Continuity equations, PDEs, probabilities, and gradient flows*

Xiaohui Chen

This version: July 5, 2020

Contents

*Working draft in progress and comments are welcome (email: xhchen@illinois.edu).

1 Continuity equation

Let $\Omega \subset \mathbb{R}^n$ be a spatial domain. Consider the *continuity equation* (CE):

$$\partial_t \mu_t + \nabla \cdot (\mu_t \mathbf{v}_t) = 0, \tag{1}$$

where μ_t is a probability measure (typically absolutely continuous with a density) on Ω, $\mathbf{v}_t : \Omega \to \mathbb{R}^n$ is a velocity vector field on Ω, and $\nabla \cdot \mathbf{v}$ is the divergence of a vector field $\mathbf{v}$.

There are several meanings of solving the continuity equation (1). Given the vector field $\mathbf{v}_t$, we can speak of a *classical (or strong) solution* as a partial differential equation (PDE) by thinking $\mu_t(x)$ as a differentiable function of two variables x and t. We can also think of a *distributional solution* by integrating against some "nice" class of test functions on (x, t). If the continuity equation (1) is satisfied, then, for any finite time point $T > 0$, we can integrate with a C^1 function $\psi : \Omega \times [0, T] \to \mathbb{R}$ with bounded support and apply integration-by-parts:

$$\begin{aligned} 0 &= \int_0^T \int_\Omega \psi \partial_t \mu_t + \int_0^T \int_\Omega \psi \nabla \cdot (\mu_t \mathbf{v}_t) && (2) \\ &= -\int_0^T \int_\Omega (\partial_t \psi) \mu_t - \int_0^T \int_\Omega \langle \nabla \psi, \mathbf{v}_t \rangle \mu_t, && (3) \end{aligned}$$

where there is no contribution from the boundary because ψ has compact support so that the divergence theorem works. Here we implicitly assumed that the first law of thermodynamics holds (i.e., mass conservation of μ_t) so that there is no mass escapes at the boundary (if Ω is bounded) or near the infinity (if Ω is not bounded). Compared with the strong solution, the distribution solution (3) does not require differentiability by moving the derivative from $\mu_t(x)$ to $\psi(x, t)$.

Suppose further we can interchange ∂_t and $\int_\Omega$ in the first integral of (2) and we keep the second integral in (3), then we can take a smaller class of test functions only in the spatial domain Ω such that the solution space is larger. Why shouldn't we interchange ∂_t and $\int_\Omega$ in the first integral of (3)? This is because if we take test functions depending only on Ω, then it does not make sense to take the time derivative as in (3), which always equals to zero.

Let $\phi : \Omega \to \mathbb{R}$ be such C^1 test function with bounded support. To make sense of differentiation outside of integration, we obviously need $t \mapsto \int_\Omega \phi \mu_t$ is absolutely continuous in t. In addition, we need the identity

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \phi \mu_t - \int_\Omega \langle \nabla \phi, \mathbf{v}_t \rangle \mu_t = 0 \tag{4}$$

to hold pointwise (in t) such that (3) equals to zero. This motivates the following definition.

Definition 1.1 (Weak solution of the continuity equation). We say that the density μ_t is a *weak solution* in the distribution sense if for any C^1 test function $\phi : \Omega \to \mathbb{R}$ with bounded support, the function $t \mapsto \int_\Omega \phi\mu_t$ is absolutely continuous in t, and for each a.e. t, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi\mu_t = \int_\Omega \langle \nabla\phi, \mathbf{v}_t\rangle \mu_t. \tag{5}$$

In these notes, we make (more) sense of dynamic behaviors of the continuity equation, and illustration its links to the classical PDEs, probabilities (such as Markov processes and stochastic differential equations), and the trajectory analysis (such as gradient flows). Specifically, we would like to understand the question that how does the weak solution of the continuity equation in an infinite-dimensional metric space (typically Wasserstein) connect with the classical solution of PDEs in a finite-dimensional space (typically Euclidean space and time)? We start from the (linear) heat equation as a concrete example.

2 Heat equation

2.1 Heat equation as a PDE in space and time

Recall that the heat equation (HE) on $\mathbb{R}^n$ is defined as:

$$(\partial_t - \Delta)u = \partial_t u - \nabla\cdot\nabla u = 0, \tag{6}$$

where $u : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ and $u(x,t)$ is a two-variable function of space and time. The fundamental solution of the heat equation (6) is given by

$$u(x,t) = H(x,0,t), \tag{7}$$

where $H(x,y,t)$ is the heat kernel (sometimes also called Green's function):

$$H(x,y,t) = (4\pi t)^{-n/2}\exp\Big(-\frac{|x-y|^2}{4t}\Big), \quad y\in\mathbb{R}^n,\, t>0. \tag{8}$$

The classical solution u is just a (positive) function of two variables x and t. One can think of $H(x,y,t)$ is the transition density from x to y in time $2t$.

Let $\Omega = \mathbb{R}^n$ and $\mu_t(x) = u(x,t)$. Clearly $\mu_t > 0$ is the probability density of a Gaussian distribution $N(y, 2tI_n)$ and the continuity equation (1) reads

$$\partial_t\mu_t - \nabla\cdot(\mu_t\frac{\nabla\mu_t}{\mu_t}) = 0. \tag{9}$$

In this case, the velocity vector field is given by

$$\mathbf{v}_t(x) = -\frac{\nabla\mu_t}{\mu_t}(x) = -\frac{\nabla u}{u}(x,t), \tag{10}$$

where the last equality is justified by the equivalence of weak solution and the classical PDE solution (because $u(\cdot,\cdot)$ is Lipschitz continuous and $\mathbf{v}_t(\cdot)$ is Lipschitz). Since

$$\nabla u = (4\pi t)^{-n/2}\exp\Big(-\frac{|x-y|^2}{4t}\Big)\Big(-\frac{x}{2t}\Big) = -u\Big(\frac{I_n}{2t}\Big)x = -u\mathbf{v}_t(x), \tag{11}$$

the velocity vector field $\mathbf{v}_t : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map that can be represented by an $n \times n$ matrix:

$$\mathbf{v}_t = \frac{I_n}{2t}. \tag{12}$$

Thus in the heat equation, the velocity vector field $\mathbf{v}_t = (2t)^{-1} I_n$ does not depend on the location x (i.e., location-free) and it dies off as $t \to \infty$. The vanishing velocity means that the particles moving according to the heat equation will eventually converge to an equilibrium distribution given by a harmonic function $\Delta u = 0$. If the boundary value is imposed (either Dirichlet problem or Neumann problem) or the heat growth at infinity is not too fast, then the solution of the harmonic function is unique.

Let $p \geqslant 1$ and $\mathcal{P}_p(\mathbb{R}^n)$ be the collection of probability measures on $\mathbb{R}^n$ such that the p-Wasserstein distance is well-defined, i.e.,

$$\mathcal{P}_p(\mathbb{R}^n) = \Big\{ \mu \in \mathcal{P}(\mathbb{R}^n) : \int_{\mathbb{R}^n} |x - x_0|^p \mu(\mathrm{d}x) < \infty \text{ for some } x_0 \in \mathbb{R}^n \Big\}, \tag{13}$$

where $\mathcal{P}(\mathbb{R}^n)$ contains all probability measures on $\mathbb{R}^n$. Let W_p be the p-Wasserstein distance

$$W_p^p(\mu, \nu) = \min \Big\{ \int |x - y|^p \gamma(\mathrm{d}x, \mathrm{d}y) : \gamma \in \Gamma \Big\}, \tag{14}$$

where Γ is the set of all couplings with marginal distributions μ and ν.

Definition 2.1 (Metric derivative). Let $(\mu_t)_{t>0}$ be an absolutely continuous curve in the Wasserstein (metric) space $(\mathcal{P}_p(\mathbb{R}^n), W_p)$. The metric derivative at time t of the curve $t \mapsto \mu_t$ w.r.t. W_p is defined as

$$|\mu'|_p(t) = \lim_{h \to 0} \frac{W_p(\mu_{t+h}, \mu_t)}{|h|}. \tag{15}$$

We write $|\mu'|(t) = |\mu'|_2(t)$.

We can compute the metric derivative of the fundamental solution curve of the heat equation. Note that μ_t is the Gaussian density $N(y, 2tI_n)$ for each $t > 0$.

Lemma 2.2 (Wasserstein distance between two Gaussians, cf. Remark 2.31 in [11]). Let $\mu_1 = N(m_1, \Sigma_1)$ and $\mu_2 = N(m_2, \Sigma_2)$. Then the optimal transport map (i.e., the Monge map) from μ_1 to μ_2 is given by

$$T(x) = m_2 + A(x - m_1), \tag{16}$$

where

$$A = \Sigma_1^{-1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Sigma_1^{-1/2}, \tag{17}$$

and the squared 2-Wasserstein distance between μ_1 and μ_2 equals to

$$W_2^2(\mu_1, \mu_2) = |m_1 - m_2|^2 + \operatorname{tr} \Big\{ \Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2} \Big\}, \tag{18}$$

where the last term is the Bures distance on positive-semidefinite matrices. In particular, if $\Sigma_1 \Sigma_2 = \Sigma_2 \Sigma_1$, then

$$W_2^2(\mu_1, \mu_2) = |m_1 - m_2|^2 + |\Sigma_1^{1/2} - \Sigma_2^{1/2}|_F^2. \tag{19}$$

In the Gaussian case, we have

$$W_2^2(\mu_{t+h},\mu_t)=|\sqrt{2(t+h)}I_n-\sqrt{2t}I_n|_F^2=2n(\sqrt{t+h}-\sqrt{t})^2, \tag{20}$$

which gives a formulas for $|\mu'|(t)$:

$$|\mu'|(t)=\lim_{h\to 0}\sqrt{2n}\Big|\frac{\sqrt{t+h}-\sqrt{t}}{h}\Big|=\sqrt{2n}\,\frac{1}{2\sqrt{t}}=\sqrt{\frac{n}{2t}}. \tag{21}$$

On the other hand, recalling (12), we get

$$\|\mathbf{v}_t\|_{L^2(\mu_t)}^2:=\int|\mathbf{v}_t(x)|^2\mu_t(\mathrm{d}x)=\frac{1}{4t^2}\int|x|^2\mu_t(\mathrm{d}x)=\frac{\mathrm{tr}(2tI_n)}{4t^2}=\frac{n}{2t}. \tag{22}$$

This implies that

$$|\mu'|(t)=\|\mathbf{v}_t\|_{L^2(\mu_t)}=\sqrt{\frac{n}{2t}}. \tag{23}$$

Equivalence in (23) is a much more general fact; see ahead Theorem 3.1 and Corollary 3.1 in Section 3. Below we shall give a heuristic interpretation based on the theory of optimal transport to see why (23) is intuitively true. Consider $p\geqslant 1$. If $(\mu_t)_{t>0}$ is an absolute continuous curve in W_p, then there is an optimal transport map $T:\Omega\to\Omega$ moving particles from μ_t to μ_{t+h} that minimizes $\int|T(x)-x|^p\mu_t(\mathrm{d}x)$, i.e.,

$$W_p^p(\mu_t,\mu_{t+h})=\int|T(x)-x|^p\mu_t(\mathrm{d}x)=\int|(T-\mathrm{id})(x)|^p\mu_t(\mathrm{d}x). \tag{24}$$

The "discrete velocity" of the particle located at x at time t (i.e., time derivative of T) is given by

$$\mathbf{v}_t(x)=\frac{T(x)-x}{h}, \tag{25}$$

so that we have

$$\|\mathbf{v}_t\|_{L^p(\mu_t)}^p=\int\Big|\frac{T(x)-x}{h}\Big|^p\mu_t(\mathrm{d}x)=\frac{W_p^p(\mu_t,\mu_{t+h})}{|h|^p}. \tag{26}$$

Letting $h\to 0$, we expect that $\|\mathbf{v}_t\|_{L^p(\mu_t)}=|\mu'|_p(t)$.

The above heuristic argument (cf. (25)) suggests consider the following *gradient flow* (GF) for moving particles in $\mathbb{R}^n$:

$$\begin{cases} y_x'(t)=\mathbf{v}_t(y_x(t)),\\ y_x(0)=x, \end{cases} \tag{27}$$

where $y_x(t)$ is the time t position of the particle initially at $y_x(0)=x$, i.e., it is the *trajectory* of the particle starting from x.

Thus the gradient flow of the heat equation is given by the following Cauchy problem:

$$y_x'(t)=(2t)^{-1}y_x(t) \tag{28}$$

with the initial datum $y_x(0)=x$. This is a first-order linear (homogeneous) ordinary differential equation with initial value problem, whose solution can be easily computed. We can take a basic solution as

$$y_x(t)=Cx\sqrt{t},\quad t>0, \tag{29}$$

for any constant $C \in \mathbb{R}$.

Let $Y_t : \Omega \to \Omega$ be the *flow* of the vector field $\mathbf{v}_t$ on Ω defined through $Y_t(x) = y_x(t)$ in (27). Note that Y_t is indeed a flow on Ω since $Y_0(x) = y_x(0) = x$ and $Y_t(Y_s(x)) = Y_t(y_x(s)) = y_{y_x(s)}(t) = y_x(s+t) = Y_{s+t}(x)$ for any $s, t \geqslant 0$, so that Y_t on Ω is a group action of additive group on $\mathbb{R}_+ = [0, \infty)$.

Then the pushforward measure $\mu_t = (Y_t)_\sharp \mu_0$ is a weak solution of the continuity equation $\partial_t \mu_t + \nabla \cdot (\mu_t \mathbf{v}_t) = 0$ in (1) because for any C^1 test function $\phi : \Omega \to \mathbb{R}$ with bounded support,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int \phi \, \mathrm{d}\mu_t &= \frac{\mathrm{d}}{\mathrm{d}t} \int \phi \, \mathrm{d}(Y_t)_\sharp \mu_0 = \frac{\mathrm{d}}{\mathrm{d}t} \int (\phi \circ Y_t) \, \mathrm{d}\mu_0 = \frac{\mathrm{d}}{\mathrm{d}t} \int \phi(y_x(t)) \, \mathrm{d}\mu_0(x) \\
&= \int \langle \nabla \phi(y_x(t)), y_x'(t) \rangle \, \mathrm{d}\mu_0(x) = \int \langle \nabla \phi(y_x(t)), \mathbf{v}_t(y_x(t)) \rangle \, \mathrm{d}\mu_0(x) \\
&= \int \langle \nabla \phi(Y_t), \mathbf{v}_t(Y_t) \rangle \, \mathrm{d}\mu_0 = \int \langle \nabla \phi, \mathbf{v}_t \rangle \, \mathrm{d}(Y_t)_\sharp \mu_0 = \int \langle \nabla \phi, \mathbf{v}_t \rangle \, \mathrm{d}\mu_t.
\end{aligned}$$

2.2 Heat equation as a (continuous) Markov semi-group of operators

Next we see how to construct the stochastic process from the heat equation. Recall that $H(x, y, t)$ is the heat kernel in (8) coming from the fundamental solution of the heat equation (6). Denote γ as the standard Gaussian measure on $\mathbb{R}^n$, i.e., the density of γ equals to $(2\pi)^{-n/2} \exp(-|x|^2/2)$. Given a measurable function $f : \mathbb{R}^n \to \mathbb{R}$, we define the (linear) integral operator

$$P_t f(x) = \int_{\mathbb{R}^n} f(z) H(z, x, t) \, \mathrm{d}z, \quad t > 0, x \in \mathbb{R}^n, \tag{30}$$

which is sometimes referred as the *forward evolution*. Then one can verify that for all $s, t > 0$,

$$P_{t+s} f = P_t(P_s f) = P_s(P_t f), \tag{31}$$

$$\lim_{t \downarrow 0} P_t f(x) = f(x). \tag{32}$$

Thus $(P_t)_{t \geqslant 0}$ forms a continuous semi-group of linear operators on $L^p(\gamma)$, which allows us to define a (homogeneous) continuous Markov process $(X_t)_{t \geqslant 0}$ on $\mathbb{R}^n$ via

$$\mathbb{E}[f(X_{s+t}) \mid X_r, r \leqslant t] = (P_s f)(X_t), \quad \gamma\text{-almost surely}, \tag{33}$$

for all bounded measurable functions f, and $s, t \geqslant 0$. This Markov process is called the *heat diffusion process*, which we show below is the rescaled Wiener process (or Brownian motion) by a factor of $\sqrt{2}$. It is also easy to check (by smooth approximation and Taylor expansion) that the generator $\mathcal{A}$ of the semi-group $(P_t)_{t \geqslant 0}$ is given by the Laplacian operator $\mathcal{A} = \Delta$ in the sense that

$$\mathcal{A} f = \lim_{t \downarrow 0} \frac{P_t f - f}{t} = \Delta f, \tag{34}$$

where the limit exists in $L^2(\gamma)$. Indeed, note that the simi-group operator can be represented as

$$P_t f(x) = \mathbb{E}_{\xi \sim \gamma}[f(x + \sqrt{2t}\xi)]. \tag{35}$$

For small enough $t > 0$ and smooth $f : \mathbb{R}^n \to \mathbb{R}$, Taylor expansion yields

$$\begin{aligned} P_t f(x) - f(x) &= \mathbb{E}_{\xi\sim\gamma}[f(x + \sqrt{2t}\xi)] - f(x) \\ &= \mathbb{E}_{\xi\sim\gamma}\left[f(x) + \sqrt{2t}\xi \cdot \nabla f(x) + \frac{1}{2}\, 2t\xi^T \mathrm{Hess}_f(x)\xi + \text{higher-order terms}\right] - f(x) \\ &= t\,\mathrm{tr}\left\{\, \mathbb{E}_{\xi\sim\gamma}[\xi\xi^T]\mathrm{Hess}_f(x)\right\} = t\,\mathrm{tr}\left(\mathrm{Hess}_f(x)\right) = t\Delta f(x). \end{aligned}$$

This is not surprising in view of the heat equation $\partial_t u = \Delta u$, where the rate of change (i.e., time derivative) in the heat diffusion process equals to the Laplacian.

As opposed to the semi-group integral operators (P_t), the generator $\mathcal{A}$ is a (linear) differential operator that unfolds P_t and generates the stochastic process X_t. The semi-group integral operators are useful for finding the stochastic differential equations (SDEs) from the evolution PDEs of probability density functions. Thus the generator does the reverse job by converting the SDEs to the evolution PDEs of densities. We shall see an example on the Fokker-Planck equation in Section 3. We remark that for the heat equation, one can do explicit calculations on P_t and directly write down the Markov process by using (33) and its SDE as following.

Using the Markov property (33) with $f(x) = \mathbf{1}(x \leqslant y)$, we get

$$\mathbb{P}(X_{s+t} \leqslant y \mid X_r, r \leqslant t) = \mathbb{P}_{\xi\sim\gamma}(X_t + \sqrt{2s}\xi \leqslant y) \quad \text{for all } y \in \mathbb{R}^n, \tag{36}$$

where $\mathbb{P}_{\xi\sim\gamma}(\cdot)$ is the probability taken only w.r.t. to ξ. This implies that

$$\mathbb{P}(X_{s+t} - X_t \leqslant y \mid X_r, r \leqslant t) = \mathbb{P}_{\xi\sim\gamma}(\sqrt{2s}\xi \leqslant y) = \Phi\Big(\frac{y}{\sqrt{2s}}\Big), \tag{37}$$

where $\Phi(y)$ is the cumulative distribution function of γ. The last display shows that the process $(X_t)_{t\geqslant 0}$ has independent increments (i.e., $X_{s+t} - X_t$ does not depend on the past values) and $X_{s+t} - X_t \sim N(0, 2sI_n)$. Thus the heat diffusion process $(X_t)_{t\geqslant 0}$ is simply $\sqrt{2}$-rescaled standard Wiener process $(W_t)_{t\geqslant 0}$ on $\mathbb{R}^n$, which we write as

$$\mathrm{d}X_t = \sqrt{2}\,\mathrm{d}W_t. \tag{38}$$

Note that the factor $\sqrt{2}$ is due to the fact that Wiener process solves a slightly different version of the heat equation $\partial_t u - \frac{1}{2}\Delta u = 0$.

