ be a random stopping time[1] corresponding to the first time the Brownian motion

[1] A stopping time τ adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is a random variable in $\mathbb{R}_{\geq 0}$, such that for all $t \in \mathbb{R}_{\geq 0}$, the event $\tau \leq t$ is measurable with respect to $\mathcal{F}_t$. In other words, we can determine whether the event defining τ has occurred using information in $\mathcal{F}_t$, e.g., the first time an $\{\mathcal{F}_t\}_{t\geq 0}$-measurable event occurs is a stopping time.

reaches a, i.e., τ is the random variable equal to $\inf\{s \mid B_s = a\}$. By independence of increments,

$$\begin{aligned}\Pr\left[\sup_{0\le s\le t} B_s \ge a\right] &= \Pr\left[\sup_{0\le s\le t} B_s \ge a \text{ and } B_t \ge a\right] + \Pr\left[\sup_{0\le s\le t} B_s \ge a \text{ and } B_t < a\right] \\ &= \Pr\left[B_t \ge a\right] + \Pr\left[\sup_{0\le s\le t} B_s \ge a \text{ and } B_t - B_\tau < 0\right] \\ &= \Pr\left[B_t \ge a\right] + \frac{1}{2}\Pr\left[\sup_{0\le s\le t} B_s \ge a\right],\end{aligned}$$

as $\sup_{0\le s\le t} B_s \ge a$ implies that $\tau \le t$, so $B_t - B_\tau < 0$ with probability $\frac{1}{2}$ regardless of the realization of $\tau \le t$. Rearranging the above display gives the claim. □

Brownian motion is an example of a continuous *martingale* with respect to $\{\mathcal{F}_t\}_{t\ge 0}$. To define this, recall that a martingale $\{\mathbf{x}_t\}_{t\ge 0}$ adapted to a filtration $\{\mathcal{F}_t\}_{t\ge 0}$ satisfies the property that

$$\mathbb{E}\left[\mathbf{x}_t \mid \mathcal{F}_s\right] = \mathbf{x}_s \text{ for all } 0 \le s \le t.$$

A useful property of martingales is that by Jensen's inequality, for all convex φ and $0 \le s \le t$,

$$\mathbb{E}\varphi(\mathbf{x}_t) = \mathbb{E}\left[\mathbb{E}\left[\varphi(\mathbf{x}_t) \mid \mathcal{F}_s\right]\right] \ge \mathbb{E}\varphi(\mathbf{x}_s).$$

One consequence of the martingale property (see Chapter 7.5, [Dur10]) is that for all bounded stopping times τ, $\mathbb{E}\mathbf{x}_\tau = \mathbf{x}_0$. We are now ready to introduce the Itô integral, which is defined with respect to a continuous process $\{\mathbf{H}_t\}_{t\ge 0} \subset \mathbb{R}^{d\times d}$ adapted to $\{\mathcal{F}_t\}_{t\ge 0}$.[2] We can view $\{\mathbf{H}_t\}_{t\ge 0}$ as a "reweighting" of Brownian motion, and we correspondingly define the Itô integral

$$\mathbf{x}_t := \int_0^t \mathbf{H}_s \mathrm{d}\boldsymbol{B}_s,$$

which is a continuous martingale adapted to $\{\mathcal{F}_t\}_{t\ge 0}$. In particular, this means that

$$\mathbb{E}\left[\int_s^t \mathbf{H}_u \mathrm{d}\boldsymbol{B}_u \mid \mathcal{F}_s\right] = 0 \text{ for all } 0 \le s \le t.$$

It is useful to view the Itô integral as a stochastic process indexed by t. For example, choosing $\mathbf{H}_t = \mathbf{I}_d$ for all t gives that $\mathbf{x}_t = \boldsymbol{B}_t$ is simply Brownian motion.

An important characteristic of a stochastic process $\{\mathbf{x}_t\}_{t\ge 0}$ is its *quadratic variation*, defined by

$$[\mathbf{x}]_t = \lim_{\|P\|_{\text{gap}}\to 0} \sum_{k=1}^{|P|} \left\|\mathbf{x}_{t_k} - \mathbf{x}_{t_{k-1}}\right\|_2^2, \tag{1}$$

the limit over meshes $P = \{t_k\}_{k\le |P|} \subset [0,t]$ where $\|P\|_{\text{gap}} = \max_{k\le|P|} |t_k - t_{k-1}|$, and $t_{|P|} = t$ and $t_0 := 0$. One can formally verify that when $\mathbf{x}_t$ is an Itô integral driven by $\{h_t\}_{t\ge 0}$, we have that

$$[\mathbf{x}]_t = \int_0^t \|\mathbf{H}_s\|_{\mathrm{F}}^2 \,\mathrm{d}s, \tag{2}$$

thought of as the "total variance" accumulated thus far throughout the process. For example, if $\{\boldsymbol{B}_t\}_{t\ge 0}$ is 1-dimensional Brownian motion then $[\boldsymbol{B}]_t = t$. The quadratic variation is useful in the context of a characterization known as the *Dambis-Dubins-Schwarz* theorem [Dam65, DS65], which says that in one dimension, all continuous martingales $\{x_t\}_{t\ge 0}$ are distributionally identical to a time-changed Brownian motion $\{B_{\tau(t)}\}_{t\ge 0}$, where we let $q(t) := [x]_t$ be the quadratic variation at time t. So, all continuous martingales in $\mathbb{R}$ are characterized by their quadratic variation.

Intuitively, the expression (2) holds because of the heuristic "$\mathrm{d}B_t^2 = \mathrm{d}t$," which makes sense in one dimension and can be extended to higher dimensions straightforwardly. This is because an infinitesimal advancement of Brownian motion (at timescale $\mathrm{d}t$) behaves like $\mathcal{N}(0, \mathrm{d}t)$, so its square

[2] Formally, the continuity of $\{\mathbf{H}_t\}_{t\ge 0}$ is not necessary, and it just needs to be *progressive*, a strengthening of being adapted that implies stopped processes are measurable. All of the Itô integrals that we encounter will be continuous and adapted processes, so we do not make this distinction for simplicity.

has expectation $\mathrm{d}t$. Formalizing this intuition and using the martingale property of Itô integrals to deal with "cross-terms" establishes equivalence of the quadratic variation definitions (1) and (2).

Relatedly, we mention that one of the most important properties of the Itô integral is the *Itô isometry*, which holds if the right-hand side below is finite:

$$\mathbb{E}\left[\left\|\int_0^t \mathbf{H}_s \mathrm{d}\boldsymbol{B}_s\right\|_2^2\right] = \mathbb{E}\left[\int_0^t \|\mathbf{H}_s\|_{\mathrm{F}}^2 \,\mathrm{d}s\right].$$

The handling of cross-terms when expanding the above expression is handled analogously to the intuition for (2), and the Itô isometry then follows by taking total expectations on both sides.

Remark 1. *Much of the previous discussion can in fact be substantially generalized via the formalism of* local martingales *and* localizing sequences. *Specifically, while martingales have $\mathbb{E}\mathbf{x}_\tau = \mathbf{x}_0$ for all stopping times τ, a local martingale $\{\mathbf{x}_t\}_{t\geq 0}$ adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ is a stochastic process satisfying the weaker property that there is a sequence of stopping times $\{\tau_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}\tau_n \to \infty$ with probability 1, where $\mathbf{x}_{\min\{t,\tau_n\}}$ is a martingale for all $n \in \mathbb{N}$.*[3] *These stopping times are called a* localizing sequence*, and are used to formalize Itô integrals, which as previously mentioned can be extended to handle* progressive *integrands, also through the machinery of localizing sequences. Localizing sequences are useful as a way to handle undesirable behavior as $t \to \infty$, for rigorously defining key concepts in stochastic calculus by using Itô integrals.*[4]

We also mention that Itô integrals are not the only method of formalizing a theory of stochastic calculus. In particular, the Stratonovich integral *is a popular alternative in physics. Loosely speaking, the distinction is that the Stratonovich integral is defined with respect to a "midpoint" rule (familiar from Riemann integration), whereas the Itô integral is defined with respect to left endpoints. The key difference useful for our purposes is that the definition of the Itô integral makes it a martingale, whereas the Stratonovich integral is not in general. For the remainder of the lecture, we specialize Itô integrals to the setting of continuous adapted $\{\mathbf{H}_t\}_{t\geq 0}$.*

Next, we define drift-diffusion processes on $\mathbb{R}^d$, adapted to the same filtration $\{\mathcal{F}_t\}_{t\geq 0}$ that Brownian motion is adapted to. Informally, drift-diffusion processes can be thought of as modeling the deterministic and stochastic components in the time-evolution of a random particle. A drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$ on $\mathbb{R}^d$ is driven by a vector-valued function $\boldsymbol{\mu} : \mathbb{R}^d \to \mathbb{R}^d$ and a matrix-valued function $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and captured by the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}(x_t)\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t. \tag{3}$$

In principle, SDEs can be defined with respect to any progressive process (not just functions $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of x_t), but (3) suffices for our purposes. We say such a stochastic process $\{\mathbf{x}_t\}_{t\geq 0}$ is a drift-diffusion process, sometimes also called an Itô process, and we write it in integral form as

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t \boldsymbol{\mu}(\mathbf{x}_s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\mathbf{x}_s)\mathrm{d}\boldsymbol{B}_s.$$

To establish existence and uniqueness of SDE solutions, the following characterization is helpful. We defer a proof to [Øk03]; the intuition for this result is similar to that for the Picard-Lindelöf theorem for establishing existence and uniqueness for the solution of ordinary differential equations (ODEs), which can be proven by defining an appropriate fixed-point iteration.

Proposition 1 (Theorem 5.2.1, [Øk03])**.** *Suppose, for a SDE of the form (3), that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ both have a finite Lipschitz constant (with respect to $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$, respectively). Then for all $t \geq 0$, there exists a unique solution $\mathbf{x}_t$ to the SDE for every realization of $\mathcal{F}_t$. Moreover, $\mathbf{x}_t$ is square-integrable in the sense that $\mathbb{E}\int_0^t \|\mathbf{x}_s\|_2^2 \,\mathrm{d}s < \infty$.*

We conclude the section with the most useful property of drift-diffusion processes when performing computations: a stochastic calculus generalization of the chain rule.

[3] One example which may help see the difference is the "sticky Brownian motion" in $\mathbb{R}$, which is Brownian motion until the first time -1 is reached, and is stuck at -1 henceforth. This is not a martingale, because the stopping time τ when -1 is first reached is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$, but $\mathbb{E}[x_\tau] = -1$. However, it is a local martingale.

[4] For instance, every bounded local martingale is a martingale, so the distinction is only due to limiting behavior.

Proposition 2 (Itô's lemma)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice-differentiable, and suppose that $\{\mathbf{x}_t\}_{t\geq 0}$ is a drift-diffusion process following the SDE (3). Then the stochastic process $\{f(\mathbf{x}_t)\}_{t\geq 0}$ is also a drift-diffusion process, and follows the SDE*

$$\mathrm{d}f(\mathbf{x}_t) = \left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}(\mathbf{x}_t)\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\sigma}(\mathbf{x}_t)^\top \right\rangle\right)\mathrm{d}t + \langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t\rangle\,.$$

More generally, if $\{\mathbf{x}_t\}_{t\geq 0}$ follows the SDE

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}_t\mathrm{d}t + \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t,$$

where $\{\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\}_{t\geq 0} \subset \mathbb{R}^d \times \mathbb{R}^{d\times d}$ are continuous stochastic processes adapted to the same filtration $\{\mathcal{F}_t\}_{t\geq 0}$ as $\{\boldsymbol{B}_t\}_{t\geq 0} \subset \mathbb{R}^d$, then $\{f(\mathbf{x}_t)\}_{t\geq 0}$ is a stochastic process following the SDE

$$\mathrm{d}f(\mathbf{x}_t) = \left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}_t\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^\top \right\rangle\right)\mathrm{d}t + \langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}_t\mathrm{d}\boldsymbol{B}_t\rangle\,.$$

We note that there is also a generalization of Proposition 2 to the more complicated setting where f is a time-dependent function, but we will not require it. The proof of Proposition 2 follows from a similar calculation as used to show the quadratic variation formula (2) holds, e.g., formalizing our aforementioned "$\mathrm{d}B_t^2 = \mathrm{d}t$" argument. For example, a Taylor expansion gives that for vanishing η, and approximating $\mathbf{x}_{t+\eta} \approx \mathbf{x}_t + \eta\boldsymbol{\mu}(\mathbf{x}_t) + \sqrt{\eta}\boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\xi}$ for $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$,[5]

$$\begin{aligned}
\mathrm{d}f(\mathbf{x}_t) &\approx f(\mathbf{x}_{t+\eta}) - f(\mathbf{x}_t) \\
&\approx \langle \nabla f(\mathbf{x}_t), \eta\boldsymbol{\mu}(\mathbf{x}_t) + \sqrt{\eta}\boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\xi}\rangle + \left\langle \nabla^2 f(\mathbf{x}_t), \eta\boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\xi}\boldsymbol{\xi}^\top\boldsymbol{\sigma}(\mathbf{x}_t)^\top \right\rangle \\
&\approx \eta\left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}(\mathbf{x}_t)\rangle + \left\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\sigma}(\mathbf{x}_t)^\top \right\rangle\right) + \langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\cdot(\sqrt{\eta}\boldsymbol{\xi})\rangle\,.
\end{aligned}$$

We used the approximation $\boldsymbol{\xi}\boldsymbol{\xi}^\top = \mathbf{I}_d$ in the last line, and dropped all terms of lower order than η. Unlike in the standard chain rule, there is an additional second-order component that persists, because $\|\sqrt{\eta}\boldsymbol{\xi}\|_2^2$ scales with η, explaining the presence of the additional term in Proposition 2.

