MAT1525 Variational Methods in Imaging and Generative Neural Networks

1 Introduction

Consider the dataset CELEBA, 200k images of size 200×200. Assume the images are grayscale, they are points in $\mathbb{R}^{40000}$. They are samples from an unknown probability measure ν. There are two questions:

1. Find an approximation of ν
2. Find a way to create/generate new samples from ν

Definition: 1.1: Push-forward Measure

Let $G : X \to Y$ be a continuous function, η be a measure on X. $(G_{\#}\eta)(B) = \eta(G^{-1}(B))$.

To generate an image, $G_\theta : \mathbb{R}^m \to \mathbb{R}^{40000}$, typically for GANs, $m = 100$. Sample from a known probability measure η (*e.g.* multivariant Gaussian). It has a push-forward measure $\mu_\theta = (G_\theta)_{\#}\eta$. We want μ_θ to be as close as possible to ν of which we have samples.

Notion of closeness of measures: GANs use Jensen-Shannon divergence $\inf_\theta JS(\mu_\theta, \nu)$, where

$$
\begin{aligned}
JS(\mu, \nu) &= \frac{1}{2} D(\mu||M) + \frac{1}{2} D(\nu||M) \\
M &= \frac{1}{2}(\mu + \nu) \\
D(\mu||M) &= \int \mu(x) \log\left(\frac{\mu(x)}{M(x)}\right) dx
\end{aligned}
$$

Issue: If μ and ν have disjoint support, then $JS(\mu, \nu) = c$ is a constant.

Definition: 1.2: Optimal Transport

Let μ be a measure on X, ν be a measure on Y, $\mu(X) = \nu(Y)$. Want to find a transport map $T : X \to Y$ with $T_{\#}\mu = \nu$ s.t. $\inf_{T:T_{\#}\mu=\nu} \int |x - Tx| d\mu(x) = W_1(\mu, \nu)$ (variational problem).

The Wasserstein distance, $W_1(\mu, \nu)$, (minimum cost of transporting μ to ν) does not have the issue, and resulted in much better training. Training $\inf_\theta W_1((G_\theta)_{\#}\eta, \nu)$ is another variational problem.

However, it is difficult to compute $W_1(\mu, \nu)$, WGANs consist of 2 neural networks. G_θ is the generator, and a neural net (the critic) is introduced for approximately computing $W_1((G_\theta)_{\#}\eta, \nu)$. The approximation is inaccurate [11]. WGAN-GP (gradient penalty) computes a congested transport cost [7].

Let x evolve according to a SDE

$$dx = f(x,t)dt + g(t)dW$$

p_t be the resulting probability distribution of $x(t)$s. $\nabla_x \log p_t(x)$ is the score. If the score is known, then we can reverse the process

$$dx = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right] dt + g(t)dW$$

In score-based diffusion model, use a neural network to approximate the score. $t \to p_t$ is a gradient flow.

Inverse Spectral Problems Consider the Dirichlet eigenfunctions

$$\begin{cases} -\Delta u = \lambda_j u_j \\ u_j|_{\partial\Omega} = 0 \end{cases}$$

Given $\lambda_1, ..., \lambda_n$, can we determine Ω? In theory, this is not achievable.

2 Wasserstein Distance

2.1 Optimal Transport Theory

Let X, Y be complete separable metric spaces (Polish spaces). Most of the time, we will have $X = Y$ bounded in $\mathbb{R}^d$. μ, ν finite Radon measures on X, Y, $\mu(X) = \nu(Y)$. Assume μ, ν are probability measures.

Monge (1781): Find a Borel measurable map $T : X \to Y$ solution of

$$\min_{T_{\#}\mu=\nu} \int_X |x - Tx| d\mu(x) \tag{1}$$

More generally, can have a cost function $c(x, y)$ continuous, $c : X \times Y \to [0, \infty)$:

$$\min_{T_{\#}\mu=\nu} \int_X c(x, Tx) d\mu(x)$$

Reasons to consider general cost functions:

1. $c(x, y) = |x - y|^p$ in $\mathbb{R}^d$. $W_p(\mu, \nu) = \left(\min_{T_{\#}\mu=\nu} \int_X |x - Tx|^p d\mu(x) \right)^{1/p}$ for $1 \leq p < \infty$. (There are WGAN papers that propose W_p-distances)
2. If $X = Y$ is a Riemannian manifold, $c(x, y) = d(x, y)^2$ is the most often used in geometry
3. WGAN-GP yield cost which is not $|x - y|$

Note: In Monge's formulation the transport map may not exist. Standard counter example of existence: Let $\mu = \delta_{x_0}$, $\nu = \frac{1}{2}(\delta_{y_0} + \delta_{y_1})$, where δ are pointmass. $X = \{x_0\}, Y = \{y_0, y_1\}$. There is no transport map $X \to Y$.

Kantorovich (1942) relaxed the problem, in lieu of a transport map T, we seek a probability measure γ on $X \times Y$. If γ has density $d\gamma = \gamma(x, y)dxdy$. Intuitively, $\gamma(x, y)$ is the fraction of the mass at x transported y. If $A \subset X, B \subset Y$, $\gamma(A \times B)$ is the mass from A to B.
γ is admissible if $(\Pi_x)_{\#}\gamma = \mu$ and $(\Pi_y)_{\#}\gamma = \nu$, where $\Pi_x(x, y) = x, \Pi_y(x, y) = y$ are projections onto corresponding spaces, so $\mu(A) = \gamma(\Pi_x^{-1}(A)) = \gamma\{(x, y) : x \in A\}$, $\nu(B) = \gamma(\Pi_y^{-1}(B)) = \gamma\{(x, y) : y \in B\}$ for $A \subset X, B \subset Y$.
If μ, ν, γ are absolutely continuous, $\mu = f(x)dx, \nu = g(y)dy, \gamma = \gamma(x, y)dxdy$, dx, dy are Lebesgue. Then $\int \gamma(x, y)dy = f(x)$, $\int \gamma(x, y)dx = g(y)$. In short, γ should have marginals μ and ν. Let $\Pi(\mu, \nu)$ be the set of admissible γ (transport plans).

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) d\gamma \tag{2}$$

Eq. 2 is the Monge-Kantorovich (MK) problem, which is a linear programming problem.

***Theorem:* 2.1:**

Assume X, Y are compact metric spaces, there is a solution to Eq. 2.

Remark 1. Solution need not be unique.

Example: $X = [0, 10] \subset \mathbb{R}, Y = [1, 11] \subset \mathbb{R}$, μ, ν are Lebesgue measures in $\mathbb{R}$, $c(x, y) = |x - y|$.
$T_1(x) = x + 1$, $W_1(\mu, \nu) = \int_X |x - T_1(x)| d\mu = 10$

$$T_2(x) = \begin{cases} x + 10, x \in [0, 1) \\ x, x \in [1, 10] \end{cases}, W_1(\mu, \nu) = \int_X |x - T_1(x)| d\mu = \int_0^1 10 dx = 10$$

2.2 Duality Theory

Key to compute $W_1(\mu, \nu)$ using a neural network is duality theory.

The dual Kandorvich (DK) problem is

$$\sup_{\phi,\psi \in C_b} \int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \text{ s.t. } \phi(x) + \psi(y) \leq c(x, y), \tag{3}$$

where C_b are continuous bounded functions.

DK has a solution and $\min \text{MK} = \max \text{DK}$.

Intuition: easier to explain:

$$\sup_{\rho,\psi} \int_X -\rho(x) d\mu(x) + \int_Y \psi(y) d\nu(y)$$

for $\rho = -\psi$. $\rho(x)$ is price we pay at source and $\psi(y)$ is price we collect at destination. The equation is to maximize the profit.
Honesty constraint: $-\rho(x) + \psi(y) \leq c(x, y)$.

Proof. (Weak Duality) $\max \text{DK} \leq \min \text{MK}$.

Let $\gamma \in \Pi(x, y)$.

$$\int_X \phi d\mu + \int_Y \psi d\mu = \int_{X \times Y} \phi d\gamma + \int_{X \times Y} \psi d\gamma \leq \int_{X \times Y} c(x, y) d\gamma$$

□

***Definition:* 2.1: c-transform**

If $\phi(x) + \psi(y) \leq c(x, y), \forall x, y$, then $\psi(y) \leq c(x, y) - \phi(x)$ and $\psi(y) \leq \inf_x(c(x, y) - \phi(x))$. Define $\phi^c(y)$ the c-transform of $\phi(x)$,

$$\phi^c(y) = \inf_x (c(x, y) - \phi(x))$$

Similarly, we define $\psi^{\bar{c}}$ by

$$\psi^{\bar{c}}(x) = \inf_y (c(x, y) - \psi(y))$$

Most often, we will have $X = Y$ and $c(x, y) = c(y, x)$, the notions will be the same.

***Definition:* 2.2: c-concave**

If $\phi = \psi^{\bar{c}}$ for some ψ, then ϕ is c-concave.

***Theorem:* 2.2:**

Assume X, Y are compact metric spaces, $c : X \times Y \to [0, \infty)$ continuous. There exists a solution of Eq. 3 with $\psi = \phi^c$. Moreover, ϕ can be taken c-concave.

$$\begin{aligned} \sup_{\psi,\psi} \left\{ \int_X \phi d\mu + \int_Y \psi d\nu : \phi(x) + \psi(y) \leq c(x, y) \right\} &= \sup_{\phi,\psi \leq \phi^c} \int_X \phi d\mu + \int_Y \psi d\nu \\ &= \sup_{\phi} \int_X \phi d\mu + \int_Y \phi^c d\nu \end{aligned}$$

***Proposition:* 2.1:**

Suppose $c(x, y)$ is a distance function on X. Then $u : X \to \mathbb{R}$ is c-concave if and only if it is 1-Lipschitz. *i.e.* $|u(x) - u(y)| \leq c(x, y)$. Moreover, for any Lip-1 function, $u^c = -u$.

Proof. ($\Rightarrow$) Assume u is c-concave, *i.e.* $u = \psi^{\overline{c}}$ for some ψ. $u(x) = \inf_y c(x, y) - \psi(y)$.

$$\begin{aligned}(c(x_1, y) - \psi(y)) - (c(x_2, y) - \psi(y)) = c(x_1, y) - c(x_2, y) &\leq c(x_1, x_2)\\ c(x_1, y) - \psi(y) &\leq c(x_1, x_2) + c(x_2, y) - \psi(y)\\ \psi^{\overline{c}}(x_1) &\leq c(x_1, x_2) + \psi^{\overline{c}}(x_2) \text{ Take inf on both sides.}\end{aligned}$$

Therefore $u(x_1) - u(x_2) \leq c(x_1, x_2)$.
Similarly, $u(x_2) - u(x_1) \leq c(x_1, x_2)$, combining both to get $|u(x_1) - u(x_2)| \leq c(x_1, x_2)$.

($\Leftarrow$) Suppose u is Lip-1.
$u(x_1) - u(x_2) \leq c(x_1, x_2) \Rightarrow u(x_1) \leq \inf_{x_2}(c(x_1, x_2) - u(x_2))$.
On the other hand, $\inf_{x_2}(c(x_1, x_2) + u(x_2)) \leq c(x_1, x_1) + u(x_1) = u(x_1)$.
Hence, $u(x_1) = \inf_{x_2}(c(x_1, x_2) + u(x_2)) = (-u)^{\overline{c}}$. Thus u is c-concave.
Apply the above to Lip-1 function $-u$, we get $-u = u^{\overline{c}} = u^c$, since c is symmetric. □

Corollary 1. $W_1(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} d(x, y) d\gamma(x, y) = \max_{u \in Lip\text{-}1} \int_X u d\mu - \int_Y u d\nu$ *(Assume $X = Y$) u is called the Kantorovich potential. This is used to approximate the Wasserstein distance.*

Recall that we want to adjust θ s.t. $\mu_\theta = (G_\theta)_\# \eta$ is as close as possible to ν.

$$\min_\theta W_1(\mu_\theta, \nu) \approx \min_\theta \max_{u \in \text{Lip-1}} \int_X u d\mu_\theta - \frac{1}{N} \sum_{i=1}^N u(x_i),$$

where $x_i \sim \nu$, and $\frac{1}{N} \sum_{i=1}^N u(x_i)$ is the empirical distribution.

Idea in the first WGAN paper: approximate a Kantorovich potential by a second neural network, $u \approx u_w$ the critic, trained (modify w) so as to maximize $W_1(\mu_\theta, \nu) = \int_X u_w d\mu_\theta + \int_Y u_w d\nu$.

However, it is difficult to train a neural network to approximate a Lip-1 function.

First attempt to produce a Lipschitz function u_w was weight clipping (cutoff weights in the neural net for u_w above a threshold)

Improved version [4] replaced the constrained optimization $u \in$ Lip-1 by a weighted unconstrained problem with a penalty term (WGAN-GP):

$$\max_u \int_X u d\mu - \int_Y u d\nu - \lambda \int_X (|\nabla u(\hat{x})| - 1)_+^2 \, d\sigma(\hat{x}),$$

where $\sigma(\hat{x})$ is sampling measure. To sample from σ, sample x from μ, y from ν and t from Unif($[0, 1]$) and let $\hat{x} = (1 - t)y + tx$.

2.3 Basic Convex Analysis

Let $\mathcal{B}$ be a Banach space and $\mathcal{B}^*$ its dual. Think of $\mathcal{B} = \mathbb{R}^N$ and $\mathcal{B}^* \cong \mathbb{R}^N$ with dual pairing $\langle x, \xi \rangle$ where ξ is a linear functional acting on x. We will have $\mathcal{B} = C(X \times Y)$ and $\mathcal{B}^*$ finite measure.

***Definition:* 2.3: Fenchel Transform**

For a functional $F : \mathcal{B} \to (-\infty, \infty]$ not necessarily convex, define its Fenchel transform (or conjugate) as

$$F^*(\beta) = \sup_{b \in \mathcal{B}} \langle \beta, b \rangle - F(b)$$

for $b \in \mathcal{B}$, $\beta \in \mathcal{B}^*$.