For $X_0 = 0$, then $\mathbb{E}[X_t] = 0$ and $\mathrm{Var}(X_t) = 2t$, which means that, after time t, the expected position of the heat diffusion process is $\sqrt{2t}$, which should be compared with the position of trajectory of $y_x(t)$ at time t given in (29).

2.3 Heat equation as a (negative) gradient flow in Wasserstein space

We have seen from (27) that the heat equation is the gradient flow of moving particles in $\mathbb{R}^n$. In this section, we show that the heat equation can also viewed as a (negative) gradient flow of the entropy in the Wasserstein space of probability measures. To do this, we need first to make sense what do we mean by a "derivative" of the entropy (or more general functionals) in terms of a density in the infinite-dimensional Wasserstein (or metric) space.

Definition 2.3 (First variation, Chapter 7 in [13]). Let ρ be a density on Ω. Given a functional $F : \mathcal{P}(\Omega) \to \mathbb{R}$, we call $\frac{\delta F}{\delta \rho}(\rho) : \mathcal{P}(\Omega) \to \mathbb{R}$, if it exists, the unique (up to additive constants) measurable function such that

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} F(\rho + h\chi) = \lim_{h\to 0} \Big| \frac{F(\rho + h\chi) - F(\rho)}{h} \Big| = \int \frac{\delta F}{\delta \rho}(\rho)\, \mathrm{d}\chi \tag{39}$$

holds for every *mean-zero* perturbation density χ such that $\rho + h\chi \in \mathcal{P}(\Omega)$ for all small enough h. The function $\frac{\delta F}{\delta \rho}(\rho) : \Omega \to \mathbb{R}$ is the *first variation* of F at ρ.

In $\mathbb{R}^n$, the first variation behaves like the directional derivatives projected to all possible directions $\chi \in \mathbb{R}^n$. For example, take $F(x) = \frac{1}{2}|x|^2$ for $x \in \mathbb{R}^n$. Then $\nabla F(x) = x$ and for any $y \in \mathbb{R}^n$,

$$\lim_{h\to 0} \Big| \frac{|x + hy|^2 - |x|^2}{2h} \Big| = \lim_{h\to 0} \Big| \langle x, y\rangle + \frac{h}{2}|y|^2 \Big| = \langle x, y\rangle = \langle \nabla F(x), y\rangle. \tag{40}$$

Comparing (39) with (40), we may interpret

$$\int \frac{\delta F}{\delta \rho}(\rho)\, \mathrm{d}\chi = \Big\langle \frac{\delta F}{\delta \rho}(\rho), \chi \Big\rangle \tag{41}$$

as an inner product in a Hilbert space. Thus first variation is an infinite-dimensional analog of the gradient in the finite-dimensional Euclidean space. Thus $\frac{\delta F}{\delta \rho}(\rho)$ can be viewed as a gradient in $\mathcal{P}(\Omega)$.

Note that first variation is defined only up to additive constants since $\int c\, \mathrm{d}\chi = 0$ for any constant $c \in \mathbb{R}$. Below are two important functionals that will be extremely useful in studying heat equation, or more generally the Fokker-Planck equation in Section 3.

Example 2.4 ("Generalized" entropy). Let $f : \mathbb{R} \to \mathbb{R}$ be a convex superlinear function and

$$F(\rho) = \int f(\rho(x))\, \mathrm{d}x = \int f \circ \rho. \tag{42}$$

Clearly

$$\frac{F(\rho + h\chi) - F(\rho)}{h} = \int \frac{f(\rho(x) + h\chi(x)) - f(\rho(x))}{h\chi(x)} \chi(x)\, \mathrm{d}x.$$

Letting $h \to 0$, we get

$$\frac{\mathrm{d}}{\mathrm{d}h}\Big|_{h=0} F(\rho + h\chi) = \int f'(\rho(x))\chi(x)\, \mathrm{d}x = \int f'(\rho)\, \mathrm{d}\chi, \tag{43}$$

which implies that

$$\frac{\delta F}{\delta \rho}(\rho) = f'(\rho). \tag{44}$$

For the special case where $f(\rho) = \rho \log \rho$ is the entropy, it first variation at ρ is given by

$$\frac{\delta F}{\delta \rho}(\rho) = 1 + \log \rho. \tag{45}$$

Example 2.5 (Potential). Let $V : \Omega \to \mathbb{R}$ be a potential function and the energy functional

$$F(\rho) = \int V\, \mathrm{d}\rho = \int V(x)\rho(x)\, \mathrm{d}x. \tag{46}$$

Compute

$$\frac{F(\rho+h\chi)-F(\rho)}{h}=\int V(x)\frac{\rho(x)+h\chi(x)-\rho(x)}{h}\,\mathrm{d}x=\int V(x)\chi(x)\,\mathrm{d}x=\int V\,\mathrm{d}\chi.$$

Thus we see that

$$\frac{\delta F}{\delta\rho}(\rho)=V \tag{47}$$

is a constant function that does not depend on ρ.

Intuitively, the first variation of a functional F (either entropy or energy) at ρ is the rate of change in distribution for moving the particles on Ω from ρ that minimizes the entropy/energy.

Given a functional $F:\mathcal{P}(\Omega)\to\mathbb{R}$, the *minimizing movement scheme* introduced by Jordan-Kinderlehrer-Otto [6] (sometimes also called the JKO scheme) is a time-discretized version of gradient flows that solves a sequence of iterated minimization problems (in the context of $(\mathcal{P}(\Omega),W_2)$):

$$\rho_{k+1}^h=\mathrm{argmin}_{\rho\in\mathcal{P}(\Omega)}\,F(\rho)+\frac{W_2^2(\rho,\rho_k^h)}{2h}. \tag{48}$$

By strong duality,

$$W_2^2(\rho,\rho_k^h)=\max_{\varphi\in\Phi_c(\Omega)}\int_\Omega\varphi\,\mathrm{d}\rho+\int_\Omega\varphi^c\,\mathrm{d}\rho_k^h, \tag{49}$$

where $\varphi^c(y)=\inf_{x\in\Omega}\{|x-y|^2-\varphi(x)\}$ is the c-transform[1]. The functions φ realizing the maximum on the right hand side of (49) is called the *Kantorovich potentials* for the transport from ρ to ρ_k^h. For each ρ_k^h, $W_2^2(\rho,\rho_k^h)$ is convex in ρ since it is a supremum of linear functionals in ρ.

Now differentiating (48) w.r.t. ρ, the first-order optimality condition is implied by

$$\frac{\delta F}{\delta\rho}(\rho_{k+1}^h)+\frac{\varphi}{h}=\text{constant}, \tag{50}$$

where φ now is the Kantorovich potential for the transport from ρ_{k+1}^h to ρ_k^h (not in the reversed direction). Here we implicitly assumed the uniqueness of c-concave Kantorovich potential.

From Brenier's polarization theorem, we know that optimal map $\widetilde{T}$ from μ_{t+h} and μ_t to and $\varphi:\Omega\to\mathbb{R}$ are linked through

$$\widetilde{T}(x)=x-\nabla\varphi(x). \tag{51}$$

So the velocity vector from time $t+h$ to t (note the reverse direction!) is given by

$$\widetilde{\mathbf{v}}_t=\frac{\widetilde{T}(x)-x}{h}=-\frac{\nabla\varphi(x)}{h}=\nabla\Big(\frac{\delta F}{\delta\rho}(\rho_{k+1}^h)\Big)(x). \tag{52}$$

Reverting the time direction ($\mathbf{v}_t=-\widetilde{\mathbf{v}}_t$) and letting $h\to0$ (using the continuity of ρ), we get

$$\mathbf{v}_t(x)=-\nabla\Big(\frac{\delta F}{\delta\rho}(\rho)\Big)(x) \tag{53}$$

[1]Here c stands for a general cost function and the c-transform in general is defined as $\varphi^c(y)=\inf_{x\in\Omega}\{c(x,y)-\varphi(x)\}$. A function φ is said to be c-concave if there exists a χ such that $\varphi=\chi^c$ and $\Phi_c(\Omega)$ is the set of all c-concave functions.

and we get the following continuity equation

$$\partial_t \rho - \nabla \cdot \Big(\rho \nabla \Big(\frac{\delta F}{\delta \rho}(\rho) \Big) \Big) = 0 \tag{54}$$

in the Wasserstein space of measures.

If we choose the entropy functional $F(\rho) = \int f(\rho(x))\,\mathrm{d}x$ with $f(\rho) = \rho \log \rho$, then

$$\frac{\delta F}{\delta \rho}(\rho) = 1 + \log \rho, \quad \text{and} \quad \nabla \Big(\frac{\delta F}{\delta \rho}(\rho) \Big) = \frac{\nabla \rho}{\rho} = \nabla \log \rho, \tag{55}$$

so that the continuity equation in (54) reads

$$\partial_t \rho - \nabla \cdot (\rho \nabla \log \rho) = 0. \tag{56}$$

Now recall that we can rewrite the heat equation as:

$$0 = (\partial_t - \Delta) u = \partial_t u - \nabla \cdot (u \frac{\nabla u}{u}) = \partial_t u - \nabla \cdot (u \nabla \log u), \tag{57}$$

where we can think of $\mu_t = u(\cdot, t)$ and $\mathbf{v}_t = \nabla \log u$ in the continuity equation (1). Thus we see that (56) and (57) are really the same continuity equation associated with the heat equation. However, they are viewed as different gradient flows in the spaces $(\mathcal{P}_2(\Omega), W_2)$ and Ω, respectively. In either case, we call it the *heat flow*.

3 Fokker-Planck equations: diffusions with drift

We start from the equivalence of the metric derivative and the velocity vector field.

Theorem 3.1. If $(\mu_t)_{t\in[0,1]}$ is an absolute continuous curve in W_p, then for any $t \in [0,1]$ a.e., there is a velocity vector field $\mathbf{v}_t \in L^p(\mu_t; \Omega)$ for $\Omega \subset \mathbb{R}^n$ such that:

1. μ_t is a weak solution of the continuity equation $\partial_t \mu_t + \nabla \cdot (\mu_t \mathbf{v}_t) = 0$ in the sense of distribution;
2. for a.e. t, we have $\|\mathbf{v}_t\|_{L^p(\mu_t)} \leqslant |\mu'|_p(t)$.

Conversely, if $(\mu_t)_{t\in[0,1]}$ are measures in $\mathcal{P}_p(\mathbb{R}^n)$ and $\mathbf{v}_t \in L^p(\mu_t; \Omega)$ for each t such that $\int_0^1 \|\mathbf{v}_t\|_{L^p(\mu_t)} \mathrm{d}t < \infty$ satisfying the continuity equation $\partial_t \mu_t + \nabla \cdot (\mu_t \mathbf{v}_t) = 0$, then

1. $(\mu_t)_{t\in[0,1]}$ is an absolute continuous curve in W_p;
2. for a.e. t, we have $|\mu'|_p(t) \leqslant \|\mathbf{v}_t\|_{L^p(\mu_t)}$.

Corollary 3.2. If $(\mu_t)_{t\in[0,1]}$ is an absolute continuous curve in W_p, then the velocity vector field given in part 1 of Theorem 3.1 must satisfy

$$|\mu'|_p(t) = \|\mathbf{v}_t\|_{L^p(\mu_t)}. \tag{58}$$

As in the heat equation, we can take two alternative perspectives on the continuity equation structure and the gradient flow in the spaces $(\mathcal{P}_2(\Omega), W_2)$ and Ω.

First, if we choose the functional

$$F(\rho) = \int f(\rho(x))\,\mathrm{d}x + \int V(x)\,\mathrm{d}\rho(x) = \int f \circ \rho + \int V\,\mathrm{d}\rho, \tag{59}$$

where $f(\rho) = \rho \log \rho$ is the entropy and $V : \Omega \to \mathbb{R}$ is a potential (independent of ρ), then the first variation of F at ρ is given by

$$\frac{\delta F}{\delta \rho}(\rho) = 1 + \log \rho + V. \tag{60}$$

Thus,

$$\nabla\Big(\frac{\delta F}{\delta \rho}(\rho)\Big) = \nabla \log \rho + \nabla V = \frac{\nabla \rho}{\rho} + \nabla V, \tag{61}$$

and the continuity equation for this entropy+potential functional F (note that F is convex in ρ) becomes

$$0 = \partial_t \rho - \nabla \cdot \Big(\rho \Big(\frac{\nabla \rho}{\rho} + \nabla V\Big)\Big) = \partial_t \rho - \Delta \rho - \nabla \cdot (\rho\, \nabla V), \tag{62}$$

or alternatively, we may write

$$\partial_t \rho = \Delta \rho - \langle \nabla \rho, \nabla V \rangle - \rho \Delta V, \tag{63}$$

which is the *Fokker-Planck equation* (FPE). Hence if we define the *drift Laplacian operator* $\mathcal{L}_V$ as

$$\mathcal{L}_V \rho = \Delta \rho + \langle \nabla \rho, \nabla V \rangle = e^V\, \nabla \cdot (e^{-V} \nabla \rho), \tag{64}$$

then the Fokker-Planck equation (63) is nothing but the *drift heat equation*:

$$\partial_t \rho - \mathcal{L}_V \rho - \rho \Delta V = 0, \tag{65}$$

which describes the time evolution of density of particles under the influence of drag forces and random forces (such as in Brownian motion or heat diffusion in the pure diffusion case.)

For a general potential V, from the continuity equation (62), the velocity vector field $\mathbf{v}_t$ is given by

$$\mathbf{v}_t = -\frac{\nabla \rho_t}{\rho_t} - \nabla V. \tag{66}$$

Thus we can write down the gradient flow of $\mathbf{v}_t$ as in (27). In particular, the Mehler flow (with the potential $V(x) = \frac{1}{4}|x|^2$) is given by

$$y_x'(t) = -\Big(\frac{\nabla \rho_t}{\rho_t} + \nabla V\Big)(y_x(t)) = -\Big(\frac{\nabla \rho_t}{\rho_t}\Big)(y_x(t)) - \frac{y_x(t)}{2} \tag{67}$$

with the initial datum $y_x(0) = x$, which again is a first-order linear (homogeneous) ordinary differential equation with initial value problem.

3.1 Mehler flow

It is interesting to note that as a special case if $V(x) = \frac{1}{4}|x|^2$, then $\nabla V = \frac{x}{2}, \Delta V = \frac{n}{2}$, and the drift heat equation (65) is the *Mehler flow*:

$$(\partial_t - L_M)u = 0, \tag{68}$$

where L_M is the Mehler operator defined as

$$L_M(\rho) = \mathcal{L}_{\frac{|x|^2}{4}}(\rho) + \frac{n}{2}\rho. \tag{69}$$

It is well-known that the Mehler flow can be obtained by the heat flow in the following sense. Let $u(x,t)$ be a smooth function and define the rescaled version

$$\widetilde{u}(x,t) = t^{n/2}u(\sqrt{t}x,t) \to c(4\pi)^{-n/2}\exp\Big(-\frac{|x|^2}{4}\Big) \quad \text{as } t \to \infty, \tag{70}$$

where $c = \int_{\mathbb{R}^n} u$ and the convergence holds by the central limit theorem (CLT). Changing the time clock unit $t = e^s$ and setting $v(x,s) = \widetilde{u}(x,e^s)$ for $s > 0$, one can verify that

$$(\partial_s - L_M)v = e^{\frac{ns}{2}+s}(\partial_t - \Delta)u. \tag{71}$$

Thus u solves the heat equation $(\partial_t - \Delta)u = 0$ if and only if v solves drift heat equation $(\partial_s - L_M)v = 0$. In particular, this implies that if we consider the fundamental solution of the heat equation given by (7) and (8):

$$u(x,t) = (4\pi t)^{-n/2}\exp\Big(-\frac{|x|^2}{4t}\Big), \quad t > 0, \tag{72}$$

then

$$v(x,s) = (4\pi)^{-n/2}\exp\Big(-\frac{|x|^2}{4}\Big) \tag{73}$$

is a constant in time. Thus $\partial_s v(x,s) = 0$ and we see that $L_M(v) = 0$, i.e., v is an L_M-harmonic function so that $v(x)$ is a static point that minimizes the Mehler energy: for $f : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$,

$$\mathcal{E}_t(f) = -\int f L_M(f) e^{\frac{|x|^2}{4}} = \int (|\nabla f|^2 - \frac{n}{2}f^2)e^{\frac{|x|^2}{4}} \geqslant 0. \tag{74}$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_t(f) = -2\langle f_t, L_M f\rangle_M = -2\mathcal{E}_t(f), \tag{75}$$

where $\langle f,g\rangle_M = \int_{\mathbb{R}^n} f(x)g(x)e^{\frac{|x|^2}{4}}\,\mathrm{d}x$ defines an inner product, we see that L_M is the twice negative gradient flow of the Mehler energy (74) in $\mathbb{R}^n$. Note that (75) implies that the exponential rate of convergence for the Mehler flow (i.e., the Fokker-Planck equation with $V(x) = \frac{1}{4}|x|^2$) in the Euclidean space $\mathbb{R}^n$: for $t \geqslant 0$,

$$\mathcal{E}_t(f) = \mathcal{E}_0(f)e^{-2t}. \tag{76}$$

In Section 3.3 below, we shall also look at the rate of convergence of the gradient flow of the Fokker-Planck equation (in particular the Mehler flow) to v in the Wasserstein space $(\mathcal{P}(\mathbb{R}^n), W_2)$.

3.2 Fokker-Planck equation as a stochastic differential equation

The Fokker-Planck equation (in the most general form) is a PDE that describes the time evolution of density of particles of a diffusion process with a (deterministic) drift and (random) noise, which is governed by the following stochastic differential equation (SDE) (sometimes referred as the *Langevin diffusion*):

$$\mathrm{d}X_t = m(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t, \tag{77}$$

where $m(X_t, t)$ is the drift coefficient vector in $\mathbb{R}^n$ and $\sigma(X_t, t)$ is an $n \times n$ matrix. Here (W_t) is again the standard Wiener process on $\mathbb{R}^n$. The SDE in (77) is understood in the integral form:

$$X_{t+s} - X_s = \int_s^{s+t} m(X_u, u)\,\mathrm{d}u + \int_s^{s+t} \sigma(X_u, u)\,\mathrm{d}W_u \tag{78}$$

as a sum of an Lebesgue integral and an Itô integral. In this model, both the random drift coefficient vector $m(X_t, t)$ and the random matrix $\sigma(X_t, t)$ are path/state and time dependent.

The (standard) n-dimensional Wiener process is the special case where $m(X_t, t) = 0$ and $\sigma(X_t, t) = I_n$. For $n = 1$, the density evolution equation of (77) is given by

$$\partial_t \rho(x, t) = -\partial_x(m(x, t)\rho(x, t)) + \partial_x^2(D(x, t)\rho(x, t)), \tag{79}$$

where $D(X_t, t) = \sigma(X_t, t)^2/2$ is the diffusion coefficient. (In Appendix A, we show how to convert the SDE of sample paths (77) to the evolution PDE of densities (79).) In the one-dimensional heat diffusion case (where there is no drift $m(x, t) = 0$) with $D(X_t, t) = 1$ (i.e., $\sigma(X_t, t) = \sqrt{2}$), the evolution of the density $\rho(x, t)$ is governed by

$$\partial_t \rho - \partial_x^2 \rho = 0, \tag{80}$$

which is just the heat equation on $\mathbb{R}$. Higher dimension analog $\partial_t u - \Delta u = 0$ can also be made by noting that the density evolution equation now becomes

$$\partial_t \rho(x, t) = -\nabla \cdot (m(x, t)\rho(x, t)) + \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}(D_{ij}(x, t)\rho(x, t)), \tag{81}$$

where $D(x, t) = \sigma(x, t)\sigma(x, t)^T/2$ is the diffusion tensor.

