UC San Diego

Lecture 18
A Geometric View of Optimal Transportation and Generative Model

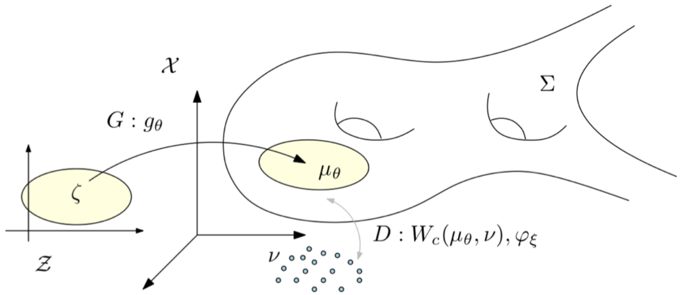
Fangchen Liu, Zhiao Huang

Jan 9, 2018

Generative Model

- Generator
  - Latent Space to image space
  - Minimize the gap between P_θ and P_r

- Discriminator
  - Maximize



$$L(D, g_\theta) = \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{x\sim\mathbb{P}_\theta}[\log(1 - D(x))]$$

Generative Model

- GAN suffered from mode collapse
  - KLD is asymmetric, unbalanced penalty for Generator when P_r -> 0
- Gradient Vanishing:
  - JSD will be a constant when two distributions are “far away”

Generative Model

- GAN suffered from mode collapse
  - KLD is asymmetric, unbalanced penalty for Generator when P_r -> 0
- Gradient Vanishing:
  - JSD will be a constant when two distributions are “far away”
- WGAN
  - Wasserstein Distance is better
    - Almost smooth and differentiable everywhere
    - Better estimation as the distance of distribution
  - Closely related to optimal transport theory

Optimal Transport Theory

- Monge's Formulation of Wasserstein distance

$$W_c(\mu, \nu) = \min_{T: X \to Y} \left\{ \int_X c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\}$$

- T here is a **Measure-preserving Map**
  - Suppose T: X -> Y, as a measure-preserving map

$$\mu(x)dx = \nu(T(x))dT(x)$$

  - We have

$$det(DT(x)) = \frac{\mu(x)}{\nu \circ T(x)}$$

Optimal Transportation Theory

- Kantorovich's Approach
  - If there is a joint measure

$$\rho(A \times Y) = \mu(A), \rho(X \times B) = \nu(B).$$

$$W_c(\mu, \nu) := \min_{\rho} \left\{ \int_{X \times Y} c(x, y) d\rho(x, y) : \pi_{x\#}\rho = \mu, \pi_{y\#}\rho = \nu \right\}$$

  - It is a relaxation of Monge's formulation
    - continuous distribution (\mu is abs. continuous measure on X)
    - And L1, L2 norm is convex. So Monge is OK!

Kantorovich Dual Formulation

- Still far away from the GAN's min-max formulation
- Consider the dual problem of Kantorovich's formulation!
- Primal:

$$
\begin{aligned}
KP(\mu,\nu) = \quad & \min_\gamma \quad \int_{X\times Y} c(x,y)\,d\gamma(x,y) \\
& s.t. \quad \int_Y d\gamma(x,y) = p(x), \quad \int_X d\gamma(x,y) = q(y) \\
& \qquad\;\; \gamma(x,y) \geq 0
\end{aligned}
$$

Kantorovich Dual Formulation

- Still far away from the GAN's min-max formulation
- Consider the dual problem of Kantorovich's formulation!
- Primal:

$$
\begin{aligned}
KP(\mu,\nu) = \quad & \min_\gamma \quad \int_{X\times Y} c(x,y)d\gamma(x,y) \\
& s.t. \quad \int_Y d\gamma(x,y) = p(x), \quad \int_X d\gamma(x,y) = q(y) \\
& \qquad\quad \gamma(x,y) \geq 0
\end{aligned}
$$

- Dual:

$$
\begin{aligned}
DP(\mu,\nu) = \quad & \max_{\phi,\psi} \quad \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) \\
& s.t. \qquad \phi(x) + \psi(y) \leq c(x,y), \quad \forall (x,y) \in X \times Y
\end{aligned}
$$

Kantorovich Dual Formulation (cont.)

$$inf_{\gamma\in\prod(X,Y)}\int_{X\times Y}c(x,y)d\gamma(x,y)$$

$$= inf_{\gamma\in M^{+}(X,Y)}\int_{X\times Y}c(x,y)d\gamma(x,y) + \left\{\begin{array}{c}0, \gamma(x,y)\in\prod(X,Y)\\ \infty, otherwise\end{array}\right.$$

$$= inf_{\gamma\in M^{+}(X,Y)}\int_{X\times Y}c(x,y)d\gamma(x,y) + sup_{\varphi,\psi}\int_{X}\varphi du + \int_{Y}\psi dv - \int_{X\times Y}(\varphi+\psi)d\gamma(x,y)$$

$$= inf_{\gamma\in M^{+}(X,Y)}sup_{\varphi,\psi}\int_{X}\varphi du + \int_{Y}\psi dv + \int_{X\times Y}(c(x,y)-\varphi(x)-\psi(y))d\gamma(x,y)$$

$$inf_{\gamma}sup_{\varphi,\psi}\geq sup_{\varphi,\psi}inf_{\gamma}$$

$$\geq sup_{\varphi,\psi}inf_{\gamma\in M^{+}(X,Y)}\int_{X}\varphi du + \int_{Y}\psi dv + \int_{X\times Y}(c(x,y)-\varphi(x)-\psi(y))d\gamma(x,y)$$

$$= sup_{\varphi,\psi}\int_{X}\varphi du + \int_{Y}\psi dv + inf_{\gamma\in M^{+}(X,Y)}\int_{X\times Y}(c(x,y)-\varphi(x)-\psi(y))d\gamma(x,y)$$

$$= sup_{\varphi,\psi}\int_{X}\varphi du + \int_{Y}\psi dv + \left\{\begin{array}{c}0, c(x,y)\geq\varphi(x)+\psi(y)\\ -\infty, otherwise\end{array}\right.$$

$$= sup_{\varphi,\psi}\int_{X}\varphi du + \int_{Y}\psi dv$$

Kantorovich Dual Formulation (cont.)