***Theorem:* 2.3: Properties of F^***

1. F^* is convex, because $\langle \beta, b \rangle - F(b)$ is affine in β and supremum of affine function is convex.
2. F^* is lower semi-continuous (l.s.c), $f(\overline{x}) \leq \liminf_{x \to \overline{x}} f(x)$, since it is the supremum of continuous functions
3. If $F(b) \leq G(b), \forall b$, then $F^*(b) \geq G^*(b)$, because of $-$ sign
4. $\langle \beta, b \rangle \leq F(b) + F^*(\beta)$
5. $F^{**}(b) \leq F(b), \forall b$, because $F^{**}(b) = \sup_{\beta \in \mathcal{B}^*} \langle \beta, b \rangle - F^*(\beta) \leq F(b) + F^*(\beta) - F^*(\beta) = F(b)$.

***Theorem:* 2.4: Fenchel-Moreau-Rockafellar**

Assume F is not identially ∞ (such an F is called *proper*). Then $F = F^{**}$ if and only if F is convex and l.s.c.

***Theorem:* 2.5:**

min(MK) = max(DK)

Proof. Define a functional H on $C(X \times Y)$ as

$$H(\xi) = -\max_{\phi, \psi} \left\{ \int_X \phi d\mu + \int_Y \psi d\nu : \phi(x) + \psi(y) \leq c(x, y) - \xi(x, y) \right\}$$

By Theorem 2.2, there exist maximizers for any ξ continuous.

Firstly, we show that H is convex.
Let (ϕ_0, ψ_0) solve Eq (3) for $c(x, y) - \xi_0(x, y)$ and (ϕ_1, ψ_1) solve Eq (3) for $c(x, y) - \xi_1(x, y)$.
Let $\xi_t = (1 - t)\xi_0 + t\xi_1$, $\phi_t = (1 - t)\phi_0 + t\phi_1$, $\psi_t = (1 - t)\psi_0 + t\psi_1$. Since $\phi_t + \psi_t \leq c - \xi_t$

$$\begin{aligned} H(\xi_t) &\leq -\left(\int \phi_t d\mu + \int \psi_t d\nu \right) \\ &= -\left\{ (1 - t) \left(\int \phi_0 d\mu + \int \psi_0 d\nu \right) + t \left(\int \phi_1 d\mu + \int \psi_1 d\nu \right) \right\} \\ &= (1 - t)H(\xi_0) + tH(\xi_1) \end{aligned}$$

So H is convex.

H is l.s.c. by Arzela-Ascoli Theorem

Let γ be a signed Radon measure,

$$\begin{aligned} H^*(\gamma) &= \sup_{\xi}(\langle \gamma, \xi\rangle - H(\xi)) \\ &= \sup_{\xi}\left[\int \xi d\gamma + \max_{\phi+\psi\le c-\xi}\left(\int \phi d\mu + \int \psi d\nu\right)\right] \\ &= \sup_{\phi,\psi}\sup_{\xi\le c-\phi-\psi}\int \xi d\gamma + \int \phi d\mu + \int \psi d\nu \end{aligned}$$

We will now apply Theorem 2.4 to $\mathcal{B} = C(X \times Y)$ and $\mathcal{B}^*$ finite Radon measure on $X \times Y$ to show $H^* * (\xi) = H(\xi)$ for all $\xi \in C(X \times Y)$.

Claim: $H^*(\gamma) = 0$ if γ is not a non-negative measure.
If γ is not ≥ 0, then $\exists \xi_0$ s.t. $\int \xi_0 d\gamma > 0$.
Consider $\psi = \psi = 0$, $\xi_N = N\xi_0 + c \le c = c - \phi - \psi$ for N large. Then

$$H^*(\gamma) \ge \int_{X\times Y} \xi_N d\gamma = N\int \xi_0 d\gamma + \int c d\gamma \to \infty \text{ as } N \to \infty.$$

If $\gamma \ge 0$, then taking the largest possible ξ gives

$$H^*(\gamma) = \sup_{\psi,\psi}\int_{X\times Y} c(x,y) - \phi(x) - \psi(y) d\gamma + \int_X \phi d\mu + \int_Y \psi d\nu$$

Let $\gamma_1 = (\Pi_X)_{\#}\gamma$, $\gamma_2 = (\Pi_Y)_{\#}\gamma$ be the marginals of γ, then $\int_{X\times Y}\phi(x)d\gamma = \int_X \phi d\gamma_1$ and $\int_{X\times Y}\psi(y)d\gamma = \int_Y \psi d\gamma_2$

$$\begin{aligned} H^*(\gamma) &= \sup_{\psi,\psi}\int_{X\times Y} c(x,y)d\gamma + \int_X \phi(d\mu - d\gamma_1) + \int_Y \psi(d\nu - d\gamma_2) \\ &= \begin{cases} \sup\limits_{\phi,\psi}\int_{X\times Y} c(x,y)d\gamma, & \text{if } \gamma_1 = \mu, \gamma_2 = \nu \\ \infty, & \text{otherwise} \end{cases} \end{aligned}$$

Note $\int c(x,y)d\gamma$ is independent of ϕ and ψ. Therefore:

$$H^*(\gamma) = \begin{cases} \int_{X\times Y} c(x,y)d\gamma, & \text{if } \gamma \in \Pi(\mu,\nu) \\ \infty, & \text{otherwise} \end{cases}$$

Recall definition of $H(\xi)$:

$$H(0) = -\sup_{\phi+\psi\le c}\int_X \phi d\mu + \int_Y \psi d\nu = -\max(\text{DK})$$

By Theorem 2.4, $\max(\text{DK}) = -H(0) - H^{**}(0)$. Then

$$H^{**}(0) = \sup_{\gamma}\langle \gamma, 0\rangle - H^*(\gamma) = -\inf_{\gamma} H^*(\gamma) = -\inf_{\gamma\in\Pi(\mu,\nu)}\int_{X\times Y} c(x,y)d\gamma = -\min(\text{MK})$$

Therefore, $\max(\text{DK}) = -H(0) - H^{**}(0) = \min(\text{MK})$. □

***Proposition:* 2.2:**

Suppose γ is an optimal transport plan (*i.e.* solution of Eq. (2)) and ϕ, ψ solution of Eq. 3. Then $\forall (x,y) \in \text{spt}(\gamma)$, $\phi(x) + \psi(y) = c(x,y)$.

Proof. By Theorem 2.5, also γ has the marginals μ, ν,

$$\int_{X\times Y} (\phi + \psi) d\gamma = \int_X \phi d\mu + \int_Y \psi d\nu = \int_{X\times Y} c(x,y) d\gamma$$

Therefore, $\int_{X\times Y} (c(x,y) - (\phi(x) + \psi(y))) d\gamma = 0$.
Also we require $c \geq \phi + \psi$ in Eq. 3. Thus $c = \phi + \psi$ for γ-a.e. (x,y).
Since c, ϕ, ψ are continuous functions, $\{(x,y) : c = \phi + \psi\}$ is closed. Hence it contains the smallest closed set with complement of zero γ measure. Therefore the set is $\text{spt}(\gamma)$. □

***Proposition:* 2.3:**

Suppose γ is a solution of Eq. (2) and ϕ is a Kantorovich potential, *i.e.* a solution of Eq. 3 with $\psi = \phi^c$. If $(x_0, y_0) \in \text{spt}(\gamma)$, and if $\phi(x)$ and $c(x, y_0)$ are differentiable at x_0, then $\nabla\phi(x_0) = \nabla_x c(x_0, y_0)$.

Proof. By Definition 2.1, $\phi^c(y) = \inf_x c(x,y) - \phi(x)$, so $\phi^c(y_0) \leq c(x, y_0) - \phi(x), \forall x$.
For $x = x_0$, $\phi^c(y_0) = c(x_0, y_0) - \phi(x_0)$ by Proposition 2.2, since $(x_0, y_0) \in \text{spt}(\gamma)$.
Therefore, x_0 is a minimizer of $c(x, y_0) - \phi(x)$. Hence $\nabla_x c(x, y_0) - \nabla_x \phi(x) = 0$ at $x = x_0$. □

***Definition:* 2.4: Sub-differential**

$\beta \in \mathcal{B}^*$ is called a subgradient of $F : \mathcal{B} \to (-\infty, \infty]$ (convex) at b_0 if $F(b) \geq F(b_0) + \langle \beta, b - b_0 \rangle, \forall b$.
The set of all subgradients of F at b_0 is called the subdifferential of F at b_0 and denoted

$$\partial F(b_0) = \{\beta \in \mathcal{B}^* : F(b) \geq F(b_0) + \langle \beta, b - b_0 \rangle, \forall b\}$$

Intuitively, if F is differentiable at b_0, $\beta = F'(b_0)$. RHS is the linearization at b_0.

Example: $\mathcal{B} = \mathbb{R}$, $F(b) = |b|$, $\mathcal{B}^* = \mathbb{R}$, $\beta \in \mathbb{R}$ is a subgradient at $b_0 = 0$ if and only if $|b| \geq 0 + \beta b \Leftarrow \beta \in [-1, 1]$.

$\partial(|b|)(0) = [-1, 1] \subset \mathbb{R}$, if $b_0 \neq 0$, $\partial(|b|)(b_0) = \begin{cases} 1, b_0 > 0 \\ -1, b_0 < 0 \end{cases}$.

***Theorem:* 2.6: Properties of Sub-differential**

1. F has a (global) minimum at $b_0 \Leftrightarrow 0 \in \partial F(b_0) \Leftrightarrow F(b) \geq F(b_0), \forall b$
2. $\langle \beta, b \rangle = F(b) + F^*(\beta) \Leftrightarrow \beta \in \partial F(b_0)$
3. Suppose F is convex and l.s.c. then $F^{**}(b) = F(b)$ (Theorem 2.4) and by 2, $\langle \beta, b_0 \rangle = F^{**}(b_0) + F^*(\beta) \Leftrightarrow b \in \partial F^*(\beta)$
4. If F is convex and l.s.c., $\beta \in \partial F(b_0) \Leftrightarrow b_0 \in \partial F^*(\beta_0)$.
   Hence $\partial F^*(\beta_0) = \arg\max \{\langle \beta_0, b \rangle - F(b)\}$.

Proof. (2)

$$\begin{aligned}\beta \in \partial F(b_0) &\Leftrightarrow F(b) \geq F(b_0) + \langle \beta, b - b_0 \rangle, \forall b \\ &\Leftrightarrow \langle \beta, b_0 \rangle - F(b_0) \geq \langle \beta, b \rangle - F(b), \forall b \\ &\Leftrightarrow \langle \beta, b_0 \rangle - F(b_0) \geq \sup_b \langle \beta, b \rangle - F(b) = F^*(\beta)\end{aligned}$$

Since the reverse inequality always hold, we get $\langle \beta, b_0 \rangle - F(b_0) = F^*(\beta)$. □

Apply it to $\mathcal{B} = C(X)$ with X compact and $\mathcal{B}^*$ =finite signed Radon measure with $\|\cdot\|_{L^\infty}$-norm.

Let H be the functional $\mathcal{B} \to (-\infty, \infty]$ defined as $H(\phi) = -\int_Y \phi^c d\nu, \phi^c(y) = \inf_x c(x,y) - \phi(x)$.

Claim: H is convex and l.s.c. In fact, H is continuous.

Proof.

$$\begin{aligned}\phi_1^c(y) &= \inf_x c(x,y) - \phi_1(x) = \inf_x c(x,y) - \phi_0(x) - \phi_1(x) + \phi_0(x) \\ &\leq \inf_x c(x,y) - \phi_0(x) + \|\phi_1 - \phi_0\|_{L^\infty}\end{aligned}$$

Since $\phi_1^c(y) - \phi_0^c(y) \leq \|\phi_1 - \phi_0\|_{L^\infty}$ by continuity of c-transform. Integrate both sides:

$$-H(\phi_1) + H(\phi_0) \leq \|\phi_1 - \phi_0\|_{L^\infty} \int d\gamma = \|\phi_1 - \phi_0\|_{L^\infty}$$

Interchanging ϕ_0 and ϕ_1, we get $|H(\phi_1) - H(\phi_0)| \leq \|\phi_1 - \phi_0\|_{L^\infty}$. *i.e.* H is continuous.

Convexity: Let $\phi_t = (1-t)\phi_0 + t\phi_1$

$$\begin{aligned}c(x,y) - \phi_t &= c(x,y) - (1-t)\phi_0(x) - t\phi_1(x) \\ &= (1-t)(c(x,y) - \phi_0(x)) + t(c(x,y) - \phi_1(x)) \\ &\geq (1-t)\phi_0^c(y) + t\phi_1^c(y) \\ \Rightarrow &\phi_t^c(y) \geq (1-t)\phi_0^c(x) + t\phi_1^c(y).\end{aligned}$$

Integrate w.r.t. $d\nu$, $-H(\phi_t) \geq -(1-t)H(\phi_0) - tH(\phi_1)$.
Therefore, $H(\phi_t) \leq (1-t)H(\phi_0) + tH(\phi_1)$, H is convex. □

Since F is continuous and convex, then

$$\partial H^*(\mu) = \arg\max_{\phi \in C(X)} \left\{ \int_X \phi d\mu - H(\phi) \right\} = \arg\max_{\phi \in C(X)} \int_X \phi d\mu + \int_Y \phi^c d\nu$$

Claim: $\max_{\phi \in C(X)} \int_X \phi d\mu + \int_Y \phi^c d\nu = \begin{cases} \mathcal{T}_c(\mu,\nu) = \min_\gamma \int c(x,y) d\gamma, \mu \in \gamma(X), \text{ more generally } \mu(X) = \nu(Y) \\ \infty, \text{ otherwise} \end{cases}$

Proof. If μ is not ≥ 0, then there exists $\phi \leq 0$ s.t. $\int \phi_0 d\mu > 0$.
Let $\lambda \in \mathbb{R}$, $(\lambda\phi_0)^c = \inf_x c(x,y) - \lambda\phi_0(x) \geq \inf_x c(x,y) \geq \underline{c}$, since $X \times Y$ is compact and c is continuous, $\underline{c}$ always exist. Then

$$\lambda \int_X \phi_0 d\mu + \int (\lambda\phi_0)^c d\nu \geq \lambda \int_X \phi_0 d\mu + \underline{c}$$

Since $\int_X \phi_0 d\mu > 0$, let $\lambda \to \infty$, we get ∞.