How does the density evolution equation (81) link to the continuity equation version (62) and (63)? With the diffusion tensor $D(x, t) = I_n$, (81) reads

$$\partial_t \rho = -\nabla \cdot (m\rho) + \Delta\rho. \tag{82}$$

Comparing the last display with the continuity equation (62):

$$\partial_t \rho - \Delta\rho - \nabla \cdot (\rho\,\nabla V) = 0, \tag{83}$$

we see that

$$m = -\nabla V, \tag{84}$$

which means that the mean drift vector m in the Fokker-Planck equation proceeds in the negative gradient direction of minimizing the potential V (and of course subject to the diffusion

effect given by the Laplacian Δ). Combining this with the fact that the heat flow proceeds in the negative gradient direction in the entropy, we see that the Fokker-Planck equation is the negative gradient flow of the entropy+potential functional

$$F(\rho) = \underbrace{\int f \circ \rho}_{\text{microscopic behavior}} + \underbrace{\int V \,\mathrm{d}\rho}_{\text{macroscopic behavior}} \tag{85}$$

in the Wasserstein space $(\mathcal{P}_2(\Omega), W_2)$.

We remark that in the special case $m(x,t) = -\frac{x}{2}$ (i.e., $V(x) = \frac{1}{4}|x|^2$) and $D(x,t) = 1$ in the Langevin diffusion (77) is often called the *Ornstein-Uhlenbeck (OU) process* in $\mathbb{R}^n$:

$$\mathrm{d}X_t = -\frac{X_t}{2}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{86}$$

which is simply a mean-reverting diffusion process. The semi-group operator of the OU process is given by

$$P_t f(x) = \mathbb{E}_{\xi\sim\pi}\left[f\Big(e^{-t}\frac{x}{2} + \sqrt{1-e^{-2t}}\xi\Big)\right], \quad t \geqslant 0, \tag{87}$$

where π is again $\sqrt{2}$-scaled standard Gaussian measure γ on $\mathbb{R}^n$, i.e., $\pi(x) = (4\pi)^{-n/2}\exp(-|x|^2/4)$. The OU semi-group $(P_t)_{t\geqslant 0}$ admits π as stationary measure, where the convergence holds in $L^2(\pi)$: for any bounded measurable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\begin{aligned}
\|P_t f - \pi f\|^2_{L^2(\pi)} &= \mathbb{E}_{\xi'\sim\pi}\Big|\,\mathbb{E}_{\xi\sim\pi}[f(e^{-t}\frac{\xi'}{2} + \sqrt{1-e^{-2t}}\xi)] - \mathbb{E}_{\xi\sim\pi}[f(\xi)]\Big|^2 \\
&\leqslant \mathbb{E}_{\xi'\sim\pi}\,\mathbb{E}_{\xi\sim\pi}\left[f(e^{-t}\frac{\xi'}{2} + \sqrt{1-e^{-2t}}\xi) - f(\xi)\right]^2 \qquad (88)\\
&= \mathbb{E}\left[f(e^{-t}\frac{\xi'}{2} + \sqrt{1-e^{-2t}}\xi) - f(\xi)\right]^2 \qquad (89)\\
&\to 0, \quad \text{as } t\to\infty, \qquad (90)
\end{aligned}$$

where (88) follows from Jensen's inequality, (89) from Fubini's theorem, and (90) from the dominated convergence theorem (since $f \in L^2(\pi)$). In addition, the generator $\mathcal{A}$ of $(P_t)_{t\geqslant 0}$ is given by

$$\mathcal{A}f = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} P_t f = -\frac{1}{2}\langle\nabla f, x\rangle + \Delta f, \tag{91}$$

which is called the Ornstein-Uhlenbeck (OU) operator (also write $\mathcal{A} = L_{OU}$). Comparing (91) with the Mehler operator in (69):

$$L_M(\rho) = \mathcal{L}_{\frac{|x|^2}{4}}(\rho) + \frac{n}{2}\rho = \Delta\rho + \frac{1}{2}\langle\nabla\rho, x\rangle + \frac{n}{2}\rho, \tag{92}$$

which gives the forward Fokker-Planck equation (cf. Appendix A for more details), the generator in (91) gives the backward Fokker-Planck equation. Integrating-by-parts w.r.t. $\mathrm{d}x$, we conclude that $L_M = \mathcal{A}^*$, which means that the Mehler operator (forward equation) is the adjoint of the generator, i.e., the OU operator (backward equation); that is, we have $L_M = L_{OU}^*$ in $L^2(\mathrm{d}x)$. This holds for a general potential V, not just $V(x) = \frac{1}{4}|x|^2$. (Here we need to be slightly careful on the reference measure: if we consider $L^2(e^{-\frac{|x|^2}{4}}\,\mathrm{d}x) = L^2(\mathrm{d}\pi)$, then $\mathcal{L}_{\frac{|x|^2}{4}} = L_{OU}^*$ in $L^2(\mathrm{d}\pi)$.)

To summarize, given a general Fokker-Planck equation:

$$0 = \partial_t \rho_t - \Delta \rho_t - \nabla \cdot (\rho_t \nabla V) \tag{93}$$

$$= \partial_t \rho_t - \mathcal{L}_V \rho_t - \rho_t \Delta V, \tag{94}$$

where (93) is the continuity equation version and (94) is the drifted heat equation version, if we assume it admits a stationary distribution $\pi(x) = \frac{1}{Z} e^{-V(x)}$ on $\mathbb{R}^n$ (cf. Section 3.3 ahead for more details), then the generator $\mathcal{A}$ of the drift heat diffusion process (i.e., the Langevin diffusion) is given by

$$\mathcal{A} = L_V^*, \tag{95}$$

where

$$L_V = \mathcal{L}_V + \rho \Delta V = \Delta \rho + \langle \nabla \rho, \nabla V \rangle + \Delta V, \tag{96}$$

and the semi-group $(P_t)_{t \geqslant 0}$ is given by

$$P_t f(x) = \mathbb{E}_{\xi \sim \pi}[f(e^{-t} \nabla V(x) + \sqrt{1 - e^{-2t}} \xi)], \quad t \geqslant 0. \tag{97}$$

Letting $t \to \infty$, we see that P_t asymptotically converges to the stationary distribution π (in $L^2(\pi)$). Thus from the evolution PDE of probability density functions, we can fully characterize the related SDEs. The reverse direction from SDEs to PDEs can be found in Appendix A, where we use Itô's formula to show that a measure solution ρ_t to the continuity equation can be seen as the law at time t of the process $(X_t)_{t>0}$ solution to the SDE. This justifies the equivalence between the PDEs and the SDEs.

3.3 Rate of convergence for Wasserstein gradient flows of the Fokker-Planck equation

If $Z = \int_{\mathbb{R}^n} e^{-V(x)} \, \mathrm{d}x < \infty$, then the Fokker-Planck equation has a stationary distribution on $\mathbb{R}^n$:

$$\pi(x) = \frac{1}{Z} e^{-V(x)}, \tag{98}$$

where $0 < Z < \infty$ is a normalization constant. To see this, recall that $F(\rho) = \int \rho \log \rho + \int V \, \mathrm{d}\rho$ and $\frac{\delta F}{\delta \rho}(\pi) = \log \pi + V$. Since $\nabla \pi = Z^{-1} e^{-V} (-\nabla V) = -\pi \nabla V$,

$$\pi \nabla \Big(\frac{\delta F}{\delta \rho}(\pi) \Big) = \pi \frac{\nabla \pi}{\pi} + \pi \nabla V = \nabla \pi + \pi \nabla V = 0. \tag{99}$$

Then,

$$\partial_t \pi = \nabla \cdot \Big(\pi \nabla \Big(\frac{\delta F}{\delta \rho}(\pi) \Big) \Big) = 0, \tag{100}$$

which implies that π is a stationary point of the Fokker-Planck continuity equation. In this case, the gradient flow of the functional F can be viewed as the gradient flow of the relative entropy between ρ and π (cf. (104) in Definition 3.3 below):

$$F(\rho) + \log Z = \int \rho \Big[\log \Big(\frac{\rho}{e^{-V}} \Big) + \log Z \Big] = \int \rho \log \Big(\frac{\rho}{\pi} \Big) = H(\rho \| \pi) \tag{101}$$

so that

$$\frac{\delta F}{\delta \rho}(\rho) = \frac{\delta (F + \log Z)}{\delta \rho}(\rho) = \frac{\delta H(\rho \| \pi)}{\delta \rho}(\rho). \tag{102}$$

Now suppose the Fokker-Planck equation admits a stationary distribution. Given an initial distribution ρ_0, we would like to ask how fast does the Wasserstein gradient flow $(\rho_t)_{t\geqslant 0}$ converge to the stationary distribution π? In a nutshell, the answer is given by the following statement.

> If the stationary measure π satisfies a logarithmic Sobolev inequality (LSI), then ρ_t converges to π exponentially fast in time t (under numerous distance or divergence measures).

Suppose π is probability measure (i.e., $0 < Z < \infty$). Since the convergence quantities involved depend only on the (first and second) derivatives of the density π, without loss of generality, we may assume $Z = 1$ and

$$\pi = e^{-V}. \tag{103}$$

Definition 3.3 (Relative entropy and relative Fisher information). Let p, q be two probability measures on $\mathbb{R}^n$ such that $q \ll p$. Then the *relative entropy* (or Kullback-Leibler divergence) between p and q is defined as

$$H(p\|q) = \int \log\left(\frac{\mathrm{d}p}{\mathrm{d}q}\right) \mathrm{d}p. \tag{104}$$

In particular, if $\mathrm{d}x \ll p$ and $\mathrm{d}x \ll q$, then

$$H(p\|q) = \int p \log\left(\frac{p}{q}\right) \mathrm{d}x. \tag{105}$$

Let π, μ be two probability measures on $\mathbb{R}^n$ such that $\pi \ll \mu$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} = \rho$, then the *relative Fisher information* between μ and π is defined as

$$I(\mu\|\pi) = \int \frac{|\nabla\rho|^2}{\rho} \mathrm{d}\pi. \tag{106}$$

We state a powerful information inequality that implies the LSI.

Theorem 3.4 (HWI++ inequality [5]). Let $(\mathbb{R}^n, |\cdot|, \pi)$ be a probability space such that

$$\mathrm{d}\pi(x) = e^{-V(x)} \mathrm{d}x, \tag{107}$$

where $V : \mathbb{R}^n \to [0, \infty)$ is a $C^\infty(\mathbb{R}^n)$ function such that $\mathrm{Hess}(V) \succeq \kappa I_n$ for some parameter $\kappa \in \mathbb{R}$. Then for any $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^n)$ such that $H(\mu_0\|\pi) < \infty$,

$$H(\mu_1\|\pi) - H(\mu_0\|\pi) \leqslant W_2(\mu_0, \mu_1)\sqrt{I(\mu_1\|\pi)} - \frac{\kappa}{2} W_2^2(\mu_0, \mu_1). \tag{108}$$

Theorem 3.4 implies several classical functional and information inequalities.

Corollary 3.5. In the setting of Theorem 3.4 and assume further $\pi \in \mathcal{P}_2(\mathbb{R}^n)$, then we have

1. *HWI inequality (Theorem 3 in [10])*: for any $\nu \in \mathcal{P}_2(\mathbb{R}^n)$,

$$H(\nu\|\pi) \leqslant W_2(\nu, \pi)\sqrt{I(\nu\|\pi)} - \frac{\kappa}{2} W_2^2(\nu, \pi). \tag{109}$$

2. *Talagrand's T_2-inequality*: for any $\nu \in \mathcal{P}_2(\mathbb{R}^n)$ with finite second moment,

$$\frac{\kappa}{2} W_2^2(\nu, \pi) \leqslant H(\nu \| \pi). \tag{110}$$

3. *Logarithmic Sobolev inequality*: if $\kappa > 0$, then for any $\nu \in \mathcal{P}(\mathbb{R}^n)$,

$$H(\nu \| \pi) \leqslant \frac{1}{2\kappa} I(\nu \| \pi). \tag{111}$$

Combining (110) and (111), we have that if $\kappa > 0$, then

$$W_2(\nu, \pi) \leqslant \frac{\sqrt{I(\nu \| \pi)}}{\kappa}, \quad \forall \pi \ll \nu. \tag{112}$$

In Appendix B and **??**, we discuss more details of the LSIs and Talagrand's transportation inequalities.

Proof of Corollary 3.5. Corollary 3.5 follows from Theorem 3.4 with specific choices.

1. Take $\mu_0 = \pi$ and $\mu_1 = \nu$.
2. Take $\mu_0 = \nu$ and $\mu_1 = \pi$.
3. Take $\mu_0 = \pi$ and $\mu_1 = \nu$ to get the HWI inequality in part 1, and then maximize its right-hand side $-\frac{\kappa}{2}x^2 + \sqrt{I(\nu\|\pi)}x$, where the maximizer occurs at $x^* = \frac{1}{\kappa}\sqrt{I(\nu\|\pi)}$. Then a standard approximation argument is sufficient.

■

There are various ways of measuring the gap between the gradient flow $(\rho_t)_{t \geqslant 0}$ as a solution of the Fokker-Planck equation and the stationary distribution π: total variation (as in Meyn-Tweedie's Markov chain approach), L^2-norm, relative entropy, Wasserstein distance, etc. Below in Theorem 3.6, we state the exponential rate of convergence under the relative entropy.

Theorem 3.6 (Exponential rate of convergence for the Fokker-Planck gradient flow: relative entropy). Let $(\rho_t)_{t \geqslant 0}$ solves the (gradient drift) Fokker-Planck equation:

$$\partial_t \rho_t = \nabla \cdot (\nabla \rho_t + \rho_t \nabla V), \tag{113}$$

where the potential $V > 0$ satisfies $V \in C^2(\mathbb{R}^n)$. If the Bakry-Émery criterion

$$\mathrm{Hess}(V) \succeq \kappa I_n \tag{114}$$

holds for some $\kappa > 0$ (i.e., $\mathrm{Hess}_V(x) \succeq \kappa I_n$ for all $x \in \mathbb{R}^n$), then

$$H(\rho_t \| \pi) \leqslant e^{-2\kappa t} H(\rho_0 \| \pi), \quad t \geqslant 0, \tag{115}$$

where $\pi(x) = e^{-V(x)}$ is the stationary distribution for the Fokker-Planck equation (113).

Note that the Bakry-Émery criterion is a strong convexity (sometimes called the κ-convexity) requirement for the potential V. From the sampling point of view, the Bakry-Émery criterion requires that the stationary distribution π is strictly log-concave to obtain exponential rate of convergence under the relative entropy.

It is known that strictly log-concave probability density satisfies an LSI, and vice versa (cf. Problem 3.17 in [15]). In the celebrated result of Bakry and Émery, a simple sufficient condition is given to ensure that the density satisfies an LSI.

Lemma 3.7 (Bakry-Émery criterion). Let $V : \mathbb{R}^n \to \mathbb{R}$ be a $C^2(\mathbb{R}^n)$ function and $\mathrm{d}\pi = e^{-V}\,\mathrm{d}x$ be a probability measure such that $\mathrm{Hess}(V) \succeq \kappa I_n$ for some $\kappa > 0$. Then π satisfies the LSI with parameter κ:

$$H(\nu\|\pi) \leqslant \frac{1}{2\kappa} I(\nu\|\pi) \tag{116}$$

for all $\nu \in \mathcal{P}(\mathbb{R}^n)$ such that $\pi \ll \nu$.

However, it is an open question whether or not all log-concave densities satisfy an LSI. It is even an open question to ask whether or not all log-concave distributions satisfy a Poincaré inequality with a *dimension-free* constant. This is the Kannan-Lovász-Simonovits (KLS) conjecture [1, 3].

Conjecture 3.8 (Kannan-Lovász-Simonovits conjecture, cf. Conjecture 1.2 in [1]). There exists an absolute constant $C > 0$ such that for any log-concave probability measure ν on $\mathbb{R}^n$, we have

$$\mathrm{Var}_\nu(f) := \mathbb{E}_\nu\,|f - \mathbb{E}_\nu[f]|^2 \leqslant C\lambda_\nu^2\,\mathbb{E}_\nu\,|\nabla f|^2 \tag{117}$$

for any locally Lipschitz (i.e., Lipschitz on any Euclidean ball) function $f \in L^2(\nu)$, where λ_ν is the square root of the largest eigenvalue of the covariance matrix $\mathbb{E}_{X\sim\nu}[XX^T]$.

Currently, the best known result is that, for an isotropic n-dimensional log-concave distribution, the dimension-dependent constant $C(n)$ is on the order $n^{1/4}$ [8].

We also note that the Bakry-Émery criterion can be replaced by a slightly stronger Otto-Villani criterion (with an additional finite second moment assumption), so that we can obtain both LSI and T_2 inequality.

Lemma 3.9 (Otto-Villani criterion, Theorem 1 in [10]). Let $V : \mathbb{R}^n \to \mathbb{R}$ be a $C^2(\mathbb{R}^n)$ function and $\mathrm{d}\pi = e^{-V}\,\mathrm{d}x$ be a probability measure with a finite second moment such that $\mathrm{Hess}(V) \succeq \kappa I_n$ for some $\kappa > 0$. Then π satisfies the LSI and Talagrand's T_2 inequality, both with parameter κ:

$$H(\nu\|\pi) \leqslant \frac{1}{2\kappa} I(\nu\|\pi) \quad \text{and} \quad \frac{\kappa}{2} W_2^2(\nu,\pi) \leqslant H(\nu\|\pi) \tag{118}$$

for all $\nu \in \mathcal{P}(\mathbb{R}^n)$ such that $\pi \ll \nu$.

With the (slightly) stronger Otto-Villani criterion, exponential rate of convergence also holds under the Wasserstein distance [2]. In fact, rate of convergence for (more general) non-gradient drift Fokker-Planck equation is established in [2].

Theorem 3.10 (Exponential rate of convergence for the Fokker-Planck gradient flow: Wasserstein distance). Let $(\rho_t)_{t\geqslant 0}$ solves the (gradient drift) Fokker-Planck equation (113) with the stationary distribution $\mathrm{d}\pi = e^{-V}\,\mathrm{d}x$ satisfying the Otto-Villani criterion for some $\kappa > 0$ (i.e., $\mathrm{d}\pi = e^{-V}\,\mathrm{d}x$ has finite second moment such that $V \in C^2(\mathbb{R}^n)$ and $\mathrm{Hess}(V) \succeq \kappa I_n$). Then,

$$W_2(\rho_t,\pi) \leqslant e^{-\kappa t} W_2(\rho_0,\pi). \tag{119}$$

(Stochastic) proof of Theorem 3.10. Using the relation between the PDE and SDE, solution to the (gradient drift) Fokker-Planck equation (113) is the density of the stochastic process $(X_t)_{t\geqslant 0}$ solving the following SDE:

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{120}$$

where $(W_t)_{t\geqslant 0}$ is the standard Wiener process on $\mathbb{R}^n$ and the initial datum has distribution ρ_0 (cf. Section 3.2 for more details of the equivalence).