A particular drift-diffusion process of significant interest is the *Langevin diffusion*. Letting $V : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function, the associated Langevin diffusion is the SDE

$$\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t, \tag{4}$$

i.e., we take $\boldsymbol{\mu} = -\nabla V$ and $\boldsymbol{\sigma} = \sqrt{2}\mathbf{I}_d$ in (3). Intuitively, (4) can be thought of as a noisy gradient flow, with a deterministic drift $-\nabla V(\mathbf{x}_t)\mathrm{d}t$ (making function value progress) and a diffusion component $\sqrt{2}\mathrm{d}\boldsymbol{B}_t$. For the SDE (4), Itô's lemma (Proposition 2) yields

$$\mathrm{d}f(\mathbf{x}_t) = (-\langle \nabla f(\mathbf{x}_t), \nabla V(\mathbf{x}_t)\rangle + \Delta f(\mathbf{x}_t))\,\mathrm{d}t + \sqrt{2}\,\langle \nabla f(\mathbf{x}_t), \mathrm{d}\boldsymbol{B}_t\rangle\,, \tag{5}$$

where Δ is the *Laplacian* operator given by the formula

$$\Delta f(\mathbf{x}) = \mathrm{Tr}\left(\nabla^2 f(\mathbf{x})\right). \tag{6}$$

Remark 2. *There is a connection between the Laplacian operator defined in (6) and the Laplacian matrix associated with a graph, defined in Definition 5, Part II. At a high level, the Laplacian $\mathbf{L}_G$ of an unweighted path graph G acts on each indicator vector $\mathbf{e}_v$ for a vertex v of G, with neighbors u, w, by computing $\mathbf{L}_G\mathbf{e}_v = 2\mathbf{e}_v - \mathbf{e}_u - \mathbf{e}_w$, i.e., it computes the difference between a vertex value and "averages" around its neighbors. Similarly, the Laplacian operator Δ computes a difference of a function value with an average in a small neighborhood, which is best-understood in one dimension:*

$$\begin{aligned}
\eta f''(x) &\approx f'(x+\eta) - f'(x) \\
&\approx f(x+\eta) - f(x) - f(x) + f(x-\eta) = f(x+\eta) + f(x-\eta) - 2f(x).
\end{aligned}$$

[5] Note that $\sqrt{\eta}\boldsymbol{\xi}$ has covariance matrix $\eta\mathbf{I}_d$, as required by Definition 1 for advancing time by η.

2 Markov semigroups

We now take a dual view on stochastic differential equations, by assuming that $\mathbf{x}_0$ is drawn from a density $\pi_0 \in \mathcal{P}(\mathbb{R}^d)$, and associating each time $t \geq 0$ with a density $\pi_t \in \mathcal{P}(\mathbb{R}^d)$ corresponding to the density of $\mathbf{x}_t$ (called its law), which evolves from $\mathbf{x}_0$ following a drift-diffusion process (3). That is, while the evolution of $\mathbf{x}_t$ is stochastic, the evolution of its *distribution* can be viewed as a deterministic process following a partial differential equation (PDE) derived from (3).

To introduce this equation and its analysis, we first give some definitions from *Markov semigroup theory*, which studies "continuous-time Markov chains" such as drift-diffusion processes through the evolution of functions. For a drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$, we define an associated *Markov semigroup* $\{P_t\}_{t\geq 0}$, which acts on an appropriate set of functions $f : \mathbb{R}^d \to \mathbb{R}$ via the definition

$$P_t f(\mathbf{x}) := \mathbb{E}\left[f(\mathbf{x}_t) \mid \mathbf{x}_0 = \mathbf{x}\right]. \tag{7}$$

Formally, when an SDE (3) has a stationary density $\pi^\star : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, i.e., drawing $\mathbf{x}_0 \sim \pi^\star$ yields $\mathbf{x}_t$ distributed according to $\pi^\star$ for all $t \geq 0$, an appropriate set of functions is

$$L^2(\pi^\star) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \int f(\mathbf{x})^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty \right\}. \tag{8}$$

For brevity, we assume that all functions discussed in this section are sufficiently smooth[6] and in $L^2(\pi^\star)$; all arguments can be formalized for a broader family of functions in $L^2(\pi^\star)$ by taking appropriate limits with an approximating sequence of smooth functions.

The Markov (memoryless) property of stochastic processes, including drift-diffusion processes, says that for $0 \leq s \leq t$, $\mathbf{x}_t$ is independent of $\mathcal{F}_s$ given $\mathbf{x}_s$. Because drift-diffusion processes are Markov, the law of iterated expectations shows that for all $s, t \geq 0$,

$$P_{t+s} f = P_t P_s f = P_s P_t f \text{ for all } f : \mathbb{R}^d \to \mathbb{R}, \tag{9}$$

and P_0 is clearly the identity operator, which justifies calling $\{P_t\}_{t\geq 0}$ a semigroup. In fact, (9) shows that it is additionally a *commutative* semigroup, which will be used later.

Next, given a Markov semigroup $\{P_t\}_{t\geq 0}$, we define the associated *generator* $\mathcal{L}$ which also acts on functions $f : \mathbb{R}^d \to \mathbb{R}$ via the following definition (which holds pointwise on $\mathbb{R}^d$):

$$\mathcal{L} f := \lim_{\eta \to 0} \frac{P_\eta f - f}{\eta}. \tag{10}$$

The generator commutes with all elements of the semigroup $\{P_t\}_{t\geq 0}$, due to the property (9).

Lemma 2. *For all $t \geq 0$ and $f : \mathbb{R}^d \to \mathbb{R}$, we have pointwise on $\mathbb{R}^d$,*

$$\frac{\partial}{\partial t} P_t f = \mathcal{L} P_t f = P_t \mathcal{L} f. \tag{11}$$

Proof. By using the commutativity property in (9),

$$\mathcal{L} P_t f = \lim_{\eta \to 0} \frac{P_\eta - P_0}{\eta} P_t f = \lim_{\eta \to 0} \frac{P_{t+\eta} f - P_t f}{\eta} = \lim_{\eta \to 0} P_t \frac{P_\eta - P_0}{\eta} f = P_t \mathcal{L} f$$

holds pointwise in $\mathbb{R}^d$. The conclusion follows since $\frac{\mathrm{d}}{\mathrm{d}t} P_t f = \lim_{\eta \to 0} \frac{P_{t+\eta} f - P_t f}{\eta}$ by definition. □

The expression (11) is sometimes called *Kolmogorov's backward equation*, a name which is perhaps best justified by also introducing *Kolmogorov's forward equation*. Let $\pi_0 \in \mathcal{P}(\mathbb{R}^d)$ be a density. Because of the definition (7), we have that for all $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E} f(\mathbf{x}_t) = \int P_t f(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int f(\mathbf{x}) P_t^* \pi_0(\mathbf{x}) \mathrm{d}x,$$

[6] When we say a function in this lecture is smooth, we mean that it has continuous derivatives of all orders.

where P_t^* is the *adjoint operator* to P_t.[7] This means that π_t, the law of $\mathbf{x}_t$, is given by $P_t^*\pi_0$, so P_t^* is formally the operator which advances π_0 forward by time t. In other words, if we want to understand the density π_t, i.e., the law of x_t, we should aim to understand P_t^*. Indeed, appropriately taking adjoints of Lemma 2 yields Kolmogorov's forward equation,

$$\frac{\partial}{\partial t}\underbrace{P_t^*\pi_0}_{=\pi_t} = \mathcal{L}^* P_t^*\pi_0 = P_t^*\mathcal{L}^*\pi_0, \tag{12}$$

so P_t^* and $\mathcal{L}^*$ also commute for all $t \geq 0$. In the case of drift-diffusion processes, there is in fact a convenient formula for P_t^*, enabling us to write a PDE for density evolution. We first remind the reader of the multivariate integration by parts formula, which holds for any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and vector field $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$ with sufficiently fast decay at infinity,[8]

$$\int \langle \mathbf{v}(\mathbf{x}), \nabla f(\mathbf{x})\rangle \, \mathrm{d}\mathbf{x} = -\int f(\mathbf{x})(\nabla \cdot \mathbf{v})(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{13}$$

where $\nabla\cdot$ is the divergence operator of a vector-valued function (for example, $\nabla \cdot \nabla = \mathrm{Tr}\nabla^2 = \Delta$).

Proposition 3 (Fokker-Planck equation)**.** *Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the drift-diffusion process (3) from $\mathbf{x}_0 \sim \pi_0 \in \mathcal{P}(\mathbb{R}^d)$. Then for all $t \geq 0$, denoting the law of $\mathbf{x}_t$ by π_t, we have*

$$\frac{\partial}{\partial t}\pi_t(\mathbf{x}) = -\nabla \cdot (\boldsymbol{\mu}(\mathbf{x})\pi_t(\mathbf{x})) + \frac{1}{2}\sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top \pi_t(\mathbf{x})\right]_{ij} \text{ for all } \mathbf{x} \in \mathbb{R}^d. \tag{14}$$

Proof. We first compute using Proposition 2 that for smooth $f : \mathbb{R}^d \to \mathbb{R}$, and $\mathbf{x}_t$ following (3),

$$\mathrm{d}\mathbb{E}f(\mathbf{x}_t) = \left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}(\mathbf{x}_t)\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\sigma}(\mathbf{x}_t)^\top\right\rangle\right)\mathrm{d}t,$$

where we drop the martingale term $\langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t\rangle$ because it vanishes in expectation. Therefore, by the definition (10) of $\mathcal{L}$, we have

$$\mathcal{L}f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \boldsymbol{\mu}(\mathbf{x})\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top\right\rangle, \tag{15}$$

so for all probability densities $\pi \in \mathcal{P}(\mathbb{R}^d)$ (which must decay to zero as $\mathbf{x} \to \infty$),

$$\begin{aligned}
\int f(\mathbf{x})\mathcal{L}^*\pi(\mathbf{x})\mathrm{d}\mathbf{x} &= \int \mathcal{L}f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int \left(\langle \nabla f(\mathbf{x}), \mu(\mathbf{x})\rangle + \frac{1}{2}\left\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top\right\rangle\right)\pi(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int \left(-f(\mathbf{x})\nabla \cdot (\boldsymbol{\mu}(\mathbf{x})\pi(\mathbf{x})) - \frac{1}{2}\left\langle \nabla f(\mathbf{x}), \nabla \cdot (\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top \pi(\mathbf{x}))\right\rangle\right)\mathrm{d}\mathbf{x} \\
&= \int f(\mathbf{x})\left(-\nabla \cdot (\boldsymbol{\mu}(\mathbf{x})\pi(\mathbf{x})) + \frac{1}{2}\sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top \pi(\mathbf{x})\right]_{ij}\right)\mathrm{d}\mathbf{x},
\end{aligned}$$

where we repeatedly applied the multivariate integration by parts formula (13). Because this holds for all test functions f and densities $\pi \in \mathcal{P}(\mathbb{R}^d)$, this means that pointwise,

$$\mathcal{L}^*\pi(\mathbf{x}) = -\nabla \cdot (\boldsymbol{\mu}(\mathbf{x})\pi(\mathbf{x})) + \frac{1}{2}\sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\left[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top \pi(\mathbf{x})\right]_{ij},$$

and the conclusion follows because $\frac{\partial}{\partial t}\pi_t = \mathcal{L}^*\pi_t$ by Kolmogorov's forward equation (12). □

[7] The adjoint to an operator P on a Hilbert space is denoted P^*, and satisfies $\langle Pf, g\rangle = \langle f, P^*g\rangle$. Here, the relevant Hilbert space is sufficiently regular functions on $\mathbb{R}^d$, and the inner product is $\langle f, g\rangle = \int f(\mathbf{x})g(\mathbf{x})\mathrm{d}\mathbf{x}$.

[8] The decay lets us discard the "boundary term" which typically arises in integration by parts formulas.

At this point, it is helpful to do an example. Consider the Langevin dynamics (4), where $\boldsymbol{\sigma}(\mathbf{x}) = \sqrt{2}\mathbf{I}_d$ for all $\mathbf{x} \in \mathbb{R}^d$, so that $\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top = 2\mathbf{I}_d$, so the formula (14) reads

$$\frac{\partial}{\partial t}\pi_t(\mathbf{x}) = \nabla \cdot (\nabla V(\mathbf{x})\pi_t(\mathbf{x})) + \Delta\pi_t(\mathbf{x}). \tag{16}$$

The contribution of the component $\Delta\pi_t(\mathbf{x})$ is the *heat equation*, and by going back through the calculations, we can check that it arose because of the presence of Brownian motion $\sqrt{2}\mathrm{d}\boldsymbol{B}_t$ in (4). Intuitively, this component is because of the dissipation of heat, where each point contributes symmetrically to its surroundings. On the other hand, the $\nabla\cdot(\nabla V(\mathbf{x})\pi_t(\mathbf{x}))$ term is the contribution of the drift, and combining characterizes the stationary distribution of the Langevin dynamics.

Theorem 1 (Langevin diffusion stationarity)**.** *Suppose that $V : \mathbb{R}^d \to \mathbb{R}$ is differentiable and $\int \exp(-V(x))\mathrm{d}x < \infty$. Then, a stationary distribution for the Langevin dynamics (4) is given by the density $\pi^\star : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ satisfying $\pi^\star \propto \exp(-V)$.*

Proof. By the definition of the stationary distribution, we have that $\frac{\partial}{\partial t}\pi_t$ vanishes pointwise when $\pi_t = \pi^\star$. By reparameterizing $U := -\log \pi^\star$, i.e., $\pi^\star \propto \exp(-U)$, (16) gives that pointwise

$$0 = -\nabla \cdot (\nabla V(\mathbf{x})\pi^\star(\mathbf{x}) + \nabla\pi^\star(\mathbf{x})) = -\nabla \cdot ((\nabla V(\mathbf{x}) - \nabla U(\mathbf{x}))\pi^\star(\mathbf{x}))\,,$$

which is solved by setting $U = V$ up to addition by a universal constant. □

In general, it is not necessarily the case that the zero vector field $\mathbf{v}$ is the only solution to $\nabla\cdot(\mathbf{v}\pi^\star) = 0$ pointwise, but we will soon give conditions (in Section 5) under which $\pi^\star \propto \exp(-V)$ is the unique stationary distribution for the Langevin dynamics. We pause here to draw some analogies between spectral graph theory, introduced in Part XIV, and the developments of this section. Intuitively, $\mathcal{L}$ plays the role of an infinitesimal transition operator from functions to functions, and its adjoint acts on probability densities. By solving Kolmogorov's forward equation (12), we should heursitically expect the distribution π_t to follow the law

$$\pi_t = \exp\left(t\mathcal{L}^*\right)\pi_0.$$

Advancing time for $t = 1$ units, this means that the state evolution of π_t roughly follows $\pi_{t+1} = \exp(\mathcal{L}^*)\pi_t \approx (\mathrm{id} + \mathcal{L}^*)\pi_t$, so $\exp(\mathcal{L}^*)$ acts like the discrete-time transition operator. We also expect that as $t \to \infty$, $\pi_t \to \pi^\star$ (the stationary distribution of the semigroup $\{P_t\}_{t\geq 0}$), so if we view $\mathcal{L}^*$ as an infinite-dimensional operator, it should have a leading eigenfunction of $\pi^\star$ with eigenvalue 0 (so its exponential has eigenvalue 1), and if $\pi^\star$ is unique, then all other eigenvalues of $\mathcal{L}^*$ should be negative (so that they rapidly decay with time after exponentiation). To formalize this spectral characterization of $\mathcal{L}$ (which in the spectral graph theory case only made sense when the transition operator was reversible), it is helpful to introduce the analogous definition for Markov semigroups.