Now we need to show μ is a probability measure.
Suppose $\mu \geq 0$, but $\mu(X) \neq \nu(Y)$. Consider $(\phi + \lambda)^c = \phi^c - \lambda$. Let $\phi = 0$. $0^c(y) = \inf_x c(x, y) \geq \underline{c}$.

$$\begin{aligned}\int_X (0+\lambda)d\mu + \int_Y (0+\lambda)^c d\nu &= \lambda \int_X d\mu - \lambda \int_Y d\nu + \int_Y 0^c(y) d\nu \\ &\geq \lambda \left(\int_X d\mu - \int_Y d\nu \right) + \underline{c}\end{aligned}$$

If $\int_X d\mu - \int_Y d\nu > 0$, take $\lambda \to \infty$, and if $\int_X d\mu - \int_Y d\nu < 0$, take $\lambda \to -\infty$. Therefore, we must have $\mu(X) = \nu(Y)$. □

History of Monge Problem:

$$\inf_{T_{\#}\mu = \nu} \int_X |x - Tx| d\mu$$

T is the transport map.
Sudakov (1979) claimed existence of an optimal transport map if $\mu \ll \mathcal{L}$ (Lebesgue measure), but there was gap in the proof.
Evans & Gangbo (1999) proved using PDE methods which requires additional hypothesis.
Cafarelli-Feldman-McCann (2002) proved existence of T if $\mu \ll \mathcal{L}$ and $\nu \ll \mathcal{L}$.
Ambrosio & Pratteli relaxed to $\mu \ll \mathcal{L}$ and fixed Sudakov's proof.

2.4 Application to Wasserstein Distance

Fact: If the subdifferential at b_0 consists of a unique element, F is Gateaux differentiable at b_0.

***Definition:* 2.5: Gateaux Derivative**

Let χ be a finite Radon measure, the Gateaux derivative is

$$\liminf_{\epsilon \to 0} \frac{\mathcal{T}_c(\mu + \epsilon\chi, \nu) - \mathcal{T}_c(\mu, \nu)}{\epsilon}$$

Suppose Eq. (3) has a unique c-concave solution ϕ upto constant, then $\mathcal{T}_c(\mu, \nu) = \int_X \phi d\mu + \int_Y \phi^c d\nu$.

$$\begin{aligned}\liminf_{\epsilon \to 0} \frac{\mathcal{T}_c(\mu + \epsilon\chi, \nu) - \mathcal{T}_c(\mu, \nu)}{\epsilon} &\geq \liminf_{\epsilon \to 0} \frac{\int_X \phi d(\mu + \epsilon\chi) + \int_Y \phi^c d\nu - \int_X \phi d\mu - \int_Y \phi^c d\nu}{\epsilon} \\ &= \liminf_{\epsilon \to 0} \frac{\epsilon \int_X \phi d\chi}{\epsilon} = \int_X \phi d\chi = \langle \phi, \chi \rangle\end{aligned}$$

Also, $\limsup_{\epsilon \to 0} \frac{\mathcal{T}_c(\mu + \epsilon\chi, \nu) - \mathcal{T}_c(\mu, \nu)}{\epsilon} \leq \langle \phi, \chi \rangle$. Therefore,

$$\lim_{\epsilon \to 0} \frac{\mathcal{T}_c(\mu + \epsilon\chi, \nu) - \mathcal{T}_c(\mu, \nu)}{\epsilon} = \int_X \phi d\chi = \langle \phi, \chi \rangle$$

LHS is the directional derivative at μ in the direction χ, also called the *first variation*. The unique Kantorovich potential ϕ is the Gateaux derivative of $\mu \mapsto \mathcal{T}_c(\mu, \nu)$ with ν fixed.

Consider the case $c(x, y) = \frac{1}{2}|x - y|^2$, $X = Y = \Omega$ compact in $\mathbb{R}^d$. $\nabla_x c(x, y) = x - y$, so if $(x_0, y_0) \in \text{spt}(\gamma)$, we will have $x_0 - y_0 = \nabla\phi(x_0)$, $y_0 = x_0 - \nabla\phi(x_0)$ is unique. This gives a transport map $T(x) = x - \nabla\phi(x)$.

Special case: $c(x,y) = |x-y|$. Since this is a distance function, we know ϕ is c-concave if and only if ϕ is Lipschitz-1 and $\phi^c = -\phi$. $\phi(x) + \phi^c(y) = c(x,y) \Leftrightarrow \phi(x) - \phi(y) = |x-y|$ given $(x,y) \in \text{spt}(\gamma)$. Also if $\mu \ll \mathcal{L}_d$, there is an optimal transport map T solution of the Monge problem.

***Lemma:* 2.1:**

Let u be a Lipschitz-1 function. If $u(x) - u(y) = |x-y|$, then $u((1-t)x + ty) = u(x) - t|x-y|$ for $0 \le t \le 1$.

Proof.

$$\begin{aligned} u(x) - u((1-t)x+ty) &\le |x - (1-t)x - ty| = t|x-y| \\ u((1-t)x+ty) - u(y) &\le |(1-t)x + ty - y| = (1-t)|x-y| \\ \Rightarrow u(x) - u(y) &\le |x-y| \end{aligned}$$

But we have assumed equality, we must have equality in all equations. *i.e.* $u(x) - u((1-t)x+ty) = t|x-y|$. Also, the equality is saturated for any point inside the segment $x \to y$:

$$u((1-\tilde{t})x + \tilde{t}y) - u((1-t)x+ty) = \left|(1-\tilde{t})x + \tilde{t}y - ((1-t)x+ty)\right|$$

□

***Lemma:* 2.2:**

If u is Lipschitz-1 and $u(x) - u(y) = |x-y|$, then u is differentiable at all points $x_t = (1-t)x + ty$, $t \in (0,1)$, and $\nabla u(x_t) = \frac{x-y}{|x-y|}$. In particular, $\nabla u(x) = \frac{x-Tx}{|x-Tx|}$ μ-a.e.

The map Tx goes in the direction $-\nabla u(x)$, but we don't know how far. The lemmas also imply that tranport rays (the segments) cannot cross each other, but potentially meet at the end point.

Usefulness of transport maps: we can use it to denoise, deblur, and translate images [6].

3 Score-based Generative Models

This section explains the basic ideas of [10], [3].

The score-based generative models can generate stunning images, but the generation process is very costly by solving reverse SDEs. WGANs are more efficient and needs less training data.

3.1 Image Denoising

Let f be a gray scale noisy image, f is a function on a square $\Omega \subset \mathbb{R}^2 \to \mathbb{R}$, $f(x)$ is gray scale level at x. We want to denoise it mathemtically.

***Definition:* 3.1: Tikhonov Regularization**

$$u_0 = \arg\min_u \int_\Omega |\nabla u|^2 + \frac{\lambda}{2} \int_\Omega |u - f|^2$$

This is like a low pass filter, penalizing high frequency features.

***Definition:* 3.2: Rudin-Osher-Fatemi**

$$u_0 = \arg\min_u \int_\Omega |Du|^2 + \frac{\lambda}{2} \int_\Omega |u - f|^2,$$

where $\int_\Omega |u-f|^2$ is the fidelity term (denoised image should be close to the original image), $\int_\Omega |Du|^2$ is the regularization term.

***Definition:* 3.3: Total Variation**

$$\int_\Omega |Du| = \sup_{\|\phi\|_\infty \le 1} \left\{ \int_\Omega u \nabla \cdot \phi : \phi = (\phi_1, ..., \phi_d), \phi \in C_0^1(\Omega, \mathbb{R}^d) \right\}$$

If $u \in W^{1,1}(\Omega)$, *i.e.* $u \in L^1$ and $\int_\Omega |\nabla u| dx < \infty$, then $\int_\Omega |Du| = \int_\Omega |\nabla u|$. This definition allows Du to be a measure.

$$BV(\Omega) = \left\{ u \in L^1 : \int_\Omega |Du| < \infty \right\}$$

Because $\int_\Omega |Du|^2$ is total variation, ROF is sometimes called TV regularization.

Example: If $A \subset \Omega$ is open, $\chi_A(x) = \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases}$, then $\chi_A \in BV(\Omega)$, but $\chi_A \notin W^{1,1}(\Omega)$, because the derivative is δ on boundary and 0 a.e. else. It is not bounded in L^1. ROF model allows for such $\chi_A \to$ sharp images.

***Theorem:* 3.1:**

There exists a unique solution of Model 3.2.

Proof. Let $E(u) = \int_\Omega |Du| + \frac{\lambda}{2}\int_\Omega |u-f|^2$ for a fixed $\lambda > 0$. Assume $f \in L^2$ (in $\mathbb{R}^2$, $BV(\Omega) \subset L^2(\Omega)$)
Since $E(u) \geq 0, \forall u$, let $\delta_0 = \inf_{u \in BV(\Omega)} E(u) \geq 0$.

We want to show that $\exists u_0 \in BV(\Omega)$ with $E(u_0) = \delta_0$.

For any $n \in \mathbb{N}$, $\exists u_n$ s.t. $\delta_0 \leq E(u_n) < \delta_0 + \frac{1}{n}$, $\{u_n\}$ is called a minimizing sequence $\lim_{n\to\infty} E(u_n) = \delta_0$.
Consider $E(0)$, $\delta_0 \leq E(0) = \frac{\lambda}{2}\int_\Omega |f|^2$.
If $E(0) = \delta_0$, then done.
Otherwise, $E(0) > \delta_0$. Let n be s.t. $\frac{1}{n} < E(0) - \delta_0$, $E(u_n) \leq \delta_0 + \frac{1}{n} < E(0) = \frac{\lambda}{2}\int_\Omega |f|^2$.

$$\|u_n\|_{L^2} \leq \|u_n - f\|_{L^2} + \|f\|_{L^2} \leq \left(\frac{2}{\lambda}E(u_n)\right)^{1/2} + \|f\|_{L^2} \leq \|f\|_{L^2} + \|f\|_{L^2} = 2\|f\|_{L^2}$$

Ball in L^2 is weakly compact, so there is a subsequence $u_{n_k} \to u_0 \in L^2$.

To show that $u \in BV$ and $E(u_0) = \delta_0$, need to show $u \mapsto E(u)$ is l.s.c. with respect to weak convergence in L^2. By definition of sup, for any $\phi \in C_0^1(\Omega, \mathbb{R}^d)$:

$$\liminf_{k\to\infty} \int_\Omega |Du_{n_k}| \geq \liminf_{k\to\infty} \int_\Omega u_{n_k} \nabla\cdot\phi = \int_\Omega u_0 \nabla\cdot\phi$$

Taking sup over $\|\phi\|_\infty \leq 1$, $\liminf_{k\to\infty} \int_\Omega |Du_{n_k}| \geq \int_\Omega |Du_0|$.

$$\delta_0 = \liminf_{k\to\infty} E(u_{n_k}) \geq \liminf_{k\to\infty} \int_\Omega |Du_{n_k}| \geq \int_\Omega |Du_0|$$

This shows $u_0 \in BV(\Omega)$. Now we need to show that $u \mapsto \|u-f\|^2$ is l.s.c.

$$\begin{aligned}
\|u_0 - f\|_2 &= \sup_{\|v\|_2 \leq 1} \int_\Omega (u_0 - f)v \\
&= \sup_{\|v\|_2 \leq 1} \liminf_{k\to\infty} \int_\Omega (u_{n_k} - f)v \\
&\leq \sup_{\|v\|_2 \leq 1} \liminf_{k\to\infty} \|u_{n_k} - f\|_2 \|v\|_2 \text{ by Cauchy-Schwarz} \\
&= \liminf_{k\to\infty} \|u_{n_k} - f\|_2
\end{aligned}$$

Combining inequalities, we get

$$\delta_0 \leq E(u_0) = E(\lim_{k\to\infty} u_{n_k}) \leq \liminf_{k\to\infty} E(u_{n_k}) \leq \lim_{k\to\infty} \delta_0 + \frac{1}{n_k} = \delta_0$$

Thus $E(u_0) = \delta_0$, so u_0 solves ROF.

Uniqueness: suppose $u_1 \in BV(\Omega)$, also satisfies $E(u_1) = \delta_0$ By Parallelogram law:

$$\left\|\frac{u_0+u_1}{2} - f\right\|_2^2 = \left\|\frac{u_0-f}{2} + \frac{u_1-f}{2}\right\|_2^2 = 2\left\|\frac{u_0-f}{2}\right\|^2 + 2\left\|\frac{u_1-f}{2}\right\|^2 - \left\|\frac{u_0-u_1}{2}\right\|^2$$

$$\begin{aligned}
E\left(\frac{u_0+u_1}{2}\right) &\leq \frac{1}{2}\int_\Omega |Du_0| + \frac{1}{2}\int_\Omega |Du_1| + \frac{\lambda}{2}\left\|\frac{u_0+u_1}{2} - f\right\|_2^2 \\
&\leq \frac{1}{2}\left\{\left(\int_\Omega |Du_0| + \frac{\lambda}{2}\|u_0 - f\|^2\right) + \left(\int_\Omega |Du_1| + \frac{\lambda}{2}\|u_1 - f\|^2\right)\right\} - \frac{\lambda}{8}\|u_0 - u_1\|^2 \\
&= \frac{1}{2}E(u_0) + \frac{1}{2}E(u_1) - \frac{\lambda}{8}\|u_0 - u_1\|^2 = \delta_0 - \frac{\lambda}{8}\|u_0 - u_1\|^2
\end{aligned}$$

Therefore,

$$\delta_0 \le E\left(\frac{u_0+u_1}{2}\right) \le \delta_0 - \frac{\lambda}{8}\|u_0-u_1\|^2$$

And $u_0 = u_1$ □

3.1.1 Tadmor-Nezzar-Vese (TNV)

Given any $f \in L^2$, choose λ_0 and let u_0 be the corresponding solution of ROF.
Let $v_0 = f - u_0$, so $f = u_0 + v_0$. u_0 can be thought of as main features and v_0 may be noise.
Let $\lambda_1 = 2\lambda_0$ for simplicity. Replace f by v_0. Let u_1 be the corresponding solution of ROF:

$$u_1 = \arg\min_{u\in BV(\Omega)} \int_\Omega |Du| + \frac{\lambda}{2}\|v_0-u\|_2^2$$

Let $v_1 = v_0 - u_1$, so $v_0 = u_1 + v_1$, and $f = u_0 + u_1 + v_1$. Repeat: Having found v_{k-1}. Let

$$u_k = \arg\min_{u\in BV(\Omega)} \int_\Omega |Du| + \frac{\lambda_k}{2}\int_\Omega \|u-v_{k-1}\|_2^2$$
$$v_k = v_{k-1} - u_k$$
$$f = u_0 + u_1 + \cdots + u_k + v_k$$

This gives a nonlinear unique decomposition of f consisting of different features.