Let μ_0 and ν_0 be two probability measures on $\mathbb{R}^n$, and (X_0, Y_0) be a coupling at the initial time with the marginal distributions $X_0 \sim \mu_0$ and $Y_0 \sim \nu_0$ such that

$$\mathbb{E}\,|X_0 - Y_0|^2 = W_2^2(\mu_0, \nu_0). \tag{121}$$

Then we run two coupled copies of the SDE with $(X_t)_{t\geqslant 0}$ (resp. $(Y_t)_{t\geqslant 0}$) as the solution to (120) starting from $X_0 \sim \mu_0$ (resp. $Y_0 \sim \nu_0$), both driven by the same Wiener process $(W_t)_{t\geqslant 0}$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\,|X_t - Y_t|^2 = -2\,\mathbb{E}\,\big\langle X_t - Y_t,\, \nabla V(X_t) - \nabla V(Y_t)\big\rangle. \tag{122}$$

Note that for any $x, y \in \mathbb{R}^n$,

$$V(x) = V(y) + \langle x - y, \nabla V(y)\rangle + \frac{1}{2}(x-y)^T \mathrm{Hess}_V(z)(x-y), \tag{123}$$

$$V(y) = V(x) + \langle y - x, \nabla V(x)\rangle + \frac{1}{2}(y-x)^T \mathrm{Hess}_V(z')(y-x), \tag{124}$$

where z, z' are on the line segment joining x and y. Adding the last two equalities, we get

$$\langle x - y, \nabla V(x) - \nabla V(y)\rangle = \frac{1}{2}(x-y)^T[\mathrm{Hess}_V(z) + \mathrm{Hess}_V(z')](x-y). \tag{125}$$

If $\mathrm{Hess}(V) \succeq \kappa I_n$ for some $\kappa > 0$, then

$$\big\langle x - y,\, \nabla V(x) - \nabla V(y)\big\rangle \geqslant \kappa |x-y|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{126}$$

Then Grönwall's lemma yields that

$$\mathbb{E}\,|X_t - Y_t|^2 \leqslant e^{-2\kappa t}\,\mathbb{E}\,|X_0 - Y_0|^2. \tag{127}$$

By definition of the Wasserstein distance,

$$W_2^2(\mu_t, \nu_t) \leqslant \mathbb{E}\,|X_t - Y_t|^2. \tag{128}$$

Now we have

$$W_2^2(\mu_t, \nu_t) \leqslant e^{-2\kappa t} W_2^2(\mu_0, \nu_0), \tag{129}$$

which gives an exponential contraction between *any* two solutions μ_t and ν_t to the Fokker-Planck equation (113). In particular, (119) follows from choosing ν_0 (and thus all $\nu_t, t \geqslant 0$) as the stationary solution $\pi = e^{-V}$. ∎

For the Mehler flow, recall that the stationary distribution is given by

$$\pi(x) = (4\pi)^{-n/2} \exp\Big(-\frac{|x|^2}{4}\Big), \tag{130}$$

which is a $\sqrt{2}$-rescaled standard Gaussian distribution γ on $\mathbb{R}^n$. By the Gaussian LSI for γ, we know that π satisfies a similar LSI and thus the Mehler flow has the exponential rate of convergence to its stationary distribution π.

Corollary 3.11 (Exponential rate of convergence for the Mehler flow). We have

$$H(\rho_t\|\pi) \leqslant e^{-t} H(\rho_0\|\pi) \quad \text{and} \quad W_2(\rho_t, \pi) \leqslant e^{-t/2} W_2(\rho_0, \pi), \quad t \geqslant 0, \tag{131}$$

where π is the stationary distribution the Mehler flow $(\rho_t)_{t\geqslant 0}$ with $V(x) = \frac{|x|^2}{4}$. In particular, ρ_t converges weakly to π (in distribution) as $t \to \infty$.

Proof of Corollary 3.11. For the Mehler flow, $V(x) = \frac{|x|^2}{4}, \nabla V = \frac{x}{2}$, and $\mathrm{Hess}(V) = \frac{1}{2} I_n$. So $\kappa = \frac{1}{2}$. Then Corollary 3.11 follows from Theorem 3.6 and Theorem 3.10. ∎

Corollary 3.11 is a quantitative version of Boltzmann's H-theorem stating that the total entropy of an isolated system can never decrease over time (i.e., the second law of thermodynamics):

$$\frac{\mathrm{d}}{\mathrm{d}t} H(\rho_t\|\pi) = -I(\rho_t\|\pi) \leqslant 0 \quad \text{with } \pi = \frac{e^{-V}}{Z}. \tag{132}$$

3.4 Problem set

Problem 3.1. Prove the central limit theorem for the Mehler flow in (70).

Problem 3.2. Derive the Mehler energy formula $\mathcal{E}_t(f)$ in (74) and show that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}_t(f) = -2\mathcal{E}_t(f). \tag{133}$$

Problem 3.3. Let $\mathcal{A}$ be the Ornstein-Uhlenbeck operator defined in (91) and L_M is the Mehler operator defined in (69).

1. Show that $L_M = \mathcal{A}^*$ in $L^2(\mathrm{d}x)$.
2. Show that $\mathcal{L}_{\frac{|x|^2}{4}} = \mathcal{A}^*$ in $L^2(e^{-\frac{|x|^2}{4}}\mathrm{d}x)$.

Problem 3.4. Prove Boltzmann's H-theorem in (132).

Problem 3.5. The KLS conjecture (Conjecture 3.8) implies a weaker *variance conjecture*: there exists an absolute constant $C > 0$ such that for any log-concave probability measure ν on $\mathbb{R}^n$, we have

$$\mathrm{Var}_{X\sim\nu}(|X|^2) \leqslant C\lambda_\nu \, \mathbb{E}_{X\sim\nu}[|X|^2], \tag{134}$$

where λ_ν is the square root of the largest eigenvalue of the covariance matrix $\Sigma = \mathbb{E}_{X\sim\nu}[XX^T]$. In particular, if $\Sigma = I_n$ is isotropic, then the variance conjecture reads

$$\mathrm{Var}_{X\sim\nu}(|X|^2) \leqslant Cn. \tag{135}$$

It is easy to see that the variance conjecture in (134) is a special case of the KLS conjecture by considering $f(x) = |x|^2$.

1. *Gaussian Poincaré inequality.* Let γ be the standard Gaussian measure on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function. Show that

$$\mathrm{Var}_\gamma(f) \leqslant \mathbb{E}_\gamma[|\nabla f|^2], \tag{136}$$

where the equality is attained for $f(x) = x_1 + \cdots + x_n$ and $x = (x_1, \ldots, x_n)$.

2. Show that the Gaussian Poincaré inequality (136) is sharp (up to a multiplicative constant) for $f(x) = x_1^2 + \cdots + x_n^2$.

3. *Exponential Poincaré inequality.* Let X be a Laplace (i.e., double-exponential) distribution on $\mathbb{R}$ (i.e., the density ν of X is given by $\nu(x) = \frac{1}{2}e^{-|x|}$ for $x \in \mathbb{R}$). Show that

$$\mathrm{Var}(f(X)) \leqslant 4\,\mathbb{E}[f'(X)^2], \tag{137}$$

4. Now using the Efron-Stein inequality (i.e., a special case of the tensorization technique), show that

$$\mathrm{Var}_{\nu^{\otimes n}}(f) \leqslant 4\,\mathbb{E}_{\nu^{\otimes n}}[|\nabla f|^2]. \tag{138}$$

4 Nonlinear parabolic equations: interacting diffusions

4.1 McKean-Vlasov equation

Consider the following nonlinear continuity equation:

$$\partial_t \rho = \Delta \rho - \nabla \cdot \Big(\rho \int_{\mathbb{R}^n} b(\cdot, y)\, \rho_t(\mathrm{d}y)\Big), \tag{139}$$

where $b : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a bounded Lipschitz function and $\rho_t(\cdot) = \rho(\cdot, t)$. Note that (139) is a nonlinear parabolic PDE, which arises from McKean's example of the *interacting diffusion process* (in statistical physics), which is described by the following SDE: starting from X_0,

$$\mathrm{d}X_t = \int_{\mathbb{R}^n} b(X_t, y)\rho_t(\mathrm{d}y)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{140}$$

where $\rho_t(\mathrm{d}y)$ is the distribution of X_t. In (140), the drift coefficient depends on the process state X_t and its marginal density ρ_t. Here we consider a simpler version of the McKean-Vlasov equations where the diffusion coefficient is constant. Note that (139) is the forward equation corresponding to time marginals of the nonlinear process (140) (cf. Appendix A). Indeed, by Itô's lemma, for $f \in C_b^2(\mathbb{R}^n)$,

$$\mathrm{d}f(X_t) = \Big[\Delta f(X_t) + \langle f(X_t), \int b(X_t, y)\rho_t(\mathrm{d}y)\rangle\Big]\,\mathrm{d}t + \langle \nabla f(X_t),\, \mathrm{d}W_t\rangle, \tag{141}$$

which implies the generator $\mathcal{A}$ of $(X_t)_{t\geqslant 0}$ is given by

$$\mathcal{A}f = \Delta f + \langle f, \int b(\cdot, y)\rho_t(\mathrm{d}y)\rangle. \tag{142}$$

Thus we get the adjoint operator

$$\mathcal{A}^*\rho = \Delta \rho - \nabla \cdot \Big(\rho \int b(\cdot, y)\rho_t(\mathrm{d}y)\Big), \tag{143}$$

which entails the forward equation (139).

Conversely, one can construct a unique nonlinear process with the time marginal distributions satisfying the continuity equation (139).

Theorem 4.1 (Theorem 1.1 in [14]). There is existence and uniqueness, trajectorial and in distribution for the solution of (139).

A construction of (a system of) the interacting diffusion processes is given in (145) below based on the interacting N-particle system.

4.2 Mean field asymptotics

Consider the N-particle system: for each particle $i = 1, \dots, N$,

$$\mathrm{d}X_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} b(X_t^{i,N}, X_t^{j,N})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i, \tag{144}$$

starting from $X_0^{i,N}$. Note that Theorem 4.1 allows us to construct the processes $(\overline{X}_t^i)_{t\geqslant 0}$ for each $i \geqslant 1$ as the solution of

$$\overline{X}_t^i = \overline{X}_0^i + \sqrt{2}W_t^i + \int_0^t \int b(\overline{X}_s^i, y)\rho_s(\mathrm{d}y)\,\mathrm{d}s, \tag{145}$$

where $\rho_s(\mathrm{d}y)$ is the distribution of $\overline{X}_s^i$. The following theorem gives a guarantee that the (weak) solution of the McKean-Vlasov equation is governed by the mean-field limit.

Theorem 4.2 (Finite-N approximation error for mean-field limit). For any $i \geqslant 1$ and $T > 0$, we have

$$\sup_N \sqrt{N}\,\mathbb{E}[\sup_{t\leqslant T} |X_t^{i,N} - \overline{X}_t^i|] < \infty. \tag{146}$$

The Vlasov equation is a special case of (144) where $b(x, x') = \nabla U(x - x')$ and $U : \mathbb{R}^n \to \mathbb{R}$ is a potential for the particle interactions. In this case, the system of Langevin equations for the N-particle system becomes

$$\mathrm{d}X_t^{i,N} = \frac{1}{N}\sum_{j=1}^{N} \nabla U(X_t^{i,N} - X_t^{j,N})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i. \tag{147}$$

If the initial distribution of X_0 is chaotic (i.e., $X_0 \sim \rho_0^{\otimes N}$) and U is smooth, then for each t, we have the mean-field limit:

$$\rho_t^{(N)} \rightsquigarrow \rho_t, \quad \text{as } N \to \infty, \tag{148}$$

where $\rho_t^{(N)} = N^{-1}\sum_{i=1}^N \delta_{X_t^{i,N}}$ is the empirical measure of $X_t^{i,N}$ and $\rightsquigarrow$ denotes weak convergence.

On the other hand, the Vlasov equation (139) can be viewed as a W_2-gradient flow for the functional

$$F(\rho) = \int \rho \log \rho + \frac{1}{2}\iint U(x - y)\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y), \tag{149}$$

because the first variation of the interaction potential

$$\mathcal{U}(\rho) = \frac{1}{2} \iint U(x-y)\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(y) \tag{150}$$

at ρ equals to $\frac{\delta \mathcal{U}}{\delta \rho}(\rho) = U * \rho$. Since $\nabla(U*\rho) = (\nabla U) * \rho$, the continuity equation in (56) with F given in (149) becomes

$$\partial_t \rho - \Delta \rho - \nabla \cdot (\rho\,(\nabla U) * \rho) = 0, \tag{151}$$

which can be thought as a nonlinear Fokker-Planck equation (62) with a state and density dependent free energy functional (see the similar comments for the SDE version (140)).

4.3 Application: landscape of two-layer neural networks

We present an application of the mean-field asymptotics to approximate the dynamics of stochastic gradient descent (SGD) for learning two-layer neural networks [9].

In supervised machine learning problems, one often wants to learn the relationship between $(x, y) \in \mathbb{R}^n \times \mathbb{R}$, where x is the feature vector and y is the label such that $(x, y) \sim \mu$. Suppose a sample of independent and identically distributed (i.i.d.) data $(x_i, y_i)_{i=1}^m$ is drawn and observed from the joint distribution μ. In a two-layer neural network, the dependence of y on x is modeled as:

$$\hat{y}(x; \boldsymbol{\theta}) = \frac{1}{N} \sum_{j=1}^{N} \sigma_*(x; \theta_j), \tag{152}$$

where N is the number of hidden units (i.e., neurons), $\sigma_* : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is an activation function, and $\theta_j \in \mathbb{R}^p$ are parameters for the j-th hidden units. Here we collectively denote all parameters as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. Let $\ell(\hat{y}, y) = (\hat{y} - y)^2$ be the square loss function. Given the activation function σ_*, the learning problem aims to find the "best" parameters $\boldsymbol{\theta}$ such that the finite-N risk (i.e., generalization error)

$$R_N(\boldsymbol{\theta}) = \mathbb{E}[\ell(\hat{y}(x; \boldsymbol{\theta}), y)] = \mathbb{E}[\hat{y}(x; \boldsymbol{\theta}) - y]^2 \tag{153}$$

is minimized. The parameter learning is often carried out by the SGD, which iteratively updates the parameter by the following rule:

$$\theta_j^{k+1} = \theta_j^k + 2 s_k (y_k - \hat{k}(x_k; \boldsymbol{\theta}^k)) \nabla_{\theta_j} \sigma_*(x_k; \theta_j^k), \tag{154}$$

where $\boldsymbol{\theta}^k = (\theta_1^k, \ldots, \theta_N^k)$ denotes the parameters after k iterations and s_k is a step size.

Note that we can decompose the finite-N risk as

$$R_N(\boldsymbol{\theta}) = R_\sharp + \frac{2}{N} \sum_{j=1}^{N} V(\theta_j) + \frac{1}{N^2} \sum_{i,j=1}^{N} U(\theta_i, \theta_j), \tag{155}$$

where $R_\sharp = \mathbb{E}[y^2]$ is the constant risk of $\hat{y} = 0$, $V(\theta_j) = -\mathbb{E}[y \sigma_*(x; \theta_j)]$, and $U(\theta_i, \theta_j) = \mathbb{E}[\sigma_*(x; \theta_i)\sigma_*(x; \theta_j)]$. Let $\rho^{(N)} = N^{-1} \sum_{j=1}^N \delta_{\theta_j}$ be the empirical measure of $\theta_1, \ldots, \theta_N$. Then R_N in (155) can be written as

$$R_N(\boldsymbol{\theta}) = R_\sharp + 2 \int V(\theta)\,\rho^{(N)}(\mathrm{d}\theta) + \iint U(\theta_1, \theta_2)\,\rho^{(N)}(\mathrm{d}\theta_1)\,\rho^{(N)}(\mathrm{d}\theta_2), \tag{156}$$

which suggests considering the minimization of a mean-field limit risk (as $N \to \infty$) defined for the probability measure $\rho \in \mathcal{P}(\mathbb{R}^p)$:

$$R(\rho) = R_\sharp + 2\int V(\theta)\,\rho(\mathrm{d}\theta) + \iint U(\theta_1,\theta_2)\,\rho(\mathrm{d}\theta_1)\,\rho(\mathrm{d}\theta_2). \tag{157}$$

Under mild assumptions,

$$\inf_{\boldsymbol{\theta}\in(\mathbb{R}^p)^{\otimes N}} R_N(\boldsymbol{\theta}) = \inf_{\rho\in\mathcal{P}(\mathbb{R}^p)} R(\rho) + O(1/N), \tag{158}$$

so that the mean-field limit provides a good approximation of the finite-N risk. In addition, if the SGD (154) has a chaotic initialization $\boldsymbol{\theta}^0 = (\theta_1^0,\dots,\theta_N^0) \sim \rho_0^{\otimes N}$ and it has step size chosen as $s_k = \varepsilon\xi(k\varepsilon)$ for some sufficiently regular function $\xi : \mathbb{R}_+ \to \mathbb{R}_+$, then the SGD for estimating $\boldsymbol{\theta}$ is also well approximated by a continuum dynamics on $\mathcal{P}(\mathbb{R}^p)$ in the sense that, for each fixed $t > 0$,

$$\rho_{t/\varepsilon}^{(N)} \rightsquigarrow \rho_t, \quad \text{as } N\to\infty,\ \varepsilon\to 0, \tag{159}$$

where $\rho_{t/\varepsilon}^{(N)} = N^{-1}\sum_{j=1}^N \delta_{\theta_j^k}$ is the empirical distribution after k SGD steps and $(\rho_t)_{t\geqslant 0}$ is a weak solution of the following *distributional dynamics*:

$$\partial_t\rho_t = 2\xi(t)\nabla_\theta\cdot(\rho_t\nabla_\theta\Psi(\theta;\rho_t)) \tag{160}$$

and

$$\Psi(\theta;\rho) = V(\theta) + \int U(\theta_1,\theta_2)\,\rho(\mathrm{d}\theta_2). \tag{161}$$

In particular, if $\xi(t)\equiv 1$, then the distributional dynamics in (160) and (161) is the Wasserstein gradient flow of the risk $R(\rho)$ in $(\mathcal{P}_2(\mathbb{R}^p), W_2)$.

Hence the mean-field approximation provides an "averaging out" perspective of the complexity of the landscape of neural networks, which can be useful to prove the convergence of the SGD algorithm or its variants.

5 Diffusions on manifold

5.1 Heat equation on manifold

In order to speak of the heat equation on manifold, we need first to extend the concept of gradient vector field ∇, divergence $\nabla\cdot$, and Laplace operator $\Delta = \nabla\cdot\nabla$ from the Euclidean space to the manifold setting.

Definition 5.1 (Divergence). Let X be a vector field on a smooth manifold $(M,\mathcal{O},\mathcal{A})$ with connection ∇. On a chart (U,x), the *divergence* $\nabla\cdot : \Gamma(TM)\to\mathbb{R}$ (sometimes also written as $\mathrm{div}(\cdot)$) is defined as

$$\mathrm{div}(X) := \nabla\cdot X = \left(\nabla_{\frac{\partial}{\partial x^i}} X\right)^i, \tag{162}$$

where $\nabla_{\frac{\partial}{\partial x^i}} X$ is the covariant derivative of X in the direction of $\frac{\partial}{\partial x^i}$ and the last expression is implied by the Einstein summation convention (cf. Appendix C.2 for the notation).

Remark 5.2. Often we also write the divergence as $\nabla_M \cdot$ or $\mathrm{div}_M(\cdot)$ to explicit denote it is a divergence on a (curved) manifold (rather than on a Euclidean space with a global coordinate system.) Moreover, even though the divergence is defined on charts, one can show that the divergence of a vector field is chart-independent. Similar comments apply to the gradient and Laplace-Beltrami operator below. ■

The next question is that how can we define an analog of gradient vector field? Recall that for a (smooth) function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient vector $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$. Extending gradient vector to gradient vector field on the domain in $\mathbb{R}^n$, we get the gradient vector field $\nabla : C^\infty(\mathbb{R}^n) \to \Gamma(T\mathbb{R}^n)$, where $\Gamma(T\mathbb{R}^n)$ is the set of all vector fields on the tangent bundle $T\mathbb{R}^n \cong \mathbb{R}^n$ (cf. Appendix C.4 for more details). Replacing the Euclidean space $\mathbb{R}^n$ by a Riemannian manifold M, we can define the gradient on such manifolds.