- The Dual Problem:

$$W_c(\mu,\nu) := \max_{\varphi,\psi}\left\{\int_X \varphi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) : \varphi(x)+\psi(y) \le c(x,y)\right\}$$

- Define c-Transform: $\varphi^c(y) = \inf_{x\in X}\left(c(x,y)-\varphi(x)\right)$
  - If a function has c-transform, then it is c-concave

- How can we guarantee the equality?

$$W_c(\mu,\nu) := \max_{\varphi}\left\{\int_X \varphi(x)d\mu(x) + \int_Y \varphi^c(y)d\nu(y)\right\}$$

Kantorovich Dual Formulation (cont.)

- The optimality gap between primal and dual

$$inf_{\gamma} sup_{\varphi,\psi} \geq sup_{\varphi,\psi} inf_{\gamma}$$

- Kantorovich proved that, if cost function is bounded by some 1-Lipschitz functions, supremum of the dual is equals to the infimum of primal

Theorem 1.29 (duality). *Let $\mu, \nu \in \mathbb{R}$ and $c\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ continuous and bouned below such that*

$$c(x, y) \leq a(x) + b(y)$$

for some $a \in L^1(\mu)$ and $b \in L^1(\nu)$. Then the minimum of the Kantorovich problem equals the supremum in the dual formulation and this supremum is attained by some couple (φ, φ^{c+}) with φ a c-concave function.

Revisit Wasserstein GAN

- If the L-1 transportation cost is $c(x, y) = |x - y|$
- We have $\varphi^c = -\varphi$
- The objective for Discriminator

$$W_c(\mu, \nu) := \max_{\varphi} \left\{ \int_X \varphi(x) d\mu(x) - \int_Y \varphi(y) d\nu(y) \right\}$$

- Note that the Kantorovich's potential should be 1-Lipsitz, so they do weight clipping

Brenier's Theorem

- What if we use L-2 transportation cost?

Theorem 3.5 (Brenier[5]) *Suppose X and Y are the Euclidean space $\mathbb{R}^n$, and the transportation cost is the quadratic Euclidean distance $c(x,y) = |x-y|^2$. If μ is absolutely continuous and μ and ν have finite second order moments, then there exists a convex function $u : X \to \mathbb{R}$, its gradient map ∇u gives the solution to the Monge's problem, where u is called Brenier's potential. Furthermore, the optimal mass transportation map is unique.*

- Gradient of a convex scalar function: curl free
- Based on the properties of measure-preserving map, we have

$$det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right)(x) = \frac{\mu(x)}{\nu \circ \nabla u(x)}.$$

Brenier's Potential

- Better solutions than compute the Hessian?
  - Yes!
- Consider a point (x_0, y_0) , under the transport map from X -> Y
  - By definition: $\varphi^c(y_0) = \inf_x c(x, y_0) - \varphi(x)$
  - Take the gradient:

$$\nabla\varphi(x_0) = \nabla_x c(x_0, y_0) = \nabla h(x_0 - y_0)$$

  - Here we assume the cost $c(x, y) = h(x - y)$ is strictly convex with h
  - Then we will have

$$y_0 = x_0 - (\nabla h)^{-1}(\nabla\varphi(x_0))$$

Brenier's Potential

- Replace $(x_{0,} y_0)$ with $(x, T(x))$:

When $c(x,y) = \frac{1}{2}|x-y|^2$, we have

$$T(x) = x - \nabla\varphi(x) = \nabla\left(\frac{x^2}{2} - \varphi(x)\right) = \nabla u(x)$$

- Which implies

$$u(x) = \frac{x^2}{2} - \varphi(x)$$

- That's the relationship between Brenier's potential and Kantorovich's potential!

Get Generator Directly

- Optimal discriminator

  ==> Kantorovich's potential

  ==> Brenier's potential

  ==> Optimal Transport Map

  ==> Optimal Generator

Get Generator Directly

- Optimal discriminator
  ==> Kantorovich's potential
  ==> Brenier's potential
  ==> Optimal Transport Map
  ==> Optimal Generator
- No adversarial training
- No mode collapse
  - Everything is derived from the closed-form solution of Wasserstein distance
  - An optimal solution under this measure

How to obtain discriminator?

- If cost is L2, Kantorovich's potential is closely related to Brenier's potential, which is know to be convex.

  ==> Convex Optimization

- Formulation is not clear

$$W_c(\mu, \nu) := \max_{\varphi} \left\{ \int_X \varphi(x) d\mu(x) - \int_Y \varphi(y) d\nu(y) \right\}$$

How to obtain discriminator?

- If cost is L2, Kantorovich's potential is closely related to Brenier's potential, which is know to be convex.