***Theorem:* 3.2: Tadmor-Nezzar-Vese**

Assume $f \in BV$ and can be generated to larger scale (*e.g.* $L^2(\Omega)$),

$$\|f\|_2^2 = \sum_{j=0}^{\infty}\left(\|u_j\|_2^2 + \frac{2}{\lambda_j}\|u_j\|_{BV}\right)$$

Registration Problem: Find ϕ a diffeomorphism to transform distribution I_0 to I_1, $\frac{\lambda}{2}\|I_1 - I_0\circ\phi\|_2^2 + d^2(\phi, e)$, where e is the identity map. ϕ can be decomposed into a sequence of diffeomorphisms.

Iterated TNV in the Denoising Direction
Consider the decomposition $f = u_0 + v_0$. In some cases, we may need to denoise only part of an image. Then we can use a weighted TV $\int_\Omega a(x)|Du|$.

$$\text{Let } \tilde{u}_1 = \arg\min\left\{\int_\Omega a(x)|Du| + \frac{\lambda_1}{2}\|u-u_0\|^2\right\} \text{ (denoise more), } \tilde{v}_1 = u_0 - \tilde{u}_1$$
$$\tilde{u}_{k+1} = \arg\min\left\{\int_\Omega a(x)|Du| + \frac{\lambda_{k+1}}{2}\|u-u_k\|^2\right\}$$
$$f = u_0 + v_0 = v_0 + \tilde{v}_1 + \tilde{u}_1 = v_0 + \tilde{v}_1 + \tilde{v}_2 + \cdots + \tilde{v}_k + \tilde{u}_k$$

We collect the noise and get clean $\tilde{u}_k$.
This is particular case of a proximal point algorithm $\tilde{u}_k \to \frac{1}{|\Omega|}\int_\Omega f$.

3.1.2 Denoising with Learned Regularizers

Suppose wee have a dataset of noisy images with distribution μ and ν. The data is not paired.

Lunz and Schoenlieb proposed to replace ROF by

$$x_0 = \arg\min_x\left\{u_0(x) + \frac{\lambda}{2}\|x-f\|^2\right\}$$

Here we are thinking of vectorized discrete images $f \in \mathbb{R}^d$, with u_0 a Kantorovich potential (Eq. (3) for $W_1(\mu,\nu)$) *i.e.*

$$u_0 = \arg\max_{u\in \text{Lip}-1} \int u d\mu - \int u d\nu$$

u_0 is large on noisy image and small on clean images.

3.2 Proximal Operators

***Definition:* 3.4: Proximal Operator**

Let X be a reflexive Banach space, $F : X \to (-\infty, \infty]$ a proper, convex l.s.c. function. Define the proximal operator by

$$\text{prox}_{\tau,F}(x) = \arg\inf_{u\in X} \left\{ F(u) + \frac{1}{2\tau}\|u-x\|_X^2 \right\}, \tau > 0$$

Definition 3.2 is a special case, with $X = L^2(\Omega)$, $x = f$, $\lambda = \frac{1}{\tau}$, $F(u) = \int |Du|$.

Another special case: $X = \mathcal{H}$ a Hilbert space, $F(x) = I_C(x) = \begin{cases} 0, x \in C \\ \infty, x \notin C \end{cases}$, where C is closed convex subset of $\mathcal{H}$. Then

$$\text{prox}_{\tau,I_C}(x) = \arg\inf_{u\in\mathcal{H}} \left\{ I_C(u) + \frac{1}{2\tau}\|u-x\|^2 \right\} = \arg\inf_{u\in C} \left\{ \frac{1}{2\tau}\|u-x\|^2 \right\} = \text{proj}_C(x_0)$$

Proximal is the generalization of projection.

***Definition:* 3.5: Moreau-Yosida Envelope/Regularization**

The Moreau-Yosida envelope/regularization of F is

$$M_{\tau,F}(x) = \inf_{u\in X} \left\{ F(u) + \frac{1}{2\tau}\|u-x\|^2 \right\}$$

Example: $X = \mathbb{R}$, $F(u) = |u|$,

$$M_{\tau,|\cdot|}(x) = \inf_{u\in\mathbb{R}} \left\{ |u| + \frac{1}{2\tau}|u-x|^2 \right\} = \begin{cases} \frac{1}{2\tau}|x|^2, |x| \le \tau \\ |x| - \frac{\tau}{2}, |x| > \tau \end{cases}$$

This is the Huber function. Note that $M_{\tau,|\cdot|}$ is differentiable, and

$$\text{prox}_{\tau,|\cdot|}(x) = \begin{cases} 0, |x| < \tau \\ x - \tau\frac{x}{|x|}, |x| > \tau \end{cases}$$

***Proposition:* 3.1: Properties of Proximal Operators**

1. $M_{\tau,F}(x)$ is convex and l.s.c., $M_{\tau,F} < \infty, \forall x \in X$
2. $\forall x \in X$, there exists a unique $x_0 = \arg\min_{u\in X} \left\{ F(u) + \frac{1}{2\tau}\|x-u\|^2 \right\}$ s.t. $M_{\tau,F}(x) = F(x_0) + \frac{1}{2\tau}\|x_0 - x\|^2$
3. $\inf_{x\in X} F(x) = \int_{x\in X} M_{\tau,F}(x)$

Proof. 2) Proof similar to Theorem 3.1

3)

$$\inf_x M_{\tau,F}(x) = \inf_x \inf_u \left\{ F(u) + \frac{1}{2\tau} \|x-u\|^2 \right\} = \inf_u \inf_x \left\{ F(u) + \frac{1}{2\tau} \|x-u\|^2 \right\} = \inf_u F(u)$$

□

Proposition: 3.2:

x^* minimizes $F \Leftrightarrow \text{prox}_{\tau,F}(x^*) = x^*$.

Proof. ($\Rightarrow$) Suppose x^* minimizes F. Then

$$F(u) + \frac{1}{2\tau} \|u-x^*\|^2 \geq F(u) \geq F(x^*)$$

Then $\inf_{u \in X} F(u) + \frac{1}{2\tau} \|u-x^*\|^2 \geq F(x^*)$, attained when $u = x^*$. Therefore,

$$\text{prox}_{\tau,F}(x^*) = \arg\inf_{u \in X} \left\{ F(u) + \frac{1}{2\tau} \|u-x^*\|_X^2 \right\} = x^*$$

($\Leftarrow$) If $x^* = \arg\inf_{u \in X} \left\{ F(u) + \frac{1}{2\tau} \|u-x^*\|^2 \right\}$, then

$$0 \in \partial \left\{ F(u) + \frac{1}{2\tau} \|u-x^*\|^2 \right\} (x^*) = \partial F(x^*) + \frac{1}{\tau}(x^*-x^*)$$

So $0 \in \partial F(x^*)$. This implies that x^* minimizes F. □

To look for minimizers of F, we look for fixed points of $\text{prox}_{\tau,F}$,

$$x_n = \text{prox}_{\tau,F}(x_{n-1})$$

This is the *Proximal Point Algorithm.*

$\text{prox}_{\tau,F}$ is, in general, not a contraction map, but it is firmly non-expansive.

Definition: 3.6: Firmly Non-expansive

An operator T is firmly non-expansive if

$$\|Tx-Ty\|^2 + \|(I-T)x-(I-T)y\|^2 \leq \|x-y\|^2$$

Recall a contraction is $\|Tx-Ty\| \leq C\|x-y\|$ for $C < 1$.

Proposition: 3.3:

T is firmly non-expansive $\Leftrightarrow$ T is resolvent of a maximally monotone operator O.

$$x_0 = \text{prox}_{\tau,F}(x) = \arg\inf_{u \in X} \left\{ F(u) + \frac{1}{2\tau} \|u-x\|^2 \right\} \Leftrightarrow 0 \in \partial F(x_0) + \frac{1}{2\tau}(x_0-x)$$
$$\Leftrightarrow x \in (I + 2\tau\partial F)(x_0), x_0 = (I+2\tau\partial F)^{-1}(x)$$

***Proposition:* 3.4:**

∂F is maximally monotone.

Therefore, $\text{prox}_{\tau,F} = (I + 2\tau\partial F)^{-1}$. Together, this gives the convergence $x_n \to x^*$. $x_n = \text{prox}_{\tau,F}(x_{n-1})$ generalizes iterated denoising:

$$u_m = \arg\inf \left\{ \int_\Omega |Du| + \frac{\lambda}{2} \|u - u_{n-1}\|^2 \right\}, u_{-1} = f$$

so u_n converges to the minimizer of $\int_\Omega |Du| + \frac{\lambda}{2} \|u - u_{n-1}\|^2$, which is $\frac{1}{|\Omega|} \int_\Omega f$.

3.3 Gradient Flows

Heuristics on Gradient Descent: Consider $x_n = x_{n-1} - \tau \nabla F(x_{n-1})$, where τ is the step size. Rearranging, $\frac{x_n - x_{n-1}}{\tau} = -\nabla F(x_{n-1})$ is the Euler discretization. The continuous limit is $\frac{dx}{dt} = -\nabla F(x(t))$. This is a gradient flow in $\mathbb{R}^d$.

For proximal point algorithm, $x_n = \text{prox}_{\tau,F}(x_{n-1})$. Since $0 \in \partial F(x_0) + \frac{1}{2\tau}(x_0 - x)$, we have $0 \in \partial F(x_n) + \frac{1}{2\tau}(x_n - x_{n-1})$, $x_n = x_{n-1} - \tau \partial F(x_n)$. This is the implicit Euler discretization of the same gradient flow, but with better convergence property. That is why proximal operators come into play when discussing gradient flows.

Di Giorgi et. al. wanted to define gradient flows in metric spaces like Wasserstein space.

***Definition:* 3.7: Minimizing Movements Scheme**

Let τ be step size. $x_\tau^n = \arg\min_x \left\{ F(x) + \frac{1}{2\tau} d^2(x, x_\tau^{n-1}) \right\}$. Define $\overline{x}_\tau(t) = x_\tau^n$ on $((n-1)\tau, n\tau]$ a piecewise constant interpolant. We say $x(t)$ is a minimizing movement of F if there is a family of $\overline{x}_\tau(t) \to x(t)$ for every t as $\tau \to \infty$.

Now suppose we are on a Riemannian manifold M^d. Gradient flow of $F : M^d \to \mathbb{R}$ is a solution $u : \mathbb{R} \to M^d$ of $\frac{du}{dt} = -\nabla F(u(t)), u(0) = u_0$. $\frac{du}{dt}$ is the velocity vector at $u(t)$, an element in the tagent space ($\frac{du}{dt} \in T_{u(t)} M^d$). ∇F is defined ad $dF(Y) = g(\nabla F, Y)$, where g is a metric on M^d (inner product on tangent space), $Y \in T_{u(t)}$, and $dF(Y)$ is the differential form.

For gradient flows in Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, probability measures in $\mathbb{R}^d$, with W_2 metric, we will need to make sense of

1. Velocity vector field of a flow $t \mapsto \mu_t$
2. Tangent space $T_{\mu(t)} \mathcal{P}_2$
3. Notion (and computations) of gradient of a functional F on $\mathcal{P}_2(\Omega)$
4. Conditions on F for convergence of a minimizing movement scheme.

3.4 Wasserstein-2 Space

***Definition:* 3.8: Hopf-Lax Formula**

Consider Hamilton-Jacobi equation

$$\begin{cases} \frac{\partial u}{\partial t} + H(\nabla u) = 0 \\ u(x, t = 0) = g(x) \end{cases}$$

It has solution:

$$u(x,t) = \min_y \left\{ g(y) + tH^* \left(\frac{x-y}{t} \right) \right\},$$

where $H^*(q) = L(q) = \sup_p \{\langle q, p\rangle - H(p)\}$, the Legendre-Fenchel transform.

In particular, for $H(p) = \frac{1}{2}|p|^2$, $L(q) = \frac{1}{2}|q|^2$ (the dual of L^2-norm is L^2-norm). Thus, $\begin{cases} \frac{\partial u}{\partial t} + \frac{1}{2}|\nabla u|^2 = 0 \\ u(x, t = 0) = g(x) \end{cases}$ has solution

$$u(x,t) = \min_y \left\{ g(y) + \frac{1}{2t}|x-y|^2 \right\} = M_{t,g}(x)$$

Recall the Wasserstein distance

$$W_2(\mu,\nu) = \min_{\gamma \in \Pi(\mu,\nu)} \int_{X\times Y} \frac{1}{2}|x-y|^2 d\gamma(x,y) = \max_{\phi \in C(\Omega)} \int_X \phi d\mu + \int_Y \phi^c d\nu$$

We know that there is an optimal plan γ, solution of MK and a c-concave ϕ solution of DK.

Claim: For $c(x,y) = \frac{1}{2}|x-y|^2$ and $\mu \ll \mathcal{L}_d$ (Lebesgue measure on $\mathbb{R}^d$), there is a unique optimal transport map T s.t. $Tx = \nabla h$ for a convex function h.