Definition 5.3 (Gradient on manifold). On a Riemannian manifold $(M, \mathcal{O}, \mathcal{A}, g)$, the *gradient* is defined as a map

$$\begin{aligned} \nabla \quad : \quad & C^\infty(M) \to \Gamma(TM), \\ & f \mapsto \nabla f := \flat^{-1}(\mathrm{d}f), \quad \text{for } f \in C^\infty(M), \end{aligned} \tag{163}$$

where $\flat := \flat_g : \Gamma(TM) \to \Gamma(T^*M)$ is the metric-induced (invertible) musical map defined in (270) (cf. Appendix C.6).

With the concept of gradient (Definition 5.3) and divergence (Definition 5.1), we can talk about the Laplace-Beltrami operator on a Riemannian manifold.

Definition 5.4 (Laplace-Beltrami operator). On a Riemannian manifold $(M, \mathcal{O}, \mathcal{A}, g)$, the *Laplace-Beltrami operator* is a linear operator $\Delta : C^\infty(M) \to C^\infty(M)$ is define as

$$\Delta f = \nabla \cdot \nabla f. \tag{164}$$

In particular, on a (positively) oriented Riemannian manifold $(M, \mathcal{O}, \mathcal{A}^\uparrow, g)$, the Laplace-Beltrami operator in a given chart $(U, x) \in \mathcal{A}^\uparrow$ can be expressed as (cf. Chapter 1 in [12])

$$\Delta f = \frac{1}{\sqrt{\det(g)}} \frac{\partial}{\partial x^i} \Big(\sqrt{\det(g)}\, g^{ij} \frac{\partial f}{\partial x^j} \Big). \tag{165}$$

The (linear) heat equation on a Riemannian manifold $(M, \mathcal{O}, \mathcal{A}, g)$ is defined as

$$(\partial_t - \Delta) u = 0, \tag{166}$$

where $u : M \times [0, \infty) \to \mathbb{R}$ is a smooth function and Δ is the Laplace-Beltrami operator (164).

5.2 Application: manifold clustering

We present an application of manifold clustering based on the idea of the heat diffusion on Riemannian manifolds [4].

For $k = 1, \dots, K$, let M_k be compact n_k-dimensional Riemannian manifolds that are disjoint. Suppose we observe a sequence of independent random variables $X_1, \dots, X_m$ taking values in $M := \sqcup_{k=1}^K M_k$. Suppose further that there exists a clustering structure $G_1^*, \dots, G_K^*$ (i.e., a partition on $[n] := \{1, \dots, n\}$ such that $[n] = \sqcup_{k=1}^K G_k^*$) such that each of the m data points belong to one of the K clusters: if $i \in G_k^*$, then $X_i \sim \mu_k$ for some probability measure μ_k supported on M_k. Given the observations $X_1, \dots, X_m$, the manifold clustering task is to develop (computationally feasible) algorithms with strong theoretical guarantees to recover the true clustering structure $G_1^*, \dots, G_K^*$.

6 Nonlinear geometric equations

6.1 Mean curvature flow

Now we turn to the *nonlinear* and *geometric* continuity equations that describe the (time) evolution of manifolds deformed by certain volume functional. Let $M := M^n \subset \mathbb{R}^N$ be an n-dimensional *closed* submanifold properly embedded in an ambient Euclidean space $\mathbb{R}^N$. For $p \in M$, denote T_pM as the tangent space to M at p, and $T_p^\perp M$ as the orthogonal complement of T_pM. Let

$$A : T_pM \times T_pM \to T_p^\perp M \tag{167}$$

be the second fundamental form that is bilinear and symmetric defined as

$$A(\mathbf{u}, \mathbf{v}) = \nabla^\perp_{\mathbf{v}} \mathbf{u}, \tag{168}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two vector fields tangent to M, and ∇ is the Euclidean covariant derivative of $\mathbf{u}$ in the direction of $\mathbf{v}$ (cf. the definition of covariant derivative in Appendix C.5). The *mean curvature* is defined as

$$H = -\operatorname{tr}(A) \tag{169}$$

such that $H(p) = -\sum_{i=1}^n \nabla^\perp_{e_i} e_i$, where $(e_i)_{i=1}^n$ is an orthonormal basis (ONB) for T_pM.

Given an infinitely differentiable (i.e., smooth), compactly supported, normal vector field $\mathbf{v}$ on M, consider the one-parameter variation

$$M_{s,\mathbf{v}} = \{x + s\mathbf{v}(x) : x \in M\}, \tag{170}$$

which gives a curve $s \mapsto M_{s,\mathbf{v}}$ in the space of submanifolds with $M_{0,\mathbf{v}} = M$. To study the geometric flow of the one-parameter family submanifolds, we need first to compute the time derivative of volume, which is given by the first variation formula of volume.

Lemma 6.1 (First variation of the volume functional). The *first variation* formula of volume is given by

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \operatorname{Vol}(M_{s,\mathbf{v}}) = \int_M \langle \mathbf{v}, H \rangle, \tag{171}$$

where H is the mean curvature vector in (169) and the integration $\int_M$ is with respect to the volume form dVol_M.

Example 6.2 (n-sphere). Consider the n-sphere $M = \{x \in \mathbb{R}^{n+1} : |x| = r\}$ with radius r embedded in $\mathbb{R}^{n+1}$ for $n \geqslant 1$ and $\mathbf{v} = x/|x|$ is the unit normal vector field (since M is a hypersurface in $\mathbb{R}^{n+1}$). Consequently the one-parameter variation of M is given by

$$M_{s,\mathbf{v}} = \left\{x + s\frac{x}{|x|} : x \in M\right\} = \left\{\left(1 + \frac{s}{r}\right)x : |x| = r\right\}. \tag{172}$$

Then,

$$\operatorname{Vol}(M_{s,\mathbf{v}}) = \left(1 + \frac{s}{r}\right)^n \operatorname{Vol}(M) \tag{173}$$

and

$$\frac{\operatorname{Vol}(M_{s,\mathbf{v}}) - \operatorname{Vol}(M)}{s} = \frac{\left(1 + \frac{s}{r}\right)^n - 1}{s} \operatorname{Vol}(M). \tag{174}$$

Letting $s \to 0$, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \mathrm{Vol}(M_{s,\mathbf{v}}) = \frac{n}{r}\,\mathrm{Vol}(M). \tag{175}$$

On the other hand,

$$\langle H, \mathbf{v}\rangle = -\sum_{i=1}^{n} \langle \nabla^{\perp}_{e_i} e_i, \mathbf{v}\rangle = \sum_{i=1}^{n} \langle e_i, \nabla_{e_i}\mathbf{v}\rangle = \nabla_M \cdot \mathbf{v} = \nabla_M \cdot \Big(\frac{x}{|x|}\Big) = \frac{n}{r}, \tag{176}$$

which implies that

$$\int_M \langle H, \mathbf{v}\rangle = \frac{n}{r}\,\mathrm{Vol}(M). \tag{177}$$

Combining (175) and (177), we see that (171) indeed holds for n-sphere.

Definition 6.3 (Mean curvature flow). A one-parameter family of n-dimensional submanifolds $M_t^n \subset \mathbb{R}^N$ is said to flow (or move by motion) by the *mean curvature flow* (MCF) if

$$x_t = -H, \tag{178}$$

where $x_t = \frac{\partial x}{\partial t}$ is the normal component of the time derivative of the position vector x. If $n = 1$, then (178) is also called the *curve-shortening flow*.

In light of the first variation formula (171), we see that the mean curvature flow (178) is the negative gradient flow of volume.

Let $M_t^n \subset \mathbb{R}^N$ be n-dimensional submanifolds flowing by the MCF. By Lemma **??**, coordinate functions $x_1, \ldots, x_N$ of the ambient Euclidean space restricted to the evolving submanifolds satisfy the (nonlinear) heat equation

$$(\partial_t - \Delta_{M_t}) x_i = 0, \quad i = 1, \ldots, N. \tag{179}$$

In fact, the heat equation (179) is a characterization of the mean curvature flow (178).

Definition 6.4 (Gaussian surface area or volume). Let $M \subset \mathbb{R}^N$ be an n-dimensional submanifold. Define the *Gaussian surface area or volume* as

$$F(M) = (4\pi)^{-n/2} \int_M e^{\frac{-|x|^2}{4}}. \tag{180}$$

6.2 Problem set

Problem 6.1. Prove the first variation formula of volume in Lemma 6.1.

Problem 6.2. Prove that the coordinate functions satisfy the nonlinear heat equation (179).

A From SDE to PDE, and back

We derive the PDE from the associated SDE. The derivation is based on Itô's calculus and a guess-and-verify trick. Let $(X_t)_{t\geqslant 0}$ be a one-dimensional diffusion process of the form

$$\mathrm{d}X_t = m(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t. \tag{181}$$

For a twice differentiable function $f(x,t)$, Itô's lemma gives

$$\mathrm{d}f(X_t,t) = (\partial_t f + m\partial_x f + D\partial_x^2 f)\,\mathrm{d}t + \sigma\partial_x f\,\mathrm{d}W_t, \tag{182}$$

where $D(X_t,t) = \frac{1}{2}\sigma(X_t,t)^2$. Equation (182) implies that the *backward equation* is given by

$$\partial_t f(x,t) + m(x,t)\partial_x f(x,t) + D(x,t)\partial_x^2 f(x,t) = 0. \tag{183}$$

Indeed, if (183) holds, then (182) implies that for any $T > t$,

$$f(X_T,T) - f(X_t,t) = \int_t^T \sigma(X_s,s)\partial_x f(X_s,s)\,\mathrm{d}W_s. \tag{184}$$

Taking the conditional expectation on X_t, we get

$$\mathbb{E}[f(X_T,T) \mid X_t] - f(X_t,t) = \mathbb{E}\left[\int_t^T \sigma(X_s,s)\partial_x f(X_s,s)\,\mathrm{d}W_s \mid X_t\right] = 0, \tag{185}$$

where the last step follows from the fact that the Itô integral of a martingale is a martingale. Now we need to verify that the solution f to the equation

$$f(X_t,t) = \mathbb{E}[f(X_T,T) \mid X_t] \tag{186}$$

should satisfy the assumed backward equation (183). (In finance, the function $f(x,t)$ is called the *value function* at time t and (186) represents the value function at a future time point $T > t$.) Using $T = t + \mathrm{d}t$ and the Itô formula trick (i.e., expanding time derivative to the first order and spatial derivative to the second order due to the quadratic variation of $(W_t)_{t\geqslant 0}$), we see that solution to (186) satisfies the backward equation (183).

Now if we define the differential operator $\mathcal{A}$ as:

$$\mathcal{A}f = m\partial_x f + D\partial_x^2 f, \tag{187}$$

then the backward equation (183) can be written as

$$\partial_t f + \mathcal{A}f = 0. \tag{188}$$

The operator $\mathcal{A}$ is called the *generator* of the diffusion process $(X_t)_{t\geqslant 0}$. Then we find the adjoint operator $\mathcal{A}^*$ in $L^2(\mathrm{d}x)$ satisfying $\langle \mathcal{A}f, \rho\rangle = \langle f, \mathcal{A}^*\rho\rangle$ for all ρ and f, where $\langle f, g\rangle = \int fg$. Integrating by parts, we see that the adjoint operator $\mathcal{A}^*$ in $L^2(\mathrm{d}x)$ is given by

$$\mathcal{A}^*\rho = -\partial_x(m\rho) + \partial_x^2(D\rho). \tag{189}$$

Thus the *forward equation* is determined by the adjoint operator:

$$\partial_t \rho = \mathcal{A}^*\rho. \tag{190}$$

Note that (190) is nothing but the one-dimensional Fokker-Planck equation given in the PDE form:

$$\partial_t \rho(x,t) = -\partial_x(m(x,t)\rho(x,t)) + \partial_x^2(D(x,t)\rho(x,t)), \tag{191}$$

where $\rho(x,t)$ is the density of particles at position x at time t. This is indeed a forward equation since it predicts $\rho(x,t)$ from the initial datum $\rho(x,0)$.

The process of deriving the SDE from the PDE in the Fokker-Planck equation case is described in Section 3.2.

B Logarithmic Sobolev inequalities and relatives

A probability measure π is said to satisfy the logarithmic Sobolev inequality (LSI) if there exists a $\kappa > 0$ such that for any $\nu \in \mathcal{P}(\mathbb{R}^n)$ and $\pi \ll \nu$,

$$H(\nu\|\pi) \leqslant \frac{1}{2\kappa} I(\nu\|\pi). \tag{192}$$

B.1 Gaussian Sobolev inequalties

Let γ be the standard Gaussian measure on $\mathbb{R}^n$, i.e., $\gamma(x) = (2\pi)^{-n/2} \exp(-|x|^2/2)$.

Lemma B.1 (Gaussian LSI: information-theoretic version). For any $\nu \in \mathcal{P}(\mathbb{R}^n)$ and $\gamma \ll \nu$,

$$H(\nu\|\gamma) \leqslant \frac{1}{2} I(\nu\|\gamma). \tag{193}$$

From Lemma B.1, we see that the standard Gaussian measure γ satisfies the LSI with $\kappa = 1$. The equality in (193) is achieved if ν is a translation of γ. To see this, take

$$\nu(x) = (2\pi)^{-n/2} \exp\Big(- \frac{|x-y|^2}{2} \Big), \quad y \in \mathbb{R}^n. \tag{194}$$

Then

$$\begin{aligned} H(\nu\|\gamma) &= \int \log\Big(\frac{\nu}{\gamma}\Big)\, \mathrm{d}\nu = \int -\frac{1}{2}(|x-y|^2 - |x|^2)\, \mathrm{d}\nu \\ &= \int \langle x, y\rangle\, \mathrm{d}\nu - \frac{1}{2}\int |y|^2\, \mathrm{d}\nu = |y|^2 - \frac{1}{2}|y|^2 = \frac{1}{2}|y|^2. \end{aligned} \tag{195}$$

On the other hand, since $\gamma \ll \nu$, we have

$$\rho(x) = \frac{\mathrm{d}\nu}{\mathrm{d}\gamma} = \exp\Big(- \frac{|x-y|^2}{2} + \frac{|x|^2}{2} \Big) = \exp(\langle x, y\rangle)\, \exp\Big(- \frac{|y|^2}{2} \Big) \tag{196}$$

and

$$\nabla\rho(x) = y \exp(\langle x, y\rangle)\, \exp\Big(- \frac{|y|^2}{2} \Big) = y\rho. \tag{197}$$

Then

$$I(\nu\|\gamma) = \int \frac{|y|^2\rho^2}{\rho}\, \mathrm{d}\gamma = \int |y|^2 \rho\, \mathrm{d}\gamma = |y|^2. \tag{198}$$

Combining (195) and (198), we conclude $H(\nu\|\gamma) = \frac{1}{2} I(\nu\|\gamma)$. In fact, translation is the only case that is currently known to achieve the equality case. Hence we pose the following conjecture.

Conjecture B.2 (Characterization of the equality case in the Gaussian LSI). The equality case $H(\nu\|\gamma) = \frac{1}{2} I(\nu\|\gamma)$ in the Gaussian LSI is achieved if and only if ν is a translation of γ, i.e., $\nu(x) = \gamma(x + y)$ for some $y \in \mathbb{R}^n$.

Lemma B.3 (Gaussian LSI: functional version). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function such that $f > 0$ and $\int f^2\, \mathrm{d}\gamma = 1$. Then

$$\int f^2 \log f\, \mathrm{d}\gamma \leqslant \int |\nabla f|^2\, \mathrm{d}\gamma. \tag{199}$$

Equivalently, we can write

$$\mathrm{Ent}_\gamma(f^2) \leqslant 2\int |\nabla f|^2 \,\mathrm{d}\gamma, \tag{200}$$

where $\mathrm{Ent}_\gamma(g) = \int g \log g \,\mathrm{d}\gamma$ is the entropy of the probability density $g > 0$.

It is easy to see that Lemma B.3 and B.1 are equivalent. Let $g = f^2$. Then $\nabla g = 2f\nabla f$ and $\nabla f = \frac{1}{2\sqrt{g}}\nabla g$. Note that (199) is equivalent to

$$\int g \log g \,\mathrm{d}\gamma \leqslant 2\int \Big|\frac{1}{2\sqrt{g}}\nabla g\Big|^2 \,\mathrm{d}\gamma = \frac{1}{2}\int \frac{|\nabla g|^2}{g}\,\mathrm{d}\gamma. \tag{201}$$

Since $g > 0$ and $\int g = 1$, we can define a probability measure μ such that $\frac{\mathrm{d}\mu}{\mathrm{d}\gamma} = g$. Then (201) can be written as

$$H(\mu\|\gamma) \leqslant \frac{1}{2} I(\mu\|\gamma). \tag{202}$$

Since g is arbitrary, the equivalence between Lemma B.3 and B.1 follows.

B.2 Euclidean Sobolev inequalties

There are Euclidean LSIs and the following Lemma B.4 is a version with sharp constant.

Lemma B.4 (Euclidean LSI). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function such that $\int f^2 \,\mathrm{d}x = 1$. Then

$$\int f^2 \log f^2 \,\mathrm{d}x \leqslant \frac{n}{2}\log\Big(\frac{2}{n\pi e}\int |\nabla f|^2 \,\mathrm{d}x\Big). \tag{203}$$

Lemma B.5 (Sobolev inequality). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function with compact support. Then for all $n \geqslant 3$,

$$\int |f|^{\frac{2n}{n-2}} \leqslant C(n)\int |\nabla f|^2, \tag{204}$$

where

$$C(n) = \frac{1}{n(n-2)}\Big(\frac{\Gamma(n)}{\Gamma(n/2)}\Big)^{\frac{2}{n}} > 2. \tag{205}$$

Note that the Gaussian LSIs (Lemma B.1 and B.3) are dimension-free inequalities, while the Euclidean LSI (Lemma B.4) and Sobolev inequality (Lemma B.5) are not. The constants in (203) and (204) are both sharp. Indeed, direct computation gives the constant $\frac{2}{n\pi e}$ from (205). Let $m = nk$ and $k \to \infty$ and take a function $f : \mathbb{R}^{nk} \to \mathbb{R}$ such that $\int f^2 \,\mathrm{d}x = 1$. Then

$$C(m) = \frac{1}{\pi nk(nk-2)}\Big(\frac{\Gamma(nk)}{\Gamma(nk/2)}\Big)^{\frac{2}{nk}} \sim \frac{2^{\frac{1}{nk}}}{\pi n^2k^2}\,\frac{2nk}{e} \sim \frac{2}{nk\pi e} = \frac{2}{m\pi e}, \tag{206}$$

where we used Stirling's approximation

$$(nk)! \sim \sqrt{2\pi nk}\Big(\frac{nk}{e}\Big)^{nk}. \tag{207}$$

C Riemannian geometry: some basics

In this section, we collect some basic facts and results about Riemannian geometry.

C.1 Smooth manifolds

Definition C.1 (Topological manifold). A topological space $(M, \mathcal{O})$ is said to be an n-dimensional *topological manifold* if for any $p \in M$, there is an open set $U \in \mathcal{O}$ containing p such that there exists a map $x : U \to x(U) \subset \mathbb{R}^n$ (equipped with the standard topology on $\mathbb{R}^n$) satisfying: (i) x is invertible, i.e., there is a map $x^{-1} : x(U) \to U$; (ii) x is continuous; (iii) x^{-1} is continuous. In other words, x is a *homeomorphism* between U and $x(U)$.

The pair (U, x) in Definition C.1 is called a *chart* and x is called the *chart map*. For $p \in U$ and $x(p) = (x^1(p), \dots, x^n(p))$, $x^i(p)$ is the i-th coordinate of $x(p)$. An *atlas* is a collection of charts $\mathcal{A} = \{(U_\alpha, x_\alpha) : \alpha \in A\}$ such that $M = \cup_{\alpha \in A} U_\alpha$.