  ==> Convex Optimization
- Formulation is not clear

$$W_c(\mu, \nu) := \max_{\varphi} \left\{ \int_X \varphi(x) d\mu(x) - \int_Y \varphi(y) d\nu(y) \right\}$$

- Get the solution from geometry
  - Magical truth: construct a convex polytope with user prescribed normals and face volumes is equivalent to solve OTM in L2

Semi-discrete Optimal Transportation

- Generator is a mapping from a fixed distribution X to the empirical distribution Y, e.g. the image manifold.
- In practice, the empirical distribution is represented by a set of data $y_1, y_2, \cdots, y_k$

- Dirac measure

$$\nu = \sum_{j=1}^{k} \nu_j \delta(y - y_j)$$

- Total mass

$$\int_\Omega d\mu(x) = \sum_{i=1}^{k} \nu_i$$

Geometric View

- Monge's formulation: $T : X \to Y$
- Metrics: $c(x, y) = \|x - y\|_2$
- Preimage of y_i decompose the space X into cells:

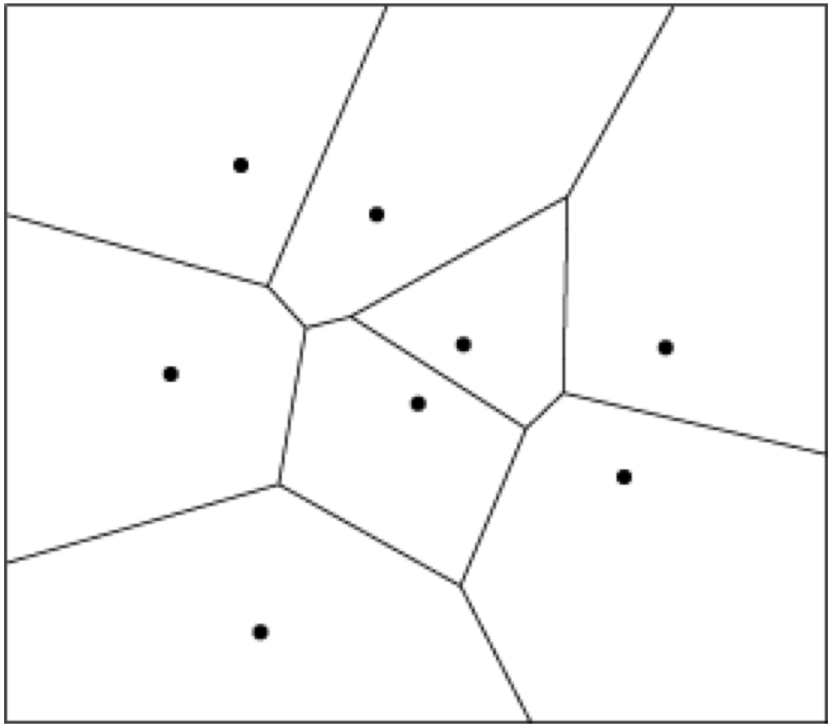
$$W_i = \{x | T(x) = y_i, x \in X\}$$

- Question:

$$\inf \left\{ \int_X c(x, T(x)) d\mu(x) \right\}$$

Voronoi Diagram

- Transport each point to its nearest neighbor!



Voronoi Diagram is not enough

- In transportation problem, we have constrains on the mass received by each point $y_1, y_2, \cdots, y_k$

$$T_*(\mu) = \nu$$

$$\Rightarrow \mu(T^{-1}(y_i)) = \nu(y_i) = \nu_i = \frac{1}{k}$$

- The area (mass) of each cell must be the same.
- The optimal transport map may not be Voronoi Diagram.

Back to Kantorovich's potential

- Define Kantorovich's potential $\psi(y)$

$$\psi_i := \psi(y_i)$$

$$\psi^c(x) = \min_{1 \leq i \leq n} c(x, y_i) - \psi_i$$

- Point x can transport to y when

$$\psi^c(x) + \psi(y) = c(x, y)$$

- The optimal transport must be a Power Diagram!

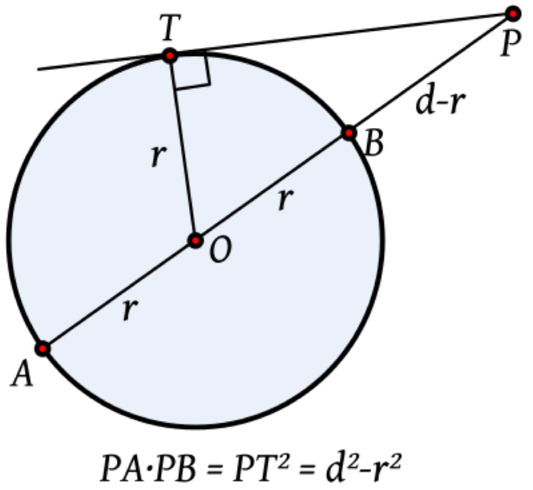
$$pow(x, y_i) = \|x - y\|^2 - \psi_i$$

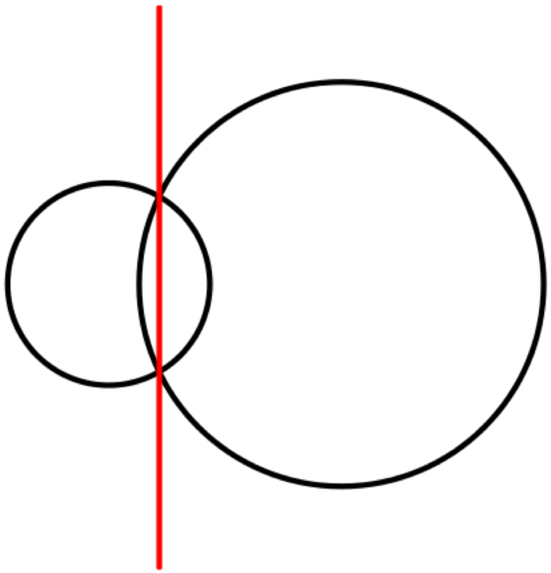
$$W_i(\psi) = \{x \in \mathbf{R}^n | \forall j, pow(x, y_i) \leq pow(x, y_j)\}$$