Definition 2. *Let $\{P_t\}_{t\geq 0}$ be a Markov semigroup with generator $\mathcal{L}$, and stationary distribution $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$. We say $\{P_t\}_{t\geq 0}$ is reversible if for all $f, g \in L_2(\pi^\star)$, following notation (8),*

$$\int \mathcal{L}f(\mathbf{x})g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = \int \mathcal{L}g(\mathbf{x})f(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}.$$

In the spectral graph theory case, the transition operator $\mathbf{P} = \mathbf{D}_G^{-1}\mathbf{A}_G$ corresponded to a degree-normalized adjacency matrix of a graph G, so reversibility corresponded to $\mathbf{D}_G\mathbf{P} = \mathbf{P}^\top\mathbf{D}_G$, or alternatively $\mathbf{D}_G^{\frac{1}{2}}\mathbf{P}\mathbf{D}_G^{-\frac{1}{2}}$ being a symmetric matrix. In this case, the stationary distribution was $\propto \mathbf{d}_G$, the degrees of G. Analogously, we can interpret the condition in Definition 2 as saying

$$\text{“}\left\langle \mathcal{L}f, \mathbf{diag}\left(\pi^\star\right) g\right\rangle = \left\langle \mathbf{diag}\left(\pi^\star\right) f, \mathcal{L}g\right\rangle \iff \mathcal{L}^*\mathbf{diag}\left(\pi^\star\right) = \mathbf{diag}\left(\pi^\star\right)\mathcal{L},\text{”}$$

which also justifies the definition being an integration with respect to $\pi^\star$, since we need to conjugate the generator by $\pi^\star$ to make it truly self-adjoint. If we let $f = \mathbf{1}_E$ and $g = \mathbf{1}_{E'}$ be indicators of subsets E and E' of $\mathbb{R}^d$, integrating Definition 2 for time t implies that

$$\begin{aligned}
&\int P_t\mathbf{1}_E(\mathbf{x})\mathbf{1}_{E'}(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = \int P_t\mathbf{1}_{E'}(\mathbf{x})\mathbf{1}_E(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\implies \Pr_{\mathbf{x}_0\sim\pi^\star}\left[\mathbf{x}_0 \in E', \mathbf{x}_t \in E\right] = \Pr_{\mathbf{x}_0\sim\pi^\star}\left[\mathbf{x}_t \in E', \mathbf{x}_0 \in E\right].
\end{aligned}$$

This holds for all events E, E', so we can verify that $(\mathbf{x}_0, \mathbf{x}_t)$ has the same joint distribution as $(\mathbf{x}_t, \mathbf{x}_0)$. This can be interpreted as the fact that the reversed-time Markov semigroup, i.e., that which follows (12) but flips time, induces the same evolutions as the original Markov semigroup.

Finally, we wish to formalize our earlier intuition of a "spectral gap" driving all functions orthogonal to $\pi^\star$ rapidly to zero, so that the transition operator only preserves the stationary distribution. By using Jensen's inequality with the definition (7), observe that for all functions $f \in L^2(\pi^\star)$,

$$(P_t f(\mathbf{x}))^2 = \mathbb{E}\left[f(\mathbf{x}_t) \mid \mathbf{x}_0 = \mathbf{x}\right]^2 \leq \mathbb{E}\left[f(\mathbf{x}_t)^2 \mid \mathbf{x}_0 = \mathbf{x}\right] = P_t(f^2)(\mathbf{x}), \text{ for all } \mathbf{x} \in \mathbb{R}^d. \tag{17}$$

Moreover, the above inequality is tight only when f is a constant function, corresponding to $\mathcal{L}$ having a constant right eigenfunction (just as graph transition operators had a right eigenfunction of $\mathbf{1}_V$). Our hope is to establish a quantitative variant of (17), where the inequality is strict for all other functions. When $\{P_t\}_{t\geq 0}$ is reversible with stationary distribution $\pi^\star$, we introduce the *carré du champ* operator Γ, and its integral $\mathcal{E}$ (called the Dirichlet energy), defined as follows:

$$\Gamma(f,g)(\mathbf{x}) := -f(\mathbf{x})\mathcal{L}g(\mathbf{x}),\ \mathcal{E}(f,g) := \int \Gamma(f,g)(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{18}$$

so that the reversibility assumption in Definition 2 gives $\mathcal{E}(f,g) = \mathcal{E}(g,f)$. As we explore in Section 5, the question of a spectral gap in (17) is really about understanding the spectrum of the symmetric operator $\mathcal{E}$. To be more concrete, we conclude the section by working out $\mathcal{E}$ explicitly when $\{P_t\}_{t\geq 0}$ is the Markov semigroup of the Langevin diffusion (4).

Lemma 3. *Following notation (18), if $\{P_t\}_{t\geq 0}$ is the Markov semigroup corresponding to the Langevin diffusion (4) and $\int \exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} < \infty$, then for $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\pi^\star \propto \exp(-V)$,*

$$\mathcal{E}(f,g) = \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle\, \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \text{ for all } f, g \in L^2(\pi^\star).$$

Proof. Recall from the calculation (15), specialized to the Langevin diffusion (4), that

$$\begin{aligned}
\mathcal{E}(f,g) &= -\int \mathcal{L}f(\mathbf{x})g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int \langle \nabla f(\mathbf{x}), \nabla V(\mathbf{x})\rangle\, g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} + \int \Delta f(\mathbf{x})g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int \langle \nabla f(\mathbf{x}), \nabla V(\mathbf{x})\rangle\, g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} + \int \langle \nabla f(\mathbf{x}), \nabla (g\cdot\pi^\star)(\mathbf{x})\rangle\, \mathrm{d}\mathbf{x} \\
&= \int \langle \nabla f(\mathbf{x}), \nabla V(\mathbf{x})\rangle\, g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} - \int \langle \nabla f(\mathbf{x}), \nabla V(\mathbf{x})\rangle\, g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\quad + \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle\, \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle\, \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x},
\end{aligned}$$

as claimed. The third line above used integration by parts (13) (with $\Delta f = \nabla\cdot(\nabla f)$), and the fourth line used that $\pi^\star(\mathbf{x}) = \frac{\exp(-V(\mathbf{x}))}{Z}$ for a universal normalizing constant Z, so that

$$-\nabla V(\mathbf{x})\pi^\star(\mathbf{x}) = -\nabla V(\mathbf{x})\cdot\frac{\exp(-V(\mathbf{x}))}{Z} = \nabla\left(\frac{\exp(-V(\mathbf{x}))}{Z}\right) = \nabla\pi^\star(\mathbf{x}).$$

□

Notice in particular that when f or g is a constant function, then $\mathcal{E}(f,g) = 0$ by Lemma 3, so the constant function is an eigenfunction of $\mathcal{E}$. The question of a spectral gap then corresponds to characterizing the eigenvalues of all eigenfunctions of $\mathcal{E}$ orthogonal to the constant function (which is exactly the space where probability densities are allowed to evolve, since they must stay densities). Lemma 3 also justifies the name "carré du champ," which is French for "square of a field." Here, we mean the vector field ∇f, which is relevant because $\mathcal{E}(f,f) = \int \|\nabla f(\mathbf{x})\|_2^2\, \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}$. Further, Lemma 3 proves that $\mathcal{E}(f,g) = \mathcal{E}(g,f)$, so the Langevin diffusion is indeed reversible.

Remark 3. *Markov semigroup theory extends to the discrete setting as well, where a continuous-time Markov chain is governed by a matrix specifying transition probabilities for a continuous-time jump process. The continuous-time analog of the "transition matrix" is a family of matrices $\{\exp(t\mathbf{L})\}_{t\geq 0}$ for a generator matrix $\mathbf{L}$, specifying transition probabilities after t time has passed.*

3 Optimal transport

In this section, we take a brief digression to introduce a notion of distance between probability measures in $\mathcal{P}(\mathbb{R}^d)$, which is important for measuring convergence rates of stochastic processes to their stationary distribution. We begin by defining the relevant distance we will consider.

Definition 3 (Wasserstein distance)**.** *Let $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$ have finite second moments, i.e., $\mathbb{E}_{\mathbf{x}\sim\mu}[\|\mathbf{x}\|_2^2] < \infty, \mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^2] < \infty$. We define the* Wasserstein distance *between μ and π by*

$$W_2(\mu,\pi) := \inf_{\gamma\in\mathcal{C}(\mu,\pi)} \sqrt{\int \|\mathbf{x}-\mathbf{y}\|_2^2 \,\gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}}, \tag{19}$$

where $\mathcal{C}(\mu,\pi)$ is the set of all couplings of μ and π (see Fact 1, Part XI).

In other words, the squared distance $W_2^2(\mu,\pi)$ is induced by the coupling γ which minimizes the expected squared distance between $(x,y) \sim \gamma$. More generally, there is also a family of p-Wasserstein distances $W_p(\mu,\pi)$, which generalize Definition 3. In fact, historically it was the W_1 distance which was first considered by Monge in the 1700s (who only considered a restricted set of deterministic mappings $\mathbf{x} \to \mathbf{y}$), and later Kantorovich during World War II (who introduced the more general coupling definition), who were interested in the transportation of resources. The study of optimization problems of the form in Definition 3 is hence called *optimal transport*. For an extended introduction to optimal transport and its applications, we recommend the text [Vil08]; for a somewhat briefer summary of some of the key technical tools, see [MG10].

One useful observation for our purposes is the fact that W_2 induces a metric on

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) \mid \int \|\mathbf{x}\|_2^2 \,\pi(\mathbf{x})\mathrm{d}\mathbf{x} < \infty\right\},$$

the set of densities with finite second moment. Hence, it yields a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

Lemma 4. *For all $\mu, \nu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, we have $W_2(\mu,\pi) \le W_2(\mu,\nu) + W_2(\nu,\pi)$.*

Proof. Let $\gamma_{\mu,\nu} \in \mathcal{C}(\mu,\nu)$ be the optimal coupling of μ,ν realizing the value $W_2(\mu,\nu)$, and similarly let $\gamma_{\nu,\pi} \in \mathcal{C}(\nu,\pi)$ realize the value $W_2(\nu,\pi)$. We first claim that there exists a joint distribution $(\mathbf{x},\mathbf{y},\mathbf{z}) \sim \gamma_{\mu,\nu,\pi}$ on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, such that the marginal distribution of $(\mathbf{x},\mathbf{y})$ is $\gamma_{\mu,\nu}$, and the marginal distribution of $(\mathbf{y},\mathbf{z})$ is $\gamma_{\nu,\pi}$. To see this, we can first draw $\mathbf{y} \sim \nu$, and then draw $\mathbf{x}$ and $\mathbf{z}$ from the conditional distributions of $\gamma_{\mu,\nu} \mid \mathbf{y}$ and $\gamma_{\nu,\pi} \mid \mathbf{y}$, respectively.

Next, the marginal distribution of $(\mathbf{x},\mathbf{z})$ for $(\mathbf{x},\mathbf{y},\mathbf{z}) \sim \gamma_{\mu,\nu,\pi}$ is a valid coupling of μ,π, so

$$\begin{aligned}
W_2(\mu,\pi) &\le \sqrt{\int \|\mathbf{x}-\mathbf{z}\|_2^2 \,\gamma_{\mu,\nu,\pi}(\mathbf{x},\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}} \\
&\le \sqrt{\int \left(\|\mathbf{x}-\mathbf{y}\|_2 + \|\mathbf{y}-\mathbf{z}\|_2\right)^2 \gamma_{\mu,\nu,\pi}(\mathbf{x},\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}} \\
&\le \sqrt{\int \|\mathbf{x}-\mathbf{y}\|_2^2 \,\gamma_{\mu,\nu,\pi}(\mathbf{x},\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}} + \sqrt{\int \|\mathbf{y}-\mathbf{z}\|_2^2 \,\gamma_{\mu,\nu,\pi}(\mathbf{x},\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}} \\
&= \sqrt{\int \|\mathbf{x}-\mathbf{y}\|_2^2 \,\gamma_{\mu,\nu}(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}} + \sqrt{\int \|\mathbf{y}-\mathbf{z}\|_2^2 \,\gamma_{\nu,\pi}(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}} = W_2(\mu,\nu) + W_2(\nu,\pi).
\end{aligned}$$

The second line used the triangle inequality, and the inequality in the third line follows by squaring both sides and bounding cross-terms. The last line used our construction of $\gamma_{\mu,\nu,\pi}$. □

We mention that while Lemma 4 shows W_2 obeys the triangle inequality, it is not necessarily the case that $W_2(\mu,\pi) = 0 \implies \mu = \pi$. However, this conclusion is true except for on a set of measure zero, so we will treat W_2 as a genuine metric under this set of equivalence classes.

We conclude the section by introducing the fundamental theorem of optimal transport, and providing a brief proof sketch. This result gives a remarkable characterization of the optimal transport

coupling γ inducing W_2, and establishes that strong duality holds for the corresponding problem. Specifically, note that $\mathcal{C}(\mu,\pi)$ is a convex subset of the probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ (since taking convex combinations does not affect marginals), and further

$$\frac{1}{2}W_2^2(\mu,\pi) = \inf_{\gamma\in\mathcal{C}(\mu,\pi)} \int \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{20}$$

is a linear optimization problem over $\mathcal{C}(\mu,\pi)$ (i.e., the objective above is linear in γ). Therefore, we may consider its Lagrangian dual, which is derived via

$$\begin{aligned}
&\inf_{\gamma\in\mathcal{C}(\mu,\pi)} \int \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= \inf_{\gamma\in\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \sup_{\substack{f\in L^1(\mu)\\ g\in L^1(\pi)}} \int \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \int f(\mathbf{x})\mu(\mathbf{x})\mathrm{d}\mathbf{x} - \int f(\mathbf{x})\gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&\qquad + \int g(\mathbf{y})\pi(\mathbf{y})\mathrm{d}\mathbf{y} - \int g(\mathbf{y})\gamma(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},
\end{aligned}$$

where $L^1(\pi)$ is the set of functions with finite expectation with respect to π (analogously to (8)). The role of f and g above is to enforce the marginal constraints in the unconstrained problem over γ, forcing $\gamma \in \mathcal{C}(\mu,\pi)$ at optimality. Now, if strong duality holds, the above expression equates to

$$\begin{aligned}
&\sup_{(f,g)\in\mathcal{D}(\mu,\pi)} \int f(\mathbf{x})\mu(\mathbf{x})\mathrm{d}\mathbf{x} + \int g(\mathbf{y})\pi(\mathbf{y})\mathrm{d}\mathbf{y}, \\
&\text{where } \mathcal{D}(\mu,\pi) := \left\{(f,g)\in L^1(\mu)\times L^1(\pi) \mid f(\mathbf{x})+g(\mathbf{y}) \le \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \text{ for all } \mathbf{x},\mathbf{y}\in\mathbb{R}^d\right\}.
\end{aligned} \tag{21}$$

We think of $\mathcal{D}(\mu,\pi)$ as the set of dual feasible potentials (f,g) inducing the dual optimization problem. Incredibly, under mild conditions strong duality does hold, and we also have a complete characterization of the optimal coupling $\gamma \in \mathcal{C}(\mu,\pi)$ and the optimal dual potentials (f,g).