For two charts (U, x) and (V, y) such that $U \cap V \neq \emptyset$, the map $y \circ x^{-1} : \mathbb{R}^n \supset x(U \cap V) \to y(U \cap V) \subset \mathbb{R}^n$ is said to be the *chart transition map*. We say (U, x) and (V, y) are $\clubsuit$-*compatible* if either $U \cap V = \emptyset$, or $U \cap V \neq \emptyset$ if the chart transition maps $y \circ x^{-1} : x(U \cap V) \to y(U \cap V)$ and $x \circ y^{-1} : y(U \cap V) \to x(U \cap V)$ has certain $\clubsuit$-property (as a map from $\mathbb{R}^n$ to $\mathbb{R}^n$). For instance, the $\clubsuit$-property can be $C^0(\mathbb{R}^n), C^1(\mathbb{R}^n), \dots, C^\infty(\mathbb{R}^n)$, and so on. An atlas $\mathcal{A}$ is $\clubsuit$-compatible if the transition maps between any two charts in $\mathcal{A}$ have the $\clubsuit$-property.

Definition C.2 (Smooth manifold). A *smooth manifold* $(M, \mathcal{O}, \mathcal{A})$ is a topological manifold $(M, \mathcal{O})$ equipped with a $C^\infty(\mathbb{R}^n)$-compatible atlas $\mathcal{A}$.

In these notes, we shall only consider smooth manifolds unless otherwise indicated.

Definition C.3 (Smooth functions on manifold). Let $(M, \mathcal{O}, \mathcal{A})$ be a smooth manifold. A function $f : M \to \mathbb{R}$ is said to be a *smooth map* (or C^∞-map) if $f \circ y^{-1} : \mathbb{R}^n \to \mathbb{R}$ is smooth for every chart y in the atlas $\mathcal{A}$.

According to Definition C.3, the coordinate functions $x^1, \dots, x^n$ of any chart (U, x) in a C^∞-compatible atlas are all smooth maps.

A map $\phi : M \to N$ is *smooth* if there is a chart (U, x) in M and a chart (V, y) in N such that $\phi(U) \subset V$ and the map $y \circ \phi \circ x^{-1} : \mathbb{R}^m \supset x^{-1}(U) \to y(V) \subset \mathbb{R}^n$ is C^∞.

Definition C.4 (Diffeomorphism). Let $(M, \mathcal{O}_M, \mathcal{A}_M)$ and $(N, \mathcal{O}_N, \mathcal{A}_N)$ are two smooth manifolds. We say that $(M, \mathcal{O}_M, \mathcal{A}_M)$ and $(N, \mathcal{O}_N, \mathcal{A}_N)$ are *diffeomorphic* if there exists a bijection $\phi : M \to N$ such that $\phi : M \to N$ and $\phi^{-1} : N \to M$ are both smooth maps.

C.2 Tensors

Let $(V, +, \cdot)$ be a vector space and $(V^*, \oplus, \odot)$ be the dual space, where

$$V^* = \{\phi : V \xrightarrow{\sim} \mathbb{R}\} =: \mathrm{Hom}(V, \mathbb{R}) \tag{208}$$

is the set of linear functionals on V. An element $\phi \in V^*$ is called a *covector*.

Definition C.5 (Tensor). Let $(V, +, \cdot)$ be a vector space and $r, s \in \mathbb{N}_0 := \{0, 1, \dots\}$. An *(r,s)-tensor* T over V is an $\mathbb{R}$-multi-linear map

$$T : \underbrace{V^* \times \cdots \times V^*}_{r \text{ times}} \times \underbrace{V \times \cdots \times V}_{s \text{ times}} \xrightarrow{\sim} \mathbb{R}. \tag{209}$$

By Definition C.5, a covector $\phi \in V^*$ is a $(0, 1)$-tensor over V. If $\dim(V) < \infty$, then $V \cong (V^*)^*$ is an isomorphism so that $v \in V$ can be identified as a linear map $V^* \xrightarrow{\sim} \mathbb{R}$, which means that v is a $(1, 0)$-tensor.

Let V be an n-dimensional vector space with an (arbitrarily chosen) basis $(e_1, \ldots, e_n)$. Then the *dual basis* $(\varepsilon^1, \ldots, \varepsilon^n)$ for V^* is uniquely determined by

$$\varepsilon^i(e_j) = \delta^i_j, \tag{210}$$

where $\delta^i_j = 1$ if $i = j$, and $\delta^i_j = 0$ if $i \neq j$.

Definition C.6 (Components of tensor). Let T be an (r, s)-tensor over an n-dimensional vector space V such that $n < \infty$. Let $(e_1, \ldots, e_n)$ be a basis of V and $(\varepsilon^1, \ldots, \varepsilon^n)$ be the dual basis of V^*. Then the *components* of T w.r.t. the chosen basis are defined as the $(r + s)^n$ real numbers (or sometimes called *coefficients*)

$$T^{i_1,\ldots,i_r}{}_{j_1,\ldots,j_s} = T(\varepsilon^{i_1}, \ldots, \varepsilon^{i_r}, e_{j_1}, \ldots, e_{j_s}) \tag{211}$$

for $i_1, \ldots, i_r, j_1, \ldots, j_s \in \{1, \ldots, n\}$.

As an example, consider a $(1, 1)$-tensor T with components given by $T^i{}_j = T(\varepsilon_i, e_j)$. Then for any $v \in V$ and $\phi \in V^*$, we can express

$$T(\phi, v) = T\Big(\sum_{i=1}^n \phi_i \varepsilon^i, \sum_{j=1}^n v^j e_j\Big) = \sum_{i=1}^n \sum_{j=1}^n \phi_i v^j T(\varepsilon^i, e_j) = \sum_{i=1}^n \sum_{j=1}^n \phi_i v^j T^i{}_j. \tag{212}$$

Thus components fully determine the tensor (given the basis). To avoid write too many sums, we typically use the *Einstein summation convention*:

$$T(\phi, v) = \phi_i v^j T^i{}_j, \tag{213}$$

where the repeated up-and-down indices i and j are summed over.

C.3 Tangent and cotangent spaces

Definition C.7 (Velocity). Let $(M, \mathcal{O}, \mathcal{A})$ be a smooth manifold and $\gamma : \mathbb{R} \to M$ be a curve (at least C^1). Let $p \in M$ and $\gamma(t_0) = p$ for some $t_0 = p$. The *velocity* of γ at p is the linear map defined as

$$\begin{aligned} v_{\gamma,p} \quad &: \quad C^\infty(M) \xrightarrow{\sim} \mathbb{R}, \\ &\qquad f \mapsto v_{\gamma,p}(f) = (f \circ \gamma)'(t_0). \end{aligned} \tag{214}$$

Note that $(C^\infty(M), \oplus, \odot)$ forms a vector space equipped with $(f \oplus g)(p) = f(p) + g(p)$ and $(\lambda \odot g)(p) = \lambda \cdot g(p)$. Note further that $(C^\infty(M), \oplus, \odot)$ is not only a vector space, it is also a *ring*. However, $(C^\infty(M), \oplus, \odot)$ is not a field!

Definition C.8 (Tangent vector space). Let $(M, \mathcal{O}, \mathcal{A})$ be a smooth manifold. For each $p \in M$,

$$T_pM := \Big\{ v_{\gamma,p} : \gamma \text{ is a smooth curve in } M \Big\} \tag{215}$$

is the *tangent space* to M at p.

Intuitively, we may understand the velocity $v_{\gamma,p}$ (i.e., a tangent vector in T_pM) as the directional derivative along the curve γ at p. The following Lemma C.9 confirms that the tangent space T_pM is indeed a vector space.

Lemma C.9. For each $p \in M$, T_pM is a vector space of linear maps from $C^\infty(M)$ to $\mathbb{R}$.

Proof of Lemma C.9. We shall define $\oplus$ and $\odot$ to make $(T_pM, \oplus, \odot)$ a vector space. For $f \in C^\infty(M)$, we define

$$\begin{aligned} \oplus \quad &: \quad T_pM \times T_pM \to \mathrm{Hom}(C^\infty(M), \mathbb{R}), \\ &\quad (v_{\gamma,p} \oplus v_{\delta,p})(f) = v_{\gamma,p}(f) + v_{\delta,p}(f), \end{aligned} \tag{216}$$

and

$$\begin{aligned} \odot \quad &: \quad \mathbb{R} \times T_pM \to \mathrm{Hom}(C^\infty(M), \mathbb{R}), \\ &\quad (s \odot v_{\gamma,p})(f) = s \cdot v_{\gamma,p}(f). \end{aligned} \tag{217}$$

We still need to show that there are smooth curves σ and τ such that $v_{\sigma,p} = v_{\gamma,p} \oplus v_{\delta,p}$ and $v_{\tau,p} = s \odot v_{\gamma,p}$. We first construct the curve τ via

$$\begin{aligned} \tau \quad &: \quad \mathbb{R} \to M, \\ &\quad t \mapsto \tau(t) = \gamma(st + t_0) =: \gamma \circ \mu_s(t), \end{aligned} \tag{218}$$

where $\mu_s(t) = st + t_0$. Now $\tau(0) = \gamma(t_0) = p$ and by the chain rule,

$$\begin{aligned} v_{\tau,p}(f) = (f \circ \tau)'(0) = (f \circ \gamma \circ \mu_s)'(0) = \mu_s'(0) \cdot (f \circ \gamma)'(\mu_s'(0)) \\ = s \cdot (f \circ \gamma)'(t_0) = s \cdot v_{\gamma,p}(f) = (s \odot v_{\gamma,p})(f). \end{aligned} \tag{219}$$

Next choose a chart (U, x) such that $p \in U$, and define

$$\begin{aligned} \sigma_x \quad &: \quad \mathbb{R} \to M, \\ &\quad t \mapsto \sigma_x(t) = x^{-1}((x \circ \gamma)(t_0 + t) + (x \circ \delta)(t_1 + t) - (x \circ \gamma)(t_0)), \end{aligned} \tag{220}$$

where $\gamma(t_0) = \delta(t_1) = p$. Then $\sigma_x(0) = p$ and by the chain rule,

$$\begin{aligned} v_{\sigma_x,p}(f) = (f \circ \sigma_x)'(0) = \big((f \circ x^{-1}) \circ (x \circ \sigma_x)\big)'(0) \\ = \big((x \circ \sigma_x)^i\big)'(0) \cdot \big(\partial_i (f \circ x^{-1})\big)(x(\sigma_x(0))) \\ = \Big(\big((x \circ \gamma)^i\big)'(t_0) + \big((x \circ \delta)^i\big)'(t_1)\Big) \cdot \big(\partial_i (f \circ x^{-1})\big)(x(p)) \\ = \Big(\big((x \circ \gamma)^i\big)'(t_0)\Big) \cdot \big(\partial_i (f \circ x^{-1})\big)(x(p)) + \Big(\big((x \circ \delta)^i\big)'(t_1)\Big) \cdot \big(\partial_i (f \circ x^{-1})\big)(x(p)) \\ = \big((f \circ x^{-1}) \circ (x \circ \gamma)\big)'(t_0) + \big((f \circ x^{-1}) \circ (x \circ \delta)\big)'(t_1) \\ = (f \circ \gamma)'(t_0) + (f \circ \delta)'(t_1) = v_{\gamma,p}(f) + v_{\delta,p}(f) = (v_{\gamma,p} \oplus v_{\delta,p})(f), \end{aligned} \tag{221}$$

which the last line does not depend on the choice of the chart (U, x). ∎

Notation. In the proof of Lemma C.9, we have seen that for a curve $\gamma : \mathbb{R} \to M$ such that $\gamma(0) = p$, where (U, x) is a chart containing p,

$$v_{\gamma,p}(f) = \underbrace{\big((x \circ \gamma)^i\big)'(0)}_{=: \dot\gamma_x^i(0)} \cdot \underbrace{\big(\partial_i (f \circ x^{-1})\big)(x(p))}_{=: \left(\frac{\partial f}{\partial x^i}\right)_p}, \quad f : M \to \mathbb{R} \text{ smooth}, \tag{222}$$

where we reserve $\partial_i g$ as the j-th Euclidean partial derivative of the function $g : \mathbb{R}^n \to \mathbb{R}$. Thus we write the velocity vector $v_{\gamma,p} \in T_pM$ as a linear map from $C^\infty(M)$ to $\mathbb{R}$ defined as

$$v_{\gamma,p} = \dot{\gamma}_x^i(0)\Big(\frac{\partial}{\partial x^i}\Big)_p, \tag{223}$$

where $\Big(\frac{\partial}{\partial x^i}\Big)_p$ is the chart-induced basis of T_pM and $\dot{\gamma}_x^i(0)$ are the components of $v_{\gamma,p}$ w.r.t. the chart-induced basis. We emphasize that $\Big(\frac{\partial f}{\partial x^i}\Big)_p$ in (223) are *not* partial derivatives of f because it does not make sense to speak of partial derivatives for a function defined on a manifold where there is no *global* canonical basis such as in $\mathbb{R}^n$. Once we see $\Big(\frac{\partial f}{\partial x^i}\Big)_p$, we need always to translate back to its definition in (222).

Lemma C.10 (Chart-induced basis of tangent space). Let $(M, \mathcal{O}, \mathcal{A})$ be a smooth manifold and (U, x) be a chart in $\mathcal{A}$ that contains p. Then

$$\Big(\frac{\partial}{\partial x^1}\Big)_p, \dots, \Big(\frac{\partial}{\partial x^n}\Big)_p \tag{224}$$

form a basis of T_pU. In particular, $\dim(T_pM) = d = \dim(M)$, where $\dim(T_pM)$ is the vector space dimension of T_pM and $\dim(M)$ is the dimension of the topological manifold $(M, \mathcal{O})$.

Proof of Lemma C.10. We have shown in (223) that every $v_{\gamma,p} \in T_pU$ can be expressed as a linear combination of $\Big(\frac{\partial}{\partial x^1}\Big)_p, \dots, \Big(\frac{\partial}{\partial x^n}\Big)_p$. It remains to show that these vectors are linearly independent. Since $\mathcal{A}$ is a $C^\infty(\mathbb{R}^n)$-compatible atlas, $x^j : U \to \mathbb{R}$ is smooth. By (222),

$$\alpha^i\Big(\frac{\partial}{\partial x^i}\Big)_p(x^j) = \alpha^i\big(\partial_i(x^j \circ x^{-1})\big)(x(p)) = \alpha^i\delta_i^j = \alpha^j, \tag{225}$$

since $x^j \circ x^{-1} : \mathbb{R}^n \to \mathbb{R}$ such that $(x^j \circ x^{-1})(c_1, \dots, c_n) = c_j$. If

$$0 = \alpha^i\Big(\frac{\partial}{\partial x^i}\Big)_p, \tag{226}$$

then $\alpha^j = 0$ for all $j = 1, \dots, n$, which means that the chart-induced basis vectors in (224) are linearly independent. ∎

Definition C.11 (Cotangent space). The *cotangent space* T_p^*M is the dual space of T_pM, i.e.,

$$T_p^*M = \{\phi : T_pM \xrightarrow{\sim} \mathbb{R}\} \tag{227}$$

contains the set of linear functionals on T_pM.

Example C.12 (Gradient is a covector in the cotangent space). Let $f \in C^\infty(M)$ and consider

$$\begin{aligned} (\mathrm{d}f)_p \quad &: \quad T_pM \xrightarrow{\sim} \mathbb{R}, \\ &\quad\; X \mapsto (\mathrm{d}f)_p(X) = Xf, \quad \text{for } X \in T_pM. \end{aligned} \tag{228}$$

We call $(\mathrm{d}f)_p$ is the *gradient* of f at $p \in M$. Clearly $(\mathrm{d}f)_p \in T_p^*M$ (i.e., $(\mathrm{d}f)_p$ is a covector) and it is a $(0,1)$-tensor over the vector space T_pM. Thus the components of $(\mathrm{d}f)_p$ w.r.t. the chart-induced basis by (U, x) (cf. (211)) is given by

$$\big((\mathrm{d}f)_p\big)_j = (\mathrm{d}f)_p\Big(\Big(\frac{\partial}{\partial x^j}\Big)_p\Big) = \Big(\frac{\partial f}{\partial x^j}\Big)_p \quad \text{for } j = 1, \dots, n. \tag{229}$$

The interpretation of gradient is as follows. The directional derivative Xf of f along a tangent vector X at p is the gradient $(\mathrm{d}f)_p$ evaluated at X. Thus gradient $(\mathrm{d}f)_p$ gives the information of directional derivatives of f along all possible tangent vectors at such point p.

Lemma C.13 (Dual basis of cotangent space). Let $(M, \mathcal{O}, \mathcal{A})$ be a smooth manifold and (U, x) be a chart in $\mathcal{A}$ that contains p, where $x^j : U \to \mathbb{R}$ is the j-th coordinate of x. Then

$$(\mathrm{d}x^1)_p, \ldots, (\mathrm{d}x^n)_p \tag{230}$$

form a dual basis of T_p^*M w.r.t. the basis $\left(\frac{\partial}{\partial x^1}\right)_p, \ldots, \left(\frac{\partial}{\partial x^n}\right)_p$ of T_pM.

Proof of Lemma C.13. The lemma follows from

$$(\mathrm{d}x^i)_p\Big(\frac{\partial}{\partial x^j}\Big)_p = \Big(\frac{\partial x^i}{\partial x^j}\Big)_p = \delta^i_j \tag{231}$$

and the dual basis definition in (210). ■

Now suppose we have two charts (U, x) and (V, y) such that $U \cap V \neq \emptyset$. Take a point $p \in U \cap V$ and a tangent vector $X \in T_pM$. Then we may express X in the two charts using different local coordinate systems:

$$X = X^i_x\Big(\frac{\partial}{\partial x^i}\Big)_p = X^j_y\Big(\frac{\partial}{\partial y^j}\Big)_p. \tag{232}$$

Let $f \in C^\infty(M)$. Applying the (multi-variable) chain rule, we compute

$$\begin{aligned}\Big(\frac{\partial}{\partial x^i}\Big)_p f = \partial_i(f \circ x^{-1})(x(p)) &= \partial_i\big((f \circ y^{-1}) \circ (y \circ x^{-1})\big)(x(p)) \\ &= \partial_i(y^j \circ x^{-1})(x(p)) \cdot \big(\partial_j(f \circ y^{-1})\big)(y(p)) \\ &= \Big(\frac{\partial y^j}{\partial x^i}\Big)_p\Big(\frac{\partial f}{\partial y^j}\Big)_p.\end{aligned} \tag{233}$$

Thus the last two displays imply that

$$X^i_x\Big(\frac{\partial y^j}{\partial x^i}\Big)_p\Big(\frac{\partial}{\partial y^j}\Big)_p = X^j_y\Big(\frac{\partial}{\partial y^j}\Big)_p. \tag{234}$$

Since $\left(\frac{\partial}{\partial y^j}\right)_p$ is a basis vector, we obtain the formula for the change of vector components under a change of chart given by

$$X^j_y = \Big(\frac{\partial y^j}{\partial x^i}\Big)_p X^i_x, \tag{235}$$

which is a linear map at the given point p. However, we should emphasize that the global chart transformation is nonlinear.

We can also derive a formula for change of covector components under a change of chart. Let $\omega \in T_p^*M$. Then we may write

$$\omega = \omega_{(x)i}(\mathrm{d}x^i)_p = \omega_{(y)j}(\mathrm{d}y^j)_p. \tag{236}$$

By similar computations in the tangent space, one can show that

$$\omega_{(y)j} = \Big(\frac{\partial x^i}{\partial y^j}\Big)_p \omega_{(x)i} \tag{237}$$

and the matrix $\left(\frac{\partial x^i}{\partial y^j}\right)_p$ is the inverse of the matrix $\left(\frac{\partial y^j}{\partial x^i}\right)_p$.

C.4 Tangent bundles and tensor fields

We have defined the tangent and cotangent spaces at a given point $p \in M$ in Section C.3. In this section, we consider the field extension of these concepts to the whole manifold based on the theory of bundles.