Power Diagram

- Weighted Voronoi diagram (by the distance to the nearest circle)

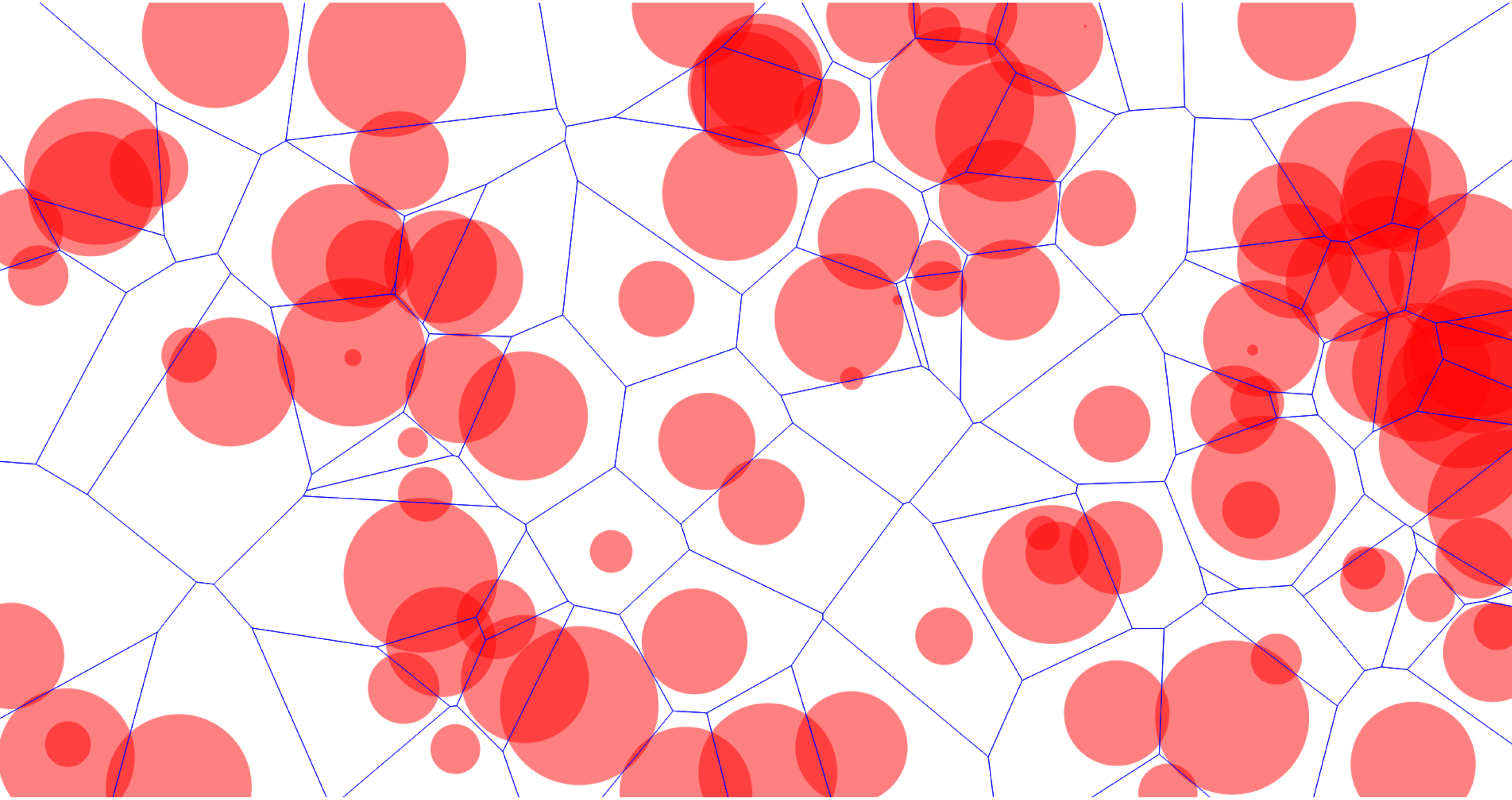
$$pow(x, y_i) = \|x - y\|^2 - \psi_i$$





$$W_i(\psi) = \{x \in \mathbf{R}^n | \forall j, pow(x, y_i) \leq pow(x, y_j)\}$$

Power Diagram



Power Diagram

- Optimal transport map can be seen as a power diagram.
- Given a set of point, how can we find such power diagram?

- First of all, does a diagram like this exist?
  - The set of the points
  - The area of the cells

