Introduction to the Otto Calculus

MAT1855 Course Project

Anton Sugolov

November 29, 2024

Contents

1 Introduction

In a series of seminal papers, Felix C. Otto introduced a geometric perspective on the normalized solutions of certain PDEs as gradient flows on the space of probability densities[6, 8]. Given an energy functional, the Otto calculus produces a PDE such that the solutions can be interpreted as gradient flows of the functional in the appropriate sense. Rather strikingly, the heat equation can be derived from the Shannon-Boltzmann entropy using this construction. In this project, we survey the necessary prerequisites to Otto's derivation and present Otto's initial results. Afterwards, we provide the derivation of the Fokker-Planck equation through this lens, and discuss the applications for stochastic processes and sampling algorithms.

2 Prerequisites

We first develop the prerequisites to understand the contributions of Otto. First, we describe the role of the continuity equation and its relationship to the Eulerian transport formulation, define the Wasserstein metric W_p, the space of L^p probability densities, and describe the Benamou-Brenier representation for the distance W_p.

2.1 Continuity equation

This fluid mechanics settings for the L^2 Monge-Kantorovich transport problem was originally discussed in the work of Benamou and Brenier [3] "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem", and motivated the dynamical perspective through the continuity equation. The continuity equation from physics describes the evolution of a quantity whose total mass is preserved

$$\frac{d}{dt}\rho_t + \operatorname{div}(\rho_t v_t) = 0\,.$$

Here, $\rho_t(x)$ is a measure that evolves in t by a time dependent vector field $v_t(x)$ such that the total mass $\int \rho_t = 1$ is conserved under flow by by v_t. The continuity equation is suited for transport in the the Eulerian formalism, in which the model is described through its density and velocity in time. For an initial ρ_0, we find the solution of the ODE given by

$$\begin{cases} y_x'(t) &= v_t(y_x(t)) \\ y_x(0) &= x \end{cases}$$

Evolving individual points with $Y_t(x) = y_x(t)$, the measure at time t is given by $\rho_t = (Y_t)_{\#}\rho_0$. For this formalism to preserve mass, ρ_t and v_t must together solve the continuity equation.

Definition 2.1.1. Consider $\Omega \subset \mathbb{R}^d$ is a bounded domain or $\Omega = \mathbb{R}^d$. A family of pairs of measures and vector fields (ρ_t, v_t) for $t \in [0,1]$ satisfying $v_t \in L^1(\rho_t; \mathbb{R}^d)$ and $\int_0^T \|v_t\|_{L_1(\rho_t)}\, dt < \infty$ solves the continuity equation in the **distributional** sense if for every $\phi \in C_c^1([0,1] \times \overline{\Omega})$

$$\int_0^T \int_\Omega \frac{d}{dt}\phi d\rho_t dt + \int_0^T \int_\Omega \nabla\phi \cdot v_t \, d\rho_t dt = 0$$

and in the **weak** sense if for every $\psi \in C_c^1(\overline{\Omega})$, $t \mapsto \int \psi \, d\rho_t$ is abs. continuous in t, and for a.e. t,

$$\frac{d}{dt}\int_\Omega \psi \, d\rho_t = \int_\Omega \nabla\psi \cdot v_t \, d\rho_t\,.$$

2.2 Wasserstein metric W_p and the space $\mathcal{P}_p(\mathbb{R}^d)$

For the sake of summary, we focus on the case of $\mathbb{R}^d$. All of the following stated results generalize to a Polish space X.

Consider the transport problem for probability measures on Ω with cost $c_p(x, y) = |x - y|^p$ for $p \in [1, \infty)$. We restrict our attention to the space of measures where the cost c_p is finite, which are those with finite L^p norm.

$$\mathcal{P}_p(\Omega) = \left\{ \mu \in \mathcal{P}(\Omega) \,\middle|\, \int_\Omega |x|^p \, d\mu < \infty \right\}.$$

Note that $p < q \implies \mathcal{P}_p(\Omega) \subset \mathcal{P}_q(\Omega)$. The **Wasserstein distance** W_p defines a metric on $\mathcal{P}_p(\Omega)$ associated with the minimal transport cost with c_p.

$$W_p(\mu, \nu) = \min \left\{ \int_{\Omega \times \Omega} |x - y|^p \, d\gamma \,\middle|\, \gamma \in \Pi(\mu, \nu) \right\}^{1/p}.$$

It can be shown that the Wasserstein distance $W_p(\mu, \nu)$ is indeed a metric on $\mathcal{P}_p(\Omega)$.