Definition C.14 (Bundle and fiber). A *bundle* is a triple (E, M, π) such that

$$\pi : E \to M, \tag{238}$$

where E is a smooth manifold ("total space"), M is a smooth manifold ("base space"), and π is a surjective smooth map ("projection map"). For $p \in M$, the *fiber* over p is the pre-image $\pi^{-1}(p)$ of $\{p\}$.

Consider a smooth n-dimensional manifold $(M, \mathcal{O}, \mathcal{A})$. Let

$$TM = \bigsqcup_{p \in M} T_p M \tag{239}$$

be a total space and $\pi : TM \to M$ be the surjective map defined via $X \mapsto p$ where p is the unique point in M such that $X \in T_pM$. We would like first to turn TM into a topological manifold such that π is continuous. We consider the coarsest topology:

$$\mathcal{O}_{TM} = \{\pi^{-1}(U) : U \in \mathcal{O}\}, \tag{240}$$

which is sometimes also referred as the *initial topology* w.r.t. π.

Next we need to construct a C^∞-atlas on TM from the C^∞-atlas on M. Let

$$\mathcal{A}_{TM} = \{(TU, \xi_x) : (U, x) \in \mathcal{A}\}, \tag{241}$$

where the chart map $\xi_x : TU \to \mathbb{R}^{2n}$ is defined as

$$X \mapsto \xi_x(X) = \big(\underbrace{(x^1 \circ \pi)(X), \ldots, (x^n \circ \pi)(X)}_{(U,x)\text{-coordinates of base point } \pi(X)}, \underbrace{(\mathrm{d}x^1)_{\pi(X)}(X), \ldots, (\mathrm{d}x^n)_{\pi(X)}(X)}_{\text{components of } X \text{ w.r.t. } (U,x)} \big). \tag{242}$$

Note that

$$\begin{aligned} \xi_x^{-1} \quad : \quad & \xi_x(TU) \to TU, \\ & (a^1, \ldots, a^n, b^1, \ldots, b^n) \mapsto \Big(b^1 \Big(\frac{\partial}{\partial x^1}\Big)_{x^{-1}(a^1,\ldots,a^n)}, \cdots, b^n \Big(\frac{\partial}{\partial x^n}\Big)_{x^{-1}(a^1,\ldots,a^n)} \Big), \end{aligned} \tag{243}$$

where $x^{-1}(a^1,\dots,a^n)=\pi(X)$ is the base point. Then we need to check that the atlas $\mathcal{A}_{TM}$ is smooth. Let (U,x) and (V,y) be two charts in $\mathcal{A}$ such that $U\cap V\neq\emptyset$. Combining (242) and (243), we can compute the chart transition map from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2n}$ in $\mathcal{A}_{TM}$ as following:

$$\begin{aligned}(\xi_y\circ\xi_x^{-1})(a^1,\dots,a^n,b^1,\dots,b^n)&=\xi_y\Big(b^j\Big(\frac{\partial}{\partial x^j}\Big)_{x^{-1}(a^1,\dots,a^n)}\Big)\\ =\Big(\dots,(y^i\circ\pi)\Big(b^j\Big(\frac{\partial}{\partial x^j}\Big)_{x^{-1}(a^1,\dots,a^n)}\Big),\dots,\dots,(\mathrm{d}y^i)_{x^{-1}(a^1,\dots,a^n)}\Big(b^j\Big(\frac{\partial}{\partial x^j}\Big)_{x^{-1}(a^1,\dots,a^n)}\Big)\dots\Big)\\ =\Big(\dots,(y^i\circ x^{-1})(a^1,\dots,a^n),\dots,\dots,b^j\Big(\frac{\partial y^i}{\partial x^j}\Big)_{x^{-1}(a^1,\dots,a^n)},\dots\Big).\end{aligned}\tag{244}$$

Note that $(y^i\circ x^{-1})(a^1,\dots,a^n)$ is just the i-th coordinate of chart transition map of $\mathcal{A}$ and

$$\Big(\frac{\partial y^i}{\partial x^j}\Big)_{x^{-1}(a^1,\dots,a^n)}=\partial_j(y^i\circ x^{-1})(x(x^{-1}(a^1,\dots,a^n)))=\partial_j(y^i\circ x^{-1})(a^1,\dots,a^n)).\tag{245}$$

Since $y\circ x^{-1}$ is smooth, it follows that $\xi_y\circ\xi_x^{-1}$ is also smooth. Thus $(TM,\mathcal{O}_{TM},\mathcal{A}_{TM})$ is a smooth manifold. Because π is a projection from TM to M, both of which are smooth manifolds, we conclude that π is a smooth map. Thus the triple (TM,M,π) has a bundle structure, which is referred as the tangent bundle.

Definition C.15 (Tangent bundle). For a smooth manifold $(M,\mathcal{O},\mathcal{A})$, the triple (TM,M,π) is called the *tangent bundle*. For simplicity, we sometimes call TM the tangent bundle.

Why do we care about tangent bundles? The reason is that any smooth vector field on a smooth manifold M can be represented as a smooth section of the tangent bundle (TM,M,π).

Definition C.16 (Section). Let (E,M,π) be a bundle. A *section* σ of the bundle is a map $\chi:M\to E$ such that $\pi\circ\chi=\mathrm{id}_M$, where id_M is the identity map on M.

Recall that $(C^\infty(M),\oplus,\odot)$ is not only a vector space, it is also a *ring*. Here we slightly abuse the notation by denoting this ring as $(C^\infty(M),+,\cdot)$. Define

$$\Gamma(TM)=\{\chi:M\to TM:\chi\text{ is a smooth section}\}.\tag{246}$$

The key idea is that *one can think of* $\Gamma(TM)$ *as a collection of smooth vector fields on* M. Note that $\chi:M\to TM$ such that $p\mapsto\chi(p)\in T_pM$. Thus given a smooth map $f\in C^\infty(M)$, we can view $\chi f:M\to\mathbb{R}$ defined via $p\mapsto\chi(p)f\in\mathbb{R}$. In words, a smooth vector field $\chi\in\Gamma(TM)$ on M acting on a smooth function $f\in C^\infty(M)$ gives another smooth function $\chi f\in C^\infty(M)$.

Now we want to give more algebraic structures to the set $\Gamma(TM)$. Let $\chi,\widetilde{\chi}\in\Gamma(TM)$ be two smooth vector fields on M and $f,g\in C^\infty(M)$. Define

$$(\chi\oplus\widetilde{\chi})(f)=\chi f+\widetilde{\chi}f,\tag{247}$$

$$(g\odot\chi)(f)=g\cdot\chi(f).\tag{248}$$

Then $(\Gamma(TM),\oplus,\odot)$ would be a vector space, if $C^\infty(M)$ is a field. However, $C^\infty(M)$ is only a ring, so this makes $(\Gamma(TM),\oplus,\odot)$ a $C^\infty(M)$-*module*. Informally, we call $\Gamma(TM)$ is "tangent field."

Similarly, one can consider the vector fields on the cotangent bundle T^*M and construct the section $\Gamma(T^*M)$ as a $C^\infty(M)$-module. In words, an element in $\Gamma(T^*M)$ takes a vector field

in $\Gamma(TM)$ and it produces a smooth function in $C^\infty(M)$. Informally, we also call $\Gamma(T^*M)$ "cotangent field."

Definition C.17 (Tensor field). An (r,s)-*tensor field* T over $\Gamma(TM)$ is a $C^\infty(M)$-multi-linear map

$$T : \underbrace{\Gamma(T^*M) \times \cdots \times \Gamma(T^*M)}_{r \text{ times}} \times \underbrace{\Gamma(TM) \times \cdots \times \Gamma(TM)}_{s \text{ times}} \xrightarrow{\sim} C^\infty(M). \tag{249}$$

By convention, a $(0,0)$-tensor field is a smooth function $C^\infty(M)$.

In the language of tensor field, an element of the cotangent field $\Gamma(T^*M)$ is a $(0,1)$-tensor field, which takes a vector field in $\Gamma(TM)$ and produces a smooth function in $C^\infty(M)$. Below we give such an example.

Example C.18 (Gradient tensor field as a cotangent field). The $C^\infty(M)$-linear map

$$\begin{aligned} \mathrm{d}f \quad &: \quad \Gamma(TM) \xrightarrow{\sim} C^\infty(M), \\ &\qquad \chi \mapsto \mathrm{d}f(\chi) := \chi f, \end{aligned} \tag{250}$$

where as before $(\chi f)(p) = \chi(p) f$ gives the directional derivative of f along the vector $\chi(p)$ at p, is the $(0,1)$-tensor field over the smooth vector fields $\Gamma(TM)$ on M. In other words, the gradient tensor field $\mathrm{d}f$ gives the gradients of f (on the whole manifold M) along all possible smooth vector fields χ, whereas we recall that the gradient $(\mathrm{d}f)_p$ gives the directional derivatives of f along all possible tangent vectors at a given point p.

C.5 Connections and curvatures

Let $X \in \Gamma(TM)$ be a smooth vector field on M. (Note that we slightly change the notation for a vector field from χ in Appendix C.4 to X.)

In this section, we wish to extend the notion of directional derivative $X : C^\infty(M) \to C^\infty(M)$ defined through $(Xf)(p) = X(p)f$ for $p \in M$ to the notion of connections on tensor fields (cf. Example C.12 and C.18), which allows us to define straight lines and eventually leads to curvatures.

Definition C.19 (Connection). A *connection* (sometimes also referred as *affine connection*) ∇ on a smooth manifold $(M, \mathcal{O}, \mathcal{A})$ is a map taking a pair of a vector (field) X on M and an (r,s)-tensor field T over $\Gamma(TM)$ and sending them to an (r,s)-tensor (field) $\nabla_X T$, satisfying:

1. $\nabla_X f = Xf$ for $f \in C^\infty(M)$ (i.e., f is a $(0,0)$-tensor);
2. Additivity (in T): for (r,s)-tensor fields T and S,
$$\nabla_X (T + S) = \nabla_X T + \nabla_X S; \tag{251}$$
3. Leibniz rule (in T): for $\omega \in \Gamma(T^*M), Y \in \Gamma(TM)$, and $(1,1)$-tensor field T,
$$\nabla_X (T(\omega, Y)) = (\nabla_X T)(\omega, Y) + T(\nabla_X \omega, Y) + T(\omega, \nabla_X Y); \tag{252}$$
4. $C^\infty(M)$-linearity (in X): for $f \in C^\infty(M)$,
$$\nabla_{fX+Z} T = f \nabla_X T + \nabla_Z T. \tag{253}$$

We say $\nabla_X T$ is the *covariant derivative* of T in the direction of X.

Remark C.20 (Remark on the Leibniz rule). First, the Leibniz rule (252) is equivalent to the tensor product form. Let T be a (p,q)-tensor field and S be an (r,s)-tensor field. The *tensor product* $T \otimes S$ of T and S is defined as the $(p+r, q+s)$-tensor field such that

$$\begin{aligned}&(T \otimes S)(\omega_1, \dots, \omega_{p+r}, X_1, \dots X_{q+s}) \\ &= T(\omega_1, \dots, \omega_p, X_1, \dots, X_q) \cdot S(\omega_{p+1}, \dots, \omega_{p+r}, X_{q+1}, \dots, X_{q+s}),\end{aligned} \tag{254}$$

for $\omega_1, \dots, \omega_{p+r} \in \Gamma(T^*M)$ and $X_1, \dots, X_{q+s} \in \Gamma(TM)$. Then the Leibniz rule (252) can be expressed in terms of the tensor product:

$$\nabla_X (T \otimes S) = (\nabla_X T) \otimes S + T \otimes (\nabla_X S). \tag{255}$$

Second, the Leibniz rule (252) can defined on higher-order tensors, which are useful to describe curvatures of smooth manifolds. For instance, for $(1,2)$-tensor T, the Leibniz rule in T becomes

$$\nabla_X (T(\omega, Y, Z)) = (\nabla_X T)(\omega, Y, Z) + T(\nabla_X \omega, Y, Z) + T(\omega, \nabla_X Y, Z) + + T(\omega, Y, \nabla_X Z). \tag{256}$$

■

Now we can equip a smooth manifold with an additional connection structure. A *smooth manifold with connection* is a quadruple of structures $(M, \mathcal{O}, \mathcal{A}, \nabla)$. Intuitively, one can think of ∇_X is an extension of the directional derivative X and ∇ is an extension of the differential d, where both extensions are seen from smooth functions to tensor fields.

How many ways can we determine the connection ∇ on a smooth manifold $(M, \mathcal{O}, \mathcal{A})$? It turns out the degree of freedom (without putting extra structures) is high. Actually there are infinitely many connections can be given on a smooth manifold.

To see this, let X and Y be vector fields on M. Take a chart (U, x) and note that $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n}$ are $(1,0)$-tensor fields (through an isomorphic identification). By the $C^\infty(M)$-linearity (253) and the Leibniz rule (255) (in the tensor product form), we may compute

$$\begin{aligned}\nabla_X Y &= \nabla_{X^i \frac{\partial}{\partial x^i}} \Big(Y^j \frac{\partial}{\partial x^j} \Big) = X^i \nabla_{\frac{\partial}{\partial x^i}} \Big(Y^j \frac{\partial}{\partial x^j} \Big) \\ &= X^i \Big(\nabla_{\frac{\partial}{\partial x^i}} Y^j \Big) \Big(\frac{\partial}{\partial x^j} \Big) + X^i Y^j \Big(\nabla_{\frac{\partial}{\partial x^i}} \Big(\frac{\partial}{\partial x^j} \Big) \Big) \\ &= X^i \Big(\frac{\partial Y^j}{\partial x^i} \Big) \frac{\partial}{\partial x^j} + X^i Y^j \underbrace{\nabla_{\frac{\partial}{\partial x^i}} \Big(\frac{\partial}{\partial x^j} \Big)}_{=: {\Gamma^k}_{ji} \frac{\partial}{\partial x^k}},\end{aligned} \tag{257}$$

where ${\Gamma^k}_{ji}$ are the (chart-dependent) *connection coefficient functions* (on M) of ∇ w.r.t. (U, x).

Definition C.21 (Christoffel symbols). Given a smooth manifold $(M, \mathcal{O}, \mathcal{A})$ and a chart $(U, x) \in \mathcal{A}$, the *Christoffel symbols* ${\Gamma^k}_{ji} := {\Gamma_{(x)}}^k{}_{ji}$ are the connection coefficient functions defining a connection on the smooth manifold via

$$\begin{aligned}{\Gamma^k}_{ji} \quad &: \quad U \to \mathbb{R}, \\ &\qquad p \mapsto {\Gamma^k}_{ji}(p) := \Big(\mathrm{d}x^k \Big(\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} \Big) \Big)(p).\end{aligned} \tag{258}$$

In one chart (U, x), the Christoffel symbols, we can write the vector field in (257) as

$$\nabla_X Y = X^i\Big(\frac{\partial Y^j}{\partial x^i}\Big)\frac{\partial}{\partial x^j} + X^i Y^j \Gamma^k{}_{ji}\frac{\partial}{\partial x^k} = X^i\Big[\Big(\frac{\partial Y^j}{\partial x^i}\Big)\frac{\partial}{\partial x^j} + Y^j\Gamma^k{}_{ji}\frac{\partial}{\partial x^k}\Big], \tag{259}$$

or we can write down its components

$$(\nabla_X Y)^m = X(Y^m) + \Gamma^m{}_{ji}\, Y^j\, X^i, \tag{260}$$

where Y^m is the m-th component of Y and the products in this expression are all pointwise products of C^∞ functions. By the duality between basis and Leibniz rule, one can show that

$$(\nabla_X \omega)_m = X(\omega_m) - \Gamma^j{}_{mi}\,\omega_j\, X^i. \tag{261}$$

Based on (260) and (261), given an n-dimensional smooth manifold $(M, \mathcal{O}, \mathcal{A})$, we can determine a connection ∇ from the n^3-many Christoffel symbols $\Gamma^k{}_{ji}$, and we are left with a huge freedom for choosing $\Gamma^k{}_{ji}$ in order to determine a manifold with smooth connection $(M, \mathcal{O}, \mathcal{A}, \nabla)$.

A first step to nail down the number of connections is to require the torsion-free structure.

Definition C.22 (Torsion). Let $(M, \mathcal{O}, \mathcal{A}, \nabla)$ be a smooth manifold with connection. The *torsion* of ∇ is the $(1,2)$-tensor field

$$T(\omega, X, Y) = \omega(\nabla_X Y - \nabla_Y X - [X, Y]), \tag{262}$$

where the Lie bracket $[X, Y]$ is the vector field defined by

$$[X, Y]f = X(Yf) - Y(Xf), \quad f \in C^\infty(M). \tag{263}$$

The manifold $(M, \mathcal{O}, \mathcal{A}, \nabla)$ (or simply the connection ∇) is said to be *torsion-free* if $T \equiv 0$.

It is easy to check that torsion $T(\omega, X, Y)$ defined in (262) is indeed a $C^\infty(M)$-multi-linear map. Note that, for chart-induced basis,

$$[\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}] = 0, \tag{264}$$

so that the components of torsion are given by $T^k{}_{ij} = \Gamma^k{}_{ji} - \Gamma^k{}_{ij}$. Thus, on a chart-induced basis, the connection ∇ is torsion-free if the Christoffel symbols are symmetric $\Gamma^k{}_{ji} = \Gamma^k{}_{ij}$.

In order to define a curvature, it is first instructive to think about how the straight lines look like in a curved smooth manifold. It is intuitively clear to speak about straight lines in Euclidean spaces. This leads to the notion of *autoparallely transported curves*.

Definition C.23 (Paralle transport). A vector field X on M is said to be *parallelly transported* along a smooth curve $\gamma : \mathbb{R} \to M$ if

$$\nabla_{v_\gamma} X = 0, \tag{265}$$

where v_γ denotes the velocity vector field along γ.

Definition C.24 (Autoparallely transported curve). A smooth curve $\gamma : \mathbb{R} \to M$ is said to be *autoparallely transported* if

$$\nabla_{v_\gamma} v_\gamma = 0, \tag{266}$$

where v_γ again denotes the velocity vector field along γ.

In words, the velocity vector field along an autoparallely transported curve is constant along the curve. We can view such curves have constant velocity, which mimics the "no-acceleration" situation of straight lines in $\mathbb{R}^n$.

Definition C.25 (Riemann curvature). The *Riemann curvature* of a connection ∇ is the $(1,3)$-tensor field

$$\mathcal{R}(\omega, Z, X, Y) = \omega(\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z). \tag{267}$$

It is also easy to check that the Riemann curvature $\mathcal{R}(\omega, Z, X, Y)$ defined in (267) is indeed a $C^\infty(M)$-multi-linear map. In a chart-induced basis $\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n}$, components of the Riemann curvature tensor can be expressed in terms of Christoffel symbols:

$$\begin{aligned} \mathcal{R}^a{}_{bcd} &= \mathcal{R}\Big(\mathrm{d}x^a, \frac{\partial}{\partial x^b}, \frac{\partial}{\partial x^c}, \frac{\partial}{\partial x^d}\Big) \\ &= \frac{\partial}{\partial x^c}\Gamma^a{}_{db} - \frac{\partial}{\partial x^d}\Gamma^a{}_{cb} + \Gamma^a{}_{cs}\Gamma^s{}_{db} - \Gamma^a{}_{ds}\Gamma^s{}_{cb}. \end{aligned} \tag{268}$$

Remark C.26 (Euclidean space with Cartesian coordinates). Consider the Euclidean space as a smooth manifold $(\mathbb{R}^n, \mathcal{O}, \mathcal{A})$ equipped with the standard topology (i.e., topology generated by open subsets in $\mathbb{R}^n$) and a smooth atlas $\mathcal{A}$. Under different atlas, we may have a Euclidean space with different coordinate systems such as Cartesian and polar coordinate systems. By default, if we assume the chart $(\mathbb{R}^n, \mathrm{id}_{\mathbb{R}^n}) \in \mathcal{A}$ and for such chart,

$$\Gamma^k{}_{ji} = \Big(\mathrm{d}x^k\Big(\nabla^e_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\Big)\Big) = 0, \tag{269}$$

where $\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n}$ is the global Cartesian basis, then we call ∇^e the Euclidean connection and the resulting the manifold with connection $(\mathbb{R}^n, \mathcal{O}, \mathcal{A}, \nabla^e)$ the n-dimensional Euclidean space. Usually, we suppress the superscript e and simply write $\nabla = \nabla^e$.

