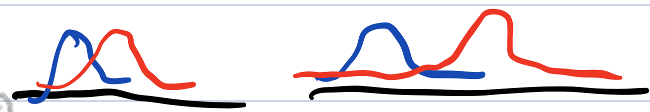
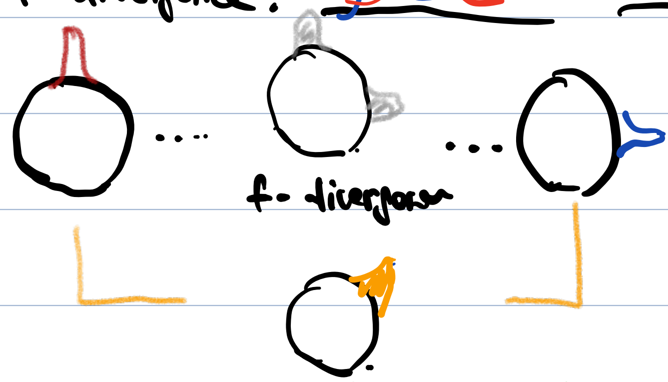
Introduction to Optimal Transport (and Particle Systems)

Motivation f-divergence :





Optimal Transport (Earth Moving distance, Wasserstein Distance)

Monge $\mu \in P(X)$

$\nu \in P(Y)$.

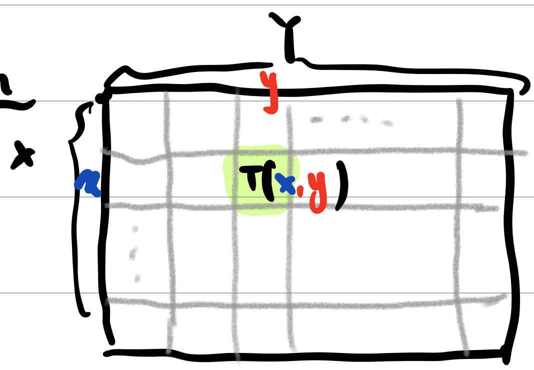
measurable map : $T : X \to Y$

① $\nu(A) = \mu(T^{-1}(A))$

② minimize. $\int_X \underline{c(x, T(x))} \, d\mu(x)$.

c(x, y) is a cost function.

Kantorovich's Problem



T(x, y) : how many products I transport from x to y

$$\min_{T \geq 0} \sum_{x,y} T(x,y)\, c(x,y) \longrightarrow \min_{T \geq 0} \int T(x,y)\, c(x,y)\, dx\, dy$$

s.t. $\sum_y T(x,y) = \mu(x)$ s.t. $\int_y T(x,y)\, dy = \mu(x) \quad \forall x$

$\sum_x T(x,y) = \nu(y)$ $\underbrace{\int_x T(x,y)\, dx = \nu(y)}_{\text{coupling}} \quad \forall y$

$\sum_x \varphi(x) \left[\sum_y T(x,y) - \mu(x) \right]$

Optimal Transport as IPM (Kantorovich Duality)

$$\min_{T\geq 0}\sup_{\varphi,\psi}\Big[\int T(x,y)\,c(x,y)\,dx\,dy + \int_{X\times Y}\varphi(x)\,T(x,y)\,dx\,dy - \boxed{\int_X \varphi(x)\,\mu(x)\,dx} + \int_{X\times Y}\psi(y)\,T(x,y)\,dx\,dy - \boxed{\int_Y \psi(x)\,\nu(y)\,dy}\Big]$$

$$= \sup_{\varphi,\psi}\min_{T\geq 0}\int_{X\times Y} T(x,y)\big[c(x,y)+\varphi(x)+\psi(y)\big]\,dx\,dy - \boxed{\int_X \varphi(x)\,\mu(x)\,dx} - \boxed{\int_Y \psi(y)\,\nu(y)\,dy}\Big]$$

$$= \sup_{\varphi(x)+\psi(y)+c(x,y)\geq 0}\int_X -\varphi(x)\,\mu(x)\,dx - \int_Y \psi(y)\,\nu(y).$$

$$\Leftrightarrow \min_{\underline{\varphi}(x)-\underline{\psi}(y)\geq c(x,y)}\int_X \varphi(x)\,\mu(x)\,dx - \int_Y \psi(y)\,\nu(y)\,dy$$

shadow price

(proof provide at Appendix)

if $c(x,y)=\|x-y\|$.