Proposition 4 (Brenier's theorem)**.** *If $\mu,\pi \in \mathcal{P}_2(\mathbb{R}^d)$, the values of (20) and (21) are equal (i.e., strong duality holds), and both values are realized, respectively by $\gamma^\star \in \mathcal{C}(\mu,\pi)$ and $(f^\star,g^\star) \in \mathcal{D}(\mu,\pi)$. Moreover, there exists a unique convex function $\varphi:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ such that*

$$f^\star(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \varphi(\mathbf{x}),\ g^\star(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \varphi^*(\mathbf{y}),$$

and the optimal coupling $\gamma^\star$ is supported only on points $(\mathbf{x},\mathbf{y})$ where $\mathbf{y}\in\partial\varphi(\mathbf{x})$.

We remark that Alexandrov's theorem tells us that convex functions are differentiable almost everywhere, so the support is almost surely of the form $(\mathbf{x},\nabla\varphi(\mathbf{x}))$. In other words, the optimal transport plan from μ to π realizing W_2^2 is actually a unique deterministic mapping $\nabla\varphi$, which sends $\mathbf{x}\sim\mu$ to $\nabla\varphi(\mathbf{x})\sim\pi$, except on a measure-zero subset. As a sanity check, we indeed have

$$\begin{aligned}
f^\star(\mathbf{x}) + g^\star(\mathbf{y}) &= \frac{1}{2}\|\mathbf{x}\|_2^2 - \varphi(\mathbf{x}) + \frac{1}{2}\|\mathbf{y}\|_2^2 - \varphi^*(\mathbf{y}) \\
&= \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 - \varphi(\mathbf{x}) - \varphi^*(\nabla\varphi(\mathbf{x})) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,
\end{aligned}$$

for all $(\mathbf{x},\mathbf{y}) = (\mathbf{x},\nabla\varphi(\mathbf{x}))$ in the support of $\gamma^\star$, where we recall from Corollary 1, Part III, that

$$\varphi(\mathbf{x}) + \varphi^*(\nabla\varphi(\mathbf{x})) = \langle\nabla\varphi(\mathbf{x}),\mathbf{x}\rangle = \langle\mathbf{y},\mathbf{x}\rangle.$$

Therefore, all of the constraints in $\mathcal{D}(\mu,\pi)$ are tight for the optimal dual potentials $(f^\star,g^\star)$. The induced convex function φ is often called a Brenier potential, and much of Proposition 4 was established by the works [Bre87, Bre91]. The proof of Proposition 4 is a bit tedious for this brief exposition; a summary of it can be found in Section 1.3 of [Che24]. However, we mention one crucial step of the proof which may be of broader interest to the reader.

Fact 1 ([Roc70])**.** *If $S\subseteq\mathbb{R}^d\times\mathbb{R}^d$ is* cyclically monotone*, i.e., for all $n\in\mathbb{N}$ and all permutations $\sigma:[n]\to[n]$, we have for all $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i\in[n]}\subseteq S$,*

$$\sum_{i\in[n]}\langle\mathbf{x}_i,\mathbf{y}_i\rangle \ge \sum_{i\in[n]}\langle\mathbf{x}_i,\mathbf{y}_{\sigma(i)}\rangle, \tag{22}$$

then there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *such that*

$$S \subseteq \left\{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \partial\varphi(x)\right\}.$$

One can check that pairs $(\mathbf{x}, \partial\varphi(\mathbf{x}))$ do indeed induce cyclically monotone sets, which follows from monotonicity of convex gradients (see Definition 2, Part IV). The characterization in Proposition 4 then follows because optimal transport plans induce cyclically monotone subsets. Indeed, if (22) did not hold for some pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]}$ in the support of the transport plan, then we could improve the W_2^2 objective by pairing up the transported points differently:

$$\sum_{i\in[n]} \langle \mathbf{x}_i, \mathbf{y}_i \rangle \le \sum_{i\in[n]} \langle \mathbf{x}_i, \mathbf{y}_{\sigma(i)} \rangle \implies \sum_{i\in[n]} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 \ge \sum_{i\in[n]} \|\mathbf{x}_i - \mathbf{y}_{\sigma(i)}\|_2^2.$$

Remark 4 (Caffarelli's contraction theorem)**.** *Under further structural assumptions on the measures* μ, π *in Definition 3, one can say more about the optimal transport map* $\partial\varphi$ *described in Proposition 4. For example, a famous result by [Caf00] states that if* $\mu \propto \exp(-V)$ *and* $\pi \propto \exp(-W)$ *where* V *is* L*-smooth*[9] *and* W *is* m*-strongly convex, then the optimal transport map is induced by a differentiable convex function* φ *which is* $\sqrt{L/\mu}$*-smooth. For a short proof of this fact and a generalization to entropic variants of optimal transport, see [CP23].*

We conclude this section with our first convergence analysis for the Langevin dynamics, albeit in continuous time (which does not give an implementable algorithm). We will later analyze a discretized version of the following result. Before stating our claim, we require one definition.

Definition 4 (Strong logconcavity)**.** *We say* $\pi \in \mathcal{P}(\mathbb{R}^d)$ *is* μ-strongly logconcave *if* $\pi \propto \exp(-V)$ *for* $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$*, such that* V *is* μ*-strongly convex.*

For instance, the multivariate normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ for $\mathbf{m} \in \mathbb{R}^d$, $\mathbf{\Sigma} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}$, whose density is $\propto \exp(-\frac{1}{2}\|\cdot - \mathbf{m}\|^2_{\mathbf{\Sigma}^{-1}})$, is μ-strongly logconcave for any value of $\mu > 0$ with $\mathbf{\Sigma}^{-1} \succeq \mu \mathbf{I}_d$.

Theorem 2 (Wasserstein convergence of Langevin dynamics)**.** *Let* $\{\mathbf{x}_t\}_{t\ge 0}$ *follow the Langevin dynamics (4) with stationary distribution* $\pi^\star$*, and assume that* $\pi^\star$ *is* μ*-strongly logconcave. For all* $t \ge 0$*, the law of* $\mathbf{x}_t$*, denoted* $\pi_t \in \mathcal{P}(\mathbb{R}^d)$*, satisfies*

$$W_2^2(\pi_t, \pi^\star) \le \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star).$$

Proof. Let γ_t be a coupling of $(\pi_t, \pi^\star)$ defined as follows. Let $\gamma_0^\star$ be the coupling of $(\pi_0, \pi^\star)$ which realizes $W_2^2(\pi_0, \pi^\star)$, draw $(\mathbf{x}_0, \mathbf{x}_0^\star) \sim \gamma_0^\star$, and advance both points using the Langevin dynamics (4), sharing the same copy of Brownian motion $\{\boldsymbol{B}_s\}_{0\le s\le t}$. It is clear that the marginals of γ_t are $(\pi_t, \pi^\star)$ respectively (the latter because $\pi^\star$ is stationary for (4)). Moreover, letting $\{(\mathbf{x}_s, \mathbf{x}_s^\star)\}_{s\in[0,t]}$ be advanced through (4) using the same copy of Brownian motion, so $(\mathbf{x}_t, \mathbf{x}_t^\star) \sim \gamma_t$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}s}(\mathbf{x}_s - \mathbf{x}_s^\star) = \nabla V(\mathbf{x}_s^\star) - \nabla V(\mathbf{x}_s) \\
\implies &\frac{\mathrm{d}}{\mathrm{d}s}\|\mathbf{x}_s - \mathbf{x}_s^\star\|_2^2 = -2\langle \nabla V(\mathbf{x}_s) - \nabla V(\mathbf{x}_s^\star), \mathbf{x}_s - \mathbf{x}_s^\star \rangle \le -2\mu \|\mathbf{x}_s - \mathbf{x}_s^\star\|_2^2 \\
\implies &\|\mathbf{x}_t - \mathbf{x}_t^\star\|_2^2 \le \exp(-2\mu t)\, \|\mathbf{x}_0 - \mathbf{x}_0^\star\|_2^2.
\end{aligned}$$

Above, the second inequality used Gronwall's inequality (Fact 1, Part II), and the first inequality used that strong convexity of V implies (e.g., by adding Eq. (9), Part II with x, x' interchanged)

$$\langle \nabla V(\mathbf{x}) - \nabla V(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \ge \mu \|\mathbf{x} - \mathbf{x}'\|_2^2 \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathrm{dom}(V).$$

Finally, because γ_t is a coupling of $(\pi_t, \pi^\star)$ (so it has no better objective value for (19) than the optimal coupling), we conclude that

$$\begin{aligned}
W_2^2(\pi_t, \pi^\star) &\le \mathbb{E}_{(\mathbf{x}_t, \mathbf{x}_t^\star)\sim\gamma_t}\left[\|\mathbf{x}_t - \mathbf{x}_t^\star\|_2^2\right] \\
&\le \exp(-2\mu t)\, \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_0^\star)\sim\gamma_0^\star}\left[\|\mathbf{x}_0 - \mathbf{x}_0^\star\|_2^2\right] = \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star).
\end{aligned}$$

□

[9] Here we mean smoothness in the sense of Definition 3, Part II.

We also give a basic initialization strategy under strong logconcavity. The following result shows that initializing at π_0 set to a point mass at $\mathbf{x}^\star$ achieves bounded $W_2^2(\pi_0, \pi^\star)$. The proof is tolerant to approximate minimizers; finding such a point typically does not dominate the cost of running a sampler (as state-of-the-art optimization rates are better than their sampling counterparts).

Lemma 5. *Let $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfy $\pi^\star \propto \exp(-V)$, and suppose $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is μ-strongly convex and minimized at $\mathbf{x}^\star$. Then,*

$$\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\|\mathbf{x} - \mathbf{x}^\star\|_2^2\right] \le \frac{2d}{\mu}.$$

Proof. Shifting V by a constant so $\int \exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} = 1$, using $\langle \nabla V(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle \ge \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|_2^2$ which follows from strong logconcavity where $V(\mathbf{x}) < \infty$,[10] and integrating by parts (using (13)),

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\|\mathbf{x} - \mathbf{x}^\star\|_2^2\right] &= \int \|\mathbf{x} - \mathbf{x}^\star\|_2^2 \exp\left(-V(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\
&\le \frac{2}{\mu}\int \langle \nabla V(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle \exp\left(-V(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\
&= \frac{2}{\mu}\int (\nabla\cdot(\cdot - \mathbf{x}^\star))(\mathbf{x}) \exp\left(-V(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\
&= \frac{2}{\mu}\int \mathrm{Tr}(\mathbf{I}_d)\exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} = \frac{2d}{\mu}.
\end{aligned}$$

□

Theorem 2 has the appealing property of giving a linear rate of convergence, but the convergence guarantee is in the Wasserstein distance W_2, which may be less flexible in downstream applications. For instance, W_2 is typically incomparable to the total variation distance D_{TV}, and when designing sampling algorithms to be used as subroutines (i.e., to be called multiple times), D_{TV} guarantees are preferable because they compose under the union bound. In Section 5, we introduce techniques for proving convergence of the Langevin dynamics and other stochastic differential equations in stronger error metrics such as the KL divergence (which implies bounds on D_{TV} by Pinsker's inequality, as well as bounds on W_2 under appropriate conditions discussed in Section 5).

4 Probability densities as a Riemannian manifold

4.1 Riemannian manifolds

In this section, we build upon our development of optimal transport to present a view of the space of square-integrable probability densities $\mathcal{P}_2(\mathbb{R}^d)$ as a differentiable (in particular, Riemannian) manifold. It is first helpful to give some brief description of what differentiable manifolds are, and the sorts of calculations we can expect to arise from manipulating them. We informally do so here, focusing on presenting only the relevant material in a way that motivates the calculations specific to $\mathcal{P}_2(\mathbb{R}^d)$. For a more extended discussion on these topics, we suggest [Vis18] as a resource for learning about Riemannian manifolds in a way that is accessible to a computer science audience, as well as [dC92] for a more rigorous treatment from a mathematical perspective.

Roughly speaking, a manifold $\mathcal{M}$ is a topological space such that every $p \in \mathcal{M}$ has a open neighborhood U_p that "looks like" Euclidean space. Formally, in the finite-dimensional setting this means that there is a homeomorphism between U_p and an open set in $\mathbb{R}^k$. In many applications, this is interesting because $\mathcal{M}$ is actually a subset of a higher-dimensional Euclidean space (e.g., $\mathbb{R}^d$ for $d \ge k$). A canonical example is the surface of the unit sphere in $\mathbb{R}^3$, which is a 2-dimensional manifold (because neighborhoods of points are homeomorphic to balls in $\mathbb{R}^2$). This definition extends straightforwardly to Hilbert spaces, which are inner product spaces (possibly of infinite dimension) that generalize Euclidean space. For instance, $L^2(\mathbb{R}^d)$, the set of all square-integrable functions in $\mathbb{R}^d$ (i.e., $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int f(\mathbf{x})^2\mathrm{d}\mathbf{x} < \infty$) is a Hilbert space equipped with the inner product

$$\langle f, g\rangle := \int f(\mathbf{x})g(\mathbf{x})\mathrm{d}\mathbf{x}.$$

[10] When $V : \mathbb{R}^d \to \mathbb{R}$ is unconstrained so that $\nabla V(x^\star) = \mathbf{0}_d$ by first-order optimality, the constant factor can be sharpened by using $\langle \nabla V(\mathbf{x}) - \nabla V(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle \ge \mu\|\mathbf{x} - \mathbf{x}^\star\|_2^2$.