Definition 2.2.1. The **Wasserstein space** of order $p \in [1, \infty)$, is the metric space

$$\mathbb{W}_p(\Omega) = (\mathcal{P}_p(\Omega), W_p).$$

The metric W_p induces a certain topology on $\mathbb{W}_p(\Omega)$. A natural question is how it is related to the topology induced by weak convergence. Convergence in the Wasserstein metric turns out to be equivalent to weak convergence if Ω is compact. The general case is summarized by the following theorem.

Theorem 1. In the space $\mathbb{W}_p(\Omega)$,

$$W_p(\mu_n, \mu) \to 0 \iff \mu_n \to \mu \text{ weakly}, \int |x|^p \, d\mu_n \to \int |x|^p \, d\mu.$$

2.3 Absolute continuity in $\mathbb{W}_p$

In order to build towards the gradient flows of Otto calculus, we must first discuss properties of curves in $\mathbb{W}_p(\Omega)$. A **curve** in $\mathbb{W}_p(\Omega)$ is defined as a mapping $\mu_t : [0, 1] \to \mathcal{P}_p(\Omega)$. The curve μ_t is **absolutely continuous** if there exists $g \in L^1([0, 1])$ such that

$$W_p(\mu_{t_0}, \mu_{t_1}) \leq \int_{t_0}^{t_1} g(s) \, ds$$

for every $0 \leq t_0 < t_1 \leq 1$. The **metric derivative** of the curve $t \mapsto \mu_t$ at t is defined as

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

The absolute continuity of μ_t is equivalent to the existence of a vector field v_t so that $(\mu_t, v_t)_t$ solve the continuity equation of Section 2.1. The equivalence of the two conditions is formally presented in the following theorem in generality, and is originally due to Ambrosio [1].

Theorem 2. Let $(\mu_t)_{t\in[0,1]}$ be an absolutely continuous curve in $\mathbb{W}_p(\Omega)$ for $p > 1$ and compact $\Omega \subset \mathbb{R}^d$. For almost every $t \in [0,1]$, there exists a vector field $v_t \in L^p(\mu_t, \mathbb{R}^d)$ such that

- (μ_t, v_t), $t \in [0,1]$ satisfy $\frac{d}{dt}\mu_t + \operatorname{div}(v_t\mu_t) = 0$ in the weak sense and
- for almost every t, $\|v_t\|_{L^p(\mu_t)} \leq |\mu'|(t)$.

Conversely, if $(\mu_t)_{t\in[0,1]}$ is a family of measures in $\mathcal{P}_p(\Omega)$, and for each t there is $v_t \in L^p(\mu_t, \mathbb{R}^d)$ with $\int_0^1 \|v_t\|_{L^p(\mu_t)}\, dt < \infty$ satisfying $\frac{d}{dt}\mu_t + \operatorname{div}(v_t\mu_t) = 0$, then μ_t is absolutely continuous and $\|v_t\|_{L^p(\mu_t)} \leq |\mu'|(t)$.

2.4 Benamou-Brenier formula

Consider the case $p = 2$ for the purpose of demonstration. In Section 2.1 we discussed the Eulerian transport perspective, and the evolution of a measure by a vector field, which satisfies the continuity equation. The **quadratic action** for a measure $\mu \in \mathcal{P}_2(\mathbb{R}^n)$ and a measurable vector field $v : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$\mathcal{A}(v, \mu) = \int_{\mathbb{R}^n} \|v\|^2 \, d\mu = \|v\|_{L^2(\mu)} \, .$$