Optimal Transport cost

$$= \sup_{\underbrace{\varphi(x)-\varphi(y)\leq\|x-y\|}_{\text{the shadow price is Lipschitz.}}}\overbrace{\int_X \varphi\,\mu(x)\,dx - \int_Y \varphi\,\mu(y)\,dy}^{\text{IPM}}$$

[hw for these week]

Optimal Transport Distance between empirical and population can be bounded by the covering number of Lipschitz function

$\Rightarrow W(\hat{\mathbb{P}}_n, \mathbb{P}_n) \propto n^{-1/d}$ "curse of dimensionality"

Shadow price has too much free dim.

§ Gradient flow in Wasserstein Space. $\boxed{\text{advance thing}}$.

What is Gradient flow:

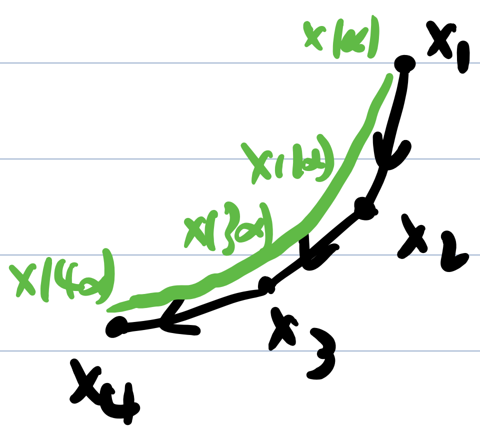
Gradient descent. $x_{t+1} = x_t - \alpha \nabla f(x_t)$

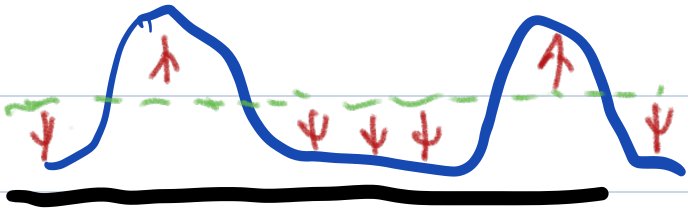
$$\Rightarrow \frac{x_{t+1} - x_t}{\alpha} = -\nabla f(x_t)$$

$x_t \approx x(\alpha t)$

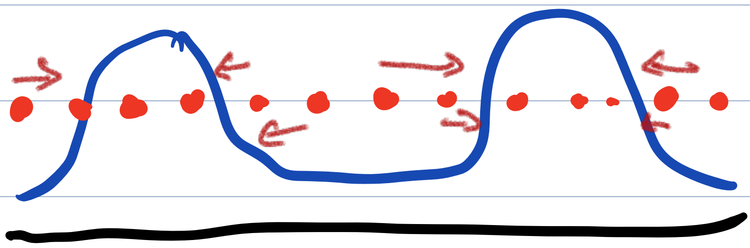
$$\frac{x(\alpha(t+1)) - x(\alpha t)}{\alpha} \approx \left.\frac{dx}{dt}\right|_{\alpha t}$$

Gradient flow: $$\frac{dx(t)}{dt} = -\nabla f(x(t))$$



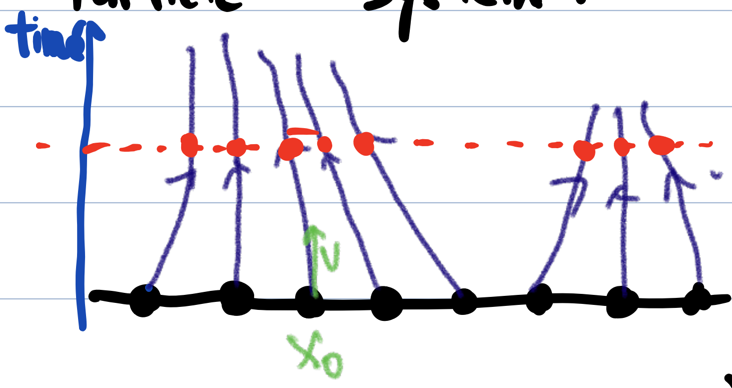


① ↑ ↓ consider in f-diver / as a vector



② Transporting the particles.