Hyper-Plane intersection

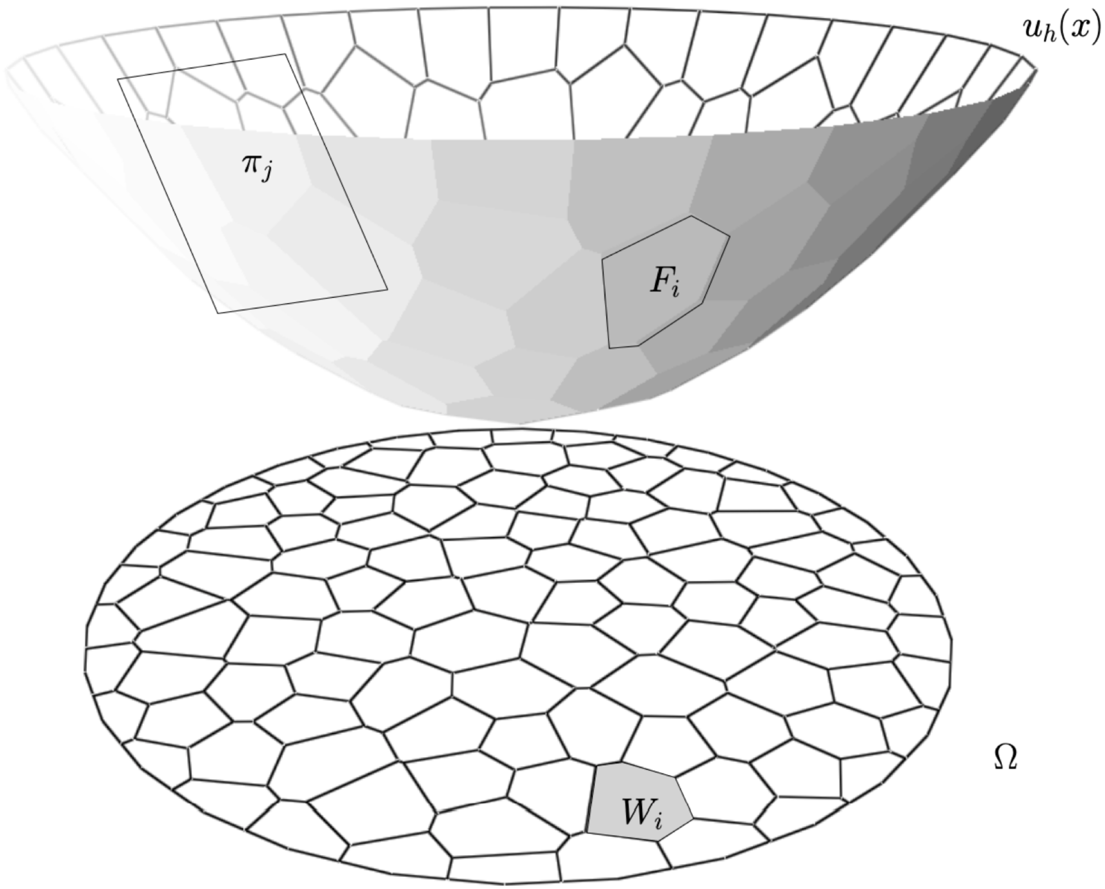
- It's well known that the power diagram is equivalent with hyper-plane intersection

$$pow(x, y_i) \leq pow(x, y_j) \Leftrightarrow \langle x, y_i \rangle + \frac{1}{2}(\psi_i - |y_i|^2) \geq \langle x, y_j \rangle + \frac{1}{2}(\psi_j - |y_j|^2)$$

$$h_i = \frac{1}{2}(\psi_i - |y_i|^2)$$

Normal	y_i
Hyperplane	$\langle x, y_i \rangle + h_i$
Hyperplane intersections	$u_h(x) = \max_i \{\langle x, y_i \rangle + h_i\}$
Power diagram	$W_i(h) = \{x \in \mathbf{R}^n \mid u_h(x) = \langle x, y_i \rangle + h_i\}$

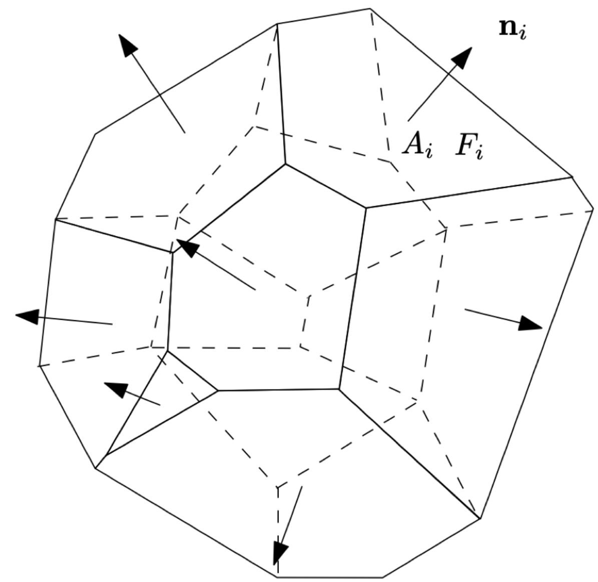
Hyper-Plane intersection



$$W_i(h) = \{x \in \mathbf{R}^n | u_h(x) = \langle x, y_i \rangle + h_i\}$$

Minkowski's theorem

- Minkowski's theorem ensures that the polytope with given normal vectors and face areas exists
- Not useful in our case