The **Benamou-Brenier Formula** is a representation theorem, showing for $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^n)$, the curves μ_t minimizing the transport cost $W_2(\mu_0, \mu_1)$ must be solutions to the continuity equation $(\mu_t, v_t)_t$:

$$W_2^2(\mu_0, \mu_1) = \min \left\{ \int_0^1 \|v_t\|_{L^2(\mu_t)} \, dt \text{ s.t. } \frac{d}{dt}\mu_t + \operatorname{div}(v_t\mu_t) = 0 \text{ in } (0,1) \times \mathbb{R}^n \right\} . \tag{1}$$

The representation formula generalizes for W_p, requiring the definition of another action functional, however the idea of the representation is the same. This formulation through the principle of least action is reminiscent of geodesics from Riemannian geometry: on a Riemannian manifold (M, g), the geodesics $\gamma(t)$ with $t \in [0,1]$ minimize the energy functional

$$E(\gamma) = \frac{1}{2} \int_0^1 g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t)) \, dt \, .$$

This hints at the special geometric structure of $\mathcal{P}_2(\Omega)$ induced by the Wasserstein metric, and provides a motivation for some of the methods of Otto calculus.

3 Otto Calculus

We first review Otto's original construction of the gradient flow of an energy functional for the porous medium equation [8], summarize the relationship to the previous theory, give a new perspective on the heat equation, and present some interesting applications to stochastic processes and statistical learning.

3.1 The Geometry of the Porous Medium Equation

In the work "The Geometry of Dissipative Evolution Equations: The Porous Medium Equation", Otto [8] describes a way to derive the weak solution of the porous medium equation as the gradient flow of a particular energy functional. Prior to this work, Newman [7] showed the existence of a Lyapunov function for the equation, while Otto's construction extends to general energy functionals. The key difference was the coupling of the tangent space with solutions to the continuity equation, hinting at the connection with absolute continuity of curves in $\mathbb{W}_p$.

The porous medium equation for $m \geq 1$ is given by

$$\frac{\partial \rho}{\partial t} - \Delta \rho^m = 0\,. \tag{2}$$

The aim is to derive this equation in terms of an evolution of ρ given by the gradient flow of an energy functional. Given a Riemannian manifold (M, g) and functional E on M, the dynamical system given by

$$\frac{\partial \rho}{\partial t} = -\nabla E(\rho)$$

is called the gradient flow on M generated by E. Crucially, we can identify tangent vectors $s \in T_\rho M$ with their co-tangent vectors through the metric tensor g,

$$g(\nabla E, s) = dE\,(s)\,.$$

Therefore the gradient flow can be represented dually through

$$g_\rho\left(\frac{\partial \rho}{\partial t}, s\right) + d_\rho\, E\,(s) = 0\,. \tag{3}$$

We work to extend this to the case of functions $\rho : \mathbb{R}^d \to \mathbb{R}$ by considering

$$M = \left\{\rho \,\middle|\, \rho \geq 0,\ \int \rho = 1\right\} \quad T_\rho M = \left\{s \,\middle|\, \int s = 0\right\}\,.$$

We define the metric tensor g_ρ. First, identify tangent space of functions s with functions p coupled with ρ heuristically through the continuity equation:

$$T_\rho M \cong \left\{p : \mathbb{R}^d \to \mathbb{R} \,\middle|\, -\operatorname{div}(\rho \nabla p) = s\right\}\,.$$

With this identification, define

$$g_\rho(s_1, s_2) = \int \rho\, \nabla p_1 \cdot \nabla p_2 = \int s_1 p_2 \tag{4}$$

where the second equality holds after integrating by parts. The functional E associated with (2) is defined as

$$E(\rho) = \begin{cases} \frac{1}{m-1}\int \rho^m & m \neq 1 \\ \int \rho \ln \rho & m = 1 \end{cases}\,.$$

Through the representation given in (4), we compute

$$d_\rho E(s) = g_\rho(\nabla E, s) = \begin{cases} \frac{m}{m-1}\int \rho^{m-1} s & m \neq 1 \\ \int (\ln \rho + 1) s & m = 1 \end{cases}. \tag{5}$$