Particle System.



assume every particle moves

$$\begin{cases} x(0) = x_0 \\ \dfrac{dx}{dt} = \vec{v}_t(x) \end{cases}$$

x is the position, v is the speed

Question! How does the density move?

$$\partial_t \rho_t = -\nabla \cdot (\vec{v}_t \rho_t)$$

$$\Uparrow$$

$$\nabla \cdot \begin{pmatrix} \vec{v}_x \\ \vec{v}_y \end{pmatrix} = \frac{d\vec{v}_x}{dx} + \frac{d\vec{v}_y}{dy}$$

★ Introduce a test function. ψ

then we keep track of $\int_X \psi(x)\, \rho_t(x)\, dx = E_{\rho_t} \psi$

$$\frac{d}{dt}\int_X \rho_t(x)\psi(x)\,dx$$ <u>write down in particle side!!</u>

$$\parallel$$

$$\int_X \frac{d}{dt}\rho_t(x)\,\psi(x)\,dx$$

$$\frac{d}{dt}\int \psi(X(t))\,\rho_0(x)\,dx$$

we first sample $X(0) \sim \rho_0(x)$

run $\frac{dX(t)}{d(t)} = \vec{v}_t(X(t))$ till times

$$= \int \underbrace{\nabla\psi(X(t))}\,\underbrace{v_t(X_t)}\,\rho_0(X_0)\,dx \,.$$

Chain rule: $\frac{d}{dt}\psi(X(t)) = \nabla\psi(X(t))\underbrace{\frac{dx}{dt}}_{= v_x}$.

$$= -\int \psi\, \nabla\cdot(v\rho)$$

Thus $\frac{d}{dt}\rho_t = -\nabla\cdot(v\rho)$.

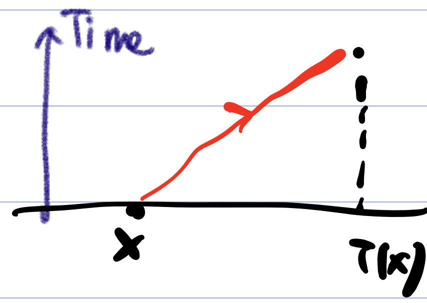
<u>**Benamou-Brenier**</u>. If $c(x,y) = \|x-y\|^2$, then optimal transport can be charaterize as

$$\min. \quad \int_0^1\int_X \|v_t(x)\|^2\rho_t(x)\,dx\,dt.$$

$$\text{s.t.} \quad \partial_t\rho_t + \nabla\cdot(v_t\rho_t) = 0.$$

$$\rho(\cdot,0) = \mu, \qquad \rho(\cdot,1) = \nu$$

$$X(t,x) = x + t\,(T(x) - x),$$



Gradient Descent in Wasserstein Space.

$F(\rho)$ is a function of distribution ρ.

$\frac{dF(\rho)}{d\rho}$ is a function f, actually satisfies.

$$F(\rho + \varepsilon \rho_0) = F(\rho) + \varepsilon \int f \cdot \rho + o(\varepsilon)$$

$\frac{d\rho}{dt} = -\frac{dF(\rho)}{d\rho}$ (If I run the gradient in vector space - not exactly Fisher - Rao Flow)

Wasserstein Gradient Descent:

Particles, $x(t)$.

$$\frac{dx}{dt} = -\nabla \left(\frac{dF(\rho)}{d\rho} \right).$$

If $F(\rho) = \int \psi(x) \rho(x)$ is a linear function of ρ

$$\frac{dF(\rho)}{d\rho} = \psi$$

$$\Rightarrow \frac{dx}{dt} = -\nabla \psi$$

[Appendix] Duality of Optimal Transport when $c(x,y) = \|x-y\|$

If g is 1-Lipschitz: then

$$\underbrace{\inf_x}\underbrace{\{\|x-y\| - g(x)\}} = -g(y)$$

must be 1-lipschitz.

if φ is the shadow price for x, then the shadow price for y is $\psi(y) = \inf_x \{\|x-y\| - \varphi(x)\}$