In particular, note that $\mathcal{P}_2(\mathbb{R}^d)$ is a subset of the space of square-integrable functions, and indeed we will view it as a manifold embedded in $L^2(\mathbb{R}^d)$. We now mention some definitions tailored to a specific family of manifolds which is convenient for calculation purposes, namely *Riemannian manifolds*. In a Riemannian manifold $\mathcal{M}$, every point $p \in \mathcal{M}$ has an associated *tangent space* $T_p\mathcal{M}$. Informally, this should be viewed as the set of all possible "local velocities" of curves on $\mathcal{M}$ passing through p. Notice that we treat points on $\mathcal{M}$ (which should be thought of as representing position) differently than elements of a tangent space $T_p\mathcal{M}$ (which should be thought of as representing velocities). Also, in general elements of $T_p\mathcal{M}$ and $T_q\mathcal{M}$ for $p, q \in \mathcal{M}$, $p \neq q$ are not directly comparable (e.g., we cannot simply add or subtract them directly, since they live in different spaces). To compare them, we should first transport $T_p\mathcal{M}$ to $T_q\mathcal{M}$ in an appropriate way along the manifold, which is done using a construction called the *Levi-Civita connection*. We will not explicitly describe this connection in the remainder of the lecture for brevity's sake, but mention this caveat to caution the reader when performing calculations on manifolds.

There are two main properties of Riemannian manifolds that are convenient for our purposes. The first is that every point $p \in \mathcal{M}$ comes with a local metric g_p which can be used to perform calculations specific to $T_p\mathcal{M}$. Specifically, there is a way to assign values $g_p(u, v)$ such that $g_p(u, v) = g_p(v, u)$ for all $u, v \in T_p\mathcal{M}$, and $g_p(\cdot, w)$ is linear in its first argument for all $w \in T_p\mathcal{M}$. To help build intuition for this, when we view $\mathbb{R}^d$ as a trivial manifold in $\mathbb{R}^d$, g_p is simply the standard Euclidean inner product $\langle \cdot, \cdot \rangle$ pointwise, i.e., it does not change depending on $p \in \mathbb{R}^d$ (because the space is not "curved"). Another interesting example is the *Hessian manifold* of a self-concordant barrier function. For a subset $\mathcal{K} \subseteq \mathbb{R}^d$ equipped with a self-concordant barrier function $\varphi : \mathcal{K} \to \mathbb{R}$ (see Definition 2, Part X), we can define associated local metrics satisfying

$$g_{\mathbf{p}}(\mathbf{u}, \mathbf{v}) := \nabla^2 \varphi(\mathbf{p})[u, v] \text{ for all } \mathbf{p} \in \mathcal{K},\ \mathbf{u}, \mathbf{v} \in \mathbb{R}^d.$$

Note that in this example, each tangent space $T_{\mathbf{p}}\mathcal{M}$ is d-dimensional (i.e., it is not actually lower-dimensional than its ambient space). The role of defining the local metrics $g_{\mathbf{p}}$ above is to "curve space" in a way that can helpfully guide algorithm design, e.g., the Newton's method we analyzed in Part X is just a discretization of a standard gradient flow from the perspective of the local metrics, which reweight the different directions in $T_p\mathcal{M}$ using $\nabla^2 \varphi(\mathbf{p})$.[11]

We can correspondingly measure the "length" and "correlation" of elements in $T_p\mathcal{M}$ by defining the norm and local inner product,

$$\|v\|_p := \sqrt{g_p(v, v)},\ \langle u, v \rangle_p := g_p(u, v),\ \text{for all } u, v \in T_p\mathcal{M}. \tag{23}$$

The second main property we need is that there is a meaningful notion of differentiability associated with functions on the manifold, $\mathcal{F} : \mathcal{M} \to \mathbb{R}$. We begin by defining geodesics, which informally are length-minimizing curves along $\mathcal{M}$ which do not leave the manifold. Concretely, given two points $p, q \in \mathcal{M}$, we denote their distance along the manifold by $\mathrm{d}_{\mathcal{M}}(p, q)$, which is defined by

$$\mathrm{d}_{\mathcal{M}}(p, q) := \inf_{\substack{\gamma:[0,1]\to\mathcal{M} \\ \gamma(0)=p, \gamma(1)=q}} \left\{ \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)} \, \mathrm{d}t \right\}. \tag{24}$$

We pause to explain the above formula. The infimum is taken with respect to all curves $\gamma : [0, 1] \to \mathcal{M}$, which travel for one unit of time along $\mathcal{M}$, starting from $\gamma(0) = p$ and ending at $\gamma(1) = q$. Moreover, $\dot{\gamma}(t) \in T_{\gamma(t)}\mathcal{M}$ is the velocity vector of the curve at time t, whose length is measured in the local metric $\|\cdot\|_{\gamma(t)}$. So, the formula (24) measures the total magnitude of the distance accumulated over the curve, a meaningful generalization of the "length" of a curve. When the infimum is achieved, we call the length-minimizing curve γ which realizes the value of (24) the *geodesic* between p and q. Under mild conditions, geodesics actually have constant speed $\|\dot{\gamma}(t)\|_{\gamma(t)}$ along their trajectories, so an equivalent definition is

$$\mathrm{d}_{\mathcal{M}}(p, q) := \inf_{\substack{\gamma:[0,1]\to\mathcal{M} \\ \gamma(0)=p, \gamma(1)=q}} \left\{ \sqrt{\int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)}^2 \, \mathrm{d}t} \right\}. \tag{25}$$

[11] A helpful example to consider is $\mathcal{M} = \mathbb{R}^d_{>0}$, equipped with the self-concordant barrier function $\varphi(\mathbf{p}) := -\sum_{i \in [d]} \log \mathbf{p}_i$. In this example, $\nabla^2 \varphi(\mathbf{p}) = \mathbf{diag}(\frac{1}{\mathbf{p}})$, so as $\mathbf{p}_i$ approaches any boundary of the orthant for any coordinate $i \in [d]$, space in $\mathcal{M}$ is curved in a way such that the weight assigned to that coordinate blows up.

For instance, geodesics in $\mathbb{R}^d$ are just straight-line paths (which clearly have a constant-speed parameterization, i.e., move along the line at a fixed speed), and geodesics along the unit sphere example mentioned earlier follow arcs in the way that one would expect.

We can now finally define our notion of differentiability. Given a function $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, the gradient of $\mathcal{F}$ at $p \in \mathcal{M}$ is denoted $\nabla_{\mathcal{M}}\mathcal{F}(p) \in T_p\mathcal{M}$, and is the element of $T_p\mathcal{M}$ such that for all curves $\{\gamma(t)\}_{t\in\mathbb{R}}$ passing through p at time $t = 0$, i.e. with $\gamma(0) = p$, it holds that

$$\frac{\partial}{\partial t}\mathcal{F}(\gamma(t)) \mid_{t=0} = \langle \nabla_{\mathcal{M}}\mathcal{F}(p), \dot{\gamma}(0)\rangle_p . \tag{26}$$

Namely, this is just the generalization of the chain rule in $\mathbb{R}^d$: if $x(t)$ is a smooth curve parameterized by time $t \in \mathbb{R}$, then we have $\frac{\mathrm{d}}{\mathrm{d}t}f(x(t)) = \left\langle \nabla f(x(t)), \frac{\mathrm{d}}{\mathrm{d}t}x(t)\right\rangle$. The formula (26) extends this definition to hold for all curves passing through p, which may not be fully comparable (due to their different trajectories) except locally at p. In the remainder of the section, we focus on describing how to do these calculations for an appropriate Riemannian manifold associated with $\mathcal{P}_2(\mathbb{R}^d)$.

4.2 Wasserstein space

The punchline of the above digression, and the developments of Section 3, is that $\mathcal{P}_2(\mathbb{R}^d)$ can naturally be viewed as a manifold $\mathcal{M}$, equipped with the distance function (see (24))

$$\mathrm{d}_{\mathcal{M}}(\mu, \pi) := W_2(\mu, \pi), \text{ for all } \mu, \pi \in \mathcal{M} := \mathcal{P}_2(\mathbb{R}^d). \tag{27}$$

We call the manifold $\mathcal{P}_2(\mathbb{R}^d)$ (referred to as $\mathcal{M}$ for the remainder of this section for brevity), equipped with the distance W_2, *Wasserstein space* for short. Importantly, there are succinct characterizations of tangent spaces, local metrics, geodesics, and gradients of functionals on $\mathcal{M}$. The goal of this section is to state these characterizations, as well as some rough sketches for their proofs. For a more extended proof overview, we defer to Section 1.3 of [Che24], and we defer formal derivations of these results to the excellent resource [AGS08].

To introduce the Riemannian structure of $\mathcal{M}$, it is helpful to first adopt a dual view on curves along Wasserstein space, just as we did in Section 2. In particular, in Section 2 we showed that every *particle flow* governed by an SDE on particles in $\mathbb{R}^d$ enduces a deterministic *density flow* along $\mathcal{M}$. In the case of Section 2, we considered stochastic particle flows, but this connection works fine for deterministic particle flows as well,[12] summarized in the following fundamental result.

Lemma 6 (Continuity equation)**.** *Let $\{\mathbf{v}_t\}_{t\in\mathbb{R}}$ be a family of vector fields on $\mathbb{R}^d$, i.e., for all $t \in \mathbb{R}$, $\mathbf{v}_t : \mathbb{R}^d \to \mathbb{R}^d$ is a corresponding vector field. Let $\mathbf{x}_0 \sim \pi_0$ and suppose that $\mathbf{x}_t$ follows the ODE $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t = \mathbf{v}_t(\mathbf{x}_t)$ for all $t \geq 0$. Then π_t, the law of $\mathbf{x}_t$, follows the following PDE pointwise on $\mathbb{R}^d$:*

$$\frac{\partial}{\partial t}\pi_t + \nabla \cdot (\pi_t \mathbf{v}_t) = 0. \tag{28}$$

Proof. For all test functions $f : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}\int f(\mathbf{x})\frac{\partial}{\partial t}\pi_t(\mathbf{x})\mathrm{d}\mathbf{x} &= \frac{\partial}{\partial t}\left(\int f(\mathbf{x})\pi_t(\mathbf{x})\mathrm{d}\mathbf{x}\right) = \frac{\partial}{\partial t}\mathbb{E}_{\mathbf{x}\sim\pi_t}[f(\mathbf{x})] \\ &= \mathbb{E}\left\langle \nabla f(\mathbf{x}_t), \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_t\right\rangle = \mathbb{E}\langle \nabla f(\mathbf{x}_t), \mathbf{v}_t(\mathbf{x}_t)\rangle \\ &= \int \langle \nabla f(\mathbf{x}), \mathbf{v}_t(\mathbf{x})\rangle \pi_t(\mathbf{x})\mathrm{d}\mathbf{x} = -\int f(\mathbf{x})\nabla \cdot (\mathbf{v}_t(\mathbf{x})\pi_t(\mathbf{x}))\mathrm{d}\mathbf{x},\end{aligned} \tag{29}$$

where we used integration by parts (13) in the last line. As this holds for all f, the claim follows. □

Tangent spaces. Lemma 6 suggests that if the connection between vector fields in $\mathbb{R}^d$ (corresponding to particle flows) and curves on $\mathcal{M}$ (corresponding to density flows) goes both ways, then every geodesic on $\mathcal{M}$ is interpretable by studying a corresponding family of vector fields which

[12] It is unsurprising in light of Proposition 4, which shows that Wasserstein distances are induced by deterministic maps in particle space, that the most interesting particle flows from a geodesic perspective are deterministic.

induces it. This viewpoint turns out to be formalizable, and in fact the relevant vector fields turn out to be those induced by gradients of functions. More precisely, for $\pi \in \mathcal{M}$, the tangent space is

$$T_\pi \mathcal{M} = \left\{ \nabla \cdot (\pi \nabla \psi) \mid \nabla \psi \in \overline{\left\{ \nabla \psi \mid \psi : \mathbb{R}^d \to \mathbb{R} \text{ is smooth}, \int \|\nabla \psi(\mathbf{x})\|_2^2 \, \pi(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty \right\}} \right\}, \quad (30)$$

where $\overline{S}$ denotes the closure of a set S.[13] In other words, when moving along the manifold $\mathcal{M}$ (following a trajectory in density space), the resulting movement in particle space is induced by a vector field given by the gradient of a smooth function, or a limit of such gradients. Moreover, each curve $\{\pi_t\}_{t \in \mathbb{R}} \in \mathcal{M}$ has a time derivative $\frac{\partial}{\partial t} \pi_t$ given by a continuity equation (28), parameterized by a family of vector fields $\{\mathbf{v}_t\}_{t \geq 0}$, such that $\mathbf{v}_t = \nabla \psi_t$ for a function $\psi_t : \mathbb{R}^d \to \mathbb{R}$ almost surely.

We have already seen two pieces of evidence suggesting this should be the case. First, vector fields in Wasserstein space yield W_2 distances between probability distributions (27), and we know that "shortest paths" (i.e., optimal transport plans) are given by gradient vector fields from Brenier's theorem, Proposition 4. Second, when performing the computation in (29) to derive the continuity equation, the vector field $\mathbf{v}_t$ only interacts with gradients of functions ∇f. So, projecting $\mathbf{v}_t$ into the subspace of function gradients[14] decreases its norm measured in the local metric, without affecting the continuity equation (28). To give an example, we showed that the Langevin dynamics (4) induce a curve in Wasserstein space given by the Fokker-Planck equation (16), i.e., $\frac{\partial}{\partial t} \pi_t = \nabla \cdot (\pi_t \nabla V) + \Delta \pi_t$ pointwise. This is consistent with (30), as it gives the continuity equation

$$\begin{aligned} \frac{\partial}{\partial t} \pi_t &= \nabla \cdot (\pi_t \nabla V + \nabla \pi_t) = \nabla \cdot (\pi_t (\nabla V + \nabla \log \pi_t)) \\ &= \nabla \cdot \left(\pi_t \nabla \log \frac{\pi_t}{\pi^\star} \right) \in T_{\pi_t} \mathcal{M}. \end{aligned} \quad (31)$$

We now mention some additional calculations relevant to Wasserstein space.

Local metrics. First, the way to define the local metric (following notation (23)) between $\nabla \cdot (\pi \nabla \psi), \nabla \cdot (\pi \nabla \varphi) \in T_\pi \mathcal{M}$ for smooth functions $\psi, \varphi : \mathbb{R}^d \to \mathbb{R}$ is the formula

$$\langle \nabla \cdot (\pi \nabla \psi), \nabla \cdot (\pi \nabla \varphi) \rangle_\pi := \int \langle \nabla \psi(\mathbf{x}), \nabla \varphi(\mathbf{x}) \rangle \, \pi(\mathbf{x}) \mathrm{d}\mathbf{x}. \quad (32)$$

For elements of $T_\pi \mathcal{M}$ which are obtained by taking the limit of smooth gradients, we similarly define their inner product in the local metric by taking limits of the above formula.