We will specify the exact computation of the variational derivative in the next section. Using the dual representation given by (3), we write $-\mathrm{div}(\rho \nabla p) = s$, and express

$$\begin{aligned} 0 = g_\rho(\rho_t, s) + dE_\rho(s) \implies & \begin{cases} 0 = \int \frac{\partial \rho}{\partial t} p + \frac{m}{m-1}\int \rho^{m-1} s & m \neq 1 \\ 0 = \int \frac{\partial \rho}{\partial t} p + \int (\ln \rho + 1) s & m = 1 \end{cases} \\ \implies & \begin{cases} 0 = \int \frac{\partial \rho}{\partial t} p - \frac{m}{m-1}\int \rho^{m-1} \mathrm{div}(\rho \nabla p) & m \neq 1 \\ 0 = \int \frac{\partial \rho}{\partial t} p - \int (\ln \rho + 1) \mathrm{div}(\rho \nabla p) & m = 1 \end{cases}. \end{aligned}$$

Integrating by parts, we find in both cases

$$\int \left(\frac{\partial \rho}{\partial t} - \Delta \rho^m \right) p = 0$$

for all p, and under appropriate conditions, this would coincide with the notion of a weak solution of Section 2.1. This construction introduces various interesting geometric relations, and in particular, the coupling to cotangent vectors through the continuity operator is related to the notion of absolutely continuous curves discussed in Section 2.3.

3.2 Wasserstein Gradient Flows

We formalize the derivation presented in the previous section and interpret it for a general energy functional.

Consider an energy function $E : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$. Suppose μ_t is a gradient flow[1]. The key connection to the space $\mathbb{W}_p$ is that this must be an absolutely continuous curve with metric derivative in L^2. By Theorem 2, μ_t satisfies the continuity equation

$$\frac{\partial}{\partial t}\mu_t + \mathrm{div}(v_t \mu_t) = 0 \tag{6}$$

for some $\|v_t\|_{L^2(\mu_t)} \in L^1_{\mathrm{loc}}(0, \infty)$. If it were possible to compute the gradient of E in the appropriate way, the gradient flow condition $v_t = -\nabla^W E(\mu_t)$ would allow us to express (6) as

$$\frac{\partial}{\partial t}\mu_t = \mathrm{div}(\nabla^W E(\mu_t)\mu_t). \tag{7}$$

[1] Its definition and properties are explored in generality on metric spaces in "Gradient Flows in Metric Spaces and in the Spaces of Probability Measures, and Applications to Fokker-Planck Equations with Respect to Log-Concave Measures" by Ambrosio [1]. For the sake of demonstration, we omit some detail.

As outlined in Otto's derivation, to represent this dually, for each $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ we must compute the action of $d_\rho E$ on every $s \in T_\rho \mathcal{P}_2(\mathbb{R}^d)$ through the derivative

$$\frac{d}{dt}\Big|_{t=0} E(\rho_t)$$

where ρ_t is an absolutely continuous curve with $\rho_0 = \rho$ and initial velocity s. Fixing arbitrary test function $\varphi \in C_c^\infty(\mathbb{R}^n)$ with vector field $\nabla\varphi = v$, we consider the curve $\rho_t = (I + tv)_\# \rho$, which in effect turns the above differentiation into a Gateaux derivative. To identify gradients as in (5), consider a general energy functional given by

$$\mathcal{U}(\rho) = \int_{\mathbb{R}^d} U(\rho)\, dx\,.$$

Then

$$\begin{aligned} d_\rho \mathcal{U}(s) = \frac{d}{dt}\Big|_{t=0} \mathcal{U}(\rho_t) &= \frac{d}{dt}\Big|_{t=0} \int_{\mathbb{R}^d} U(\rho_t)\, dx \\ &= -\int_{\mathbb{R}^d} U'(\rho_t)\, \mathrm{div}(\rho_t v)\, dx \\ &= \int_{\mathbb{R}^d} (\nabla U'(\rho_t) \cdot v)\, \rho_t\, dx\,. \qquad \text{(by parts)} \end{aligned}$$

