Gradient Flow

Chang Liu

Tsinghua University

April 24, 2017

Contents

Introduction

Introduction

Definition (Gradient Flow in Linear Space)

X is a linear space, and $F : X \to \mathbb{R}$ is smooth. Gradient flow (or steepest descent curve) is a smooth curve $x : \mathbb{R} \to X$ such that

$$x'(t) = -\nabla F(x(t)).$$

What shall we consider next and where can it be applied?

1 Existence and uniqueness of the solution
Since many PDEs are in the form of a gradient flow, the analysis can be applied to them.

Example

For $X = L^2(\mathbb{R}^n)$, a Hilbert space, and for Dirichlet energy $F(u) = \frac{1}{2}\int |\nabla u(x)|^2 \mathrm{d}x$, the Heat Equation $\partial_t u = \nabla^2 u$ is a gradient flow problem.

Introduction

Definition (Gradient Flow in Linear Space)

X is a linear space, and $F : X \to \mathbb{R}$ is smooth. Gradient flow (or steepest descent curve) is a smooth curve $x : \mathbb{R} \to X$ such that

$$x'(t) = -\nabla F(x(t)).$$

What shall we consider next and where can it be applied?

2 Numerical methods and their convergence
Since gradient flow gradually minimizes $F(x)$, so many optimization methods are related to it, e.g. gradient descent, proximal descent methods, mirror descent.

Introduction

What shall we consider next and where can it be applied?

3 Generalization to the gradient flow on general metric space.

- The need of viewing PDEs as gradient flows on general metric spaces, thus wider applicability.

Example

- PDEs in the continuity equation form $\partial_t \rho - \nabla \cdot (\rho v) = 0$, where $v = \nabla[\delta F / \delta \rho]$, can be cast as a gradient flow on the space of probabilities with Wasserstein distance.
- Heat Equation can also be viewed as a gradient flow in the Wasserstein space.

- The need of minimizing functionals on metric space.

Example

Optimization w.r.t. probability distributions, e.g. $\min_q \mathrm{KL}(q||p)$. Optimization without parameterization is possible! (e.g. Stein Variational Gradient Descent)

Gradient Flow in the Euclidean Space

Gradient Flow in the Euclidean Space

Variants of Gradient Flow in the Euclidean Space

Existence, Uniqueness and Variants

- Variant 0: $F : \mathbb{R}^n \to \mathbb{R}$ is differentiable (Cauchy Problem):

$$\begin{cases} x'(t) = -\nabla F(x(t)), \text{for } t > 0, \\ x(0) = x_0. \end{cases}$$

Theorem

$\exists!$ *solution if* ∇F *is Lipschitz.*

Existence, Uniqueness and Variants

- Variant 1: F is convex and unnecessarily differentiable:

$$\begin{cases} x'(t) \in -\partial F(x(t)), \text{for a.e. } t > 0, \\ x(0) = x_0, \end{cases}$$

where x is an absolutely continuous curve, and
$\partial F(x) = \{p \in \mathbb{R}^n : \forall y \in \mathbb{R}^n, F(y) \geq F(x) + p \cdot (y - x)\}$.

Theorem

Any two solutions x_1, x_2 of the above problem with different initial conditions satisfy $|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|$.

Corollary

For a given initial condition, the above problem has one unique solution.

Existence, Uniqueness and Variants

- Variant 2: F is semi-convex (λ convex)

Definition (λ-convex function)

F is λ-convex ($\lambda \in \mathbb{R}$) if $F(x) - \frac{\lambda}{2}|x|^2$ is convex.

$$\begin{cases} x'(t) \in -\partial F(x(t)), \text{for a.e. } t > 0, \\ x(0) = x_0, \end{cases}$$

where x is an absolutely continuous curve, and
$\partial F(x) = \{p \in \mathbb{R}^n : \forall y \in \mathbb{R}^n, F(y) \geq F(x) + p \cdot (y - x) + \frac{\lambda}{2}|y - x|^2\}$.

Theorem

Any two solutions x_1, x_2 of the above problem with different initial conditions satisfy $|x_1(t) - x_2(t)| \leq e^{-\lambda t}|x_1(0) - x_2(0)|$.

Existence, Uniqueness and Variants

- Variant 2: F is semi-convex (λ-convex)

Theorem

Any two solutions x_1, x_2 of the above problem with different initial conditions satisfy $|x_1(t) - x_2(t)| \leq e^{-\lambda t}|x_1(0) - x_2(0)|$.

Corollary

- *For a given initial condition, the above problem has one unique solution.*
- *If $\lambda > 0$ (strong convex), F has a unique minimizer x^*. $x(t) \equiv x^*$ is a solution, so for any solution $x(t)$, $|x(t) - x^*| \leq e^{-\lambda t}|x(0) - x^*|$.*

Gradient Flow in the Euclidean Space

Approximating Curves

Definition (MMS)

Minimizing Movement Scheme (MMS): for a fixed small time step τ, define a sequence $\{x_k^\tau\}_k$ by

$$x_{k+1}^\tau \in \arg\min_x F(x) + \frac{|x - x_k^\tau|^2}{2\tau}.$$

Importance:

- Practical numerical method for approximating the curve.
- Easier generalization to metric space, than $x' = -\nabla F(x)$ itself.

Properties:

- Existence of solution for mild F (e.g. Lipschitz and lower bounded by $C_1 - C_2|x|^2$).
- $\frac{x_{k+1}^\tau - x_k^\tau}{\tau} \in -\partial F(x_{k+1}^\tau)$: implicit Euler scheme (more stable but hard than explicit one: gradient descent)

Convergence:

- Define $v_{k+1}^\tau \triangleq (x_{k+1}^\tau - x_k^\tau)/\tau$, and $v^\tau(t) = v_{k+1}^\tau, t \in (k\tau, (k+1)\tau]$.
  Define two kinds of interpolations:
  1) $x^\tau(t) = x_k^\tau, t \in (k\tau, (k+1)\tau]$;
  2) $\tilde{x}^\tau(t) = x_k^\tau + (t - k\tau)v_{k+1}^\tau, t \in (k\tau, (k+1)\tau]$.
- $\tilde{x}^\tau$ is continuous and $(\tilde{x}^\tau)' = v^\tau$;
  x^τ is not continuous, but $v^\tau(t) \in -\partial F(x^\tau(t))$.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R}; \mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R};\mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Details:

1 L^p space

- For a measure space (S, Σ, μ), first define $\mathcal{L}(S;\mathbb{R}^n) \triangleq \{f : S \to \mathbb{R}^n | \int_S |f|^p \mathrm{d}\mu < \infty\}$. It is a linear space.
- Define $L^p(S;\mathbb{R}^n) \triangleq \mathcal{L}(S;\mathbb{R}^n)/\{f | f = 0 \ \mu\text{-a.e.}\}$ to be a quotient space (i.e. treat all functions that are equal μ-a.e. as one same element in L^p). Define $\|f\|_p \triangleq \left(\int_S |f|^p \mathrm{d}\mu\right)^{1/p}$, then for $1 \le p \le \infty$ it is a Banach space.
- Only $L^2(S;\mathbb{R}^n)$ can be a Hilbert space, with inner product $\langle f, g \rangle_{L^2(S;\mathbb{R}^n)} \triangleq \int_S fg \mathrm{d}\mu$.
- $L^p(S) \triangleq L^p(S;\mathbb{R})$.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R};\mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Details:

2 Weak convergence in a Hilbert space $\mathcal{H}$:

- $x_n \in \mathcal{H}, n \geq 1, x \in \mathcal{H}$, $x_n \rightharpoonup x$ is defined as:
  $\forall f \in \mathcal{H}'$, $f(x_n) \to f(x)$.
  $\Longleftrightarrow$
  $\forall y \in \mathcal{H}, \langle x_n, y\rangle_{\mathcal{H}} \to \langle x, y\rangle_{\mathcal{H}}$.
- $x_n \to x \Longrightarrow x_n \rightharpoonup x$.
  $x_n \rightharpoonup x, \|x_n\| \to \|x\| \Longrightarrow x_n \to x$.
  If $\dim(\mathcal{H}) \leq \infty$, $x_n \rightharpoonup x \Longleftrightarrow x_n \to x$.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R};\mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Details:

3 $H^k(\Omega)$ space ($\Omega \subset \mathbb{R}^n$)

- Weak derivative. For $u \in C^k(\Omega)$ and $\phi \in C_c^\infty(\Omega)$ ($\cdot_c$ for compact support),

$$\int_\Omega u D^\alpha \phi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega \phi D^\alpha u \mathrm{d}x, \text{(Integral by parts)}$$

  where $D^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$, and $|\alpha| = \sum_{i=1}^n \alpha_i$ is fixed as k. So define the *weak α-th partial derivative* of u as v:

$$\int_\Omega u D^\alpha \phi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega \phi v \mathrm{d}x, \forall \phi \in C_c^\infty(\Omega).$$

  If it exists, it is uniquely defined a.e.