Proof. When $c(x,y) = \frac{1}{2}|x-y|^2$. Let $u_\phi(x) = \frac{1}{2}|x|^2 - \phi(x)$,

$$\begin{aligned} \phi^c(y) &= \inf_x \left\{ \frac{1}{2}|x-y|^2 - \phi(x) \right\} = \inf_x \left\{ \frac{1}{2}|x|^2 - x\cdot y + \frac{1}{2}|y|^2 - \phi(x) \right\} \\ &= \frac{1}{2}|y|^2 + \inf_x \{-x\cdot y + u_\phi(x)\} \\ &= \frac{1}{2}|y|^2 - \sup_x \{x\cdot y - u_\phi(x)\} \\ &= \frac{1}{2}|y|^2 - u_\phi^*(y) \end{aligned}$$

Similarly, $\psi^{\bar{c}}(x) = \inf_y \left\{ \frac{1}{2}|x-y|^2 - \phi(y) \right\} = \frac{1}{2}|x|^2 - u_\psi^*(x)$.

ϕ is c-concave if $\phi = \psi^{\bar{c}}$ for some ψ, *i.e.*, $\phi(x) = \frac{1}{2}|x|^2 - u_\psi^*(x) \Leftrightarrow \frac{1}{2}|x|^2 - \phi(x) = u_\psi^*(x)$.
Because Fenchel transform is convex and l.s.c., then $\frac{1}{2}|x|^2 - \phi(x)$ is convex and l.s.c.

Let γ be a solution of MK and ϕ a c-concave Kantorovich potential. By Proposition 2.2 and 2.3, $c(x,y_0) - \phi(x)$ is minimal at x_0, and if ϕ is differentiable at x_0, $\nabla\phi(x_0) = \nabla_x \left(\frac{1}{2}|x_0 - y_0|^2\right) = x_0 - y_0$, $y_0 := Tx_0 = x_0 - \nabla\phi(x_0) = \nabla\left(\frac{1}{2}|x|^2 - \phi\right)(x_0)$. $h(x) = \frac{1}{2}|x|^2 - \phi(x)$ is convex.

Assumptions yield ϕ is differentiable a.e. (locally Lipschitz). □

Theorem: 3.3:

For Ω not necessarily bounded, say $\Omega = \mathbb{R}^d$. Let μ, ν be probability measures on $\mathbb{R}^d$. Suppose $\mu \ll \mathcal{L}_d$, $\int |x|^2 d\mu < \infty$ and $\int |y|^2$. Then there exists a unique optimal transport map T and it is of the form $T = \nabla h$ with h convex. The set of such measures are $\mathcal{P}_2(\Omega)$.

Theorem: 3.4:

$W_2(\mu, \nu)$ is a distance on $\mathcal{P}_2(\Omega)$, also true for $(\mathcal{P}_p(\Omega), W_p)$ for $1 \leq p < \infty$.

Definition: 3.9: Continuity Equation

Let ρ_t be the probability density, v_t be the flow velocity vector field, then

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$$

We say (ρ_t, v_t) solve the continuity equation, if ρ_t is a family of measures and v_t is time-dependent vector feild s.t.

$$\int_0^T \|v_t\|_{L^1(\rho_t)} \, dt = \int_0^T \int_\Omega |v_t| d\rho_t dt < \infty$$

The equation is satisfied in the distribution sense, *i.e.*

$$\int_0^T \int_\Omega (\partial_t \phi) d\rho_t dt + \int_0^T \int_\Omega \nabla \phi \cdot v_t d\rho_t dt = 0, \forall \phi \in C_C^1([0,T] \times \overline{\Omega}) \tag{4}$$

If $\phi(t,x)$ is supported in the interior $(0,T) \times \Omega$, then

$$\int_0^T \int_\Omega (\partial t \phi) d\rho_t dt - \int_\Omega \phi \nabla \cdot v_t = 0$$
$$\int \nabla \phi \cdot v_t = \int \nabla \cdot (\phi v_t) - \int \phi \nabla \cdot v_t = \int \phi v_t \cdot n = 0,$$

where n is the outward normal vector of Ω, so $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ in the interior.

Now allow ϕ non-zero on $\partial\Omega$, but $\int_{\partial\Omega} \phi v_t \cdot n = 0$. This shows that Eq. (4) in the sense of distribution includes the Neumann boundary condition $v_t \cdot n = 0$, and the entire flow is contained in Ω.

Another way to interpret Eq. (4) in the weak sense:

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

The two notions are essentially equivalent.

Definition: 3.10: Lagrangian Coordinates

Assume v_t is Lipschitz, uniformly in t.

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

***Theorem:* 3.5:**

$Y_t(x) = y_x(t) \in \Omega$ in view of Neumann conditions on boundary, where $y_x(t)$ is the position at time t if we start from x, and $Y_t(x)$ is the transformation of x at time t. Then for any $\rho_0 \in \mathcal{P}(\Omega)$, the 1-parameter family of measures $\rho_t = (Y_t)_\#(\rho_0)$ solves the continuity equation in Definition 3.9 with $\rho_t|_{t=0} = \rho_0$. Moreover, the equation admits a unique solution. [9]

***Definition:* 3.11: Metric Derivative**

Let (X, d) be a metric space, suppose $\omega : [0,1] \to X$ is Lipschitz, *i.e.* $d(w(t_1), w(t_2)) \leq \rho \|t_1 - t_2\|$, $\omega'(t)$ need not have a meaning (if X is not a vector space). However, we can define metric derivative:

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

***Theorem:* 3.6:**

Suppose $\omega : [0,1] \to X$ is a Lipschitz continuous curve, then $|\omega'(t)|$ exists a.e. Moreover, for $t_1 < t_2$,

$$d(\omega(t_1), \omega(t_2)) \leq \int_{t_1}^{t_2} |\omega'(s)| ds$$

***Definition:* 3.12: Absolute Continuous**

A curve $\omega : [0,1] \to X$ is absolutely continuous (AC) if $\exists g \in L^1([0,1])$ s.t. $d(\omega(t_1), \omega(t_2)) \leq \int_{t_1}^{t_2} g(s) ds$ for every $t_1 < t_2$, $g \geq 0$ a.e.

In particular, any Lipschitz continuous curves are absolutely continuous. Moreover, we can reparametrize an AC curve to make it Lipschitz continuous.

Let $G(t) = \int_0^t g(s) ds$, $S_\epsilon(t) = \epsilon t + G(t)$, S_ϵ is strictly increasing with $\epsilon \nearrow 0$. Then $\tilde{\omega}(t) = \omega(S_\epsilon^{-1}(t))$ is Lipschitz, so ω is AC, ω has a metric derivative.

***Theorem:* 3.7:**

Let $\Omega \subset \mathbb{R}^d$ compact. If Ω is unbounded, need to assume finite second moments $\int |x|^2 d\mu < \infty$. $\mathcal{P}(\Omega) = \{\mu : \mu \text{ a probability measure}\}$, $\mathcal{P}_2(\Omega)$ is $\mathcal{P}(\Omega)$ with $W_2(\mu, \nu)$ metric. Then $W_2(\mu, \nu)$ satisfies the properties of a distance function.

***Theorem:* 3.8:**

From [1].

1. Let $\{\mu_t\}_{t \in [0,1]}$ be an absolutely continous curve in $\mathcal{P}_p$ ($\mathcal{P}(\Omega)$ with W_p metric). Then for a.e. $t \in [0,1]$, there is a vector field $v_t \in L^p(\mu_t; \mathbb{R}^d)$ s.t.
   (a) $\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0$
   (b) $\|v_t\|_{L^p(\mu_t)} \leq |\mu'(t)|$
2. If we are given v_t and μ_t with $v_t \in L^p(\mu_t; \mathbb{R}^d)$, $\int_0^1 \|v_t\|_{L^p(\mu_t)} dt < \infty$ solving $\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0$. Then μ_t is absolutely continous in $\mathcal{P}_p(\Omega)$ and $|\mu'(t)| \leq \|v_t\|_{L^p(\mu_t)}$.

Combining the above, $\|v_t\|_{L^p(\mu_t)} = |\mu'(t)|$.

Definition: 3.13: Curve Length

Let (X, d) be a metric space, $\omega : [0, 1] \to X$, then

$$\text{length}(\omega) = \sup \left\{ \sum_{k=1}^{n} d(\omega(t_{k-1}), \omega(t_k)), 0 = t_0 < \cdots < t_n = 1 \right\}$$

Proposition: 3.5:

For any $\omega \in$AC, $\text{length}(\omega) = \int_0^1 |\omega'(t)|dt$

Definition: 3.14: Geodesic

A curve ω is a geodesic from $x_0 \in X$ to $x_1 \in X$ if it minimizes length among all curves with $\omega(0) = x_0$, and $\omega(1) = x_1$.
A curve ω is a constant speed geodesic if $d(\omega(t), \omega(s)) = |t - s|d(\omega(0), \omega(1))$.

Theorem: 3.9:

The following are equivalent

1. ω is a constant speed geodesic
2. $\omega \in \text{AC}(X)$ and $|\omega'(t)| = d(\omega(0), \omega(1))$
3. ω is a solution of

$$\min \left\{ \int_0^1 |\omega'(t)|^p dt, \omega(0) = x_0, \omega(1) = x_1 \right\} = \int_0^1 d(\omega(0), \omega(1))^p dt = d(\omega(0), \omega(1))^p$$

Definition: 3.15: Geodesic Space

(X, d) is a geodesic space if

$$d(x_0, x_1) = \min \{\text{length}(\omega) : \omega \in \text{AC}, \omega(0) = x_0, \omega(1) = x_1\}$$

Theorem: 3.10:

If Ω is convex, then $\mathcal{P}_2(\Omega)$ ($\mathcal{P}_p(\Omega)$ for any $p \geq 1$) is a geodesic space.

Proof. Given μ, ν, we need to produce a geodesic from μ to ν.
More precisely, let $\mu, \nu \in \mathcal{P}_p(\Omega)$ and γ be an optimal plan for the cost $c(x, y) = |x - y|^p$, *i.e.*

$$W_p^p(\mu, \nu) = \int_{\Omega \times \Omega} |x - y|^p d\gamma(x, y), \gamma \in \Pi(\mu, \nu)$$

Define $\Pi_t : \Omega \times \Omega \to \Omega$ by $\Pi_t(x, y) = (1 - t)x + ty$. Then $\mu_t = (\Pi_t)_{\#}\gamma$ is a constant speed geodesic from μ to ν. If γ is given by an optimal map T (*e.g.* when $\mu \ll \mathcal{L}_d$, $\Omega \subset \mathbb{R}^d$), then $\mu_t = ((1 - t)Id + tT)_{\#}\mu$. When $t = 1, \mu_1 = \nu$. μ_t is probability distribution at time t. This is also called *displacement interpolation*. $\mu_t = (\Pi_t)_{\#}\gamma$ says:

$$\int_\Omega \phi(x) d\mu_t = \int_{\Omega \times \Omega} \phi(\Pi_t(x, y)) d\gamma(x, y) = \int_{\Omega \times \Omega} \phi((1 - t)x + ty) d\gamma(x, y)$$

Then we need to show that $t \mapsto \mu_t$ is a geodesic from μ to ν. For it to be a geodesic, we need $W_p(\mu_t, \mu_s) = |t-s|W_p(\mu,\nu)$ (asusming $s > t$)

It sufficies to show $W_p(\mu_t, \mu_s) \leq |t-s|W_p(\mu,\nu)$, because in this case:

$$\begin{aligned} W_p(\mu,\nu) &\leq W_p(\mu,\mu_t) + W_p(\mu_t,\mu_s) + W_p(\mu_s,\nu) \\ &\leq tW_p(\mu,\nu) + (s-t)W_p(\mu,\nu) + (1-s)W_p(\mu,\nu) = W_p(\mu,\nu) \end{aligned}$$

and this gives the equality.

Now we show the claim $W_p(\mu_t, \mu_s) \leq |t-s|W_p(\mu,\nu)$.
Let $\gamma_t^s = (\Pi_t, \Pi_s)_{\#}\gamma$, where $(\Pi_t, \Pi_s) : \Omega \times \Omega \to \Omega \times \Omega$. Then

$$\int_{\Omega\times\Omega} \psi(x,y) d\gamma_t^s = \int_{\Omega\times\Omega} \psi((1-t)x+ty, (1-s)x+sy) d\gamma(x,y)$$

Claim: $\gamma_t^s \in \Pi(\mu_t, \mu_s)$, *i.e.* it is a marginal from μ_t to μ_s. Consider the marginal:

$$\int_{\Omega\times\Omega} \phi(x) d\gamma_t^s = \int_{\Omega\times\Omega} \phi((1-t)x+ty) d\gamma(x,y) = \int_\Omega \phi d\mu_t$$

Similarly, the other marginal is μ_s.