Through identifying $\nabla^W \mathcal{U}(\rho) = \nabla U'(\rho)$, we obtain a tangent vector $v_t = \nabla U'(\rho_t)$ which gives a weak solution to the continuity equation (2.1.1). Heuristically, this gives the correct action of $d_\rho \mathcal{U}$ on $T_\rho \mathcal{P}_2(\mathbb{R}^d)$ through the identification with the tangent space. From (7) we can show that the PDE generated by choosing $E = \mathcal{U}$ becomes

$$\frac{\partial}{\partial t}\mu_t = \mathrm{div}(\nabla U'(\mu_t)\, \mu_t)\,. \tag{8}$$

To summarize, we began with a general energy functional $\mathcal{U}$, and identified its variational derivative with solutions $(\mu_t, \nabla^W \mathcal{U}(\mu_t))$ of the continuity equation. In particular, the curves μ_t can be interpreted as a $\mathcal{U}$-maximizing flow in $\mathcal{P}_2(\mathbb{R}^d)$. This has two surprising interpretations for solutions to several known PDEs.

Example 1. Consider $U(\rho) = \rho \log \rho$. The corresponding $\mathcal{U}(\rho)$ is the Shannon-Boltzmann logarithmic entropy

$$S(\rho) = \int_{\mathbb{R}^d} \rho \log \rho\, dx\,.$$

We can compute $\nabla U'(\rho) = \frac{1}{\rho}\nabla\rho$. Substituting into 8, we recover the heat equation!

$$\frac{\partial \rho}{\partial t} = \mathrm{div}\left(\left(\frac{1}{\rho}\nabla\rho\right)\rho\right) = \mathrm{div}\,(\nabla\rho) = \Delta\rho\,.$$

As stated in C. Villani [12] (p. 438), this can be interpreted through the following sentence:

The gradient of Boltzmann's entropy is the Laplace operator.

Solutions to the heat equation can be thought of as producing entropy-maximizing flows in $\mathcal{P}_2(\mathbb{R}^d)$.

Example 2. Consider a potential $V : \mathbb{R}^d \to \mathbb{R}$ such that $\int e^{-V(x)}\,dx < \infty$. Define the functional

$$\mathcal{U}(\rho) = S(\rho) + \int_{\mathbb{R}^d} V(x)\rho(x)\,dx$$

with corresponding $U(\rho) = \rho\log\rho + V\rho$. We can compute

$$\nabla U'(\rho) = \frac{1}{\rho}\nabla\rho + \nabla V.$$

Substituting into 8, we recover the Fokker-Planck equation.

$$\frac{\partial\rho}{\partial t} = \operatorname{div}\left(\left(\frac{1}{\rho}\nabla\rho + \nabla V\right)\rho\right) = \Delta\rho + \operatorname{div}\left((\nabla V)\rho\right).$$

In sum, through the Otto calculus we can generate PDEs whose solutions give paths in $\mathcal{P}_2(\mathbb{R}^d)$ which are correspond to the maximization of a given functional.

3.3 Reformulation of Flows in $\mathbb{W}_2$

In the past section, we heuristically made the correct identification for the gradient flow to coincide with notions of Riemannian geometry. That is for function on manifold M, $\Phi : M \to \mathbb{R}$, we made the analogous identification

$$d_x\Phi(v) = \langle\nabla_x\Phi, v\rangle_x.$$

To extend these ideas, we want to connect the Otto calculus with the past theory developed for absolutely continuous curves on $\mathcal{P}_2(\Omega)$. To this end, we first consider the more general setting of a **geodesic space**, which is a type of metric space $(\mathcal{X}, d)$, in which every $x, y \in \mathcal{X}$ have a distance minimizing curve.

Definition 3.3.1. Consider an energy function $\Phi : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$, $X \in C([0,T);\mathcal{X})\cap \mathrm{AC}_{\mathrm{loc}}((0,T);\mathcal{X})$. X is a **trajectory** of the gradient flow of Φ if for all $t > 0$ $\Phi(X(t)) < \infty$ and for any $y \in \mathcal{X}$, there is a geodesic $\gamma_s : [0,1] \to \mathcal{X}$ with $\gamma_0 = X(t)$ and $\gamma_1 = y$.

$$\frac{d^+}{dt}\left(\frac{d(X(t),y)^2}{2}\right) \le \frac{d^+}{ds}\Big|_{s=0}\Phi(\gamma_s)$$

where d^+/dt denotes the lim sup of difference quotients.