3 $H^k(\Omega)$ space ($\Omega \subset \mathbb{R}^n$)

- Weak derivative. For $u \in C^k(\Omega)$ and $\phi \in C_c^\infty(\Omega)$ ($\cdot_c$ for compact support),

$$\int_\Omega u D^\alpha \phi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega \phi D^\alpha u \mathrm{d}x, \text{(Integral by parts)}$$

where $D^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$, and $|\alpha| = \sum_{i=1}^n \alpha_i$ is fixed as k. So define the *weak α-th partial derivative* of u as v:

$$\int_\Omega u D^\alpha \phi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega \phi v \mathrm{d}x, \forall \phi \in C_c^\infty(\Omega).$$

If it exists, it is uniquely defined a.e.

- Sobolev space $W^{k,p}(\Omega)$ for $k \in \mathbb{N}$ and $p \in [1, \infty]$:

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega), \forall |\alpha| \le k\},$$

with norm:

$$\|u\|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha| \le k} \|D^\alpha u\|_{L^p(\Omega)}^p\right)^{1/p}, & 1 \le p < +\infty, \\ \max_{|\alpha| \le k} \|D^\alpha u\|_{L^\infty(\Omega)}, & p = +\infty. \end{cases}.$$

$W^{k,p}(\Omega)$ is a Banach space.

- $H^k(\Omega) \triangleq W^{k,2}(\Omega)$. They are Hilbert spaces.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R};\mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Details:

4 Up to a subsequence
There exists a sequence $\tau_j \to 0$ s.t. $\tilde{x}^{\tau_j}$ and x^{τ_j} uniformly converge and v^{τ_j} weakly converge.

Theorem

If $F(x_0) < +\infty$ and $\inf F > -\infty$, then up to a subsequence $\tau_j \to 0$, both $\tilde{x}^{\tau_j}$ and x^{τ_j} converge uniformly to a same curve $x \in H^1(\mathbb{R}^n)$ and v^{τ_j} weakly converges in $L^2(\mathbb{R};\mathbb{R}^n)$ to a vector function v s.t. $x' = v$ and
1) $v(t) \in \partial F(x(t))$ a.e., if F is λ-convex;
2) $v(t) = -\nabla F(x(t)), \forall t$, if F is C^1.

Proof sketch:

$$\frac{|x^\tau_{k+1}-x^\tau_k|^2}{2\tau} \le F(x^\tau_k) - F(x^\tau_{k+1})$$

$\Longrightarrow \sum_{k=0}^{\ell} \frac{|x^\tau_{k+1}-x^\tau_k|^2}{2\tau} \le \left(F(x^\tau_0) - F(x^\tau_{\ell+1})\right) \le C$ for $F(x_0) < +\infty$ and $\inf F > -\infty$

$\Longrightarrow \int_0^T \frac{1}{2}|(\tilde{x}^\tau)'(t)|^2 \mathrm{d}t \le C$

$\Longrightarrow \tilde{x}^\tau$ is bounded in H^1 and v^τ in L^2, and the injection $H^1 \subset C^{0,1/2}$ gives an equicontinuity bound on $\tilde{x}^\tau$ of the form $|\tilde{x}^\tau(t) - \tilde{x}^\tau(s)| \le C|t-s|^{1/2}$

$\Longrightarrow$ According to the AA theorem, x^τ has a uniformly converging subsequence.

Gradient Flow in the Euclidean Space

Characterizing Properties

Motivation

- $x' = -\nabla F(x)$ (or $x' \in -\partial F(x)$) is hard to generalize to metric space! There is nothing but distance in metric space, so $\nabla F(x)$ or $\partial F(x)$ cannot be defined! (different from manifold)
- Use two properties of gradient flow that can characterize it and can be generalized to metric space.

Two charactering properties of gradient flow in $\mathbb{R}^d$:

- *Energy Dissipation Equality* (EDE) for $F \in C^1(\Omega), \Omega \subset \mathbb{R}^n$:

$$F(x(s)) - F(x(t)) = \int_s^t \left(\frac{1}{2}|x'(r)|^2 + \frac{1}{2}|\nabla F(x(r))|^2 \right) \mathrm{d}r, \forall 0 \le s < t \le 1$$

  is equivalent to $x' = -\nabla F(x)$. Note it is equivalent even for "$\geq$" (i.e. "$\geq$" $\iff$ "$=$").

- *Evolution Variational Inequality* (EVI) for λ-convex function F:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}|x(t) - y|^2 \le F(y) - F(x(t)) - \frac{\lambda}{2}|x(t) - y|^2, \forall y \in X$$

  is equivalent to $x'(t) \in -\partial F(x(t))$.

  - Sometimes also denoted as EVI_λ.
  - It is important for establishing the uniqueness and stability of gradient flow.

Gradient Flow in Metric Spaces

Gradient Flow in Metric Spaces

Generalization of Basic Concepts

For metric space (X, d),

Definition

Metric derivative of a curve $\omega : [0, 1] \to X$

$$|\omega'|(t) = \lim_{h \to 0} \frac{d(\omega(t+h), \omega(t))}{|h|},$$

if the limit exists.

- If ω is Lipschitz, $|\omega'|(t)$ exists for a.e. $t \in [0, 1]$.
- $d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} |\omega'|(s)\mathrm{d}s$.

For metric space (X, d),
In (X, d), ω' cannot be defined, but $|\omega'|$ can.

Definition

$\omega : [0, 1] \to X$ is *absolutely continuous* if $\exists g \in L^1([0, 1])$ s.t.

$$d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(s)\mathrm{d}s, \forall t_0 < t_1.$$

Let AC(X) be the set of such curves.

- AC $\Rightarrow$ Lipschitz
- AC $\Rightarrow$ Metric derivative exists a.e.

For metric space (X, d),

Definition

Length of the curve $\omega : [0, 1] \to X$:

$$\text{Length}(\omega) \triangleq \sup \left\{ \sum_{k=0}^{n-1} d(\omega(t_k), \omega(t_{k+1})) : n \geq 1, 0 = t_0 < \cdots < t_n = 1 \right\}.$$

- If $\omega \in \text{AC}(X)$, $\text{Length}(\omega) = \int_0^1 |\omega'|(t)\mathrm{d}t$.

For metric space (X, d),

Definition

Geodesic between x_0 and x_1 in X: a curve ω s.t. $\omega(0) = x_0$, $\omega(1) = x_1$, and $\text{Length}(\omega) = \min_{\tilde{\omega}}\{\text{Length}(\tilde{\omega}) : \tilde{\omega}(0) = x_0, \tilde{\omega}(1) = x_1\}$.

This is the generalization of straight lines in $\mathbb{R}^n$, and is used to extend convexity.

Definition

- *Length space*: metric space (X, d) s.t. $\forall x, y \in X, d(x, y) = \inf_{\omega \in \text{AC}(X)}\{\text{Length}(\omega) : \omega(0) = x, \omega(1) = y\}$.
- *Geodesic space*: length space and geodesic exists for any pair of points.

Riemann manifolds are geodesic spaces.