Geodesics. For two probability densities in Wasserstein space, $\pi, \mu \in \mathcal{M}$, we can characterize the geodesic $\gamma : [0, 1] \to \mathcal{M}$ joining $\pi = \gamma(0)$ and $\mu = \gamma(1)$ as follows. Let $(\mathbf{x}_0, \mathbf{x}_1)$ be drawn from the optimal coupling c inducing $W_2^2(\pi, \mu)$, which exists and is unique by Proposition 4. Then, the density $\gamma(t) \in \mathcal{M}$ is given by the law of $(1 - t)\mathbf{x}_0 + t\mathbf{x}_1$. This result states that the geodesic γ is (in particle space) traced out by the straight-line interpolation within the optimal coupling, and a simple calculation using the formulas (25), (32) shows that indeed,

$$d_{\mathcal{M}}(\pi, \mu) = \sqrt{\int \|\mathbf{x}_0 - \mathbf{x}_1\|_2^2 \, c(\mathbf{x}_0, \mathbf{x}_1) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_1} = W_2(\pi, \mu),$$

as claimed. This geodesic is also called McCann's displacement interpolation.

Gradients. Finally, for a functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$,[15] we compute its Wasserstein gradient $\nabla_{\mathcal{M}} \mathcal{F}(\pi)$ (i.e., its gradient over the differentiable manifold $\mathcal{M}$ in the sense of (26)), at a point $\pi \in \mathcal{M}$, as follows. Let $\{\pi_t\}_{t \in \mathbb{R}} \subset \mathcal{M}$ be an arbitrary smooth curve satisfying $\pi_0 = \pi$. By the discussion following Lemma 6, there is a family of functions $\{\psi_t : \mathbb{R}^d \to \mathbb{R}\}_{t \in \mathbb{R}}$, such that

$$\frac{\partial}{\partial t} \pi_t = \nabla \cdot (\pi_t \nabla \psi_t), \text{ for all } t \in \mathbb{R}. \quad (33)$$

[13] Formally, the closure here is with respect to the vector topology induced by the Hilbert space $L^2(\pi)$.

[14] Square-integrable gradients of smooth functions form a subspace within the Hilbert space $L^2(\pi)$, because this space is closed under linear combinations (i.e., $a \nabla f + b \nabla g$ is the gradient of $af + bg$).

[15] Following the literature, we call functions defined over $\mathcal{M}$ "functionals," as they are functions of functions. Indeed, points in $\mathcal{M}$ are probability densities, which are functions from $\mathbb{R}^d \to \mathbb{R}$.

Moreover, the element $\nabla_{\mathcal{M}}\mathcal{F}(\pi)$ is an element of $T_\pi\mathcal{M}$, so it is identifiable with a gradient $\nabla\psi$ for $\psi : \mathbb{R}^d \to \mathbb{R}$ following (30). Let us further suppose that the functional $\mathcal{F}$ satisfies the following assumption: for any smooth curve $\{\pi_t\}_{t\in\mathbb{R}} \in \mathcal{M}$, we have

$$\frac{\partial}{\partial t}\mathcal{F}(\pi_t) \mid_{t=0} = \int (\delta\mathcal{F}(\pi)(\mathbf{x})) \left(\frac{\partial}{\partial t}\pi_t(\mathbf{x}) \mid_{t=0}\right) \mathrm{d}\mathbf{x}, \tag{34}$$

for a scalar-valued function $\delta\mathcal{F}(\pi) : \mathbb{R}^d \to \mathbb{R}$. We will see an example of a functional $\mathcal{F}$ satisfying (34) shortly; the function $\delta\mathcal{F}(\pi)$ is called the *first variation* of $\mathcal{F}$ at π. Continuing, we see that

$$\begin{aligned}\frac{\partial}{\partial t}\mathcal{F}(\pi_t) \mid_{t=0} &= \int (\delta\mathcal{F}(\pi)(\mathbf{x})) \left(\nabla \cdot (\pi\nabla\psi_0)(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\ &= -\int \langle \nabla\delta\mathcal{F}(\pi)(\mathbf{x}), \nabla\psi_0(\mathbf{x})\rangle \, \pi(\mathbf{x})\mathrm{d}\mathbf{x} = \langle \nabla\cdot(-\pi\nabla\delta\mathcal{F}(\pi)), \nabla\cdot(\pi\nabla\psi_0)\rangle_\pi .\end{aligned}$$

The first equation used (33), the second was integration by parts (13), and the last used our local metric formula (32). Comparing the above display with the definition of the manifold gradient in (26), we see that whenever (34) holds for a functional $\mathcal{F}$, the manifold gradient is simply given by

$$\nabla_{\mathcal{M}}\mathcal{F}(\pi) = \nabla\cdot(-\pi\nabla\delta\mathcal{F}(\pi)) \equiv -\nabla\delta\mathcal{F}(\pi). \tag{35}$$

We use the above notation to mean $-\nabla\delta\mathcal{F}(\pi)$ is the vector field followed in particle space pointwise, inducing a density flow of the form $\frac{\partial}{\partial t}\pi_t \mid_{t=0} = \nabla\cdot(-\pi\nabla\delta\mathcal{F}(\pi))$ for curves $\{\pi_t\}_{t\in\mathbb{R}}$ passing through $\pi_0 = \pi$ (see Lemma 6).

To give an example, suppose that $V : \mathbb{R}^d \to \mathbb{R}$ satisfies $\int \exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} < \infty$, and let $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ be the density with $\pi^\star \propto \exp(-V)$. Finally, consider the functional

$$\mathcal{F}(\pi) := D_{\mathrm{KL}}\left(\pi \| \pi^\star\right), \text{ for all } \pi \in \mathcal{M}. \tag{36}$$

We begin by computing the first variation of $\mathcal{F}$, in the sense of (34). Observe that

$$\begin{aligned}\frac{\partial}{\partial t}\mathcal{F}(\pi_t) &= \frac{\partial}{\partial t}\left(\int \pi_t(\mathbf{x}) \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\mathrm{d}\mathbf{x}\right) = \int \left(\log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) + 1\right)\left(\frac{\partial}{\partial t}\pi_t(\mathbf{x})\right)\mathrm{d}\mathbf{x} \\ &= \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\partial}{\partial t}\pi_t(\mathbf{x})\right)\mathrm{d}\mathbf{x} + \frac{\partial}{\partial t}\int \pi_t(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\partial}{\partial t}\pi_t(\mathbf{x})\right)\mathrm{d}\mathbf{x},\end{aligned} \tag{37}$$

since $\int \pi_t(\mathbf{x})\mathrm{d}\mathbf{x} = 1$ for all $\pi_t \in \mathcal{M}$. Therefore, comparing (37) to (34) shows that

$$\nabla\delta\mathcal{F}(\pi)(\mathbf{x}) = \nabla\log\left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) = \nabla V(\mathbf{x}) + \nabla\log\pi(\mathbf{x}), \tag{38}$$

up to a universal additive constant. In other words, following (35), the negated manifold gradient of the functional $\mathcal{F} = D_{\mathrm{KL}}(\cdot\|\pi^\star)$ at $\pi \in \mathcal{M}$ is the vector field given pointwise by $\nabla V + \nabla\log\pi$.

Perhaps surprisingly (as first observed by the landmark result of [JKO98]), this is the same gradient field induced by the continuity equation of the Langevin dynamics, as computed in (31)! Therefore, following the gradient flow along $\mathcal{M}$, i.e., defining $\frac{\partial}{\partial t}\pi_t = -\nabla_{\mathcal{M}}\mathcal{F}(\pi_t)$, is equivalent to a particle flow governed by the Langevin dynamics. We summarize this observation as follows.

Theorem 3 ([JKO98]). *Let $\pi^\star \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\pi^\star \propto \exp(-V)$ for $V : \mathbb{R}^d \to \mathbb{R}$, and let $\mathcal{F}(\pi) := D_{\mathrm{KL}}(\pi\|\pi^\star)$ for $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. Drawing $\mathbf{x}_0 \sim \pi_0$ for a density $\pi_0 \in \mathcal{P}(\mathbb{R}^d)$, and letting π_t be the law of $\mathbf{x}_t$ for $\{\mathbf{x}_t\}_{t\geq 0}$ following the Langevin dynamics (4), the curve $\{\pi_t\}_{t\geq 0}$ follows the curve*

$$\frac{\partial}{\partial t}\pi_t = -\nabla_{\mathcal{M}}\mathcal{F}(\pi_t),$$

where $\nabla_{\mathcal{M}}$ is the manifold gradient defined in (26), and $\mathcal{M}$ is the Riemannian manifold given by $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the distance function W_2. In other words, the Langevin dynamics in particle space induce the gradient flow of the KL divergence to $\pi^\star$ in the space of measures.

We conclude this section with one additional manifold gradient formula, applied to the functional

$$\mathcal{F}(\pi) := \frac{1}{2} W_2^2(\pi, \pi^\star),$$

for a target density $\pi^\star \in \mathcal{M}$. The manifold gradient in this case is induced by the vector field

$$\begin{gathered} \nabla_{\mathcal{M}} \mathcal{F}(\pi)(\mathbf{x}) \equiv \mathbf{x} - \nabla\varphi(\mathbf{x}), \\ \text{where } (\mathbf{x}, \nabla\varphi(\mathbf{x})) \in \mathcal{C}(\pi, \pi^\star) \text{ for } \mathbf{x} \sim \pi \text{ is the coupling inducing } W_2^2(\pi, \pi^\star). \end{gathered} \tag{39}$$

This is intuitive along geodesics, given our earlier discussion of McCann's displacement interpolation and following (34); more generally we defer proving (39) to [Vil08], Theorem 23.9.

5 Functional inequalities

In this section, we introduce *functional inequalities*, which characterize certain spectral properties of a Markov semigroup. As a first example, we define and interpret the Poincaré inequality.

Definition 5 (Poincaré inequality)**.** *We say that $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a* Poincaré inequality with constant C_{PI} *if for all differentiable $f \in L^2(\pi^\star)$,*

$$\mathrm{Var}_{\pi^\star}[f] \le C_{\mathrm{PI}} \int \|\nabla f(\mathbf{x})\|_2^2 \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{40}$$

where

$$\mathrm{Var}_{\pi^\star}[f] := \int \left(f(\mathbf{x}) - \int f(\mathbf{y}) \pi^\star(\mathbf{y}) \mathrm{d}\mathbf{y} \right)^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

As a first example, suppose that $\pi^\star$ satisfies a Poincaré constant with constant C_{PI}. Then plugging in each $f(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle$, and defining $\mathrm{Cov}_{\pi^\star} := \mathbb{E}_{\mathbf{x} \sim \pi^\star}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_{\mathbf{x} \sim \pi^\star}[\mathbf{x}] \mathbb{E}_{\mathbf{x} \sim \pi^\star}[\mathbf{x}]^\top$, (40) reads

$$\mathbf{v}^\top \mathrm{Cov}_{\pi^\star} \mathbf{v} \le C_{\mathrm{PI}} \|\mathbf{v}\|_2^2 \text{ for all } \mathbf{v} \in \mathbb{R}^d \implies \mathrm{Cov}_{\pi^\star} \preceq C_{\mathrm{PI}} \mathbf{I}_d.$$

Thus, Poincaré inequalities imply covariance bounds, suggesting that they are related to concentration of $\pi^\star$. In fact, as we will see, much stronger concentration follows from Poincaré inequalities.

We now explain how Definition 5 has a natural interpretation as providing a spectral gap on the Dirichlet energy operator $\mathcal{E}$ defined in (18). Recall from Lemma 3 that when $\mathcal{E}$ corresponds to the Langevin dynamics (4) with stationary density $\pi^\star$, we have $\mathcal{E}(f, g) = \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x}) \rangle \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$ for $f, g \in L^2(\pi^\star)$, so that the right-hand side of (40) is simply $C_{\mathrm{PI}} \cdot \mathcal{E}(f, f)$. Moreover, we know that constant functions are vanishing eigenfunctions of $\mathcal{E}$, since they have pointwise zero gradients. Hence, viewing $L^2(\pi^\star)$ as a Hilbert space with the inner product $\langle f, g \rangle = \int f(\mathbf{x}) g(\mathbf{x}) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$ (and defining $\|f\|^2 := \langle f, f \rangle$), the condition (40) is equivalent to the statement that

$$\frac{\mathcal{E}(f, f)}{\|f\|^2} \ge \frac{1}{C_{\mathrm{PI}}} \text{ for all } f \in L^2(\pi^\star) \text{ with } \langle f, c \rangle = 0,$$

where c is any constant-valued function. This is precisely imposing a spectral gap on the operator $\mathcal{E}$, since it bounds its second-smallest eigenvalue. As we will see, we can formally use Poincaré inequalities to bound the decay of any eigenfunctions orthogonal to constant functions to derive convergence to $\pi^\star$ in the χ^2 distance. We also mention that our definition (40) is specialized to the Dirichlet energy operator $\mathcal{E}$ arising from the Langevin dynamics. Indeed, other reversible Markov semigroups induce alternative compatible formulations of the Poincaré inequality, by setting the right-hand side of (40) to the generalization of $\mathcal{E}$ derived in Lemma 3.

In general, the χ^2 divergence between probability distributions (whose convergence proofs are facilitated by Poincaré inequalities, as statements about "variance decay") dominates the KL divergence D_{KL}. However, in important cases of interest the initial χ^2 divergence can be exponentially larger than D_{KL} (see discussion at the end of Section 2.2, Part XII), so we would like a means to directly argue about the convergence of a semigroup in D_{KL}. This "entropy decay" strengthening of the Poincaré inequality is the log-Sobolev inequality, which we next define.

Definition 6 (Log-Sobolev inequality)**.** *We say that $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a* log-Sobolev inequality with constant C_{LSI} *if for all differentiable $f \in L^2(\pi^\star)$,*

$$\mathrm{Ent}_{\pi^\star}[f^2] \le 2C_{\mathrm{LSI}} \int \|\nabla f(\mathbf{x})\|_2^2 \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{41}$$

where

$$\mathrm{Ent}_{\pi^\star}[f] := \int f(\mathbf{x}) \log(f(\mathbf{x})) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(\int f(\mathbf{x}) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \log \left(\int f(\mathbf{x}) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} \right).$$

We show that Definition 6 is indeed a strengthening of Definition 5.