When Φ is λ-convex we have the **Evolution Variational Inequality** (EVI_λ)

$$\frac{d^+}{dt}\left(\frac{d(X(t),y)^2}{2}\right) \le \Phi(y) - \Phi(X(t)) - \lambda\frac{d(X(t),y)^2}{2}$$

and it can be shown[2] [2] that EVI_λ can be used to define a gradient flow X for λ-convex energy Φ; in this case, the definitions are equivalent.

Now, we consider the particular case of curves on $\mathcal{P}_2(\Omega)$. Consider two absolutely continuous curves $\mu_t, \hat{\mu}_t$ 2.3. Recall that as a consequence of Theorem 2 they must satisfy the continuity equation (2.1.1). That is, they are solutions $(\mu_t, v_t), (\hat{\mu}_t, \hat{v}_t)$ with $v_t, \hat{v}_t \in L^2((t_1, t_2) \times \Omega)$ with respect to their respective $\mu_t, \hat{\mu}_t$. In this case, we can formulate the derivative of the Wasserstein distance through their flows

$$\frac{d^+}{dt}\left(\frac{W_2(\mu_t, \hat{\mu}_t)^2}{2}\right) = -\int_M \langle \tilde{\nabla}\psi_t, v_t\rangle \, d\mu_t - \int_M \langle \tilde{\nabla}\hat{\psi}_t, \hat{v}_t\rangle \, d\mu_t \tag{9}$$

which is shown in Chapter 23 of [12]. Here, $\tilde{\nabla}$ denotes the subgradient of $(d^2/2)$-convex functions $\psi_t, \hat{\psi}_t$ that give the solution of the Monge problem

$$\exp(\tilde{\nabla}\psi_t)_{\#}\mu_t = \hat{\mu}_t, \qquad \exp(\tilde{\nabla}\hat{\psi}_t)_{\#}\hat{\mu}_t = \mu_t.$$

If we guess $\hat{\mu}_t$ to be a constant target measure, and μ_t to evolve towards $\hat{\mu}_t$ through a gradient flow, then heuristically (from the Otto calculus) we expect $v_t = -\nabla U'(\mu_t)$ for μ_t belonging to $\mu_t \in W^{1,1}(\Omega)$, and is true under some regularity conditions.

We further extend this by describing certain convexity inequalities where the above may be used in. Let Ω have a Gibbs-style reference measure $\nu \propto \exp(-V)$ with $V \in C^2$. Let $\mu_0, \mu_1 \in \mathcal{P}_2(\Omega)$ which are absolutely continuous with respect to ν. Define $\rho = \frac{d\mu_0}{d\nu}$ be the Radon-Nikodym density, and let ψ be a $(d^2/2)$-convex function with $T = \exp(\tilde{\nabla}\psi)$ the Monge transport $\mu_0 \to \mu_1$. For $t \in [0, 1]$, let $\mu_t = (\exp(t\tilde{\nabla}\psi))_{\#}\mu$. It can be shown[3][12] that for the Shannon-Boltzmann entropy $\mathcal{U}_\nu(\rho) = \int \rho \log \rho \, d\nu$ can be writen with the *tangent formulation*

$$\mathcal{U}_\nu(\mu_1) \geq \mathcal{U}_\nu(\mu_0) + \int \langle \tilde{\nabla}\psi, \nabla\rho\rangle \, d\nu + K\frac{W_2(\mu_0, \mu_1)^2}{2}$$

for some constant K. By Cauchy-Schwarz, we can further express

$$\mathcal{U}_\nu(\mu_1) \geq \mathcal{U}_\nu(\mu_0) - \left(\int \rho \left|\tilde{\nabla}\psi\right|^2 d\nu\right)^{\frac{1}{2}} \left(\underbrace{\int |\nabla\rho|^2/\rho \, d\nu}_{\text{Fisher information}}\right)^{\frac{1}{2}} + K\frac{W_2(\mu_0, \mu_1)^2}{2}.$$