For geodesic space (X, d),

Definition

- *Geodesic convexity*: in a geodesic metric space, a function $F : X \to \mathbb{R}$ that is convex along geodesics:

$$F(x(t)) \leq (1-t)F(x(0)) + tF(x(1)),$$

  where $x(t)$ is a geodesic joining $x(0)$ and $x(1)$.
- *λ-geodesic convexity* in a geodesic metric space, a function $F : X \to \mathbb{R}$ that is λ-convex along geodesics:

$$F(x(t)) \leq (1-t)F(x(0)) + tF(x(1)) - \lambda \frac{t(1-t)}{2} d^2(x(0), x(1)).$$

For metric space (X, d),

Definition

- $g : X \to \mathbb{R}$ is an *upper gradient* of $F : X \to \mathbb{R}$: for every Lipschitz curve x,
$$|F(x(0)) - F(x(1))| \leq \int_0^1 g(x(t))|x'|(t)\mathrm{d}t.$$
- *Local Lipschitz constant* of F:
$$|\nabla F|(x) = \limsup_{y \to x} \frac{|F(x) - F(y)|}{d(x, y)}.$$
- *Descending slope* (or just *slope*) of F:
$$|\nabla^- F|(x) = \limsup_{y \to x} \frac{[F(x) - F(y)]_+}{d(x, y)}.$$

If F is Lipschitz, $|\nabla F|$ is an upper gradient.

Gradient Flow in Metric Spaces

Generalization of Gradient Flow to Metric Spaces

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (EDE-GF)

Let (X, d) be a metric space, $F : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ is an upper gradient of F. EDE-GF is a curve $x : [0, 1] \to X$ with metric derivative a.e. such that:

$$F(x(s)) - F(x(t)) = \int_s^t \left(\frac{1}{2}|x'(r)|^2 + \frac{1}{2}g(x(r))^2 \right) \mathrm{d}r, \forall 0 \leq s < t \leq 1.$$

- Existence is easy to guarantee.
- Not enough to guarantee uniqueness.

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (EVI-GF)

Let (X, d) be a geodesic space, $F : X \to \mathbb{R}$ is λ-geodesically convex. EVI-GF is a curve $x : [0, 1] \to X$ such that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}d(x(t), y)^2 \leq F(y) - F(x(t)) - \frac{\lambda}{2}d(x(t), y)^2, \forall y \in X.$$

- EVI-GF $\Rightarrow$ EDE-GF
- Uniqueness and contractivity: for two EVI-GFs $x(t)$ and $y(s)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}d(x(t), y(s))^2 \leq F(y(s)) - F(x(t)) - \frac{\lambda}{2}d(x(t)), y(s))^2,$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{2}d(x(t), y(s))^2 \leq -F(y(s)) + F(x(t)) - \frac{\lambda}{2}d(x(t)), y(s))^2.$$

Define $E(t) = \frac{1}{2}d(x(t), y(t))^2$, then $\frac{\mathrm{d}}{\mathrm{d}t}E(t) \leq -2\lambda E(t)$
$\Rightarrow d(x(t), y(t)) \leq e^{-\lambda t}d(x(0), y(0))$, which gives uniqueness for a given initial condition and exponential convergence for $\lambda > 0$.

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (EVI-GF)

Let (X, d) be a geodesic space, $F : X \to \mathbb{R}$ is λ-geodesically convex. EVI-GF is a curve $x : [0, 1] \to X$ such that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}d(x(t), y)^2 \leq F(y) - F(x(t)) - \frac{\lambda}{2}d(x(t), y)^2, \forall y \in X.$$

- A strong condition; existence is hard to guarantee.
- A sufficient condition for the existence: Compatible Convexity along Generalized Geodesics (C^2G^2):
  $\forall x_0, x_1 \in X$, $\forall y \in X$, $\exists x : [0, 1] \to X$ s.t. $x(0) = x_0, x(1) = x_1$ and
  $$F(x(t)) \leq (1 - t)F(x_0) + tF(x_1) - \lambda\frac{t(1 - t)}{2}d^2(x_0, x_1),$$
  $$d^2(x(t), y) \leq (1 - t)d^2(x_0, y) + td^2(x_1, y) - t(1 - t)d^2(x_0, x_1),$$
  i.e. λ-convexity of F and 2-convexity of $x \mapsto d^2(x, y)$ along a same curve (not necessarily geodesic).

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (Generalized MMS)

Generalization of Minimizing Movement Scheme in a metric space (X, d): for Lipschitz $F : X \to \mathbb{R} \cup \{+\infty\}$, define

$$x_{k+1}^{\tau} \in \arg\min_{x} F(x) + \frac{d(x, x_k^{\tau})^2}{2\tau}.$$

Define two kinds of interpolations in a similar way:
1) Define $x^{\tau}(t) = x_k^{\tau}, t \in (k\tau, (k+1)\tau]$;
2) Define $\tilde{x}^{\tau}(t), t \in (k\tau, (k+1)\tau]$ to be the constant-speed geodesic between x_k^{τ} and x_{k+1}^{τ}.

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Define two kinds of interpolations in a similar way:

1) Define $x^\tau(t) = x_k^\tau, t \in (k\tau, (k+1)\tau]$;

2) Define $\tilde{x}^\tau(t), t \in (k\tau, (k+1)\tau]$ to be the constant-speed geodesic between x_k^τ and x_{k+1}^τ. (So we require X to be a length space?)

Definition

Constant-speed geodesic: in a *length space*, a curve $\omega : [t_0, t_1] \to X$ s.t.

$$d(\omega(t), \omega(s)) = \frac{|t-s|}{t_1 - t_0} d(\omega(t_0), \omega(t_1)), \forall t, s \in [t_0, t_1].$$

- Constant-speed geodesics are geodesics: $\text{Length}(\omega) = \int_{t_0}^{t_1} \frac{d(\omega(t_0),\omega(t_1))}{t_1 - t_0} \mathrm{d}t = d(\omega(t_0), \omega(t_1))$.
- The followings are equivalent:
  1. $\omega : [t_0, t_1] \to X$ is a constant-speed geodesic joining x_0 and x_1;
  2. $\omega \in \text{AC}(X)$ and $|\omega'|(t) = \frac{d(\omega(t_0),\omega(t_1))}{t_1 - t_0}$ a.e.;
  3. $\omega \in \arg\min\{\int_{t_0}^{t_1} |\omega'|(t)^p \mathrm{d}t : \omega(t_0) = x_0, \omega(t_1) = x_1\}, \forall p > 1$.

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

- Define two kinds of interpolations in a similar way:
  1) Define $x^\tau(t) = x_k^\tau, t \in (k\tau, (k+1)\tau]$;
  2) Define $\tilde{x}^\tau(t), t \in (k\tau, (k+1)\tau]$ to be the constant-speed geodesic between x_k^τ and x_{k+1}^τ. (So we require X to be a length space?)
- Define v^τ. On metric (length) spaces, only its the norm can be defined: set $|v^\tau|$ as the piecewise constant speed of $\tilde{x}^\tau$,

$$|v^\tau|(t) = d(x_{k+1}^\tau, x_k^\tau)/\tau, t \in (k\tau, (k+1)\tau].$$

Definition (MMS-GF)

Let (X, d) be a metric space (not necessarily length space). A curve $x : [0, T] \to X$ is called Generalized Minimizing Movements (GMM) (I would call it MMS-GF) if there exists a sequence $\tau_j \to 0$ s.t. x^{τ_j} uniformly converges to x in (X, d).

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (MMS-GF)

Let (X, d) be a metric space (not necessarily length space). A curve $x : [0, T] \to X$ is called (by me) MMS-GF if there exists a sequence $\tau_j \to 0$ s.t. x^{τ_j} uniformly converges to x in (X, d).

Existence analysis:

- Condition for the existence of x_k^τ:
  The sub-level set $\{x : F(x) \le c\}$ is compact in X, and F is Lipschitz.
  (The corresponding topology is either the one induced by d, or a weaker topology s.t. d is lower semi-continuous w.r.t. it.)
- Condition for the existence of limit curves (i.e. MMS-GF):
  Existence of x_k^τ is enough!
  Due to $\frac{d(x_{k+1}^\tau, x_k^\tau)^2}{2\tau} \le F(x_k^\tau) - F(x_{k+1}^\tau)$, we have $d(x^\tau(t), x^\tau(s)) \le C(|t-s|^{1/2} + \sqrt{\tau})$, i.e. $\{x^\tau\}_\tau$ are equi-Hölder continuous with exponent $1/2$ (up to a negligible error of order $\sqrt{\tau}$). So by AA theorem, the set $\{x^\tau\}_\tau$ has uniformly converging subsequences, i.e. MMS-GF. But not unique and no relation with F (EDE or EVI) is obtained.