Since $W_p(\mu_t, \mu_s)$ is infimum,

$$\begin{aligned} W_p(\mu_t,\mu_s) &\leq \left(\int_{\Omega\times\Omega} |x-y|^p d\gamma_t^s \right)^{1/p} \\ &= \left(\int_{\Omega\times\Omega} |(1-t)x+ty-(1-s)x-sy|^p d\gamma \right)^{1/p} \\ &= \left(\int_{\Omega\times\Omega} |(s-t)(x-y)|^p d\gamma \right)^{1/p} \\ &= |s-t| \int_{\Omega\times\Omega} |x-y|^p d\gamma \\ &= |s-t| W_p(\mu,\nu), \text{ since } \gamma \text{ is an optimal plan.} \end{aligned}$$

Consider the case when γ is given by a transport map T, *i.e.*

$$\gamma = (id, T)_{\#}\mu \Leftrightarrow \int \psi(x,y) d\gamma = \int \psi(x, Tx) d\mu$$

In this case:

$$\int \phi(x) d\mu_t = \int_{\Omega\times\Omega} \phi((1-t)x+ty) d\gamma(x,y) = \int_\Omega \phi((1-t)x+tTx) d\mu$$

This shows that $\mu_t = ((1-t)id + tT)_{\#}\mu$. □

Theorem: 3.11: Benamou-Brenier Formula

Assume Ω is convex and compact $\mathbb{R}^d$,

$$W_p^p(\mu,\nu) = \min_{\rho_t, v_t} \left\{ \int_0^1 \|v_t\|_{L^p(\rho_t)}^p \, dt : \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \rho_0 = \mu, \rho_1 = \nu \right\}$$

Proof. Intuition: $p = 2$. Let v_t be a nice vector field. Solve $\frac{dy_x(t)}{dt} = v(t, y_x(t)), y_x(0) = 0, Y_t(x) = y_t(x)$. Let $\rho_t = (Y_t)_{\#}\rho_0$, so $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$. Assume $\rho_1 = \nu$.

$$\begin{aligned}
W_2^2(\rho, \nu) &\leq \int_\Omega |y_x(1) - x|^2 d\mu = \int_\Omega |y_x(1) - y_x(0)|^2 d\mu \\
&= \int_\Omega \left| \int_0^1 \frac{dy_x(t)}{dt} dt \right|^2 d\mu \\
&\leq \int_\Omega \int_0^1 \left| \frac{dy_x(t)}{dt} \right|^2 dt d\mu \\
&= \int_0^1 \int_\Omega \left| \frac{dy_x(t)}{dt} \right|^2 d\mu dt \\
&= \int_0^1 \int_\Omega |v_t|^2 \, d\mu dt = \int_0^1 \|v_t\|^2_{L^2(\mu_t)} \, dt
\end{aligned}$$

Since $\mathcal{P}_p(\Omega)$ is a geodesic space, there is a geodesic ρ_t from μ to ν, which we may take to be constant speed. Then

$$\begin{aligned}
W_p^p(\mu, \nu) &= \min \left\{ \int_0^1 |\rho'(t)|^p dt, \rho_0 = \mu, \rho_1 = \nu \right\} \\
&= \min_{\rho_t, v_t} \left\{ \int_0^1 \|v_t\|^p_{L^p(\rho_t)} \, dt : \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \rho_0 = \mu, \rho_1 = \nu \right\}
\end{aligned}$$

□

We can change to a more general interval $[0, \tau]$:

$$W_2^2(\rho_0, \rho_\tau) = \tau \min_{v, \rho} \left\{ \int_0^\tau \|v_t\|^2_{L^2(\rho_t)} \, dt : \partial_t \rho + \nabla \cdot (\rho v) = 0, \rho(0, \cdot) = \rho_0, \rho(\tau, \cdot) = \rho_\tau \right\}$$

Informally, suppose v is any reasonable vector field, then $\partial_t \rho + \nabla \cdot (\rho v) = 0$, $\rho(0, \cdot) = \rho_0$ yields $\rho(\tau, \cdot) = \rho_\tau$. Let $y_x(t)$ solve $\frac{dy_x}{dt} = v(t, y_x(t))$, $y_x(0) = x$. Let $Y_t(x) = y_x(t)$, $\rho_t = (Y_t)_{\#}\rho_0$ satisfies the continuity equation, so

$$W_2^2(\rho_0, \rho_\tau) \leq \int |y_x(\tau) - x|^2 d\rho_0,$$

since the map $Y_\tau : x \mapsto y_x(\tau)$, satisfies $(Y_\tau)_{\#}\rho_0 = \rho_\tau$.

$$\begin{aligned}
\int |y_x(\tau) - x|^2 d\rho_0 &= \int_\Omega \left| \int_0^\tau \frac{dy}{dt}(t, x) dt \right|^2 d\rho_0 \text{ (By FTC)} \\
&= \tau^2 \int_\Omega \left| \frac{1}{\tau} \int_0^\tau \frac{dy}{dt}(t, x) dt \right|^2 d\rho_0 \\
&\leq \tau^2 \int \frac{1}{\tau} \int_0^\tau \left| \frac{dy}{dt}(t, x) \right|^2 d\rho_0 \text{ By Jensens' inequality with } f = \frac{dy}{dt}, \phi(x) = x^2 \\
&= \tau \int_0^\tau \int_\Omega |v(t, y_x(t))|^2 d\rho_0 \\
&= \tau \int_0^\tau \int_\Omega |v(t, x)|^2 d\rho_t, \text{ since } \rho_t = (Y_t)_{\#}\rho_0 \\
&= \tau \int_0^\tau \|v(t, \cdot)\|^2_{L^2(\rho_t)} \, dt
\end{aligned}$$

When ρ_t is a constant speed geodesic from ρ_0 to ρ, we get equality.

Assume $\rho_0 \ll \mathcal{L}_d$, there is an optimal map T, and $\rho_t = \left(x + \frac{t}{\tau}(Tx - x)\right)_{\#} \rho_0$.
Let $Y_t(x) = x + \frac{t}{\tau}(Tx - x)$, $y_x(t) = x + \frac{t}{\tau}(Tx - x)$, $\frac{dy_x}{dt} = \frac{Tx - x}{\tau} = v(x)$ independent of t.

Recall that proximal operators yield implicit Euler discretization of gradient flows. Suppose $F : \mathcal{P}_2(\Omega) \to (-\infty, \infty]$,

$$\text{Wprox}_{\tau,F}(\rho_0) = \arg\min_{\rho} \left\{ F(\rho) + \frac{W_2^2(\rho, \rho_0)}{2\tau} \right\}$$

By Definition 3.7, as $\tau \to 0$, if convergent, we have a flow $t \mapsto \rho_t$ in $\mathcal{P}_2(\Omega)$, which is the gradient flow of F w.r.t. Wasserstein distance.

Question: Given F, how can we find v?

We can consider two formulations: Definition 3.7 and Theorem 3.8.

A few additional basic facts about $\mathcal{P}_p(\Omega)$:

1. If Ω is compact, $p \in [1, \infty)$, $W_p(\mu_n, \mu) \to 0 \Leftrightarrow \mu_n \to \mu$ weakly. *i.e.* $\int \phi d\mu_n \to \int \phi d\mu$ for ϕ continuous.
2. W_p is separable: finitely supported measures with rational weights on a dense subset of Ω
3. If Ω is compact, $\mathcal{P}_p(\Omega)$ is also compact w.r.t. topology induced by the metric.

Using these facts, we can show that $\rho_{k+1}^\tau = \arg\min_{\rho} \left\{ F(\rho) + \frac{W_2^2(\rho, \rho_k^\tau)}{2\tau} \right\}$ is solvable if $F(\rho)$ is l.s.c. w.r.t. topology induced by the metric and bounded below.

Optimality Condition (for Definition 3.7):
How do we define $\frac{\delta F}{\delta \rho}$ if exists for $F : \mathcal{P}_2(\Omega) \to (-\infty, \infty]$?

***Definition:* 3.16: Regular**

We say ρ is regular for F if $F((1-\epsilon)\rho + \epsilon\tilde{\rho}) < \infty$ for every $\epsilon \in [0, 1]$ and any $\tilde{\rho} \in \mathcal{P}(\Omega) \cap L_C^\infty(\Omega)$ absolutely continuous.

***Definition:* 3.17: First Variation**

If ρ is regular for F, we define $\frac{\delta F}{\delta \rho}(\rho)$ (first variation) if exists to be a measurable function s.t.

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} F(\rho + \epsilon\chi) = \int \frac{\delta F}{\delta \rho}(\rho) d\chi$$

for all $\chi = \tilde{\rho} - \rho$, $\tilde{\rho} \in \mathcal{P}(\Omega) \cap L_C^\infty(\Omega)$, so $\rho + \epsilon\chi = (1-\epsilon)\rho + \epsilon\tilde{\rho}$.

Remark 2. $\int d\chi = \int d\tilde{\rho} - \int d\rho = 1 - 1 = 0$, so $\frac{\delta F}{\delta \rho}$ is only defined upto constants.

Therefore, solution of Definition 3.7 satisfies:

$$\frac{\delta F}{\delta \rho}(\rho) + \frac{1}{2\tau} \frac{\delta W_2^2(\rho, \rho_k^\tau)}{\delta \rho} = \text{constant}$$

Assume the c-concave Kantorovich potential ϕ is unique upto constants. (True if e.g. $\text{supp}(\rho) = \Omega$)

Recall the relation of W_2^2 to be the dual problem:

$$W_2^2(\rho,\nu) = 2\max_{\phi} \int \phi d\rho + \int \phi^c d\nu$$

Let ϕ be the minimizer, then

$$\begin{aligned}
\frac{\delta W_2^2}{\delta \rho}(\rho,\nu) &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} W_2^2(\rho+\epsilon\chi,\nu) \\
&\geq \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} 2\left(\int \phi d(\rho+\epsilon\chi) + \int \phi^c d\nu\right) \\
&= 2\int \phi d\chi
\end{aligned}$$

We can in fact show that the equality holds [9]. $\frac{\delta W_2^2}{\delta \rho}(\rho,\nu) = 2\phi$, where ϕ is the Kantorovich potential from ρ to ν. So the optimality condition is:

$$\frac{\delta F}{\delta \rho}(\rho) + \frac{1}{\tau}\phi = \text{constant}$$

Differentiate both size:

$$\nabla_x \frac{\delta F}{\delta \rho}(\rho) + \frac{1}{\tau}\nabla\phi = 0$$

Recall for $c(x,y) = \frac{1}{2}|x-y|^2$, we proved that there is an optimal transportmap T s.t. $Tx = x - \nabla\phi(x)$ in the beginning of this section, so

$$\nabla_x \frac{\delta F}{\delta \rho} = -\frac{1}{\tau}\nabla\phi(x) = \frac{1}{\tau}(Tx - x)$$

gives an optimal transport from ρ_{k+1}^τ to ρ_k^τ.

As $\tau \to 0$, we expect $\nabla_x \frac{\delta F}{\delta \rho} \to -v$, the velocity field. Therefore, $v = -\nabla_x \frac{\delta F}{\delta \rho}$.

From the continuity condition, we see that the gradient flow of $F(\rho)$ can be derived from the PDE:

$$\rho_t - \nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho}\right) = 0$$

Example: Negative entropy: $F(\rho) = \int \rho \log \rho$. More rigorously, $F(\rho) = \begin{cases} \int \rho \log \rho, \text{ if } \rho \ll \mathcal{L}^d \\ \infty, \text{ otherwise} \end{cases}$.

$$\begin{aligned}
\frac{\delta F}{\delta \rho} &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int (\rho+\epsilon\chi)\log(\rho+\epsilon\chi) \\
&= \int \chi \log(\rho+\epsilon\chi) + (\rho+\epsilon\chi)\frac{\chi}{\rho+\epsilon\chi}|_{\epsilon=0} \\
&= \int \chi \log\rho + \rho\frac{\chi}{\rho} = \int (\log\rho + 1)\chi \\
&= \log\rho + 1
\end{aligned}$$

The continuity equation becomes the heat equation:

$$\begin{aligned}
\rho_t - \nabla\cdot(\rho\nabla(\log\rho+1)) &= 0 \\
\rho_t - \nabla\cdot\left(\rho\frac{\nabla\rho}{\rho}\right) &= 0 \\
\rho_t &= \Delta\rho
\end{aligned}$$

Example: Kullback-Leibler divergence (relative entropy)

$$\begin{aligned} F(\rho) = KL(\rho||\pi) &= \int \rho \log\left(\frac{\rho}{\pi}\right) = \int \rho \log \rho - \rho \log \pi \\ \frac{\delta F}{\delta \rho} &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int (\rho + \epsilon\chi) \log(\rho + \epsilon\chi) - (\rho + \epsilon\chi) \log \pi \\ &= \log \rho + 1 - \log \pi = \log\left(\frac{\rho}{\pi}\right) + \text{constant} \end{aligned}$$

The continuity equation becomes Fokker-Planck equation:

$$\rho_t = \nabla \cdot \left(\rho \nabla \log\left(\frac{\rho}{\pi}\right)\right)$$

These provides the foundation of *forward process* of score-based diffusion models [3]. For the *backward process*, we need to sample from Gaussian noise and solve reversed diffusion equation with fine-grained step size. However, if we can properly treat the reversed diffusion with Wasserstein distance, we can use a larger step.

4 Congested Transport

Recall the dual Kantorovich problem (3). WGAN uses a neural network to compute an approximation of u (critic). However Lip-1 is hard to enforce. WGAN-GP applies a penalty term:

$$\sup \int u(d\mu - d\nu) - \frac{\lambda}{2} \int_\Omega (|\nabla u| - 1)_+^2 \sigma(\hat{x}) d\hat{x},$$

where $\sigma(\hat{x})$ is the sampling density. This still does not compute $W_1(\mu, \nu)$ exactly, but yields better generated images [11, 7].

Wardrop considers congested transport in a discrete setting [12]. Carlier, Jimenez, Santambrogio extended to continuous version [2].

Classical optimal transport only considers starting points and ending points. For congested transport, we will consider all possible paths and amount of traffic on a given path.

Definition: 4.1: Traffic Plans

Let $\mathcal{C}$ be the space of absolutely continuous curves $\omega : [0, 1] \to \Omega$ for Ω compact in $\mathbb{R}^d$ with non-empty interiors. Consider the probability measures $\mathcal{Q}$ on $\mathcal{C}$, compatible with μ, ν, analogous to requiring $\gamma \in \Pi(\mu, \nu)$ in optimal transport. Call $\mathcal{Q}$ the *traffic plans*.
Define $e_t : \mathcal{C} \to \Omega$ s.t. $e_t(\omega) = \omega(t)$. $\mu_t = (e_t)_\# \mathcal{Q}$ is a probability flow s.t. $\mu_0 = (e_0)_\# \mathcal{Q}$, $\mu_1 = (e_1)_\# \mathcal{Q}$.

$$\int \phi d\mu_t = \int \phi(e_t(\omega)) d\mathcal{Q}(\omega)$$

Let $\mathcal{T}(\mu, \nu)$ denote the traffic plans satisfying $\mu_0 = \mu, \mu_1 = \nu$.