Denoting the Fisher information as a function $\mathcal{I}_\nu$, and by Equation 9,

$$\mathcal{U}_\nu(\mu_1) \geq \mathcal{U}_\nu(\mu_0) - \sqrt{\mathcal{I}_\nu(\mu_0)}W_2(\mu_0, \mu_1) + K\frac{W_2(\mu_0, \mu_1)^2}{2}$$

[2] Chapter 11, Theorem 11.15. Ambrosio, Brué, and Semola [2]
[3] Theorem 23.14, Chapter 23. Villani [12]

This example of a HWI inequality gives a familiar looking convexity expression. Note that the fisher information plays an analogical role to slope, controlling the linear change between μ_0, μ_1 when transport is a gradient flow.

3.4 Application: Stochastic Differential Equations

The last example gives a variational formulation to the evolution of laws of certain stochastic processes [6]. For a particle at position $X(t)$ evolving according to the Ito stochastic differential equation

$$dX(t) = -\nabla V(X(t)) + \sqrt{2\beta^{-1}} dW(t)$$

The particle is acted upon by potential V with random fluctuations, introduced the Wiener process $W(t)$. The parameter $\beta^{-1} \propto T$ models the increased amount of noise W with greater temperature T. It is known [9] that the probability law $\rho(t, x)$ governing $X(t)$ must satisfy the Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \beta^{-1} \Delta \rho + \text{div } ((\nabla V)\rho) \ .$$

The interpretation provided by Section 3.2 suggests that the laws $\rho(x, t)$ are maximizing the free energy functional

$$\mathcal{U}(\rho) = \beta^{-1} S(\rho) + \int_{\mathbb{R}^d} V(x)\rho(x)\, dx \, .$$

The evolution of a particle $X(t)$ can be thought of as its probability $\rho(x, t)$ changing in this way.

3.5 Application: Statistical Learning

The applications of Otto calculus has given rise to recent interesting tools for the analysis of sampling algorithms in generative models [11], [13]. The flow of maximizing an information functional is a key interpretation for the evolution of points in a generative model. This is possible since many algorithms employ a discretized version of Langevin Monte Carlo, which parametrizes a potential function V over empirical data [11], and samples with the dynamics outlined in Section 3.4. To demonstrate this, we described the relationship to the KL-divergence with a target density [4].

The KL-divergence is a statistical functional, which measures the difference in information between two distributions, namely P, P^* with densities ρ, ρ^*. Define

$$D_{\text{KL}}(P \,||\, P^*) = \int \rho \log \left(\frac{\rho}{\rho^*} \right) \ .$$

Let ρ^* be the target data density, given in the Gibbs form

$$\rho^*(x) = \exp(-V(x)) \, .$$

We can therefore express

$$D_{\mathrm{KL}}(P \,||\, P^*) = \int \rho \log \rho + \int -\log(\rho^*)\,\rho = S(\rho) + \int V \rho\,.$$

In particular, we arrive to the Fokker-Planck form of Section 3.4. The gradient flow structure of the above equation might suggest that sampling points through Langevin Monte Carlo maximize the speed of convergence to an equilibrium. In fact, the tools of Otto calculus can be used to state and prove certain inequalities about the speed of convergence of sampling, in particular the exponential convergence of Langevin Monte Carlo. To give a flavour of these results, we state one convergence theorem, omitting much detail.

Theorem 3. The target measure ρ^* satisfies a log-Sobolev inequality if and only if for all $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, ρ_0 absolutely continuous with respect to Lebesgue measure, and Langevin dynamics $t \mapsto \rho_t$ [4] satisfy the bound

$$D_{\mathrm{KL}}(\rho_t \,||\, \rho^*) \leq \exp(-2\alpha t) D_{\mathrm{KL}}(\rho_0 \,||\, \rho^*)\,.$$

A log-Sobolev inequality for the target ρ^* guarantees that for some $C > 0$ and all smooth f,

$$\mathbb{E}_{\rho^*}\left[f^2 \ln(f^2/\mathbb{E}_{\rho^*}[f^2])\right] \leq 2C\,\mathbb{E}_{\rho^*}\left[\|\nabla f\|^2\right].$$

That is, exponential convergence to the target is guaranteed with LMC if the log-Sobolev inequality of the target is satisfied. The above theorem can be interpreted as a domination condition on the Wasserstein gradient of a Langevin diffusion. A comprehensive mathematical analysis of statistical samping tools are a growing area of research [5], since they provide guarantees for many widely used learning algorithms. A thorough treatment of current results is provided in "Log-Concave Sampling" by S. Chewi [4].