Three ways to generalize gradient flow to metric space: EDE-GF, EVI-GF, MMS-GF.

Definition (MMS-GF)

Let (X, d) be a metric space (not necessarily length space). A curve $x : [0, T] \to X$ is called (by me) MMS-GF if there exists a sequence $\tau_j \to 0$ s.t. x^{τ_j} uniformly converges to x in (X, d).

To relate MMS-GF to F and other generalizations:

- If in addition to "$\{x : F(x) \leq c\}$ is compact in X, F is Lipschitz", F and $|\nabla^- F|$ are lower-semicontinuous, we have $\frac{1}{2}\int_0^t |x'|(r)^2 \mathrm{d}r + \frac{1}{2}\int_0^t |\nabla^- F(x(r))|^2 \mathrm{d}r \leq F(x(0)) - F(x(t)), \forall 0 \leq t \leq T$. (not EDE)
- If additionally, the slope $|\nabla^- F|$ is an upper gradient of F, we have EDE: $\frac{1}{2}\int_s^t |x'|(r)^2 \mathrm{d}r + \frac{1}{2}\int_s^t |\nabla^- F(x(r))|^2 \mathrm{d}r \leq F(x(s)) - F(x(t)), \forall 0 \leq s < t \leq T$.
- If F is λ-geodesically convex, all the conditions are met.

Conclusion for now

Table: Conclusion of extentions of gradient flow to metric space

Extension	Requirement	Existence	Uniqueness and Contractivity
EVI-GF	X geodesic space, F λ-geod. convex	Hard. C^2G^2 is a sufficient condition	Guaranted
EDE-GF	X metric space	Easy	Not guaranteed
MMS-GF	X metric space	Relatively easy. "$\{x : F(x) \leq c\}$ compact and F Lipschitz" or "F λ-geod. convex" suffices	Not guaranteed

- EVI-GF $\subset$ EDE-GF
- MMS-GF $\subset$ EDE-GF if "$\{x : F(x) \leq c\}$ compact, F Lipschitz, F and $|\nabla^- F|$ lower-semicont., $|\nabla^- F|$ is an upper grad. of F" or "F λ-geod. convex"

Gradient Flows on Wasserstein Spaces

Gradient Flows on Wasserstein Spaces

Recap. of Optimal Transport Problems

Recap. of Optimal Transport Problems

- Settings
  Let X, Y be two measurable spaces, $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ are fixed measures, Let $c : X \times Y \to \mathbb{R}$ be a cost function.

Definition (push-forward of a measure)

For a measurable function $T : X \to Y$ and a measure $\mu \in \mathcal{P}(X)$, define the push-forward of μ under T, $T_{\#}\mu$, to be a measure on Y s.t.

$$T_{\#}\mu(A) = \mu(T^{-1}(A)), \forall A \in \sigma\text{-algebra of } Y.$$

Example

For $X = Y = \mathbb{R}^n$ and T invertible, then in terms of p.d.f.,
$T_{\#}\mu = (\mu \circ T^{-1})|\det(\nabla T^{-1})|$, i.e. rule of change of variables.

Recap. of Optimal Transport Problems

- Monge's Problem:

$$(MP) \inf_{T_{\#}\mu=\nu} \int_X c(x, T(x))\mathrm{d}\mu(x).$$

  - (Optimal) T is called a (optimal) transport map.
  - The problem may not be feasible.

- Kantorovich's Problem:

$$(KP) \inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y),$$

  where $\Pi(\mu,\nu) \triangleq \{\gamma|(\pi_X)_{\#}\gamma = \mu, (\pi_Y)_{\#}\gamma = \nu\}$.
  - (Optimal) γ is called a (optimal) transport plan.
  - The problem is always feasible.

- MP is a special case of KP, where γ is restricted to the form $\gamma = (\mathrm{id} \times T)_{\#}\mu$. If T^* exists, $\gamma^* = (\mathrm{id} \times T^*)_{\#}\mu$ is also optimal.

Recap. of Optimal Transport Problems

- Dual Kantorovich Problem:
  - Direct form:

$$(DKP) \sup_{\substack{\phi \in L^1(X), \psi \in L^1(Y), \\ \phi(x)+\psi(y) \leq c(x,y)}} \int_X \phi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu.$$

  - Reformulation:

Definition

- c-transform (c-conjugate) of $\chi : X \to \bar{\mathbb{R}}$, $\chi^c : Y \to \bar{\mathbb{R}}$, is defined as $\chi^c(y) \triangleq \inf_{x \in X} c(x, y) - \chi(x)$.
- $\Psi_c(X) \triangleq \{\chi^c | \chi : X \to \bar{\mathbb{R}}\}$. $\psi : Y \to \bar{\mathbb{R}}$ is c-concave if $\psi \in \Psi_c(X)$.

$$(DKP') \sup_{\phi \in \Psi_c(X)} \int_X \phi \mathrm{d}\mu + \int_Y \phi^c \mathrm{d}\nu.$$

Recap. of Optimal Transport Problems

- Dual Kantorovich Problem:
  - Reformulation:

$$(DKP') \sup_{\phi \in \Psi_c(X)} \int_X \phi \mathrm{d}\mu + \int_Y \phi^c \mathrm{d}\nu.$$

Definition (Kantorovich potential)

The optimal ϕ of $(DKP0')$ is called Kantorovich potential, denoted by φ.

When c is uniformly continuous (e.g. when c is continuous and X is compact), then the existence of Kantorovich potential φ can be proven (by AA theorem).

Remark

Strong duality holds: $KP(\mu, \nu) = DKP(\mu, \nu)$.

Recap. of Optimal Transport Problems

- Dual Kantorovich Problem:
  - Special case 1: $X = Y$, $c(x, y) = d(x, y)$ is a distance:

$$(DKP1) \sup_{\phi \in \mathrm{Lip}_1} \int_X \phi \mathrm{d}\mu - \int_X \phi \mathrm{d}\nu.$$

  - Special case 2: $X = Y = \Omega \subset \mathbb{R}^n$ and $c(x, y) = \frac{1}{2}|x - y|^2$:

Theorem

- *For quadratic cost and $\Omega \subset \mathbb{R}^n$ close, bounded and connected, $\exists!$ optimal transport plan γ^* for (KP).*
- *Additionally, if μ is absolutely continuous, optimal transport map T^* exists and $\gamma^* = (id, T^*)_{\#}\mu$. Moreover, there exists a Kantorovich potential φ s.t. $\nabla\varphi$ is unique μ-a.e, and $T = \nabla u$ a.e., where $u(x) \triangleq \frac{x^2}{2} - \phi(x)$ is convex.*

Recap. of Optimal Transport Problems

- Dual Kantorovich Problem:
  - Special case 2: $X = Y = \Omega \subset \mathbb{R}^n$ and $c(x, y) = \frac{1}{2}|x - y|^2$:

Theorem

- *For quadratic cost and $\Omega \subset \mathbb{R}^n$ close, bounded and connected, $\exists!$ optimal transport plan γ^* for (KP).*
- *Additionally, if μ is absolutely continuous, optimal transport map T^* exists and $\gamma^* = (id, T^*)_{\#}\mu$. Moreover, there exists a Kantorovich potential φ s.t. $\nabla\varphi$ is unique μ-a.e, and $T = \nabla u$ a.e. where $u(x) \triangleq \frac{x^2}{2} - \phi(x)$ is convex.*

Corollary

- *Under the same condition, any gradient of a convex function is an optimal map between μ and its image measure.*
- *Optimal transport map uniquely exists for $c(x, y) = h(x - y)$ with h strictly convex. (e.g. $|x - y|^p, p > 1$).*

Gradient Flows on Wasserstein Spaces

The Wasserstein Space

The Wasserstein Space

Definition

On metric space (X, d), for $p \geq 1$ and a fixed point $x_0 \in X$, define $m_p(\mu) \triangleq \int_X d(x, x_0)^p \mathrm{d}\mu(x)$, and $\mathcal{P}_p(X) \triangleq \{\mu \in \mathcal{P}(X) : m_p(\mu) < +\infty\}$, which is independent of the choice of x_0.