Definition: 4.2: Traffic Intensity

Define the traffic intensity, a measure $i_\mathcal{Q}$ on Ω by

$$\int_\Omega \phi d i_\mathcal{Q} = \int_\mathcal{C} \int_\omega \phi ds d\mathcal{Q}(\omega),$$
$$\text{where } \int_\omega \phi ds = \int_0^1 \phi(\omega(t)) |\omega'(t)| dt$$

RHS is the line integral on ω of ϕ, averaged over all curves $\omega \in \mathcal{C}$. Intuitively, if $A \subset \Omega$ and $\phi = \chi_A$, then $i_\mathcal{Q}(A)$ =mass of all traffic through A.

Lemma: 4.1: Calier, Jimenez, Santambrogio

Given $\mathcal{Q} \in \mathcal{T}(\mu, \nu)$, there is $\gamma \in \Pi(\mu, \nu)$ and for γ-a.e. (x, y), a probability measure $\mathcal{Q}^{x,y}$ on curves $\mathcal{C}$ supported on $\mathcal{C}^{x,y} = \{\omega \in \mathcal{C} : \omega(0) = x, \omega(1) = y\}$, *i.e.* $\mathcal{Q}^{x,y}(\mathcal{C}^{x,y}) = 1$, s.t.

$$\int_\mathcal{C} \phi(\omega) d\mathcal{Q}(\omega) = \int_{\Omega \times \Omega} \left(\int_{\mathcal{C}^{x,y}} \phi(\omega) d\mathcal{Q}^{x,y} \right) d\gamma(x, y)$$

This comes from disintegration of measures: we can split integration over $\mathcal{C}$ into integration over all curves from specific starting point x and ending point y, and then integrate over all x and y.

Definition: 4.3: Congested Transport Problem

Let $H(x,z)$ be a congestion cost function.
Let $\mathcal{T}^p(\mu,\nu) = \{\mathcal{Q} \in \mathcal{T}(\mu,\nu) : i_\mathcal{Q} = i_\mathcal{Q}(x)dx, i_\mathcal{Q}(x) \in L^p(\Omega)\}$, *i.e.* $i_\mathcal{Q} \ll \mathcal{L}$, and $i_\mathcal{Q}(x)$ is the traffic intensity through a single point x.
The congested transport cost for a given traffic plan $\mathcal{Q} \in \mathcal{T}^p(\mu,\nu)$ is $\int_\Omega H(x, i_\mathcal{Q}(x))dx$. The congested transport problem (CTP) is

$$\inf_{\mathcal{Q}\in\mathcal{T}^p(\mu,\nu)} \int_\Omega H(x, i_\mathcal{Q}(x))dx \tag{5}$$

Assumptions on H:

1. H is a Caratheodory function:
   (a) $x \mapsto H(x,z)$ is measurable in x for all z
   (b) $z \mapsto H(x,z)$ is continuous in z for a.e. x
2. $c(z^p - 1) \le H(x,z) \le c(z^p + 1)$, *i.e.* $H(x,z)$ grows like z^p for large z.

Example: $H(x,z) = \frac{1}{p}z^p + \lambda z$

1. $\frac{\partial H}{\partial z}(i_\mathcal{Q}) = i_\mathcal{Q}^{p-1} + \lambda$ is the incremental cost for large traffic intensity
2. $\frac{\partial H}{\partial z}(0) = \lambda > 0$ implies we cannot travel at infinite speed even if road is empty

Definition: 4.4: Wardrop Equilibrium

Given a traffic plan $\mathcal{Q}$, let

$$L_\mathcal{Q}(\omega) = \int_0^1 \frac{\partial H}{\partial z}(\omega(t), i_\mathcal{Q}(\omega(t)))|\omega'(t)|dt$$

This is the length of a curve w.r.t. a conformal Riemannian metric determined by $\mathcal{Q}$ and H. Define the Riemannian metric:

$$d_\mathcal{Q}(x,y) = \inf_{\omega\in\mathcal{C}^{x,y}} L_\mathcal{Q}(\omega),$$

$\omega(t)$ is a geodesic if $L_\mathcal{Q}(\omega) = d(\omega(0), \omega(1))$.
$\mathcal{Q}$ is called a Wardrop equilibrium if $\mathcal{Q}$ is supported on geodesics for $d_\mathcal{Q}$.

Theorem: 4.1: Calier, Jimenez, Santambrogio

1. $\mathcal{Q}$ solves Eq. 5 if and only if it is a Wardrop equilibrium, and $\gamma = (e_0, e_1)_\# \mathcal{Q}$ is a solution of the optimal transport problem $\inf_{\gamma\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} d_\mathcal{Q}(x,y)d\gamma(x,y)$. Congested cost defines a new distance measure for the optimal transport problem.
2. If H satisfies all assumptions, there exists a solution $\mathcal{Q} \in \mathcal{T}^p(\mu,\nu)$.

Let H^1 denote the Sobolev spaces s.t.

$$H^1(\Omega) = \left\{ u : u \in L^2(\Omega), \int |\nabla u|^2 < \infty \right\}$$
$$H^1(\Omega, \sigma) = \left\{ u : \int (|u|^2 + |\nabla u|^2)\sigma dx < \infty \right\}$$

Define:

$$\text{GP}_\lambda = \sup_{u \in H^1(\Omega)} \int u(d\mu - d\nu) - \frac{\lambda}{2} \int_\Omega (|\nabla u| - 1)_+^2 \sigma(\hat{x}) d\hat{x}$$
$$\tilde{\text{GP}}_\lambda = \sup_{u \in H^1(\Omega, \sigma)} \int u(d\mu - d\nu) - \frac{\lambda}{2} \int_\Omega (|\nabla u| - 1)_+^2 \sigma(\hat{x}) d\hat{x}$$

To integrate w.r.t. $\sigma(\hat{x})$, define $\pi(t, x, y) = (1-t)x + ty$,

$$\int \phi(x) d\sigma(x) = \iiint \phi((1-t)x + ty) dt d\mu(x) d\nu(y) = \iiint \pi(t, x, y) dt d\mu d\nu,$$

so $\sigma = \pi_\#(U[0,1], \mu, \nu)$.

***Proposition:* 4.1:**

If $\mu \ll \mathcal{L}^d$, $\nu \ll \mathcal{L}^d$, then $\sigma \ll \mathcal{L}^d$, $d\sigma = \sigma(x)dx$.

***Theorem:* 4.2:**

$$\text{GP}_\lambda = \text{CTP with } H_\lambda(x, z) = \begin{cases} \frac{1}{2\lambda\sigma(x)} z^2 + z, \text{ if } \sigma > 0 \\ \infty, \sigma = 0, z \neq 0 \\ 0, \sigma = z = 0 \end{cases}$$

Also, $\exists C > 0$ s.t. $\forall \lambda > 0$,

$$\sup(\text{GP}_\lambda) \geq W_1(\mu, \nu) \left(1 + \frac{C}{\lambda} W_1(\mu, \nu) \right)$$

Here σ acts as a speed limit.

Question: What if $\sigma(x)$ line segments cross? [5]

Example: Let $\Omega = [-2, 2]^2 \subset \mathbb{R}^2$, take μ, ν as uniform distributions on rectangles on $[0, 0.1] \times [-2, 2]$ and $[1.9, 2] \times [-2, 2]$. Take $T_0 : \mathbb{R}^2 \to \mathbb{R}^2$ with $T_0(x_1, x_2) = (x_1 + 1.9, x_2)$ an optimal transport map. $(T_0)_\# \mu = \nu$. The Kantorovich potential is $u^*(x_1, x_2) = -x_1$.

$$\int |x - T_0 x| d\mu = 1.9$$
$$\int (-x) d\mu - \int (-x) d\nu = -\frac{0.1}{2} + \frac{1.9 + 2}{2} = 1.9$$

The points sampled from σ peaks at $(1, 0)$. Optimal traffic flow concentrate mass on $(1, 0)$. Note that in $H(x, z)$, cost is low when $\sigma(x)$ is high. If WGAN-GP computes W_1 explicitly, then critic should be $u(x, y) = -x$. The level sets of u will be all verticle and transport paths will be horizontal. In reality, all transport paths bend towards where $\sigma(x)$ is large.

5 Additional Topics

5.1 Entropic Regularization

This follows Chapter 4 of [8].

***Definition:* 5.1: Discrete Optimal Transport**

Let $\mu = \sum_{i=1}^{N} a_i \delta_{x_i}, \nu = \sum_{j=1}^{N} b_j \delta_{y_j}$ be point mass and $\sum a_i = \sum b_j = 1$. Let cost c_{ij} be $|x_i - y_j|$ or $|x_i - y_j|^2$. Consider a transport plan $\gamma_{ij} \geq 0$

$$\min_{\gamma} \left\{ \sum_{ij} c_{ij}\gamma_{ij} : \gamma_{ij} \geq 0, \sum_{i} \gamma_{ij} = b_j, \sum_{j} \gamma_{ij} = a_i \right\} \tag{6}$$

This is a linear programming problem, but simplex method is expensive.

***Definition:* 5.2: Entropic Regularization**

Pick $\epsilon > 0$, solve for

$$\gamma^{\epsilon} = \arg\min_{\gamma} \left\{ \sum_{ij} c_{ij}\gamma_{ij} + \epsilon\gamma_{ij}\log\gamma_{ij} : \sum_{i} \gamma_{ij} = b_j, \sum_{j} \gamma_{ij} = a_i \right\} \tag{7}$$

$\sum_{ij} \gamma_{ij}\log\gamma_{ij}$ is strictly convex, so the problem is strictly convex.
As $\epsilon \to \infty$, $\gamma^{\epsilon}_{ij} = a_i b_j$. We try to maximize entropy by pairing every x_i with every y_j. As $\epsilon \to 0$, $\gamma^{\epsilon} \to$ optimal plan for Eq. 6 of maximal entropy.

Rewrite $c_{ij}\gamma_{ij} + \epsilon\gamma_{ij}\log\gamma_{ij} = \epsilon\gamma_{ij}\log\left(\frac{\gamma_{ij}}{\eta_{ij}}\right)$, where $\eta_{ij} = \exp\left(-\frac{c_{ij}}{\epsilon}\right)$. Then

$$\gamma^{\epsilon} = \arg\min_{\gamma} \left\{ D_{KL}(\gamma, \eta) : \sum_{i} \gamma_{ij} = b_j, \sum_{j} \gamma_{ij} = a_i \right\},$$

where η is fixed and γ is unknown.

Let $C^b = \left\{ \sum_{i} \gamma_{ij} = b_j \right\}, C^a = \left\{ \sum_{j} \gamma_{ij} = a_i \right\}$, γ is a (KL) projection to $C^a \cap C^b$.

Idea: Iteratively project onto C^a and C^b until convergence. Projection is cheap to compute.
This alternate projection is called Kaczmarz algorithm and does converge for KL.

Projection of $\overline{\gamma}$ on C^a is

$$\min \left\{ \sum \gamma_{ij} \log \frac{\gamma_{ij}}{\overline{\gamma}_{ij}} : \sum_{j} \gamma_{ij} = a_i \right\}$$

This is a separate problem for each i:

$$\min \left\{ x_j \log \frac{x_j}{\overline{x}_j} : \sum x_j = a \right\}$$

Lagrange multiplier gives

$$\log x_j - 1 - \log \overline{x}_j + \lambda = 0 \Rightarrow \log x_j = \log \overline{x}_j + \text{const},$$

$x_j = p\overline{x}_j$ for some p constant for any j.

Solution is $\gamma_{ij} = p_i \overline{\gamma}_{ij}$, $\sum_j \gamma_{ij} = \sum_j p_i \overline{\gamma}_{ij} = a_i$, so $p_i = \frac{a_i}{\sum_j \overline{\gamma}_{ij}}$.

Sinkhorn Algorithm:

$$\gamma^0 = \eta, \gamma_{ij}^{2k+1} = \frac{a_i}{\sum_j \gamma_{ij}^{2k}} \gamma_{ij}^{2k}, \gamma_{ij}^{2k+2} = \frac{b_j}{\sum_i \gamma_{ij}^{2k+1}} \gamma_{ij}^{2k+1}$$

It is fast, simple, parallelizable and preserves positivity constraints with log. However, it deos not give a transport map. The convergence deteriorates for small ϵ, so it has limited accuracy.

5.2 High Dimensional Banach Space Theory and Adversarial Examples

At the end of Section 4, we show that the transport paths will cluster around the center. However, in high dimensions, the chance of transport paths crossing is lower.

Statistically, it is cheaper (in Wasserstein distance) to go from a sample to average than from samples to samples [11]. If Wasseerstein distance is computed correctly, it approaches the average. However, in congested tranport, it is equivalent to pushing all traffic flows through the same point, which is penalized.

Consider n-dimensional Banach space, *e.g.* $\mathbb{R}^n$. An image can be represented as a point in the unit cube $[0,1]^n$. For simplicity, consider unit sphere S^{n-1}.

Let μ_1 be the area measure normalized to 1, H the hemisphere. For any $A \subset S^{n-1}$, let

$$A(\epsilon, d) = \left\{x \in S^{n-1} : d(x,y) \leq \epsilon \text{ for some } y \in A\right\}$$

Essentially, $A(\epsilon, d)$ extend A by a distance ϵ. Assume that d is the standard geodesic distance on S^{n-1}.

Theorem: 5.1: Isoperimetric Inequality

For any $n > 2$. If $\mu_1(A) > \frac{1}{2}$, then $\mu_1(A_\epsilon) \geq \mu_1(H_\epsilon)$. *i.e.* Hemispheres are the least expanded.

Theorem: 5.2: Milman-Schechtman 1986

After ϵ-expansion,

$$\mu_1(H_\epsilon) \geq 1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$

As $n \to \infty$, H_ϵ tends to full measure.