Lemma 7. *If $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a log-Sobolev inequality with constant C, it also satisfies a Poincaré inequality with constant C.*

Proof. Let $f \in L^2(\pi^\star)$, and without loss of generality suppose that $\mathbb{E}_{\pi^\star}[f] = \int f(\mathbf{x}) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} = 0$, else we may subtract an appropriate constant from f pointwise which does not affect the calculation (40). Next, consider taking $f \leftarrow 1 + \epsilon f$ in (41) for a small constant ϵ. We compute that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \mathrm{Ent}_{\pi^\star}[(1+\epsilon f)^2] = 2 \int f(\mathbf{x})^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x},$$

which follows by Lebesgue's dominated convergence theorem, the assumption $\mathbb{E}_{\pi^\star}[f] = 0$, and

$$(1+\epsilon c)^2 \log((1+\epsilon c)^2 = 2(1+2\epsilon c) \left(\epsilon c - \frac{\epsilon^2 c^2}{2} \right) + O(\epsilon^3) = 2\epsilon c + 3\epsilon^2 c^2,$$
$$(1+\epsilon^2 c) \log\left(1 + \epsilon^2 c\right) = \epsilon^2 c + O(\epsilon^3).$$

Meanwhile, it is clear that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \int \|\epsilon \nabla f(\mathbf{x})\|_2^2 \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \|\nabla f(\mathbf{x})\|_2^2 \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Combining in (41) and taking limits yields the conclusion. □

We also mention one alternative formulation of (41), which is easier to apply in certain settings.

Lemma 8. *If $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a log-Sobolev inequality with constant C_{LSI}, then for any $\pi \in \mathcal{P}(\mathbb{R}^d)$ such that $\frac{\pi}{\pi^\star} < \infty$ almost surely,*

$$D_{\mathrm{KL}}\left(\pi \| \pi^\star\right) \le \frac{C_{\mathrm{LSI}}}{2} \int \left\| \nabla \log \left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right) \right\|_2^2 \pi(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{42}$$

where

$$D_{\mathrm{KL}}\left(\pi \| \pi^\star\right) := \int \pi(\mathbf{x}) \log \left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right) \mathrm{d}\mathbf{x}.$$

Proof. By plugging in $f = \sqrt{\frac{\pi}{\pi^\star}}$ into (41), the left-hand side reads

$$\mathrm{Ent}_{\pi^\star}[f^2] = \int \frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \log \left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} = D_{\mathrm{KL}}\left(\pi \| \pi^\star\right),$$

since $\log(\int \pi(\mathbf{x}) \mathrm{d}\mathbf{x}) = \log 1 = 0$. Moreover, the right-hand side of (41) is

$$\begin{aligned} 2C_{\mathrm{LSI}} \int \|\nabla f(\mathbf{x})\|_2^2 \, \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} &= \frac{C_{\mathrm{LSI}}}{2} \int \left\| \nabla \frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right\|_2^2 \frac{\pi^\star(\mathbf{x})^2}{\pi(\mathbf{x})} \mathrm{d}\mathbf{x} \\ &= \frac{C_{\mathrm{LSI}}}{2} \int \left\| \nabla \log \left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right) \right\|_2^2 \pi(\mathbf{x}) \mathrm{d}\mathbf{x}, \end{aligned}$$

and combining yields the desired (42). □

The term $\int \|\nabla \log(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})})\|_2^2 \pi(\mathbf{x})\mathrm{d}\mathbf{x}$ in the right-hand side of (42) is sometimes called the *Fisher information*, an important quantity in statistics. For now, we mention that under our earlier calculations (35) and (38), letting $\mathcal{F}(\pi) := D_{\mathrm{KL}}(\pi\|\pi^\star)$ so that $\mathcal{F}(\pi^\star) = 0$, (42) alternatively reads

$$\mathcal{F}(\pi) - \mathcal{F}(\pi^\star) \leq \frac{C_{\mathrm{LSI}}}{2} \left\langle \nabla_{\mathcal{M}} \mathcal{F}(\pi), \nabla_{\mathcal{M}} \mathcal{F}(\pi) \right\rangle_\pi . \tag{43}$$

This is the functional analog of the gradient domination condition in Corollary 3, Part II which allowed us to prove linear convergence rates for gradient flow. We will show shortly that log-Sobolev inequalities establish a similar linear rate of decay for the KL divergence.

5.1 Convergence from functional inequalities

We now formalize our earlier statements that the Poincaré inequality implies a linear rate of decay on an appropriate notion of variance under the Langevin dynamics, and that the log-Sobolev inequality implies a similar decay on the KL divergence. We begin by proving variance decay, reminding the reader of the definition of the χ^2 divergence between distributions:

$$\chi^2\left(\pi\|\pi^\star\right) := \mathrm{Var}_{\pi^\star}\left[\frac{\pi}{\pi^\star}\right] = \int \left(\frac{\pi(x)}{\pi^\star(x)}\right)^2 \pi^\star(x)\mathrm{d}x - 1.$$

Lemma 9. *Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin dynamics (4) with stationary distribution $\pi^\star$, and assume that $\pi^\star$ satisfies a Poincaré inequality with constant C_{PI}. For all $t \geq 0$, the law of $\mathbf{x}_t$, denoted $\pi_t \in \mathcal{P}(\mathbb{R}^d)$, satisfies*

$$\chi^2\left(\pi_t\|\pi^\star\right) \leq \exp\left(-\frac{2t}{C_{\mathrm{PI}}}\right) \chi^2\left(\pi_0\|\pi^\star\right).$$

Proof. It suffices to show that $\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\pi_t\|\pi^\star) \leq -\frac{2}{C_{\mathrm{PI}}}\chi^2(\pi_t\|\pi^\star)$, at which point the conclusion follows from Grönwall's inequality (Fact 1, Part II). We compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\pi_t\|\pi^\star) &= \frac{\partial}{\partial t}\int \left(\frac{\pi_t(\mathbf{x})^2}{\pi^\star(\mathbf{x})^2} - 1\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= 2\int \left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\partial}{\partial t}\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= 2\int \left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\mathcal{L}^*\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = -2\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \frac{\pi_t}{\pi^\star}\right),
\end{aligned}$$

where we used Kolmogorov's backward equation (11) in the third equality, which implies $\frac{\partial}{\partial t}\pi_t(x) = \mathcal{L}^*\pi_t(\mathbf{x})$. Finally, using the definition (40), we have the desired

$$-2\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \frac{\pi_t}{\pi^\star}\right) \leq -\frac{2}{C_{\mathrm{PI}}}\mathrm{Var}_{\pi^\star}\left[\frac{\pi_t}{\pi^\star}\right] = -\frac{2}{C_{\mathrm{PI}}}\chi^2(\pi_t\|\pi^\star),$$

where we recalled the formula for $\mathcal{E}$ specialized to the Langevin diffusion, stated in Lemma 3. □

We prove a similar convergence for the KL divergence under a log-Sobolev inequality.

Lemma 10. *Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin dynamics (4) with stationary distribution $\pi^\star$, and assume that $\pi^\star$ satisfies a log-Sobolev inequality with constant C_{LSI}. For all $t \geq 0$, the law of $\mathbf{x}_t$, denoted $\pi_t \in \mathcal{P}(\mathbb{R}^d)$, satisfies*

$$D_{\mathrm{KL}}\left(\pi_t\|\pi^\star\right) \leq \exp\left(-\frac{2t}{C_{\mathrm{LSI}}}\right) D_{\mathrm{KL}}\left(\pi_0\|\pi^\star\right).$$

Proof. As in the proof of Lemma 9, it suffices to show that $\frac{\mathrm{d}}{\mathrm{d}t}D_{\mathrm{KL}}(\pi_t\|\pi^\star) \leq -\frac{2}{C_{\mathrm{LSI}}}D_{\mathrm{KL}}(\pi_t\|\pi^\star)$. We previously computed in (38) that the KL divergence satisfies

$$\frac{\partial}{\partial t}D_{\mathrm{KL}}\left(\pi_t\|\pi^\star\right) = \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\mathcal{L}^*\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = -\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \log\left(\frac{\pi_t}{\pi^\star}\right)\right).$$

The conclusion follows analogously to Lemma 9, using (41), upon realizing

$$\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \log\left(\frac{\pi_t}{\pi^\star}\right)\right) = \int \left\langle \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right), \nabla\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\rangle \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}$$
$$= \int \left\|\nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\|_2^2 \pi(\mathbf{x})\mathrm{d}\mathbf{x} \geq \frac{2}{C_{\mathrm{LSI}}} D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right).$$

□

5.2 Functional inequalities from strong logconcavity

We have established in Lemmas 9 and 10 that functional inequalities give us a way of proving convergence of Markov semigroups (in relative variance, i.e., χ^2, or relative entropy, i.e., D_{KL}). These results can be thought of as continuous-time generalizations of the spectral gap arguments we used in the discrete-time setting, e.g., Parts XIV and XV. However, we have not yet given examples of densities which actually satisfy Definitions 5 or 6, which is the purpose of this section.

We first establish a "Hessian-reweighted" variant of (40) for logconcave densities, which streamlines our presentation. Specifically, recall the Prékopa-Leindler inequality (Theorem 4, Part I), which says (changing function names from $(f, g, h) \to (u, v, w)$ for convenience) that for $\lambda \in [0, 1]$ and $u, v, w : \mathbb{R}^d \to \mathbb{R}$ with $w((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \geq u(\mathbf{x})^{1-\lambda} v(\mathbf{x}')^\lambda$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$\int w(\mathbf{x})\mathrm{d}\mathbf{x} \geq \left(\int u(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{1-\lambda} \left(\int v(\mathbf{x})\mathrm{d}\mathbf{x}\right)^{\lambda}. \tag{44}$$

The following result is known as the *Brascamp-Lieb inequality* [BL76], and our proof is based on an elegant strategy from [BL00] which applies the Prékopa-Leindler inequality (44).

Proposition 5 (Brascamp-Lieb inequality)**.** *Let $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ be logconcave, and let $\pi^\star \propto \exp(-V)$ where $V : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and twice-continuously differentiable. Then for all twice-differentiable $f \in L^2(\pi^\star)$,*

$$\mathrm{Var}_{\pi^\star}[f] \leq \int \|\nabla f(\mathbf{x})\|^2_{(\nabla^2 V(\mathbf{x}))^{-1}} \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Proof. Throughout the proof, assume without loss that $\int \exp(-V(\mathbf{x}))\mathrm{d}x = 1$, and that there exists $\alpha > 0$ such that $\nabla^2 V \succeq \alpha \mathbf{I}_d$ pointwise; this latter assumption can be removed using a limiting argument. Our strategy is to apply the Prékopa-Leindler inequality (44) with $\lambda = \frac{1}{2}$, and

$$u(\mathbf{x}) := \exp\left(2\delta f(\mathbf{x}) - V(\mathbf{x})\right), \; v(\mathbf{x}) := \exp\left(-V(\mathbf{x})\right), \; w(\mathbf{x}) := \exp\left(g^{\delta f}(\mathbf{x}) - V(\mathbf{x})\right),$$
$$\text{where } \delta \to 0 \text{ and } g^{\delta f}(\mathbf{y}) := \sup_{\mathbf{y} = \frac{1}{2}(\mathbf{x}+\mathbf{x}')} \left\{\delta f(\mathbf{x}) - \left(\frac{1}{2}V(\mathbf{x}) + \frac{1}{2}V(\mathbf{x}') - V(\mathbf{y})\right)\right\}.$$

By the definition of $g^{\delta f}$, $w(\frac{1}{2}(\mathbf{x}+\mathbf{x}')) \geq u(\mathbf{x})^{\frac{1}{2}} v(\mathbf{x}')^{\frac{1}{2}}$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, so (44) gives

$$\int \exp\left(2\delta f(\mathbf{x})\right) \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \leq \left(\int \exp\left(g^{\delta f}(\mathbf{x})\right) \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}\right)^2. \tag{45}$$

Next, we claim that as $\delta \to 0$, we have that for all $\mathbf{y} \in \mathbb{R}^d$,

$$g^{\delta f}(\mathbf{y}) = \delta f(\mathbf{y}) + \frac{\delta^2}{2}\left(\nabla^2 V(\mathbf{y})\right)^{-1}\left[\nabla f(\mathbf{y}), \nabla f(\mathbf{y})\right] + o(\delta^2). \tag{46}$$

To see this, essentially the idea is to treat δf as a small perturbation of the strongly convex function V, and see how the minimizer varies. For fixed $\delta, \mathbf{y}$, let us parameterize the optimization problem governing $g^{\delta f}(\mathbf{y})$ by $\mathbf{h} \in \mathbb{R}^d$ such that $\mathbf{x} = \mathbf{y} + \mathbf{h}$ and $\mathbf{x}' = \mathbf{y} - \mathbf{h}$. Then a necessary condition for the optimal $\mathbf{h}$ is

$$\delta \nabla f(\mathbf{y} + \mathbf{h}) - \frac{1}{2}\nabla V(\mathbf{y} + \mathbf{h}) + \frac{1}{2}V(\mathbf{y} - \mathbf{h}) = 0.$$

Moreover, under our assumption that $\nabla^2 V$ is pointwise lower bounded, this gives

$$\delta \nabla f(\mathbf{y}) + \delta \nabla^2 f(\mathbf{y})\mathbf{h} = \nabla^2 V(\mathbf{y})\mathbf{h} + \mathbf{r} \text{ for } \|\mathbf{r}\|_2 = o(\delta^2) \implies \|\mathbf{h}\|_2 = O(\delta),$$

where the asymptotic notation above hides local smoothness parameters of f, V at $\mathbf{y}$, that vanish as $\delta \to 0$. Plugging this estimate in above, we have shown that at optimality,

$$\mathbf{h} = \delta \nabla^2 V(\mathbf{y})^{-1} \nabla f(\mathbf{y}) + \mathbf{r} \text{ for } \|\mathbf{r}\|_2 = o(\delta).$$

Finally, we have from a second-order Taylor expansion of V that

$$\delta f(\mathbf{y}+\mathbf{h}) - \left(\frac{1}{2}V(\mathbf{y}+\mathbf{h}) + \frac{1}{2}V(\mathbf{y}-\mathbf{h}) - V(\mathbf{y})\right) = \delta f(\mathbf{y}) + \frac{1}{2}\nabla^2 V(\mathbf{y})[\mathbf{h},\mathbf{h}] + o(\delta^2),$$

so plugging in our optimal $\mathbf{h}$, we have established (46). Now, expanding (45),

$$\begin{aligned}
&1 + 2\delta \int f(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} + 2\delta^2 \int f(\mathbf{x})^2 \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} - o(\delta^2) \\
&= \int \exp(2\delta f(\mathbf{x}))\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\
&\le \left(\int \exp\left(\delta f(\mathbf{x}) + \frac{\delta^2}{2}\left(\nabla^2 V(\mathbf{x})\right)^{-1}[\nabla f(\mathbf{x}), \nabla f(\mathbf{x})]\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}\right)^2 + o(\delta^2) \\
&\le 1 + 2\delta \int f(\mathbf{x})\pi^\star(\mathbf{x}) + \delta^2 \int f(\mathbf{x})^2\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} + \delta^2\left(\int f(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}\right)^2 \\
&\quad + \delta^2 \int \left(\nabla^2 V(\mathbf{x})\right)^{-1}[\nabla f(\mathbf{x}), \nabla f(\mathbf{x})]\,\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} + o(\delta^2).
\end{aligned}$$

where the second line used a second-order Taylor expansion of exp, the third line combined (45) and (46), and the last line also used a second-order Taylor expansion of exp and $(1+\cdot)^2$. At this point, cancelling like terms and normalizing by δ^2 for $\delta \to 0$ gives the claimed result:

$$\left(\int f(\mathbf{x})^2\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} - \left(\int f(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}\right)^2\right) \le \int \left(\nabla^2 V(\mathbf{x})\right)^{-1}[\nabla f(\mathbf{x}), \nabla f(\mathbf{x})]\,\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}.$$

□

Proposition 5 implies a Poincaré inequality for *strongly logconcave* densities (Definition 4).