Theorem

$W_p(\mu, \nu) \triangleq \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_X d(x, y)^p \mathrm{d}\gamma(x, y)\right)^{1/p}$ *is a distance on* $\mathcal{P}_p(X)$

Definition (Wasserstein space)

$\mathbb{W}_p(X) \triangleq (\mathcal{P}_p(X), W_p)$.

The Wasserstein Space

Definition (Wasserstein space)

$\mathbb{W}_p(X) \triangleq (\mathcal{P}_p(X), W_p)$.

Theorem

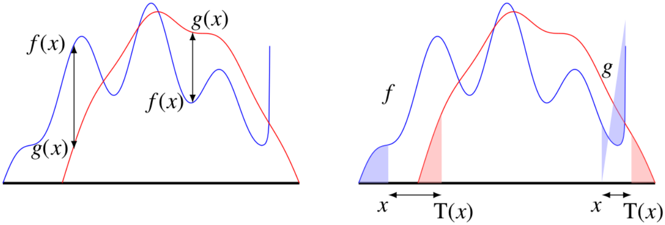
In $\mathbb{W}_p(X)$ with $p \geq 1$, given $\mu, \mu_n \in \mathcal{P}_p(X), n \in \mathbb{N}$, the followings are equivalent:

- $\mu_n \to \mu$ *w.r.t.* W_p;
- $\mu_n \rightharpoonup \mu$ *and* $m_p(\mu_n) \to m_p(\mu)$;
- $\int_X \phi \mathrm{d}\mu_n \to \int_X \phi \mathrm{d}\mu, \forall \phi \in \left\{\phi \in C^0(X) : \exists A, B \in \mathbb{R} \text{ s.t. } |\phi(x)| \leq A + Bd(x, x_0)^p, \forall x, x_0 \in X\right\}$.

The Wasserstein Space

Special cases:

- Case 1: (X, d) is compact.
  - $\mathcal{P}(X) = \mathcal{P}_p(X), \forall p \geq 1$.
  - $\mu_n \to \mu$ w.r.t. $W_p \iff \mu_n \rightharpoonup \mu$.
- Case 2: $X = \Omega \subset \mathbb{R}^d$ and $p \in [1, +\infty)$. $c(x, y) = \|x - y\|_p$.



- L^p distance between p.d.f.s of two measures: "vertical" distance. W_p distance between two measures: "horizontal" distance.
- $p_1 \leq p_2 \implies W_{p_1} \leq W_{p_2}$. If Ω is bounded, $W_{p_1} \leq W_{p_2} \implies p_1 \leq p_2$.

The Wasserstein Space

Geodesic on $\mathbb{W}_p(\Omega)$:

Theorem (McCann's displacement interpolation)

- *If* $\Omega \in \mathbb{R}^d$ *is convex, then* $\mathbb{W}_p(\Omega)$ *is a length space, and for* $\mu, \nu \in \mathbb{W}_p(\Omega)$ *and* γ *as optimal transport plan from* μ *to* ν*, then*

$$\mu^\gamma(t) \triangleq (\pi_t)_\# \gamma, \text{ where } \pi_t(x, y) \triangleq (1-t)x + ty,$$

  is a constant-speed geodesic.
- *If* $p > 1$*, then all the constant-speed geodesics are of this form.*
- *If additionally* μ *is absolutely continuous, then there is only one geodesic, whose form is*

$$\mu_t = (T_t)_\# \mu, \text{ where } T_t \triangleq (1-t)\mathrm{id} + tT,$$

  where T *is the optimal transport map from* μ *to* ν*.*

The Wasserstein Space

Geodesic convexity in $\mathbb{W}_2(\Omega)$ (displacement convexity):

- Definition is given by the general gradient flow theory.
- Important examples:

Definition (Important functionals on $\mathbb{W}_2(\Omega)$)

For $f : \mathbb{R} \to \mathbb{R}$ convex, $V : \Omega \to \mathbb{R}$, $W : \mathbb{R}^d \to \mathbb{R}$ symmetric ($W(x) = W(-x)$), define

$$\mathcal{F}(\rho) = \int f(\rho(x))\mathrm{d}x, \mathcal{V}(\rho) = \int V(x)\mathrm{d}\rho, \mathcal{W} = \frac{1}{2}\iint W(x-y)\mathrm{d}\rho(x)\mathrm{d}\rho(y).$$

Theorem

- *λ-convexity on Ω of V (or W) $\Longrightarrow$ λ-geodesic convexity on $\mathbb{W}_2(\Omega)$ of $\mathcal{V}$ (or $\mathcal{W}$).*
- *$f(0) = 0$ and $s^d f(s^{-d})$ is convex and decreasing, Ω is convex, $1 < p < \infty$ $\Longrightarrow$ $\mathcal{F}$ is geodesically convex in $\mathbb{W}_2(\Omega)$.*

Gradient Flows on Wasserstein Spaces

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

Curves/flows on $\mathbb{W}_p(\Omega), \Omega \subset \mathbb{R}^n$

Continuity equation:
What is special for $\mathbb{W}_p(\Omega)$, is that it is of probability distributions. The curve/flow/dynamics in $\mathbb{W}_p(\Omega)$, μ_t, represents the evolution of distributions. This evolution can be associated with (viewed as a result of) an evolution/dynamics in $\mathbb{R}^n$, represented by vector field v_t. The typical relation between them is the *continuity equation*:

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0.$$

Curves/flows on $\mathbb{W}_p(\Omega), \Omega \subset \mathbb{R}^n$

Theorem

Let $p > 1$, $\Omega \subset \mathbb{R}^d$ open, bounded and connected.

- *Let $\{\mu_t\}_{t\in[0,1]}$ be an AC curve in $\mathbb{W}_p(\Omega)$. Then for a.e. $t \in [0,1]$ there exists a vector field $v_t \in L^p(\mu_t; \mathbb{R}^d)$ s.t. 1) $\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$ is satisfied in the sense of distributions; 2) for a.e. $t \in [0,1]$, $\|v_t\|_{L^p(\mu_t)} \leq |\mu'|(t)$.*
- *Conversely, if $\{\mu_t\}_{t\in[0,1]} \subset \mathcal{P}_p(\Omega)$ and $\forall t$ we have a vector field $v_t \in L^p(\mu_t; \mathbb{R}^d)$ with $\int_0^1 \|v_t\|_{L^p(\mu_t)} \mathrm{d}t < +\infty$ solving $\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$, then $\{\mu_t\}_{t\in[0,1]}$ is AC in $\mathbb{W}(\Omega)$ and for a.e. $t \in [0,1], |\mu'|(t) \leq \|v_t\|_{L^p(\mu_t)}$.*
- *Thus in both cases, the conclusion can be strengthened with $|\mu'|(t) = \|v_t\|_{L^p(\mu_t)}$.*

(I guess $v_t^i : \Omega \to \mathbb{R}, 1 \leq i \leq d$ satisfies $|v_t^i|^p$ is μ_t-integrable, and $\|v_t\|_{L^p(\mu_t)} = \left(\sum_{i=1}^d \int_\Omega |v_t^i(x)|^p \mathrm{d}\mu_t(x) \right)^{1/p}$.)