Let $\mathcal{C}$ be a classifier function, $\mathcal{C} : S^{n-1} \to \{1, 2, ..., m\}$. $\mathcal{C}$ partitions S^{n-1} into m disjoint measurable sets $C_j = \{x : C(x) = j\}, 1 \leq j \leq m$. x admits an ϵ-adversarial example if $\exists \hat{x}$ with $d(x, \hat{x}) \leq \epsilon$ and $C(\hat{x}) \neq C(x)$. Safe points in C_c should be away from the boundary.

Theorem: 5.3: Existence of Adversarial Examples

Assume the m classes are distributed over $S^{n-1} \subset \mathbb{R}^n$ with density functions $\{\rho_j\}_{j=1}^m$ s.t. $\rho_j = \rho_j(x) d\mu_1$. Define $V_c = \|\rho_c\|_{L^\infty}$. Suppose c is a class with $\mu_1\{x : C(x) = c\} \leq \frac{1}{2}$. Sample x from ρ_c. Then with probability $\geq 1 - V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$, x admits an ϵ-adversarial example.

Proof. Let C_c be the points in class c. $C_c = \{x : C(x) = c\}$. Let $A = \overline{C_c} = \{x : x \notin C_c\}$.

By assumption $\mu_1(A) \geq \frac{1}{2}$. Then by Theorem 5.1, $\mu_1(A_\epsilon) \geq \mu_1(H_\epsilon)$.

By Theorem 5.2, $\mu_1(A_\epsilon) \geq \mu_1(H_\epsilon) \geq 1 - \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$.

Safe points in C_c are in $\overline{A_\epsilon}$ (more than ϵ away from the boundary).

$$\mu_1(\overline{A_\epsilon}) \leq \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$
$$\Rightarrow \rho_c(\overline{A_\epsilon}) \leq V_c \mu_1(\overline{A_\epsilon}) = V_c \left(\frac{\pi}{8}\right)^{\frac{1}{2}} \exp\left(-\frac{n-1}{2}\epsilon^2\right)$$

Close to 0 for large n. □

5.3 Score-based Diffusion Mode

This section partially explains the paper [13].

In score-based diffusion model. Let $p(t, x)$ be the density function at time t. The forward process is a Fokker Planck equation

$$\partial_t p + \nabla \cdot (pf) - \frac{g^2(t)}{2}\Delta p = 0, \qquad p(0, x) = \pi(x),$$

where $\pi(x)$ is the probability density we wish to determine, of which we have samples (training data).

The reverse/generating process is

$$\partial_t \rho + \nabla \cdot (\rho(f - g^2 \nabla \log p(T - t, \cdot))) = \frac{g^2(t)}{2}\Delta\rho, \qquad \rho(0, \cdot) = \rho_0$$

Special case: $f = 0$, $g = \beta$:

$$\partial_t p = \frac{\beta^2}{2}\Delta p, \qquad p(0, \cdot) = \pi(x)$$
$$\partial_t \rho + \beta^2 \nabla \cdot (\rho \nabla \log p(T - t, \cdot)) = \frac{\beta^2}{2}\Delta\rho, \qquad \rho(0, \cdot) = \rho_0$$

We will show that $\pi = \rho(T, \cdot)$.

We interpret the problem as a solution of regularized Wasserstein proximal problem.

Let $V(\rho) = \int V(x)\rho(x)dx$ be a functional. Specifically, $V(\rho) = \beta^2 H(\rho, \pi)$, where $H(\rho, \pi)$ is the cross entropy:

$$H(\rho, \pi) = \mathbb{E}_\rho(-\log \pi) = -\int \log \pi(x)\rho(x)dx.$$

Recall the KL-divergence:

$$KL(\rho||\pi) = \int \rho \log \frac{\rho}{\pi} = \int \rho \log \rho - \int \rho \log \pi$$
$$H(\rho, \pi) = KL(\rho||\pi) - \int \rho \log \rho = KL(\rho||\pi) + \mathcal{E}(\rho),$$

where $\mathcal{E}(\rho) = -\int \rho \log \rho$ is the entropy. Therefore, $V(x) = -\beta^2 \log \pi(x)$.

The Wasserstein proximal is

$$\rho_T = \arg\min_{\tilde{\rho}} V(\tilde{\rho}) + \frac{W_2^2(\rho_0, \tilde{\rho})}{2T}$$

The Benamou-Brenier formula (Theorem 3.11) gives:

$$\frac{W_2^2(\rho_0, \tilde{\rho})}{2T} = \inf_{\rho \geq 0, v} \left\{ \int_0^T \int_\Omega \frac{1}{2} |v(t,x)|^2 \rho(t,x) dx dt : \partial_t \rho + \nabla \cdot (\rho v) = 0, \rho(0,x) = \rho_0, \rho(T,x) = \tilde{\rho} \right\}.$$

Regularized Wasserstein proximal replaced the continuity equation with $\partial_t \rho + \nabla \cdot (\rho v) = \epsilon \Delta \rho$, and the regularized proximal problem becomes

$$\rho_T = \arg\min_{\rho \geq 0, v, \tilde{\rho}} \left\{ \int_\Omega V(x)\tilde{\rho}(x) dx + \int_0^T \int_\Omega \frac{1}{2} |v|^2 \rho : \partial_t \rho + \nabla \cdot (\rho v) = \epsilon \Delta \rho, \rho(0,x) = \rho_0, \rho(T,x) = \tilde{\rho} \right\}$$

Toy problem of Lagrange multiplier: suppose we want to minimize $f : \mathbb{R}^N \to \mathbb{R}$, subject to k constraints $g(x) = 0$,$g : \mathbb{R}^N \to \mathbb{R}^k$. Lagrange multipliers give

$$\min_{x \in \mathbb{R}^N} \sup_{\lambda \in \mathbb{R}^k} f(x) + \lambda \cdot g(x) = \min_{x \in \mathbb{R}^N, g(x) = 0} f(x),$$

because $\sup_{\lambda \in \mathbb{R}^k} \lambda \cdot g(x) = \begin{cases} 0, g(x) = 0 \\ \infty, \text{otherwise} \end{cases}$. Define the Lagrangian $L(x, \lambda) = f(x) + \lambda \cdot g(x)$. We look for a saddle point of L.

For the regularized Wasserstein proximal operator, define a Lagrange multiplier $\phi(t,x)$, the Lagrangian is:

$$\begin{aligned} L(\rho, \tilde{\rho}, v, \phi) &= \int_\Omega V(x)\tilde{\rho}(x) dx + \frac{1}{2} \int_0^T \int_\Omega |v|^2 \rho dx dt + \int_0^T \int_\Omega \phi(t,x)(\partial_t \rho + \nabla \cdot (\rho v) - \epsilon \Delta \rho) dx dt \\ &\text{Apply IBP on } t \\ &= \int_\Omega V(x)\tilde{\rho}(x) dx + \frac{1}{2} \int_0^T \int_\Omega |v|^2 \rho dx dt + \int_\Omega \phi(T,x)\rho(T,x) dx - \int_\Omega \phi(0,x)\rho(0,x) dx \\ &- \iint \rho(\partial_t \phi + \nabla \phi \cdot v - \epsilon \Delta \phi) dx dt \end{aligned}$$

Recall that the continuity equation is interpreted in the weak sense, and we apply divergence theorem and Green's theorem to reach the equality.

The optimality condition gives:

$$\begin{aligned} &\frac{\delta L}{\delta \tilde{\rho}} = 0 \Rightarrow V(x) + \phi(T,x) = 0 \\ &\frac{\delta L}{\delta v} = 0 \Rightarrow v(x) + \nabla \phi(x) = 0 \\ &\frac{\delta L}{\delta \rho} = 0 \Rightarrow \frac{1}{2}|v|^2 - (\partial_t \phi + \nabla \phi \cdot v - \epsilon \Delta \phi) = 0 \end{aligned}$$

Also we have the constraint: $\partial_t \rho + \nabla \cdot (\rho v) - \epsilon \Delta \rho = 0$

From $\frac{\delta L}{\delta v}$ and $\frac{\delta L}{\delta \rho}$, we get:

$$\Rightarrow \frac{1}{2}|v|^2 - \partial_t \phi - |\nabla \phi|^2 + \epsilon \Delta \phi = \partial_t \phi + \frac{1}{2}|\nabla \phi|^2 + \epsilon \Delta \phi = 0$$

From $\frac{\delta L}{\delta v}$ and the constraint, we get:

$$\partial_t \rho - \nabla \cdot (\rho \nabla \phi) - \epsilon \Delta \rho = 0$$

The following system gives $\rho_T = \rho(T, \cdot)$.

$$\begin{aligned}
&\partial_t \phi + \frac{1}{2}|\nabla \phi|^2 + \epsilon \Delta \phi = 0 \\
&\partial_t \rho - \nabla \cdot (\rho \nabla \phi) - \epsilon \Delta \phi = 0 \\
&\rho(0, \cdot) = \rho_0 \\
&\phi(T, x) = -V(x) = \beta^2 \log \pi(x)
\end{aligned}$$

It is a system of forward (ρ) and backward (ϕ) equations. Let $u(t, x) = \phi(T - t, x)$ with $u(0, x) = \beta^2 \log \pi$. We change the backward equation to

$$-\partial_t u + \frac{1}{2}|\nabla u|^2 + \epsilon \Delta u = 0$$

Cole-Hopf Transformation:
Let $p = h(u)$, $h : \mathbb{R} \to \mathbb{R}$ to be determined,

$$\frac{\partial p}{\partial t} = h'(u)\frac{\partial u}{\partial t}; \quad \frac{\partial p}{\partial x_j} = h'(u)\frac{\partial u}{\partial x_j}; \quad \frac{\partial^2 p}{\partial x_j^2} = h''(u)\left(\frac{\partial u}{\partial x_j}\right)^2 + h'(u)\frac{\partial^2 u}{\partial x_j^2}.$$

Then

$$\begin{aligned}
\Delta p &= h''(u)|\nabla u|^2 + h'(u)\Delta u \\
\epsilon \Delta p - \epsilon h''(u)|\nabla u|^2 &= \epsilon h'(u)\Delta u \\
\frac{\partial p}{\partial t} &= h'(u)\frac{\partial u}{\partial t} \\
&= h'(u)\left(\frac{1}{2}|\nabla u|^2 + \epsilon \Delta u\right) \\
&= \epsilon \Delta p + \left(-\epsilon h''(u) + \frac{1}{2}h'(u)\right)|\nabla u|^2
\end{aligned}$$

Choose h so that $\epsilon h'' = \frac{1}{2}h'$, $p = h(u) = \exp\left(\frac{1}{2\epsilon}u\right)$ is the Cole-Hopf transform.

Then $\frac{\partial p}{\partial t} = \epsilon \Delta p$. This converts the original HJB equation to a heat equation.

If $\epsilon = \frac{\beta^2}{2}$, $p(0, x) = \exp\left(\frac{1}{2\epsilon}u(0, \cdot)\right) = \exp\left(\frac{1}{\beta^2}u(0, \cdot)\right) = \exp\left(\frac{\beta^2}{\beta^2}\log \pi\right) = \pi$.
Also, $\log p = \frac{1}{2\epsilon}u$, so $u = \beta^2 \log p$.

$$\phi(t, x) = u(T - t, x) = \beta^2 \log p(T - t, \cdot)$$

The equation for ρ becomes:

$$\partial_t \rho - \beta^2 \nabla \cdot (\rho \nabla \log p(T - t, \cdot)) = \frac{\beta^2}{2}\Delta \rho$$

This converts the problem into forward/backward heat equation which has a solution with Green's kernel.

References

[1] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008. ISBN: 9783764387228. URL: https://books.google.ca/books?id=rCDK9JA5BAEC.

[2] Guillaume Carlier, Cristian Jimenez, and Filippo Santambrogio. "Optimal Transportation with Traffic Congestion and Wardrop Equilibria". In: *SIAM J. Control. Optim.* 47 (2006), pp. 1330–1350. URL: https://api.semanticscholar.org/CorpusID:809007.

[3] Sandesh Ghimire et al. *Geometry of Score Based Generative Models*. 2023. arXiv: 2302.04411 [cs.LG]. URL: https://arxiv.org/abs/2302.04411.

[4] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG]. URL: https://arxiv.org/abs/1704.00028.

[5] Tristan Milne. "Optimal Transport, Congested Transport, and Wasserstein Generative Adversarial Networks". English. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-08. PhD thesis. 2022, p. 199. ISBN: 9798357551146. URL: http://myaccess.library.utoronto.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fdissertations-theses%2Foptimal-transport-congested-wasserstein%2Fdocview%2F2744607011%2Fse-2%3Faccountid%3D14771.

[6] Tristan Milne, Étienne Bilocq, and Adrian Nachman. *A new method for determining Wasserstein 1 optimal transport maps from Kantorovich potentials, with deep learning applications*. 2022. arXiv: 2211.00820 [math.OC]. URL: https://arxiv.org/abs/2211.00820.

[7] Tristan Milne and Adrian Nachman. *Wasserstein GANs with Gradient Penalty Compute Congested Transport*. 2022. arXiv: 2109.00528 [cs.LG]. URL: https://arxiv.org/abs/2109.00528.

[8] Gabriel Peyre and Marco Cuturi. *Computational Optimal Transport*. 2020. arXiv: 1803.00567 [stat.ML]. URL: https://arxiv.org/abs/1803.00567.

[9] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Vol. 87. Progress in Nonlinear Differential Equations and Their Applications. Springer, 2015. DOI: 10.1007/978-3-319-20828-2.

[10] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG]. URL: https://arxiv.org/abs/2011.13456.

[11] Jan Stanczuk et al. *Wasserstein GANs Work Because They Fail (to Approximate the Wasserstein Distance)*. 2021. arXiv: 2103.01678 [stat.ML]. URL: https://arxiv.org/abs/2103.01678.

[12] J. G. Wardrop. "Some Theoretical Aspects of Road Traffic Research". In: 1952. URL: https://api.semanticscholar.org/CorpusID:110090080.

[13] Benjamin J. Zhang et al. *Wasserstein proximal operators describe score-based generative models and resolve memorization*. 2024. arXiv: 2402.06162 [stat.ML]. URL: https://arxiv.org/abs/2402.06162.