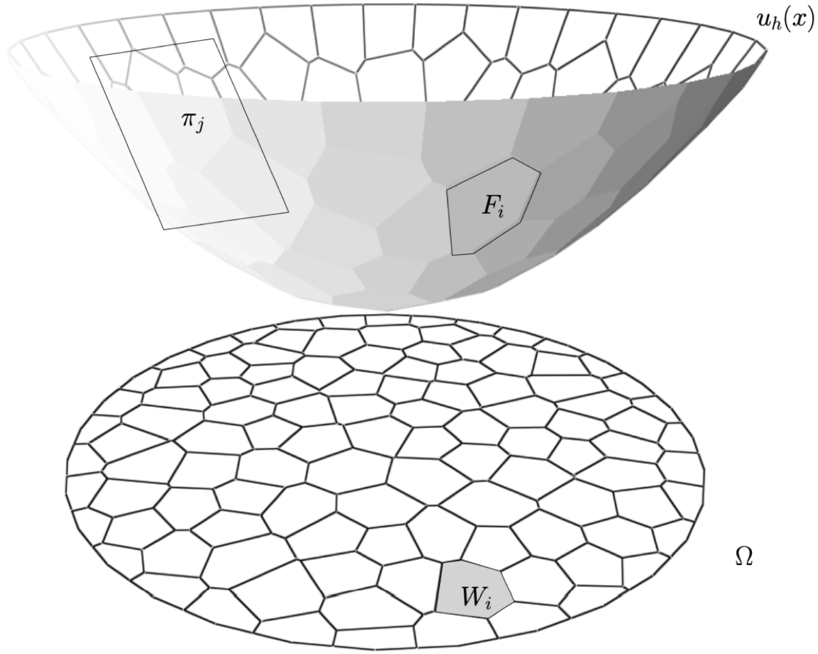


(a) Minkowski theorem

Theorem 4.1 (Minkowski) *Suppose $n_1, ..., n_k$ are unit vectors which span $\mathbb{R}^n$ and $\nu_1, ..., \nu_k > 0$ so that $\sum_{i=1}^{k} \nu_i n_i = 0$. There exists a compact convex polytope $P \subset \mathbb{R}^n$ with exactly k codimension-1 faces $F_1, ..., F_k$ so that n_i is the outward normal vector to F_i and the volume of F_i is ν_i. Furthermore, such P is unique up to parallel translation.*

Alexandrov's theorem

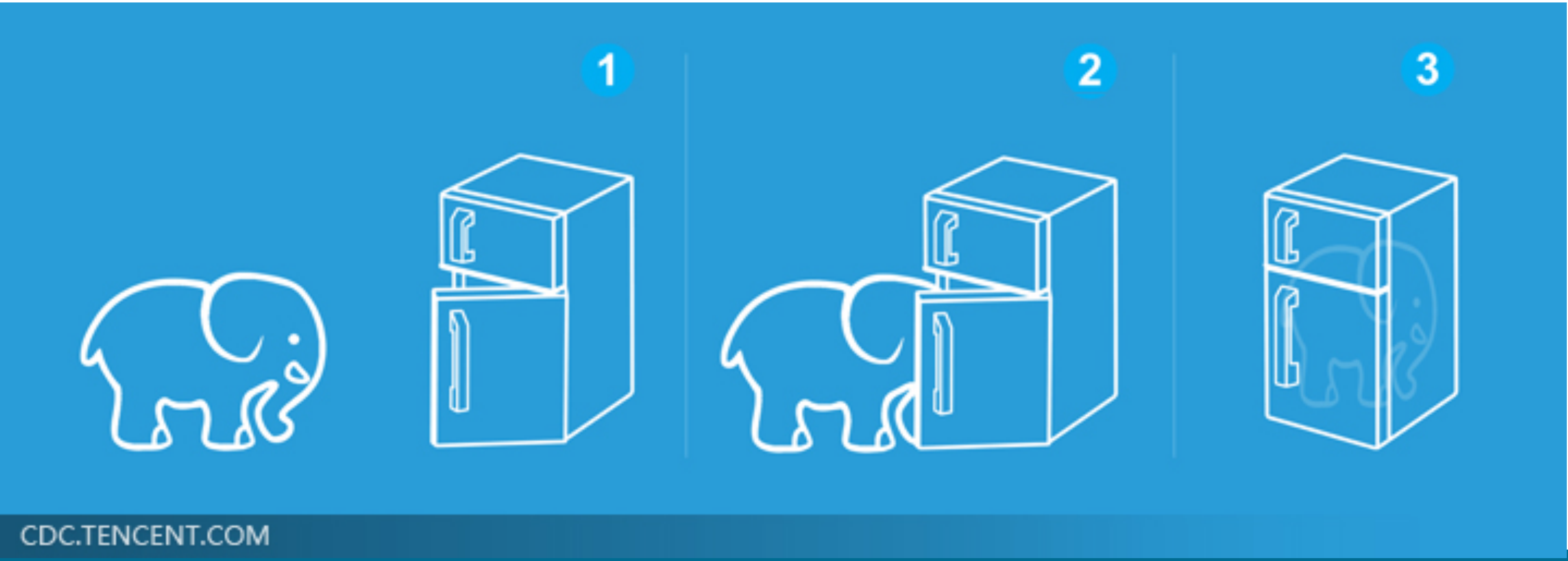
- Exactly what we want!



Theorem 4.2 (Alexandrov[2]) *Suppose Ω is a compact convex polytope with non-empty interior in $\mathbb{R}^n$, $n_1, ..., n_k \subset \mathbb{R}^{n+1}$ are distinct k unit vectors, the $(n+1)$-th coordinates are negative, and $\nu_1, ..., \nu_k > 0$ so that $\sum_{i=1}^{k} \nu_i = vol(\Omega)$. Then there exists convex polytope $P \subset \mathbb{R}^{n+1}$ with exact k codimension-1 faces$F_1, \ldots, F_k$ so that n_i is the normal vector to F_i and the intersection between Ω and the projection of F_i is with volume ν_i. Furthermore, such P is unique up to vertical translation.*

Three steps to find the transport map

- How do you find the optimal transport map?
  1. Suppose you have found the half-plane intersection with Alexandrov’s theorem.
  2. Project the polytope to find the power diagram.
  3. Use power diagram to find the map.