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

- We only consider absolutely continuous measures, denoted by ρ, so that distribution density can be accessed.
- Let $F : \mathbb{W}_2(\Omega) \to \bar{\mathbb{R}}$ be a functional on $\mathbb{W}_w(\Omega)$. Use MMS-GF to define the gradient flow w.r.t. F:

$$\rho_{k+1}^{\tau} \in \arg\min_{\rho} F(\rho) + \frac{W_2^2(\rho, \rho_k^{\tau})}{2\tau}$$

- General existence conditions apply, e.g. $\{\rho : F(\rho) \leq c\}$ compact and F Lipschitz, or F λ-geodesically convex.
- Special result:

Theorem

Let $F : \mathbb{W}_2(\Omega) \to \bar{\mathbb{R}}$ be λ-geodesically convex, then MMS-GF w.r.t. F exists. Let ρ_t^0, ρ_t^1 be two solutions, and define $E(t) \triangleq \frac{1}{2} W_2^2(\rho_t^0, \rho_t^1)$. Then $E(t) \leq e^{-\lambda t} E(0)$, which implies uniqueness for a given initial condition, and stability and exponential convergence for $\lambda > 0$.

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

- To relate F and the vector field v_t, we need the notion of first variation.

Definition (First Variation)

First variation of a functional $G : \mathcal{P}(\Omega) \to \mathbb{R})$ is defined as $\frac{\delta G}{\delta \rho}(\rho) : \Omega \to \mathbb{R}$ s.t. $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} G(\rho + \varepsilon\chi)|_{\varepsilon=0} = \int \frac{\delta G}{\delta \rho}(\rho)(x)\mathrm{d}\chi(x), \forall \chi \in \{\chi : \exists \varepsilon_0$ s.t. $\forall \varepsilon \in [0, \varepsilon_0], \rho + \varepsilon\chi \in \mathcal{P}(\Omega)\}$.

(Recall that on $\mathbb{R}^d$, $\nabla F \in \mathbb{R}^d$ s.t. $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(x + \varepsilon v)|_{\varepsilon=0} = (\nabla F, v), \forall v \in \mathbb{R}^d$.)

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

- To relate F and the vector field v_t, we need the notion of first variation.

Definition (Important functionals on $\mathbb{W}_2(\Omega)$)

For $f : \mathbb{R} \to \mathbb{R}$ convex, $V : \Omega \to \mathbb{R}$, $W : \mathbb{R}^d \to \mathbb{R}$ symmetric ($W(x) = W(-x)$), define

$$\mathcal{F}(\rho) = \int f(\rho(x))\mathrm{d}x, \mathcal{V}(\rho) = \int V(x)\mathrm{d}\rho, \mathcal{W} = \frac{1}{2}\iint W(x-y)\mathrm{d}\rho(x)\mathrm{d}\rho(y).$$

Theorem

$\frac{\delta \mathcal{F}}{\delta \rho} = f'(\rho), \frac{\delta \mathcal{V}}{\delta \rho} = V, \frac{\delta \mathcal{W}}{\delta \rho} = W * \rho$ *(convolution)*

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

- To relate F and the vector field v_t, we need the notion of first variation.

Theorem

The first variation of Wasserstein distance with cost function c: $\frac{\delta W_c(\rho,\nu)}{\delta \rho} = \varphi$, *if* ρ, ν *are defined on* $\Omega \subset \mathbb{R}^d$, $c : \Omega \times \Omega \to \mathbb{R}$ *continuous, and Kantorovich potential* φ *is unique and c-concave.*

Gradient Flows on $\mathbb{W}_2(\Omega), \Omega \subset \mathbb{R}^n$

- Relate F and the vector field v_t.

Theorem

For the Minimizing Movement Scheme $\rho_{k+1}^\tau \in \arg\min_\rho F(\rho) + \frac{W_2^2(\rho,\rho_k^\tau)}{2\tau}$, the optimality condition is:

$$\frac{\delta F}{\delta \rho}(\rho_{k+1}^\tau) + \frac{\varphi}{\tau} = const.$$

where φ is the Kantorovich potential from ρ_{k+1}^τ to ρ_k^τ.

- Relation between T^* and φ: $T^*(x) = x - \nabla\varphi$,
  relation between v_t and T: $v_t(x) = (x - T(x))/\tau$,
  so in the limit $\tau \to 0$, the gradient flow w.r.t. F induces a flow in $\mathbb{R}^n$:
  $$v_t(x) = -\nabla(\frac{\delta F}{\delta \rho}(\rho_t))(x),$$
  and the flow ρ_t in $\mathbb{W}_2(\Omega)$ is:
  $$\partial_t \rho_t - \nabla \cdot \left(\rho_t \nabla \left(\frac{\delta F}{\delta \rho}(\rho_t)\right)\right) = 0.$$

Gradient Flows on Wasserstein Spaces

Numerical methods from the JKO scheme

Numerical methods from the JKO scheme

- JKO: Jordan-Kinderleherer-Otto.
- Solve the problem of the form $\min\{F(\rho) + \frac{1}{2}W_2^2(\rho, \nu) : \rho \in \mathcal{P}(\Omega)\}$ (τ is included in F.)
- Two recent methods:
  1) based on the Benamou-Brenier formula, for convex $F(\rho)$;
  2) based on methods from semi-discrete optimal transport, for geodesically convex F. (involving techniques in computing geometry; not covered in this slide)

Benamou-Brenier formula

Theorem (McCann's displacement interpolation)

- *If* $\Omega \in \mathbb{R}^d$ *is convex, for* $\mu, \nu \in \mathbb{W}_p(\Omega)$ *and* γ *as optimal transport plan from* μ *to* ν*, then*

$$\mu^\gamma(t) \triangleq (\pi_t)_\# \gamma, \text{ where } \pi_t(x, y) \triangleq (1 - t)x + ty,$$

  is a constant-speed geodesic.
- *If* $p > 1$*, then all the constant-speed geodesics are of this form.*
- *If additionally* μ *is absolutely continuous, then there is only one geodesic, whose form is*

$$\mu_t = (T_t)_\# \mu, \text{ where } T_t \triangleq (1 - t)\mathrm{id} + tT,$$

  where T *is the optimal transport map from* μ *to* ν*.*

Benamou-Brenier formula

From this theorem, we can see:

- For the cost $c(x, y) = |x - y|^p$, find an optimal transport $\iff$ find constant-speed geodesic in $\mathbb{W}_p$, since they are closely related and (when $p > 1$ and μ absolutely continuous) they are one-to-one.
- Find constant-speed geodesic: $\min_{\mu_t} \int_0^1 |\mu'|(t)^p \mathrm{d}t$.
- In $\mathbb{W}_p$, we have $|\mu'|(t) = \|v_t\|_{L^p(\mu_t)} = \left(\int_\Omega |v_t|^p \mathrm{d}\mu_t\right)^{1/p}$, where v_t is the velocity field solving the continuity equation.

So, we get the Benamou-Brenier formula (Time-dependent Kantorovich Problem):

$$(TKP1) \min_{\substack{(\rho_t, v_t):\rho_0=\mu,\rho_1=\nu, \\ \partial_t\rho_t+\nabla\cdot(v_t\mu_t)=0}} \int_0^1 \int_\Omega |v_t|^p \mathrm{d}\rho_t \mathrm{d}t.$$

- It is a kinetic energy minimization problem.
- It selects constant-speed geodesics connecting μ to ν.
- It is non-convex for (ρ_t, v_t).

Benamou-Brenier formula

$$(TKP1) \min_{\substack{(\rho_t, v_t): \rho_0=\mu, \rho_1=\nu, \\ \partial_t \rho_t + \nabla \cdot (v_t \mu_t) = 0}} \int_0^1 \int_\Omega |v_t|^p \mathrm{d}\rho_t \mathrm{d}t.$$

Transform it to convex: let $E_t = v_t \rho_t$, and use (ρ_t, E_t) as arguments:

$$(TKP2) \min_{\substack{(\rho_t, E_t): \rho_0=\mu, \rho_1=\nu, \\ \partial_t \rho_t + \nabla \cdot E_t = 0}} \int_0^1 \int_\Omega \frac{|E_t|^p}{\rho_t^{p-1}} \mathrm{d}x \mathrm{d}t.$$

Benamou-Brenier formula

$$(TKP2) \min_{\substack{(\rho_t, E_t): \rho_0 = \mu, \rho_1 = \nu, \\ \partial_t \rho_t + \nabla \cdot E_t = 0}} \int_0^1 \int_\Omega \frac{|E_t|^p}{\rho_t^{p-1}} \mathrm{d}x \mathrm{d}t.$$

Further transformation:

- $K_q \triangleq \{(a, b) \in \mathbb{R} \times \mathbb{R}^d : a + \frac{1}{q}|b|^q \leq 0\}$ for $q = p/(p-1)$ conjugate of p. It is convex in $\mathbb{R} \times \mathbb{R}^d$.
- For $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$, define

$$f_p(t, x) \triangleq \sup_{(a,b) \in K_q} (at + b \cdot x) = \begin{cases} \frac{1}{p} \frac{|x|^p}{t^{p-1}}, & \text{if } t > 0 \\ 0, & \text{if } t = 0, x = 0 \\ +\infty, & \text{if } t = 0, x \neq 0, \text{or } t < 0. \end{cases}$$

So the optimization problem can be reformulated as

$$(TKP3) \min_{\substack{(\rho_t, E_t): \rho_0 = \mu, \rho_1 = \nu, \\ \partial_t \rho_t + \nabla \cdot E_t = 0}} \sup_{\substack{(a,b) \in \\ C(\Omega \times [0,1]; K_q)}} \iint a \mathrm{d}\rho + \iint b \cdot \mathrm{d}E,$$

where $\iint$ indicates integral w.r.t. both space and time.