[4] In particular the evolution represented by the previous gradient flow formulation.

A Notes on Referenced Works

The following chapters and articles were used as a central reference in the above.

- Chapter 4, 5 of Santambrogio [10], *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling* in Section 2.1-2.3.
- Chapter 11, 17, 18 of Ambrosio, Brué, and Semola [2], *Lectures on Optimal Transport* in Section 2.4, 3.2.
- Otto [8], "The Geometry of Dissipative Evolution Equations: The Porous Medium Equation" in Section 3.1.
- Jordan, Kinderlehrer, and Otto [6], "The Variational Formulation of the Fokker–Planck Equation" in Section 3.3.
- Chewi [4], "Log-Concave Sampling" in Section 3.4.

References

[1] Luigi Ambrosio. "Gradient Flows in Metric Spaces and in the Spaces of Probability Measures, and Applications to Fokker-Planck Equations with Respect to Log-Concave Measures". eng. In: *Bollettino dell'Unione Matematica Italiana* 1.1 (Feb. 2008), pp. 223–240. URL: http://eudml.org/doc/290477.

[2] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*. Springer International Publishing, 2021. ISBN: 9783030721626. DOI: 10.1007/978-3-030-72162-6. URL: http://dx.doi.org/10.1007/978-3-030-72162-6.

[3] Jean-David Benamou and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". In: *Numerische Mathematik* 84.3 (Jan. 2000), 375–393. ISSN: 0945-3245. DOI: 10.1007/s002110050002. URL: http://dx.doi.org/10.1007/s002110050002.

[4] Sinho Chewi. "Log-Concave Sampling". In: *Personal Website* (). URL: https://chewisinho.github.io/main.pdf.

[5] Sinho Chewi et al. *Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev*. 2024. arXiv: 2112.12662 [math.ST]. URL: https://arxiv.org/abs/2112.12662.

[6] Richard Jordan, David Kinderlehrer, and Felix Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (Jan. 1998), 1–17. ISSN: 1095-7154. DOI: 10.1137/s0036141096303359. URL: http://dx.doi.org/10.1137/S0036141096303359.

[7] William I. Newman. "A Lyapunov functional for the evolution of solutions to the porous medium equation to self-similarity. I". In: *Journal of Mathematical Physics* 25.10 (Oct. 1984), 3120–3123. ISSN: 1089-7658. DOI: 10.1063/1.526028. URL: http://dx.doi.org/10.1063/1.526028.

[8] Felix Otto. "The Geometry of Dissipative Evolution Equations: The Porous Medium Equation". In: *Communications in Partial Differential Equations* 26.1–2 (Jan. 2001), 101–174. ISSN: 1532-4133. DOI: 10.1081/pde-100002243. URL: http://dx.doi.org/10.1081/PDE-100002243.

[9] Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer Berlin Heidelberg, 1996. ISBN: 9783642615443. DOI: 10.1007/978-3-642-61544-3. URL: http://dx.doi.org/10.1007/978-3-642-61544-3.

[10] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015. ISBN: 9783319208282. DOI: 10.1007/978-3-319-20828-2. URL: http://dx.doi.org/10.1007/978-3-319-20828-2.

[11] Yang Song and Stacey F. Bent. "Learning to Generate Data by Estimating Gradients of the Data Distribution". AAI29756014. PhD thesis. Stanford, CA, USA, 2022. ISBN: 9798357507792.

[12] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN: 9783540710509. DOI: 10.1007/978-3-540-71050-9. URL: http://dx.doi.org/10.1007/978-3-540-71050-9.

[13] Max Welling and Yee Whye Teh. "Bayesian learning via stochastic gradient langevin dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, 681–688. ISBN: 9781450306195.