Life can be easier…

- Alexandrov's proof is non-constructive, so we need

Theorem 4.3 (Gu-Luo-Sun-Yau[12]) *Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$ and μ a probability measure on Ω. Then for any $\nu_1, ..., \nu_k > 0$ with $\sum_{i=1}^k \nu_i = \mu(\Omega)$, there exists $h = (h_1, ..., h_k) \in \mathbb{R}^k$, unique up to adding a constant $(c, ..., c)$, so that $w_i(h) = \nu_i$, for all i. The vectors h are exactly maximum points of the concave function*

$$E(h) = \sum_{i=1}^k h_i \nu_i - \int_0^h \sum_{i=1}^k w_i(\eta) d\eta_i \tag{19}$$

on the open convex set

$$H = \{h \in \mathbb{R}^k | w_i(h) > 0, \forall i\}.$$

Furthermore, ∇u_h minimizes the quadratic cost

$$\int_\Omega |x - T(x)|^2 d\mu(x)$$

among all transport maps $T_\# \mu = \nu$, where the Dirac measure $\nu = \sum_{i=1}^k \nu_i \delta_{y_i}$.

Brenier's Potential

- In one word, we can find the polytope by maximizing

$$E(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i$$

- And the gradient of $u_h(x) = \max_i \{\langle x, y_i \rangle + h_i\}$

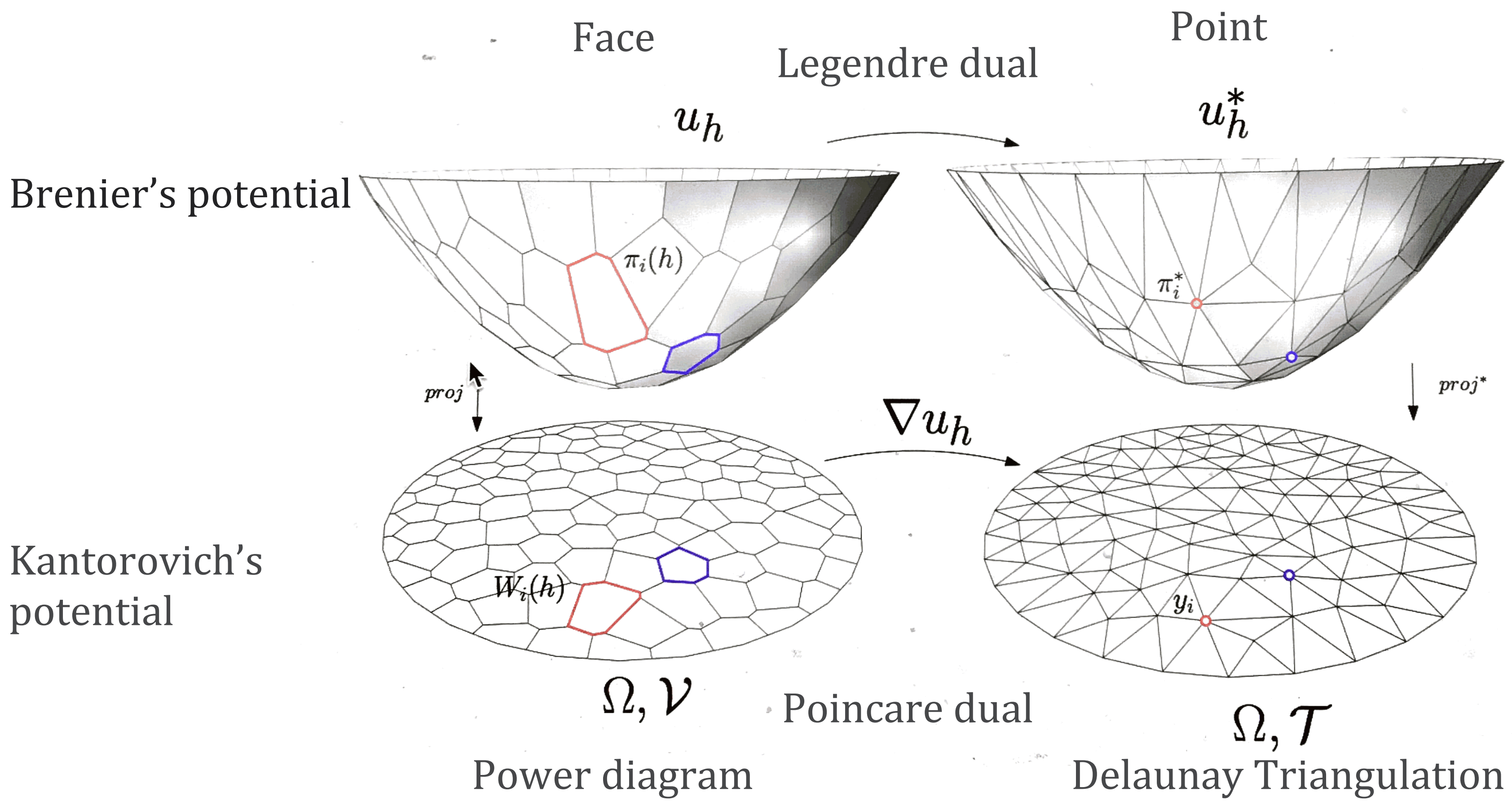
$$\forall x \in W_i(h), \nabla u_h(x) = y_i$$

is the transport map.

- This looks familiar to us

Theorem 3.5 (Brenier[5]) *Suppose X and Y are the Euclidean space $\mathbb{R}^n$, and the transportation cost is the quadratic Euclidean distance $c(x,y) = |x-y|^2$. If μ is absolutely continuous and μ and ν have finite second order moments, then there exists a convex function $u : X \to \mathbb{R}$, its gradient map ∇u gives the solution to the Monge's problem, where u is called Brenier's potential. Furthermore, the optimal mass transportation map is unique.*

Commutative Diagram



Optimization

- Since E is convex, one can find the maximum by convex optimization

$$E(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i$$

$$\nabla E(h) = (\nu_1 - w_1(h), \nu_2 - w_2(h), \cdots, \nu_k - w_k(h))^T$$

Why

- We now have two ways to do optimal transport
- **Kantorovich's approach**
  - find $\psi(y)$ to maximize

$$E_D(\psi) = \int_X \psi^c d\mu + \int_Y \psi d\nu$$

- **Brenier's approach**
  - find $h = (h_1, ..., h_k) \in \mathbb{R}^k$ to maximize

$$E(h) = \sum_{i=1}^{k} h_i \nu_i - \int_0^h \sum_{i=1}^{k} w_i(\eta) d\eta_i$$

The two approaches are equivalent!