Benamou-Brenier formula

$$(TKP3) \min_{\substack{(\rho_t, E_t):\rho_0=\mu,\rho_1=\nu,\\ \partial_t\rho_t+\nabla\cdot E_t=0}} \sup_{\substack{(a,b)\in\\ C(\Omega\times[0,1];K_q)}} \iint a\mathrm{d}\rho + \iint b\cdot \mathrm{d}E,$$

Utilizing

$$\sup_{\phi\in C^1([0,1]\times\Omega)} -\iint \partial_t\phi\mathrm{d}\rho - \iint \nabla\phi\cdot\mathrm{d}E + \int \phi_1\mathrm{d}\nu - \int \phi_0\mathrm{d}\mu$$

$$= \begin{cases} 0, \text{ if } \rho_0=\mu, \rho_1=\nu, \partial_t\rho_t+\nabla\cdot E_t=0, \\ +\infty, \text{ otherwise.} \end{cases},$$

we get

$$(TKP4) \min_{(\rho_t, E_t)} \sup_{\substack{(a,b)\in C(\Omega\times[0,1];K_q),\\ \phi\in C^1([0,1]\times\Omega)}} \iint (a-\partial_t\phi)\mathrm{d}\rho + \iint (b-\nabla\phi)\cdot\mathrm{d}E$$

$$+\int \phi_1\mathrm{d}\nu - \int \phi_0\mathrm{d}\mu.$$

Benamou-Brenier formula

$$(TKP4) \min_{(\rho_t, E_t)} \sup_{\substack{(a,b) \in C(\Omega \times [0,1]; K_q), \\ \phi \in C^1([0,1] \times \Omega)}} \iint (a - \partial_t \phi) \mathrm{d}\rho + \iint (b - \nabla \phi) \cdot \mathrm{d}E$$

$$+ \int \phi_1 \mathrm{d}\nu - \int \phi_0 \mathrm{d}\mu.$$

To simplify notation, let $m = (\rho, E)$, $A = (a, b)$, $m \cdot A = \int a \mathrm{d}\rho + \int b \cdot \mathrm{d}E$, $\nabla_{t,x}\phi = (\partial_t \phi, \nabla \phi)$, $G(\phi) = \int \phi_1 \mathrm{d}\nu - \int \phi_0 \mathrm{d}\mu$, $I_{K_p}(\cdot)$ be indicator function, then

$$(TKP4') \min_m \sup_{A,\phi} L(m, (A, \phi)) \triangleq m \cdot (A - \nabla_{t,x}\phi) - I_{K_p}(A) + G(\phi).$$

This is a mini-max problem.

Benamou-Brenier formula

$$(TKP4') \min_{m} \sup_{A,\phi} L(m,(A,\phi)) \triangleq m \cdot (A - \nabla_{t,x}\phi) - I_{K_p}(A) + G(\phi).$$

$L(m,(A,\phi))$ is the Lagrangian of the form $L(X,Y) = X \cdot \Lambda Y - H(Y)$, where Λ is a linear operator. Its optimality condition

$$\begin{cases} \Lambda Y = 0 \\ \Lambda^* X - \nabla H(Y) = 0 \end{cases}$$

is the same as the one of the *augmented Lagrangian* $\tilde{L}(X,Y) = X \cdot \Lambda Y - H(Y) - \frac{r}{2}|\Lambda Y|^2$:

$$\begin{cases} \Lambda Y = 0 \\ \Lambda^* X - \nabla H(Y) - r\Lambda^*\Lambda Y = 0 \end{cases},$$

for any $r > 0$, and Λ^* is its adjoint w.r.t. the inner product. So finally,

$$(TKP5) \min_{m} \sup_{A,\phi} m \cdot (A - \nabla_{t,x}\phi) - I_{K_p}(A) + G(\phi) - \frac{r}{2}\|A - \nabla_{t,x}\phi\|^2.$$

Benamou-Brenier formula

$$(TKP5) \min_{m} \sup_{A,\phi} m \cdot (A - \nabla_{t,x}\phi) - I_{K_p}(A) + G(\phi) - \frac{r}{2}\|A - \nabla_{t,x}\phi\|^2.$$

To solve this,

- Optimize ϕ: minimize a quadratic functional in calculus of variations, e.g. solving a Poisson equation
- Optimize A: a pointwise minimization problem, specifically a projection on the convex set K_q
- Optimize m: gradient descent. $m \leftarrow m - r(A - \nabla_{t,x}\phi)$

Application

Application

To be continued... :(

My Remarks

My Remarks

Given a functional $F(\rho)$ on $\mathbb{W}_2(\Omega)$ with $\Omega \subset \mathbb{R}^n$, if we want to minimize it, we can find a gradient flow on $\mathbb{W}_2(\Omega)$ defined by F, which gradually minimizes F, by:

1. the MMS discretization with step size τ: we get $\{\rho_k^\tau\}_k$, where
$$\rho_{k+1}^\tau \in \arg\min_\rho F(\rho) + \frac{W_2^2(\rho, \rho_k^\tau)}{2\tau}.$$
In this case we directly get a sequence of distributions DIRECTLY, e.g. in terms of pdf.
2. simulating a dynamics/flow on Ω, which is associated with the gradient flow on $\mathbb{W}_2(\Omega)$ (or which is the cause/reason of the evolution of the distribution described by the gradient flow on $\mathbb{W}_2(\Omega)$). The dynamics/flow on Ω is governed by
$$\frac{\mathrm{d}}{\mathrm{d}t}\xi_t(x) = v_t(x), v_t(x) = -\nabla \left(\frac{\delta F}{\delta \rho}(\rho_t) \right)(x).$$
In this case the distribution is embodied as samples from it. We will

My Remarks on SVGD

- Afterwards, we will only consider the second approach to get the gradient flow.
- Take $F(\rho) = \mathrm{KL}(\rho || p)$, for a fixed distribution p.
- Compare the results of gradient flow and variation calculus. (Omit $\cdot_t$ temporarily)