Corollary 1. *Any μ-strongly logconcave density satisfies a Poincaré inequality with constant $\frac{1}{\mu}$.*

Proof. Apply Proposition 5 with $(\nabla^2 V(\cdot))^{-1} \preceq \frac{1}{\mu}\mathbf{I}_d$, where V is the negative log-density. □

Although one may naturally next conjecture that the "Hessian-reweighted" variant of the log-Sobolev inequality (41) holds, i.e.,

$$\mathrm{Ent}_{\pi^\star}[f^2] \le 2\int \|\nabla f(\mathbf{x})\|^2_{(\nabla^2 V(\mathbf{x}))^{-1}}\,\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x},$$

it turns out that there are one-dimensional counterxamples (Section 2, [BL00]). Nonetheless, we have quite a substantial generalization of Corollary 1.

Theorem 4 (Proposition 3.1, [BL00])**.** *Let $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ be μ-strongly logconcave with respect to $\|\cdot\|$, i.e., $\pi^\star \propto \exp(-V)$ where V is μ-strongly convex with respect to $\|\cdot\|$ (Definition 4, Part II). Then for all differentiable $f \in L^2(\pi^\star)$,*

$$\mathrm{Ent}_{\pi^\star}\left[f^2\right] \le \frac{2}{\mu}\int \|\nabla f(\mathbf{x})\|_*^2\,\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}.$$

We do not prove Theorem 4 here, but mention that it follows from a variant of our proof of Proposition 5, instead applied with $\lambda \to 0$. The key fact used is that

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{1+\lambda}\,|_{\lambda=0} = \lim_{\lambda\to 0}\frac{x^{1+\lambda}-x}{\lambda} = x\log(x).$$

and much of the rest of the proof is the same. Thus, Theorem 4 implies a log-Sobolev generalization of Corollary 1 holds, for all norms (not just ℓ_2). This also shows that the Poincaré variant in Corollary 1 holds in all norms as well, using the reduction in Lemma 7.

5.3 Further consequences

In this section, we mention a few further consequences of functional inequalities which are often useful in applications. First, we show that they imply concentration of Lipschitz functions. The following proof strategy is attributed by [Led99] to Ira Herbst in an unpublished letter to Leonard Gross, and hence has come to be known as "Herbst's argument."

Lemma 11 (Proposition 2.3, [Led99]). *If $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a log-Sobolev inequality with constant C_{LSI}, then for any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we have that*

$$\Pr_{\mathbf{x}\sim\pi^\star}\left[f(\mathbf{x}) \geq \mathbb{E}_{\pi^\star}[f] + c\right] \leq \exp\left(-\frac{c^2}{2C_{\mathrm{LSI}}}\right) \text{ for all } c > 0.$$

Proof. The basic idea is to apply the log-Sobolev inequality with $f^2 \leftarrow \exp(\lambda f)$, the *moment-generating function* of f, treated as a random variable. Let

$$H(\lambda) := \mathbb{E}_{\pi^\star}\left[\exp\left(\lambda f\right)\right] \implies \lambda H'(\lambda) = \mathbb{E}_{\pi^\star}\left[\lambda f \exp\left(\lambda f\right)\right].$$

Then, by applying the log-Sobolev inequality,

$$\begin{aligned}\lambda H'(\lambda) - H(\lambda)\log H(\lambda) &\leq 2C_{\mathrm{LSI}}\int\left\|\nabla \exp\left(\frac{\lambda}{2}f(\mathbf{x})\right)\right\|_2^2 \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \frac{\lambda^2 C_{\mathrm{LSI}}}{2}\int \|\nabla f(\mathbf{x})\|_2^2 \exp\left(\lambda f(\mathbf{x})\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \leq \frac{\lambda^2 C_{\mathrm{LSI}}}{2}H(\lambda).\end{aligned} \tag{47}$$

Now consider the function $K(\lambda) := \frac{1}{\lambda}\log H(\lambda)$, where we define

$$K(0) := \lim_{\lambda\to 0}\frac{1}{\lambda}\log\left(\mathbb{E}_{\pi^\star}\left[\exp\left(\lambda f\right)\right]\right) = \lim_{\lambda\to 0}\frac{H(\lambda)-H(0)}{\lambda} = H'(0) = \mathbb{E}_{\pi^\star}\left[f\right],$$

using $\log(1+x) = x + o(x^2)$ for $x \to 0$. Moreover,

$$K'(\lambda) = -\frac{1}{\lambda^2}\log H(\lambda) + \frac{H'(\lambda)}{\lambda H(\lambda)} = \frac{\lambda H'(\lambda) - H(\lambda)\log H(\lambda)}{\lambda^2 H(\lambda)} \leq \frac{C_{\mathrm{LSI}}}{2},$$

where the last inequality used (47). Therefore, for all $\lambda \geq 0$, by integrating K we have

$$K(\lambda) \leq K(0) + \frac{\lambda C_{\mathrm{LSI}}}{2},$$

which yields the moment-generating function bound

$$\mathbb{E}_{\pi^\star}\left[\exp\left(\lambda f\right)\right] = H(\lambda) = \exp\left(\lambda K(\lambda)\right) \leq \exp\left(\lambda\mathbb{E}_{\pi^\star}\left[f\right] + \frac{\lambda^2 C_{\mathrm{LSI}}}{2}\right).$$

The remainder of the proof follows analogously to Theorem 1, Part VI:

$$\begin{aligned}\Pr_{\mathbf{x}\sim\pi^\star}\left[f(\mathbf{x}) \geq \mathbb{E}_{\pi^\star}[f] + c\right] &= \Pr_{\mathbf{x}\sim\pi^\star}\left[\exp(\lambda f(\mathbf{x})) \geq \exp(\lambda\mathbb{E}_{\pi^\star}[f] + \lambda c)\right] \\ &\leq \exp\left(-\lambda c + \frac{\lambda^2 C_{\mathrm{LSI}}}{2}\right) = \exp\left(-\frac{c^2}{2C_{\mathrm{LSI}}}\right), \text{ for } \lambda = \frac{c}{C_{\mathrm{LSI}}}.\end{aligned}$$

We remark that the same arguments extend straightforwardly to any norm $\|\cdot\| \neq \|\cdot\|_2$. □

It turns out that with some additional effort, one can also derive tail bounds from a Poincaré inequality. The proof is similar to Lemma 11 and we omit it.

Lemma 12 (Proposition 2.13, [Led99]). *If $\pi^\star \in \mathcal{P}(\mathbb{R}^d)$ satisfies a Poincaré inequality with constant C_{PI}, then for any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we have that*

$$\Pr_{x\sim\pi^\star}\left[f(\mathbf{x}) \geq \mathbb{E}_{\pi^\star}[f] + c\right] \leq 3\exp\left(-\frac{c}{\sqrt{C_{\mathrm{PI}}}}\right) \text{ for all } c > 0.$$

There is an interesting qualitative difference between Lemmas 11 and 12: random variables satisfying log-Sobolev inequalities exhibit *sub-Gaussian* concentration of Lipschitz functions, whereas if they satisfy Poincaré inequalities, then Lipschitz functions are merely *sub-exponential*.

To provide some justification, recall from the localization lemma (Proposition 4, Part XV) that extreme examples of logconcave densities are exponential, which induce piecewise-linear log-densities (whose Hessians do not have curvature). Intuitively, this is consistent with our understanding of Poincaré and log-Sobolev inequalities: all logconcave densities exhibit weak forms of Poincaré inequalities, via Proposition 5, but we have only claimed that log-Sobolev holds under some strict amount of positive curvature, e.g., strong logconcavity (Theorem 4).

Our final consequence is that a log-Sobolev inequality for $\pi^\star$ implies that $D_{\mathrm{KL}}(\cdot\|\pi^\star)$ grows quadratically with respect to $W_2(\cdot, \pi^\star)$. This is known as the Otto-Villani theorem [FO00], and is essentially a *quadratic growth* bound, a weakening of strong convexity (Remark 3, Part II).

Lemma 13. *If $\pi^\star \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies a log-Sobolev inequality with constant C_{LSI}, then*

$$\frac{1}{2C_{\mathrm{LSI}}} W_2^2(\pi, \pi^\star) \le D_{\mathrm{KL}}\left(\pi\|\pi^\star\right) \text{ for all } \pi \in \mathcal{P}_2(\mathbb{R}^d).$$

Proof. We instead prove that if $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the gradient domination condition

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 \text{ for all } \mathbf{x} \in \mathbb{R}^d, \tag{48}$$

where $\mathbf{x}^\star = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ is unique, then the quadratic growth bound

$$\frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^\star\|_2^2 \le f(\mathbf{x}) - f(\mathbf{x}^\star) \text{ for all } \mathbf{x} \in \mathbb{R}^d \tag{49}$$

also holds. The lemma statement follows by appropriately generalizing this proof strategy to the functional $f \leftarrow D_{\mathrm{KL}}(\cdot\|\pi^\star)$ on Wasserstein space, and applying the observation (43). This is done using the formalism of *Otto calculus* [FO00, Ott01], which extends Section 4; we defer further proof details to Proposition 1 of [FO00], as well as an alternative perspective in [GLRT20].

To show that gradient domination implies quadratic growth on $\mathbb{R}^d$, let $\mathbf{x}_0 \leftarrow \mathbf{x}$, and define a trajectory $\{\mathbf{x}_t\}_{t\ge 0}$ that follows the gradient flow ODE $\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)\mathrm{d}t$. We argued in Section 4, Part II that under the gradient domination condition (48), $\mathbf{x}_t$ converges exponentially fast in function value, so $\mathbf{x}_t \to \mathbf{x}^\star$ as $t \to \infty$. Next, consider the potential

$$\Phi_t := \|\mathbf{x}_t - \mathbf{x}_0\|_2 + \sqrt{\frac{2(f(\mathbf{x}_t) - f(\mathbf{x}^\star))}{\mu}},$$

applied to this trajectory. Note that

$$\begin{aligned}\frac{\mathrm{d}\Phi_t}{\mathrm{d}t} &\le \|\nabla f(\mathbf{x}_t)\|_2 + \sqrt{\frac{2}{\mu}} \cdot \frac{\mathrm{d}}{\mathrm{d}t} \sqrt{f(\mathbf{x}_t) - f(\mathbf{x}^\star)} \\ &\le \|\nabla f(\mathbf{x}_t)\|_2 - \frac{\|\nabla f(\mathbf{x}_t)\|_2^2}{\sqrt{2\mu(f(\mathbf{x}_t) - f(\mathbf{x}^\star))}} \le 0,\end{aligned}$$

where the first line used the triangle inequality, and the second applied the assumption (48). Thus we have that $\Phi_0 \ge \Phi_t$ for all $t \ge 0$, and rearranging after taking $t \to \infty$ gives the claim (49). □

The conclusion of Lemma 13 is sometimes called (Talagrand's) transportation inequality, so log-Sobolev inequalities imply transportation inequalities. Interestingly, transportation inequalities in turn imply Poincaré inequalities via a linearization argument similar to Lemma 7 (see Section 3, [CE17]), showing that these three properties of a density $\pi^\star$ form a natural hierarchy.

Source material

Portions of this lecture are based on reference material in [Øk03, AGS08, Vil08, Dur10, MG10, vH16, Vis18, Che24], as well as the author's own experience working in the field. We would like to mention that much of the exposition is patterned off of the excellent resource [Che24], which we highly recommend as a deeper dive into the topics covered in this lecture.

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. 2008.

[BL76] Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

[BL00] Sergey G Bobkov and Michel Ledoux. From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities. *GAFA, Geometric and Functional Analysis*, 10:1028–1052, 2000.

[Bre87] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C.R. Acad. Sci. Paris Sér I Math.*, 305:805–808, 1987.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44:375–417, 1991.

[Caf00] Luis A. Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

[CE17] Dario Cordero-Erausquin. Transport inequalities for log-concave measures, quantitative forms, and applications. *Canadian Journal of Mathematics*, 69(3):481–501, 2017.

[Che24] Sinho Chewi. *Log-Concave Sampling*. 2024.

[CP23] Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *Reports. Mathematical*, 361:1471–1482, 2023.

[Dam65] K. E. Dambis. On the decomposition of continuous submartingales. *Theor. Probab. Appl.*, 10:401–410, 1965.

[dC92] Manfredo P. do Carmo. Riemannian geometry. 1992.

[DS65] Lester E. Dubins and Gideon Schwarz. On continuous martingales. *Proceedings of the National Academy of Sciences*, 53(5):913–916, 1965.

[Dur10] Rick Durrett. *Probability: Theory and Examples*. 2010.

[FO00] Cédric Villani Felix Otto. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[GLRT20] Ivan Gentil, Christian Léonard, Luigia Ripani, and Luca Tamanini. An entropic interpolation proof of the hwi inequality. *Stochastic Processes and their Applications*, 130(2):907–923, 2020.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[Led99] Michel Ledoux. *Concentration of measure and logarithmic Sobolev inequalities*. Seminaire de probabilities XXXIII, 1999.

[MG10] Robert J. McCann and Nestor Guillen. *Five Lectures on Optimal Transportation: Geometry, Regularity and Applications*. 2010.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. in Partial Differential Equations*, 26(1–2):101–174, 2001.

[Roc70] R. T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33:209–216, 1970.

[vH16] Ramon van Handel. *Probability in High Dimension.* 2016.

[Vil08] Cédric Villani. *Optimal transport, old and new.* 2008.

[Vis18] Nisheeth K. Vishnoi. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *CoRR*, abs/1806.06373, 2018.

[Øk03] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications.* 2003.