Kantorovich's dual approach

$$\psi : Y \to \mathbb{R}, \psi(y_j) = \psi_j \qquad \psi^c(x) = \min_{1\leq j\leq k}\{c(x,y_j) - \psi_j\}$$

$$W_i(\psi) = \{x \in X | c(x,y_i) - \psi_i \leq c(x,y_j) - \psi_j, \forall 1 \leq j \leq k\}$$

- Integrate the potential piece by piece:

$$E_D(\psi) = \int_X \psi^c d\mu + \int_Y \psi d\nu$$

$$= \int_X \min_{1\leq j\leq k}\{c(x,y_j) - \psi_j\} d\mu + \sum_{i=1}^{k} \psi_i \nu_i$$

$$= \sum_{j=1}^{k} \int_{x\in W_j(\psi)} (c(x,y_j) - \psi_j)\, d\mu + \sum_{i=1}^{k} \psi_i \nu_i$$

$$= \sum_{j=1}^{k} \psi_i(\nu_i - w_i(\psi)) + \sum_{j=1}^{k} \int_{W_j(\psi)} c(x,y_j) d\mu$$

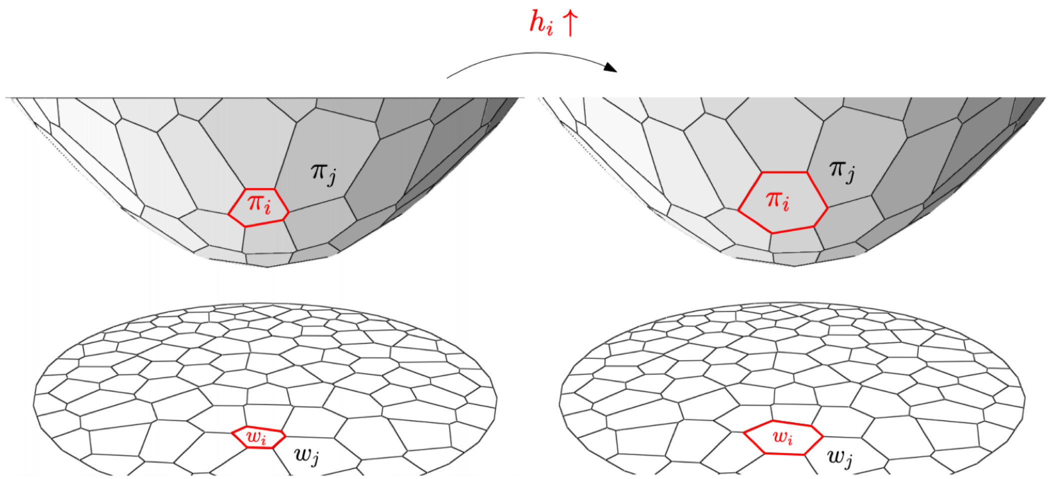
Variational Method

- Transportation cost

$$\mathcal{C}(\psi) = \sum_{j=1}^{k} \int_{W_j(\psi)} c(x, y_j) d\mu.$$

- By variational methods, it's easy to show

$$d\mathcal{C} = \sum_{i=1}^{k} \psi_i dw_i$$



Integration by parts

- Transportation cost

$$dC = \sum_{i=1}^{k} \psi_i dw_i \Rightarrow C(w) = \int^{w} \sum_{i=1}^{k} \psi_i dw_i$$

$$\int^{w} \sum_{i=1}^{k} \psi_i dw_i + \int^{\psi} \sum_{i=1}^{k} w_i d\psi_i = \sum_{i=1}^{k} w_i \psi_i$$

- If $\psi_i = h_i + 1/2|y_i|^2, d\psi_i = dh_i$

$$\int^{h} \sum_{i=1}^{k} w_i dh_i = \int^{\psi} \sum_{i=1}^{k} w_i d\psi_i + const$$

$$\Rightarrow \int^{h} \sum_{i=1}^{k} w_i(\eta) d\eta + \sum_{j=1}^{k} \int_{W_j(\psi)} c(x, y_j) d\mu = \sum_{i=1}^{k} \psi_i w_i(\psi) + const$$

Equivalence

- Put them together

$$E_D(\psi) = \sum_{i=1}^{k} \psi_i(\nu_i - w_i(\psi)) + \sum_{j=1}^{k} \int_{W_j(\psi)} c(x, y_j) d\mu.$$

$$E_B(h) = \sum_{i=1}^{k} h_i \nu_i - \int^{h} \sum_{i=1}^{k} w_i(\eta) d\eta.$$

Lemma 5.2 *Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$. Given μ a probability measure on Ω, $\nu = \sum_{i=1}^{k} \nu_i \delta_{y_i}$, with $\sum_{i=1}^{k} \nu_i = \mu(\Omega)$. If $c(x, y) = 1/2|x - y|^2$, then*

$$h_i = \psi_i - \frac{1}{2}|y_i|^2, \ \forall i$$

and

$$E_D(\psi) - E_B(h) = Const$$

Summary

- A geometric view of semi-discrete optimal transportation.