By Gradient Flow

$F(\rho) = \int_\Omega \rho \log \frac{\rho}{p} \mathrm{d}x$, $\frac{\delta F}{\delta \rho} = \log \rho - \log p + 1$, so:

$$v(x) = \nabla \log p(x) - \nabla \log \rho(x).$$

My Remarks on SVGD

By Variation Calculus

- Find the "directional derivative" $G(v,\rho)$ of $F(\rho)$ w.r.t. the dynamics $\frac{\mathrm{d}}{\mathrm{d}t}\xi_t(x) = v_t(x)$:
$$G(v,\rho) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(\rho_{[\xi^{(\varepsilon)}]})|_{\varepsilon=0}, \xi^{(\varepsilon)}(x) = x + \varepsilon v(x),$$
$$\rho_{[\xi^{(\varepsilon)}]}(x) = \rho({\xi^{(\varepsilon)}}^{-1}(x))|\mathrm{Jac}{\xi^{(\varepsilon)}}^{-1}| \approx \rho(x - \varepsilon v(x))|\mathrm{Jac}(x - \varepsilon v(x))|.$$
For $F(\rho) = \mathrm{KL}(\rho||p)$, by my written notes on SVGD or the electronic notes on R-SVGD, $G(v,\rho) = \int_\Omega \rho[\nabla \log p \cdot v + \nabla \cdot v]\mathrm{d}x$.
- Find $v(x)$ s.t. it maximizes $G(v,\rho)$: $\max_v G(v,\rho)$, s.t. $\|v\| = 1$. If we take the norm as $\|v\| = \frac{1}{2}\sum_{i=1}^n \int_\Omega v_i^2(x)\rho(x)\mathrm{d}x$ and introduce Lagrange multiplier λ,
$$\min_\lambda \max_v G(v,\rho) + \frac{\lambda}{2}\sum_{i=1}^n \int_\Omega v_i^2(x)\rho(x)\mathrm{d}x - \lambda.$$
For $F(\rho) = \mathrm{KL}(\rho||p)$, take the first variation w.r.t. v_i, i.e. let $\frac{\partial L}{\partial v_i} - \sum_{j=1}^n \partial_j \left(\frac{\partial L}{\partial(\partial_j v_i)}\right) = 0$:
$$\rho\partial_i \log p - \partial_i\rho + \lambda\rho v_i = 0, v_i \propto \partial_i \log p - \partial_i \log \rho$$

My Remarks on SVGD

By Variation Calculus

- Find $v(x)$ s.t. it maximizes $G(v,\rho)$: $\max_v G(v,\rho)$, s.t. $\|v\| = 1$. If we take the norm as $\|v\| = \frac{1}{2}\sum_{i=1}^n \int_\Omega v_i^2(x)\rho(x)\mathrm{d}x$ and introduce Lagrange multiplier λ,

$$\min_\lambda \max_v G(v,\rho) + \frac{\lambda}{2}\sum_{i=1}^n \int_\Omega v_i^2(x)\rho(x)\mathrm{d}x - \lambda.$$

For $F(\rho) = \mathrm{KL}(\rho||p)$, take the first variation w.r.t. v_i, i.e. let $\frac{\partial L}{\partial v_i} - \sum_{j=1}^n \partial_j \left(\frac{\partial L}{\partial(\partial_j v_i)}\right) = 0$:

$$\rho\partial_i \log p - \partial_i\rho + \lambda\rho v_i = 0, v_i \propto \partial_i \log p - \partial_i \log \rho,$$

as the same as the result by gradient flow.

However, in SVGD neither is adopted. It uses v in the space of vector-valued RKHS and turn the objective as an inner product in it.

My Remarks on General Results

The general equivalence of Gradient Flow and Variation Calculus?

$$
\begin{aligned}
G(v,\rho) =& \frac{\mathrm{d}}{\mathrm{d}\varepsilon} F\Big(\rho(x-\varepsilon v(x))|\mathrm{Jac}(x-\varepsilon v(x))|\Big)\Big|_{\varepsilon=0} \\
=& \lim_{\varepsilon\to 0} \int_\Omega \frac{\delta F}{\delta\rho}\Big(\rho(x-\varepsilon v(x))|\mathrm{Jac}(x-\varepsilon v(x))|\Big) \\
& \cdot \Big[- v\cdot\nabla\rho(x-\varepsilon v)|\mathrm{Jac}(x-\varepsilon v)| \\
& \quad + \rho(x-\varepsilon v)\mathrm{Tr}\big(\mathrm{Jac}(x+\varepsilon v)\mathrm{Jac}(v)\big)\Big]\mathrm{d}x \\
=& \int_\Omega \frac{\delta F}{\delta\rho}(\rho(x))\big[- v\cdot\nabla\rho(x) + \rho(x)\nabla\cdot v\big]\mathrm{d}x.
\end{aligned}
$$

But this result cannot even recover the case of $F(\rho) = \mathrm{KL}(\rho||p)$! Nor can it deduce the result of Gradient Flow $v = -\nabla(\frac{\delta F}{\delta\rho})$ by $\min_\lambda \max_v G(v,\rho) + \lambda\|v\| - \lambda$ using variation calculus. Why? I would prefer that there is something wrong in the above deduction of $G(v,\rho)$.

Appendix

Compactness

- A topological space X is compact if each of its open covers has a finite subcover.
- If X is additionally a metric space, then "X is compact" is equivalent to:
    - X is sequentially compact: every sequence in X has a convergent subsequence (the limit is in X, of course).
    - X is complete and totally bounded ($\forall \varepsilon > 0$, X is a subset of the union of FINITE open balls of radius ε).
    - X is limit point compact: every infinite subset of X has at least one limit point in X.

Weak convergence of measures

Let X be a measurable space.
$\mu_n \rightharpoonup \mu$: for any bounded function $f : X \to \mathbb{R}$, $\int f \mu_n \to \int f \mu$.

Lower semicontinuity

- On a topological space X, $f : X \to \mathbb{R} \cup \{-\infty, \infty\}$ is lower semicontinuous at $x_0 \in X$ if $\forall \varepsilon > 0$, $\exists U$ a neighbourhood of x_0 s.t. $\forall x \in U, f(x) \geq f(x_0) - \varepsilon$ when $f(x_0) < +\infty$, and $\lim_{x \to x_0} f(x) = +\infty$ when $f(x_0) = +\infty$.
- In metric space, this is equivalent to $\liminf_{x \to x_0} f(x) \leq f(x_0)$.

Original notion of absolute continuity

$I = [a, b]$ is a compact interval of $\mathbb{R}$ (when I is not compact AC can also be defined, in a more general way). A function $f : I \to \mathbb{R}$ is absolutely continuous on I if there exists a Lebesgue integrable function g on I s.t. $f(x) = f(a) + \int_a^x g(t)\mathrm{d}t, \forall x \in I$.

Hölder space

- Hölder condition: on $\mathbb{R}^d$, $|f(x) - f(y)| \leq C\|x - y\|^\alpha$, with exponent α.
- Hölder space $C^{k,\alpha}(\Omega)$: functions on Ω with continuous derivatives up to order k and kth partial derivatives are Hölder continuous with exponent $0 < \alpha \leq 1$.
- The larger $\alpha > 0$ the stronger condition. So weaker than Lipschitz ($\alpha = 1$).
- Compact inclusion $C^{0,\beta}(\Omega) \to C^{0,\alpha}(\Omega)$, for $0 < \alpha < \beta \leq 1$.

Equicontinuity

Let X and Y be two metric spaces, and F a family of functions from X to Y. The family F is *equicontinuous* at a point $x_0 \in X$ if $\forall \varepsilon > 0$, $\exists \delta > 0$ s.t. $d(f(x0), f(x)) < \varepsilon, \forall f \in F, \forall x : d(x_0, x) < \delta$.

Concept	δ depends on
Continuity	ε, x_0, f
Uniform continuity	ε, f
Pointwise equicontinuity	ε, x_0
Uniform equicontinuity	ε

Ascoli-Arzelà's theorem (AA theorem)

X: a compact Hausdorff space. $C(X)$: the space of continuous functions on X.

- Typical statement: for a sequence of real-valued continuous functions $\{f_n\}_n$ on a closed and bounded interval $[a, b]$, 1) $\exists$ uniformly converging subsequence $\{f_{n_k}\}_k \Rightarrow \{f_n\}_n$ is uniformly bounded and equicontinuous; 2) every subsequence $\{f_{n_k}\}_k$ has a uniformly convergent subsequence $\Rightarrow \{f_n\}_n$ is uniformly bounded and equicontinuous.
- General statement: a subset of $C(X)$ is compact $\Leftrightarrow$ it is closed, pointwise bounded and (uniformly) equicontinuous.
- Very general statement: a subset F of $C(X)$ is relatively compact in the topology induced by the uniform norm $\Leftrightarrow$ it is equicontinuous and pointwise bounded.
- Corollary: a sequence in $C(X)$ is uniformly convergent $\Leftrightarrow$ it is (uniformly) equicontinuous and converges pointwise to a function (not necessarily continuous a-priori).

Thanks!