C.6 Riemannian manifolds and geodesics

Now we consider a smooth manifold with a metric structure and construct a connection on such metric manifold. On a metric manifold, we can speak of metric quantities such as speed or length of curves. It turns out, with an additional metric-compatibility requirement (plus the torsion-free requirement), one can uniquely determine a connection on the metric manifold such that the straight lines are the same as length-minimizing curves.

Let $g : \Gamma(TM) \times \Gamma(TM) \to C^\infty(M)$ be a symmetric $(0,2)$-tensor field (i.e., $g(X,Y) = g(Y,X)$ for all vector fields X and Y in $\Gamma(TM)$). We first define a bundle isomorphism called the *musical map*

$$\begin{aligned} \flat \;:\; & \Gamma(TM) \to \Gamma(T^*M), \\ & X \mapsto \flat(X) \text{ such that } \flat(X)(Y) = g(X,Y). \end{aligned} \tag{270}$$

By construction, the musical map $\flat := \flat_g$ depends on the metric tensor g.

For a vector field $X \in \Gamma(TM)$, $\flat(X) \in \Gamma(T^*M)$ is a $(0,1)$-tensor with components given by

$$\begin{aligned} (\flat(X))_a &= \flat(X)\Big(\frac{\partial}{\partial x^a}\Big) = g\Big(X, \frac{\partial}{\partial x^a}\Big) \\ &= g\Big(X^m \frac{\partial}{\partial x^m}, \frac{\partial}{\partial x^a}\Big) = X^m g\Big(\frac{\partial}{\partial x^m}, \frac{\partial}{\partial x^a}\Big) = X^m g_{am}. \end{aligned} \tag{271}$$

where we used $C^\infty(M)$-multi-linearity of g.

Definition C.27 (Metric). A (non-degenerate) metric g on a smooth manifold $(M, \mathcal{O}, \mathcal{A})$ is a symmetric $(0,2)$-tensor field such that the musical map $\flat$ is a C^∞-isomorphism (i.e., $\flat$ is invertible) between the tangent field $\Gamma(TM)$ and the cotangent field $\Gamma(T^*M)$.

Given a (symmetric and non-degenerate) metric g, one can define its inverse g^{-1} as the symmetric $(2,0)$-tensor field by

$$\begin{aligned} g^{-1} \quad : \quad & \Gamma(T^*M) \times \Gamma(T^*M) \to C^\infty(M), \\ & (\omega, \sigma) \mapsto \omega(\flat^{-1}(\sigma)). \end{aligned} \tag{272}$$

For a covector field $\omega \in \Gamma(T^*M)$, $\flat^{-1}(\omega) \in \Gamma(TM)$ is a $(1,0)$-tensor as a vector field (via an isomorphic identification) with components given by

$$\begin{aligned} (\flat^{-1}(\omega))^a = \mathrm{d}x^a(\flat^{-1}(\omega)) = g^{-1}(\mathrm{d}x^a, \omega) \\ = g^{-1}(\mathrm{d}x^a, \omega_m \mathrm{d}x^m) = \omega_m g^{-1}(\mathrm{d}x^a, \mathrm{d}x^m) = \omega_m g^{am}. \end{aligned} \tag{273}$$

where we used $C^\infty(M)$-multi-linearity of g^{-1} and we denote g^{am} as the components of g^{-1}.

Informally, a $(1,1)$-tensor field can be represented by a symmetric matrix (say in a chart), and we know that a symmetric matrix has an eigendecomposition with real eigenvalues. For a general (r,s)-metric tensor field g, it does not have an eigendecomposition. Rather, it only has a *signature*.

Definition C.28 (Riemannian metric). A metric is called *Riemannian* (or sometimes called *positive-definite*) if its signature is $(+, \ldots, +)$. Metrics with all other signatures are called *pseudo-Riemannian* metric.

Definition C.29 (Riemannian manifold). A Riemannian manifold $(M, \mathcal{O}, \mathcal{A}, g)$ is a smooth manifold $(M, \mathcal{O}, \mathcal{A})$ equipped with a Riemannian metric tensor g.

Because a Riemannian metric is positive-definite, it defines an inner product structure on the tangent field $\Gamma(TM)$. In particular, on a Riemannian metric manifold $(M, \mathcal{O}, \mathcal{A}, g)$, the *speed* $s(t)$ of a curve at $\gamma(t)$ is defined as

$$s(t) = \sqrt{\Big(g(v_\gamma, v_\gamma)\Big)_{\gamma(t)}}, \tag{274}$$

where v_γ is the velocity vector field along the curve (i.e., $v_{\gamma,t}$ is the velocity of curve γ at t).

Definition C.30 (Length). Let $\gamma : [0,1] \to M$ be a smooth curve. Then the *length of* γ is defined as

$$L(\gamma) = \int_0^1 s(t)\, \mathrm{d}t = \int_0^1 \sqrt{\Big(g(v_\gamma, v_\gamma)\Big)_{\gamma(t)}}\, \mathrm{d}t \tag{275}$$

Lemma C.31 (Length is preserved under reparametrization). Let $\gamma : [0,1] \to M$ be a smooth curve and $\sigma : [0,1] \to [0,1]$ be a smooth, bijective, and increasing function. Then

$$L(\gamma) = L(\gamma \circ \sigma). \tag{276}$$

Definition C.32 (Geodesic). A curve $\gamma : [0,1] \to M$ is called a *geodesic* on a Riemannian manifold $(M, \mathcal{O}, \mathcal{A}, g)$ if it is a *stationary* curve w.r.t. the length functional L in (275).

Recall that the signature of a Riemannian manifold is $(+,\dots,+)$. Thus, given two end points on M, a geodesic corresponding to the stationary point of L is the length-minimizing curve on the manifold. Nevertheless, we should note that geodesic does not always exist: consider two points $(+1,+1)$ and $(-1,-1)$ in $\mathbb{R}^2\setminus\{(0,0)\}$ with the Euclidean metric on $\mathbb{R}^2$. This problem can be fixed by considering the *admissible curve*, which is a piecewise smooth curve segment. Setting

$$d(p,q)=\inf\big\{L(\gamma):\gamma:[0,1]\to M \text{ is admissible},\gamma(0)=p,\gamma(1)=q\big\},\tag{277}$$

then (M,d) is a metric space [7, Chapter 2] (sometimes referred as a *length space*). If (M,d) is a complete metric space, then it is a *geodesically complete manifold/space* or simply *geodesic space*.

Theorem C.33 (Levi-Civita connection). On a Riemannian manifold $(M,\mathcal{O},\mathcal{A},g)$, there is a unique connection ∇ that is torsion-free $T=0$ and metric-compatible $\nabla g=0$. This connection is called the *Levi-Civita connection* (or sometimes called the *Riemiannian* connection).

For any vector fields X,Y,Z in the tangent bundle TM, the Leibniz rule (252) reads

$$\nabla_X g(Y,Z)=(\nabla_X g)(Y,Z)+g(\nabla_X Y,Z)+g(Y,\nabla_X Z).\tag{278}$$

Since $g(Y,Z)\in C^\infty(M)$ and $\nabla_X g=0$, metric compatibility can be equivalently expressed as

$$X(g(Y,Z))=g(\nabla_X Y,Z)+g(Y,\nabla_X Z).\tag{279}$$

From now on, on a Riemannian manifold $(M,\mathcal{O},\mathcal{A},g)$, the connection endowed will always be the Levi-Civita connection (without mentioning further). In such case, the metric compatibility implies that a straight line between two points on the manifold is the same as a geodesic that minimizes the length functional L in (275). That is, once we work on a Riemannian manifold, the Riemman curvature tensor will be (implicitly) induced from the Levi-Civita connection. However, we should highlight that the Riemann curvature tensor can be defined through any connection without referring to the Riemannian metric.

Lemma C.34 (Existence and uniqueness of geodesics). Let $(M,\mathcal{O},\mathcal{A},g)$ be a Riemannian manifold. For every point $p\in M$ and $v\in T_pM$, there is a unique maximal geodesic $\gamma:=\gamma_v:I\to M$ with $\gamma(0)=p$ and $\gamma'(p)=v$, defined on some open interval I containing 0.

Proof of Lemma C.34 can be found in Theorem 4.27 and Corollary 4.28 in [7].

Definition C.35 (Exponential map). Let $(M,\mathcal{O},\mathcal{A},g)$ be a Riemannian manifold and $p\in M$. The *exponential map* $\exp_p:T_pM\to M$ is defined by

$$\exp_p(v)=\gamma_v(1),\tag{280}$$

where γ_v is the unique maximal geodesic defined on an open interval containing $[0,1]$ (cf. Lemma C.34).

Note that for a smooth manifold $(M,\mathcal{O},\mathcal{A})$, without additional connection ∇ or metric g structures, we cannot speak of the manifold shape. Given two end points, the straight line determined by the autoparallely transported curves coincides to the geodesic determined by the metric g if coefficient functions of the Levi-Civita connection satisfy that

$$\Gamma^i{}_{jk}=\frac{1}{2}g^{iq}\Big(\frac{\partial g_{kq}}{\partial x^j}+\frac{\partial g_{jq}}{\partial x^k}-\frac{\partial g_{jk}}{\partial x^q}\Big),\tag{281}$$

which comes from the *geodesic equation*.

Example C.36 (Round metric on 2-sphere). Consider the 2-sphere manifold $(\mathbb{S}^2, \mathcal{O}, \mathcal{A})$ and a chart map $x(p) := (x^1(p), x^2(p)) = (\theta, \varphi)$ such that $\theta \in (0, \pi)$ and $\varphi \in (0, 2\pi)$. We now determine the shape of the smooth manifold 2-sphere by the *round metric*. Under the chart map x, we may equip the 2-sphere smooth manifold $(\mathbb{S}^2, \mathcal{O}, \mathcal{A})$ with the round metric whose components are given by

$$\big(g_{ij}(x^{-1}(\theta, \varphi))\big)_{i,j=1,2} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2(\theta) \end{pmatrix}. \tag{282}$$

It is straightforward to check that the metric g in (282) corresponds to a connection ∇^{round} with the connection coefficients in (281) given by

$$\Gamma^1{}_{22}(x^{-1}(\theta, \varphi)) = -\sin(\theta)\cos(\theta), \quad \Gamma^1{}_{12} = \Gamma^1{}_{21} = \cot(\theta), \tag{283}$$

and all other 5 Christoffel symbols are all zeros. Consider the equator curve γ of the round 2-sphere parameterized by

$$\theta(t) := (\theta \circ \gamma)(t) = (x^1 \circ \gamma)(t) = \frac{\pi}{2}, \tag{284}$$

$$\varphi(t) := (\varphi \circ \gamma)(t) = (x^2 \circ \gamma)(t) = 2\pi t^3. \tag{285}$$

Obviously, $\theta'(t) = 0$ and $\varphi'(t) = 6\pi t^2$. Then the length of γ can be computed by using the components of the round metric tensor

$$\begin{aligned} L(\gamma) &= \int_0^1 \sqrt{g_{ij}(x^{-1}(\theta(t), \varphi(t)))(x^i \circ \gamma)'(t)(x^j \circ \gamma)'(t)}\, \mathrm{d}t \\ &= \int_0^1 \sqrt{1 \cdot 0 + \sin^2(\pi/2) \cdot 36\pi^2 t^4}\, \mathrm{d}t \qquad (286) \\ &= 6\pi \int_0^1 t^2\, \mathrm{d}t = 2\pi, \qquad (287) \end{aligned}$$

which is the same as the length of the equator curve under the constant speed parametrizaiton $\varphi(t) = 2\pi t$. In the special case of round sphere, this calculation verifies the length maintenance under reparametrization stated in Lemma C.31. ■

From the metric tensor g, we can define more curvature tensors on a Riemannian manifold via *contraction*.

Definition C.37 (Curvature tensors). Let $(M, \mathcal{O}, \mathcal{A}, g)$ be a Riemannian manifold.

1. The *Riemann-Christoffel curvature* is a $(0, 4)$-tensor defined by
$$R_{abcd} = g_{am}\, \mathcal{R}^m{}_{bcd}. \tag{288}$$
2. The *Ricci curvature* is a $(0, 2)$-tensor defined by
$$R_{ab} = \mathcal{R}^m{}_{amb}. \tag{289}$$
3. The *scalar curvature* is defined by
$$R = (g^{-1})^{ab}\, R_{ab}. \tag{290}$$
4. The *Einstein curvature* is a $(0, 2)$-tensor defined by
$$G_{ab} = R_{ab} - \frac{1}{2} R\, g_{ab}. \tag{291}$$

C.7 Volume forms

Consider the problem of integrating functions over a smooth manifold $(M, \mathcal{O}, \mathcal{A})$.

Definition C.38 (Volume form). On an n-dimensional smooth manifold $(M, \mathcal{O}, \mathcal{A})$, a $(0, n)$-tensor field Ω (over $\Gamma(TM)$) is called a *volume form* if

1. Non-vanishing: $\Omega \neq 0$ on M;
2. Totally anti-symmetric: $\Omega(\dots, X_i, \dots, X_j, \dots) = -\Omega(\dots, X_j, \dots, X_i, \dots)$ for all $X_i, X_j \in \Gamma(TM)$ and $i, j = 1, \dots, n$.

We can construct a (natural) volume form from a Riemannian manifold. Let $(M, \mathcal{O}, \mathcal{A}, g)$ be a Riemannian manifold. Take an arbitrary chart (U, x) and we let the components of the tensor field Ω be given by

$$\Omega_{i_1,\dots,i_n} = \sqrt{g_{ij}}\, \epsilon_{i_1,\dots,i_n}, \tag{292}$$

where $(i_1, \dots, i_n)$ is a permutation of $(1, \dots, n)$ such that $\epsilon_{1,\dots,n} = 1$ and $\epsilon_{i_1,\dots,i_n} = \epsilon_{[i_1,\dots,i_n]}$ for any anti-symmetric bracket $[\dots]$. The $\epsilon_{i_1,\dots,i_n}$'s are called the *Levi-Civita symbols*. One can check that $\Omega_{i_1,\dots,i_n}$ is well-defined if and only if for any pair of charts (U, x) and (U, y),

$$\det\left(\frac{\partial y}{\partial x}\right) = \det(\partial_i(y^j \circ x^{-1})) > 0. \tag{293}$$

Condition (293) means that the chart transition maps are *oriented*.

Definition C.39 (Oriented atlas). Let $\mathcal{A}$ be a smooth atlas. Then $\mathcal{A}^\uparrow \subset \mathcal{A}$ is called a (positively) *oriented sub-atlas* of $\mathcal{A}$ if for any two charts $(U, x), (V, y) \in \mathcal{A}^\uparrow$, the chart transition maps $y \circ x^{-1}$ and $x \circ y^{-1}$ are oriented:

$$\det\left(\frac{\partial y}{\partial x}\right) > 0 \quad \text{or} \quad \det\left(\frac{\partial x}{\partial y}\right) > 0 \quad \text{on } U \cap V. \tag{294}$$

Let $(M, \mathcal{O}, \mathcal{A}^\uparrow)$ be a smooth oriented manifold and $(U, x) \in \mathcal{A}^\uparrow$ be an oriented chart. Given a volume form Ω, we can define a *scalar density* on M by

$$\omega(p) = \Omega_{i_1,\dots,i_n}(p)\, \epsilon^{i_1,\dots,i_n} \quad \text{for } p \in M, \tag{295}$$

where $\epsilon^{i_1,\dots,i_n} = \epsilon_{i_1,\dots,i_n}$. Note that $\Omega_{i_1,\dots,i_n}$ is chart-dependent, so is ω. Moreover, the change of scalar density under a change of chart is given by

$$\omega_{(y)} = \det\left(\frac{\partial x}{\partial y}\right) \omega_{(x)} \tag{296}$$

for any two charts $(U, x), (U, y) \in \mathcal{A}^\uparrow$.

Let $(M, \mathcal{O}, \mathcal{A}^\uparrow, g)$ be an oriented metric manifold. On one chart domain (U, x), we can define the integration of a function $f : U \to \mathbb{R}$ as

$$\int_U f := \int_{x(U)} \sqrt{\det(g_{ij})(x^{-1}(\alpha))}(f \circ x^{-1})(\alpha)\, \mathrm{d}\alpha. \tag{297}$$

One can check that (297) is well-defined and does not depend on the chart map on the same chart domain U. To extend the integration to the whole manifold, we would then need

partition of unity. Let $(U_i, x_i) \in \mathcal{A}^\uparrow$ and $\varrho_i : U_i \to \mathbb{R}, i = 1, \dots, N$, be a finite collection of continuous functions such that for any $p \in M$, we have $\sum_i \varrho_i(p) = 1$ where the sum is taken over i such that $p \in U_i$. Then we define the integration of a function $f : M \to \mathbb{R}$ as

$$\int_M f = \sum_{i=1}^N \int_{U_i} (\varrho_i f). \tag{298}$$

Combining (297) and (298), we can define the (natural) volume form of a Riemannian manifold as following.

Definition C.40 (Volume form of oriented Riemannian manifold). Let $(M, \mathcal{O}, \mathcal{A}^\uparrow, g)$ be an oriented Riemmanian manifold and $(U, x) \in \mathcal{A}^\uparrow$ be an oriented chart. The *volume form* of the manifold M in an oriented chart (i.e., local coordinates) is defined as

$$\mathrm{dVol} = \sqrt{\det(g)}\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n, \tag{299}$$

where $\wedge$ denotes the wedge product. The integration of a function $f : M \to \mathbb{R}$ is defined as

$$\int_M f := \int_M f(p)\, \mathrm{dVol}(p). \tag{300}$$

Acknowledgement

This lecture note was carried out when the author was visiting the Institute for Data, System, and Society (IDSS) at Massachusetts Institute of Technology. The author would like to thank Sinho Chewi, Tobias Colding, Thibaut Le Gouic, Philippe Rigollet, Austin Stromme, Yun Yang for their helpful comments and feedbacks. This work is supported in part by an NSF CAREER award DMS-1752614 and a Simons Fellowship in Mathematics.

References

[1] David Alonso-Gutiérrez and Jesús Bastero. *Approaching the Kannan-Lovász-Simonovits and Variance Conjectures.* Springer, Cham, 2015.

[2] Francois Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in wasserstein distance for fokker-planck equations. *Journal of Functional Analysis*, 263(8):2430 – 2457, 2012.

[3] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of Isotropic Convex Bodies*, volume 196 of *Mathematical Surveys and Monographs.* American Mathematical Society, 2014.

[4] Xiaohui Chen and Yun Yang. Diffusion k-means clustering on manifolds: provable exact recovery via semidefinite relaxations. *Applied and Computational Harmonic Analysis (arXiv:1903.04416)*, 2020.

[5] Ivan Gentil, Christian Léonard, Luigia Ripani, and Luca Tamanini. An entropic interpolation proof of the hwi inequality. *Stochastic Processes and their Applications*, 2019.

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

[7] John Lee. *Introduction to Riemannian Manifolds*. Graduate Texts in Mathematics. Springer International Publishing, 2 edition, 2018.

[8] Y. T. Lee and S. S. Vempala. Eldan's stochastic localization and the kls hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007, Oct 2017.

[9] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

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

[11] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Now Foundations and Trends, 2019.

[12] Steven Rosenberg. *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*. London Mathematical Society Student Texts. Cambridge University Press, 1997.

[13] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians - Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, 2015.

[14] Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[15] Ramon van Handel. Probability in high dimension. APC 550 Lecture Notes, Princeton University, 2016.